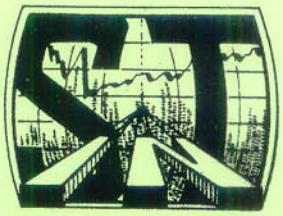


STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 12, Number 1, August 2011

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 1—5

FROM THE EDITOR

A set of fourteen articles composing this Journal's issue is arranged in three major groups: comparative surveys - sampling methods and estimation - other articles. A new section is added, conferences, with information about two important international meetings that are under preparation - one organized by the Central Statistical Office of Poland in cooperation with the World Bank, other by the Polish Statistical Association which celebrates in 2012 its one hundredth anniversary.

The two papers included in the first section address similar aspects of comparative surveys in that they both share the focus on comparison of the proposed class of estimators; however, derived from data collected in different ways.

In ***Subsampling the nonrespondents in cluster sampling on sampling on two successive occasions,*** **H. P. Singh** and **S. Kumar** discuss the problem of estimation of finite population mean for current occasion in the context of cluster sampling on two successive occasions when there is non-response on both the occasions. Estimators for the current occasion are derived as the particular case when there is non-response on first occasion and second occasion only. A comparison between variances of the estimates is studied, using Hansen and Hurwitz (1946) technique for estimation of population mean for current occasion in the context of cluster sampling over sampling on two successive occasions. Three different possible cases when there is non-response (i) on both the occasions, (ii) only on the first occasion, and (iii) only on the second occasion, are discussed. Authors recommend the approach they developed in this study for when there is a need to correct for non-response in cluster sampling over two occasions. They illustrate the performance of the proposed strategy empirically.

Some rotation patterns in two-phase sampling are analyzed by **G. N. Singh** and **S. Prasad,** focusing on the estimation of population mean on the current occasion using two-phase successive (rotation) sampling on two occasions has been considered. Two-phase ratio, regression and chain-type estimators for estimating the population mean on current (second) occasion have been proposed. Properties of the proposed estimators have been studied and their respective optimum replacement policies are discussed. Estimators are compared with the sample mean estimator, when there is no matching and the natural optimum estimator, which is a linear combination of the means of the matched and unmatched portions of the sample on the current occasion. Results are demonstrated through empirical means of comparison and suitable recommendations are made.

The following six papers compose the sampling methods and estimation section.

Paper by **B. B. Khare** and **R. R. Sinha,** ***Estimation of population mean using multi-auxiliary characters with subsampling the nonrespondents,*** is aimed at suggesting a class of two phase sampling estimators for population mean using multi-auxiliary characters in presence of non-response on study character. The expressions for bias, mean square error and the condition for attaining the minimum mean square error of the proposed class of estimators have been obtained, along with the optimum values of the size of first phase sample, second phase sample and the sub sampling fraction of non-responding group. Those have been determined for the fixed cost and for the specified precision. A comparison of the proposed class of estimators has been carried out with an empirical study.

N. A. Montes De Oca in ***Estimation of average income in Cuban municipalities*** applies a small area statistics approach to capture a new kind of spatial differentiation among the territorial units (municipalities) resulting from recently implemented policies towards increasing economic activity in the country (expanding international tourism and joint ventures, legalization of the possession of US dollars; permitting self-employment, agricultural markets etc.). The focus is on finding small area estimates which are more precise than the direct estimates of monthly mean income for people aged 15 and over at a municipal level (all the 169 Cuban municipalities are included). Though the empirical results obtained so far provide only a rough estimates (given the limited income-relevant data), the estimates are still more precise than the direct estimates for small areas/domains.

In the next article, ***Bootstrap method with calibration for standard error estimators of income poverty measures,*** **A. Zięba-Pietrzak, J. Kordos** and **R. Wieczorkowski** discuss the calibration approach in sample surveys in reference to the Eurostat recommended approach. The authors focus on the indicators of poverty and social exclusion as a major monitoring tool for policy purposes, and they use data from European Statistics on Income and Living Conditions (EU-SILC) for empirical illustration. Given complexity of the EU-SILC sample design, a system of weights for estimates of population parameters and approximate methods of standard error estimation needed to be developed. In the study, the McCarthy and Snowden (1985) bootstrap method for standard errors estimation of income poverty measures is employed. Subsequently, the reweighting of bootstrap weights was applied, and results of such calibration are discussed.

Ahmed Hurairah in ***The beta Pareto distribution,*** introduce a generalization—referred to as the beta Pareto distribution, generated from the logit of a beta random variable. A comprehensive treatment of the mathematical properties of the beta Pareto distribution is given, including expressions for the *kth* moments of the distribution, variance, skewness, kurtosis, mean deviation about the mean, mean deviation about the median, Rényi and Shannon entropyies. Also, the estimation procedures by the methods of moments and maximum likelihood are

provided. It is shown that beta Pareto distributions are the most tractable of all the known distributions out of the certain class of distributions (as introduced by Jones (2004)). According to author's view, the approach presented in this paper can be used as a reference to obtain the corresponding results for other distributions belonging to such a type of generalization.

In the paper ***Robustness of the confidence interval for At-Risk-of-Poverty Rate,*** **W. Zieliński** discusses the problem of robustness of the confidence level of the confidence interval for binomial probability focusing on an earlier (Zieliński, 2009) introduced a nonparametric interval for a*t-risk-of-poverty rate*. Since it appeared that the confidence level of the interval depends on the underlying distribution of the income, for some distributions (e.g. lognormal, gamma, Pareto) the confidence level shown to be smaller than the nominal one. The question about the largest deviance from the nominal level is being extended by the above mentioned issue of robustness. The worst distribution is derived as well as the smallest true confidence level is calculated. Some asymptotic characteristics (sample size tends to infinity) are also specified.

Paper by **J. L. Wywiał,** ***Estimation of domain means on the basis of strategy dependent on depth function of auxiliary variables' distribution*** deals with the problem of estimation of a domain means in a finite and fixed population. It is assumed that observations of a multidimensional auxiliary variable are known in the population. The proposed estimation strategy consists of the Horvitz-Thompson estimator and the non-simple sampling design dependent on a synthetic auxiliary variable observations of which are equal to the values of a depth function of the auxiliary variable distribution. Both spherical and Mahalanobis depth functions are considered. A sampling design is proportionate to the maximal order statistic determined on the basis of the synthetic auxiliary variable observations in a simple sample drawn without replacement. A computer simulation analysis leads to the conclusion that the proposed estimation strategy is more accurate for domain means than the well known simple sample means.

In the *other articles* section included are papers presenting results of methodological investigation or (and) empirical analyses of policy-relevant issues, across a wide spectrum of realms.

In ***Remarks about the generalizations of the Fisher index,*** **J. Białek** discusses the Fisher index (defined as a geometric mean of Laspeyres and Paasche indexes), its theory and methodological characteristics, as the crossing of formulas and weights was considered the most suitable way to derive an "ideal" index formula. Despite that it satisfies most of the postulates of the axiomatic index theory, it has some limitations – e.g., the *time reversibility* test is satisfied only in some specific cases. In order to overcome them author presents a generalized Fisher index and develops some more general class of indexes (including the generalized Fisher index and other indices, like Laspeyres, Paasche, Fisher or Marhall-Edgeworth). As a pragmatic corollary to the employed strategy, author suggests that it is easier to prove that a given index belongs to a particular class than to verify that the axioms in question are satisfied. For

illustrative purposes, the general formula is applied to data from the whole time interval, with intention to show that it can be used in research, for instance, on pension or investment funds.

Paper by **J. Grosman** and **M. Kowerski,** ***Multiple–equation models of ordered dependent variables in exploration of the results of rehabilitation of locomotive organ disorders*** is focused on the factors determining patient's self-service during the admission and release from hospital . A two-equation model of ordered dependent variables is proposed, considered especially useful when the results of rehabilitation of locomotive organ disorders are not described by means of exact values obtained by mechanical measurements, but are described by means of qualitative valuation (ranking) made by a therapist, given that the distances between neighbouring ranks are not known. The advantages of the proposed model were presented on the basis of the results of estimation based on data of 4063 patients of hospitals from Mazowieckie and Warmińsko-Mazurskie provinces. This model allows for simulating the probabilities of the patient's self-service status both at the admission and at the release from a hospital, depending on various factors describing a patient.

The main objective of **T. Khan's** paper ***Identifying an appropriate forecasting model for forecasting total import of Bangladesh*** is to select an appropriate model for time series forecasting of total import (in taka crore) of Bangladesh. The presented study concerns mainly with seasonal autoregressive integrated moving average model (SARIMA), Holt-Winters' trend and seasonal model with seasonality modeled additively, and vector autoregressive model with some other relevant variables. An attempt was made to derive a unique and suitable forecasting model of total import of Bangladesh that would allow for making forecasts with minimum forecasting error. However, more research needs to be done to handle seasonality in a better way and to find more relevant variables that might be useful for forecasting total import of Bangladesh.

B. Liberda and **M. Pęczkowski** ask the question ***Does a change of occupation lead to higher earnings***? in their paper aimed at identifying whether and how the mobility between different types of broadly defined occupation - hired work, self-employment in industry, services and agriculture or social security beneficiaries – affects personal income of individuals. The Markov transition matrices are applied to the panel data from the Polish Household Budget Surveys (30,540 individuals), for years 2007-2008. According to the chief hypothesis, a change of occupation has an effect on individual capability to earn income, controlling for the occupation a person quits and the occupation a person starts, and for age, education level and a character of work (permanent or temporary). The hypothesis was tested using the regression analysis showing that the inter-occupational mobility matters mostly for those quitting hired work, for self-employment, for the better educated, and for respondents above 60 years of age.

In ***Neonatal mortality by gestational age and birth weight in Italy, 1998-2003: a record linkage study,*** **C. Marini** and **A. Nuccitelli** discuss neonatal

mortality rates by gestational age and birth weight category as particularly important indicators of maternal and child health and care quality. Since these specific rates have not been calculated in Italy since 1999, the main aim of their work is to assess the possibility of retrieving information on neonatal mortality by the linkage between records related to live births and records related to infant deaths within the first month of life, with reference to 2003 and 2004 birth cohorts. The focus is on some critical aspects of the most used record linkage approach as specific problems may arise from the choice of records to be linked if there are consistency constraints between pairs (in this context, one death record can be linked to at most one birth record). However, given the actual quality of the starting data, the retrieval of information on neonatal mortality by gestational age and birth weight is limited to Northern Italy. Specific neonatal mortality rates are provided with reference to 2003 and discussed with particular emphasis on quality issues in the data collection processes.

Włodzimierz OKRASA
Editor-in-Chief

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 7

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, August 2011
Vol. 12, No. 1, pp. 9—24

SUBSAMPLING THE NONRESPONDENTS IN CLUSTER SAMPLING ON SAMPLING ON TWO SUCCESSIVE OCCASIONS

Housila P. Singh, Sunil Kumar[1]

ABSTRACT

The problem of estimation of finite population mean for current occasion in the context of cluster sampling on two successive occasions when there is non-response on both the occasions. Estimators for the current occasion are derived as the particular case when there is non-response on first occasion and second occasion only. A comparison between variances of the estimates is studied. An empirical study is made to study the performance of the proposed strategy.

Key words: Non-response, Successive sampling; Mail surveys; Study variate; Auxiliary variate.

1. Introduction

Cluster or area sampling is widely practiced in sample surveys because of its low cost and time saving device to conduct large scale and complicated surveys. In those sample surveys where there is no list of the elements in the finite population or units of the population are widely scattered, it is required to take repeated observations on the selected units. It is well known that the cluster sampling is better than the simple random sampling when the intra class correlation within the same cluster is negative and smaller than $-(M-1)^{-1}$, where M denotes the size of the cluster. It is noted that the relative efficiency of the cluster sampling with respect to simple random sampling depends upon M, the size of the cluster and the intra-class correlation coefficient, it decreases if the size of the cluster M increases considerably. In practice, the intra-class correlation coefficient is generally non-negative and decreases as M increases, but the rate of decrease is small relative to the rate of increase in M, so that ordinarily, increase in the size of a cluster leads to substantial increase in the sampling variance of the sample estimate, see Sukhatme and Sukhatme (1970).

[1] School of Studies in Statistics, Vikram University, Ujjain – 456010, M. P., India. E-mail: hpsujn@rediffmail.com; sunilbhougal06@gmail.com

Zarkovich and Krane (1965) demonstrated that the correlation between two characters in cluster sampling with clusters as sampling units is expected to be higher than correlation coefficient in element sampling.

It is well known that the use of auxiliary at the estimation stage improves the efficiency of the estimators. Out of many ratio, product and regression methods of estimation are good illustrations in this context. If the survey is repetitive in nature, past values of the variable under investigation may be used as an auxiliary to improve on the precision of the current estimate. In such repetitive surveys, a fraction of the original sample units may be retained for use at the current occasion, while the remaining fraction is selected afresh. This procedure is known as sampling on successive occasions. The papers by Jessen (1942), Ware and Cunia (1962), Frayer and Furnival (1967), Singh (1968), Kathuria (1975), Raj (1979), Arnab (1980), Singh et al (1992) and Singh and Vishwakarma (2007). Recently Pradhan (2004) considered the problem of estimating the mean on second occasion over sampling on two successive occasions when the sampling units are clusters and the observations on the first occasion are regarded as ancillary information for the observations on the second or current occasion. These authors have carried out their studies under the supposition that there is complete response from all the sample units.

In surveys covering human populations this information is usually not obtained from all the sample units even after callbacks. Hansen and Hurwitz (1946) were the first to deal with the problem of incomplete samples in mail surveys. This method has been applied by various authors including Cochran (1977), Rao (1986), Khare and Srivastava (1993, 1995, 1997), Okafor and Lee (2000), Tabasum and Khan (2004, 2006), Choudhary et al. (2004), Okafor (2001, 2005) and Singh and Kumar (2008 a, b).

In this paper the theory for use of Hansen and Hurwitz (1946) technique for estimation of population mean for current occasion in the context of cluster sampling over sampling on two successive occasions has been developed. The results obtained are depicted with the help of numerical illustration.

2. Estimation of population mean for current occasion in the presence of nonresponse on both the occasions

Suppose that the population is composed of N clusters of M elements each, and that a simple random sample of n clusters is drawn without replacement from it. Let (x_{ij}, y_{ij}), $(i = 1,2,...,N;\ j = 1,2,...,M)$ be the values of the characteristic on first and second occasions for the j^{th} unit of the i^{th} cluster, respectively. We assume that the population can be divided into two classes, those who will respond at the first attempt and those who will not. Let the sizes of these two classes be N_1 and N_2 clusters respectively. In the second occasion $m\ (= n\lambda)$ of the n clusters selected on the first occasion retained at random and the remaining

$u = n\mu = (n - m)$ of the clusters are replaced by a fresh selection. The characters x and y are supposed to be correlated when they are observed (repeatedly). We assume that in the unmatched portion of the sample on the two occasions u_1 clusters (i.e. $M u_1$ elements) respond and u_2 clusters (i.e. $M u_2$ elements) do not. Similarly, in the matched portion m_1 clusters (i.e. $M m_1$ elements) respond and m_2 clusters (i.e. $M m_2$ elements) do not. Let $m_{h_2} = m_2/k$, $k > 1$ denote the size of the sub sample drawn from the non-response class from the matched portion of the sample on the two occasions for collecting information through personal interview. Similarly, denote by $u_{h_2} = u_2/k$, $k > 1$ the size of the sub sample drawn from the non-response class from the unmatched portion of the sample on the two occasions. Further, we define:

- $\overline{X}_{i.} = \sum_{j=1}^{M} x_{ij}/M$ and $\overline{Y}_{i.} = \sum_{j=1}^{M} y_{ij}/M$ are means of the i^{th} cluster on the first and second occasion respectively,
- $\overline{X} = \sum_{i=1}^{N} \overline{x}_{i.}/N$ and $\overline{Y} = \sum_{i=1}^{N} \overline{y}_{i.}/N$ are cluster population means of x and y respectively.
- $\overline{X}_{NM} = \sum_{i=1}^{N}\sum_{j=1}^{M} x_{ij}/NM$ and $\overline{Y}_{NM} = \sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}/NM$ are the population means of x and y per element on the first and second occasions respectively.
- $S_x^2 = \sum_{i=1}^{N}\sum_{j=1}^{M} \left(x_{ij} - \overline{X}_{NM}\right)^2 / \left(NM - 1\right)$: the population mean square between elements on the first occasion.
- $S_y^2 = \sum_{i=1}^{N}\sum_{j=1}^{M} \left(y_{ij} - \overline{Y}_{NM}\right)^2 / \left(NM - 1\right)$: the population mean square between elements on the second occasion,
- $\rho_x = \sum_{i}^{N}\sum_{j<k}^{M} \left(x_{ij} - \overline{X}_{NM}\right)\left(x_{jk} - \overline{X}_{NM}\right) / \left\{(M-1)(NM-1)S_x^2\right\}$: the intra-class correlation coefficient between elements of a cluster on first occasion,
- $\rho_y = \sum_{i}^{N}\sum_{j<k}^{M} \left(y_{ij} - \overline{Y}_{NM}\right)\left(y_{jk} - \overline{Y}_{NM}\right) / \left\{(M-1)(NM-1)S_y^2\right\}$: the intra-class correlation coefficient between elements of a cluster on second occasion,

- $\rho_b = \sum_{i=1}^{N}\left(\bar{Y}_{i.}-\bar{Y}_{NM}\right)\left(\bar{X}_{i.}-\bar{X}_{NM}\right)\Big/\left\{\sum_{i=1}^{N}\left(\bar{Y}_{i.}-\bar{Y}_{NM}\right)^2\sum_{i=1}^{N}\left(\bar{X}_{i.}-\bar{X}_{NM}\right)^2\right\}^{1/2}$: the simple correlation coefficient between cluster means on both occasions.

Let an equivalent sample of nM elements be selected from the population of NM elements by simple random sampling. Then the mean per element on the first and second occasions respectively be defined by $\bar{x}_{nM}=\sum_{l=1}^{nM}x_l/nM$ and $\bar{y}_{nM}=\sum_{l=1}^{nM}y_l/nM$. Similarly, we define $\bar{x}_{uM}=\sum_{l=1}^{uM}x_l/uM$ and $\bar{y}_{uM}=\sum_{l=1}^{uM}y_l/uM$ as sample means based on a simple random sample of uM units.

Further, we denote

- $S^2_{x(2)}=\sum_{i=1}^{N_2}\sum_{j=1}^{M}\left(x_{ij}-\bar{X}_{N_2M}\right)^2/\left(N_2M-1\right)$: the population mean square between the elements pertaining to the non-response class on the first occasion,
- $S^2_{y(2)}=\sum_{i=1}^{N_2}\sum_{j=1}^{M}\left(y_{ij}-\bar{Y}_{N_2M}\right)^2/\left(N_2M-1\right)$: the population mean square between the elements pertaining to the non-response class on the second occasion,
- $\rho_{x(2)}=\sum_{i=1}^{N_2}\sum_{j<k}^{M}\left(x_{ij}-\bar{X}_{N_2M}\right)\left(x_{jk}-\bar{X}_{N_2M}\right)\Big/\left\{\left(M-1\right)\left(N_2M-1\right)S^2_{x(2)}\right\}$: the intra-class correlation coefficient between elements of a cluster on first occasion pertaining to non-response class,
- $\rho_{y(2)}=\sum_{i=1}^{N_2}\sum_{j<k}^{M}\left(y_{ij}-\bar{Y}_{N_2M}\right)\left(y_{jk}-\bar{Y}_{N_2M}\right)\Big/\left\{\left(M-1\right)\left(N_2M-1\right)S^2_{y(2)}\right\}$: the intra-class correlation coefficient between elements of a cluster on second occasion pertaining to non-response class,
- $\rho_{b(2)}=\sum_{i=1}^{N_2}\left(\bar{Y}_{i.}-\bar{Y}_{N_2M}\right)\left(\bar{X}_{i.}-\bar{X}_{N_2M}\right)\Big/\left\{\sum_{i=1}^{N_2}\left(\bar{Y}_{i.}-\bar{Y}_{N_2M}\right)^2\sum_{i=1}^{N_2}\left(\bar{X}_{i.}-\bar{X}_{N_2M}\right)^2\right\}^{1/2}$: the simple correlation coefficient between cluster means on both the occasions pertaining to the non-response class,

- $\bar{X}_{N_2M} = \sum_{i=1}^{N_2}\sum_{j=1}^{M} x_{ij} / N_2 M$ and $\bar{Y}_{N_2M} = \sum_{i=1}^{N_2}\sum_{j=1}^{M} y_{ij} / N_2 M$ are the means of x and y per element on the first and second occasions respectively pertaining to the non-response class.

Let $\bar{x}^*_{mM}$ and $\bar{x}^*_{uM}$ denote the Hansen and Hurwitz estimator for matched and unmatched portion of the sample on the first occasion. Let the corresponding estimator for the second occasion be denoted by $\bar{y}^*_{mM}$ and $\bar{y}^*_{uM}$. Thus, we have the following set-up:

I[st] Occasion	$\bar{x}^*_{uM}$	$\bar{x}^*_{mM}$
2[nd] Occasion		$\bar{y}^*_{mM}$ $\quad$ $\bar{y}^*_{uM}$

where

$\bar{y}^*_{mM} = \left(m_1 M\, \bar{y}_{m_1 M} + m_2 M\, \bar{y}_{m_{h_2} M}\right) / mM$, $\bar{y}^*_{uM} = \left(u_1 M\, \bar{y}_{u_1 M} + u_2 M\, \bar{y}_{u_{h_2} M}\right) / uM$,

$\bar{x}^*_{mM} = \left(m_1 M\, \bar{x}_{m_1 M} + m_2 M\, \bar{x}_{m_{h_2} M}\right) / mM$, $\bar{x}^*_{uM} = \left(u_1 M\, \bar{x}_{u_1 M} + u_2 M\, \bar{x}_{u_{h_2} M}\right) / uM$,

$\bar{y}_{m_1 M} = \sum_{i=1}^{m_1}\sum_{j=1}^{M} y_{ij} / m_1 M$, $\bar{y}_{m_{h_2} M} = \sum_{i=1}^{m_{h_2}}\sum_{j=1}^{M} y_{ij} / m_{h_2} M$, $\bar{x}_{m_1 M} = \sum_{i=1}^{m_1}\sum_{j=1}^{M} x_{ij} / m_1 M$,

$\bar{x}_{m_{h_2} M} = \sum_{i=1}^{m_{h_2}}\sum_{j=1}^{M} x_{ij} / m_{h_2} M$, $\bar{y}_{u_1 M} = \sum_{i=1}^{u_1}\sum_{j=1}^{M} y_{ij} / u_1 M$, $\bar{y}_{u_{h_2} M} = \sum_{i=1}^{u_{h_2}}\sum_{j=1}^{M} y_{ij} / u_{h_2} M$,

$\bar{x}_{u_1 M} = \sum_{i=1}^{u_1}\sum_{j=1}^{M} x_{ij} / u_1 M$ and $\bar{x}_{u_{h_2} M} = \sum_{i=1}^{u_{h_2}}\sum_{j=1}^{M} x_{ij} / u_{h_2} M$.

With these notations and background we define a generalized estimator of the population mean $\bar{Y}$ on second (current) occasion as

$$\hat{D}_{12} = a\bar{x}^*_{uM} + b\bar{x}^*_{mM} + c\bar{y}^*_{mM} + d\bar{y}^*_{uM}, \tag{2.1}$$

where *a, b, c* and *d* are suitably chosen constants. We have

$$E\left(\hat{D}_{12}\right) = (a+b)\bar{Y}_{NM} + (c+d)\bar{X}_{NM} \tag{2.2}$$

$$= (a+b)\bar{Y} + (c+d)\bar{X} \text{ (since clusters are of equal size).}$$

In order that $\hat{D}_{12}$ is an unbiased estimator of $\bar{Y}_{NM}\left(or\ \bar{Y}\right)$, we have

$(a+b) = 0$ and $(c+d) = 1$.

Substituting the values of *a* and *d* we obtain

$$\hat{D}_{12} = a(\bar{x}^*_{uM} - \bar{x}^*_{mM}) + c\bar{y}^*_{mM} + (1-c)\bar{y}^*_{uM}. \tag{2.3}$$

The variance of $\hat{D}_{12}$ is given by

$$Var\left(\hat{D}_{12}\right)=\left[a^2Var(\bar{x}^*_{uM})+a^2Var(\bar{x}^*_{mM})+c^2Var(\bar{y}^*_{mM})+(1-c)^2Var(\bar{y}^*_{uM})\right.$$
$$\left.-2acCov(\bar{x}^*_{mM},\bar{y}^*_{mM})\right] \tag{2.4}$$

Assuming that N is sufficiently large, other covariance terms being zero, the variances and covariance involved in (2.4) are given by

$$Var\left(\bar{x}^*_{uM}\right)=\frac{1}{uM}\left\{\rho^*_x S^2_x+W_2(k-1)\rho^*_{x(2)}S^2_{x(2)}\right\},$$

$$Var\left(\bar{y}^*_{uM}\right)=\frac{1}{uM}\left\{\rho^*_y S^2_y+W_2(k-1)\rho^*_{y(2)}S^2_{y(2)}\right\},$$

$$Var\left(\bar{x}^*_{mM}\right)=\frac{1}{mM}\left\{\rho^*_x S^2_x+W_2(k-1)\rho^*_{x(2)}S^2_{x(2)}\right\},$$

$$Var\left(\bar{y}^*_{mM}\right)=\frac{1}{mM}\left\{\rho^*_y S^2_y+W_2(k-1)\rho^*_{y(2)}S^2_{y(2)}\right\},$$

$$Cov\left(\bar{x}^*_{mM},\bar{y}^*_{mM}\right)=\frac{1}{mM}\left\{\rho_b\sqrt{\rho^*_x\rho^*_y}S_yS_x+W_2(k-1)\rho_{b(2)}\sqrt{\rho^*_{x(2)}\rho^*_{y(2)}}S_{y(2)}S_{x(2)}\right\},$$

where $\rho^*_x=\{1+\rho_x(M-1)\}$, $\rho^*_{x(2)}=\{1+\rho_{x(2)}(M-1)\}$, $\rho^*_y=\{1+\rho_y(M-1)\}$ and $\rho^*_{y(2)}=\{1+\rho_{y(2)}(M-1)\}$.

Putting the above variances and covariance expressions in (2.4) we get the variance of $\hat{D}_{12}$ as

$$Var\left(\hat{D}_{12}\right)=\left[\frac{1}{M}\left(\frac{1}{u}+\frac{1}{m}\right)\left\{\left(\rho^*_x S^2_x+W_2(k-1)\rho^*_{x(2)}S^2_{x(2)}\right)a^2+\left(\rho^*_y S^2_y+W_2(k-1)\rho^*_{y(2)}S^2_{y(2)}\right)c^2\right\}\right.$$
$$+\frac{1}{uM}(1-2c)\left(\rho^*_y S^2_y+W_2(k-1)\rho^*_{y(2)}S^2_{y(2)}\right)$$
$$\left.-\frac{2ac}{mM}\left(\rho_b\sqrt{\rho^*_x\rho^*_y}S_yS_x+W_2(k-1)\rho_{b(2)}\sqrt{\rho^*_{x(2)}\rho^*_{y(2)}}S_{y(2)}S_{x(2)}\right)\right]. \tag{2.5}$$

which is minimized for

$$\left.\begin{aligned} a&=\frac{\lambda\mu V_yV_{xy}}{\left(V_xV_y-\mu^2V^2_{xy}\right)}\\ c&=\frac{\lambda V_yV_x}{\left(V_xV_y-\mu^2V^2_{xy}\right)}\end{aligned}\right\} \tag{2.6}$$

where

$$V_y=\left(\rho^*_y S^2_y+W_2(k-1)\rho^*_{y(2)}S^2_{y(2)}\right),$$
$$V_x=\left(\rho^*_x S^2_x+W_2(k-1)\rho^*_{x(2)}S^2_{x(2)}\right),$$

$$V_{xy} = \left(\rho_b \sqrt{\rho_x^* \rho_y^*} S_y S_x + W_2 (k-1) \rho_{b(2)} \sqrt{\rho_{x(2)}^* \rho_{y(2)}^*} S_{y(2)} S_{x(2)} \right).$$

Substituting the values of a and c the minimum linear unbiased estimator for the population mean on the second occasion is given by

$$\hat{D}_{12}^* = \left[\frac{\lambda V_y V_x}{\left(V_x V_y - \mu^2 V_{xy}^2 \right)} \left\{ \left(\frac{\mu V_{xy}}{V_x} \right) \left(\bar{x}_{uM}^* - \bar{x}_{mM}^* \right) + \bar{y}_{mM}^* \right\} + \left(\frac{\mu V_x V_y - \mu^2 V_{xy}^2}{V_x V_y - \mu^2 V_{xy}^2} \right) \bar{y}_{uM}^* \right]. \tag{2.7}$$

The variance of $\hat{D}_{12}^*$ is obtained as

$$Var\left(\hat{D}_{12}^* \right) = \frac{\left(\mu V_x V_y - \mu^2 V_{xy}^2 \right) V_y}{nM\mu \left(V_x V_y - \mu^2 V_{xy}^2 \right)} = \frac{\left(V_x V_y - \mu V_{xy}^2 \right) V_y}{nM \left(V_x V_y - \mu^2 V_{xy}^2 \right)}. \tag{2.8}$$

In case there is no-response then $Var\left(\hat{D}_{12}^* \right)$ reduces to

$$Var\left(\hat{D}_0^* \right) = \frac{\left(1 - \mu \rho_b^2 \right) V_y}{nM \left(1 - \mu^2 \rho_b^2 \right)} \tag{2.9}$$

which is the same as obtained by Pradhan (2004) for the optimum estimator of the population mean on the current occasion in the context of sampling on two occasion when there is non-response.

$$\hat{D}_0^* = \left[\frac{S_y}{S_x} \left\{ \frac{1 + (M-1)\rho_y}{1 + (M-1)\rho_x} \right\}^{1/2} \frac{\mu(1-\mu)\rho_b}{(1 - \mu^2 \rho_b^2)} \left(\bar{X}_{uM} - \bar{X}_{mM} \right) \right.$$

$$\left. + \frac{(1-\mu)}{(1 - \mu^2 \rho_b^2)} \bar{Y}_{uM} + \left\{ 1 - \frac{(1-\mu)}{(1 - \mu^2 \rho_b^2)} \right\} \bar{Y}_{mM} \right]. \tag{2.10}$$

Minimizing the variance of $\hat{D}_{12}^*$ given at (2.8) yields the optimum fraction of unmatched and matched portion of the sample as

$$\mu_{opt} = \frac{V_x V_y}{V_x V_y + \sqrt{V_x^2 V_y^2 - V_x V_y V_{xy}^2}}, \tag{2.11}$$

$$\lambda_{opt} = \frac{\sqrt{V_x^2 V_y^2 - V_x V_y V_{xy}^2}}{V_x V_y + \sqrt{V_x^2 V_y^2 - V_x V_y V_{xy}^2}}. \tag{2.12}$$

Thus, the resulting minimum value of the variance $\hat{D}_{12}^*$ is given by

$$\min . Var\left(\hat{D}_{12}^* \right) = \frac{V_{xy}^2 V_y}{2nM \left(V_x V_y - \sqrt{V_x^2 V_y^2 - V_x V_y V_{xy}^2} \right)}. \tag{2.13}$$

Remark 2.1

When there is non-response only on first occasion, then minimum variance unbiased linear estimator (MVULE) for the estimator for the population mean on current occasion can be obtained as follows:

$$\hat{D}_1 = \left\{a\left(\bar{x}^*_{uM} - \bar{x}^*_{mM}\right) + c\bar{y}_{mM} + (1-c)\bar{y}_{uM}\right\}. \tag{2.14}$$

The variance of $\hat{D}_1$ is given by

$$Var\left(\hat{D}_1\right) = \frac{1}{M}\left[\left\{\frac{1}{u}(1-2c) + \left(\frac{1}{m}+\frac{1}{u}\right)c^2\right\}V_{y(1)} - 2ac\frac{V_{xy(1)}}{m} + a^2\left(\frac{1}{m}+\frac{1}{u}\right)V_x\right], \tag{2.15}$$

where $V_{y(1)} = \rho^*_y S^2_y$ and $V_{xy(1)} = \rho_b S_y S_x \sqrt{\rho^*_y S^2_x}$.

The variance of $\hat{D}_1$ is minimized for

$$\left.\begin{aligned} a &= \frac{\lambda\mu V_{y(1)} V_{xy(1)}}{V_x V_{y(1)} - \mu^2 V^2_{xy(1)}} \\ c &= \frac{\lambda V_x V_{y(1)}}{V_x V_{y(1)} - \mu^2 V^2_{xy(1)}} \end{aligned}\right\}. \tag{2.16}$$

Thus, the MVULE for the current occasion in this case is given by

$$\hat{D}^*_1 = \left[\left(\frac{\lambda V_x V_{y(1)}}{V_x V_{y(1)} - \mu^2 V^2_{xy(1)}}\right)\left\{\left(\frac{\mu V_{xy(1)}}{V_x}\right)\left(\bar{x}^*_{uM} - \bar{x}^*_{mM}\right) + \bar{y}^*_{mM}\right\} + \left(\frac{\mu V_x V_{y(1)} - \mu^2 V^2_{xy(1)}}{V_x V_{y(1)} - \mu^2 V^2_{xy(1)}}\right)\bar{y}^*_{uM}\right] \tag{2.17}$$

Substituting (2.16) in (2.15) we get the variance of $\hat{D}^*_1$ as

$$Var\left(\hat{D}^*_1\right) = \frac{V_x V_{y(1)} - \mu V^2_{xy(1)}}{nM\left(V_x V_{y(1)} - \mu^2 V^2_{xy(1)}\right)}. \tag{2.18}$$

The optimum fraction to be unmatched is given by

$$\left.\begin{aligned} \mu_{opt} &= \frac{V_x V_{y(1)}}{V_x V_{y(1)} + \sqrt{V^2_x V^2_{y(1)} - V_x V_{y(1)} V^2_{xy(1)}}} \\ \lambda_{opt} &= \frac{\sqrt{V^2_x V^2_{y(1)} - V_x V_{y(1)} V^2_{xy(1)}}}{V_x V_{y(1)} + \sqrt{V^2_x V^2_{y(1)} - V_x V_{y(1)} V^2_{xy(1)}}} \end{aligned}\right\}. \tag{2.19}$$

Thus, the resulting minimum value of $Var\left(\hat{D}_1^*\right)$ at (2.18) is given by

$$\min .Var\left(\hat{D}_1^*\right)=\frac{V_{xy(1)}^2 V_{y(1)}}{2nM\left(V_x V_{y(1)}-\sqrt{V_x^2 V_{y(1)}^2-V_x V_{y(1)} V_{xy(1)}^2}\right)}. \tag{2.20}$$

Remark 2.2

When there is non-response only on second occasion, the MVULE for population mean on the current occasion can be obtained as follows:

$$\hat{D}_2=\left\{a\left(\bar{x}_{uM}-\bar{x}_{mM}\right)+c\,\bar{y}_{mM}^*+(1-c)\bar{y}_{uM}^*\right\}. \tag{2.21}$$

The variance of $\hat{D}_2$ is given by

$$Var\left(\hat{D}_2\right)=\frac{1}{M}\left[\frac{1}{mu}\left\{c^2+(1-2c)\lambda\right\}V_y-2ac\frac{V_{xy(1)}}{n\lambda}+\frac{a^2}{mu}V_{x(1)}\right], \tag{2.22}$$

where $V_{x(1)}=\rho_x^* S_x^2$.

The variance of $\hat{D}_2$ is minimized for

$$\left.\begin{aligned} a&=\frac{\lambda\mu V_y V_{xy(1)}}{V_{x(1)}V_y-\mu^2 V_{xy(1)}^2}\\ c&=\frac{\lambda V_{x(1)}V_y}{V_{x(1)}V_y-\mu^2 V_{xy(1)}^2}\end{aligned}\right\}. \tag{2.23}$$

Thus, the MVULE in this case is given by

$$\hat{D}_2^*=\left[\left(\frac{\lambda V_{x(1)}V_y}{V_{x(1)}V_y-\mu^2 V_{xy(1)}^2}\right)\left\{\left(\frac{\mu V_{xy(1)}}{V_{x(1)}}\right)\left(\bar{x}_{uM}-\bar{x}_{mM}\right)+\bar{y}_{mM}^*\right\}+\left(\frac{\mu V_{x(1)}V_y-\mu^2 V_{xy(1)}^2}{V_{x(1)}V_y-\mu^2 V_{xy(1)}^2}\right)\bar{y}_{uM}^*\right] \tag{2.24}$$

Substituting (2.23) in (2.22) we get the variance of $\hat{D}_2^*$ as

$$Var\left(\hat{D}_2^*\right)=\frac{V_{x(1)}V_y-\mu V_{xy(1)}^2}{nM\left(V_{x(1)}V_y-\mu^2 V_{xy(1)}^2\right)}. \tag{2.25}$$

The optimum fraction to be unmatched is given by

$$\left.\begin{aligned} \mu_{opt}&=\frac{V_{x(1)}V_y}{V_{x(1)}V_y+\sqrt{V_{x(1)}^2V_y^2-V_{x(1)}V_yV_{xy(1)}^2}}\\ \lambda_{opt}&=\frac{\sqrt{V_{x(1)}^2V_y^2-V_{x(1)}V_yV_{xy(1)}^2}}{V_{x(1)}V_y-\sqrt{V_{x(1)}^2V_y^2-V_{x(1)}V_yV_{xy(1)}^2}}\end{aligned}\right\}. \tag{2.26}$$

Thus, the resulting minimum value of $Var\left(\hat{D}_2^*\right)$ at (2.25) is given by

$$\min.Var\left(\hat{D}_2^*\right)=\frac{V_{xy(1)}^2 V_y}{2nM\left(V_{x(1)}V_y-\sqrt{V_{x(1)}^2V_y^2-V_{x(1)}V_yV_{xy(1)}^2}\right)}. \tag{2.27}$$

3. Comparison between variances of the estimator

3.1. Comparison between variance of $\hat{D}_{12}^*$ and $\hat{D}_2^*$.

We note from (2.13) and (2.27) that

$\min.Var\left(\hat{D}_{12}^*\right) > \min.Var\left(\hat{D}_2^*\right)$ if

$$\frac{V_{xy}^2}{\left(V_xV_y-\sqrt{V_x^2V_y^2-V_xV_yV_{xy}^2}\right)} > \frac{V_{xy(1)}^2}{\left(V_{x(1)}V_y-\sqrt{V_{x(1)}^2V_y^2-V_{x(1)}V_yV_{xy(1)}^2}\right)}. \tag{3.1}$$

3.2. Comparison between variance of $\hat{D}_{12}^*$ and $\hat{D}_1^*$.

We note from (2.13) and (2.20) that

$\min.Var\left(\hat{D}_{12}^*\right) > \min.Var\left(\hat{D}_1^*\right)$ if

$$\frac{V_y}{V_{y(1)}}\left(\frac{V_{xy}}{V_{xy(2)}}\right)^2 > \frac{\left(V_{x^2}V_y-\sqrt{V_x^2V_y^2-V_xV_yV_{xy}^2}\right)}{\left(V_xV_{y(1)}-\sqrt{V_x^2V_{y(1)}^2-V_xV_{y(1)}V_{xy(1)}^2}\right)}. \tag{3.2}$$

For the sake of convenience we assume that

$\rho_x=\rho_y=\rho\left(say\right)\Rightarrow\rho_x^*=\rho_y^*=\left\{1+\rho\left(M-1\right)\right\}=\rho^*\left(say\right)$,

$\rho_{x(2)}=\rho_{y(2)}=\rho_{(2)}\left(say\right)\Rightarrow\rho_{x(2)}^*=\rho_{y(2)}^*=\left\{1+\rho_{(2)}\left(M-1\right)\right\}=\rho_{(2)}^*\left(say\right)$,

$S_x^2=S_y^2=S^2\left(say\right)$, $S_{x(2)}^2=S_{y(2)}^2=S_{(2)}^2\left(say\right)$.

Thus $Var\left(\hat{D}_0^*\right)$, $Var\left(\hat{D}_1^*\right)$, $Var\left(\hat{D}_2^*\right)$ and $Var\left(\hat{D}_{12}^*\right)$ respectively in (2.9), (2.16), (2.23) and (2.8) reduce to:

$$Var\left(\hat{D}_0^*\right)=\frac{\left(1-\mu\rho_b^2\right)V}{nM\left(1-\mu^2\rho_b^2\right)}, \tag{3.3}$$

$$Var\left(\hat{D}_1^*\right) = Var\left(\hat{D}_2^*\right) = \frac{VV_{(1)} - \mu V_{(1)}^*}{nM\left(VV_{(1)} - \mu^2 V_{(1)}^{*2}\right)}, \tag{3.4}$$

$$Var\left(\hat{D}_{12}^*\right) = \frac{\left(V^2 - \mu V^{*2}\right)V}{nM\left(V^2 - \mu^2 V^{*2}\right)}, \tag{3.5}$$

where $V = V_y = V_x = \left\{\rho^* S^2 + W_2(k-1)\rho_{(2)}^* S_{(2)}^2\right\}$,

$V^* = V_{xy} = \left\{\rho_b \rho^* S^2 + W_2(k-1)\rho_{b(2)}\rho_{(2)}^* S_{(2)}^2\right\}$,

$V_{(1)} = V_{y(1)} = V_{x(1)} = \rho^* S^2$ and $V_{(1)}^* = V_{xy(1)} = \rho_b S^2 \sqrt{\rho^* S^2}$.

Hence, the expression for the percentage loss in precision of $\hat{D}_{12}^*$ over $\hat{D}_0^*$ is given by

$$L_{12} = \left\{\frac{Var\left(\hat{D}_{12}^*\right)}{Var\left(\hat{D}_0^*\right)} - 1\right\} \times 100, \tag{3.6}$$

where $Var\left(\hat{D}_{12}^*\right)$ and $Var\left(\hat{D}_0^*\right)$ are respectively given by (3.6) and (3.3).

The percentage loss in precision of $\hat{D}_2^*$ (or $\hat{D}_1^*$) over $\hat{D}_0^*$ can be achieved by replacing $Var\left(\hat{D}_{12}^*\right)$ with $Var\left(\hat{D}_2^*\right)$(or $Var\left(\hat{D}_1^*\right)$) given by (3.4). The percentage loss in precision of $\hat{D}_{12}^*, \hat{D}_2^*$ (or $\hat{D}_1^*$) over $\hat{D}_0^*$ for different values of W_2, $(k-1)$, M, μ, ρ, $\rho_{(2)}, \rho^*, \rho_{(2)}^*, \rho_b, \rho_{b(2)}, S^2$ and $S_{(2)}^2$ shown in Table 1 and 2. It is assumed that $N = 300$ and $n = 50$.

Table 1. Percentage loss in precision of $\hat{D}_{12}^*, \hat{D}_2^*$ (or $\hat{D}_1^*$) over $\hat{D}_0^*$ for different values of W_2, $(k-1)$, μ, ρ, $\rho_{(2)}$, ρ_b, $\rho_{b(2)}$, S^2 and $S_{(2)}^2$ when $M = 2$.

W_2	$(k-1)$	μ	ρ	$\rho_{(2)}$	ρ^*	$\rho_{(2)}^*$	ρ_b	$\rho_{b(2)}$	S^2	$S_{(2)}^2$	L_{12}	L_1
$\rho < \rho_{(2)}$												
0.7	2.5	0.5	0.1	0.8	1.1	1.8	0.8	0.2	0.4	0.2	17.04	12.62
0.7	2.5	0.5	0.2	0.7	1.2	1.7	0.8	0.2	0.4	0.2	16.37	12.12
0.7	2.5	0.5	0.3	0.6	1.3	1.6	0.8	0.2	0.4	0.2	15.68	11.61
0.7	2.5	0.5	0.4	0.5	1.4	1.5	0.8	0.2	0.4	0.2	14.95	11.10
$\rho > \rho_{(2)}$												
0.7	2	0.5	0.8	0.5	1.8	1.5	0.7	0.2	0.4	0.2	7.94	0.17
0.7	2	0.5	0.8	0.4	1.8	1.4	0.7	0.2	0.4	0.2	7.67	2.64
0.7	2	0.5	0.8	0.3	1.8	1.3	0.7	0.2	0.4	0.2	7.37	5.24
0.7	2	0.5	0.8	0.2	1.8	1.2	0.7	0.2	0.4	0.2	7.05	7.98
$\rho = \rho_{(2)}$												
0.4	2	0.4	0.7	0.7	1.7	1.7	0.8	0.2	0.5	0.2	7.33	3.76
0.4	2	0.4	0.7	0.7	1.7	1.7	0.8	0.2	0.5	0.2	7.33	3.76
0.4	2	0.4	0.7	0.7	1.7	1.7	0.8	0.2	0.5	0.2	7.33	3.76
0.4	2	0.4	0.7	0.7	1.7	1.7	0.8	0.2	0.5	0.2	7.33	3.76
$\rho_b < \rho_{b(2)}$												
0.7	1.5	0.3	0.5	0.8	1.5	1.8	0.4	0.5	0.8	0.2	*	*
0.7	1.5	0.3	0.5	0.8	1.5	1.8	0.3	0.5	0.8	0.2	*	*
0.7	1.5	0.3	0.5	0.8	1.5	1.8	0.2	0.5	0.8	0.2	*	*
0.7	1.5	0.3	0.5	0.8	1.5	1.8	0.1	0.5	0.8	0.2	*	*
$\rho_b > \rho_{b(2)}$												
0.6	2.5	0.7	0.5	0.7	1.5	1.7	0.7	0.4	0.4	0.2	6.58	2.59
0.6	2.5	0.7	0.5	0.7	1.5	1.7	0.7	0.3	0.4	0.2	8.22	2.59
0.6	2.5	0.7	0.5	0.7	1.5	1.7	0.7	0.2	0.4	0.2	9.64	2.59
0.6	2.5	0.7	0.5	0.7	1.5	1.7	0.7	0.1	0.4	0.2	10.87	2.59
$\rho_b = \rho_{b(2)}$												
0.6	1.5	0.5	0.4	0.5	1.4	1.5	0.2	0.2	0.5	0.3	2.22	*
0.6	1.5	0.5	0.4	0.5	1.4	1.5	0.2	0.2	0.5	0.3	2.22	*
0.6	1.5	0.5	0.4	0.5	1.4	1.5	0.2	0.2	0.5	0.3	2.22	*
0.6	1.5	0.5	0.4	0.5	1.4	1.5	0.2	0.2	0.5	0.3	2.22	*
$S^2 < S_{(2)}^2$												
0.6	1.5	0.4	0.5	0.7	1.5	1.7	0.7	0.2	0.1	0.6	12.59	7.21
0.6	1.5	0.4	0.5	0.7	1.5	1.7	0.7	0.2	0.2	0.5	11.37	7.15
0.6	1.5	0.4	0.5	0.7	1.5	1.7	0.7	0.2	0.3	0.4	9.83	6.85
0.6	1.5	0.4	0.5	0.7	1.5	1.7	0.7	0.2	0.4	0.3	7.95	6.29

$S^2 > S^2_{(2)}$												
0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
$S^2 = S^2_{(2)}$												
0.5	0.5	0.8	0.2	0.5	1.2	1.5	0.8	0.4	0.7	0.7	6.88	3.63
0.5	0.5	0.8	0.2	0.5	1.2	1.5	0.8	0.4	0.7	0.7	6.88	3.63
0.5	0.5	0.8	0.2	0.5	1.2	1.5	0.8	0.4	0.7	0.7	6.88	3.63
0.5	0.5	0.8	0.2	0.5	1.2	1.5	0.8	0.4	0.7	0.7	6.88	3.63
W_2												
0.7	0.5	0.6	0.5	0.8	1.5	1.8	0.8	0.1	0.7	0.2	5.49	*
0.6	0.5	0.6	0.5	0.8	1.5	1.8	0.8	0.1	0.7	0.2	4.85	0.12
0.5	0.5	0.6	0.5	0.8	1.5	1.8	0.8	0.1	0.7	0.2	4.16	1.56
0.4	0.5	0.6	0.5	0.8	1.5	1.8	0.8	0.1	0.7	0.2	3.42	3.05
μ												
0.8	0.5	0.8	0.6	0.7	1.6	1.7	0.8	0.2	0.5	0.5	11.07	3.51
0.8	0.5	0.7	0.6	0.7	1.6	1.7	0.8	0.2	0.5	0.5	11.93	5.62
0.8	0.5	0.6	0.6	0.7	1.6	1.7	0.8	0.2	0.5	0.5	11.51	5.70
0.8	0.5	0.5	0.6	0.7	1.6	1.7	0.8	0.2	0.5	0.5	10.35	4.42
$(k-1)$												
0.8	2	0.4	0.5	0.8	1.5	1.8	0.8	0.1	0.2	0.3	18.34	3.09
0.8	1.5	0.4	0.5	0.8	1.5	1.8	0.8	0.1	0.3	0.3	16.20	8.48
0.8	1	0.4	0.5	0.8	1.5	1.8	0.8	0.1	0.4	0.4	14.38	0.46
0.8	0.5	0.4	0.5	0.8	1.5	1.8	0.8	0.1	0.5	0.5	10.61	4.84

Table 1 exhibit that for all cases, i.e. $\rho(<,=,>)\rho_{(2)}$, $\rho_b(<,=,>)\rho_{b(2)}$, $S^2(<,=,>)S^2_{(2)}$, W_2, μ and $(k-1)$, when $M=2$, the percentage loss in precision ver $\hat{D}_0^*$ is maximum in $\hat{D}_{12}^*$ while it is least in $\hat{D}_1^*$ (or $\hat{D}_2^*$). It is noted that in some cases the percentage loss in precision of $\hat{D}_{12}^*$ over $\hat{D}_0^*$ is less than that of $\hat{D}_1^*$ (or $\hat{D}_2^*$) which contravene the condition (3.1). For the cases $\rho < \rho_{(2)}$, $S^2 < S^2_{(2)}$ and $S^2 > S^2_{(2)}$, the percentage loss in precision of all the estimators over $\hat{D}_0^*$ decreases. For the case $\rho > \rho_{(2)}$, the percentage loss in precision decreases in $\hat{D}_{12}^*$ while it increases in $\hat{D}_1^*$ (or $\hat{D}_2^*$). However, for the cases $\rho = \rho_{(2)}$, $\rho_b = \rho_{b(2)}$ and $S^2 = S^2_{(2)}$, the percentage loss in precision remains constant for all the estimators. For $\rho_b > \rho_{b(2)}$, the percentage loss in precision increases in $\hat{D}_{12}^*$ and remains constant in $\hat{D}_1^*$ (or $\hat{D}_2^*$). For different values of W_2,

the percentage loss in precision of all the estimators over $\hat{D}_0^*$ increases in $\hat{D}_{12}^*$ while it decreases in $\hat{D}_1^*$ (or $\hat{D}_2^*$). For the case of μ, the percentage loss in precision of all the estimators over $\hat{D}_0^*$ first increases and then decreases. For different values of $(k-1)$, the percentage loss in precision decreases in case of $\hat{D}_{12}^*$ and in $\hat{D}_1^*$ (or $\hat{D}_2^*$) it first increases and then decreases.

4. Conclusions

In sampling on two occasions we have considered the estimation of population mean on second occasion when the sampling units are clusters and the observations on the first occasion are regarded as ancillary information for the observations on the second or current occasion. In many surveys, information is usually not obtained from all the sample units even after callbacks. The technique of Hansen and Hurwitz (1946) for estimating population mean for current occasion in the context of cluster sampling over sampling on two successive occasions has been developed. Three different possible cases when there is non-response (i) on both the occasions, (ii) only on the first occasion, and (iii) only on the second occasion, have been discussed. The loss in precision by using the estimators $\hat{D}_{12}^*$ (optimum estimate in case (i)), $\hat{D}_1^*$ (optimum estimate in case (ii)) and $\hat{D}_2^*$ (optimum estimate in case (iii)) over direct estimator $\bar{y}^*$ of the population mean $\bar{Y}$ on the second occasion using no information gathered on the first occasion, have been computed. It is envisaged numerically that the loss in precision due to $\hat{D}_{12}^*$ over $\bar{y}^*$ is maximum as compared to $\hat{D}_1^*$ and $\hat{D}_2^*$ for $m =$ 2 and 4. Thus, the present study is recommended when there is need to correct for non-response in cluster sampling over two occasions.

Acknowledgement

The Authors express sincere thanks to the editor and the referee for their useful suggestions which have helped in no small way to the improvement of the quality of this paper.

REFERENCES

ARNAB, R. (1980): Two-stage sampling over two occasions. *Aust. J. Statist., 22,* 349-357.

COCHRAN, W. G. (1977): *Sampling Techniques*, 3rd edition, John Wiley and Sons. New York.

CHOUDHARY, R. K., BATHLA, H. V. L. and SUD, U. C. (2004): On non-response in sampling on two occasions. *J. Indian Soc. Agricultural Statist., 58,* 331-343.

FRAYER, W. E. and FURNIVAL, G. M. (1967): Area change estimates from sampling with partial replacement. *Forest Science, 13,* 72-77.

HANSEN, M. H. and HURWITZ, W. N. (1994): The problem of the non-response in sample surveys. *J. Amer. Statist. Assoc., 41,* 517-529.

JESSEN, R. K. (1942): Statistical investigation of a sample survey for obtaining farm Facts. *Iowa Agr. Exp. Stat. Res. Bull.,* 304.

KATHURIA, O. P. (1975): Some estimators in two-stage sampling on successive occasions with partial matching at both stages. *Sankhya, 37,* 147-162.

KHARE, B. B. and SRIVASTAVA, S. (1993): Estimation of population mean using auxiliary character in presence of nonresponse. *Nat. Acad. Sc. Letters, India, 16(3),* 111-114.

KHARE, B. B. and SRIVASTAVA, S. (1995): Study of conventional and alternative two -phase sampling ratio, product and regression estimators in presence of nonresponse. *Proc. Nat. Acad. Sci. India, 65(A), II,* 195-203.

KHARE, B. B. and SRIVASTAVA, S. (1997): Transformed ratio type estimators for the population mean in the presence of nonresponse. *Comm. Statist. - Theory Methods, 26(7),* 1779-1791.

OKAFOR, F. C. (2001): Treatment of nonresponse in successive sampling, *Statistica, LXI, (2),* 195-204.

OKAFOR, F. C. (2005): Sub-sampling the nonrespondents in two-stage sampling over two successive occasions. *Jour. Ind. Statist. Assoc., 43,* 33-49.

OKAFOR, F. C. and LEE, H. (2000): Double sampling for ratio and regression estimation with sub-sampling the nonrespondent. *Survey Methodology, 26,* 183-188.

PRADHAN, B. K. (2004): On efficient of cluster sampling on sampling on two Occasions. *Statistica, anno LXIV, 1,* 183-191.

RAJ, D. (1979): *Sampling Theory*, Tata Mc Graw Hill, New Delhi, 152-162.

RAO, P. S. R. S. (1986): Ratio estimation with sub sampling the non-respondents. *Survey methodology, 12(2),* 217-230.

SINGH, D. (1968): Estimates in successive sampling using a multistage design,. *J. Amer. Statist. Assoc., 63,* 99-112.

SINGH, H. P. and KUMAR, S. (2008 a): Estimation of population product in presence of non-response in successive sampling. *Statistical Papers, 51(4),* 975-996.

SINGH, H. P. and KUMAR, S. (2008 b): Effect of non-response in sampling over two successive occasions using auxiliary information. Statistics in Transition – new series, 9, (2), 273-296.

SINGH, H. P., SINGH, HARI P. and SINGH, V. P. (1992): A generalized efficient class of estimators of population mean in two-phase and successive sampling. *Int. Jour. Management and Systems, 8, (2),* 173-183.

SINGH, H. P. and VISHWAKARMA, G. K. (2007): A general class of estimators in successive sampling. *METRON, LXV, (2),* 201-227.

SUKHATME, P. V. and SUKHATME, B. V. (1970): *Sampling theory of surveys with Applications.* Food and Agriculture Organization, Rome, Second Edition.

TABASUM, R.. and KHAN, I. A. (2004): Double sampling for ratio estimation with non Response. *Jour. Ind. Soc. Agril. Statist., 58(3),* 300-306.

TABASUM, R.. and KHAN, I. A. (2006): Double Sampling Ratio Estimator for the population mean in Presence of Non-Response. *Assam Statist. Review, 20(1),* 73-83.

WARE, K. D. and CUNIA, T. (1962): Continuous forest inventory with partial replacement of samples. *Forest Science Monograph No. 3, Society of American Foresters, Bethesda.*

ZARKOVICH, S. S. and KRANE, J. (1965): Some efficient ways of cluster sampling. *Proceedings of 35th session of International Statistical Institute, Belgrade*.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 25—44

SOME ROTATION PATTERNS IN TWO-PHASE SAMPLING

G. N. Singh[1], Shakti Prasad[1]

ABSTRACT

A problem related to the estimation of population mean on the current occasion using two-phase successive (rotation) sampling on two occasions has been considered. Two-phase ratio, regression and chain-type estimators for estimating the population mean on current (second) occasion have been proposed. Properties of the proposed estimators have been studied and their respective optimum replacement policies are discussed. Estimators are compared with the sample mean estimator, when there is no matching and the natural optimum estimator, which is a linear combination of the means of the matched and unmatched portions of the sample on the current occasion. Results are demonstrated through empirical means of comparison and suitable recommendations are made.

Key words: Two-phase, successive sampling, auxiliary information, chain-type, bias, mean square error, optimum replacement policy.

Mathematics Subject Classification: 62D05

1. Introduction

In many social, demographic, industrial and agricultural surveys, the same population is sampled repeatedly and the same study variable is measured on each occasion, so that development over time can be followed. For example, labour force surveys are conducted monthly to estimate the number of people in employment, data on prices of goods are collected monthly to determine the consumer price index, political opinion surveys are conducted at regular intervals to know the voter's preferences, etc. In such studies, successive (rotation) sampling plays an important role to provide the reliable and the cost effective estimates of real life (practical) situations at different successive points of time (occasions). It also provides the effective (in terms of cost and precision) estimates of the patterns of change over a period of time.

[1] Department of Applied Mathematics, Indian School of Mines, Dhanbad-826004, India.
E-mail: gnsingh_ism@yahoo.com

The problem of successive (rotation) sampling with a partial replacement of sampling units was first considered by Jessen (1942) in the analysis of survey data related to agriculture farm. He pioneered using the entire information collected in previous investigations (occasions). The theory of successive (rotation) sampling was further extended by Patterson (1950), Rao and Graham (1964), Gupta (1979), Das (1982) and Chaturvedi and Tripathi (1983), among others. Sen (1971) applied this theory with success in designing the estimator for the population mean on the current occasion using information on two auxiliary variables available on previous occasion. Sen (1972, 1973) extended his work for several auxiliary variables. Singh *et al.* (1991) and Singh and Singh (2001) used the auxiliary information available only on the current occasion and proposed estimators for the current population mean in two-occasion successive (rotation) sampling. Singh (2003) generalized his work for h-occasions successive sampling. Feng and Zou (1997) and Biradar and Singh (2001) used the auxiliary information on both the occasions for estimating the current population mean in successive sampling.

In many situations, information on the auxiliary variable may be readily available on the first as well as on the second occasion; for example, tonnage (or seat capacity) of each vehicle or ship is known in survey sampling of transportation, number of beds in different hospitals may be known in hospital surveys, number of polluting industries and vehicles are known in environmental surveys, nature of employment status, educational status, food availability and medical aids of a locality are well known in advance for estimating the various demographic parameters in demographic surveys. Many other situations in life sciences could also be explored to show the benefits of the present study. Utilizing the auxiliary information on both the occasions, Singh (2005), Singh and Priyanka (2006, 2007, 2008), Singh and Karna (2009a,b) have proposed several estimators for estimating the population mean on current (second) occasion in two-occasion successive (rotation) sampling. It is worth to be mentioned that almost all the above recent works have assumed that the population means of the auxiliary variables are known, which may not often be the case. In such situations, it is more generously advisable to go for two-phase successive (rotation) sampling. Two-phase sampling is a well tested scheme to provide the estimates of unknown population parameters related to the auxiliary variables in first-phase sample. Motivated with this argument and utilizing the information on a stable auxiliary variable with unknown population mean, we have proposed some estimators under two-phase sampling scheme for estimating the current population mean in two-occasion successive (rotation) sampling. Behaviours of the proposed estimators are examined through empirical means of comparison and subsequently suitable recommendations are made.

2. Sample structures and notations on two occasions

Let $U = (U_1, U_2, - - -, U_N)$ be the finite population of N units and the one which has been sampled over two occasions. The character under study is denoted by x (y) on the first (second) occasion respectively. It is assumed that the information on an auxiliary variable z (stable over occasion) whose population mean is unknown on both the occasions is available and it is closely related (positively correlated) to x and y on the first and second occasions respectively. To furnish a good estimate of the population mean of the auxiliary variable z on the first occasion, a preliminary sample of size n' is drawn from the population by the method of simple random sampling without replacement (SRSWOR) and information on z is collected. Further, a second-phase sample of size n ($n' > n$) is drawn from the first-phase (preliminary) sample by the method of SRSWOR and henceforth the information on study character x is gathered. A random sub-sample of size $m = n\lambda$ is retained (matched) from the second-phase sample selected on the first occasion for its use on the second occasion. Once again to furnish a fresh estimate of the population mean of the auxiliary variable z on the second occasion, a preliminary (first-phase) sample of size u' is drawn from the non-sampled units of the population by the method of SRSWOR and information on z is collected. A second-phase sample of size $u = (n-m) = n\mu$ ($u' > u$) is drawn from the first-phase (preliminary) sample by the method of SRSWOR and the information on study variable y is gathered. It is obvious that the sample size on the second occasion is also n. λ and μ ($\lambda+\mu =1$) are the fractions of the matched and fresh samples, respectively, on the second (current) occasion. Hence onward, we consider the following notations for their further use:

$\overline{X}, \overline{Y}, \overline{Z}$: The population mean of the variables x, y, z respectively.

$\overline{x}_n, \overline{x}_m, \overline{y}_u, \overline{y}_m, \overline{z}_u, \overline{z}_n, \overline{z}_m$: The sample means of the respective variables based on the sample sizes shown in suffices $\overline{z}'_n, \overline{z}'_u$: The sample means of the auxiliary variable z and based on the first-phase samples of sizes n' and u' respectively $\rho_{yx}, \rho_{yz}, \rho_{xz}$: The correlation coefficients between the variables shown in suffices.

$S_X^2 = (N-1)^{-1} \sum_{i=1}^{N} (x_i - \overline{X})^2$: Population mean square of x.

S_Y^2, S_Z^2 : Population mean squares of y and z respectively.

3. Formulation of the estimators

To estimate the population mean $\bar{Y}$ on the current (second) occasion, two different sets of estimators are considered. One set of estimators $S_u = \{T_{1u}, T_{2u}\}$ based on sample of size u (= nμ) drawn afresh on the current (second) occasion and the second set of estimators $S_m = \{T_{1m}, T_{2m}\}$ based on the matched sample of size m (= nλ), which is common with both the occasions. Estimators of the sets S_u and S_m are defined as:

$$T_{1u} = \frac{\bar{y}_u}{\bar{z}_u}\bar{z}'_u \quad , T_{2u} = \bar{y}_u + b_{yz}(u)\left(\bar{z}'_u - \bar{z}_u\right), \quad T_{1m} = \frac{\bar{y}_m}{\bar{x}_m}\frac{\bar{x}_n}{\bar{z}_n}\bar{z}'_n,$$

$$T_{2m} = \bar{y}^*_m + b_{yx}(m)\left(\bar{x}^*_n - \bar{x}^*_m\right)$$

where $\bar{y}^*_m = \bar{y}_m + b_{yz}(m)\left(\bar{z}'_n - \bar{z}_m\right)$, $\bar{x}^*_n = \bar{x}_n + b_{xz}(n)\left(\bar{z}'_n - \bar{z}_n\right)$ and $\bar{x}^*_m = \bar{x}_m + b_{xz}(m)\left(\bar{z}'_n - \bar{z}_m\right)$.

$b_{yz}(u)$, $b_{yx}(m)$, $b_{yz}(m)$, $b_{xz}(n)$ and $b_{xz}(m)$ are the sample regression coefficients between the variables shown in suffices and based on the sample sizes shown in braces.

Combining the estimators of sets S_u and S_m, we have the following estimators of population mean $\bar{Y}$ on the current (second) occasion:

$$T_{ij} = \varphi_{ij}T_{iu} + \left(1 - \varphi_{ij}\right)T_{jm} \quad (i, j = 1, 2) \qquad (1)$$

where $\varphi_{ij}(i, j = 1, 2)$ are the unknown constants to be determined under certain criterion.

Remark 3.1.

For estimating the population mean on each occasion the estimators T_{iu} (i = 1, 2) are suitable, which implies that more belief on T_{iu} could be shown by choosing $\varphi_{ij}(i, j = 1, 2)$ as 1 (or close to 1), while for estimating the change over the occasions, the estimators $T_{jm}(j = 1, 2)$ could be more useful and hence φ_{ij} might be chosen as 0 (or close to 0). For asserting both the problems simultaneously, the suitable (optimum) choices of φ_{ij} are required.

4. Properties of the estimators t_{ij} (i, j = 1, 2)

Since $T_{i\,u}(i = 1, 2)$ and $T_{j\,m}(j = 1, 2)$ are ratio, simple linear regression, chain-type ratio and regression estimators, they are biased for the population mean $\bar{Y}$, therefore the resulting estimators T_{ij} (i, j = 1, 2) defined in equation (1) are also biased estimators of $\bar{Y}$. The bias B (.) and mean square errors M (.) are derived up to $o(n^{-1})$ under the large sample approximations and using the following transformations:

$\bar{y}_u = \bar{Y}(1+e_1)$, $\bar{y}_m = \bar{Y}(1+e_2)$, $\bar{x}_m = \bar{X}(1+e_3)$, $\bar{x}_n = \bar{X}(1+e_4)$,

$\bar{z}_u = \bar{Z}(1+e_5)$, $\bar{z}_u' = \bar{Z}(1+e_6)$,

$\bar{z}_n = \bar{Z}(1+e_7)$, $\bar{z}_n' = \bar{Z}(1+e_8)$, $s_{yz}(u) = S_{yz}(1+e_9)$, $s_z^2(u) = S_z^2(1+e_{10})$,

$s_{yx}(m) = S_{yx}(1+e_{11})$, $s_x^2(m) = S_x^2(1+e_{12})$, $s_{yz}(m) = S_{yz}(1+e_{13})$,

$s_z^2(m) = S_z^2(1+e_{14})$, $s_{xz}(n) = S_{xz}(1+e_{15})$, $s_z^2(n) = S_z^2(1+e_{16})$,

$s_{xz}(m) = S_{xz}(1+e_{17})$ and $\bar{z}_m = \bar{Z}(1+e_{18})$ such that $E(e_k) = 0$ and $|e_k| < 1 \;\forall\; k = 1, 2, - - -, 18.$

Under the above transformations $T_{i\,u}$ (i = 1, 2) and $T_{j\,m}$ (j = 1, 2) take the following forms:

$$T_{1u} = \bar{Y}(1+e_1)(1+e_6)(1+e_5)^{-1} \tag{2}$$

$$T_{2u} = \bar{Y}(1+e_1) + \bar{Z}\,\beta_{yz}(1+e_9)(e_6-e_5)(1+e_{10})^{-1} \tag{3}$$

$$T_{1m} = \bar{Y}(1+e_2)(1+e_4)(1+e_8)(1+e_3)^{-1}(1+e_7)^{-1} \tag{4}$$

and

$$T_{2m} = \bar{Y}(1+e_2) + \bar{Z}\beta_{yz}(1+e_{13})(e_8-e_{18})(1+e_{14})^{-1} + \beta_{yx}(1+e_{11})(1+e_{12})^{-1}(I_1 - I_2) \tag{5}$$

where

$$I_1 = \bar{X}(1+e_4) + \bar{Z}\,\beta_{xz}(1+e_{15})(1+e_{16})^{-1}(e_8 - e_7)$$

$$I_2 = \bar{X}(1+e_3) + \bar{Z}\,\beta_{xz}(1+e_{17})(1+e_{14})^{-1}(e_8 - e_{18})$$

Thus, we have the following theorems:

Theorem 4.1. Bias of the of estimators T_{ij} (i, j = 1, 2) to the first order of approximations are obtained as

$$B(T_{ij}) = \varphi_{ij}B(T_{iu})+(1-\varphi_{ij})B(T_{jm}) \; ; \; (i, j = 1, 2) \tag{6}$$

where

$$B(T_{1u}) = \bar{Y}\left(\frac{1}{u}-\frac{1}{u'}\right)\left(C_z^2-\rho_{yz}C_yC_z\right) \tag{7}$$

$$B\left(T_{2u}\right) = \beta_{yz}\left(\frac{1}{u}-\frac{1}{u'}\right)\left(\frac{\alpha_{003}}{S_z^2}-\frac{\alpha_{102}}{S_{yz}}\right) \tag{8}$$

$$B\left(T_{1m}\right) = \bar{Y}\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left(C_x^2-\rho_{yx}C_yC_x\right)+\left(\frac{1}{n}-\frac{1}{n'}\right)\left(C_z^2-\rho_{yz}C_yC_z\right)\right] \tag{9}$$

$$B\left(T_{2m}\right) = \beta_{yz}\left(\frac{1}{m}-\frac{1}{n'}\right)\left(\frac{\alpha_{003}}{S_z^2}-\frac{\alpha_{102}}{S_{yz}}\right)+\beta_{yx}\left(\frac{1}{m}-\frac{1}{n}\right)\left\{\left(\frac{\alpha_{030}}{S_x^2}-\frac{\alpha_{120}}{S_{yx}}\right)+\beta_{xz}\left(\frac{\alpha_{012}}{S_{xz}}-\frac{\alpha_{003}}{S_z^2}-\frac{\alpha_{021}}{S_x^2}+\frac{\alpha_{111}}{S_{yx}}\right)\right\} \tag{10}$$

where

$$\alpha_{rst} = E\left[\left(y-\bar{Y}\right)^r\left(x-\bar{X}\right)^s\left(z-\bar{Z}\right)^t\right]; \; (r, s, t \geq 0) \text{ are integers.}$$

Proof: The bias of the estimators T_{ij} (i, j = 1, 2) are given by

$$B(T_{ij}) = E\left[T_{ij}-\bar{Y}\right] = \varphi_{ij}E(T_{iu}-\bar{Y}) + (1-\varphi_{ij})\,E(T_{jm}-\bar{Y})$$

$$= \varphi_{ij}B(T_{iu})+\left(1-\varphi_{ij}\right)B(T_{jm}) \tag{11}$$

where

$$B(T_{iu}) = E\left[T_{iu}-\bar{Y}\right] \text{ and } B(T_{jm}) = E\left[T_{jm}-\bar{Y}\right].$$

Substituting the values of T_{1u}, T_{2u}, T_{1m} and T_{2m} from equations (2) – (5) in the equation (11), expanding the terms binomially and taking expectations up to o (n^{-1}), we have the expressions for the bias of the estimators T_{ij} (i, j = 1, 2) as described in equation (6).

Theorem 4.2. Mean square errors of the estimators T_{ij} (i, j = 1, 2) to the first order of approximations are obtained as

$$M(T_{ij}) = \varphi_{ij}^2 M(T_{iu}) + (1-\varphi_{ij})^2 M(T_{jm}) + 2\varphi_{ij}(1-\varphi_{ij})C_{ij} \; ; (i, j = 1, 2) \tag{12}$$

where

$$M(T_{1u}) = \left[2\left(\frac{1}{u}-\frac{1}{N}\right)(1-\rho_{yz}) - \left(\frac{1}{u'}-\frac{1}{N}\right)(1-2\rho_{yz})\right] S_y^2 \tag{13}$$

$$M(T_{2u}) = \left[\left(\frac{1}{u}-\frac{1}{N}\right) - \left(\frac{1}{u}-\frac{1}{u'}\right)\rho_{yz}^2\right] S_y^2 \tag{14}$$

$$M(T_{1m}) = \left[\left(\frac{1}{m}-\frac{1}{N}\right)2(1-\rho_{yx}) + \left(\frac{1}{n}-\frac{1}{N}\right)2(\rho_{yx}-\rho_{yz}) - \left(\frac{1}{n'}-\frac{1}{N}\right)(1-2\rho_{yz})\right] S_y^2$$

$$M(T_{2m}) = \left[\left(\frac{1}{m}-\frac{1}{N}\right) - \left(\frac{1}{m}-\frac{1}{n'}\right)\rho_{yz}^2 + \left(\frac{1}{m}-\frac{1}{n}\right)\rho_{yx}\left\{\rho_{yz}^2(2-\rho_{yx}) - \rho_{yx}\right\}\right] S_y^2$$

(16)

$$C_{11} = -\frac{S_y^2}{N} \tag{17}$$

$$C_{12} = -\frac{S_y^2}{N} \tag{18}$$

$$C_{21} = -\frac{S_y^2}{N} \tag{19}$$

and

$$C_{22} = -\frac{S_y^2}{N} \tag{20}$$

Remark 4.1.

The above results are derived under the assumptions that the coefficients of variations of the variables x, y and z are approximately equal and $\rho_{xz} = \rho_{yz}$, which

are intuitive assumptions considered by Cochran (1977) and Feng and Zou (1997).

Proof: It is obvious that the mean square errors of the estimators T_{ij} (i, j = 1, 2) are given by

$$M(T_{ij}) = E\left[T_{ij}-\bar{Y}\right]^2 = E\left[\varphi_{ij}\left(T_{iu}-\bar{Y}\right)+\left(1-\varphi_{ij}\right)\left(T_{jm}-\bar{Y}\right)\right]^2$$

$$= \varphi_{ij}^2 M(T_{iu})+\left(1-\varphi_{ij}\right)^2 M\left(T_{jm}\right)+2\varphi_{ij}\left(1-\varphi_{ij}\right)C_{ij} \qquad (21)$$

where

$$M\left(T_{iu}\right)=E\left[T_{iu}-\bar{Y}\right]^2,\ M(T_{jm}) = E\left[T_{jm}-\bar{Y}\right]^2 \text{ and } C_{ij} = E\left[\left(T_{iu}-\bar{Y}\right)\left(T_{jm}-\bar{Y}\right)\right]$$

Substituting the expressions of T_{iu} (i = 1, 2) and T_{jm} (j = 1, 2) given in equations (2)-(5) in equation (21), expanding the terms binomially and taking expectations up to $o(n^{-1})$, we have the expressions of the mean square errors of T_{ij} as given in equation (12).

5. Minimum mean square errors of the estimators t_{ij} (i, j = 1, 2)

Since the mean square errors of the estimators T_{ij} (i, j = 1, 2) in equation (12) are the functions of the unknown constants φ_{ij} (i, j = 1, 2), therefore they are minimized with respect to φ_{ij} and subsequently the optimum values of φ_{ij} are obtained as

$$\varphi_{ij_{opt}} = \frac{M(T_{jm})-C_{ij}}{M(T_{iu})+M(T_{jm})-2C_{ij}};\ (i, j = 1, 2) \qquad (22)$$

Now, substituting the values of $\varphi_{ij\,opt}$ in equation (12), we get the optimum mean square errors of T_{ij} as

$$M(T_{ij})_{opt} = \frac{M(T_{iu}).M(T_{jm})-C_{ij}^2}{M(T_{iu})+M(T_{jm})-2C_{ij}};\ (i, j = 1, 2) \qquad (23)$$

Further, substituting the values from equations (13)-(20) in equations (22) and (23), the simplified values of $\varphi_{ij_{opt}}$ and $M(T_{ij})_{opt}$ are given as:

$$\varphi_{11_{opt}} = \frac{\mu_{11}\left[A_7+\mu_{11}A_8\right]}{\left[A_1+\mu_{11}A_{10}+\mu_{11}^2A_{11}\right]} \tag{24}$$

$$M(T_{11})_{opt} = \frac{1}{n}\left[\frac{A_{13}+\mu_{11}A_{14}+\mu_{11}^2A_{15}}{A_1+\mu_{11}A_{10}+\mu_{11}^2A_{11}}\right]S_y^2 \tag{25}$$

$$\varphi_{12_{opt}} = \frac{\mu_{12}\left[A_{23}+\mu_{12}A_{24}\right]}{\left[A_1+\mu_{12}A_{25}+\mu_{12}^2A_{26}\right]} \tag{26}$$

$$M(T_{12})_{opt} = \frac{1}{n}\left[\frac{A_{30}+\mu_{12}A_{31}+\mu_{12}^2A_{32}}{A_1+\mu_{12}A_{25}+\mu_{12}^2A_{26}}\right]S_y^2 \tag{27}$$

$$\varphi_{21_{opt}} = \frac{\mu_{21}\left[A_7+\mu_{21}A_8\right]}{\left[A_{18}+\mu_{21}A_{16}+\mu_{21}^2A_{17}\right]} \tag{28}$$

$$M(T_{21})_{opt} = \frac{1}{n}\left[\frac{A_{19}+\mu_{21}A_{20}+\mu_{21}^2A_{21}}{A_{18}+\mu_{21}A_{16}+\mu_{21}^2A_{17}}\right]S_y^2 \tag{29}$$

$$\varphi_{22_{opt}} = \frac{\mu_{22}\left[A_{23}+\mu_{22}A_{24}\right]}{\left[A_{18}+\mu_{22}A_{33}+\mu_{22}^2A_{34}\right]} \tag{30}$$

$$M(T_{22})_{opt} = \frac{1}{n}\left[\frac{A_{38}+\mu_{22}A_{39}+\mu_{22}^2A_{40}}{A_{18}+\mu_{22}A_{33}+\mu_{22}^2A_{34}}\right]S_y^2 \tag{31}$$

where

$A_1 = 2\left(1-\rho_{yz}\right),\ A_2 = \left(1-2\rho_{yz}\right),\ A_3 = 2\left(1-\rho_{yx}\right),\ A_4 = 2\left(\rho_{yx}-\rho_{yz}\right),\ A_5 = \rho_{yz}^2$

$A_6 = A_3+A_4-A_2-1,\ A_7 = A_3+A_4-f'A_2-f\,A_6,\ A_8 = f'A_2-A_4+f\,A_6,$

$A_9 = A_1-2A_2+A_3+A_4-2,\ A_{10} = A_3-A_1-A_2\left(t+f'\right)+A_4-fA_9,$

$A_{11} = A_2\left(t+f'\right)-A_4+f\,A_9,\ A_{12} = \left(A_1-A_2\right)\left(A_3+A_4-A_2\right)-1,$

$A_{13} = A_1\left(A_3+A_4-f'A_2\right)-f\,A_1\left(A_3+A_4-A_2\right),$

$A_{14} = A_1\left(f'A_2-A_4\right)-\left(tA_2+f\left(A_1-A_2\right)\right)\left(A_3+A_4-f'A_2\right)+f\left(A_3+A_4-A_2\right)\left(A_1+tA_2\right)+f^2A_{12}$

$A_{15} = \left(tA_2+f\left(A_1-A_2\right)\right)\left(A_4-f'A_2\right)- f\left(tA_2\left(A_3+A_4-A_2\right)-f A_{12}\right)$,

$A_{16}= A_7-1+A_5\left(t+1\right)$,

$A_{17}= A_8- tA_5$, $A_{18}=1-A_5$, $A_{19}= A_{18}\left(A_7-f\right)$,

$A_{20}=A_{18}\left(A_8+f\right)+tA_5\left(A_7-f\right)-fA_7$, $A_{21}= tA_5\left(A_8-f\right)-fA_8$,

$A_{22}= \rho_{yx}\left\{\rho_{yz}^2\left(2-\rho_{yx}\right)-\rho_{yx}\right\}$, $A_{23}= A_{18}+A_5f'$, $A_{24}= A_{22}- f' A_5$,

$A_{25}= A_{23}-A_1-tA_2-\left(A_1-A_2-1\right)f$, $A_{26}= tA_2+A_{24}+(A_1-A_2-1)f$, $A_{27}= A_1A_{23}$,

$A_{28}= A_1A_{24}-tA_2A_{23}$, $A_{29}= -tA_2A_{24}$, $A_{30}= A_{27}-fA_1$,

$A_{31}= A_{28}-f\left\{\left(A_1-A_2\right)A_{23}-\left(tA_2+A_1\right)\right\}+f^2\left(A_1-A_2-1\right)$,

$A_{32}= A_{29}-f\left\{\left(A_1-A_2\right)A_{24}+tA_2\right\}-f^2\left(A_1-A_2-1\right)$, $A_{33}= tA_5-A_{18}+A_{23}$,

$A_{34}= A_{24}-tA_5$, $A_{35}= A_{18}\left(A_{18}+f'A_5\right)$, $A_{36}= A_{18}\left(A_{22}-f'A_5+tA_5\right)+f't\,A_5^2$,

$A_{37}= tA_5\left(A_{22}-f'A_5\right)$, $A_{38}= A_{35}-f\,A_{18}$, $A_{39}= A_{36}- f\,A_5\left(f'+t\right)$,

$A_{40}= A_{37}-f\left\{A_{22}-A_5\left(t+f'\right)\right\}$, $f= \frac{n}{N}$, $f'= \frac{n}{n'}$ and $t= \frac{n}{u'}$.

6. Optimum replacement policy

To determine the optimum values of μ_{ij} (i, j =1, 2) (fraction of samples to be drawn afresh on the current (second) occasion) so that population mean $\overline{Y}$ may be estimated with the maximum precision, we minimize mean square errors of T_{ij} (i, j =1, 2) given in equations (25), (27), (29) and (31) respectively with respect to μ_{ij}, which result in quadratic equations in μ_{ij} and respective solutions of μ_{ij} , say $\hat{\mu}_{ij}$ (i, j =1, 2), are given below:

$$Q_1\mu_{11}^2+2Q_2\mu_{11}+Q_3=0 \tag{32}$$

$$\hat{\mu}_{11}= \frac{-Q_2 \pm \sqrt{Q_2^2-Q_1Q_3}}{Q_1} \tag{33}$$

$$Q_7\mu_{12}^2 + 2Q_8\mu_{12} + Q_9 = 0 \tag{34}$$

$$\hat{\mu}_{12} = \frac{-Q_8 \pm \sqrt{Q_8^2 - Q_7Q_9}}{Q_7} \tag{35}$$

$$Q_4\mu_{21}^2 + 2Q_5\mu_{21} + Q_6 = 0 \tag{36}$$

$$\hat{\mu}_{21} = \frac{-Q_5 \pm \sqrt{Q_5^2 - Q_4Q_6}}{Q_4} \tag{37}$$

$$Q_{10}\mu_{22}^2 + 2Q_{11}\mu_{22} + Q_{12} = 0 \tag{38}$$

$$\hat{\mu}_{22} = \frac{-Q_{11} \pm \sqrt{Q_{11}^2 - Q_{10}Q_{12}}}{Q_{10}} \tag{39}$$

where

$Q_1 = A_{10}A_{15} - A_{14}A_{11}$, $Q_2 = A_1A_{15} - A_{13}A_{11}$, $Q_3 = A_1A_{14} - A_{13}A_{10}$,

$Q_4 = A_{21}A_{16} - A_{20}A_{17}$, $Q_5 = A_{21}A_{18} - A_{19}A_{17}$, $Q_6 = A_{18}A_{20} - A_{19}A_{16}$,

$Q_7 = A_{25}A_{32} - A_{31}A_{26}$, $Q_8 = A_1A_{32} - A_{30}A_{26}$, $Q_9 = A_1A_{31} - A_{30}A_{25}$,

$Q_{10} = A_{33}A_{40} - A_{39}A_{34}$, $Q_{11} = A_{18}A_{40} - A_{38}A_{34}$, $Q_{12} = A_{18}A_{39} - A_{38}A_{33}$

From equations (33), (35), (37) and (39), it is obvious that the real values of $\hat{\mu}_{ij}$ (i, j =1, 2) exist if the quantities under square roots are greater than or equal to zero. For any combinations of correlations ρ_{yx} and ρ_{yz}, which satisfy the conditions of real solutions, two real values of $\hat{\mu}_{ij}$ are possible. Hence, while choosing the values of $\hat{\mu}_{ij}$, it should be remembered that $0 \le \hat{\mu}_{ij} \le 1$. All the other values of μ_{ij} (i, j =1, 2) are inadmissible. Substituting the admissible values of $\hat{\mu}_{ij}$, say $\mu_{ij}^{(0)}$ (i, j =1, 2), from equations (33), (35), (37) and (39) into equations (25), (27), (29) and (31) respectively, we have the following optimum values of mean square errors of T_{ij} (i, j =1, 2).

$$M(T_{11}^0)_{opt} = \frac{1}{n}\left[\frac{A_{13} + \mu_{11}^{(0)}A_{14} + \mu_{11}^{(0)^2}A_{15}}{A_1 + \mu_{11}^{(0)}A_{10} + \mu_{11}^{(0)^2}A_{11}}\right]S_y^2 \tag{40}$$

$$M(T_{12}^0)_{opt} = \frac{1}{n}\left[\frac{A_{30} + \mu_{12}^{(0)}A_{31} + \mu_{12}^{(0)^2}A_{32}}{A_1 + \mu_{12}^{(0)}A_{25} + \mu_{12}^{(0)^2}A_{26}}\right]S_y^2 \tag{41}$$

$$M(T_{21}^{0})_{opt} = \frac{1}{n}\left[\frac{A_{19}+\mu_{21}^{(0)}A_{20}+\mu_{21}^{(0)^2}A_{21}}{A_{18}+\mu_{21}^{(0)}A_{16}+\mu_{21}^{(0)^2}A_{17}}\right]S_y^2 \tag{42}$$

$$M(T_{22}^{0})_{opt} = \frac{1}{n}\left[\frac{A_{38}+\mu_{22}^{(0)}A_{39}+\mu_{22}^{(0)^2}A_{40}}{A_{18}+\mu_{22}^{(0)}A_{33}+\mu_{22}^{(0)2}A_{34}}\right]S_y^2 \tag{43}$$

7. Efficiency comparison

The percent relative efficiencies of the estimators $T_{ij}(i, j = 1, 2)$ with respect to (i) $\bar{y}_n$, when there is no matching and (ii) $\hat{\bar{Y}}=\varphi^{*}\bar{y}_u+(1-\varphi^{*})\bar{y}'_m$, when no auxiliary information is used on any occasion, where $\bar{y}'_m=\bar{y}_m+\beta_{yx}\left(\bar{x}_n-\bar{x}_m\right)$, have been obtained for different choices of ρ_{yx} and ρ_{yz}. Since $\bar{y}_n$ and $\hat{\bar{Y}}$ are unbiased estimators of $\bar{Y}$, therefore, following Sukhatme *et al.* (1984), the variance of $\bar{y}_n$ and the optimum variance of $\hat{\bar{Y}}$ are given by

$$V(\bar{y}_n) = \left(\frac{1}{n}-\frac{1}{N}\right)S_y^2 \tag{44}$$

$$V\left(\hat{\bar{Y}}\right)_{opt^{*}} = \left[1+\sqrt{1-\rho_{yx}^2}\right]\frac{S_y^2}{2n}-\frac{S_y^2}{N} \tag{45}$$

For $N = 5000$, $n = 100$, $u' = 500$ and $n' = 500$, the percent relative efficiencies $E_{ij}^{(l)}\left(i, j, l = 1, 2\right)$ of the estimators $T_{ij}(i, j = 1, 2)$ are computed with respect to the estimators $\bar{y}_n$ and $\hat{\bar{Y}}$ and shown in the tables 1- 4; where

$$E_{ij}^{(1)} = \frac{V(\bar{y}_n)}{M\left(T_{ij}^{0}\right)_{opt}}\times 100 \text{ and } E_{ij}^{(2)} = \frac{V\left(\hat{\bar{Y}}\right)_{opt^{*}}}{M\left(T_{ij}^{0}\right)_{opt}}\times 100$$

Table 1. Optimum values of μ_{11} and percent relative efficiencies of T_{11} with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$

ρ_{yz}	ρ_{yx}	0.4	0.5	0.6	0.7	0.8	0.9
0.4	$\mu_{11}^{(o)}$	*	0.5228	0.5505	0.5858	0.6340	0.7101
	$E_{11}^{(1)}$	-	**	**	**	108.12	121.75
	$E_{11}^{(2)}$	-	**	**	**	**	**
0.5	$\mu_{11}^{(o)}$	0.4772	*	0.5279	0.5635	0.6126	0.6910
	$E_{11}^{(1)}$	**	-	105.69	112.99	123.08	139.28
	$E_{11}^{(2)}$	**	-	**	**	**	**
0.6	$\mu_{11}^{(o)}$	0.4495	0.4721	*	0.5359	0.5858	0.6667
	$E_{11}^{(1)}$	110.12	115.67	-	131.29	143.51	163.33
	$E_{11}^{(2)}$	105.43	107.76	-	112.14	114.22	116.32
0.7	$\mu_{11}^{(o)}$	0.4142	0.4365	0.4641	*	0.5505	0.6340
	$E_{11}^{(1)}$	131.67	138.55	147.05	-	173.46	198.70
	$E_{11}^{(2)}$	126.06	129.08	132.05	-	138.06	141.51
0.8	$\mu_{11}^{(o)}$	0.3660	0.3874	0.4142	0.4495	*	0.5858
	$E_{11}^{(1)}$	167.11	176.18	187.44	202.08	-	256.93
	$E_{11}^{(2)}$	160.00	164.14	168.31	172.61	-	182.98
0.9	$\mu_{11}^{(o)}$	0.2899	0.3090	0.3333	0.3660	0.4142	*
	$E_{11}^{(1)}$	242.00	255.47	272.22	294.11	325.12	-
	$E_{11}^{(2)}$	231.70	238.00	244.44	251.21	258.77	-

Note: "" indicate* $\mu_{11}^{(0)}$ *do not exist and "**" indicate no gain.*

Table 2. Optimum values of μ_{12} and percent relative efficiencies of T_{12} with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$

ρ_{yz}	ρ_{yx}	0.6	0.7	0.8	0.9
0.4	$\mu_{12}^{(o)}$	0.0583	0.2875	0.4441	0.5938
	$E_{12}^{(1)}$	115.10	117.89	124.79	138.70
	$E_{12}^{(2)}$	103.35	100.70	**	**
0.5	$\mu_{12}^{(o)}$	0.1270	0.3257	0.4631	0.5992
	$E_{12}^{(1)}$	126.05	129.91	138.10	154.13
	$E_{12}^{(2)}$	113.19	110.96	109.92	109.77
0.6	$\mu_{12}^{(o)}$	0.1797	0.3538	0.4741	0.5980
	$E_{12}^{(1)}$	142.57	147.57	157.37	176.31
	$E_{12}^{(2)}$	128.02	126.05	125.25	125.57
0.7	$\mu_{12}^{(o)}$	0.2223	0.3729	0.4760	0.5879
	$E_{12}^{(1)}$	168.37	174.77	186.81	210.05
	$E_{12}^{(2)}$	151.19	149.28	148.68	149.59
0.8	$\mu_{12}^{(o)}$	0.2712	0.3836	0.4643	0.5626
	$E_{12}^{(1)}$	212.72	221.18	236.74	267.06
	$E_{12}^{(2)}$	191.02	188.93	188.43	190.20
0.9	$\mu_{12}^{(o)}$	0.3344	0.3720	0.4190	0.4998
	$E_{12}^{(1)}$	308.63	320.57	342.77	386.86
	$E_{12}^{(2)}$	277.13	273.81	272.81	275.52

*Note: "**" indicate no gain.*

Table 3. Optimum values of μ_{21} and percent relative efficiencies of T_{21} with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$

ρ_{yz}	ρ_{yx}	0.4	0.5	0.6	0.7	0.8	0.9
0.4	$\mu_{21}^{(o)}$	*	*	0.8737	0.7693	0.7444	0.7717
	$E_{21}^{(1)}$	-	-	116.12	121.27	130.14	145.31
	$E_{21}^{(2)}$	-	-	104.27	103.58	103.58	103.49
0.5	$\mu_{21}^{(o)}$	*	*	0.9061	0.7457	0.7150	0.7471
	$E_{21}^{(1)}$	-	-	126.37	132.06	142.42	160.21
	$E_{21}^{(2)}$	-	-	113.48	112.80	113.36	114.10
0.6	$\mu_{21}^{(o)}$	0.1373	*	*	0.7230	0.6805	0.7180
	$E_{21}^{(1)}$	118.57	-	-	147.91	160.34	181.85
	$E_{21}^{(2)}$	113.52	-	-	126.34	127.62	129.51
0.7	$\mu_{21}^{(o)}$	0.2730	0.1813	*	0.7091	0.6383	0.6825
	$E_{21}^{(1)}$	140.53	146.69	-	172.42	187.91	215.13
	$E_{21}^{(2)}$	134.54	136.66	-	147.28	149.56	153.21
0.8	$\mu_{21}^{(o)}$	0.3022	0.2906	0.1686	0.7479	0.5831	0.6363
	$E_{21}^{(1)}$	174.83	184.07	194.49	214.67	235.01	272.04
	$E_{21}^{(2)}$	167.39	171.49	174.65	183.36	187.05	193.74
0.9	$\mu_{21}^{(o)}$	0.2631	0.2678	0.2402	0.7922	0.4931	0.5662
	$E_{21}^{(1)}$	247.85	261.68	278.64	305.50	336.31	394.07
	$E_{21}^{(2)}$	237.29	243.79	250.21	260.95	267.68	280.65

Note: "" indicate* $\mu_{21}^{(0)}$ *do not exist.*

Table 4. Optimum values of μ_{22} and percent relative efficiencies of T_{22} with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$

ρ_{yz}	ρ_{yx}	0.4	0.5	0.6	0.7	0.8	0.9
0.4	$\mu_{22}^{(o)}$	0.5089	0.5210	0.5390	0.5656	0.6065	0.6786
	$E_{22}^{(1)}$	119.31	122.17	126.41	132.68	142.33	159.37
	$E_{22}^{(2)}$	114.23	113.82	113.51	113.33	113.29	113.50
0.5	$\mu_{22}^{(o)}$	*	0.5109	0.5279	0.5536	0.5941	0.6667
	$E_{22}^{(1)}$	-	132.60	136.98	143.62	154.04	172.68
	$E_{22}^{(2)}$	-	123.54	123.00	122.67	122.60	122.98
0.6	$\mu_{22}^{(o)}$	0.4870	0.4962	0.5118	0.5365	0.5764	0.6495
	$E_{22}^{(1)}$	145.58	148.27	152.82	160.01	171.58	192.64
	$E_{22}^{(2)}$	139.38	138.13	137.23	136.67	136.56	137.20
0.7	$\mu_{22}^{(o)}$	0.4671	0.4741	0.4880	0.5113	0.5504	0.6240
	$E_{22}^{(1)}$	170.47	172.88	177.66	185.70	199.05	223.93
	$E_{22}^{(2)}$	163.21	161.06	159.53	158.61	158.43	159.48
0.8	$\mu_{22}^{(o)}$	0.4345	0.4386	0.4500	0.4715	0.5092	0.5834
	$E_{22}^{(1)}$	213.86	215.65	220.73	230.16	246.55	277.97
	$E_{22}^{(2)}$	204.75	200.91	198.21	196.59	196.23	197.97
0.9	$\mu_{22}^{(o)}$	0.3705	0.3707	0.3789	0.3973	0.4323	0.5054
	$E_{22}^{(1)}$	308.76	308.89	314.40	326.56	349.07	393.64
	$E_{22}^{(1)}$	295.61	287.77	282.31	278.93	277.83	280.35

Note: "" indicates $\mu_{22}^{(o)}$ does not exist.*

8. Conclusion

From Table 1, it is clear that

(a) For fixed values of $\rho_{yx,}$ the values of $\mu_{11}^{(0)}$ are decreasing while the values of $E_{11}^{(1)}$ and $E_{11}^{(2)}$ are increasing with the increasing values of ρ_{yz}. This behaviour is highly desirable, since it concludes that if highly correlated auxiliary character is available, it pays in terms of enhanced precision of estimates as well as it reduces the cost of survey.

(b) For fixed values of $\rho_{yz,}$ the values of $\mu_{11}^{(0)}$, $E_{11}^{(1)}$ and $E_{11}^{(2)}$ are increasing with the increasing values of $\rho_{yz.}$

(c) The minimum value of $\mu_{11}^{(0)}$ is 0.2899, which indicates that only about 29 percent of the total sample size is to be replaced on the current (second) occasion for the corresponding choices of correlations.

From Table 2, it is visible that

(a) For fixed values of ρ_{yx}, the values of $\mu_{12}^{(0)}$ are increasing for some choices of ρ_{yz} while decreasing pattern may also be seen for few choices of ρ_{yz} and the values of $E_{12}^{(1)}$ and $E_{12}^{(2)}$ are increasing with increase in the values of ρ_{yz}.

(b) For fixed values of ρ_{yz}, the values of $\mu_{12}^{(0)}$ and $E_{12}^{(1)}$ are increasing with the increasing values of ρ_{yx} while the values of $E_{12}^{(2)}$ are decreasing for some choices of $\rho_{yx.}$ This behaviour is in agreement with Sukhatme et al. (1984) results, which explains that the more value of ρ_{yx}, the more fraction of fresh sample is required on the current (second) occasion.

(c) The minimum value of $\mu_{12}^{(0)}$ is 0.0503, which indicates that the fraction of fresh sample to be replaced on current occasion is as low as about 5 percent of the total sample size, which leads to an appreciable reduction in the cost of the survey.

From Table 3, it is observed that

(a) For fixed values of ρ_{yx}, $\mu_{21}^{(0)}$ do not follow any definite pattern when the value of ρ_{yz} is increased while $E_{21}^{(1)}$ and $E_{21}^{(2)}$ are increasing with the increasing values of ρ_{yz}.

(b) For fixed values of ρ_{yz}, $\mu_{21}^{(0)}$ and $E_{21}^{(2)}$ do not follow any definite trend when the value of ρ_{yx} is increased while $E_{21}^{(1)}$ are increasing with the increase in the values of ρ_{yx}.

(c) The minimum value of $\mu_{21}^{(0)}$ is 0.1373, which indicates that the fraction of the fresh sample to be replaced on current occasion is as low as about 14 percent of the total sample size, leading to an appreciable reduction in the cost.

From Table 4, it can be seen that

(a) For fixed values of $\rho_{yx,}$ the values of $\mu_{22}^{(0)}$ are decreasing while the values of $E_{22}^{(1)}$ and $E_{22}^{(2)}$ are increasing with the increasing values of ρ_{yz}. This behaviour is highly desirable, since it concludes that if highly correlated auxiliary character is available, it pays in terms of enhanced precision of estimates as well as it reduces the cost of the survey.

(b) For fixed values of $\rho_{yz,}$ the values of $\mu_{22}^{(0)}$ and $E_{22}^{(1)}$ are increasing with the increasing values of ρ_{yx} while the values of $E_{22}^{(2)}$ are decreasing for some choices of $\rho_{yx.}$ This behaviour is in agreement with Sukhatme et al. (1984) results, which explains that the more value of ρ_{yx} , the more fraction of fresh sample is required on the current (second) occasion.

(c) The minimum value of $\mu_{22}^{(0)}$ is 0.3705, which indicates that the fraction of fresh sample to be replaced on current occasion is as low as about 37 percent of the total sample size, which leads to an appreciable reduction in the cost of the survey.

Thus, it is clear that the use of information on the auxiliary variable is highly rewarding in terms of the proposed estimators. It is also clear that if a highly correlated auxiliary variable is used, only a relatively smaller fraction of the sample on the current (second) occasion is required to be replaced by a fresh sample, which reduces the cost of the survey.

Acknowledgement

Authors are thankful to the referee for his valuable and inspiring suggestions. Authors are also thankful to the Indian School of Mines, Dhanbad for providing financial and infrastructural support to carry out the present work.

REFERENCES

BIRADAR, R. S. AND SINGH, H. P. (2001): Successive sampling using auxiliary information on both occasions. *Cal. Statist. Assoc. Bull*. 51, 243-251.

COCHRAN, W.G., (1977): Sampling Techniques. 3rd ed. New York: Wiley.Chaturvedi, D. K. and Tripathi, T. P. (1983): Estimation of population ratio on two occasions using multivariate auxiliary information. *Jour. Ind. Statist. Assoc*., 21, 113-120.

DAS, A. K. (1982): Estimation of population ratio on two occasions. *Jour. Ind. Soc. Agric. Statist.* 34, 1-9.

FENG, S. AND ZOU, G. (1997): Sample rotation method with auxiliary variable. *Communications in Statistics-Theory and Methods*, 26, 6, 1497-1509.

GUPTA, P. C. (1979): Sampling on two successive occasions*. Jour. Statist. Res*. 13, 7-16.

JESSEN, R.J. (1942): Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin No. 304, Ames, Iowa, U. S. A.,* 1-104.

PATTERSON, H. D. (1950): Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society,* 12, 241-255.

RAO, J. N. K. and GRAHAM, J. E. (1964): Rotation design for sampling on repeated occasions. *Jour. Amer. Statist. Assoc.* 59, 492-509.

SEN, A. R. (1971): Successive sampling with two auxiliary variables. *Sankhya*, 33, Series B, 371-378.

SEN, A. R. (1972): Successive sampling with p $(p \geq 1)$ auxiliary variables. *Ann. Math. Statist.*, 43, 2031-2034.

SEN, A. R. (1973): Theory and application of sampling on repeated occasions with several auxiliary variables. *Biometrics* 29, 381-385.

SINGH, G. N. (2003): Estimation of population mean using auxiliary information on recent occasion in h-occasion successive sampling. *Statistics in Transition,* 6, 523-532.

SINGH, G. N. (2005): On the use of chain-type ratio estimator in successive sampling. *Statistics in Transition*, 7, 21-26.

SINGH, G. N. and SINGH, V. K. (2001): On the use of auxiliary information in successive sampling. *Jour. Ind. Soc. Agric. Statist*., 54 (1), 1-12.

SINGH, G. N. and PRIYANKA, K. (2006): On the use of chain-type ratio to difference estimator in successive sampling. *IJAMAS*, 5 (S06) 41-49.

SINGH, G. N. and PRIYANKA, K**.** (2007): On the use of auxiliary information in search of good rotation patterns on successive occasions. *Bulletin of Statistics and Economics,* 1 (A07) 42 – 60.

SINGH, G. N. and PRIYANKA, K. (2008): Search of good rotation patterns to improve the precision of estimates at current occasion. *Communications in Statistics- Theory and Methods,* 37(3), 337-348.

SINGH, V. K., SINGH, G. N. and SHUKLA, D. (1991): An efficient family of ratio-cum-difference type estimators in successive sampling over two occasions. *Jour. Sci. Res*. 41 C, 149-159.

SINGH, G. N. and KARNA, J. P (2009, a): Estimation of population mean on current occasion in two occasion successive sampling. *METRON*, 67(1), 69-85.

SINGH, G. N. and KARNA, J. P (2009, b): Search of effective rotation patterns in presence of auxiliary information in successive sample over two-occasions. *Statistics in Transition- New series* 10(1), 59-73.

SUKHATME, P. V., SUKHATME, B. V., SUKHATME, S. and ASOK, C. (1984): *Sampling theory of surveys with applications*. *Iowa State University Press, Ames, Iowa (USA) and Indian Society of Agricultural Statistics, New Delhi (India).*

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 45—56

ESTIMATION OF POPULATION MEAN USING MULTI-AUXILIARY CHARACTERS WITH SUBSAMPLING THE NONRESPONDENTS

B. B. Khare[1], R. R. Sinha[2]

ABSTRACT

The aim of this paper is to suggest a class of two phase sampling estimators for population mean using multi-auxiliary characters in presence of non-response on study character. The expressions for bias and mean square error are obtained. The condition for minimum mean square error of the proposed class of estimators has been given. The optimum values of the size of first phase sample, second phase sample and the sub sampling fraction of non-responding group have been determined for the fixed cost and for the specified precision. A comparative study of the proposed class of estimators has been carried out with an empirical study.

Key words: Population mean; Bias; Mean square error; Multi-auxiliary characters.

1. Introduction

Hansen and Hurwitz (1946) suggested that the effect of non-response while conducting a sample survey can be reduced by using the method of subsampling from non-responding units. Further, the improvement in reducing the effect of non-response was considered by El-Badry (1956) and Foradori (1961). Rao (1986, 90) and Khare and Srivastava (1996, 97, 2000) proposed some estimators for population mean by using the auxiliary character with known population mean in presence of non-response while Khare and Sinha (2009) proposed some classes of estimators for population mean by using multi-auxiliary characters in presence of non-response.

But sometimes it has been observed that the population means of the auxiliary characters are not known due to the change in scenario. In such case we draw a

[1] Department of Statistics, Banaras Hindu University, Varanasi, India.
E-mail: bbkhare56@yahoo.com.
[2] Department of Mathematics, Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, India. E-mail: raghawraman@gmail.com.

first phase sample to observe the auxiliary characters which is used in estimating the population mean of the auxiliary characters. Khare (1992) proposed two phase sampling regression estimators for population mean in presence of non-response. Further, Khare and Srivastava (1993, 1995) proposed two phase sampling ratio and product estimators and made a comparative study for two phase sampling ratios, product and regression type estimators in presence of non-response while Khare and Sinha (2002) proposed general classes of two phase sampling estimators for population mean using an auxiliary character in presence of non-response.

In this paper, we have suggested a class of two phase sampling estimators (T) for population mean using multi-auxiliary characters with unknown population means in presence of non-response on study character only. The expressions for bias, mean square error and the condition for attaining the minimum mean square error of the proposed class of estimators have been obtained. The optimum values of the size of the first phase sample (n'), second phase sample (n) and the subsampling fraction $(1/k)$ of non-responding group have been determined for the fixed cost and for the specified precision. The merits of the suggested class of estimators have been judged through an empirical study.

2. The suggested class of estimator

In most of the situations it usually happens that the list of the units is available but the population means $\bar{X}_1, \bar{X}_2, \ldots\ldots, \bar{X}_p$ of the auxiliary characters $x_1, x_2, \ldots\ldots, x_p$ respectively are not known. In such situations, we select a larger sample of size n' from N in the first phase by using simple random sampling without replacement (SRSWOR) method of sampling and collect information regarding the auxiliary characters and estimate $\bar{X}_j$ $(j = 1,2, \ldots, p)$ based on n' units by $\bar{x}_j'$ $(j = 1,2, \ldots, p)$. Again, select a second phase sample of size n $(< n')$ from selected n' first phase units by using SRSWOR method of sampling and observe the study character y. We observe that n_1 units are responding and n_2 units are not responding in the sample of size n for the study character y. Further, from n_2 non-responding units, we select a subsample of size r $\left(r = \frac{n_2}{k}, k > 1\right)$ using SRSWOR method of sampling by making extra effort and observe the study character y.

Now, we have information on $(n_1 + r)$ selected units for the study character y. Using Hansen and Hurwitz (1946) technique, the estimator for $\bar{Y}$ based on $(n_1 + r)$ units is given by

$$\bar{y}^* = \frac{n_1}{n}\bar{y}_1' + \frac{n_2}{n}\bar{y}_2'', \tag{2.1}$$

where $\bar{y}_1'$ and $\bar{y}_2''$ are the sample means of the character y based on n_1 and r units respectively.

The estimator $\bar{y}^*$ is unbiased and has the variance given by

$$V(\bar{y}^*) = \frac{1-f}{n} S_0^2 + \frac{W_2(k-1)}{n} S_{0(2)}^2 , \tag{2.2}$$

where S_0^2 and $S_{0(2)}^2$ are the population mean square of y for the entire population and for the non-responding part of the population respectively and $W_i = N_i/N\,, (i = 1,2),\ f = {}^{n}/_{N}$.

Let $\bar{x}_j^{'}$ and $\bar{x}_j$ denote the sample means of the auxiliary characters $x_j (j = 1,2,\dots,p)$ based on $n^{'}$ and n units respectively. Let $Y_l, X_{1l}, X_{2l}, \dots X_{pl}$ be the $l^{th} (l = 1,2,\dots,N)$ unit of the characters $y, x_1, x_2, \dots, x_p$ respectively in the population of size N and are assumed to be non-negative.

In case when population means $\bar{X}_1, \bar{X}_2, \dots\dots, \bar{X}_p$ of the auxiliary characters $x_1, x_2, \dots \dots, x_p$ respectively are not known then they are estimated by $\bar{x}_1^{'}, \bar{x}_2^{'}, \dots, \bar{x}_p^{'}$ which are based on larger first phase of size $n^{'}$. In such situation, when we have incomplete information on the study character but complete information on the auxiliary characters from the sample of size n, we propose some generalized two phase sampling estimators with suitably chosen constants $\upsilon_{1j}, \upsilon_{2j}, \upsilon_{3j}$ and β_{1j} for population mean $\bar{Y}$

$$T_{01}^{**} = v\, exp\Big[\textstyle\sum_{j=1}^{p} \upsilon_{1j}\, log\, z_j\Big], \tag{2.3}$$

$$T_{02}^{**} = v \textstyle\sum_{j=1}^{p} W_j\, z_j^{(\upsilon_{2j}/W_j)}, \qquad \sum_{j=1}^{p} W_j = 1 \tag{2.4}$$

and

$$T_{03}^{**} = \textstyle\sum_{j=1}^{p} \Big[W_j\, z_j^{(\upsilon_{3j}/W_j)}\Big]\big[v + \beta_{1j}(\, z_j - 1)\big], \tag{2.5}$$

which are the members of our suggested class of two phase sampling estimators T given by

$$T = f(v, \boldsymbol{z}') \tag{2.6}$$

such that

$$f(\bar{Y}, \boldsymbol{e}') = \bar{Y}\ , \quad f_1(\bar{Y}, \boldsymbol{e}') = \left(\frac{\partial}{\partial v} f(v, \boldsymbol{z}')\right)_{(\bar{Y}, \boldsymbol{e}')} = 1 \tag{2.7}$$

where $\boldsymbol{z}$ and $\boldsymbol{e}$ denote the column vectors $(z_1, z_2, \dots, z_p)^{'}$ and $(1,1,\dots,1)^{'}$ respectively and $v = \bar{y}^*$, $z_j = \frac{\bar{x}_j}{\bar{x}_j^{'}}\,, (j = 1,2,\dots,p)$.

The function $f(v, \boldsymbol{z}')$ also satisfies the following conditions:

(i) For any sampling design, whatever be the sample chosen, $(v, \boldsymbol{z}')$ assumes values in a bounded closed convex subset D, of $p+1$ dimensional real space containing the point $(\bar{Y}, \boldsymbol{e}')$.

(ii) In D, the function $f(v, \boldsymbol{z}')$ and its first and second derivatives exist and are continuous and bounded.

Here, $f_1(v, \boldsymbol{z}')$ and $f_2(v, \boldsymbol{z}')$ denote the first partial derivatives of $f(v, \boldsymbol{z}')$ with respect to v and $\boldsymbol{z}'$ respectively. The second partial derivative of $f(v, \boldsymbol{z}')$ with respect to $\boldsymbol{z}'$ is denoted by $f_{22}(v, \boldsymbol{z}')$ and the first partial derivative of $f_2(v, \boldsymbol{z}')$ with respect to v is denoted by $f_{12}(v, \boldsymbol{z}')$.

On account of the regularity conditions imposed on $f(v, \boldsymbol{z}')$, it may be seen that the bias and mean square error of the estimator T will always exist.

3. Bias and mean square error (mse)

Expand $f(v, \boldsymbol{z}')$ about the point $(\bar{Y}, \boldsymbol{e}')$ by using Taylor's series up to second order partial derivatives and using (2.7), the expressions for bias and mean square error of T up to the terms of order n^{-1} for any sampling design are as follows:

$$B(T) = E(v - \bar{Y})(\boldsymbol{z} - \boldsymbol{e})' f_{12}(v^*, \boldsymbol{z}'^*) + \frac{1}{2} E(\boldsymbol{z} - \boldsymbol{e})' f_{22}(v^*, \boldsymbol{z}'^*)(\boldsymbol{z} - \boldsymbol{e}) \tag{3.1}$$

and
$$M(T) = V(\bar{y}^*) + 2E(v - \bar{Y})(\boldsymbol{z} - \boldsymbol{e})' f_2(\bar{Y}, \boldsymbol{e}') + E\left(f_2(\bar{Y}, \boldsymbol{e}')\right)'(\boldsymbol{z} - \boldsymbol{e})(\boldsymbol{z} - \boldsymbol{e})' f_2(\bar{Y}, \boldsymbol{e}') \tag{3.2}$$

where $v^* = \bar{Y} + \emptyset(v - \bar{Y})$, $\boldsymbol{z}^* = \boldsymbol{e} + \emptyset_{\boldsymbol{1}}(\boldsymbol{z} - \boldsymbol{e})'$ and $\emptyset_1 = diag\left[\emptyset_{11}, \emptyset_{12}, \dots, \emptyset_{1p}\right]_{p \times p}$.

such that $0 < \emptyset, \emptyset_{1j} < 1, \quad \forall\, j = 1,2, \dots, p$

The optimum value of $f_2(\bar{Y}, \boldsymbol{e}')$ for which the $M(T)$ will be minimum is given by

$$f_2(\bar{Y}, \boldsymbol{e}') = -\left(E(\boldsymbol{z} - \boldsymbol{e})(\boldsymbol{z} - \boldsymbol{e})'\right)^{-1} E(v - \bar{Y})(\boldsymbol{z} - \boldsymbol{e}) \tag{3.3}$$

and the minimum value of $M(T)$ is given by

$$M(T)_{min.} = V(\bar{y}^*) - E(v - \bar{Y})(\boldsymbol{z} - \boldsymbol{e})' \left(E(\boldsymbol{z} - \boldsymbol{e})(\boldsymbol{z} - \boldsymbol{e})'\right)^{-1} E(v - \bar{Y})(\boldsymbol{z} - \boldsymbol{e}) \tag{3.4}$$

Considering simple random sampling without replacement (SRSWOR) method, let $\boldsymbol{A}_0 = \left[a_{0jj'}\right]$ be a $p \times p$ positive definite matrix and $\boldsymbol{b} = (b_1, b_2, \dots, b_3)'$ is a column vector such that

$$a_{0jj'} = \rho_{jj'} C_j C_{j'}, \qquad j \neq j' = 1,2, \dots, p$$

and
$$b_j = \rho_j C_j, \qquad j = 1,2, \dots, p,$$

where $C_j^2 = \frac{S_j^2}{\bar{X}_j^2}$, $S_j^2 = \frac{1}{N-1}\sum_{l=1}^{N}\left(X_{jl} - \bar{X}_j\right)^2$ and $\rho_{jj'}$ is the correlation coefficient between the auxiliary characters $(x_j, x_{j'})$ while ρ_j is the correlation coefficient between y and x_j for the entire group of the population.

Now, the expressions for bias and mean square error of T in the case of simple random sampling without replacement (SRSWOR) method of sampling are given by

$$B(T) = \gamma\left(\left(S_0 \boldsymbol{b}'\right) f_{12}(v^*, \boldsymbol{z}'^*) + \frac{1}{2}\, trace \boldsymbol{A}_0 f_{22}(v^*, \boldsymbol{z}'^*)\right) \tag{3.5}$$

and
$$M(T) = V(\bar{y}^*) + \gamma\left(\left(f_2(\bar{Y}, \boldsymbol{e}')\right)'. \boldsymbol{A}_0. f_2(\bar{Y}, \boldsymbol{e}') + 2S_0 \boldsymbol{b}' f_2(\bar{Y}, \boldsymbol{e}')\right) \tag{3.6}$$

wher $\gamma = \frac{1}{n} - \frac{1}{n'}$ and $S_0^2 = \frac{1}{N-1}\sum_{l=1}^{N}(Y_l - \bar{Y})^2$.

The mean square error $M(T)$ will attain minimum value if

$$f_2(\bar{Y}, \boldsymbol{e}') = -S_0 \boldsymbol{A}_0^{-1} \boldsymbol{b} \tag{3.7}$$

and the minimum value of $M(T)$ is given by

$$M(T)_{min.} = V(\bar{y}^*) - \gamma S_0^2 \boldsymbol{b}' \boldsymbol{A}_0^{-1} \boldsymbol{b} \tag{3.8}$$

The suggested class of two phase sampling estimators has a wider class of estimators and all the members of this class attain minimum value of mean square error given in (3.8) if the condition (3.7) is applied. Now, using the conditions (3.7) in case of SRSWOR method of sampling, the constants involved in T_{01}^{**}, T_{02}^{**} and T_{03}^{**} can be evaluated. The condition (3.7) is sometimes obtained in the form of constants along with some parameters and sometimes in the form of some conditions between parameters. The later one is difficult to realize in practice and rarely used. In former case, the value of constants in the form of parameters can be computed on the basis of past data. Reddy (1978) showed that such values are stable over time and region. However, if no guess value is available from the past data then one can estimate it on the basis of sample observations without any loss in efficiency. Srivastava and Jhajj (1983) showed that up to the terms of order (n^{-1}), the efficiency of such type of estimators does not decrease if we replace the optimum values of the constants by their estimates based on sample values.

4. Determination of n', n and k for fixed cost $C \leq C_0$

Let C_0 be the total cost (fixed) of the survey apart from overhead cost. The cost function C′ for the cost incurred on the survey apart from overhead expenses can be expressed by

$$C' = C_1' n' + C_1 n + C_2 n_1 + C_3 \frac{n_2}{k} \tag{4.1}$$

Since C' will vary from sample to sample, so we consider the expected cost C to be incurred in the survey apart from overhead expenses, which is given by

$$C = E(C') = C_1' n' + n\left[C_1 + C_2 W_1 + C_3 \frac{W_3}{k}\right], \tag{4.2}$$

where

C_1': The cost per unit of identifying and observing auxiliary characters,

C_1: The cost per unit of mailing questionnaire/visiting the unit at the second phase,

C_2: The cost per unit of collecting and processing data for the study character y obtained from n_1 responding units and

C_3: The cost per unit of obtaining and processing data for the study character y (after extra efforts) from the subsampled units.

The $M(T)$ can be expressed in terms of the notations V_{01}, V_{11}, V_{21} which is given as

$$M(T) = \frac{1}{n}V_{01} + \frac{1}{n'}V_{11} + \frac{k}{n}V_{21} + terms\ independent\ of\ n, n'\ and\ k \quad (4.3)$$

where V_{01}, V_{11} and V_{21} are the coefficients of the terms of $\frac{1}{n}$, $\frac{1}{n'}$ and $\frac{k}{n}$ respectively in the expressions of $M(T)$.

Now, for minimizing the $M(T)$ for the fixed cost $C \leq C_0$ and to obtain the optimum values of n', n and k, we define a function ψ given as

$$\psi = M(T) + \lambda\left\{C_1' n' + n\left(C_1 + C_2 W_1 + C_3 \frac{W_2}{k}\right) - C_0\right\} \quad (4.4)$$

where λ is Lagrange's multiplier.

Now, differentiating ψ with respect to n', n and k and equating to zero, we have

$$n' = \sqrt{\frac{V_{11}}{\lambda C_1'}} \quad (4.5)$$

$$n = \sqrt{\frac{V_{01} + kV_{21}}{\lambda\left(C_1 + C_2 W_1 + C_3 \frac{W_2}{k}\right)}} \quad (4.6)$$

and

$$k_{opt.} = \sqrt{\frac{C_3 W_2 V_{01}}{(C_1 + C_2 W_1)\, V_{21}}} \quad (4.7)$$

Using the value of k from (4.7) and putting the values of n' and n from (4.5) and (4.6) in (4.2), we have

$$\sqrt{\lambda} = \frac{1}{C_0}\left[\sqrt{C_1' V_{11}} + \sqrt{\left(V_{01} + k_{opt.} V_{21}\right)\left(C_1 + C_2 W_1 + C_3 \frac{W_2}{k_{opt.}}\right)}\right]. \quad (4.8)$$

It has also been observed that the determinant of the matrix of second order derivative of ψ with respect to n', n and k is positive for the optimum values of n', n and k, which shows that the solutions for n', n given by (4.5), (4.6) and the optimum value of k under the condition $C \leq C_0$ minimize the variance of T. It is also important to note here that the subsampling fraction $1/k_{opt.}$ will decrease as $\sqrt{C_3/(C_1 + C_2 W_1)}$ increases.

The minimum value of $M(T)$ can be obtained by putting the optimum values of n', n and k in the expression (4.3). Hence, we have

$$M(T)_{min.} = \frac{1}{C_0}\left[\sqrt{C_1' V_{11}} + \sqrt{\left(V_{01} + k_{opt.} V_{21}\right)\left(C_1 + C_2 W_1 + C_3 \frac{W_2}{k_{opt.}}\right)}\right]^2 - \frac{S_0^2}{N}. \quad (4.9)$$

In case of $\bar{y}^*$, the expected total cost is given by

$$C = E(C') = n\left(C_1 + C_2W_1 + C_3\frac{W_2}{k}\right). \tag{4.10}$$

For fixed cost C_0, the expression for $M(\bar{y}^*)_{min.}$ is given by

$$M(\bar{y}^*)_{min.} = \frac{1}{C_0}\left[\sqrt{V_0(C_1 + C_2W_1)} + \sqrt{(C_3W_2V_2)}\right]^2 - \frac{S_0^2}{N} \tag{4.11}$$

where V_0 and V_2 are the coefficients of the terms of $\frac{1}{n}$ and $\frac{k}{n}$ respectively in the expression (2.2).

5. Determination of n', n and k for the specified variance $V = V_0''$

Let V_0'' be the variance of the estimator T fixed in advance and we have

$$V_0'' = \frac{1}{n}V_{01} + \frac{1}{n'}V_{11} + \frac{k}{n}V_{21} - \frac{S_0^2}{N}. \tag{5.1}$$

For minimizing the average total cost C for the specified variance of the estimator T (i.e. $M(T) = V_0''$), we define a function ψ' which is given as

$$\psi' = C_1'n' + n\left(C_1 + C_2W_1 + C_3\frac{W_2}{k}\right) - \mu\left(M(T) - V_0''\right) \tag{5.2}$$

where μ is Lagrange's multiplier.

Now, for obtaining the optimum values of n', n and k, differentiate ψ' with respect to n', n and k and equating to zero, we have

$$n' = \sqrt{\frac{\mu V_{11}}{C_1'}} \tag{5.3}$$

$$n = \sqrt{\frac{\mu(V_{01} + k\,V_{21})}{\left(C_1 + C_2W_1 + C_3\frac{W_2}{k}\right)}} \tag{5.4}$$

and

$$k_{opt.} = \sqrt{\frac{V_{01}C_3W_2}{V_{21}(C_1 + C_2W_1)}}. \tag{5.5}$$

Again, by putting the values of n' and n from (5.3) and (5.4) utilizing the optimum value of k in (5.1), we get

$$\sqrt{\mu} = \frac{\left[\sqrt{C_1'V_{11}} + \sqrt{(V_{01} + k_{opt.}V_{21})\left(C_1 + C_2W_1 + C_3\frac{W_2}{k_{opt.}}\right)}\right]}{\left[V_0'' + \frac{S_0^2}{N}\right]}. \tag{5.6}$$

The minimum expected total cost incurred in attaining the specified variance V_0'' by the estimator T is then given by

$$C(T)_{min.} = \frac{\left[\sqrt{C_1' V_{11}} + \sqrt{(V_{01} + k_{opt.} V_{21})\left(C_1 + C_2 W_1 + C_3 \frac{W_2}{k_{opt.}}\right)}\right]^2}{\left[V_0'' + \frac{S_0^2}{N}\right]}. \tag{5.7}$$

6. An empirical study

96 village wise population of rural area under Police-station -Singur, District – Hooghly, West Bengal has been taken under the study from District Census Handbook 1981. The first 25% villages (i.e. 24 villages) have been considered as non-response group of the population. Here, we have taken the number of agricultural labours in the village as the study character (y) while the number of literate persons (x_1), the area (in hectares) of the village (x_2) and the number of cultivators in the village (x_3) are used as auxiliary characters. The values of the parameters of the population under study are given below:

$\bar{Y}$= 137.9271	$\bar{X}_1$= 955.8750	$\bar{X}_2$= 144.8720	$\bar{X}_3$= 185.2188
S_0 = 182.5012	C_1 = 1.1066	C_2 = 0.8115	C_3 = 1.0529
$S_{0(2)}$= 148.6390	ρ_1 = 0.705	ρ_2= 0.773	ρ_3= 0.786
ρ_{12}= 0.772	ρ_{13} = 0.770	ρ_{23} = 0.819	

In this problem, we have considered the estimator

$$T_{01}^{**} = v\, exp\left[\sum_{j=1}^{p} v_{1j}\, log\, z_j\right]$$

as a member of the proposed class of estimator T to study the relative efficiency with respect to $\bar{y}^*$ and other relevant estimators. The optimum values of the constant v_{1j} involved in T_{01}^{**} can be calculated by using the condition (3.7), which are as follows:

$$T_{01}^{**} \quad \begin{cases} T_{01}^{**}(p=1): v_{11} = -0.8430 \\ T_{01}^{**}(p=2): v_{11} = -0.3205,\ v_{12} = -0.9232 \\ T_{01}^{**}(p=3): v_{11} = -0.1530,\ v_{12} = -0.5507,\ v_{13} = -0.5163 \end{cases}$$

The mean square error and relative efficiency (R. E.) of T_{01}^{**} with respect to $\bar{y}^*$ using one, two and three auxiliary characters for different values of the subsampling fraction $(1/k)$ in case of fixed n and n' are given in Table 6.1. The relative efficiency of the proposed class of estimators with respect to $\bar{y}^*$ for fixed cost (C_0) and their expected cost for specified precision (V_0'') are given in Table 6.2 and Table 6.3 respectively.

Table 6.1. Relative efficiency R. E. $(\cdot)$ in % with respect to $\bar{y}^*$ (for fixed $n' = 60$ and $n = 40$ for different values of k)

Estimators	Auxiliary character(s)	$1/k$		
		1/4	1/3	1/2
$\bar{y}^*$	–	100.00 (899.8656)*	100.00 (761.7809)	100.00 (623.6962)
T_{01}^{**}	x_1	118.10 (761.9478)	122.11 (623.8631)	128.39 (485.7784)
T_{01}^{**}	x_1, x_2	123.95 (726.0187)	129.57 (587.9340)	138.65 (449.8493)
T_{01}^{**}	x_1, x_2, x_3	126.24 (712.8124)	132.55 (574.7277)	142.84 (436.6430)

*Figures in parenthesis give $M(\cdot)$.

On comparing the $R.E.(.)$, it is clear from Table 6.1 that the estimator T_{01}^{**} is more efficient than $\bar{y}^*$ in case of one auxiliary character for the different values of k. The mean square error of T_{01}^{**} decreases as the subsampling fraction $(1/k)$ increases. It has also been observed that the relative efficiency of T_{01}^{**} with respect to $\bar{y}^*$ is increasing by increasing the subsampling fraction $(1/k)$ and the number of auxiliary characters used. It means that all the numbers of the suggested class of two phase sampling estimators T in presence of non-response attain minimum mean square error for every number of auxiliary characters used if the condition (3.7) is applied.

Table 6.2. Relative efficiency R. E. $(\cdot)$ with respect to $\bar{y}^*$ (for fixed cost $C_0 =$ Rs. 200.00)

Estimators	Auxiliary character(s)	$C_1' = Rs.0.70$ $C_1 = Rs.2.00$ $C_2 = Rs.5.00$ $C_3 = Rs.25.00$				
		$k_{opt.}$	$n'_{opt.}$ $(approx.)$	$n_{opt.}$ $(approx.)$	$M(.)$	$R.E.(.)$ in %
$\bar{y}^*$	–	2.3383	-	24	1367.0545	100.00
T_{01}^{**}	x_1	1.4864	56	16	1151.9815	118.67
T_{01}^{**}	x_1, x_2	1.1664	68	14	933.4605	146.45
T_{01}^{**}	x_1, x_2, x_3	1.0240	74	13	835.0246	163.71

From the Table 6.2, we observe that in case of one auxiliary character $M(T_{01}^{**})$ is less than $M(\bar{y}^*)$ and the relative efficiency of T_{01}^{**} with respect to $\bar{y}^*$ increases by increasing the number of the auxiliary characters for fixed cost.

Table 6.3. Expected cost for different estimators for specified precision $V_0'' =$ 800.00

Estimators	Auxiliary character(s)	$C_1' = Rs.0.70$ $C_1 = Rs.2.00$ $C_2 = Rs.5.00$ $C_3 = Rs.25.00$			
		$k_{opt.}$	$n'_{opt.}$ $(approx.)$	$n_{opt.}$ $(approx.)$	$Expected\ cost$ $(in\ Rs.)$
$\bar{y}^*$	–	2.3383	-	36	298.88
T_{01}^{**}	x_1	1.4864	73	21	261.38
T_{01}^{**}	x_1, x_2	1.1664	76	15	223.72
T_{01}^{**}	x_1, x_2, x_3	1.0240	76	13	206.11

Further, for the specified precision, the Table 6.3 shows that the expected cost incurred in T_{01}^{**} is less than the cost incurred for $\bar{y}^*$ and the expected cost for attaining the specified precision of T_{01}^{**} decreases by increasing the number of auxiliary characters.

Acknowledgement

Authors are very much grateful to the referee and the editor for their valuable suggestions.

REFERENCES

EL-BADRY, M. A. (1956). A sampling procedure for mailed questionnaires, *Jour. Amer. Statist. Assoc*, 51, 209-227.

FORADORI, G. T. (1961). Some non-response sampling theory for two stage designs, *Mimeo*, 297, North Carolina State College, Raleigh.

HANSEN, M. H. and HURWITZ, W. N. (1946). The problem of non-response in sample surveys, *J. Amer. Stat. Assoc*, 41, 517-529.

KHARE, B. B. (1992). Nonresponse and the estimation of population mean in its presence in sample surveys, *Seminar proceedings. 12th orientation course (Science Stream) sponsored by U.G.C. Academic Staff College*, *B.H.U.*, India.

KHARE, B. B. and SINHA, R. R. (2002). General class of two phase sampling estimators for the population mean using an auxiliary character in the presence of non-response, *Proceedings of the 5th International Symposium on Optimization and Statistics*, 233-245.

KHARE, B. B. and SINHA, R. R. (2009). On class of estimators for population mean using multi-auxiliary character in presence of non-response, *Statistics in Transition-new series*, 10, 3-14.

KHARE, B. B. and SRIVASTAVA, S. (1993). Estimation of population mean using auxiliary character in presence of nonresponse, *Nat. Acad. Sc. Letters*, 16(3), 111-114.

KHARE, B. B. and SRIVASTAVA, S. (1995). Study of conventional and alternative two phase sampling ratio, product and regression estimators in presence of non-response, *Proc. Nat. Acad. Sci*., India, 65(a) II, 195-203.

KHARE, B. B. and SRIVASTAVA, S. (1996). Transformed product type estimators for population mean in presence of softcore observations, *Proc. Math. Soc. BHU*, 12, 29-34.

KHARE, B. B. and SRIVASTAVA, S. (1997). Transformed ratio type estimators for the population mean in the presence of non-response, *Commun. Statis.Theory Math*, 26(7), 1779-1791.

KHARE, B. B. and SRIVASTAVA, S. (2000). Generalised estimators for population mean in presence of nonresponse*, Internal. J. Math. & Statist. Sci*, 9(1), 75-87.

RAO, P. S. R. S. (1986). Ratio estimation with subsampling the nonrespondents, *Survey Methodology*, 12(2), 217-230.

RAO, P. S. R. S. (1990). Regression estimators with subsampling of nonrespondents, *In-Data Quality Control, Theory and Pragmatics, (Eds.) Gunar E. Liepins and V.R.R. Uppuluri*, *Marcel Dekker*, New York, 191-208.

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

SRIVASTAVA, S. K. and JHAJJ, H. S. (1983). A class of estimators of the population mean using multi-auxiliary information, *Cal. Statist. Assoc. Bull*, 32, 47-56.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 57—80

ESTIMATION OF AVERAGE INCOME IN CUBAN MUNICIPALITIES

Nestor Arcia Montes De Oca[1]

ABSTRACT

This article asserts that diversification to produce small area statistics for different fields in Cuban society should be a priority for the National Statistics Office. The key question about small area estimation is how to obtain reliable local statistics when the sample data contain too few observations for statistical inference of adequate precision. The social research presented here is focused on finding small area estimates which are more precise than the direct estimates of monthly mean income for people aged 15 and over at a municipal level. In this case, all 169 Cuban municipalities are considered small areas of interest.

The empirical results obtained from this application are only intended to provide a first impression of the usefulness of applying small area estimation methods in Cuba. This study yields more precise estimates than the direct estimates for small areas/domains, even though in Cuba, as in any other developing country, the search for suitable auxiliary variables is used to "borrow strength" from neighbouring areas or domains may frequently be an important limitation.

Key words: small area estimation, borrow strength.

1. Introduction

In recent years, the academic world has taken an increasing interest in the analysis of regional economic disparities which represent a serious challenge to achieving national economic growth and, thus, social cohesion. In Cuba, this analysis was particularly apposite after the collapse of the Soviet Union in the late 1980s and early 1990s, when the Cuban economy experienced a severe crisis. As part of its overall response to the crisis, the Cuban government initiated a series of monetary and market reforms which were designed to provide material incentives for economic activities. These included: expanding international tourism and joint

[1] National Statistics Office. Email: nestor@one.cu.

ventures; legalising the possession of US dollars - making it possible for Cubans to receive money from relatives or friends living abroad; permitting self-employment, agricultural markets and other markets (e.g. handicrafts) and others. While these reforms contributed to the subsequent economic recovery, they also resulted in an increase in economic inequality, even though salary scales in Cuba still remain fairly egalitarian. This fact has increased interest in producing regional statistical information and has stimulated research on income distribution, poverty and social exclusion at the small area level in Cuba.

In this respect, this article proposes a new small-area application intended to produce small area estimates which are more precise than direct estimates of *monthly mean income for those aged 15 and over at a municipal level*. The 169 Cuban municipalities are considered as the small areas of interest. These direct estimates are obtained from the second national survey on health risk factors and non-communicable chronic ailments (HRF) which was conducted during November-December 2000 by the National Institute of Hygiene, Epidemiology and Microbiology and the National Statistics Office (ONE) which, in turn, was represented by the Centre for Population and Development Studies (CEPDE, Spanish acronym) and the Social Statistics Department. In this survey, out of the 169 Cuban municipalities, 158 are represented in the sample and the remaining 11 are not sampled at all. Likewise, the direct estimates for some of these 158 sampled municipalities cannot be produced because they have unacceptably large variances. The challenge is to find plausible strategies for improving the direct estimates for the 158 sampled municipalities, when needed, and also to find acceptable estimates for the 11 non-sampled municipalities. Such a strategy may consist of using the same national sample size as the HRF survey and producing the municipal estimates by small area estimation (SAE) methods. If this strategy proves easy to apply, policy makers could use it to acquire more accurate information at a municipal level and this could be considered as an initial finding to apply to other indicators in future surveys.

In addition to the direct estimator, two small area model-based estimators are proposed in this article: (*a*) the synthetic-regression estimator; and (*b*) the empirical best linear unbiased prediction (EBLUP) estimator. The basic unit level model (Rao, 2003) is used to derive each of these two small area estimators. Using HRF survey data, it illustrates how the small area estimates derived from such a model could potentially be useful to improve the efficiency of the direct small area estimates. The MSE estimator of each estimator is used as a measure of uncertainty. Since the real population data are not available, two simulation experiments are used - a model-based simulation and a design-based simulation - to compare the performance of these three estimation methods. The performances of the MSE estimators for the best small area estimator are also compared to each other.

This article is structured as follows. Section 2 describes the sampling plan and the method of estimation. Section 3 describes the SAE methods (synthetic-

regression estimator and EBLUP) derived from the basic unit level model, which will be compared with the direct estimator. The results obtained from the basic unit level model are presented in Section 4. The comparison between the three estimators is based on the results obtained from model and design-based Monte Carlo simulation studies (Section 5). Section 6 presents the results obtained from the application using HRF data. Section 7 gives some final remarks and suggestions for further research.

2. Sampling design

This section describes the sampling design of the HRF survey, as well as the direct estimator of monthly mean income and its MSE estimator. The performance of this estimator will be compared to the two SAE estimators described later in this article (Section 3).

The HRF sample was drawn from the Cuban sampling frame for population surveys (ONE, 2004). A three-stage cluster probability sample was selected from urban areas. Fourteen provinces and the special Isle of Youth municipality were used as strata. Primary sampling units (PSU) were the sampled geographical areas (AGEM with an average 180 housing units (ONE, 2004), drawn with probability proportional to size, using the total number of houses with permanent residents within each stratum as size measure. The second sampling units (SSU) were blocks, which have at least 30 housing units and they were also selected with probability proportional to size using the total number of houses with permanent residents within each AGEM as size measure. Tertiary and last sampling units (TSU) were the sections (5 housing units on average), which were drawn with equal probability. A block is drawn within each selected AGEM, and finally two sections were drawn in each selected block, giving 10 houses in each AGEM. All adults aged 15 or over were interviewed in each selected house. The sampling design assured equal probability within each province, which allowed the acquisition of accurate and reliable information of the main health indicators at provincial and national level. Assuming a 95% confidence interval and 5% survey non-response rate, it is guaranteed a proper sample size to obtain the core outputs from the survey at provincial and national level. The province allocation was made with probability proportional to size that is given by the population of people aged 15 or older in each province. This sampling design leads to a sample size distribution for the 158 sampled municipalities that ranges between 16 and 1148 individuals, with no individuals for the 11 non-sampled municipalities (Figure 1A).

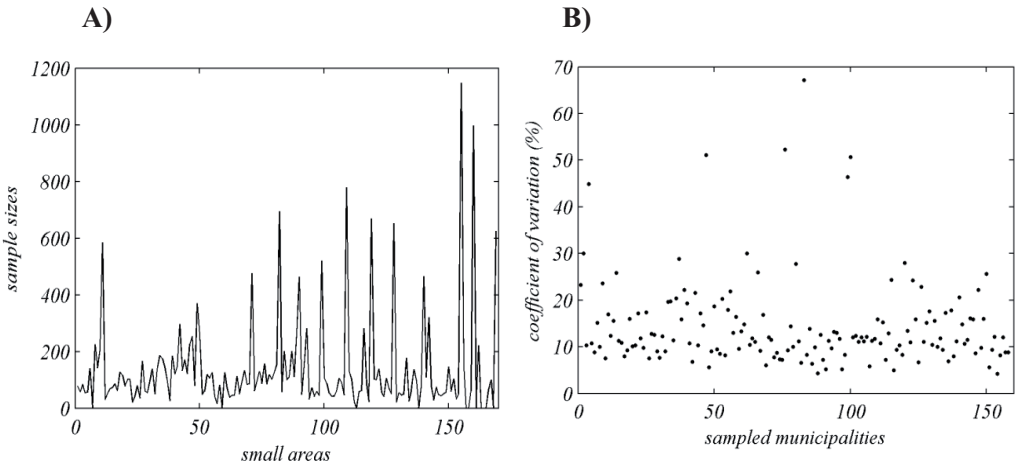


Figure 1. A) Sample size distribution between the small areas Cuban municipalities. B) Sample size vs. the coefficient of variation of the direct estimates for each sampled municipality.

Figure 1B shows that the direct estimates are not acceptable for some municipalities, particularly for those with a coefficient of variation greater than 20% (Sarndal et al., 1992). This confirms the need to find methods of producing more precise estimates than the direct estimates for such municipalities. The sampling plan yields a self-weighting sample within the stratum (province) which means that every individual has the same probability of being included in the sample within each province. Thus, all the municipalities within a specific province will also have the same individual sample weight. Taking into account the self-weighting sample property, the direct estimator of the monthly mean income in the municipality *i* is given by:

$$\hat{\bar{y}}_i = \frac{\sum_{j,k,l,h\in S_i} w_t y_{jklh}}{\sum_{j,k,l\in S} w_t} = \frac{\sum_{j,k,l,h\in S_i} y_{jklh}}{n_i} \tag{1}$$

where n_i is the sample size of the municipality *i* and $j,k,l,h \in S_i$ denotes the sampled individual *h* within the section *l* within the block *k* within the AGEM *j* within the municipality *i*. By using this sampling plan, the direct estimator in each municipality coincides exactly with the sample mean. This estimator will be compared with the two small area estimators that are used later in this article.

The calculation of the variance estimate of the direct means under the HRF sampling design is difficult since it requires the knowledge of the first and second-order inclusion probabilities. For simplicity, the following variance estimator is used

$$\hat{MSE}\left(\hat{\bar{y}}_i\right) = \frac{n_t\left(N_t - n_t\right)\left(N_i - 1\right)\left(n_i - 1\right)}{{n_i}^2 N_t\left(N_t - 1\right)b_i} s_{yi}^2 \tag{2}$$

where N_t and n_t are the population and sample size in the province t respectively, n_i and N_i are the sample and population size of the municipality i respectively, $b_i = \frac{n_t}{N_t}\left(N_i - 1\right)\left[1 - \frac{\left(N_t - n_t\right)}{\left(N_t - 1\right)n_i}\right]$, and $s_{yi}^2 = \frac{1}{\left(n_i - 1\right)} \sum_{j,k,l,h \in S_i} \left(y_{jklh} - \hat{\bar{y}}_i\right)^2$.

The estimator (2) coincides with the MSE estimator of the direct mean for each municipality i when a simple random sampling without replacement (SRSWOR) is considered in the province t. The implication of ignoring clustering may lead to an underestimation of the true MSE and consequently the confidence intervals will be narrower than expected. In this application we accepted to take such a risk because the MSE estimates of the direct means obtained under the HRF sampling design will always be greater than the MSE estimates of the direct means under a SRSWOR plan, i.e. the design effects will always be greater than 1. Therefore, if the precision of a small area estimate is found to be better than the precision of the direct mean measured under SRSWOR plan, it will also be better than the precision of the direct mean measured under the HRF sampling design.

3. Small area estimators

This section aims to describe the SAE methods and their MSE. The direct estimator of monthly mean income in each of the 158 sampled municipalities (1) will be compared with the two small area estimators derived from the basic unit level model (synthetic-regression and EBLUP estimators).

The direct estimator and its MSE were both described in Section 2. By avoiding the notation system employed for each sampling unit, the direct estimator (sample mean) can be expressed as: $\hat{\bar{Y}}_{iD} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$, where the subscript i and j identify the municipality and the individual, respectively. The MSE estimator ($mse\left(\hat{\bar{Y}}_{iD}\right)$) was given in (2). The general synthetic estimator can be

expressed as: $\hat{\bar{Y}}_{iSR} = \bar{X}_i^t \hat{\beta}$ (3), where $\bar{X}_i$ is the *ith* population proportion of x_{ij}'s, and $\hat{\beta}$ may be estimated by using any likelihood estimation methods or a moment method. The synthetic estimator (3) can be considered a model-based estimator and the MSE of $\bar{X}_i^t \hat{\beta}$ can then be estimated by adopting this approach. Using the assumptions of the basic unit level model, the unknown true mean $\bar{Y}_i$ can be considered itself as a random variable with variance σ_υ^2 and expectation $\bar{X}_i^t \hat{\beta}$. The synthetic estimator $\bar{X}_i^t \hat{\beta}$ can also be treated as an independent random variable with the same expectation. Then, the MSE of $\bar{X}_i^t \hat{\beta}$, under such assumptions, is given by Goldring et al. (2005) as:

$$MSE\left(\bar{X}_i^t \hat{\beta}\right) = E\left(\bar{X}_i^t \hat{\beta} - \bar{Y}_i\right)^2 = \bar{X}_i^t V\left(\hat{\beta}\right) \bar{X}_i + \sigma_\upsilon^2 \tag{4}$$

An estimate of the MSE of $\bar{X}_i^t \beta$ can be obtained by substituting the restricted iterative generalized least squares (RIGLS) estimates $\hat{\sigma}_\upsilon^2$ and $\hat{V}\left(\hat{\beta}\right)$ for σ_υ^2 and $V\left(\hat{\beta}\right)$ respectively (4) (Goldstein, 1989). The EBLUP estimator is a weighted average of the "survey regression" estimator $\left[\hat{\bar{Y}}_{iD} + \left(\bar{X}_i - \bar{x}_i\right)^t \hat{\beta}\right]$ and the synthetic estimator $\bar{X}_i^t \hat{\beta}$. It is given by (Rao, 2003):

$$\hat{\bar{Y}}_i^{EBLUP} = \hat{\gamma}_i \left[\hat{\bar{Y}}_{iD} + \left(\bar{X}_i - \bar{x}_i\right)^t \hat{\beta}\right] + \left(1 - \hat{\gamma}_i\right) \bar{X}_i^t \hat{\beta} \tag{5}$$

where $\hat{\gamma}_i = \dfrac{\hat{\sigma}_\upsilon^2}{\hat{\sigma}_\upsilon^2 + \dfrac{\hat{\sigma}_e^2}{n_i}}$, $\left(\hat{\bar{Y}}_{iD}, \bar{x}_i\right)$ are the *i-th* area sample means, and $\hat{\sigma}_\upsilon^2$, and $\hat{\sigma}_e^2$ are the variances at level-2 and level-1 respectively. They are estimated using the RIGLS method. A set of MSE estimators can be used for the EBLUP estimator (5). Two such MSE estimators, the MSE estimator given by Prasad and Rao

(1990) (PR EBLUP MSE) and the MSE estimator proposed by Jiang, Lahiri and Wan (2002) (JLW EBLUP MSE), will be used in this article. The leading term in both MSE estimators is $g_{1i}\left(\hat{\sigma}_v^2,\hat{\sigma}_e^2\right)=\hat{\gamma}_i\left(\frac{\hat{\sigma}_e^2}{n_i}\right)$ (Prasad and Rao, 1990; Jiang, Lahiri and Wan, 2002).

4. Basic unit level model

Both the synthetic estimator and EBLUP estimator are derived from the basic unit level model. The basic unit level model was selected due to the continuous nature of the response variable (total individual monthly income) and because the auxiliary information is available at a small area level, which is an important prerequisite to produce small area estimates derived from this model. Of the many possible covariates available in the HRF questionnaire, those for which area means were available from the 2002 census results were chosen: sex, age, marital status and educational level. These four covariates are included into the model by applying a backward elimination procedure. The RIGLS method is used to estimate the fixed and random parameters of the selected model. Table 1 gives the summary statistics for three multilevel models of income behaviour in Cuba. The four level-1 predictors - sex, marital status, education and age - are specified as a set of six dummy indicator variables. The reference person is a single woman aged between 15 and 34 years old who lives in Havana City and has a low educational level (less than a high school education).

Table 1. Multilevel estimates for models of income behaviour

	Estimates
Fixed effects	
Level 1	
Intercept	122.760(14.808)
Female	-
Male	82.279(6.977)
Single	-
Married	22.943(7.381)
Less than high school	-
High school or above	125.383(7.532)
15-34	-
35-54	75.411(8.225)
55-74	48.322(10.078)
75+	29.086(15.663)
Level 2	
Havana City	-
Pinar del Rio	-73.122(19.735)
Havana Province	-96.729(19.127)
Matanzas	-101.810(20.413)
Villa Clara	-97.031(20.304)
Cienfuegos	-76.086(21.453)
Sancti Spiritus	-82.827(20.915)
Ciego de Avila	-98.394(21.365)
Camaguey	-87.239(21.291)
Las Tunas	-123.486(21.758)
Holguin	-99.007(20.712)
Granma	-136.161(20.837)
Santiago de Cuba	-106.606(22.646)
Guantanamo	-126.880(23.836)
Isla de la Juventud	-84.514(34.142)
Random effects	
Level 1	
Intercept, σ_e^2	276050.400(2589.162)
Level 2	
Intercept, σ_υ^2	562.424(268.720)

Standard deviation for each regression estimates are shown in brackets.

To test normality, a normal Q-Q plot and the Shapiro-Wilk's test for standardised residuals of the fitted model were examined. From the figure below, neither standardised municipal nor individual residuals are close to the $y = x$

line, which suggests that the normality assumption for the residuals does not hold. The Shapiro-Wilk's test also rejected the null hypothesis of normality for both residuals (p-value=0.0000).

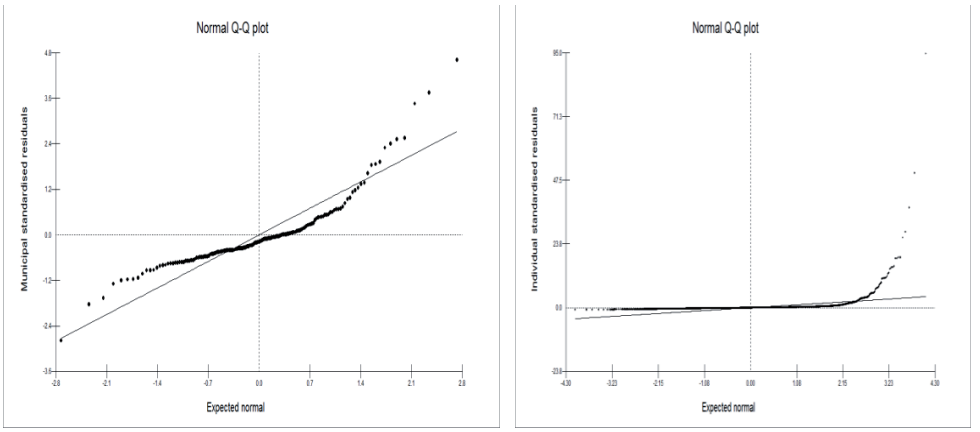


Figure 2. Municipal standardised residuals vs. expected normal (left-side) and individual standardised residuals vs. expected normal (right-side).

We relax the usual normality assumption for residuals in order to examine the robustness of the small area estimates derived from such a model. This means that the effects of the residuals' non-normality may have an impact on the small area estimators and such an impact may be even worse for the MSE small area estimators. This will be examined later in the simulation study (Section 5). Finally, the logarithmic transformation is widely used in models for income because the logarithms of values generated from a positive asymmetric distribution are generally more "normal" than the raw values. However, it was decided not to use such a transformation because almost 30% of individuals in the data set have reported a total income equal to zero, which might introduce a bias into the estimation of the small area estimators. Note that the logarithmic transformation effectively controls the influence of raw-scale outliers, but is then susceptible to log-scale outliers (e.g. either values near zero or variables that contain a significant proportions of zeros).

5. Simulation studies

This section details the results from two simulation experiments that enable the comparison of the three SAE methods described in Section 3. These simulation experiments also allow the performance of the MSE estimators for the best small area estimator to be evaluated. The first is a model-based simulation in which a small area population (municipality population) and sample data are generated from a distribution model. In this case, small area population and

sample data are based on the basic unit level model, assuming normal distributions for the two random effects. The second is a design-based simulation in which a fixed population is generated from existing samples and sample data are drawn from this fixed population. In this case, the fixed population is generated using the HRF sample data. The way to generate the population for the model-based and design-based simulation experiments are outlined in Section 5.1 and 5.2 respectively. The multilevel for Window (MLwiN) statistical software is used to implement all the small area estimators and their MSE estimators (see Appendice 1).

Since the residuals do not show evidence of normality (see the Normal Q-Q plot in Section 4), if there are different estimators' patterns for these two different simulation experiments, those found in the design-based simulation will be assumed to be more accurate and realistic than the other one. Therefore, the results from the model-based simulation will only be used for a simple empirical purpose. These, along with those from the design-based simulation, will enable an appraisal of the extent to which the small area estimators and their MSE estimators change under different population assumptions.

The Average Absolute Relative Bias (*AARB*) and the Average Mean Root Square Error (*ARMSE*) are used to evaluate the estimators' performance in the two simulation processes:

$$AARB = m^{-1}\sum_{i=1}^{m}\left|R^{-1}\sum_{r=1}^{R}\left(\frac{\hat{\bar{Y}}_{i(r)}}{\bar{Y}_{i(r)}}-1\right)\right| \qquad ARMSE = m^{-1}\sum_{i=1}^{m}\sqrt{R^{-1}\sum_{r=1}^{R}\left(\hat{\bar{Y}}_{i(r)}-\bar{Y}_{i(r)}\right)^{2}}$$

where $\hat{\bar{Y}}_{i(r)}$ and $\bar{Y}_{i(r)}$ are the small area estimator and the true mean for the municipality *i* at the simulation *r*, respectively. The gain in efficiency connected to each small area estimator is evaluated using the ratio of its *ARMSE* to the *ARMSE* of certain estimator that is used as benchmark. In particular, all estimators are compared with the direct estimator, and this ratio is denoted as $AEFF_{dir}$. Empirical results on model-based and design-based simulation processes are given in Sections 5.1 and 5.2 respectively.

5.1 Model-based simulation

This section compares the performance of the three small area estimators (direct, synthetic, and EBLUP) described earlier in Section 3. This comparison is made through a model-based simulation, by using the two evaluation measures (*AARB* and *ARMSE*) mentioned in Section 5. The population values $y_{ij(r)}$ were independently generated for each of the 158 municipalities at the simulation *r* from the following basic unit level model:

$$y_{ij(r)} = 122.76 + 82.279x_{ij1(r)} + 22.943x_{ij2(r)} + 125.383x_{ij3(r)} + 75.411x_{ij4(r)} + 48.322x_{ij5(r)}$$
$$+ 29.086x_{ij6(r)} - 73.122x_{i7} - 96.729x_{i8} - 101.81x_{i9} - 97.031x_{i10} - 76.086x_{i11} - 82.827x_{i12}$$
$$- 98.394x_{i13} - 87.239x_{i14} - 123.486x_{i15} - 99.007x_{i16} - 136.161x_{i17} - 106.606x_{i18}$$
$$- 126.880x_{i19} - 84.514x_{i20} + \upsilon_i + e_{ij}$$

where the area effects υ_i and individual effects e_{ij} were independently drawn from $N(0,\sigma_\upsilon^2)$ and $N(0,\sigma_e^2)$ respectively, the values $\sigma_\upsilon^2 = 562.424$ and $\sigma_e^2 = 276050.4$ were obtained from the model fitted to the HRF data (Table 1). For each simulation *r*, the level-1 variables $\left(x_{ij1(r)}, x_{ij2(r)}, x_{ij3(r)}, x_{ij4(r)}, x_{ij5(r)} \text{ and } x_{ij6(r)}\right)$ are generated from Bernoulli distribution with probability equal to their corresponding population proportions and using a population size of the municipality *i* equal to N_i. This population size is obtained from the individual's sample weight, $N_i = n_i \frac{N_t}{n_t}$, with n_i the sample size of the municipality *i*, and N_t and n_t are the population and sample size in the province *t* (the province where the individual lives), respectively. The level-2 variables $\left(x_{i7}, ..., x_{i20}\right)$ indicate which province each individual belongs to. The population mean for each municipality at the simulation *r* was then calculated as $\bar{Y}_{i(r)} = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij(r)}$. This manner of calculating the population size for each small area guarantees that the population size of the pseudo-population coincides with the real population size of the population of interest (Cuban population aged 15 and over). The simulation itself was model-based, so the population values $y_{ij(r)}$ were independently regenerated at each simulation and an independent sample for each municipality was then generated from these population values each time. Within each municipality, a SRSWOR design was used to generate the sample with the sample size n_i equal to the sample size obtained from the HRF survey. The same procedure was repeated 200 times ($r = 1,...,200$). For each simulation, the three estimators (direct, synthetic, and EBLUP) were computed, along with their MSE estimators. Over the 200 simulations, the *AARB* and *ARMSE* were calculated as in Section 5. Table 2 contains the values of *AARB*, *ARMSE*, and $AEFF_{dir}$ obtained for the direct estimator, the synthetic estimator, and the EBLUP estimator:

Table 2. Performance indicators of the three small area estimates.

Estimators	*AARB*	*ARMSE*	$AEFF_{dir}$ (%)
Direct	0.019	57.90	100
Synthetic	0.018	27.08	46.77
EBLUP	0.015	25.56	44.14

Table 2 shows that the synthetic and EBLUP estimates perform significantly better than the direct estimates, leading to less than 100% $AEFF_{dir}$ values. This gain in efficiency of about 55% ($\approx 100 - 44.14$) may be due to the use of different auxiliary information from the Cuban census. In terms of bias, all three small area estimators perform very well and there is no significant difference between them. The performance of each small area estimate can be also seen graphically for each of the 158 sampled municipalities (Figure 3A). The sampled municipalities are arranged in ascending order, according to the true mean of the monthly mean income $\overline{Y}_i = \frac{1}{200}\sum_{r=1}^{200} \overline{Y}_{i(r)}$. The true mean for each municipality is also plotted. In general, there is little difference between the three small area estimates.

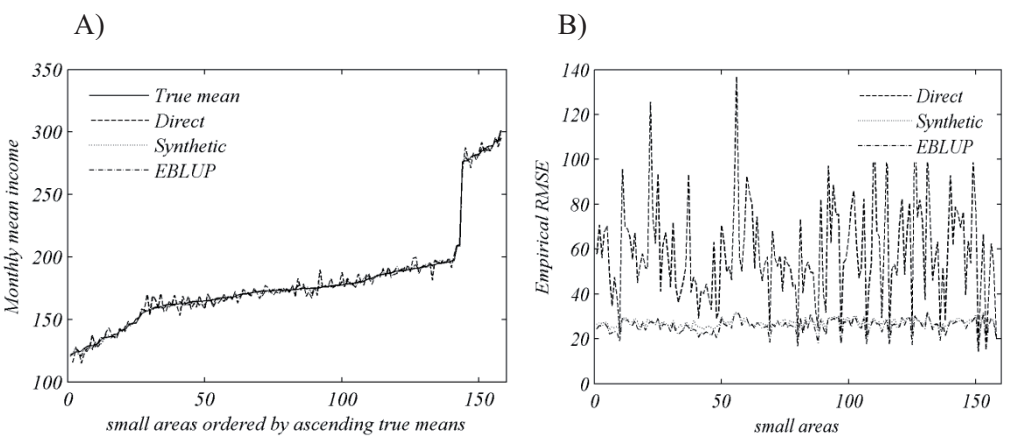


Figure 3. A) Estimated means of each small area estimates and the true mean for each of the 158 sampled small areas. B) Empirical RMSE of each small area estimates for each of the 158 sampled small areas.

Figure 3B represents the performance of all small area estimators in terms of its accuracy. The *RMSE* of the EBLUP estimate for each small area is smaller than the direct *RMSE* and slightly better than the synthetic *RMSE*. Therefore, the EBLUP estimator is more efficient than the direct estimator and is slightly more efficient than the synthetic estimator.

5.1.1 Assesing the perfromance of the eblup mse estimators

Since the EBLUP estimator emerged as the best performer in the previous section, this section aims to assess the performance of the EBLUP MSE estimators. It will compare the performance of the two EBLUP MSE estimators described in Section 3, the PR EBLUP MSE estimator and the JLW EBLUP MSE estimator. The following two measures are used to evaluate the performance of the EBLUP MSE estimators:

$$AARB = m^{-1}\sum_{i=1}^{m}\left| R^{-1}\sum_{r=1}^{R}\left(\frac{mse_{*}\left(\hat{\bar{Y}}_{i(r)}^{EBLUP}\right)}{mse_{EMP}\left(\hat{\bar{Y}}_{i}^{EBLUP}\right)} - 1 \right) \right|$$

$$ARMSE = m^{-1}\sum_{i=1}^{m}\sqrt{R^{-1}\sum_{r=1}^{R}\left(mse_{*}\left(\hat{\bar{Y}}_{i(r)}^{EBLUP}\right) - mse_{EMP}\left(\hat{\bar{Y}}_{i}^{EBLUP}\right)\right)^{2}}$$

where $mse_{EMP}\left(\hat{\bar{Y}}_{i}^{EBLUP}\right) = R^{-1}\sum_{r=1}^{R}\left(\hat{\bar{Y}}_{i(r)}^{EBLUP} - \bar{Y}_{i(r)}\right)^{2}$ is the empirical MSE and will be used as benchmark to compare the performance of each EBLUP MSE estimator. The symbol * refers to the type of EBLUP MSE estimator, i.e., PR, and JLW. As in the case of the small area estimator, the gain in efficiency connected to EBLUP MSE estimator is also evaluated, using the ratio between the corresponding *ARMSE*, that is $AEFF_{JLW} = \frac{ARMSE_{PR}}{ARMSE_{JLW}}$. The Average Relative Bias (*ARB*) is also used to provide a better understanding of whether the given EBLUP MSE estimators tend to underestimate the MSE or not,

$$ARB = m^{-1}\sum_{i=1}^{m} R^{-1}\sum_{r=1}^{R}\left(\frac{mse_{*}\left(\hat{\bar{Y}}_{i(r)}^{EBLUP}\right)}{mse_{EMP}\left(\hat{\bar{Y}}_{i}^{EBLUP}\right)} - 1 \right)$$

Table 3 shows the values of *AARB*, *ARB*, and $AEFF_{JLW}$ for each of the EBLUP MSE estimators. In terms of bias, both the mse_{PR} and the mse_{JLW} have a good performance. The mse_{PR} slightly underestimates the true MSE (4.1%), whereas the mse_{JLW} overestimates the true MSE in only a 4.8%. This similar behaviour for both EBLUP MSE estimators is also corroborated for each small area (Figure 4). It can be seen that the values of both EBLUP MSE estimators are concentrated on the straight line.

Table 3. Performance of each EBLUP MSE estimators.

Estimators	*AARB*	*ARB*	$AEFF_{JLW}$ (%)
mse_{JLW}	0.048	0.01	100
mse_{PR}	0.041	-0.02	90.80

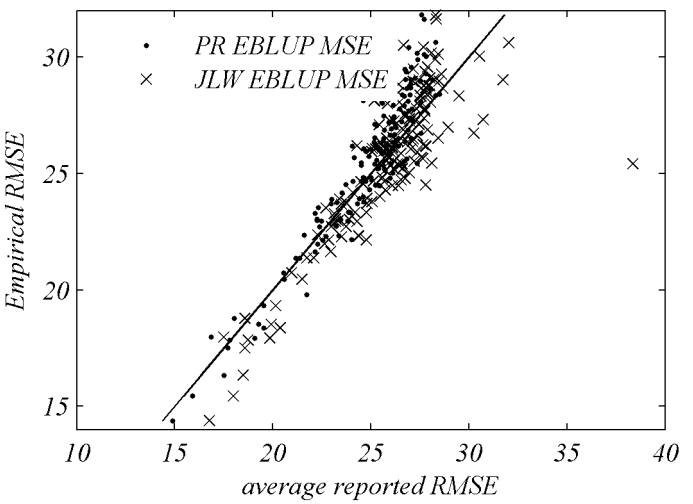


Figure 4. Empirical *RMSE* of the EBLUP estimator versus averaged reported *RMSE* of the EBLUP MSE estimators (mse_{PR} and mse_{JLW}) for each small area.

In terms of accuracy ($AEFF_{JLW}$), the mse_{PR} performs slightly better than the mse_{JLW}, offering a gain in efficiency of about 10% ($\approx 100 - 90.80$). For this reason, the mse_{PR} emerges as the most appropriate of the two EBLUP MSE estimators. However, this conclusion must be taken with caution because we assume normality for both residuals to generate the population, even though the residuals obtained from the model fitted to the HRF data did not show evidence of normality (see the Normal Q-Q plot in Section 4).

5.2 Design-based simulation

This section aims to compare the performance of three small area estimators through a design-based simulation and will also assess the performance of the MSE estimators for the best small area estimator. Due to the fact that individual income is not measured by the Cuban census, a pseudo-population was generated

using the HRF data and then samples were drawn using the original sampling design. The population values y_{ij} were obtained by bootstrapping the HRF data - which means that a re-sampling with replacement was carried out in each small area (158 sampled municipalities). The population size N_i for each small area is determined by the individual's sample weights, i.e. $N_i = n_i \frac{N_t}{n_t}$. In contrast to the model-based population, the pseudo-population in this case is fixed for each simulation. The true mean for the small area *i* is calculated as: $\overline{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$. Two hundred simulations were carried out. Two hundred independently stratified random samples, with sample size n_i equal to the original sample, were selected without replacement within each small area. Each small area estimator (direct, synthetic and EBLUP), and their corresponding MSE were computed from each of the selected areas. Table 4 contains the values of *AARB*, *ARMSE*, and $AEFF_{dir}$ obtained for the direct estimator, the synthetic estimator, and the EBLUP estimator.

Table 4. Performance indicators of the three small area estimates.

Estimators	*AARB*	*ARMSE*	$AEFF_{dir}$ (%)
Direct	0.012	29.07	100
Synthetic	0.203	39.61	136.25
EBLUP	0.128	30.39	104.55

In terms of accuracy, the direct estimates (29.07) perform very similarly to the EBLUP estimates (30.39), whereas the average accuracy of the synthetic estimates increases to 39.61. According to the bias criterion (*AARB*), the EBLUP estimates have the second best performance, whereas the synthetic estimates have the worst performance (Table 4). The excellent performance of the direct estimates in this case could be attributed to the fact that reliable information can be produced using only the direct estimates for an important numbers of municipalities. In other words, the small areas with smaller coefficients of variation may have a considerable influence on the overall performance of the direct estimates over the 158 small areas. For this reason, it was decided to compare the overall estimators' performance by classifying the municipalities in three main groups: 1) 75% of the municipalities with the lowest direct *RMSE* ("*lowest 75%"*), 2) 10% of the municipalities with the highest direct *RMSE* ("*highest 10%"*), and 3) the remaining 15% of municipalities ("*75-90%"*). As a result of this, 119 out of 158 municipalities are in the *"lowest 75%"* group, 23 are

in the "*75-90%"* group, and 16 are in the "*highest 10%"* group. Table 5 gives the results obtained for each of these three groups of municipalities:

Table 5. Performance indicators for each of the three groups of municipalities.

Estimators	*AARB*			*ARMSE*			$AEFF_{dir}$ (%)		
	lowest 75%	***75-90%***	***highest 10%***	***lowest 75%***	***75-90%***	***highest 10%***	***lowest 75%***	***75-90%***	***highest 10%***
Direct	0.008	0.017	0.035	15.30	38.22	115.43	100	100	100
Synthetic	0.197	0.149	0.327	29.96	36.90	114.39	195.81	96.54	99.09
EBLUP	0.123	0.096	0.209	20.80	32.24	97.68	135.94	84.35	84.62

The Table shows that the direct estimator performs better than the synthetic and the EBLUP estimators (both in terms of bias and accuracy) for 75% of the 158 sampled municipalities (the "*lowest 75%*" group). However, for the remaining 25% of the sampled municipalities (the "*75-90%*" group, and the "*highest 10%*" group), the EBLUP estimator is better than the direct estimator, offering a gain in efficiency of about 16% (84.35% for the "*75-90%*" group, and 84.62% for the "*highest 10%*" group). The EBLUP still performs well but relative to the direct there is not a remarkable change between the two groups of municipalities, i.e. the "*75-90%*" group and the "*highest 10%*" group, because both the MSE of the EBLUP estimator and the variance of the direct estimator increase at approximately the same rate for both groups of municipalities.

5.2.1 Assesing the perfromance of the eblup mse estimators

Since the EBLUP estimator consistently performs better than the direct estimator for those municipalities with the highest direct *RMSE* (the "*75-90%*" group, and the "*highest 10%*" group) under the *ARMSE* criterion (Table 5), this section will evaluate the performance of two EBLUP MSE estimators - the mse_{PR} and the mse_{JLW} - for each of these two groups of municipalities. Table 6 presents the results of the performance of the EBLUP MSE estimators:

Table 6. Performance of each EBLUP MSE estimator.

Estimators	*AARB*		*ARB*		$AEFF_{JLW}$ **(%)**	
	75-90%	***highest 10%***	***75-90%***	***highest 10%***	***75-90%***	***highest 10%***
mse_{JLW}	0.141	0.415	0.598	-0.306	100	100
mse_{PR}	0.115	0.565	0.339	-0.535	92.61	141.92

For the "*75-90%"* group, both MSE estimators overestimate the actual MSE, although mse_{PR} overestimates less than the mse_{JLW} (11.5% for the mse_{PR} and 14.1% for the mse_{JLW}). This slight overestimation by both EBLUP MSE estimators may be because, for the vast majority of the municipalities in the "*75-90%*" group, $\hat{\sigma}_e^2$ largely overstates the actual variation in the data (see the left side of Figure 6) which may yield an overestimation of the leading term of both EBLUP MSE estimators $g_{1i}\left(\hat{\sigma}_v^2,\hat{\sigma}_e^2\right)=\hat{\gamma}_i\left(\frac{\hat{\sigma}_e^2}{n_i}\right)$. In terms of accuracy, the mse_{PR} is slightly better than the mse_{JLW} ($AEFF_{JLW}(\%)=92.61$), leading to a gain in efficiency of about 8%. For the "*highest 10%*" group, the results are different. In terms of bias, both EBLUP MSE estimators severely underestimate the actual MSE, although the mse_{JLW} underestimates less than the mse_{PR} (Figure 6). This underestimation is of 41.5% for the mse_{JLW} and 56.5% for the mse_{PR}. This is because, for the majority of these municipalities, $\hat{\sigma}_e^2$ severely underestimates the actual variation in the data (see the right side of Figure 5). In terms of accuracy, the performance of the mse_{JLW} is much better than the performance of the mse_{PR} ($AEFF_{JLW}(\%)=141.92$).

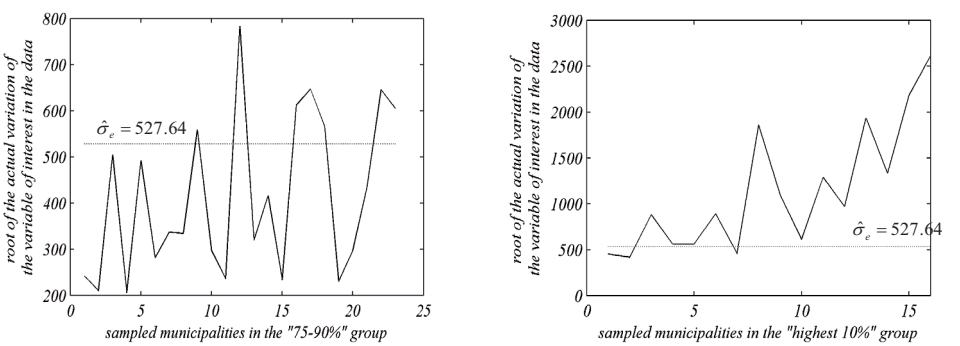


Figure 5. Root of the actual variability of the variable of interest in the data versus the root of the within area variance ($\hat{\sigma}_e$).

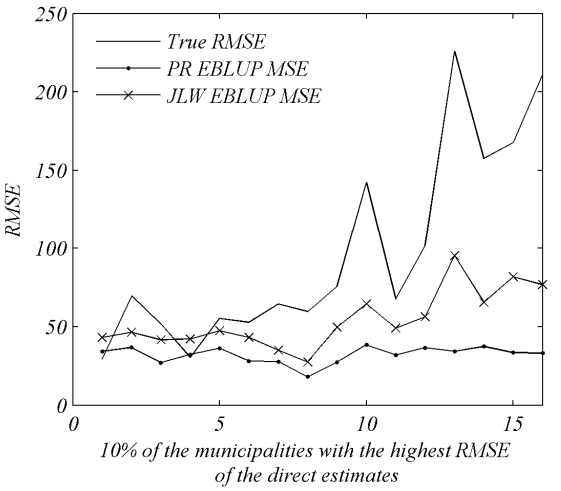


Figure 6. *RMSE* for 16 sampled municipalities in the "*highest 10%"* group.

These findings are consistent with both theoretical predictions and the simulation results as described in Jiang et al. (2002). This overestimation (underestimation) of both EBLUP MSE estimators for the municipalities in the "*75-90%*" group (the "*highest 10%*" group) is probably due to the fact that the failure of normality of the residuals causes the overestimation (underestimation) of $\hat{\sigma}_e^2$. In general, since there is a slight difference in the performance of both EBLUP MSE estimators for the municipalities that are in the "*75-90%*" group and this difference increases significantly for the municipalities that are in the "*highest 10%*" group, it is evident that the mse_{PR} are more sensitive to the non-normality of the residuals than the mse_{JLW}. Similar results were found by Fabrizi et al. (2007).

To conclude, it is suggested that the EBLUP estimates should be used to produce official statistics instead of the direct estimates for 25% of the 158 sampled municipalities (the "*75-90%*" group and the "*highest 10%*" group). Furthermore, for the 23 municipalities in the "*75-90%*" group, the mse_{PR} should be used as a measure to evaluate the quality of the information, and the mse_{JLW} should be used for the 16 municipalities in the "*highest 10%*" group. These findings may be closer to reality than the results from the model-based simulation.

6. Small area application

The core idea of this section aims to produce more precise estimates than the direct estimates of the monthly mean income at a municipal level. The 169 Cuban municipalities were considered as small areas, 158 of which were represented in the HRF sample and the remaining 11 were non-sampled municipalities. As posited in section 5 above, the design-based simulation results may be more realistic than the model-based simulation results. It was therefore decided to build the following final estimate (FE_i) for each *ith* small area (municipality):

$$FE_i = \begin{cases} \hat{\bar{Y}}_{iD} & \text{if the small area } i \text{ is in the "lowest 75\%" group} \\ \hat{\bar{Y}}_i^{EBLUP} & \text{if the small area } i \text{ is in either the "75-90\%" group or the "highest 10\%" group} \\ \hat{\bar{Y}}_{iSR} & \text{if the small area } i \text{ is a non-sampled area} \end{cases}$$

and the MSE estimator of FE_i is given by:

$$mse\left(FE_i\right) = \begin{cases} mse\left(\hat{\bar{Y}}_{iD}\right) & \text{if the small area } i \text{ is in the "lowest 75 \%" group} \\ mse_{PR}\left(\hat{\bar{Y}}_i^{EBLUP}\right) & \text{if the small area } i \text{ is in the "75 - 90 \%" group} \\ mse_{JLW}\left(\hat{\bar{Y}}_i^{EBLUP}\right) & \text{if the small area } i \text{ is in the "higest 10 \%" group} \\ mse\left(\hat{\bar{Y}}_{iSR}\right) & \text{if the small area } i \text{ is a non - sampled area} \end{cases}$$

where $\hat{\bar{Y}}_{iD}$, $\hat{\bar{Y}}_{iSR}$, and $\hat{\bar{Y}}_i^{EBLUP}$ are given in (1), (3), and (5) respectively; and $mse\left(\hat{\bar{Y}}_{iD}\right)$, $mse\left(\hat{\bar{Y}}_{iSR}\right)$, $mse_{PR}\left(\hat{\bar{Y}}_i^{EBLUP}\right)$, and $mse_{JLW}\left(\hat{\bar{Y}}_i^{EBLUP}\right)$ are described in section 3. The "*lowest 75%*" group, the "*75-90 %*" group, and the "*highest 10 %*" group were defined previously in Section 5.2. The final proposal is that the estimator, FE_i, along with its MSE estimator, $mse\left(FE_i\right)$, should be published in the future rather than the traditional direct estimator and its MSE estimator. They also permit provision of information of the monthly mean income, not only for the 158 sampled municipalities, but also for the 11 non-sampled municipalities. The gain in efficiency connected to each final municipality estimator FE_i is evaluated using the ratio of its root MSE to the root MSE of the direct estimator. Figure 7 shows the relative efficiency for the 23 municipalities that are in the "*75-*

90%" group (left side), and the 16 municipalities that are in the "*highest 10%*" group (right side). For the remaining 119 sampled municipalities in the "*lowest 75%*" group, the relative efficiency is equal to 1.

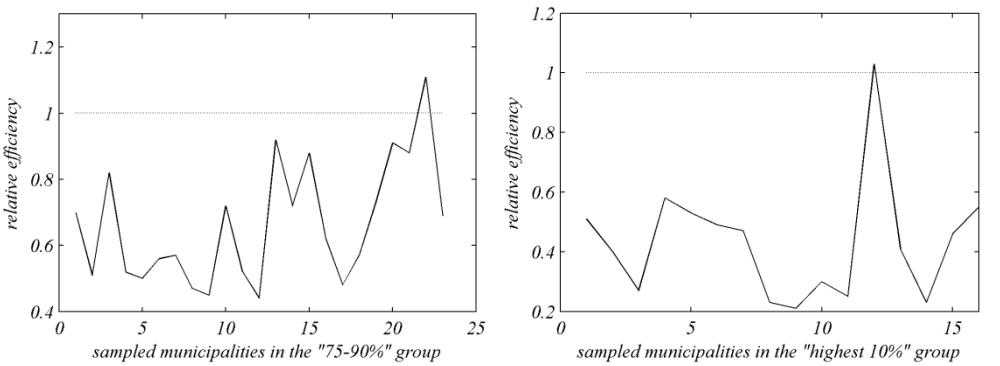


Figure 7. Relative efficiency of the "Final estimator", FE_i, with respect to the direct estimator, $\hat{\bar{Y}}_{iD}$.

The precision of the 119 sampled municipalities that are in the "*lowest 75%*" group could not be improved because the direct estimates of the monthly mean income perform acceptably well, i.e. the coefficient of variation of most of these municipalities is less than 20%. For the sampled municipalities in the "*75-90%*" group, the performance of the final estimates is observed to be better than the direct estimates in 22 out of 23 municipalities in this group. The final estimates are still better than the direct estimates even though the $mse_{PR}\left(\hat{\bar{Y}}_i^{EBLUP}\right)$, which the EBLUP MSE estimator used for these municipalities, overestimates the actual EBLUP MSE. For the 16 municipalities in the "*highest 10%*" group (right side of Figure 8), the final estimates perform better than the direct estimates in 15 of the 16, i.e. the relative efficiencies of the final estimates with respect to the direct estimates are smaller than 1. However, this assessment must be regarded with caution, considering that the design-based simulation results showed that the $mse_{JLW}\left(\hat{\bar{Y}}_i^{EBLUP}\right)$, which the EBLUP MSE estimator used for these municipalities, underestimates the actual EBLUP MSE, and therefore this relative efficiency is also underestimated. It is therefore recommended that the behaviour of other EBLUP MSE estimators should be investigated, particularly in cases where the normality assumption of the residuals does not hold. For example, different EBLUP MSE estimators can be considered to experiment with first, such as those given by the following authors: Buttar and Lahiri (2003), Pfeffermann and Glickman (2004), and Hall and Maiti (2006).

7. Final remarks and further developments

Some criticisms could be made of the results discussed in this article. The production of SAE methods depend upon two main factors, which are both addressed to find suitable models that fit the data well. The first factor relates to the quality and relevance of each of the sample data variables used to fit the model, and the second factor relates to the availability and quality of different alternative data sources, which the auxiliary information related to the target variable is selected from. With respect to the first factor, no consideration was made of the presence of outliers of the monthly total income of each individual that may affect the production of small area estimates under the basic unit level model. It is well documented that the least squares estimator of β in the model and ML or REML estimators for the variance components are sensitive to outliers (Richardson and Welsh, 1995). In order to obtain robust small area estimation with the presence of outliers we recommend analysing the feasibility of applying the methods proposed by Sinha and Rao (2009), and Chambers and Tzavidis (2006) to the HRF data. The second reason mentioned above is linked to the availability of the auxiliary information related to the variable of interest. This article chose only those variables for which municipality means were available from the 2002 Cuban census results: sex, age, education level and marital status. However, taking into account the complexity of the national Cuban monetary system (two currencies), it is advisable to keep searching for more appropriate supplementary variables – specifically economic indicators – which would improve the direct estimators, such as the proportion of workers at a small area level. It would be useful to list all the plausible variables related to income and make an inventory of the different information sources where these variables appear. The Ministry of Labour and Social Security of Cuba may be the main provider of this type of information at a small area level.

Another relevant criticism is related to the variance estimates of the direct means. In this application, the variance estimator (2) was used to estimate the precision of the direct means. This variance estimator ignores the probabilistic three-stage cluster design employed in HRF survey which may lead to a severe underestimation of the true variance. A more accurate procedure to overcome this problem could have been either to use the variance estimates obtained by applying the ultimate cluster method of variance estimation given by Hansen, Hurwitz and Madow (1953), or to develop methods to adjust the variance estimator for design effects by considering information from previous studies. A further important criticism is linked to the EBLUP MSE estimators. In this article, neither of the two EBLUP MSE estimators (mse_{PR}, and mse_{JLW}) appear to be close to the true variability when the normality assumption of the residuals does not hold. It would therefore be advisable to keep working on this topic in order to locate more robust and accurate MSE estimators. For example, as noted above, the parametric double

bootstrap technique posited by Hall and Maiti (2006) may be a possible starting point.

Despite these criticisms, this article proposes that this SAE application could constitute an important first step for extending these techniques in Cuba. The results are only meant to provide a first impression of the utility of SAE methods. More research is needed to develop a generic SAE methodology in Cuba. Finally, the implementation of the MLwiN code for this application (Appendix 1) shows the feasibility of using this statistical software in small area applications. This statistical code may be generalised to a standardized subroutine which serves to obtain different small area estimates derived from a basic unit level model and their MSE estimates.

Appendix 1. Main macro implemented in MlwiN software to calculate the two SAE methods (synthetic, EBLUP) and their corresponding MSE estimators (synthetic MSE estimator, PR EBLUP MSE estimator, and JLW EBLUP MSE estimator).

```
note running the model
resp c647
batc 1
stop
tole 2
star
note covariate matrix
obey d:/simfr/covmatrix.txt
note population mean covariate matrix
obey d:/simfr/popmatrix.txt
note sample mean covariate matrix multiply by sigma
pick 1 c1096 b1
pick 2 c1096 b2
calc c614=b1/(b1+(b2/c1105))
obey d:/simfr/samplematrix.txt
note EBLUP=EB estimates
calc c620=c1098
matr c620 21 1
calc c621=c600*c614
matr c621 158 1
calc c622=c621+(c619-c615)*.c620
note synthetic estimates
calc c623=c619*.c620
note synthetic MSE estimates
eras c599
pick 1 c1096 b1
calc c1100=sym(c1099)
obey d:/simfr/synthetic.txt
note PR EBLUP MSE estimates
note calcucale g1
calc c625=(1-c614)*b1
note calculate g2
gene 1 3318 1 c626
calc c627=c619-c615
eras c632
loop b55 1 158
calc c628=c626
obey d:/simfr/g2.txt
endl
note calculate g3
obey d:/simfr/g3.txt
note JLW EBLUP MSE estimates
put 158 0 c1051; put 158 0 c1049
loop b49 1 158
calc c5000=c1050==b49
omit 1 c5000 c5 c4999 c1020
omit 1 c5000 c14 c4999 c1021
omit 1 c5000 c15 c4999 c1022
omit 1 c5000 c18 c4999 c1023
omit 1 c5000 c21 c4999 c1024
omit 1 c5000 c22 c4999 c1025
omit 1 c5000 c23 c4999 c1026
omit 1 c5000 c26 c4999 c1027
omit 1 c5000 c27 c4999 c1028
omit 1 c5000 c28 c4999 c1029
omit 1 c5000 c29 c4999 c1030
omit 1 c5000 c30 c4999 c1031
omit 1 c5000 c31 c4999 c1032
omit 1 c5000 c32 c4999 c1033
omit 1 c5000 c33 c4999 c1034
omit 1 c5000 c34 c4999 c1035
omit 1 c5000 c35 c4999 c1036
omit 1 c5000 c36 c4999 c1037
omit 1 c5000 c37 c4999 c1038
omit 1 c5000 c38 c4999 c1039
omit 1 c5000 c39 c4999 c1040
omit 1 c5000 c1050 c4999 c1043
pick b49 c1105 b1
calc b2=22851-b1
put b2 1 c1041
gene 1 b2 1 c1042
note build the model with the deleted observations
clea
resp c1020
iden 2 c1043
iden 1 c1042
expl 1 c1041
setv 2 c1041
setv 1 c1041
```

```
note calc PR EBLUP MSE estimates = g1+g2+2*g3
calc c634=c625+c632+2*c633
```

```
addt 'c1021'; addt 'c1022'; addt 'c1023';
addt 'c1024'; addt 'c1025'; addt 'c1026';
addt 'c1027'; addt 'c1028'; addt 'c1029';
addt 'c1030'; addt 'c1031'; addt 'c1032';
addt 'c1033'; addt 'c1034'; addt 'c1035';
addt 'c1036'; addt 'c1037'; addt 'c1038';
addt 'c1039'; addt 'c1040'
note running the model
tole 2
meth 0
batc 1
star
pick 1 c1096 b3
pick 2 c1096 b4
calc c1044=b3/(b3+(b4/c1105))
calc c1045=(1-c1044)*b3
calc c1051=c1051+(c1045-c625)
calc c614=c1044
obey d:/simfr/samplematrix.txt
calc c1046=c1098
matr c1046 21 1
calc c1047=c600*c614
matr c1047 158 1
calc c1048=c1047+(c619-c615)*.c1046
calc c1049=c1049+(c1048-c622)^2
endl
note JLW EBLUP MSE estimates
calc c1051=(157/158)*c1051
calc c1049=(157/158)*c1049
calc c1053=c625-c1051+c1049
```

REFERENCES

BUTTAR, F., AND LAHIRI, P. (2003). On measures of uncertainty of empirical bayes small-area estimators. *Journal of Statistical Planning and Inference,* 112, 63-76.

CHAMBERS, R., and TZAVIDIS, N. (2006). "M-quantile models for small area estimation". *Biometrika* 93(2):255-268.

FABRIZI, E., FERRANTE, M.R., and PACEI, S. (2007). Small area estimation of average household income based on unit level models for panel data. *Survey Methodology*, vol. 33, No. 2, pp. 187-198. Statistics Canada, Catalogue no.12-001-X.

GOLDRING, S., LONGHURST, J., and CRUDDAS, M. (2005). Model-Based Estimates of Income for wards, 2001/2002. Technical Report. *Office for National Statistics* (ONS).

GOLDSTEIN, H. (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika*, 76: 622-623.

HALL, P., and MAITI, T. (2005). Nonparametric estimation of mean-squared prediction error in nested-error regression models. Sep. http://arxiv.org/abs/math/0509493.

HANSEN, M.H., HURWITZ, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory. *New York, Wiley*.

JIANG, J., LAHIRI, P., AND WAN, S.M. (2002). A unified jackknife theory for empirical best prediction with *M*-estimation. *The Annals of Statistics*, 30, 1782-1810.

ONE. (2005). Diseño Muestral General del sistema de encuesta de hogares." Havana. *Oficina Nacional de Estadísticas*.

PFEFFERMAN, D., AND GLICKMAN, H. (2004). Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap. *ASA Proceedings of the Survey Research Methods Section.*

PRASAD, N.G.N., AND RAO, J.N.K (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

RAO, J.N.K. (2003). Small area estimation. *Wiley series in survey methodology*.

RICHARDSON, A.M., and WELSH, A.H. (1995). Robust Restricted Maximum Likelihood in Mixed Linear Models. *Biometrics*, 51, 1429-1439.

SARNDAL, C., SWENSSOn B., and WRETMAN J. (1992). Model assisted survey sampling. *Springer Series in Statistics*.

SINHA, S.K., and RAO, J.N.K. (2009). Robust Small Area Estimation. *The Canadian Journal of Statistics*, to appear.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 81—96

BOOTSTRAP METHOD WITH CALIBRATION FOR STANDARD ERROR ESTIMATORS OF INCOME POVERTY MEASURES

Agnieszka Zięba-Pietrzak[1], Jan Kordos[2], Robert Wieczorkowski[3]

ABSTRACT

The authors begin with calibration approach in sample surveys, focussing on the Eurostat approach. Next, the indicators of poverty and social exclusion are discussed as an essential tool for monitoring progress in the reduction of these problems. Most of these indicators are calculated according to the Eurostat recommendations, using data from European Statistics on Income and Living Conditions (EU-SILC). Complex sample design of the EU-SILC requires weighted analyses for estimates of population parameters and approximate methods of standard error estimation. In our study McCarthy and Snowden (1985) bootstrap method for standard errors estimation of income poverty measures is presented. In the next step the reweighting of bootstrap weights is applied and results of such calibration are discussed.

Key words: Auxiliary information; Bootstrap, Calibration, Complex sample surveys, EU-SILC, Poverty indicators; Weighting.

1. Introduction

Calibration provides a systematic way to incorporate auxiliary information in the estimation procedure. Calibration has established itself as an important methodological instrument in large-scale sample surveys. Several national statistical agencies have developed software designed to compute weights, usually calibrated to auxiliary information available in administrative registers and other accurate sources. In this paper calibration approach has been applied to estimation of standard errors of poverty measures, using calibration approach accepted by Eurostat (2002). This approach is generally presented later.

Standard error estimation indicates precision of estimators of unknown parameters and it is one of the most important issues in statistical inference.

[1] WSE. E-mail: a.zieba.pietrzak@gmail.com.
[2] WSE. E-mail: jan1kor2@aster.pl.
[3] GUS. E-mail: R.Wieczorkowski@stat.gov.pl.

Derivation of suitable estimator of the variance of the given statistic is difficult in case of non-linear structure of estimated measures and complex survey design (Sitter, 1992). Income poverty indicators are such kind of measures and are calculated according to the Eurostat recommendations, using data from European Statistics on Income and Living Conditions EU-SILC (Eurostat, 2009). Using this survey it is necessary to take into consideration that the selection of the survey sample is done by two-stage stratified sampling with different selection probabilities at the first stage. In this case it is not possible to obtain a closed-form algebraic expression for the estimated standard error. For such measures approximate standard error estimation is required. Various methods for obtaining variance estimates have been proposed in the literature (Kordos, Zięba-Pietrzak, 2010). One of the most common methods is bootstrap – a method, which has become an important tool in sample surveys since first research article about it (Efron, 1979).

Bootstrap methodology for sample surveys application has been extended for stratified sampling, involving a large number of strata with relatively few primary sampling units within strata when the parameter of interest is a nonlinear function (Rao, Wu, 1984). For stratified multistage designs McCarthy and Snowden proposed procedure which is an asymptotically valid method in assessing the variability of direct estimators (McCarthy, Snowden, 1985). This procedure has been applied in Polish official statistics since 2003 in the Labour Force Survey (Popiński, 2006). It is also proper for standard error estimation for poverty measures and it has been applied in the Polish EU-SILC survey since 2005 (GUS, 2009; Zieba & Kordos, 2010).

In this study McCarthy and Snowden (1985) bootstrap method for standard error estimators of income poverty measures is presented. In the next step the reweighting of bootstrap weights is applied and results of this integrated calibration method are discussed.

2. Income poverty measures

Poverty indicators used to monitor social exclusion phenomena are calculated on the basis of data from EU-SILC. Basic aim of this survey is the study, both at national and European level, of households' living conditions mainly in relation to their income. Four most common poverty measures and two additional are being focused on in this work. All of them are calculated using *equivalised income distribution* and therefore are called *income poverty measures* (Eurostat, 2005).

Equivalised disposable income (EQ_INC*i*) is defined as the household's total disposable income divided by its "equivalent size" to take account of the size and composition of the household, and attributed to each household member (Eurostat 2005). The equivalised household size is defined according to the modified OECD (Organisation for Economic Co-operation and Development) scale which

gives a weight of 1.0 to the first adult, 0.5 to other household members aged 14 or over and 0.3 to household members aged less than 14. Finally, equivalised household size (*EQ_SS*) is calculated as:

$$EQ\_SS = 1 + 0{,}5 \cdot (HM_{14+} - 1) + 0{,}3 \cdot HM_{13-} \qquad (1)$$

where:

HM_{14+} – number of household members aged 14 and over
HM_{13-} – number of household members aged 13 or less.

The equivalence scales are the parameters which allow comparing the conditions of households of different sizes and different demographic structures.

The formulas for selected poverty indicators

1) MED (MED_{EQINC}) – *national median equivalised income* – it is a median in distribution of yearly equivalent net disposable income in households.
2) ARPR – *At-risk-of-poverty rate after social transfers* is defined as a share of poor persons in all persons. A poor person is defined as the one with an equivalised disposable income below the poverty threshold (ARPT)

$$ARPR = \frac{\sum_{i|EQINC_i \le ARPT} w_i}{\sum_{i=1}^{n} w_i} \qquad (2)$$

where:
w_i – the survey weight attached to i^{th} sample unit (individual),
ARPT – At-risk-of-poverty threshold after social transfers.

The poverty threshold ARPT equals 60% of the national median equivalised disposable income

$$ARPT = 0{,}6 \cdot MED_{EQINC} \qquad (3)$$

Disposable income is defined as a sum of net (after social transfers) monetary incomes gained by all the household members reduced by: property tax, inter-household cash transfers paid and statements for the Treasury Office.

3) RRPG – *Relative median poverty risk gap* is defined as a difference between the median equivalised disposable income of persons below the at-risk-of poverty threshold (MED_{POOR}) and the threshold itself (ARPT), expressed as a percentage of the at-risk-of poverty threshold:

4)

$$RRPG = \frac{ARPT - MED_{POOR}}{ARPT} \tag{4}$$

5) GINI – *Gini coefficient* measures inequality of income distribution and this is the relationship between cumulative shares of the population arranged according to the level of income and the cumulative share of the equivalised total income received by them.

To calculate this indicator persons have to be sorted according to EQINC from the lowest to the highest value. If w_i denote the sample weight of a unit *i-th sample unit* in the ascending income sorted distribution, then we have:

$$GINI = 100 \cdot \left(\frac{2 \cdot \sum_{i=first\_person}^{last\_person} \left(w_i \cdot EQINC_i \cdot \sum_{first\_person}^{person\_i} w_i \right) - \sum_{i=first\_person}^{last\_person} \left(w_i^2 \cdot EQINC_i \right)}{\left(\sum_{i=first\_person}^{last\_person} w_i \right) \cdot \sum_{i=first\_person}^{last\_person} \left(w_i \cdot EQINC_i \right)} - 1 \right) \tag{5}$$

where:

w_i – the survey weight attached to *i-th sample unit* (individual).

6) S80/S20 – *S80/S20 income quintile share ratio* measures inequality of income distribution – it is the ratio of sum of equivalised disposable income received by the 80% of the country's population with the highest income (top quintile) to that received by the 20% of the country's population with the lowest income (lowest quintile):

7)

$$S80/S20 = \frac{\sum_{i \mid EQINC_i \le Q_{80}} w_i EQINC_i}{\sum_{i \mid EQINC_i \le Q_{20}} w_i EQINC_i} \tag{6}$$

where:

w_i – the survey weight attached to *i-th sample unit* (individual).

Due to the way quintiles are calculated, the number of people with $EQINC_i \leq Q_{20}$ can differ from the number of people with $EQINC_i \geq Q_{80}$, in which case the following correction is necessary:

$$S80/S20 = \frac{\sum_{i|EQINC_i \geq Q_{80}} w_i EQINC_i \Big/ \sum_{i|EQINC_i \geq Q_{80}} w_i}{\sum_{i|EQINC_i \leq Q_{20}} w_i EQINC_i \Big/ \sum_{i|EQINC_i \leq Q_{20}} w_i} \tag{7}$$

8) MEAN_EQINC – *mean equivalised income* – it is an average yearly equivalent net disposable income in households.

9)

$$MEAN\_EQINC = \frac{\sum_{i=1}^{n} wiEQINC_i}{\sum_{i=1}^{n} wi} \tag{8}$$

where:
w_j – the survey weight attached to *i-th sample unit* (individual).

3. Sample design in the Polish EU-SILC

The Polish EU-SILC sample is two-stage stratified sampling with unequal selection probabilities at the first stage. The first-stage sampling units (primary sampling units – PSU) are enumeration census areas, while at the second stage dwellings are selected. All the households from the selected dwellings are supposed to enter the survey. The basis for creating strata is the size of a locality (measured by the number of dwellings), and strata are built up within a voivodship. A voivodship is NUTS2 level – a territorial system of classification worked out by Eurostat. According to the sample design, proper weights are constructed. Every statistic and its variability are estimated using these weights.

The selection of the EU-SILC sample is done by two-stage stratified sampling with unequal selection probabilities at the first stage. The first-stage sampling units (primary sampling units – PSU) are enumeration census areas, while at the second stage dwellings are selected. All the households from the selected

dwellings are supposed to enter the survey. The basis for creating strata is the size of a locality (measured by the number of dwellers), and strata are built up within a voivodship. A voivodship is NUTS2 level at The Nomenclature of Territorial Units for Statistics (NUTS) – a territorial system of classification worked out by Eurostat. There are 16 NUTS2 units in Poland.

In EU-SILC carried out by Central Statistical Office of Poland in 2008 the final sample contained 13 984 households with 41 200 persons (33 801 persons aged 16 and more were interviewed) from 5093 PSUs (primary sampling units) allocated in 211 strata.

Weights construction in Polish EU-SILC 2008

According to EU-SILC sample, structure proper weights are calculated. Final weights, DB090 (for households) and RB050 (for household members), are constructed in complex and multi-stage process (for more detail see GUS, 2009). During this process the *integrated calibration method* recommended by Eurostat is applied (Eurostat, 2004). It means that final weights are calculated with the use of demographic data from other sources. This method ensures consistency of the generalized results with the external demographic data available. Every statistic and its variability are estimated using these weights.

4. The bootstrap method (McCarthy and Snowden, 1985)

In Polish practice bootstrap method (McCarthy and Snowden, 1985) has been used in Polish official statistics since 2003 till present time (Kordos & Zieba-Pietrzak, 2010). This method makes possible treating different estimators uniformly and eliminates the necessity of derivation of complicated analytical formulas (Popiński, 2006). It is applied separately in each stratum and it is proper for stratified multistage sample design. For such design this bootstrap procedure is an asymptotically valid method in assessing the variability of direct estimators (Shao, 2003).

Applying this method, every bootstrap sample is obtained drawing with-replacement random sample of $a_h - 1$ PSUs out the a_h, sampled in each stratum h ($h = 1,2,...,L$). All items (secondary sampling units) from sampled PSUs are included in the new bootstrap sample. After every resampling, the original weights are properly rescaled:

$$w_j(b) = w_j \frac{a_h}{a_h - 1} m_j(b) \tag{9}$$

where:

a_h – number of PSUs in stratum *h*

$w_j(b)$ – weight for person from *j-th* household in *b-th* bootstrap sample,

w_j – original weight for person from *j-th* household,

$m_j(b)$ – number of times PSU from *j-th* household is included in b^{th} bootstrap sample ($b=1,2,...,B$).

The bootstrap variance estimate of the corresponding indicator is obtained by the usual Monte Carlo approximation based on the independent bootstrap replicates (the sampling procedure is replicated B times).

At national level variance of direct estimator is estimated in the following way:

$$V(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}^*)^2 \tag{10}$$

where:

$\hat{\theta}$ – the direct estimator of indicator

$\hat{\theta}_b^*$ – the bootstrap estimator obtained using modified weights from b^{th} bootstrap sample

$\hat{\theta}^*$ – is computed as: $\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$.

Finally, the estimator of standard error for direct estimator is defined as:

$$SE(\hat{\theta}) = \sqrt{V(\hat{\theta})}\,. \tag{11}$$

5. Bootstrap method with calibration of sampling weights

As it has been stressed at the beginning, calibration in its several forms has become an essential and popular tool in sample surveys. There are many reasons why the accuracy of the estimates should be improved by applying it. One of these reasons is that particular estimates of the calibrated data should coincide with known population totals (Gambino, 1999). In EU-SILC this issue is crucial also in case of subpopulations. The original totals, based on demographic variables such as age, sex, geographic location and household size should be estimated without error. The "sampling weights" are not enough for safeguarding this accuracy and calibration is necessary.

A mainstream practice is the use of auxiliary information (e.g. from the census) for the strengthening of the survey results (Eurostat, 2002). By applying calibration (reweighting of the sample weights) agreement with known population margins results is provided. This method is also called *post-stratification*. Appropriate calculations are made and each statistical unit (individual, household or business) is re-assigned a unique weight. Calibrated estimates are subsequently based on these weights.

5.1. Calibration algorithm used in EU-SILC

The principle of the calibration is the following: starting from an initial weighting variable, and auxiliary variables, final weighting variable is constructed, which will perfectly estimate the totals on whichever of the auxiliary variable in such a way that its weight values for each unit in the sample are as close as possible to the initial weights (Eurostat, 2004).

In basic idea of calibration it is assumed that the totals of the auxiliary variables (called calibration variables) over the whole population are known and come from sufficiently reliable external data sources such as, for instance, census.

The calibration procedure is to calculate new household weights (depending on the sample of households), called "calibrated" (or final) weights, such that two following properties are satisfied:

the "calibrated" weights are "not very different" from the initial weights (ratios between the final weights and the initial weights, called the g-weights, is "not very different" from 1),

the totals of the auxiliary variables are estimated *with zero variance* from the final weights (Eurostat, 2004).

From the mathematical point of view, a *"function of distance"* G (.) is considered. This function will control that the *g-weights* do not vary too much from 1. G is supposed to be (Deville et al., 1993):

positive,

continuously differentiable in a neighborhood of 1,

strictly convex,

- G'(*1*)=*0*.

Under these assumptions, the "calibrated" weights form (i.e. different methods) the following problem (*P*) of optimization subject to constraints:

$$(P)\begin{cases} \min \sum_{k \in S} d_k \cdot G\left(\frac{w_k}{d_k}\right) \\ \text{subject\_to}: \sum_{k \in S} w_k \cdot x_{rk} = X_r \quad for \;\; r = 1,\ldots, p \end{cases} \tag{12}$$

where:

d_k – the *k-th* household design weight (before calibration),

w_k – the *k-th* household final weight (after calibration),

$x_1, x_2 \ldots x_p$ – auxiliary variables,

x_{rk} – the value of the auxiliary variable *xr* (*r=1...p*) on the household *k*,

X_r – total of *r-th* auxiliary variable in population,

S – sample of households.

In order to derive calibration weights the significant number of matrix calculations should be made and specialized software is needed. For this need the calibration algorithm was implemented and applied using SAS software (GUS,

2007). The software finds a solution in iterative way, function G can take different form (different method can be used: the linear method, the raking ratio method, the "logit" method, the linear truncated method, hyperbolic sinus, etc.) (Eurostat, 2004).

In this work the hyperbolic sinus method was applied. In practice this method yields final weights very close to initial ones. The formula of function G is as follows:

$$G_\alpha(x) = \frac{1}{2\alpha}\int_1^x sh\left[\alpha\left(\frac{1-t}{t}\right)\right]dt \tag{13}$$

where:

α – parameter which is positive (it allows to control a degree of dispersion of calibrated weights in relation to original weights, its initial value equals 1)

sh(x) – the hyperbolic sinus function: $sh(x) = \frac{\exp(x) - \exp(-x)}{2}$.

5.2. Integrated calibration method

The idea of integrated calibration method, recommended by Eurostat, is to use calibration variables defined at both household and individual levels. The individual variables are aggregated at the household level by calculating household totals (such as the number of males/females in the household, the number of persons aged of 16 and over, etc.). After that, the calibration is done at the household level using the household variables and the individual variables in their aggregate form. This technique ensures consistency between households and individual estimates because (see the next step) all household members will have the same household cross-sectional weight as personal cross-sectional weight.

The weights obtained after this procedure of integrated calibration are the household cross-sectional weights (target variable DB090). The personal cross-sectional weight RB050 is equal for all the members of a given household and to the household cross-sectional weight DB090: (Eurostat, 2004).

Auxiliary information in Polish EU-SILC

In EU-SILC calibration procedure the additional variables comprised data on the number of households according to 4 size classes (1-person, 2-person, 3-person and 4 and more person household) in voivodship (NUTS2 regions) breakdown and place-of-residence (urban/rural). As regards individuals, the data were presented as the number of persons by gender and 14 age groups (under 16 years, 16-19 years, eleven 5-year groups, 75 years and over) in voivodship breakdown.

These additional variables were derived from the current demographic estimates and the 2002 Census, and were specific for each year of the survey.

Integrated calibration method in bootstrap method for standard error estimation

In bootstrap algorithm of standard error estimation procedure the construction of every resample will in general mean that the post-strata marginal totals are no longer satisfied. It means that every bootstrap sample should be reweighted. Bootstrap data may be reweighted by applying to it the usual algorithm with the original population marginal totals (Canty, Davison, 1999). The results of such calibration are presented below.

5.3. Empirical results - bootstrap method with calibration

Integrated calibration method for Poland

Every bootstrap sample was reweighted using additional variables, derived from the current demographic estimates (the same as original weights). After that standard error estimation was conducted. Table 1 presents results before and after weights calibration.

To draw a comparison a coefficient of variation was used:

$$cv = \frac{SE(\hat{\theta})}{\hat{\theta}} \tag{14}$$

where:

$SE(\hat{\theta})$ – standard error of a statistic under the given sample design,

$\hat{\theta}$ – the direct estimator of indicator.

Table 1. Comparison of coefficients of variation for Poland using bootstrap (B=1000) before and after weights calibration.

Indicator	Estimate	Without calibration		With calibrated weights	
		SE	cv (in %)	SE_calib	cv_calib (in %)
ARPR (%)	16,88	0,4	2,38	0,39	2,33
RRPG (%)	20,54	0,72	3,51	0,72	3,48
Gini (%)	32,02	0,42	1,32	0,41	1,27
S80_S20	5,11	0,1	1,93	0,1	1,88
Mean_Inc (EUR)	4936,68	42,74	0,87	39,72	0,8
Median (EUR)	4153,96	31,74	0,76	28,9	0,7

Source: Derived from data provided by GUS: European Statistics on Income and Living Conditions 2008

Comparing results one can observe that values of standard error after weights calibration are lower than before but differences are very small. It means that at country level reweighting has statistically insignificant effect for the bootstrap.

Integrated calibration method at NUTS2 level in Poland

At NUTS2 level weights calibration does not improve the outcomes – for some voivodships standard error decreases and in other cases it increases. That is why average values of coefficient of variation show very similar performance before and after reweighting for every measure (Table 2).

Table 2. Comparison of average values of estimates, standard errors and coefficient of variation calculated for six measures for Poland (16 NUTS2 units) using bootstrap (B=1000) before and after weights calibration.

Indicator		Estimate	Without calibration		With calibrated weights	
			SE	CV (in %)	SE_calib	CV_calib (in%)
ARPR (%)	MIN	12.4	0.9	7.2	0.9	6.8
	MAX	27.6	3.5	21.5	3.2	19.4
	MEAN	17.9	2.1	11.6	2.0	11.1
RRPG (%)	MIN	10.1	1.7	8.6	1.8	9.0
	MAX	24.7	9.4	39.6	8.9	44.1
	MEAN	20.0	3.1	16.1	3.1	16.5
Gini (%)	MIN	27.4	0.7	2.4	0.7	2.4
	MAX	38.5	2.2	8.0	2.7	9.9
	MEAN	30.2	1.3	4.3	1.3	4.3
S80_S20	MIN	4.1	0.2	4.0	0.2	4.0
	MAX	6.6	0.5	10.8	0.6	13.5
	MEAN	4.8	0.3	6.7	0.3	6.7
Mean_Inc (EUR)	MIN	3 997.8	92.9	1.8	83.5	1.6
	MAX	6 223.9	278.6	5.4	333.2	6.5
	MEAN	4 760.9	150.3	3.1	147.3	3.1
Median (EUR)	MIN	3 472.8	74.1	1.8	70.1	1.7
	MAX	4 618.6	228.2	5.9	221.2	5.7
	MEAN	4 089.2	137.5	3.4	131.6	3.2

Source: Derived from data provided by GUS: European Statistics on Income and Living Conditions 2008

There is no one tendency in differences between average values of coefficient of variation before and after weights calibration – in case of ARPR, Mean_Inc and Median it decreases, in case of RRPG it increases (Table 3).

Table 3. Comparison of average values of coefficient of variation for Poland (16 NUTS2 units) using bootstrap (B=1000) before and after weights calibration.

	mean_cv (%)	mean_cv_calib (%)	cv-cv_calib (%)
ARPR (%)	11.65	11.07	0.6
RRPG (%)	16.06	16.52	- 0.5
Gini (%)	4.28	4.29	- 0.0
S80_S20	6.68	6.70	- 0.0
Mean_Inc (EUR)	3.14	3.06	0.1
Median (EUR)	3.37	3.23	0.1

Source: Derived from data provided by GUS: European Statistics on Income and Living Conditions 2008 (EU-SILC), version of 01.03.2010

There are some significant differences in case of estimation at NUTS2 level – mainly in region with smaller sample size.

6. Concluding remarks

We would like to add that two important measurement concepts related to calibration are *precision* and *accuracy*. *Precision* refers to the variability in the parameter depending on size of sample, while *accuracy* refers to the actual amount of error that exists in calibration. All measurement processes used for calibration are subject to various sources of error. It is common practice to classify them as random or systematic errors. When measurement is repeated many times, the results will exhibit random statistical fluctuations which may or may not be significant, depending on the size of a sample. Systematic errors are offsets from the true value of a parameter and, if they are known, corrections are generally applied, eliminating their effect on the calibration. If they are not known, they can have an adverse effect on the accuracy of the calibration. High-accuracy calibrations are usually accompanied by an analysis of the sources of an error and a statement of the uncertainty of the calibration. Uncertainty indicates how much the accuracy of the calibration could be degraded as a result of the combined errors. In our case we estimated precision of the estimates.

In conclusion:

1. Reweighting (calibration of sampling weights), in our case, has not significant effect for the bootstrap estimates at country level.
2. At NUTS2 level weights calibration does not always reduce standard error estimates – in some cases standard error decreases and in other cases it increases. However, these differences are statistically not significant.
3. Calibration is time-consuming because of repeating calibration algorithm as many times as the number of bootstrap replicates (for example $b=1000$)

Acknowledgment

The presented work was done under the S.A.M.P.L.E. project (Small Area Methods for Poverty and Living Condition Estimates). This research programme was funded by the European Commission under the Seventh Framework (FP7) Programme of the European Union. (http://www.sample-project.eu/).

The authors would like to thank the Central Statistical Office of Poland for the production and provision of the survey data used, Mr. Bronisław Lednicki for sampling consultations, and Dr. Waldemar Popinski, the referee, for constructive corrections and comments.

Data sources provided by GUS come from European Statistics on Income and Living Conditions (EU-SILC 2008, version of 01.03.2010; EU-SILC 2007, version of 01.08.2009). The European Commission does not take responsibility for conclusions in this work.

REFERENCES

CANTY A.J., DAVISON A.C.,(1999), *Resampling-based Variance Estimation for Labour Force Survey*, Journal of the Royal Statistical Society. Series D (The Statistician), Vol. 48, No. 3 (1999), pp. 379-391.

DEVILLE, J.-C. and SARNDAL, C.E. and Sautory, O. (1993), *Generalized Ranking Procedures in Survey Sampling*, Journal of the American Statistical Association, Vol. 88, No. 423, pp. 1013-1020.

EFRON B., (1979) *Bootstrap methods: another look at the jackknife*, "Annals of statistics" 7, 1-26.

EUROSTAT, (2002), *Estimation Techniques in Statistics – A Methodological Note on Calibration.*

EUROSTAT, (2004), *Description of target variables: Cross-sectional and Longitudinal*, EU-SILC 065/04, pp. 31 – 36.

EUROSTAT, (2005), *Continuity of indicators between end-ECHP and start-SILC Algorithms to compute cross-sectional indicators of poverty and social inclusion adopted under the open method of coordination.*

EUROSTAT, (2009) *Methodological studies and quality assessment of EU-SILC*, Report SILC.04 15 April 2009, SAS programs for variance estimation of the measures required for Intermediate Quality Report.

GAMBINO J., (1999), *Discussion of "Issues in Weighting Household and Business Surveys"*, Proceedings of the 52th International Statistical Institute Session, Finland.

GUS, (2009), *Incomes and Living Conditions of the Population in Poland,* report from the EU-SILC survey of 2007 and 2008, Warsaw.

KORDOS J. , ZIĘBA-PIETRZAK A., (2010), *Development of standard errors estimation methods in complex household sample surveys in Poland*, Statistics in Transition –new series, Vol. 11, No.2, pp. 231–253.

MCCARTHY P. J., SNOWDEN C. B., (1985) T*he Bootstrap and Finite Population Sampling*, "Vital and Health Statistics", Series 2, no. 95, Public Health Service Publication 85-1369, U.S. Government Printing Office, Washington DC.

POPIŃSKI W.,(2006), *Development of the Polish Labour Force Survey*, Statistics in Transition, Vol. 7, No. 5, pp.. 1009-1030.

RAO J.N.K., WU C.F.J. (1984), *Bootstrap Inference for Sample Surveys*, Proceedings of the American Statistical Association Survey Research Methods Section, 106-112.

SHAO J., (2003) *Impact of the Bootstrap on Sample Surveys*, "Statististical Science" Vol. 18, No 2, 191-198.

SITTER R. R., (1992), *Comparing Three Bootstrap Methods for Survey Data*, The Canadian Journal of Statistics, Vol. 20, No. 2, 135-154.

ZIĘBA, A., KORDOS, J. (2010), Comparing Three Methods of Standard Error Estimation for Poverty Measures. In: J. Wywiał, W. Gamrot (Eds), Research in Social and Economic Surveys, Katowice, University of Economics.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 97—114

THE BETA PARETO DISTRIBUTION

AHMED HURAIRAH[1]

ABSTRACT

In this paper, we introduce a generalization—referred to as the beta Pareto distribution, generated from the logit of a beta random variable. We provide a comprehensive treatment of the mathematical properties of the beta Pareto distribution. We derive expressions for the *kth* moments of the distribution, variance, skewness, kurtosis, mean deviation about the mean, mean deviation about the median, Rényi entropy, Shannon entropy. We also discuss simulation issues, estimation of parameters by the methods of moments and maximum likelihood.

Key words: Beta Pareto distribution; Kurtosis; Rényi entropy

1. Introduction

The generalization is motivated by the following general class: If *G* denotes the cumulative distribution function (cdf) of a random variable then a generalized class of distributions can be defined by

$$F(x) = I_{G(x)}(a, b) \tag{1}$$

for $a > 0$ and $b > 0$, where

$$I_{y(a,b)} = \frac{B_y(a,b)}{B(a,b)},$$

denotes the incomplete beta function ratio, and

$$B_y(a, b) = \int_0^y x^{a-1}(1-x)^{b-1} \ dx$$

[1] Department of Statistics, Sana'a University, Yemen. E-mail: Hurairah69@yahoo.com.

denotes the incomplete beta function. This class of distributions came to prominence after the recent paper by Jones (2004). Eugene et al. (2002) introduced what is known as the beta normal distribution by taking *G* in (1) to be the cdf of the normal distribution with parameters μ and σ. The only properties of the beta normal distribution known are some first moments derived by Eugene et al. (2002) and some more general moment expressions derived by Gupta and Nadarajah (2004). More recently, Natarajahand and Kotz (2004) introduced what is known as the beta Gumbel distribution by taking *G* in (1) to be the cdf of the Gumbel distribution with parameters μ and σ. This distribution is a little more tractable than the beta normal distribution in that Natarajah and Kotz (2004) were able to provide closed-form expressions for the moments, the asymptotic distribution of the extreme order statistics and the estimation procedure. Another distribution that happens to belong to (1) class is the log *F* (or beta logistic) distribution. This distribution appears reasonably tractable partly because it has been around for over 20 years [5], but it did not originate directly from (1) idea. In this paper we will introduce the beta Pareto (BP) distribution by taking *G* in (1) to be the cdf of Pareto distribution with parameter λ. The cdf of the BP distribution becomes

$$F(x) = I_{\left(\frac{\lambda}{x}\right)^c}(a,b) \tag{2}$$

for $x \geq \lambda$, $\lambda > 0$, $c > 0$, $a > 0$ and $b > 0$. The corresponding probability density function (pdf) and the hazard rate function associated with (2) are:

$$f(x) = \frac{c\,x^{-1}}{B(a,b)}\left(\frac{\lambda}{x}\right)^{ac}\left(1-\left(\frac{\lambda}{x}\right)^{c}\right)^{b-1}, \quad x \geq \lambda \tag{3}$$

and

$$h(x) = \frac{c\,x^{-1}\left(\frac{\lambda}{x}\right)^{ac}\left(1-\left(\frac{\lambda}{x}\right)^{c}\right)^{b-1}}{B_{\left(\frac{\lambda}{x}\right)^c}(a,b)} \tag{4}$$

The shapes of the probability density function (3) and the hazard rate function (4) for the beta Pareto distributions are given in Figs. 1 and 2, respectively.

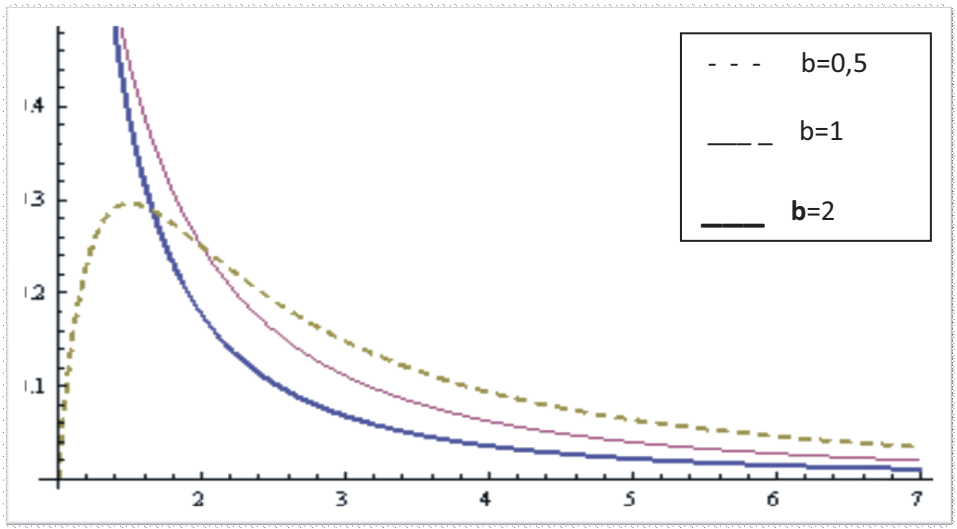


Figure. 1. The beta Pareto pdf (3) for $a = 1,\ \lambda = 1\ and\ c = 1$

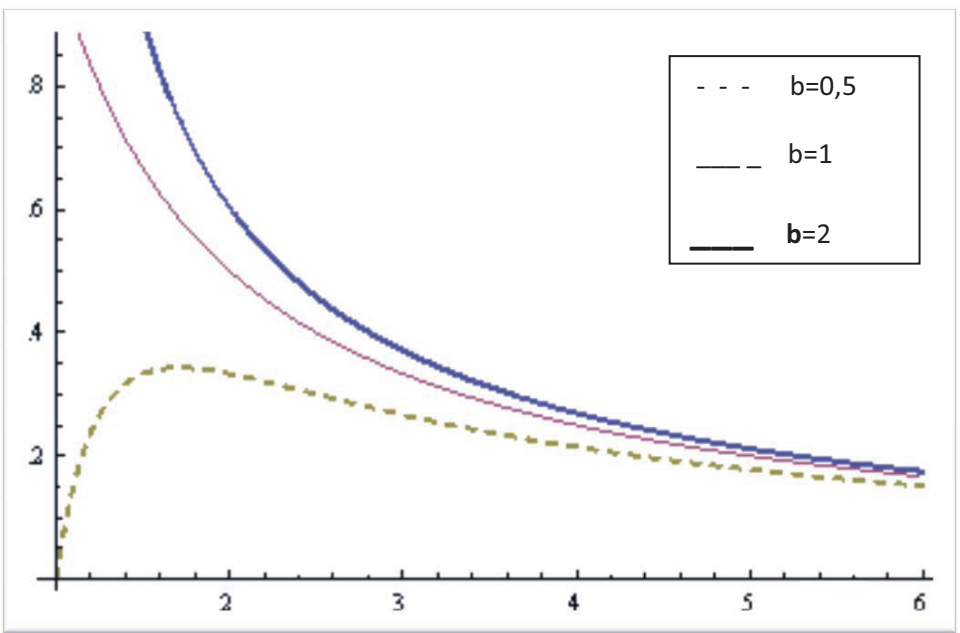


Figure. 2. The beta Pareto hazard rate function pdf (4)
for $a = 1,\ \lambda = 1\ and\ c = 1$

2. Moments

The *kth* moments of x can be written as

$$E\left(x^k\right) = \frac{\lambda^k \Gamma(a+b)\Gamma\left(a-\frac{k}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{k}{c}\right)} \tag{5}$$

In particular, the first four moments can be worked out as

$$E(x) = \frac{\lambda\, \Gamma(a+b)\, \Gamma\left(a-\frac{1}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{1}{c}\right)} \tag{6}$$

$$E(x^2) = \frac{\lambda^2 \Gamma(a+b)\Gamma\left(a-\frac{2}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{2}{c}\right)} \tag{7}$$

$$E(x^3) = \frac{\lambda^3 \Gamma(a+b)\Gamma\left(a-\frac{3}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{3}{c}\right)} \tag{8}$$

and

$$E(x^4) = \frac{\lambda^4 \Gamma(a+b)\Gamma\left(a-\frac{4}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{4}{c}\right)} \tag{9}$$

The first three entral moments, skewness and the kurtosis of x can be given by

$$Var(x) = E(x^2) - [E(x)]^2$$
(10)

$$Var(x) = \frac{\lambda^2 \Gamma(a+b)\Gamma\left(a-\frac{2}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{2}{c}\right)} - \left[\frac{\lambda\, \Gamma(a+b)\Gamma\left(a-\frac{1}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{1}{c}\right)}\right]^2 \tag{11}$$

$$E[x - E(x)]^3 = \frac{\lambda^3 \Gamma(a+b)\Gamma\left(a-\frac{3}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{3}{c}\right)} + \frac{2\lambda^3 \Gamma(a+b)^3 \Gamma\left(a-\frac{1}{c}\right)^3}{\Gamma(a)^3 \Gamma\left(a+b-\frac{1}{c}\right)^3} - \frac{3\lambda^3 \Gamma(a+b)^2 \Gamma\left(a-\frac{1}{c}\right)\Gamma\left(a-\frac{2}{c}\right)}{\Gamma(a)^2 \Gamma\left(a+b-\frac{1}{c}\right)\Gamma\left(a+b-\frac{2}{c}\right)}$$

(12)

$$E[x-E(x)]^4 = \frac{\lambda^4\Gamma(a+b)\Gamma\left(a-\frac{4}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{4}{c}\right)} - \frac{3\lambda^4\Gamma(a+b)^4\Gamma\left(a-\frac{1}{c}\right)^4}{\Gamma(a)^4\Gamma\left(a+b-\frac{1}{c}\right)^4} + \frac{6\lambda^4\Gamma(a+b)^3\Gamma\left(a-\frac{2}{c}\right)\Gamma\left(a-\frac{1}{c}\right)^2}{\Gamma(a)^3\Gamma\left(a+b-\frac{2}{c}\right)\Gamma\left(a+b-\frac{1}{c}\right)^2} - \frac{4\lambda^4\Gamma(a+b)^2\Gamma\left(a-\frac{3}{c}\right)\Gamma\left(a-\frac{1}{c}\right)}{\Gamma(a)^2\Gamma\left(a+b-\frac{3}{c}\right)\Gamma\left(a+b-\frac{1}{c}\right)} \tag{13}$$

$$Skewness(x) = \frac{\lambda^3\Gamma(a+b)}{\Gamma(a)^3}\left(\frac{\Gamma(a)^2\Gamma\left(a-\frac{3}{c}\right)}{\Gamma\left(a+b-\frac{3}{c}\right)} + \frac{2\Gamma(a+b)^2\Gamma\left(a-\frac{1}{c}\right)^3}{\Gamma\left(a+b-\frac{1}{c}\right)^3} - \left(3\Gamma(a)\Gamma(a+b)\Gamma\left(a-\frac{1}{c}\right)\Gamma\left(a-\frac{2}{c}\right)\right) \Big/ \left(\Gamma\left(a+b-\frac{2}{c}\right)\left(\frac{\lambda^2\Gamma(a+b)}{\Gamma(a)^2}\Gamma(a+b)\left(\frac{\Gamma(a)\Gamma\left(a-\frac{2}{c}\right)}{\Gamma\left(a+b-\frac{2}{c}\right)} - \frac{\Gamma(a+b)\Gamma\left(a-\frac{1}{c}\right)^2}{\Gamma\left(a+b-\frac{1}{c}\right)^2}\right)^{\frac{3}{2}}\Gamma\left(a+b-\frac{1}{c}\right)\right)\right) \tag{14}$$

$$Kurtosis(x) = \frac{1}{\Gamma(a)^4}\left(\frac{\left[\frac{\lambda^4\Gamma(a)^3\Gamma(a+b)\Gamma\left(a-\frac{4}{c}\right)}{\Gamma\left(a+b-\frac{4}{c}\right)} - \frac{3\lambda^4\Gamma(a+b)^4\Gamma\left(a-\frac{1}{c}\right)^4}{\Gamma\left(a+b-\frac{1}{c}\right)^4} + \frac{6\lambda^4\Gamma(a)\Gamma(a+b)^3\Gamma\left(a-\frac{2}{c}\right)\Gamma\left(a-\frac{1}{c}\right)^2}{\Gamma\left(a+b-\frac{2}{c}\right)\Gamma\left(a+b-\frac{1}{c}\right)^2} - \frac{4\lambda^4\Gamma(a)^2\Gamma(a+b)^2\Gamma\left(a-\frac{3}{c}\right)\Gamma\left(a-\frac{1}{c}\right)}{\Gamma\left(a+b-\frac{3}{c}\right)\Gamma\left(a+b-\frac{1}{c}\right)}\right]}{\left[\left(\frac{\lambda^2\Gamma(a)\Gamma(a+b)\Gamma\left(a-\frac{2}{c}\right)}{\Gamma\left(a+b-\frac{2}{c}\right)} - \frac{\lambda^2\Gamma(a+b)^2\Gamma\left(a-\frac{1}{c}\right)^2}{\Gamma\left(a+b-\frac{1}{c}\right)^2}\right)^2\right]}\right) \tag{15}$$

Note that the skewness and kurtosis measures depend on a, b, λ and c .

3. Mean deviations

The amount of scatter in a population is evidently measured to some extent by the totality of deviations from the mean and median. These are known as the mean deviation about the mean and the mean deviation about the median defined by

$$\delta_1(x) = \int_{\lambda}^{\infty} |x-\mu| f(x)dx$$

and

$$\delta_2(x) = \int_{\lambda}^{\infty} |x - M| f(x) dx$$

where $\mu = E(x)$ and M denotes the median. These measures can be calculated using the relationships that

$$\delta_1(x) = \int_{\lambda}^{\mu} (\mu - x) f(x) dx + \int_{\mu}^{\infty} (x - \mu)\, f(x) dx$$

$$= 2 \int_{\lambda}^{\mu} (\mu - x) f(x) dx$$

$$= 2\{\mu F(\mu) - \int_{\lambda}^{\mu} x f(x) dx\} \tag{16}$$

and

$$\delta_2(x) = \int_{\lambda}^{M} (M - x) f(x) dx + \int_{M}^{\infty} (x - M)\, f(x) dx$$

$$= 2MF(M) - M - \int_{\lambda}^{M} x f(x) dx + \int_{M}^{\infty} x\, f(x) dx$$

$$= E(x) + 2MF(M) - M - 2 \int_{\lambda}^{M} x f(x) dx \tag{17}$$

$$\int_{\lambda}^{m} x\, f(x) dx = \frac{c}{B(a,b)} \int_{\lambda}^{m} \left(\frac{\lambda}{X}\right)^{ac} \left(1 - \left(\frac{\lambda}{X}\right)^{c}\right)^{b-1} dx$$

$$= \frac{1}{B[a,b]} \lambda \left(B_1\left[a - \frac{1}{c}, b\right] - \lambda^{ac} \left(\left(\frac{1}{m}\right)^{c}\right)^{-a+\frac{1}{c}} \left(\frac{1}{\lambda\, m}\right)^{ac} m\, B_{\left(\frac{\lambda}{m}\right)^{c}}\left[a - \frac{1}{c}, b\right]\right)$$

(18)

Substituting (18) into (16) and (17), one obtains the following expressions for the mean deviations:

$$\delta_1(x) = 2\left\{\mu\, I_{\left(\frac{\lambda}{\mu}\right)^{c}}(a,b) - \frac{\lambda B_1\left[a - \frac{1}{c}, b\right]}{B[a,b]} + \frac{\lambda^{ca+1} \left(\left(\frac{1}{\mu}\right)^{c}\right)^{-a+\frac{1}{c}} \left(\frac{1}{\lambda\, \mu}\right)^{a\, c} \mu\, B_{\left(\frac{\lambda}{\mu}\right)^{c}}\left[a - \frac{1}{c}, b\right]}{B[a,b]}\right\}$$

and

$$\delta_2(x) = \frac{\lambda\,\Gamma(a+b)\,\Gamma\left(a-\frac{1}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{1}{c}\right)} + 2MI_{\left(\frac{\lambda}{M}\right)^c}(a,b) - M - 2\left\{\frac{\lambda B_1\left[a-\frac{1}{c},b\right]}{B[a,b]} - \frac{\lambda\left(\left(\frac{1}{\mu}\right)^c\right)^{-a+\frac{1}{c}}\left(\frac{1}{\lambda\,\mu}\right)^{a\,c}\mu\, B_{\left(\frac{\lambda}{\mu}\right)^{\alpha}}\left[a-\frac{1}{c},b\right]}{c}\right\}$$

It can be verified that in the particular case $a = 1$ and $b = 1$, $\delta_1(x)$ reduces to $2c\lambda(c-1)^{-1}\left(1-\frac{\lambda}{\mu}\right)^{c-1}$, which is the mean deviation for the Pareto distribution.

4. Rényi and Shannon entropies

An entropy of a random variable x is a measure of variation of the uncertainty. Rényi entropy is defined by

$$\mathcal{J}_R(\gamma) = \frac{1}{1-\gamma}\, log[\int f^{\gamma}(x)dx] \tag{19}$$

where $\gamma > 0$ and $\gamma \neq 1$ Rényi (1961). For the beta Pareto pdf given by (3)

$$\int_{\lambda}^{\infty} f^{\gamma}(x)dx = \frac{c^{\gamma}}{B^{\gamma}(a,b)}\int_{\lambda}^{\infty} x^{-\gamma}\left(\frac{\lambda}{x}\right)^{\gamma ac}\left(1-\left(\frac{\lambda}{x}\right)^c\right)^{\gamma b-\gamma} dx$$

$$= \frac{\left(\frac{c}{\lambda}\right)^{\gamma-1} B\left(\gamma b-\gamma+1, \frac{\gamma ac+\gamma-1}{c}\right)}{B^{\gamma}(a,b)}$$

The Rènyi entropy, (19) takes the expression

$$\mathcal{J}_R(\gamma) = -\log\left(\frac{c}{\lambda}\right) + \frac{1}{1-\gamma}\, log\left[\frac{\left(\frac{c}{\lambda}\right)^{\gamma-1} B\left(\gamma b-\gamma+1, \frac{\gamma ac+\gamma-1}{c}\right)}{B^{\gamma}(a,b)}\right] \tag{20}$$

Shannon entropy defined by $E[-logf(x)]$ is the particular case of (19) for $\gamma \uparrow 1$. Limiting $\gamma \uparrow 1$ in (20) and using L'Hospital's rule, one obtains

$$E[-logf(x)] = -logc + logB(a,b) + log\left(\frac{\lambda\,\Gamma(a+b)\,\Gamma\left(a-\frac{1}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{1}{c}\right)}\right)(c(a-b+1)+1) - c\, log\lambda(a-b+1)$$

Song (2001) observed that the gradient of the Rényi entropy $\mathcal{J}_R^{'}(\gamma) = \left(\frac{d}{d\gamma}\right)\mathcal{J}_R(\gamma)$ is related to the log likelihood by $\mathcal{J}_R^{'}(1) = -\left(\frac{1}{2}\right)Var\left[log\left(f(x)\right)\right]$. This equality and the fact that the quantity $-\acute{\mathcal{J}}_R(1)$ remains invariant under location and scale transformations motivated Song to propose $-2\,\acute{\mathcal{J}}_R(1)$ as a measure of the shape of a distribution. From (20), the first derivative is

$$\mathcal{J}_R^{'}(\gamma) = \Bigg(c\,B\left(1+(b-1)\gamma, \tfrac{\gamma ac+\gamma-1}{c}\right)\Bigg(B^{\gamma}(a,b)\Bigg((\gamma-1)log\left(\tfrac{c}{\lambda}\right) + log\left[\frac{\left(\frac{c}{\lambda}\right)^{\gamma-1}B\left(1+(b-1)\gamma, \frac{\gamma ac+\gamma-1}{c}\right)}{B^{\gamma}(a,b)}\right]\Bigg) + (\gamma-1)\left(B^{\gamma}\log\left(B\right)(a,b)\right) - (\gamma-1)B^{\gamma}(a,b)\Bigg((ac+1)B^{(0,1)}\left(1+(b-1)\gamma, \tfrac{\gamma ac+\gamma-1}{c}\right) + (b-1)cB^{(1,0)}\left(1+(b-1)\gamma, \tfrac{\gamma ac+\gamma-1}{c}\right)\Bigg)\Bigg)\Bigg) / \left(c(\gamma-1)^2 B\left(1+(b-1)\gamma, \tfrac{\gamma ac+\gamma-1}{c}\right)B^{\gamma}(a,b)\right)$$

Using L'Hospital's rule again, one gets the expression

$$-2\,\acute{\mathcal{J}}_R(1) = -log\left(\tfrac{B(b,a)}{B(a,b)}\right)\Big(c\,B(b,a)\big(B\;log(B)\big)(a,b) + B(a,b)\Big((1+ac)B^{(0,1)}(b,a) + (b-1)c\,B^{(1,0)}(b,a)\Big)\Big) \qquad (21)$$

for the measure proposed by Song (2001). When f(x) has a finite fourth moment, this measure plays a similar role as the kurtosis measure in comparing the shapes of various densities and measuring heaviness of tails. Song (2001) provided a detailed discussion including examples to show in general that his measure is better: "it measures more than what kurtosis measures. Kurtosis is often regarded as a measure of tail heaviness of a distribution relative to that of the normal distribution and a measure of deviation from normality depending on the relative frequency of values either near the mean or far from it to values an intermediate distance from the mean. However, one of the problems with the kurtosis measure is that in many situations of interest when the tails of a

distribution are heavy, the fourth moments may not exist or, even worse, the mean may not exist. This illustrates the limitation of the moments as indicators of distributional shape. However, our shape parameter is almost always applicable in virtually all situations encountered in practice. It does not require restrictions like symmetry assumption and may be applied to distributions with heavy tails such as Cauchy, Cramer, Levy distributions.

5. Estimation of the parameters

Here, we consider estimation by two methods: the method of moments and the method of maximum likelihood.

5.1. Method of moments

Let $x_1, x_2, \ldots, x_n$ be a random sample from (3). Equating $\mathrm{E}(x)$, $Var(x)$ and $E[x - E(x)]^3$ in (6), (10) and (11), respectively, with the corresponding sample estimates

$$S_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$S_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - S_1)^2$$

and

$$S_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - S_1)^3$$

respectively, one obtains the system of equations

$$S_1 = \frac{\lambda\,\Gamma(a+b)\,\Gamma\left(a-\frac{1}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{1}{c}\right)} \tag{22}$$

$$S_2 = \frac{\lambda^2\Gamma(a+b)\Gamma\left(a-\frac{2}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{2}{c}\right)} - \left[\frac{\lambda\,\Gamma(a+b)\Gamma\left(a-\frac{1}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{1}{c}\right)}\right]^2 \tag{23}$$

$$S_3 = \frac{\lambda^3\Gamma(a+b)\Gamma\left(a-\frac{3}{c}\right)}{\Gamma(a)\Gamma\left(a+b-\frac{3}{c}\right)} + \frac{2\lambda^3\Gamma(a+b)^3\Gamma\left(a-\frac{1}{c}\right)^3}{\Gamma(a)^3\Gamma\left(a+b-\frac{1}{c}\right)^3} - \frac{2\lambda^3\Gamma(a+b)^2\Gamma\left(a-\frac{1}{c}\right)\Gamma\left(a-\frac{2}{c}\right)}{\Gamma(a)^2\Gamma\left(a+b-\frac{1}{c}\right)\Gamma\left(a+b-\frac{2}{c}\right)} \tag{24}$$

Combining (22) with (23) and (22) with (24), one obtains the equations

$$\frac{s_2}{s_1^2} = -1 + \frac{\Gamma(a)\,\Gamma\left(a-\frac{2}{c}\right)\Gamma\left(a+b-\frac{1}{c}\right)^2}{\Gamma(a+b)\,\Gamma\left(a+b-\frac{2}{c}\right)\Gamma\left(a-\frac{1}{c}\right)^2}$$

$$\frac{s_3}{s_1^3} = 2 + \frac{\left(\Gamma(a)\,\Gamma\left(a+b-\frac{1}{c}\right)^2\left(-\frac{3\Gamma(a+b)\,\Gamma\left(a-\frac{2}{c}\right)\Gamma\left(a-\frac{1}{c}\right)}{\Gamma\left(a+b-\frac{2}{c}\right)}+\frac{3\Gamma(a)\,\Gamma\left(a-\frac{3}{c}\right)\Gamma\left(a+b-\frac{1}{c}\right)}{\Gamma\left(a+b-\frac{3}{c}\right)}\right)\right)}{\Gamma(a+b)^2\,\Gamma\left(a-\frac{1}{c}\right)^3}$$

which can be solved simultaneously to give estimates for a, b and c. The estimate for λ can then be obtained directly from (22). Note that if $b = 1$ then (22) gives $\hat{\lambda} = \frac{s_1\,(ac-1)}{a\,c}$, which is the usual estimator for λ under the Pareto model.

5.2. Maximum likelihood estimator

The logarithm of the likelihood function for a random sample $x_1, x_2, \ldots, x_n$ from (3) can be expressed as

$$\log L(a,b,c,\lambda) = n\log c - n\log B(a,b) - \sum_{i=1}^{n} log x_i + a\,c\sum_{i=1}^{n} log\left(\frac{\lambda}{x_i}\right) + (b-1)\sum_{i=1}^{n} log\left(1-\left(\frac{\lambda}{x_i}\right)^c\right) \quad (25)$$

To obtain the normal equations for the unknown parameters, we differentiates (25) partially with respect to the parameters a, b, λ and c and equating to zero. The resulting equations are given below in (26), (27), (28) and (29), respectively.

$$\frac{\partial \log L}{\partial a} = -n[\Psi(a) - \Psi(a+b)] + c\sum_{i=1}^{n} log\left(\frac{\lambda}{x_i}\right) \quad (26)$$

$$\frac{\partial \log L}{\partial b} = -n[\Psi(b) - \Psi(a+b)] + \sum_{i=1}^{n} log\left(1-\left(\frac{\lambda}{x_i}\right)^c\right) \quad (27)$$

$$\frac{\partial \log L}{\partial \lambda} = \frac{anc}{\lambda} - (b-1)\sum_{i=1}^{n} \frac{c\left(\frac{\lambda}{x_i}\right)^{c-1}}{\left(1-\left(\frac{\lambda}{x_i}\right)^c\right)x_i} \quad (28)$$

$$\frac{\partial \log L}{\partial c} = \frac{n}{c} + a\sum_{i=1}^{n} log\left(\frac{\lambda}{x_i}\right) - (b-1)\sum_{i=1}^{n} \frac{\left(\frac{\lambda}{x_i}\right)^c \log\left(\frac{\lambda}{x_i}\right)}{1-\left(\frac{\lambda}{x_i}\right)^c} \quad (29)$$

The second order partial derivatives of the log-likelihood function $L(\theta/x)$ with respect to the parameters a, b, λ and c are given by:

$$\frac{\partial^2 \log L}{\partial a^2} = -n[\Psi(a) - \Psi(a+b)] \tag{30}$$

$$\frac{\partial^2 \log L}{\partial a \partial b} = n\,\Psi(a+b) \tag{31}$$

$$\frac{\partial^2 \log L}{\partial a \partial \lambda} = \frac{nc}{\lambda} \tag{32}$$

$$\frac{\partial^2 \log L}{\partial a \partial c} = \sum_{i=1}^{n} \log\left(\frac{\lambda}{x_i}\right) \tag{33}$$

$$\frac{\partial^2 \log L}{\partial b^2} = -n[\Psi(b) - \Psi(a+b)] \tag{34}$$

$$\frac{\partial^2 \log L}{\partial b \partial \lambda} = -\sum_{i=1}^{n} \frac{c\left(\frac{\lambda}{x_i}\right)^{c-1}}{x_i\left(1-\left(\frac{\lambda}{x_i}\right)^{c}\right)} \tag{35}$$

$$\frac{\partial^2 \log L}{\partial b \partial c} = -\sum_{i=1}^{n} \frac{\left(\frac{\lambda}{x_i}\right)^{c} \log\left(\frac{\lambda}{x_i}\right)}{1-\left(\frac{\lambda}{x_i}\right)^{c}} \tag{36}$$

$$\frac{\partial^2 \log L}{\partial \lambda^2} = -\frac{a\,nc}{\lambda^2} + (b-1)\sum_{i=1}^{n}\left(-\frac{c(c-1)\left(\frac{\lambda}{x_i}\right)^{c-2}}{x_i^2\left(1-\left(\frac{\lambda}{x_i}\right)^{c}\right)} - \frac{c^2\left(\frac{\lambda}{x_i}\right)^{2c-2}}{x_i^2\left(1-\left(\frac{\lambda}{x_i}\right)^{c}\right)^2}\right) \tag{37}$$

$$\frac{\partial^2 \log L}{\partial \lambda \partial c} = \frac{a\,n}{\lambda} + (b-1)\sum_{i=1}^{n}\left(-\frac{\left(\frac{\lambda}{x_i}\right)^{c-1}}{x_i\left(1-\left(\frac{\lambda}{x_i}\right)^{c}\right)} - \frac{c\left(\frac{\lambda}{x_i}\right)^{c-1}\log\left(\frac{\lambda}{x_i}\right)}{x_i\left(1-\left(\frac{\lambda}{x_i}\right)^{c}\right)} - \frac{c\left(\frac{\lambda}{x_i}\right)^{2c-1}\log\left(\frac{\lambda}{x_i}\right)}{x_i\left(1-\left(\frac{\lambda}{x_i}\right)^{c}\right)^2}\right) \tag{38}$$

$$\frac{\partial^2 \log L}{\partial c^2} = -\frac{n}{c^2} + (b-1)\sum_{i=1}^{n}\left(-\frac{\left(\frac{\lambda}{x_i}\right)^{c}\log\left(\frac{\lambda}{x_i}\right)^2}{\left(1-\left(\frac{\lambda}{x_i}\right)^{c}\right)} - \frac{\left(\frac{\lambda}{x_i}\right)^{2c}\log\left(\frac{\lambda}{x_i}\right)^2}{\left(1-\left(\frac{\lambda}{x_i}\right)^{c}\right)^2}\right) \tag{39}$$

where $\Psi(x)$ is the digamma function defined by $\Psi(x) = d\log\Gamma(x)/dx$, and $\Gamma(x)$ is the Gamma function. For interval estimation of $(a, b, \lambda$ and $c)$ and tests of hypothesis, one requires the Fisher information matrix. The elements of the Fisher information matrix can be easily derived using the equations (30-39) and results in (6) and (7). The equations (26 -29) cannot be solved analytically. An

iterative process such as Newton-Raphson method has to be adopted to solve this system of equations for $(a, b, \lambda \text{ and } c)$. To study the properties of the estimators of the beta Pareto distribution and their performances. There are many measures that can be used to get information about the performance of the estimators. The bias, the t value, mean square error (MSE), finite sample variance (FSV) and asymptotic variance (ASV) are such useful measures.

5.3. Simulation study

The outline of the simulation is as follows:

1. Note from (2) that if U be a uniform $\left(0, 1 \right)$ random number and by equating $F(x) = U$, then $x = \frac{\lambda}{U^{1/c}}$, which is the formula used for generating random numbers.
2. We took the sample size as $n = 10, 20, 30, 50, 80, 100,$ and the parameters a, b, λ and c as $1.2, 1.3, 1.4, 1.5$. Without loss of generality.
3. We generated 2000 samples from the beta Pareto distribution based on the cumulative distribution function described above. The sample observations $x_1, x_2, \cdots, x_n$ are generated from $x = \frac{\lambda}{U^{1/c}}$, where $u_1, u_2, \cdots, u_n$ random numbers are uniformly distributed on the interval $(0,1)$.
4. Then solve the likelihood equations (26 -29) using Newton-Raphson method. The solutions that we obtained are the maximum likelihood estimators of a, b, λ and c say $\hat{a}, \hat{b}, \hat{\lambda}$ and $\hat{c}$ for the samples obtained.
5. Then compute the following.

$$Bias\left(\widehat{\theta_\iota}\right) = E\left(\widehat{\theta_\iota} - \theta\right) = E\left(\widehat{\theta_\iota}\right) - \theta \ , \quad i = 1,2,3, \dots, n \tag{40}$$

where $\theta = a, b, \lambda$ and, c $\hat{\theta} = \hat{a}, \hat{b}, \hat{\lambda}$ and $\hat{c}$.

The t value

$$t = \sqrt{n}\frac{\left(\bar{\hat{\theta}} - \theta\right)}{\sqrt{var(\hat{\theta})}}, \tag{41}$$

$\bar{\hat{\theta}} = \frac{1}{r}\sum_{i=1}^{r} \widehat{\theta_\iota}$, r number of replications.

The mean square error (MSE)

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \tag{42}$$

The asymptotic Variance (ASV)

$$ASV(\hat{\theta}) = I^{-1}(\theta) \tag{43}$$

where the asymptotic variance (ASV), obtained from the inverse of the observed information matrix, and the finite sample variance (FSV),

$$FSV(\hat{\theta}) = E\left(\hat{\theta} - \bar{\hat{\theta}}\right)^2 \tag{44}$$

6. Repeat steps (2), (3), (4), and (5), 2000 times.

Where $\hat{\theta} = (\hat{a}, \hat{b}, \hat{\lambda}, \hat{c})$ is the maximum likelihood estimators of $\theta = (a, b, \lambda, \mathrm{c})$. The t value measures the significance of the bias. The bias is considered significant if $|t| > 1.96$. The asymptotic variance of $\theta = (a, b, \lambda, \mathrm{c})$ is calculated from the asymptotic approximation of the variance given by the inverse of the observed information matrix for the i -th simulated sample (Hinkley (1978) and Sprott (1980)). The finite sample variance is calculated from the values of the estimators of the distribution in the simulation study, and they represent the actual variance of the estimators. The approximation is valid because asymptotic variance decreases as sample size increases, also because the finite sample variance and asymptotic variance agrees closely with one another (Jennings, (1986)). In determining a good estimator it is also important to study the bias and variances of the estimators. Bias is much less important than the variance, and the main contribution to the mean squared error comes from the variance.

Table 1. Statistical properties of the MLE for the parameters of the beta Pareto distribution

Sample Size				
10	a	b	λ	c
Bias	1.1897473	1.1987473	0.7765018	-0.3039930
t	3.7685034	5.0027368	2.6693100	-0.0146470
ASV	0.2962686	0.0633647	-4.0702400	0.0100182
FSV	0.0154190	0.0028167	0.0341668	0.0080601
MSE	1.2036930	1.4349639	0.6759438	0.1004721
20				
Bias	1.0820847	1.1960311	0.7721289	-0.3051750
t	2.6329270	3.5852405	0.8101258	-0.0104680
ASV	0.1203950	0.0281434	-4.8003100	0.0029462
FSV	0.0130163	0.0007492	0.0267757	0.0052063
MSE	1.1839237	1.4312397	0.6681753	0.0983379
30				
Bias	1.0659877	1.1956768	0.7681117	-0.2907280
t	2.0869101	2.9306000	0.7573450	-0.0091850
ASV	0.0958127	0.0266895	-3.2867700	0.0028018
FSV	0.0146839	0.0008887	0.0283523	0.0057960
MSE	1.1783918	1.4292709	0.6371219	0.0960453
50				
Bias	1.0550919	1.1956061	0.7648447	-0.2856880
t	1.5906622	2.2851957	0.4007003	-0.0068300
ASV	0.0632573	0.0147032	-0.6622140	0.0016095
FSV	0.0116556	0.0007036	0.0215396	0.0045580
MSE	1.1650871	1.4281414	0.6229589	0.0949251
80				
Bias	1.0493616	1.1942572	0.7108231	-0.2015460
t	1.3102052	1.7892695	0.4022143	-0.0046030
ASV	0.0387269	0.0085875	-4.8407600	0.0009300
FSV	0.0133702	0.0004587	0.0185098	0.0051151
MSE	1.1510137	1.4267089	0.6183479	0.0903200
100				
Bias	1.0473837	1.1932963	0.7040942	-0.2003250
t	1.1624449	1.6043640	0.2860630	-0.0040780
ASV	0.0266321	0.0053432	-5.8658000	0.0005939
FSV	0.0119602	0.0005376	0.0216078	0.0047301
MSE	1.1248745	1.4183154	0.6065270	0.0861756

The results given in Table 1 indicate that.

1. The estimates of the c parameter have insignificant bias for all sample sizes. The estimates of the a parameter are significantly biased for small samples ($n < 50$), while it is insignificant for samples greater than 30. The estimates of the b parameter are significantly biased for small samples ($n \leq 50$), while it is insignificant for samples greater than 50. The estimates of the λ parameter are insignificantly biased for all sample sizes except ($n = 10$). These are indicated by the t values for the estimators. The bias is negative for the c parameter, while it is positive for the other parameters. c estimator has the smallest bias compared to the other estimators. In general, the bias tends to decrease as the values of the sample size increases.

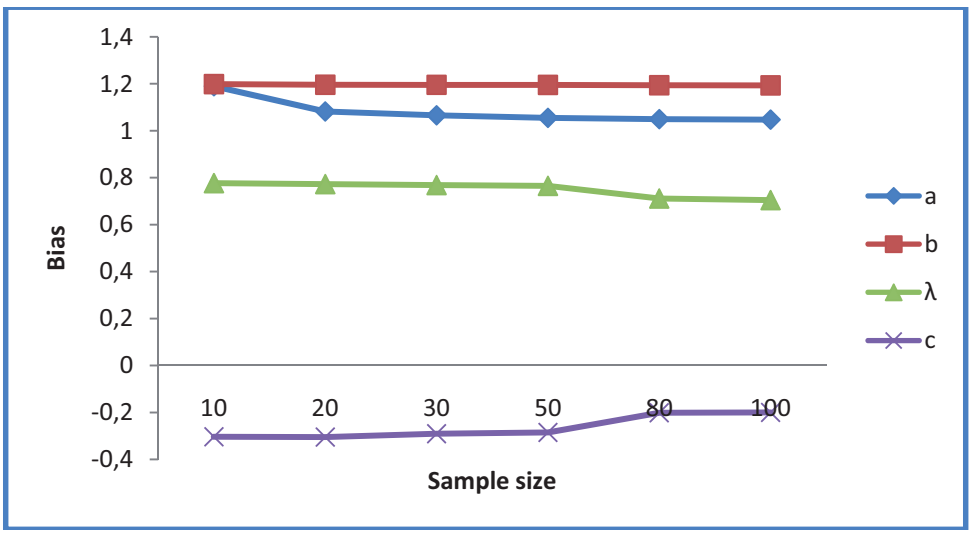


Figure 3. Relationship between the bias of the estimators and sample size.

2. The asymptotic variance of the a, b, λ and c estimators generally decreases as the sample size increases. λ estimator has the smallest ASV for all sample sizes. For the a estimator, the asymptotic variance is consistently higher.

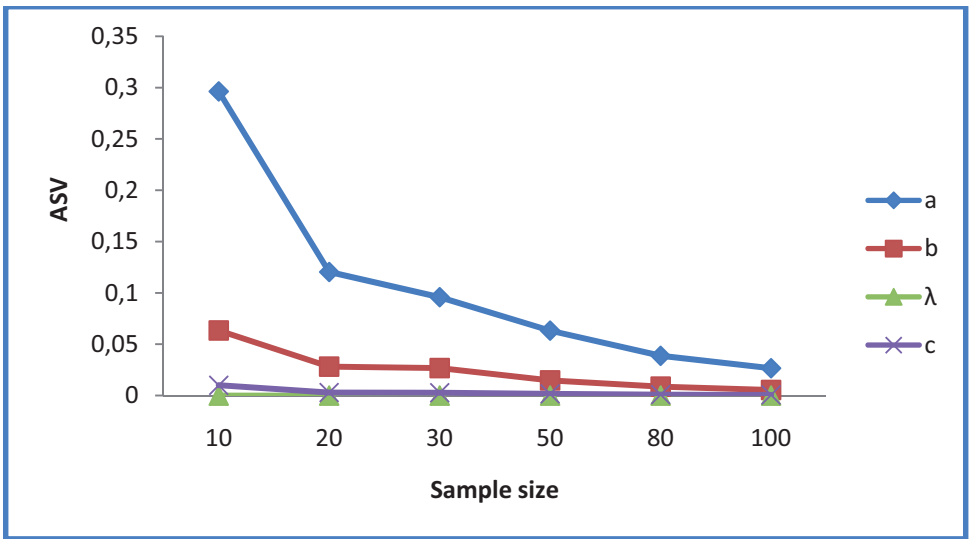


Figure 4. Relationship between the ASV of the estimators and sample size.

3. In general, the mean square error of the a, b, λ and c estimators decreases as the sample size increases as shown in Figure 5. c estimator has the smallest MSE values compared to the other estimators.

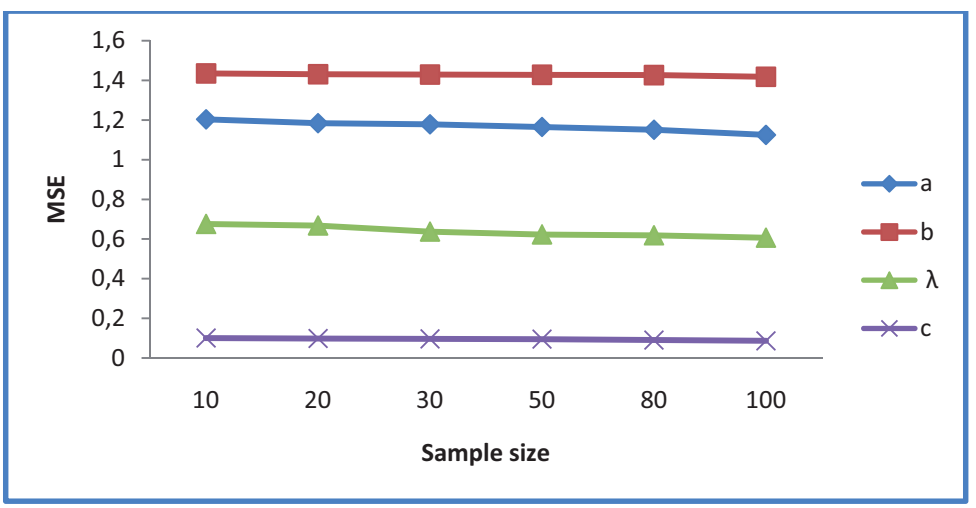


Figure 5. Relationship between the MSE of the estimators and sample size.

8. Conclusions

We have studied the beta Pareto distribution which arise by taking G in (1) to be Pareto. We have derived various properties of the beta Pareto distribution, including the mean, variance, skewness, kurtosis, the mean deviation about the mean, the mean deviation about the median, Rényi and Shannon entropies, the estimation procedures by the methods of moments and maximum likelihood. We have shown that beta Pareto distributions are the most tractable of all the known distributions out of (1). The results presented in this paper can be used as a reference to obtain the corresponding results for other distributions belonging to (1).

Acknowledgements

The author would like to thank the referee and the Editor-in-Chief for carefully reading the paper and for their great help in improving the paper.

REFERENCES

BROWN, B.W, SPEARS, F.M and LEVY, L.B. (2002): The log F: a distribution for all seasons. Comput. Statist., 17, 47–58.

EUGENE, N., LEE, C. and FAMOYE, F. (2002): Beta-normal distribution and its applications. Commun Statist—Theory Methods. 31, 497–512.

GUPTA, A.K. and NADARAJAH, S. (2004): On the moments of the beta normal distribution. Commun Statist—Theory Methods. 33, 1–13.

HINKLEY, D. (1978): Likelihood inference about location and scale parameters. Biometrika. 65(2), 253-261.

JENNINGS, D. (1986): Judging inference adequacy in logistic regression. JASA. 81(394), 471-476.

JONES, M. C. (2004): Families of distributions arising from distributions of order statistics. Test. 13, 1–43.

NADARAJAH, S. and KOTZ, S. (2004): The beta Gumbel distribution. Math Probab Eng. 10, 323–332.

RÉNYI, A. (1961): On measures of entropy and information. In: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability. vol. 1. Berkeley: University of California Press. p. 547–61.

SONG, K. S. (2001): RÉNYI information, log likelihood and an intrinsic distribution measure. J Statisst Plan Inference. 93, 51–69.

SPROTT, D. (1980): Maximum likelihood in small samples: Estimation in the presence of a nuisance parameters. Biometrija. 67(3), 515-523.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 115—126

ROBUSTNESS OF THE CONFIDENCE INTERVAL FOR AT-RISK-OF-POVERTY RATE

Wojciech Zieliński[1]

ABSTRACT

Zieliński (2009) constructed a nonparametric interval for *At-Risk-of-Poverty Rate*. It appeared that the confidence level of the interval depends on the underlying distribution of the income. For some distributions (e.g. lognormal, gamma, Pareto) the confidence level may be smaller than the nominal one. The question is, what is the largest deviance from the nominal level?

In the paper, a more general problem is considered, i.e. the problem of robustness of the confidence level of the confidence interval for binomial probability. The worst distribution is derived as well as the smallest true confidence level is calculated. Some asymptotic remarks (sample size tends to infinity) are also given.

Key words: ARPR, binomial distribution, confidence interval, robustness.

1. Confidence interval for ARPR

In the European Commission Eurostat document Doc. IPSE/65/04/EN page 11, the *„at-risk-of-poverty rate"* (ARPR) is defined as the percentage of population with the equivalised disposable income below 60% of calculated median value, i.e. ARPR is defined as

$$ARPR = F(0.6 \cdot Q(0.5)),$$

where F denotes the distribution of the population income ($Q(0.5)$ stands for the median of the distribution F).

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. Let $X_{1:n} \le \cdots \le X_{n:n}$ denote the ordered sample. As an estimator of the population median we take $X_{M:n}$,

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

where $M = \mathrm{d}n/2\mathrm{t}+1$ ($\mathrm{d}a\mathrm{t}$ is the greatest integer not greater than a). Let ξ be the number of observations not greater than $0.6 \cdot X_{M:n}$:

$$\xi = \#\{X_i \le 0.6 \cdot X_{M:n}\}.$$

Here, $\#S$ denotes the cardinality of the set S. The estimator $\widehat{ARPR}$ is defined:

$$\widehat{ARPR} = \frac{\xi}{n}.$$

The properties of the above estimator depends strongly on the distribution F of population income. Zieliński (2009) showed that the distribution of ξ is almost binomial with parameters $M-1$ and $2 \cdot ARPR$. Zieliński (2006) showed that the estimator is almost unbiased, i.e.

$$E_F \widehat{ARPR} \approx ARPR$$

for all continuous F. He also calculated its variance:

$$D_F^2 \widehat{ARPR} \approx \frac{1}{n} ARPR(1-ARPR).$$

The nonparametric interval for *ARPR* at a confidence level $\gamma \in (0,1)$ is based on the random variable ξ (Zieliński 2009):

$$\left(0.5B^{-1}\left(\xi, M-\xi+1; \frac{1-\gamma}{2}\right); 0.5B^{-1}\left(\xi+1, M-\xi; \frac{1-\gamma}{2}\right)\right), (*)$$

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

The above confidence interval is a special case of the Clopper and Pearson (1934) confidence interval for binomial proportion. In the binomial statistical model

$$\left(\{0,1,\ldots,n\}, \{Bin(n,q), 0<q<1\}\right),$$

where $Bin(n,q)$ denotes binomial distribution, the confidence interval for q is of the form

$$\left(\beta^{-1}\left(\eta, n-\eta+1; \frac{1-\gamma}{2}\right); \beta^{-1}\left(\eta+1, n-\eta; \frac{1+\gamma}{2}\right)\right).$$

Here, η denotes the random variable distributed as $Bin(n,q)$.

2. Robustness of the confidence level

Suppose now that η is not distributed as $Bin(n,q)$, but its distribution is given by the cdf function F from a class F_q such that $F \in F_q$ iff

- F is discrete: $P_F(i) = p_i \geq 0$ for $i \in \{0,1,\ldots,n\}$;
- $\sum_{i=0}^{n} p_i = 1$;
- $\sum_{i=0}^{n} i p_i = nq$;
- $\sum_{i=0}^{n} i^2 p_i = nq(1-q) + (nq)^2$.

In the class F_q there are distributions with the same expected value and variance as binomial $Bin(n,q)$. We want to calculate

$$\inf_{q\in[0,1]} \inf_{F\in F_q} P_F\left\{q \in \left(\beta^{-1}\left(\eta, n-\eta+1; \frac{1-\gamma}{2}\right); \beta^{-1}\left(\eta+1, n-\eta; \frac{1+\gamma}{2}\right)\right)\right\},$$

i.e.

$$\inf_{q\in[0,1]} \inf_{F\in F_q} \sum_{i=\lfloor x_1(q)\rfloor}^{\lfloor x_2(q)+1\rfloor} p_i,$$

where $\lfloor x_1(q)\rfloor$ and $\lfloor x_2(q)\rfloor$ are such that

$$\beta(x_1(q)+1, n-x_1(q); q) = \frac{1+\gamma}{2} \quad \text{and} \quad \beta(x_2(q), n-x_2(q)+1; q) = \frac{1-\gamma}{2}.$$

For any given q (because of the symmetry of the problem, we consider $0<q<0.5$), we have a linear programming problem with respect to $p_0, p_1, \ldots, p_n$:

$$\begin{cases} \sum_{i=\lfloor x_1(q)\rfloor}^{\lfloor x_2(q)+1\rfloor} p_i = \min! \\ \sum_{i=0}^{n} p_i = 1 \\ \sum_{i=0}^{n} i p_i = nq \\ \sum_{i=0}^{n} i^2 p_i = nq(1-q) + (nq)^2 \\ 0 \leq p_i \leq 1, i = 0,1,\ldots,n \end{cases}$$

The solution of this problem is a one of vertices of the appropriate simplex, i.e. the point such that exactly three probabilities are not zero and others equal zero. Each vertex is a solution of the system of linear equations

$$\begin{bmatrix} 1 & 1 & 1 \\ a & b & c \\ a^2 & b^2 & c^2 \end{bmatrix} \begin{bmatrix} p_a \\ p_b \\ p_c \end{bmatrix} = \begin{bmatrix} 1 \\ w \\ z \end{bmatrix},$$

where *a, b, c* are integers such that $0 \le a < b < c \le n$, $w = nq$, $z = nq(1-q) + (nq)^2$. The solution is

$$p_a(a,b,c) = -\frac{-bc + (b+c)w - z}{(a-b)(a-c)},$$

$$p_b(a,b,c) = -\frac{ac - (a+c)w + z}{(a-b)(b-c)},$$

$$p_c(a,b,c) = -\frac{-ab + (a+b)w - z}{(a-c)(b-c)},$$

Among all solutions we have to choose one such that $0 < p_a(a,b,c), p_b(a,b,c), p_c(a,b,c) < 1$. It is enough to find *a, b, c* such that $p_a(a,b,c), p_b(a,b,c), p_c(a,b,c) > 0$. Solving the appropriate system of inequalities we obtain:

$$\begin{cases} 0 \le a < b \le (n-1)q \;\&\; nq\left(1 + \dfrac{1-q}{nq-a}\right) < c < nq\left(1 + \dfrac{1-q}{nq-b}\right) \\ \text{or} \\ 0 \le a \le (n-1)q \;\&\; (n-1)q < b \le nq\left(1 + \dfrac{1-q}{nq-a}\right) \;\&\; nq\left(1 + \dfrac{1-q}{nq-a}\right) < c \le n \end{cases} \quad (*)$$

We are looking for values *a, b, c* satisfying (*), such that, if F is concentrated in those points, then

$$P_F\left\{q \in \left(\beta^{-1}\left(\eta, n-\eta+1; \frac{1-\gamma}{2}\right); \beta^{-1}\left(\eta+1, n-\eta; \frac{1+\gamma}{2}\right)\right)\right\}, \quad (P)$$

is the smallest. Note that

$$(P)=\begin{cases} 1 \quad if\ \lfloor x_1(q)t\rfloor \le a<b<c\le \lfloor x_2(q)t\rfloor+1 & A\\ \mathrm{p_a}(\mathrm{a},\ \mathrm{b},\ \mathrm{c})\ +\ \mathrm{p_b}(\mathrm{a},\ \mathrm{b},\ \mathrm{c}), \quad if\ \lfloor x_1(q)t\rfloor \le a<b\le \lfloor x_2(q)t\rfloor+1<c & B\\ \mathrm{p_a}(\mathrm{a},\ \mathrm{b},\ \mathrm{c}), \quad if\ \lfloor x_1(q)t\rfloor \le a\le \lfloor x_2(q)t\rfloor+1<b<c, & C\\ \mathrm{p_b}(\mathrm{a},\ \mathrm{b},\ \mathrm{c})\ +\ \mathrm{p_c}(\mathrm{a},\ \mathrm{b},\ \mathrm{c}), \quad if\ a<\lfloor x_1(q)t\rfloor \le b<c\le \lfloor x_2(q)t\rfloor+1, & D\\ \mathrm{p_b}(\mathrm{a},\ \mathrm{b},\ \mathrm{c}), \quad if\ a<\lfloor x_1(q)t\rfloor \le b\le \lfloor x_2(q)t\rfloor+1<c, & E\\ \mathrm{p_c}(\mathrm{a},\ \mathrm{b},\ \mathrm{c}), \quad if\ a<b<\lfloor x_1(q)t\rfloor \le c\le \lfloor x_2(q)t\rfloor+1, & F\\ 0, \quad elswhere \end{cases}$$

Case (A) is not interesting. Case (G) does not hold because $\lfloor x_1(q)t\rfloor < nq < \lfloor x_2(q)t\rfloor+1$ and hence at least one inequality holds

$$\lfloor x_1(q)t\rfloor < a < \lfloor x_2(q)t\rfloor+1, \quad \lfloor x_1(q)t\rfloor < b < \lfloor x_2(q)t\rfloor+1, \quad \lfloor x_1(q)t\rfloor < c < \lfloor x_2(q)t\rfloor+1.$$

It easy to see that the minimum of (P) is achieved in case (E), i.e. for *a, b, c* such that

$$a<\lfloor x_1(q)\rfloor \le b\le \lfloor x_2(q)t\rfloor+1<c$$

and $q\in(0,0.5)$ such that

$$\lfloor x_1(q)t\rfloor \ge 1.$$

So we consider probabilities q from the interval $(q(n),0.5)$, where $q(n)$ is the solution of

$$\beta(2,n-1;q(n))=\frac{1+\gamma}{2} \quad \text{i.e.} \quad \frac{1}{B(2,n-1)}\int_0^{q(n)} u(1-u)^{n-2}du=\frac{1+\gamma}{2}.$$

Exemplary values of $q(n)$ are given in the Table 1.

Table 1.

n	10	50	100	500	1000
q(n)	0.482497	0.108542	0.054997	0.011115	0.005564

The distribution minimizing (P) for a given q is given in theorem 1.

Theorem 1. Let $q \in (q(n), 0.5)$. Then

$$P_F\left\{q \in \left(\beta^{-1}\left(\eta, n-\eta+1; \frac{1-\gamma}{2}\right); \beta^{-1}\left(\eta+1, n-\eta; \frac{1+\gamma}{2}\right)\right)\right\}$$

reaches minimum at the distribution F concentrated in points

$$a = \mathrm{dx}_1(q)\mathrm{t} - 1,\ b = \frac{\mathrm{dx}_1(q)\mathrm{t} + \mathrm{dx}_2(q)\mathrm{t} + 1}{2},\ c = \mathrm{dx}_2(q)\mathrm{t} + 2$$

Proof. Proof of the theorem is given in the Appendix.

For a given q the distribution F minimizing probability (P) (or maximizing probability $prob(a,c)$) may be found. This distribution is concentrated in points

$$a = \max\left\{d \in \mathbf{N}: \beta^{-1}\left(d+1, n-d; \frac{1+\gamma}{2}\right) \le q\right\}, c = \min\left\{d \in \mathbf{N}: \beta^{-1}\left(d, n-d+1; \frac{1-\gamma}{2}\right) \ge q\right\}$$

and $b = \frac{a+c}{2}$. It is clear that (P) does not depend only on F but also on q. Of course, it is because of discreetness of considered distributions. For example, let $n = 20$, $a = 0$ and $c = 8$. Then, $b = 4$, $\mathrm{dx}_1(q)\mathrm{t} = 1$ and $\mathrm{dx}_2(q)\mathrm{t} + 1 = 7$ and for $(\gamma = 0.95)$

$$q \in \left(\beta^{-1}\left(1, 20; \frac{1+\gamma}{2}\right), \beta^{-1}\left(7, 13; \frac{1-\gamma}{2}\right)\right) = (0.16843, 0.19119)$$

confidence level of the confidence interval equals

$$\sum_{i=1}^{7} p_i = p_4(0,4,8) = q\frac{35-95q}{4}.$$

Values of the above probability are calculated in the Table 2.

Table 2.

q	p_4(0,4,8)	prob(0,8)	q	p_4(0,4,8)	prob(0,8)
0.169	0.80043	0.19957	0.181	0.80568	0.19432
0.170	0.80113	0.19888	0.182	0.80581	0.19420
0.171	0.80178	0.19822	0.183	0.80589	0.19411
0.172	0.80238	0.19762	0.184	0.80592	0.19408
0.173	0.80294	0.19706	0.185	0.80591	0.19409
0.175	0.80391	0.19609	0.186	0.80585	0.19416
0.176	0.80432	0.19568	0.187	0.80574	0.19426
0.177	0.80469	0.19531	0.188	0.80558	0.19442
0.178	0.80501	0.19500	0.189	0.80538	0.19462
0.179	0.80528	0.19472	0.190	0.80513	0.19488
0.180	0.80550	0.19450	0.191	0.80483	0.19517

It may be seen that minimal confidence level, i.e. maximal $prob(a,c)$, is attained at the $q = \beta^{-1}\left(1,20;\frac{1+\gamma}{2}\right)$. This fact is generalized in the theorem 2.

Theorem 2. Let a be given and let

$$q_1 = \beta^{-1}\left(a+1, n-a; \frac{1+\gamma}{2}\right) \quad \text{and} \quad q_2 = \beta^{-1}\left(c, n-c+1; \frac{1-\gamma}{2}\right),$$

where $c = \min\left\{d \in \mathbf{N} : \beta^{-1}\left(d, n-d+1; \frac{1-\gamma}{2}\right) \geq q_1\right\}$. Then, $prob(a,c)$ is reached for q_1.

Proof. See Appendix

In order to find maximum of the probability $prob(a,c)$ it must be calculated for

$$a = 0,1,\ldots,\frac{n-1}{2}, \beta^{-1}\left(a+1, n-a; \frac{1+\gamma}{2}\right) \text{and c given in Theorem 2.}$$

For given n and γ it may be found by numerical solution. In the Table 3 there are calculated the minimal confidence level (*min*), nominal (*nom*) confidence level (i.e. in binomial model) and $percent = min / nom$.

Table 3.

n	min	nom	percent		n	min	nom	percent
10	0.8118	0.9625	0.8434		100	0.7645	0.9504	0.8044
15	0.7935	0.9517	0.8338		200	0.7573	0.9500	0.7972
20	0.7988	0.9580	0.8338		300	0.7547	0.9502	0.7942
30	0.7833	0.9506	0.8240		400	0.7540	0.9503	0.7935
40	0.7756	0.9503	0.8162		500	0.7534	0.9500	0.7930
50	0.7777	0.9533	0.8157		600	0.7532	0.9503	0.7926
60	0.7699	0.9503	0.8101		700	0.7531	0.9501	0.7926
70	0.7691	0.9508	0.8089		800	0.7524	0.9502	0.7918
80	0.7687	0.9513	0.8080		900	0.7520	0.9502	0.7914
90	0.7671	0.9512	0.8065		1000	0.7518	0.9501	0.7913

For large n, the binomial distribution for $q \in (q(n), 0.5)$ is almost normal $N(nq, nq(1-q))$ and

$$a, c \approx nq \pm u\sqrt{nq(1-q)},$$

where u is the $\frac{1+\gamma}{2}$ quantile of $N(0,1)$. It is easy to check that

$$prob(a,c) = \frac{1}{u^2}$$

and asymptotic minimal confidence level is $1-\frac{1}{u^2}$. For example, if $\gamma = 0.95$ then $u = 1.96$, so asymptotic minimal confidence level is 0.73969.

3. Conclusions

The confidence level of the Clopper-Pearson confidence interval for probability in binomial model depends on the distribution from the class F_q. The minimal probability of coverage was calculated. It appeared that the most unfavourable distribution is a three point one.

From practical point of view it may be interesting to consider another family of distributions, namely a class $F_q^{\varepsilon} \subset F_q$: a distribution $F \in F_q^{\varepsilon}$ iff $|F(x) - q(x)| < \varepsilon$ for all $x = 0,1,\ldots,n$. Here, ε is a given number. Such considerations are in progress.

4. Appendix

The proof of the Theorem 1 is given in two following lemmas.

Lemma 1. For given *a, c* satisfying $(*)$, the function $b \to p_b(a,b,c)$ achieves minimum for $b = \frac{a+c}{2}$.

Proof. Assume that b is real. Then

$$\frac{\partial p_b(a,b,c)}{\partial b} = \frac{(a-2b+c)(a(c-w)-cw+z)}{(a-b)^2(b-c)^2}.$$

For *a, c* from $(*)$

$$a(c-w)-cw+z = (a-w)c+z-aw < 0.$$

Hence $\frac{\partial p_b(a,b,c)}{\partial b} < 0$ for $b < \frac{a+c}{2}$ and $\frac{\partial p_b(a,b,c)}{\partial b} > 0$ for $b > \frac{a+c}{2}$.

Lemma 2. Let $b = \frac{a+c}{2}$. Then

$\max_{a,c}(p_a(a,b,c)+p_c(a,b,c)) = p_a(a^*,b,c^*)+p_c(a^*,b,c^*)$ for

$a^* = \max\{a < \frac{nw-z}{n-w}$ and $c^* = \min\{c > \frac{aw-z}{a-w}$.

Proof. Assume for a while that a and c are real numbers. We have

$$prob(a,c) = p_a(a,b,c)+p_c(a,b,c) = \frac{a^2+c^2+2a(c-2w)-4cw+4z}{(c-a)^2} = \frac{(a+c-2w)^2+4(z-w^2)}{(c-a)^2}.$$

The proof of the Lemma will be given in two steps

- for all $c > \frac{aw-z}{a-w}$ function $prob(a,c)$ is increasing with respect to a;

$*$ for all $a < \frac{nw-z}{n-w}$ function $prob(a,c)$ for $c > \frac{aw-z}{a-w}$ achieves maximum for the smallest c.

Proof of $\bullet$: we will show that

$$\frac{\partial}{\partial a} prob(a,c) = -\frac{4(c^2 + a(c-w) - 3cw + 2z)}{(a-c)^3} > 0 \quad \text{for} \quad 0 < a < \frac{nw-z}{n-w}.$$

Because $a<c$, it is enough to show that the nominator is positive. The nominator takes on the form $a(c-w) + (c^2 - 3cw + 2z)$. Because $c > \frac{aw-z}{a-w} > \frac{z}{w} > w$, one has to show that $c^2 - 3cw + 2z > 0$. Determinant $\Delta = 9w^2 - 8z = 8w(1-q) + w^2 > 0$. So

$$\frac{3w + \sqrt{9w^2 - 8z}}{2} < \frac{z}{w} \Rightarrow \sqrt{9w^2 - 8z} < \frac{2z - 3w^2}{w} \Rightarrow z(z - w^2) > 0.$$

Proof of $*$: we have

$$\frac{\partial}{\partial c} prob(a,c) = \frac{4(a^2 + a(c-3w) - cw + 2z)}{(a-c)^3} = 4\frac{c(a-w) + a^2 - 3aw + 2z}{(a-c)^3}.$$

Because $a<c$ hence the denominator is negative, and the sign of the derivative depends on the sign of the nominator. The nominator is a linear function of c. It is enough to look at the sign of the nominator at the points $c = \frac{aw-z}{a-w}$ and $c = n$

For $c = \frac{aw-z}{a-w}$ the nominator equals

$$a^2 + a(n-3w) + 2z - nw.$$

This a quadratic function of a with zeros at

$$a_{1,2} = \frac{(3w-n) \pm \sqrt{(n-w)^2 + 8(w^2 - z)}}{2}.$$

We show that

$$\frac{nw-z}{n-w} - 1 < a_2 < \frac{nw-z}{n-w}.$$

We have

$$\frac{nw-z}{n-w}-a_2=\frac{nw-z}{n-w}-\frac{(3w-n)+\sqrt{(n-w)^2+8(w^2-z)}}{2}$$
$$=\frac{n-w}{2}\left[1+2\frac{w^2-z}{(n-w)^2}-\sqrt{1+8\frac{w^2-z}{(n-w)^2}}\right]$$
$$=\frac{n(1-q)}{2}\left[1-2\frac{q}{n(1-q)}-\sqrt{1-8\frac{q}{n(1-q)}}\right].$$
$$(\acute{\imath})$$

It is easy to see that $0<(\acute{\imath})<1$ for $n>\left[\frac{q(1+q)}{1-q}\right]^2$, i.e. for $n\geq 2$.

Hence, for $c=n$ the nominator is negative, so the derivative is positive. The function $prob(a,c)$ attains its maximum at $c=\frac{aw-z}{a-w}$ or $c=n$. We show that for $a=0,1,\ldots,\lfloor\frac{aw-z}{a-w}\rfloor-1$ and $q>\beta^{-1}\left((a+1)+1,n-(a+1);\frac{1+\gamma}{2}\right)$ holds $prob\left(a,\frac{aw-z}{a-w}\right)>prob(a,n)$. We have

$$prob\left(a,\frac{aw-z}{a-w}\right)-prob(a,n)=\frac{4n(1-q)((n-1)q-a)}{(a-n)^2}.$$

For ``reasonable" γ we have $\beta^{-1}\left((a+1)+1,n-(a+1);\frac{1+\gamma}{2}\right)>\frac{a}{n-1}$. Hence

$$prob\left(a,\frac{aw-z}{a-w}\right)>prob(a,n).$$

Numbers a and c are natural, so we obtain thesis from • and $*$.

Proof of the Theorem 2. For all $q\in(q_1,q_2)$ bounds $\lfloor x_1(q)\rfloor$ and $\lfloor x_2(q)\rfloor+1$ of confidence interval remains the same. Hence, the probability $prob(a,c)$ depends only on q. Because $w=nq$, $z=nq(1-q)+(nq)^2$, so

$$prob(a,c) = \frac{(a+c-2w)^2 + 4(z-w^2)}{(c-a)^2} = \frac{(a+c-2nq)^2 + 4nq(1-q)}{(c-a)^2}$$
$$= \frac{(a+c)^2 - 4n(a+c-1)q + 4n(n-1)q^2}{(c-a)^2}.$$

This is a quadratic function of q and it achieves its maximum at q_1 or q_2.

REFERENCES

CLOPPER C. J., PEARSON E. S. (1934), The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial, Biometrika, 26, 404-413.

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

ZIELIŃSKI W. (2009), A nonparametric confidence interval for At-Risk-of-Poverty-Rate, Statistics in Transition new series, 10, 437-444.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 127—138

ESTIMATION OF DOMAIN MEANS ON THE BASIS OF STRATEGY DEPENDENT ON DEPTH FUNCTION OF AUXILIARY VARIABLES' DISTRIBUTION

Janusz L. Wywiał[1]

ABSTRACT

The paper deals with the problem of estimation of a domain means in a finite and fixed population. We assume that observations of a multidimensional auxiliary variable are known in the population. The proposed estimation strategy consists of the well known Horvitz-Thompson estimator and the non-simple sampling design dependent on a synthetic auxiliary variable whose observations are equal to the values of a depth function of the auxiliary variable distribution. The well known spherical and Mahalanobis depth functions are considered. A sampling design is proportionate to the maximal order statistic determined on the basis of the synthetic auxiliary variable observations in a simple sample drawn without replacement. A computer simulation analysis leads to the conclusion that the proposed estimation strategy is more accurate for domain means than the well known simple sample means.

Key words: sampling design, order statistic, auxiliary variable, sampling scheme, Horvitz-Thompson estimator, small area estimation, area sampling, depth function

1. Sampling strategy

Let $U=(1,2,...,N)$ be a fixed population of the size N. An observation of a variable under study (an auxiliary variable) attached to the i-th population element will be denoted by y_i (z_i), $i=1,...,N$. The sample of the size n, drawn without replacement from the population, will be denoted by s. The sampling design is denoted by $P(s)$ and inclusion probabilities of the first and second orders – by π_k, for k=1,....,N and $\pi_{k,t}$ for k≠t, k=1,...N, t=1,...N, respectively. Let $\boldsymbol{S}$ be the sample space of the samples of size n, drawn without replacement. We are

[1] Department of Statistics, Katowice University of Economics, Poland.

going to consider the following sampling designs of simple samples drawn without replacement: $P_0(s)=\binom{N}{n}^{-1}$ for all $s\in\boldsymbol{S}$.

There are several sampling designs on the value of a non-negative auxiliary variable. Some reviews of them are presented, e.g. by Tillé (2006). We are going to consider the following one analysed by Wywiał (2007, 2008, 2009).

Let $Z_{(r)}$ be the r-th order statistic from a simple sample drawn without replacement where the auxiliary variable's values are observed. Let $\alpha\in(0;1)$ and $[n\alpha]$ be the integer part of the value $n\alpha$. The sample quantile of the order α is defined as $Q_{s,\alpha}=Z_{(r)}$ where $r=[n\alpha]+1$ and $(r-1)/n\le\alpha<r/n$. Let $G(i,r)=\{s: Z_{(r)}=x_i\}$ be the set of all samples s whose r-th order statistic of the auxiliary variable is equal to x_i where $r\le i\le N-n+r$. Let $U_1=(1,...,i-1)$ be a subpopulation of the population U and let s_1 be the simple sample of the size $(r-1)$, drawn without replacement from U_1. Similarly, let $U_2=(i+1,...,N)$ be a subpopulation of the population U and let s_2 be a simple sample of the size $(n-r)$, drawn without replacement from U_2. The sample space of the samples of the type s_1 will be denoted by $\mathbf{S}(U_1,s_1)$ and the sample space of the samples of the type s_2 will be denoted by $\mathbf{S}(U_2,s_2)$. Let $s=(s_1\cup\{i\}\cup s_2)$ be such a sample that the value of the r-th order statistic of an auxiliary variable - observed in the sample - equals x_i. Hence, the sample space of such a sample s is as follows: $\mathbf{S}_{r,i}=\mathbf{S}(U_1,s_1)\times\{i\}\times\mathbf{S}(U_2,s_2)$, where $\times$ is the symbol of the Cartesian product. So, the sample space for all $i=r,...,N-n+r$ is as follows $\mathbf{S}=\bigcup_{i=r}^{N-n+r}\mathbf{S}_{r,i}$.

Wywiał (2008) proposed the following sampling design proportional to the z_i value of the $Z_{(r)}$ statistic.

$$P_1(s\mid r)=\frac{Z_{(r)}}{\sum_{j=r}^{N-n+r}\binom{j-1}{r-1}\binom{N-j}{n-r}z_j}\qquad\text{for } s\in\boldsymbol{S},\qquad(1)$$

or

$$P_1(s\mid r)=\frac{z_i}{\sum_{j=r}^{N-n+r}\binom{j-1}{r-1}\binom{N-j}{n-r}z_j}\qquad\text{for } s\in\boldsymbol{S}_{r,i},\ i=r,...,N-n+r.$$

The conditional version of the sampling design is as follows.

$$P_1(s \mid z_u \le Z_{(r)} \le z_v) = P_1(s \mid r;u,v) = \frac{Z_{(r)}}{\sum_{j=u}^{v} \binom{j-1}{r-1}\binom{N-j}{n-r} z_j} \qquad \text{for } s \in \boldsymbol{S}_{r/u,v},$$

where $\mathbf{S}_{\mathbf{r}/u,v} = \bigcup_{i=u}^{v} \mathbf{S}_{r,i}$. Equivalently,

$$P_1(s \mid r;u,v) = \frac{z_i}{\sum_{j=u}^{v} \binom{j-1}{r-1}\binom{N-j}{n-r} z_j} \quad \text{for } s \in \boldsymbol{S}_{r,i},\ i=r,\ldots,N-n+r. \tag{2}$$

for $s \in G(i,r)$ and $r \le u \le i \le v \le N-n+r$.

Let us define such a function $\delta(t)$ that if $t \le 0$ ($t>0$) then $\delta(t)=0$ ($\delta(t)=1$). Moreover, let

$$a_r(u,v) = \sum_{j=u}^{v} z_j \binom{j-1}{r-1}\binom{N-j}{n-r} \tag{3}$$

The inclusion probabilities of the first order are as follows:

$$\pi_k = \frac{\delta(u-k)\delta(r-1)\delta(v-1)\delta(u-1)}{a_r(u,v)} \sum_{i=u}^{v} \binom{i-2}{r-2}\binom{N-i}{n-r} z_i +$$

$$+ \frac{\delta(k-u+1)\delta(v-k+1)}{a_r(u,v)} \left(\delta(n-r)\delta(k-u)\delta(k-1) \sum_{i=u}^{k-1} \binom{i-1}{r-1}\binom{N-i-1}{n-r-1} z_i + \right.$$

$$\left. + \binom{k-1}{r-1}\binom{N-k}{n-r} z_k + \delta(r-1)\delta(v-k) \sum_{i=k+1}^{v} \binom{i-2}{r-2}\binom{N-i}{n-r} z_i \right) +$$

$$\frac{\delta(k-v)\delta(n-r)\delta(N-v)}{a_r(u,v)} \sum_{i=u}^{v} \binom{i-1}{r-1}\binom{N-i-1}{n-r-1} z_i \tag{4}$$

for k=1,...,N. The probabilities of the second order have been derived by Wywiał (2008).

The sampling scheme implementing the $P_2(s|r;u,v)$ conditional sampling design is as follows. Population elements are ordered according to the increasing values of the auxiliary variable. Next, the *i*-th element of the population is drawn with this probability:

$$p_1(i \mid r;u,v) = \frac{\binom{i-1}{r-1}\binom{N-i}{n-r} z_i}{a_r(u,v)}, \qquad \text{i=u,...,v} \tag{5}$$

Next, the simple sample s_1 of the size $(r-1)$ is drawn from the subpopulation U_1 and the simple sample s_2 of the size $(n-r)$ – from the subpopulation U_2. Let us note that the expression (6) shows the truncated distribution of the order statistic $Z_{(r)}$ from the sample drawn according to the sampling design $P_1(s|r;u,v)$. Hence, $p_1(i \mid r;u,v) = P_1(Z_{(r)} = z_i \mid z_u \le Z_{(r)} \le z_v)$.

The well known Horvitz-Thompson (1952) estimator is as follows.

$$\bar{y}_{HTS} = \frac{1}{N}\sum_{k=1}^{N} \frac{I_k y_k}{\pi_k} \tag{6}$$

If $I_k = 1$, if the *k*-th population element was drawn into a sample. When $I_k = 0$, the *k*-th element was not drawn into the sample. It is well known that the strategy $(\bar{y}_{HTS}, P(s))$ is unbiased for the population mean when all inclusion probabilities are positive. The expression for the variance of the Horvitz-Thompson is well known and it can be found in all survey sampling textbooks. The strategy $(\bar{y}_{HTS}, P_0(s)) = (\bar{y}_S, P_0(s))$ where $\bar{y}_S = \frac{1}{n}\sum_{i \in S} y_i$ is called a simple sample mean and its variance is: $D^2(\bar{y}_S, P_0(s)) = \frac{N-n}{Nn} v$,

2. Synthetic auxiliary variable

Now, we are going to consider an auxiliary variable with possible negative values which will be denoted by x. Let us define the following transformation of the auxiliary variable.

$$z_i = \frac{1}{1+(x_i-\bar{x})^2} \quad , \text{i=1,…,N} \tag{7}$$

So, the obtained variable z can be treated as a new synthetic auxiliary variable. Let us note that the observations of a symmetric variable x, close to its mean value $\bar{x}$, are transformed into the large values of the variable z. Let the two-dimensional variable (y,x), where y denotes the variable under study, be symmetric. So, when the variables are highly correlated, we can expect the sampling design (proportionate to values of the variable z) more frequently gives the sample with observations of the variable y close to its mean value.

Now, we are going to construct more general synthetic auxiliary variables. A multidimensional auxiliary variable will be denoted by x and its value attached to the i-th population element will be denoted by $\boldsymbol{x}_{i*}=[x_{i1}, \ldots x_{iq}]$, $i=1,\ldots,N$.

The matrix of all observations of the auxiliary variable is $\mathbf{x}=\begin{bmatrix}\mathbf{x}_{1*}\\ \ldots \\ \mathbf{x}_{N*}\end{bmatrix}=\begin{bmatrix}\mathbf{x}_{*1} & \ldots. & \mathbf{x}_{*q}\end{bmatrix}$ where $\mathbf{x}_{*j}=\begin{bmatrix}\mathbf{x}_{11}\\ \ldots \\ \mathbf{x}_{Nj}\end{bmatrix}$, $i=1,\ldots,N$. Let $\bar{\mathbf{x}}=\begin{bmatrix}\bar{\mathbf{x}}_1 & \ldots. & \bar{\mathbf{x}}_q\end{bmatrix}=\frac{1}{N}\mathbf{J}_N^T\mathbf{x}$ where $\boldsymbol{J}_N$ be a unit column-vector whose all N-elements are equal to one. Moreover, let $\mathbf{d}=\mathbf{x}-\mathbf{J}_N\bar{\mathbf{x}}$ and $\mathbf{c}(x)=\frac{1}{N-1}\mathbf{d}^T\mathbf{d}$ be the population variance-covariance matrix of the auxiliary variables. Similarly, $\mathbf{c}(x,y)=\frac{1}{N-1}\mathbf{d}^T(\mathbf{y}-\mathbf{J}_N\bar{y})$ where yT =[y1…yN]. The sample auxiliary variable observation matrix is denoted by $\boldsymbol{x}_s$ and it consists of such rows $\boldsymbol{x}_{i*}$,i∈s. So, the sample variance-covariance matrix is $\mathbf{c}_s(x)=\frac{1}{n-1}\mathbf{d}_s^T\mathbf{d}_s$, $\mathbf{c}_s(x,y)=\frac{1}{n-1}\mathbf{d}_s{}^T(\mathbf{y}_s-\mathbf{J}_N\bar{y}_s)$ and y_s is the column vector of values of the variable under study observed in the sample.

There are many definitions of depth functions, see e.g. Liu, Parelius and Singh (1999). We will consider the so called spherical depth function and the one based on the well known Mahalanobis (1936) distance. They are defined by the following equations respectively.

$$z(x_i)=\frac{1}{1+\mathbf{d}_i\mathbf{d}_i{}^T}, \; z(x_i)=\frac{1}{1+\mathbf{d}_i\mathbf{c}^{-1}(x)\mathbf{d}_i{}^T}, \qquad \text{i=1,…,N} \tag{8}$$

So, the maximal depth of an observation of the auxiliary variable is equal to one. Let us note that the synthetic variable given by the expression (8) is a one-dimensional version of the spherical depth function.

Let us suppose that the distribution of the variables $(y,\boldsymbol{x})$ where $\boldsymbol{x}=[x_1,x_2,\ldots,x_q]$ is symmetric. Moreover, let the multiple correlation coefficient between the

variable under study **y** and the *q*-dimensional auxiliary one $\boldsymbol{x}$ be high. So, the observations of the auxiliary variable $\boldsymbol{x}$, close to the mean values $\bar{\mathbf{x}}$, are transformed into high values of the synthetic variable *z*. Hence, similarly to the one-dimensional auxiliary variable, when the variables are highly correlated, we can expect the sampling design (proportionate to values of the variable *z*) more frequently gives the sample with observations of the variable y close to its mean value. Let us note that the foregoing analysis shows that the symmetry of a multidimensional auxiliary variable distribution is not necessary.

3. Accuracy comparison of estimation strategies

The accuracy of the $(t_S,P(s))$ estimation strategy was measured by means of the relative efficiency (*deff)*:

$$e = \frac{D^2(\bar{y}_{HTS}, P(s))}{D^2(\bar{y}_S, P_0(s))}$$

The parameter $D^2(\bar{y}_{HTS}, P(s))$ is assessed as the variance of the estimator's values calculated on the basis of the values of the variables observed in replicated sampling according to the sampling design *P(s)*.

Let *es* and *eM* denote the relative efficiency for the strategy based on the synthetic auxiliary variable evaluated on the basis of spherical depth function and the Mahanalobis one, respectively. Moreover, we assume that r=n, because for $r<n$ the relative efficiency coefficients are usually not greater than in the case when $n=r$, as it was shown by Wywiał (2007).

The simulation analysis was based on three–dimensional observations which were obtained on the basis of the generator of random values of three-dimensional normal distribution. The synthetic auxiliary variable was evaluated on the basis of the spherical or Mahanalobis depth functions defined by the expression (8). The distribution considered was the mixture of the three-dimensional normal random variables with the same weights equal to *1/3*. The normal distribution was dented by $N(m_1, m_2, m_3, V)$, where m_1 is the mean value of the variable under study, m_2, m_3 are mean values of auxiliary variables, and finally V is the variance-covariance matrix. The parameters of the distributions were $N(4,4,4,V_1)$, $N(4,8,8,V_3)$ and $N(4,12,4,V_3)$, where

$$V_1 = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 0.6 & 0.5 \\ 0.5 & 0.5 & 0.6 \end{bmatrix}, \qquad V_2 = \begin{bmatrix} 1 & 0.95 & 0.9 \\ 0.95 & 1 & 0.8 \\ 0.9 & 0.8 & 1 \end{bmatrix}, \quad V_3 = \begin{bmatrix} 1 & -0.9 & 0.97 \\ -0.9 & 1 & -0.8 \\ 0.97 & -0.8 & 1 \end{bmatrix}$$

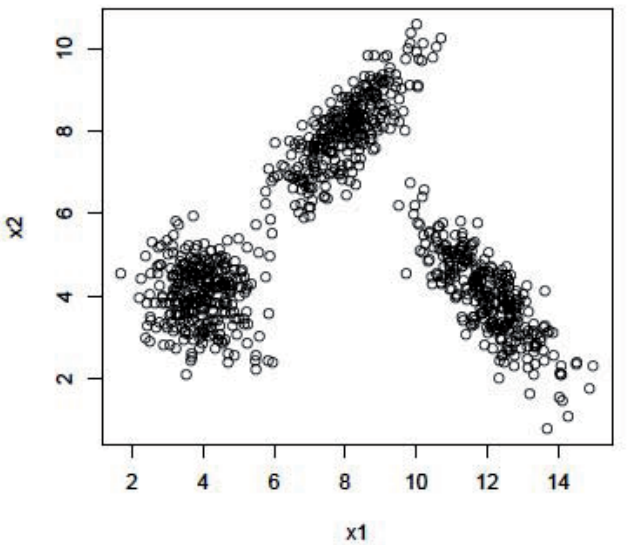


Figure 1. The spread function of the auxiliary variables.

The multiple correlation coefficients between variable under study and the auxiliary variables were calculated on the basis of the expression: $\rho_i = 1 - \det(V_i) / diag(V_i)$, i=1,2,3. Their values are ρ_1=0.913, ρ_2=0.978, ρ_3=0.992.

According to the probability distribution considered, 900 observations were generated by means of a standard program available in the package R.

The two-dimensional distribution of the auxiliary variable is the marginal distribution of the three-dimensional distribution has been just defined. It is the mixture of the following normal distributions: $N(4,4,V_{x1})$, $N(8,8,V_3)$ and $N(12,4,V_{x3})$, where

$$V_{x1} = \begin{bmatrix} 0,6 & 0,5 \\ 0,5 & 0,6 \end{bmatrix}, \quad V_{x2} = \begin{bmatrix} 1 & 0,9 \\ 0.9 & 1 \end{bmatrix}, \quad V_3 = \begin{bmatrix} 1 & 0.97 \\ 0.97 & 1 \end{bmatrix}.$$

Figure 1 shows the spread of the auxiliary variable's observations.

Table 1. The relative efficiency coefficients. The spherical depth function. The size population: N=900. The sampling scheme for r=n=u.

n	e_0	e_1	e_2	e_3
90	0,027	0,033	0,027	0,026
81	0,029	0,036	0,030	0,029
72	0,032	0,040	0,032	0,031
63	0,036	0,044	0,037	0,035
54	0,042	0,052	0,042	0,041
45	0,050	0,062	0,051	0,048
36	0,062	0,075	0,063	0,061
27	0,083	0,099	0,085	0,080
18	0,121	0,148	0,125	0,116
9	0,227	0,280	0,227	0,222

Source: Author's own calculations.

Table 2. The relative efficiency coefficients. Mahalanobis depth function. The population of the size N=900. The sampling scheme for r=n=u.

n	e_0	e_1	e_2	e_3
90	0,027	0,033	0,026	0,027
81	0,029	0,035	0,029	0,028
72	0,033	0,041	0,033	0,032
63	0,038	0,045	0,037	0,037
54	0,042	0,053	0,043	0,041
45	0,050	0,061	0,052	0,048
36	0,062	0,073	0,062	0,060
27	0,081	0,098	0,082	0,079
18	0,121	0,154	0,121	0,118
9	0,223	0,274	0,216	0,220

Source: Author's own calculations.

The analysis of the Tables 1 and 2 lead to conclusion that there is not any significant difference between the relative efficiency coefficients of the estimation strategy based on auxiliary variable dependent on spherical depth functions and the appropriate relative efficiency coefficients of the estimation strategy based on an auxiliary variable dependent on the Mahanalobis depth functions. The Tables 1 and 2, as well as Figure 2 show that those coefficients decrease when the sample size increases.

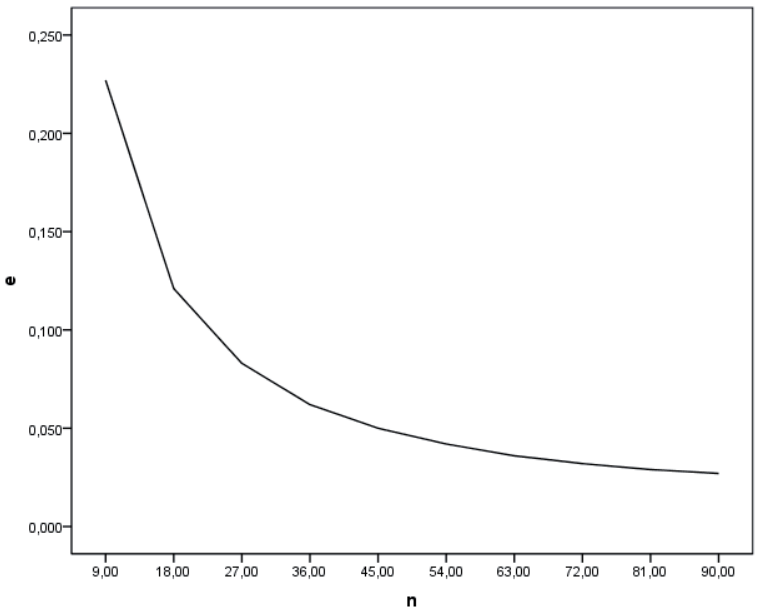


Figure 2. The relative efficiency coefficients for an increasing sample size.

Table 3. The relative efficiency coefficients. The spherical depth function. The population of the size $N=900$. The sampling scheme for $r=n=9$ and $u \geq 9$.

u	e_0	e_1	e_2	e_3
9	0,227	0,280	0,227	0,222
50	0,225	0,280	0,226	0,219
100	0,221	0,272	0,228	0,212
150	0,224	0,277	0,226	0,217
200	0,224	0,277	0,224	0,218
300	0,224	0,279	0,225	0,218
400	0,225	0,278	0,230	0,216
500	0,225	0,287	0,231	0,215
600	0,225	0,286	0,226	0,218
700	0,225	0,288	0,230	0,213
800	0,235	0,294	0,234	0,229
850	0,255	0,298	0,247	0,253

Source: Author's own calculations.

Table 4. The relative efficiency coefficients.
The Mahanalobis depth function. The population of the size *N=900*. The sampling scheme for *r=n=9* and *u≥9*.

u	e_0	e_1	e_2	e_3
9	0,223	0,274	0,216	0,220
50	0,222	0,277	0,216	0,218
100	0,220	0,283	0,214	0,215
150	0,224	0,286	0,219	0,220
200	0,222	0,276	0,216	0,218
250	0,221	0,277	0,217	0,217
300	0,224	0,282	0,216	0,221
400	0,224	0,279	0,219	0,221
500	0,225	0,281	0,219	0,221
600	0,225	0,283	0,219	0,0221
700	0,229	0,288	0,215	0,228
800	0,233	0,288	0,219	0,234
850	0,233	0,295	0,223	0,232

Source: Author's own calculations.

The Tables 1 and 2 consist of the relative efficiency coefficients of the conditional strategy $(\bar{y}_{HTS}, P(s \mid n,u,N))$. The analysis of these tables let us say that the conditional variant of the sampling design does not significantly lead to improving efficiency of the estimation.

4. Conclusions

The simulation analysis leads to the conclusion that the direct Horvitz-Thompson estimator of the domain mean under the proposed sampling designs dependent on depth functions of auxiliary variables is evidently more accurate than the simple sample mean. The sampling designs dependent on the spherical depth function and the Mahanalobis one approximately lead to the same accuracy of the estimation. The conditional versions of the sampling designs do not significantly improve the accuracy of the estimation. So, in practice, it is sufficient to use the direct Horvitz-Thompson estimator of the domain mean under the sampling designs dependent on spherical depth functions of auxiliary variables.

Acknowledgement

The research was supported by the grant number N N111 434137 from the Polish Ministry of Science and Higher Education.

REFERENCES

HORVITZ D. G., THOMPSON D. J. (1952). A generalization of sampling without replacement from finite universe. *Journal of the American Statistical Association*, vol. 47, 663-685.

LIU R.Y., PARELIUS J.M., SINGH K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, vol. 27, No. 3, 783-858.

MAHALANOBIS P. C. (1936). On the generalized distance in statistics. *Proc. Nat. Acad. Sci. India* vol. 12, 49-55.

Tillé Y. (2006). *Sampling Algorithms*. Springer.

WYWIAŁ J. L. (2007). Simulation analysis of accuracy estimation of population mean on the basis of strategy dependent on sampling design proportionate to the order statistic of an auxiliary variable. *Statistics in Transition-new series* vol. 8, No. 1, 125-137.

WYWIAŁ J. L. (2008). Sampling design proportional to order statistic of auxiliary variable. *Statistical Papers,* vol. 49, No. 2/April, 277-289.

WYWIAŁ J. L. (2009). Sampling design proportional to positive function of order statistics of auxiliary variable. *Studia Ekonomiczne-Zeszyty Naukowe,* 2009, 53, 35-60.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 139—156

REMARKS ABOUT THE GENERALIZATIONS OF THE FISHER INDEX

Jacek Białek[1]

ABSTRACT

In Fisher's index theory the crossing of formulas and weights was regarded as the most pertinent method to derive an "ideal" index formula. The well known Fisher index is a geometric mean of Laspeyres and Paasche indexes and it satisfies most of the postulates coming from the axiomatic index theory. In this paper we consider the generalized Fisher index. The purpose of the study is to propose and discuss some more general class of indexes including the generalized Fisher index.

Key words: Price index, generalized Fisher index, Laspeyres index, Paasche index, Marhall-Edgeworth index

JEL classification: C43

1. Introduction

The history of price indices is quite long – Dutot[2] presented his index in 1738, M. W. Drobisch published his formulas in 1871, Laspeyres and Paasche indices have been known since the 19th century. Depending on the type of an economic problem we may also use one of the following indices: the Fisher ideal index (see Fisher (1922)), the Törnqvist index (Törnqvist (1936)), the Marhall-Edgeworth index and a number of other indices (see Shell (1998), von der Lippe (2007)). Statistical indices are very useful in economy (see Moutlon (1999), Seskin (1998)). Balk (1995) wrote about the axiomatic price index theory, Diewert (1978) showed that the Törnqvist index and the Fisher ideal index approximate each other. But it is really hard to choose the best one from the statistical indices (Dumagan (2002), von der Lippe (2007)). The choice of an index depends on the information we want to obtain. For example, if we are interested in dynamics of money in time we can use Fisher or Marhall-Edgeworth indices.

[1] University of Lodz, Chair of Statistical Methods, Lodz, Poland. E-mail: jbialek@uni.lodz.pl.
[2] „*Reflexions politiques sur les finances et le commerce*", The Hague 1738.

From a theoretical point of view, a proper index should satisfy a group of postulates (tests) coming from the axiomatic index theory (see Balk (1995)). A system of minimum requirements of an index comes from Marco Martini (1992). According to the mentioned system a price index should satisfy at least three conditions: *identity*, *commensurability* and *linear homogeneity*. German index theoreticians – Eichhorn and Voeller (1976) – introduced a more generally acceptable system (EV) of five, and later also of four axioms: *strict monotonicity, price dimensionality*, *commensurability, identity* and (optionally) *linear homogeneity* (see also Białek (2010), von der Lippe (2007)). These five axioms imply other tests like *proportionality* (*identity* plus *linear homogeneity*) or *quantity dimensionality* (*price dimensionality* plus *commensurability*) – see von der Lippe (2007). In the literature we can also meet other systems – for example Bernhard Olt (1996) examined several systems that provide less restrictive requirements than EV-systems.

The main purpose of the study is to present and discuss the proposition of the price index formula being even more general than the generalized Fisher index. We consider the problem of the construction of some general price index that would not only satisfy at least axioms from EV-system, but also include most of useful index formulas (like Laspeyres, Paasche or Fisher). In this paper we give a study of a family of price indices sharing good axiomatic properties (like proportionality, commensurability, monotonicity). An additional purpose of the paper is to extend the considerations to the case of more than two moments of observations.

2. Fisher ideal index and the generalized Fisher index

Let us consider a group of N components observed at times s, t and let us denote[3]:

$P^s = [p_1^s, p_2^s, ..., p_N^s]'$ - a vector of components' prices at time s;

$P^t = [p_1^t, p_2^t, ..., p_N^t]'$ - a vector of components' prices at time t;

$Q^s = [q_1^s, q_2^s, ..., q_N^s]'$ - a vector of components' quantities at time s;

$Q^t = [q_1^t, q_2^t, ..., q_N^t]'$ - a vector of components' quantities at time t.

We consider only price indices but the whole discussion can be generalized to the case of quantity indices. A pair of the most popular and often used price indices are Paasche and Laspeyres indices.

Using the above denotations the Paasche price index can be defined as follows:

[3] The time moment s we consider as the *basis*, i.e. the reference situation, for the comparison.

$$I_{Pa}^{P}(Q^t, P^s, P^t) = \frac{Q^t \circ P^t}{Q^t \circ P^s}, \tag{1}$$

and the Laspeyres price index:

$$I_{La}^{P}(Q^s, P^s, P^t) = \frac{Q^s \circ P^t}{Q^s \circ P^s}, \tag{2}$$

where "$\circ$" denotes an outer product of two vectors.

As it is known, the indices (1) and (2) are very important from a practical point of view and they do not satisfy some of the axioms. For example, *time reversibility* for the price index I^P, described by the formula (3)

$$I^P(Q^s, Q^t, P^s, P^t) = \frac{1}{I^P(Q^t, Q^s, P^t, P^s)}, \tag{3}$$

is not satisfied. But none of the *reversal tests* (time and factor reversal test) or the *circular test* is mentioned in the EV-systems. Taking into consideration the reversal tests or the circular test, we would exclude even useful formulas such as Laspeyres and Paasche indices. Fisher proposed another definition based on Paasche and Laspeyres formulas (see Fisher (1922)). His formula I_F^p is a geometric mean of Paasche and Laspeyres indices:

$$I_F^P(Q^s, Q^t, P^s, P^t) = \sqrt{I_{La}^{P}(Q^s, P^s, P^t) I_{Pa}^{P}(Q^t, P^s, P^t)}. \tag{4}$$

The Fisher's definition is called an "ideal formula", because it satisfies the factor reversal test.

In the literature we can also meet the more general version of formula I_F^p (see von der Lippe (2007)), namely the weighted geometric mean of Laspeyres and Paasche indexes[4]:

$$\tilde{I}_F^P(\alpha) = (I_{La}^P)^{\alpha} (I_{Pa}^P)^{1-\alpha}, \text{ for } \alpha \in [0,1]. \tag{5}$$

Certainly we have: $\tilde{I}_F^P(0) = I_{Pa}^P, \tilde{I}_F^P(0,5) = I_F^P, \tilde{I}_F^P(1) = I_{La}^P$. (6)

[4] Sometimes we use the shorter notation I^P instead of $I^P(Q^s, Q^t, P^s, P^t)$.

In a purely formal manner Fisher's index as unweighted geometric mean of I_{La}^{P} and I_{Pa}^{P} can also be expressed as a weighted arithmetic mean (Köves (1983)) as follows:

$$I_F^P = \frac{I_F^P - I_{Pa}^P}{I_{La}^P - I_{Pa}^P} I_{La}^P + \frac{I_{La}^P - I_F^P}{I_{La}^P - I_{Pa}^P} I_{Pa}^P = I_{A1}^P \left(\frac{I_F^P - I_{Pa}^P}{I_{La}^P - I_{Pa}^P}\right), \tag{7}$$

where

$$I_{A1}^P(\beta) = \beta I_{La}^P + (1-\beta) I_{Pa}^P, \tag{8}$$

is a weighted arithmetic mean of I_{La}^{P} and I_{Pa}^{P}.

For the generalized formula (5) we can also write for any α (the proof is omitted):

$$\widetilde{I}_F^P(\alpha) = I_{A1}^P \left(\frac{(I_{La}^P)^{\alpha}(I_{Pa}^P)^{1-\alpha} - I_{Pa}^P}{I_{La}^P - I_{Pa}^P}\right) = I_{A1}^P \left(\frac{\widetilde{I}_F^P(\alpha) - I_{Pa}^P}{I_{La}^P - I_{Pa}^P}\right). \tag{9}$$

We have the analogous relations for other known types of mean, like the weighted harmonic (I_{A2}^{P}), quadratic (I_{A3}^{P}) and exponential (I_{A4}^{P}) mean, which can be expressed as follows:

$$I_{A2}^P(\beta) = \frac{I_{La}^P \cdot I_{Pa}^P}{\beta I_{La}^P + (1-\beta) I_{Pa}^P}, \tag{10}$$

$$I_{A3}^P(\beta) = \sqrt{\beta (I_{La}^P)^2 + (1-\beta)(I_{Pa}^P)^2}, \tag{11}$$

$$I_{A4}^P(\beta) = \ln[\beta \exp(I_{La}^P) + (1-\beta)\exp(I_{Pa}^P)]. \tag{12}$$

In fact, we can write (the proof is omitted)

$$\widetilde{I}_F^P(\alpha) = I_{A2}^P \left(\frac{I_{La}^P - \widetilde{I}_F^P(1-\alpha)}{I_{La}^P - I_{Pa}^P}\right), \tag{13}$$

$$\widetilde{I}_F^P(\alpha) = I_{A3}^P \left(\frac{(I_{Pa}^P)^2((I_{La}^P)^{2\alpha}(I_{Pa}^P)^{-2\alpha} - 1)}{(I_{La}^P)^2 - (I_{Pa}^P)^2}\right), \tag{14}$$

$$\widetilde{I}_F^P(\alpha) = I_{A4}^P \left(\frac{\exp(\widetilde{I}_F^P(\alpha)) - \exp(I_{Pa}^P)}{\exp(I_{La}^P) - \exp(I_{Pa}^P)}\right). \tag{15}$$

It is interesting that the generalized Fisher index fulfils the time reversability given by (3) only for $\alpha = 0.5$. In fact, we have[5]:

$$\widetilde{I}_F^P(s,t)\cdot\widetilde{I}_F^P(t,s) = [I_{La}^P(s,t)]^{\alpha}\cdot[I_{La}^P(s,t)]^{1-\alpha}\cdot[I_{La}^P(t,s)]^{\alpha}\cdot[I_{Pa}^P(t,s)]^{1-\alpha} =$$

$$= (\frac{Q^s\circ P^t}{Q^s\circ P^s})^{\alpha}(\frac{Q^t\circ P^t}{Q^t\circ P^s})^{1-\alpha}(\frac{Q^t\circ P^s}{Q^t\circ P^t})^{\alpha}(\frac{Q^s\circ P^s}{Q^s\circ P^t})^{1-\alpha} =$$

$$= (\frac{Q^s\circ P^s}{Q^s\circ P^t})^{1-2\alpha}(\frac{Q^t\circ P^t}{Q^t\circ P^s})^{1-2\alpha} = [\frac{(Q^s\circ P^s)(Q^t\circ P^t)}{(Q^s\circ P^t)(Q^t\circ P^s)}]^{1-2\alpha} =$$

$$= [\frac{\sum_{i=1}^{N} q_i^s p_i^s\cdot\sum_{i=1}^{N} q_i^t p_i^t}{\sum_{i=1}^{N} q_i^s p_i^t\cdot\sum_{i=1}^{N} q_i^t p_i^s}]^{1-2\alpha} = 1 \Leftrightarrow \alpha = \frac{1}{2}. \tag{16}$$

Let us also notice that the elasticity of the $\tilde{I}_F^P(\alpha)$ formula (α is a variable) is as follows:

$$E_{\alpha}\tilde{I}_F^P(\alpha) = \frac{d\tilde{I}_F^P(\alpha)}{d\alpha}\cdot\frac{\alpha}{\tilde{I}_F^P(\alpha)} = \alpha\ln\frac{I_{La}^P}{I_{Pa}^P}. \tag{17}$$

Hence, for small differences between Laspeyres and Paasche indexes, that can be written as

$$\frac{I_{La}^P}{I_{Pa}^P} = 1+\Delta, \text{ with } \Delta\approx 0, \tag{18}$$

from the known approximation: $\ln(1+x)\approx x$,which is true for small values of x, we get

$$E_{\alpha}\tilde{I}_F^P(\alpha)\approx\alpha\Delta\le\Delta. \tag{19}$$

Thus, in the case of small differences between I_{La}^P and I_{Pa}^P, the $\tilde{I}_F^P(\alpha)$ formula is almost non-elastic function of α.

[5] In the case of $I_{...}^P(s,t)$ the time moment s we consider as the *basis*, i.e. the reference situation, for the comparison.

3. The second step of generalization

Let $f_j : R_+^N \times R_+^N \to R_+^N$, for $j = 1,2,...,m$ with some fixed m, be such functions that for any $X = [x_1, x_2, ..., x_N]'$ and $Y = [y_1, y_2, ..., y_N]'$, it holds:

$$f_j(\lambda X, \lambda Y) = \lambda f_j(X, Y), \tag{20}$$

where λ is a $N \times N$ diagonal matrix with elements $\lambda_1, \lambda_2, ..., \lambda_N$.

The condition (20) is crucial for commensurability (see Białek (2011)).

Certainly the set of functions satisfying (20) is not empty – for example we could assume $f_j(X,Y) = c_j(X+Y)$ for $j = 1,2,...,m$ and some $c_j \in R_+$. We present the following general formula for price indexes (see Bialek (2011)):

$$\widetilde{I}^P(Q^s, Q^t, P^s, P^t) = \prod_{j=1}^{m} (\widetilde{I}_j^P(Q^s, Q^t, P^s, P^t))^{\gamma_j}, \tag{21}$$

where partial indexes $\widetilde{I}_j^P$ are defined as follows:

$$\widetilde{I}_j^P(Q^s, Q^t, P^s, P^t) = \frac{f_j(Q^s, Q^t) \circ P^t}{f_j(Q^s, Q^t) \circ P^s}, \text{ for } j = 1,2,...,m, \tag{22}$$

and for positive γ_j it holds $\sum_{j=1}^{m} \gamma_j = 1$. (23)

It is easy to prove that the following theorems are true:

Theorem 1

Let $\widetilde{I}^P$ be an aggregative price index with the structure described by (21) with additional condition (20). Then, the $\widetilde{I}^P$ index satisfies tests coming from the EV-system (*strict monotonicity, price dimensionality*, *commensurability, identity* and *linear homogeneity)*.

The proof. Let us notice that for some $k > 0$ we have:

$$\widetilde{I}^{P}(Q^s, Q^t, P^s, kP^t) = \prod_{j=1}^{m} \left(\frac{f_j(Q^s, Q^t) \circ (kP^t)}{f_j(Q^s, Q^t) \circ P^s}\right)^{\gamma_j} = \prod_{j=1}^{m} k^{\lambda_j} \left(\frac{f_j(Q^s, Q^t) \circ P^t}{f_j(Q^s, Q^t) \circ P^s}\right)^{\gamma_j} =$$

$$= k^{\sum_{j=1}^{m} \gamma_j} \prod_{j=1}^{m} \left(\frac{f_j(Q^s, Q^t) \circ P^t}{f_j(Q^s, Q^t) \circ P^s}\right)^{\gamma_j} = k\widetilde{I}^{P}(Q^s, Q^t, P^s, P^t). \tag{24}$$

Thus, *linear homogeneity* is satisfied. Certainly, if $P^s = P^t$ then $\widetilde{I}_j^{P}(Q^s, Q^t, P^s, P^t) = 1$, thus *identity* is satisfied (*identity* plus *linear homogeneity* imply *proportionality*).

Let us also notice that for any diagonal matrix λ *commensurability* from (20) is fulfilled:

$$\begin{aligned} &\widetilde{I}^{P}(\lambda^{-1}Q^s, \lambda^{-1}Q^t, \lambda P^s, \lambda P^t) = \prod_{j=1}^{m} \left(\frac{f_j(\lambda^{-1}Q^s, \lambda^{-1}Q^t) \circ (\lambda P^t)}{f_j(\lambda^{-1}Q^s, \lambda^{-1}Q^t) \circ (\lambda P^s)}\right)^{\gamma_j} = \\ &= \prod_{j=1}^{m} \left(\frac{\lambda^{-1}\lambda \cdot f_j(Q^s, Q^t) \circ P^t}{\lambda^{-1}\lambda \cdot f_j(Q^s, Q^t) \circ P^s}\right)^{\gamma_j} = \widetilde{I}^{P}(Q^s, Q^t, P^s, P^t). \end{aligned} \tag{25}$$

Directly from the definition of the $\widetilde{I}^{P}$ index we have *strict monotonicity and price dimensionality* satisfied. For example, when at least one of the prices $\tilde{P}^t$ (say $\tilde{p}_i^t$) in the current period t is higher compared to the price p_i^t (it means that $\tilde{p}_i^t = p_i^t + p_i^{\Delta}$ and $\tilde{P}^t = P^t + P^{\Delta}$ where $p_i^{\Delta} > 0$) then the index I^P reflects this rise:

$$\widetilde{I}^{P}(Q^s, Q^t, P^s, \tilde{P}^t) = \prod_{j=1}^{m} \left(\frac{f_j(Q^s, Q^t) \circ (\tilde{P}^t)}{f_j(Q^s, Q^t) \circ P^s}\right)^{\gamma_j} = \prod_{j=1}^{m} \left(\frac{f_j(Q^s, Q^t) \circ (P^t + P^{\Delta})}{f_j(Q^s, Q^t) \circ P^s}\right)^{\gamma_j} =$$

$$= \prod_{j=1}^{m} \left(\frac{f_j(Q^s, Q^t) \circ P^t + f_j(Q^s, Q^t) \circ P^{\Delta})}{f_j(Q^s, Q^t) \circ P^s}\right)^{\gamma_j} > \prod_{j=1}^{m} \left(\frac{f_j(Q^s, Q^t) \circ P^t}{f_j(Q^s, Q^t) \circ P^s}\right)^{\gamma_j} = \widetilde{I}^{P}(Q^s, Q^t, P^s, P^t).$$

(26)

Theorem 2

If we additionally assume that for $\gamma_j = \frac{1}{m}$ any for $X, Y, Z \in R_+^N$ and $j = 1,2,...,m$ it holds

$$\sum_{j=1}^{m} \ln \frac{f_j(X,Y) \circ Z}{f_j(Y,Z) \circ Z} = 0 \text{ or equivalently } \prod_{j=1}^{m} \frac{f_j(X,Y) \circ Z}{f_j(Y,X) \circ Z} = 1, \tag{27}$$

then we have *time reversibility* (3) satisfied.
The proof is immediate – under (27) we get: (28)

$$\widetilde{I}^P(Q^t,Q^s,P^t,P^s) \cdot \widetilde{I}^P(Q^s,Q^t,P^s,P^t) = [\prod_{j=1}^{m} \frac{f_j(Q^t,Q^s) \circ P^s}{f_j(Q^t,Q^s) \circ P^t}]^{\frac{1}{m}} [\prod_{j=1}^{m} \frac{f_j(Q^s,Q^t) \circ P^t}{f_j(Q^s,Q^t) \circ P^s}]^{\frac{1}{m}} =$$

$$= [\prod_{j=1}^{m} \frac{f_j(Q^t,Q^s) \circ P^s}{f_j(Q^s,Q^t) \circ P^s} \cdot \frac{f_j(Q^s,Q^t) \circ P^t}{f_j(Q^t,Q^s) \circ P^t}]^{\frac{1}{m}} = [\prod_{j=1}^{m} \frac{f_j(Q^t,Q^s) \circ P^s}{f_j(Q^s,Q^t) \circ P^s}]^{\frac{1}{m}} [\prod_{j=1}^{m} \frac{f_j(Q^s,Q^t) \circ P^t}{f_j(Q^t,Q^s) \circ P^t}]^{\frac{1}{m}} = 1^{\frac{2}{m}} = 1$$

Let us notice that although the condition (27) does not seem to be strong, the *time reversibility* seems to be too restrictive test. Many authors claim that *time reversibility* is even unnecessary because it rules out a lot of reasonable and useful index functions like Laspeyres or Paasche (see also Discussion and Conclusion). Now, we present some known index formulas belonging to the class $\widetilde{I}^P$.

Remark 1

Let us consider the case of $m = 2$ and define

$$\gamma_1 = \gamma_2 = \frac{1}{2}, \tag{29}$$

$$f_1(X,Y) = X, \tag{30}$$

$$f_2(X,Y) = Y, \tag{31}$$

where $X = (x_1, x_2, .., x_N)$, $Y = (y_1, y_2, .., y_N)$.

Hence, we have for any $Z \in R_+^N$ it holds

$$\prod_{j=1}^{2}[\frac{f_j(X,Y)\circ Z}{f_j(Y,X)\circ Z}]^{\gamma_j} = (\frac{f_1(X,Y)\circ Z}{f_1(Y,X)\circ Z})^{\frac{1}{2}} \cdot (\frac{f_2(X,Y)\circ Z}{f_2(Y,X)\circ Z})^{\frac{1}{2}} = \sqrt{\frac{X\circ Z}{Y\circ Z}\cdot\frac{Y\circ Z}{X\circ Z}} = 1 \tag{32}$$

Thus, the functions described in (30) and (31) satisfy the assumption (20) and (27). Let us also notice that in the considered case we get:

$$\tilde{I}^P(Q^s,Q^t,P^s,P^t) = \prod_{j=1}^{2}[\frac{f_j(Q^s,Q^t)\circ P^t}{f_j(Q^s,Q^t)\circ P^s}]^{\frac{1}{2}} = \sqrt{\frac{Q^s\circ P^t}{Q^s\circ P^s}\cdot\frac{Q^t\circ P^t}{Q^t\circ P^s}} =$$

$$= \sqrt{I_{La}^P(Q^s,P^s,P^t)I_{Pa}^P(Q^t,P^s,P^t)} = I_F^P(Q^s,Q^t,P^s,P^t). \tag{33}$$

The formula (33) leads to the following conclusion: the Fisher index is the special case of the general formula defined in (21). Moreover, the generalized Fisher index, presented by (5) has a structure described by (21), where $\gamma_1 = \alpha$, $\gamma_2 = 1-\alpha$, $\tilde{I}_1^P = I_{La}^P$, $\tilde{I}_2^P = I_{Pa}^P$.

Certainly, not only Fisher index is a particular case of the general formula $\tilde{I}^P$. For example, for $m=1$, and $f_1(X,Y) = \frac{1}{2}(X+Y)$ we get

$$\tilde{I}^P(Q^s,Q^t,P^s,P^t) = \frac{f_1(Q^s,Q^t)\circ P^t}{f_1(Q^s,Q^t)\circ P^s} = \frac{\frac{1}{2}(Q^s+Q^t)\circ P^t}{\frac{1}{2}(Q^s+Q^t)\circ P^s} = \frac{\sum_{i=1}^{N}\frac{q_i^s+q_i^t}{2}p_i^t}{\sum_{i=1}^{N}\frac{q_i^s+q_i^t}{2}p_i^s} \tag{34}$$

The formula presented in (34) is known in the literature as the Marhall-Edgeworth index I_{ME}^P (von der Lippe (2007)). Moreover, taking $m=1$ and functions from (30) – (31) we get Laspeyres and Paasche formulas. It can be easily proved that the Walsh and Geary-Khamis indexes can be also obtained as special cases of $\tilde{I}^P$ formula.

Theorem 3

Let us signify by $\Lambda(m)$ the class of price indices defined by (20) and (21) for some fixed $m \in N$. Let us consider a group of indices $\tilde{I}_1^P$, $\tilde{I}_2^P$, ... , $\tilde{I}_M^P$ and let us assume that

$$\forall i \in \{1,2,...,M\} \quad \tilde{I}_i^P \in \Lambda(m_i), \quad \text{for some } m_i \in N. \tag{35}$$

Then, we have

$$\tilde{I}_G^P = \sqrt[M]{\prod_{i=1}^{M} \tilde{I}_i^P} \in \Lambda(m), \quad \text{where } m = M\prod_{i=1}^{M} m_i. \tag{36}$$

Proof (see Białek (2011))

Firstly, let us notice that from the assumption (35) we get for any $i \in \{1,2,...,M\}$

$$\tilde{I}_i^P = \tilde{I}_i^P(Q^s, Q^t, P^s, P^t) = [\prod_{j=1}^{m_i} \frac{f_j^i(Q^s, Q^t) \circ P^t}{f_j^i(Q^s, Q^t) \circ P^s}]^{\frac{1}{m_i}}, \tag{37}$$

and

$$f_j^i(\lambda Q^s, \lambda Q^t) = \lambda f_j^i(Q^s, Q^t), \tag{38}$$

for any positive $Q^s, Q^t, P^s, P^t \in R_+^N$.

Let us define for the given $Q^s, Q^t, P^s, P^t \in R_+^N$

$$\theta(f) = \frac{f(Q^s, Q^t) \circ P^t}{f(Q^s, Q^t) \circ P^s}. \tag{39}$$

Hence, we can write

$$\tilde{I}_G^P = \sqrt[M]{\prod_{i=1}^{M} \tilde{I}_i^P} = [\prod_{i=1}^{M} (\prod_{j=1}^{m_i} \frac{f_j^i(Q^s, Q^t) \circ P^t}{f_j^i(Q^s, Q^t) \circ P^s})^{\frac{1}{m_i}}]^{\frac{1}{M}} = [\prod_{i=1}^{M} (\prod_{j=1}^{m_i} \theta(f_j^i))^{\frac{1}{m_i}}]^{\frac{1}{M}}. \tag{40}$$

Let us signify by $k = \prod_{s=1}^{M} m_s$ and $k_i = \frac{k}{m_i}$. We get from (40)

$$\tilde{I}_G^P = [(\prod_{j=1}^{m_1} \theta(f_j^1))^{\frac{1}{m_1}} \cdot (\prod_{j=1}^{m_2} \theta(f_j^2))^{\frac{1}{m_2}} \cdot \ldots \cdot (\prod_{j=1}^{m_M} \theta(f_j^M))^{\frac{1}{m_M}}]^{\frac{1}{M}} =$$

$$= [(\prod_{j=1}^{k_1 m_1} \theta(\tilde{f}_j^1))^{\frac{1}{k_1 m_1}} \cdot (\prod_{j=1}^{k_2 m_2} \theta(\tilde{f}_j^2))^{\frac{1}{k_2 m_2}} \cdot \ldots \cdot (\prod_{j=1}^{k_M m_M} \theta(\tilde{f}_j^M))^{\frac{1}{k_M m_M}}]^{\frac{1}{M}}, \quad (41)$$

where

$$\tilde{f}_j^i = f^i_{j-\left\lfloor \frac{j-1}{m_i} \right\rfloor m_i}, \text{ for } j = 1,2,....,k_i m_i. \quad (42)$$

Knowing that $k = k_i m_i$, we get from (41)

$$\tilde{I}_G^P = [(\prod_{j=1}^{k} \theta(\tilde{f}_j^1)) \cdot (\prod_{j=1}^{k} \theta(\tilde{f}_j^2)) \cdot \ldots \cdot (\prod_{j=1}^{k} \theta(\tilde{f}_j^M))]^{\frac{1}{kM}} = [\prod_{p=1}^{kM} \theta(\hat{f}_p)]^{\frac{1}{kM}}, \quad (43)$$

where

$$\hat{f}_p = \tilde{f}^{\,1+\left\lfloor \frac{p-1}{k} \right\rfloor}_{p-\left\lfloor \frac{p-1}{k} \right\rfloor k}, \text{ for } p = 1,2,...,kM. \quad (44)$$

It is easy to check that in the case of the formula (43) the assumption (20) is satisfied.

From (43) we have the final conclusion that

$$\tilde{I}_G^P = \sqrt[kM]{\prod_{i=1}^{kM} \hat{I}_i^P} \in \Lambda(kM), \quad (45)$$

where $\hat{I}_i^P = \frac{\hat{f}_i(Q^s, Q^t) \circ P^t}{\hat{f}_i(Q^s, Q^t) \circ P^s}$ and $k = \prod_{s=1}^{M} m_s$.

Remark 3

By using Theorem 3 we can show that $\sqrt{I_{ME}^{P} I_{F}^{P}} \in \Lambda(4)$. In fact, we have $I_{ME}^{P} \in \Lambda(1)$ and $I_{F}^{P} \in \Lambda(2)$. Taking $m_1 = 1$, $m_2 = 2$ we get from the above theorem that $\sqrt{I_{ME}^{P} I_{F}^{P}} \in \Lambda(2m_1 m_2) = \Lambda(4)$. Moreover, Theorem 3 seems to be an argument for using the geometric mean of $\tilde{I}^{P}$ indices (see formula (46) and (52)).

4. The third step of generalization

Let us consider the situation when we intend to measure the average price dynamics on the whole time interval $[T_1, T_2]$. We observe prices and quantities of N components at moments $T_1, T_1 + 1, ..., T_2$, and we collect the data: $P = [p_i^j]_{i=1,2,...,N}^{j=T_1,...,T_2}$, $Q = [q_i^j]_{i=1,2,...,N}^{j=T_1,...,T_2}$. We propose the following general price index that can be used to measure the average (one-period) price dynamics on the time interval $[T_1, T_2]$:

$$\hat{I}^{P}(T_1, T_2) = \prod_{t=T_1}^{T_2-1} [\tilde{I}^{P}(Q^t, Q^{t+1}, P^t, P^{t+1})]^{\delta_t(P,Q)}, \tag{46}$$

where

$\tilde{I}^{P}$ - is the formula defined in (21),

$Q^t = [q_1^t, q_2^t, ..., q_N^t]'$ - a vector of components' quantities at time t,

$P^t = [p_1^t, p_2^t, ..., p_N^t]'$ - a vector of components' prices at time t,

$\delta_t(P,Q)$ - are positive numbers, where $\sum_{t=T_1}^{T_2-1} \delta_t(P,Q) = 1$.

For example, it would be quite natural (analogically to the Törnqvist index) to assume

$$\delta_t(P,Q) = \frac{1}{2}\left(\frac{v(P^t, Q^t)}{V_1(P,Q)} + \frac{v(P^{t+1}, Q^{t+1})}{V_2(P,Q)}\right), \tag{47}$$

where

$$v(P^s, Q^s) = \sum_{i=1}^{N} p_i^s q_i^s, \quad \text{for } s = T_1, T_1 + 1, ..., T_2, \tag{48}$$

$$V_1(P,Q) = \sum_{t=T_1}^{T_2-1} v(P^t, Q^t), \tag{49}$$

$$V_2(P,Q) = \sum_{t=T_1}^{T_2-1} v(P^{t+1}, Q^{t+1}) . \tag{50}$$

From (21) and (35) we have:

$$\hat{I}^P(T_1,T_2) = \prod_{t=T_1}^{T_2-1} \{ \prod_{j=1}^{m} [\frac{f_j(Q^t, Q^{t+1}) \circ P^{t+1}}{f_j(Q^t, Q^{t+1}) \circ P^t}]^{\gamma_j} \}^{\delta_t(P,Q)} , \tag{51}$$

where functions f_j are described in (20).

Certainly, considering only two moments of observations $T_1 = s$ and $T_2 = t$ we get $\hat{I}^P(T_1,T_2) = \widetilde{I}^P$. Moreover, if we assumed *circurality*[6] for $\widetilde{I}^P$ and $\delta_t(P,Q) = const$, we would get:

$$\hat{I}^P(T_1,T_2) = \sqrt[T_2-T_1]{\widetilde{I}^P(Q^{T_1}, Q^{T_2}, P^{T_1}, P^{T_2})} . \tag{52}$$

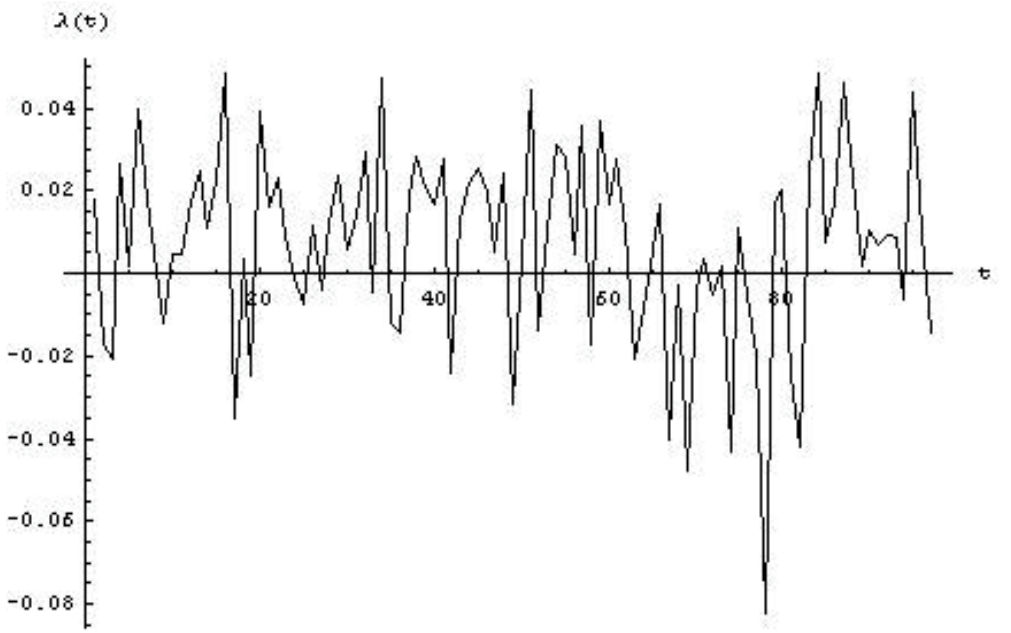
5. Worked Example

There are situations in practice when we could use the formula (46). For example, let us assume that we are interested in measuring the average, one-month dynamics of units' prices of Open Pension Funds. If we had monthly data (about values of units' prices $P^t = [p_1^t, p_2^t, ..., p_N^t]'$ and numbers of units $Q^t = [q_1^t, q_2^t, ..., q_N^t]'$) for the given time interval, we could use the formula (46) and evaluate "the average, monthly financial condition" of the group of funds. This kind of information could be also useful for a comparison with a pension fund (see fig. 1).

We consider the period of January 30, 2003 – December 30, 2010 (96 monthly observations) for $N = 14$ Polish pension funds. The financial condition of the whole group of funds can be measured by using the following function of time: $\lambda(t) = \hat{I}^P(t, 12+t) - 1$, with $\delta_t(P,Q)$ described in (47) - see fig. 1. We can see on the graph that the Polish funds are effective in the majority of months (values over zero) except for the world financial crisis ($t \in \{61,...,82\}$).

[6] The circular test is one of the most restrictive postulates in the price index theory (see von der Lippe (2007)).

Figure 1. Function $\lambda(t)$ for Polish Open Pension Funds and time period January 30, 2003 – December 30, 2010



Source: own calculations in Mathematica 6.0.

6. Discussion and Conclusion

There is no doubt that price index numbers play an important role in economic and business decision-making (see Fisher (1972)). Thus, it is very important to use well constructed formulas of indexes. The literature on axiomatic index theory is very wide (see Balk (1995), von der Lippe (2007)). Fisher's index seems to be "ideal" because it fulfils most of mentioned tests including *time reversibility*. It is possible to generalize this formula, but the *time reversibility* test is satisfied only in some specific cases. The class of indexes, considered in the paper, is even more general than the generalized Fisher index.

The Theorem 1 shows that the presented formula satisfies all the postulates coming from the EV-system (*strict monotonicity, price dimensionality, commensurability, identity and linear homogeneity)*. Thus, we have a practical conclusion: it is easier to prove that a given index belongs to the considered class than verify that the axioms in question are satisfied. We show that some of known price indices (like Laspeyres, Paasche, Fisher or Marhall-Edgeworth formulas) are particular elements of the considered class (see Remark 1). Let us

also notice that the conclusion from the theorem 2 seems to be an analogous to the conclusion from (16).

The Theorem 2 explains why the generalized Fisher index satisfies *time reversibility* only for $\alpha = \frac{1}{2}$. However, *time reversibility* seems to be too restrictive and relatively unimportant. Firstly, it rules out many reasonable and useful index functions like Laspeyres or Paasche. Secondly, the history takes one direction only. And finally, from an economic point of view, there is no need for "symmetry" described by (27) and *time reversibility.* There are many dissimilarities between intertemporal and interregional price comparisons. With regard to two countries, *A* and *B,* there is no reason to prefer one of them to the other, so it is reasonable to treat them symmetrically in the sense of country reversibility. But with regard to points in time, it should be recognized that s and t are not just two points in time but s is rather a single, fixed period and t is some variable (a multitude of points in time - see von der Lippe (2007)).

The main result – Theorem 3 – shows that the geometric mean of the indexes from the considered, general class also belongs to this class. Thus, using the geometric mean of the mentioned, discussed and known indices (like Paasche, Laspeyres, Fisher, Marhall-Edgewort, Walsh, Geary-Khamis, etc.) we still have their properties from EV-system. This theorem can be useful for constructing new index formulas with (that should be verified) better statistical properties, like dispersion.

Finally, we present the general formula based on data from the whole time interval which can be used (for example) in researching pension or investment funds. This is the third and the last step of the generalization when the index formula takes into account each time moments from T_1 to T_2 and that resembles the idea of chain indices (see von der Lippe (2007)).

7. Appendix

Below we present the formal definitions of major postulates (tests) coming from the axiomatic index theory and used in the theorem 1. Let us consider the price index formula $I^P(Q^s, Q^t, P^s, P^t)$. Let us also signify by λ any $N \times N$ diagonal matrix with elements $\lambda_1, \lambda_2, ..., \lambda_N$ and by k some positive, real number.

- *Identity* means that

$$I^P(Q^s, Q^t, P^s, P^s) = 1. \tag{53}$$

- *Proportionality* can be described by the following condition:

$$I^P(Q^s, Q^t, P^s, kP^s) = k. \tag{54}$$

- *Commensurability* can be expressed as follows:

$$I^P(\lambda^{-1}Q^s, \lambda^{-1}Q^t, \lambda P^s, \lambda P^t) = I^P(Q^s, Q^t, P^s, P^t). \quad (55)$$

- *Linear homogeneity* has the following form:

$$I^P(Q^s, Q^t, P^s, kP^t) = kI^P(Q^s, Q^t, P^s, P^t). \quad (56)$$

- *Price dimensionality* can be expressed as follows:

$$I^P(Q^s, Q^t, kP^s, kP^t) = I^P(Q^s, Q^t, P^s, P^t). \quad (57)$$

- *Strict monotonicity* is defined as follows:

$$I^P(Q^s, Q^t, P^s, \widetilde{P}^t) > I^P(Q^s, Q^t, P^s, P^t), \text{ if } \widetilde{P}^t \geq P^t \quad (58)$$

and

$$I^P(Q^s, Q^t, P^s, \widetilde{P}^t) < I^P(Q^s, Q^t, P^s, P^t), \text{ if } \widetilde{P}^t \leq P^t, \quad (59)$$

where $\widetilde{P}^t \geq P^t$ means that at least one element of the nonnegative vector $\widetilde{P}^t$ is greater than the corresponding element of the vector P^t (the relation $\widetilde{P}^t \leq P^t$ is defined analogously).

REFERENCES

BALK M. (1995). *Axiomatic Price Index Theory: A Survey*, International Statistical Review 63, 69-95.

BIAŁEK J. (2010). The generalized formula for aggregative price indexes, Statistics in Transiotion – new series, vol. 11, 145-154, GUS, Warszawa.

BIAŁEK J. (2011). Proposition of the general formula for price indices, Communications in Statistics: Theory and Methods (in press).

DIEWERT W. (1978). Superlative Index Numbers and Consistency in Aggregation, „Econometrica" 46, 883-900.

DOMAŃSKI CZ. (2001). Metody statystyczne. Teoria i zadania, Wydawnictwo Uniwersytetu Łódzkiego, Lodz, Poland.

DUMAGAN J. (2002). Comparing the superlative Törnqvist and Fisher ideal indexes, Economic Letters 76, 251-258.

EICHHORN W., VOELLER J. (1976). Theory of the Price Index. Fisher's Test Approach and Generalizations, Berlin, Heidelberg, New York: Springer-Verlag.

FISHER I. (1922). The Making of Index Numbers, Boston: Houghton Mifflin.

FISHER F.M. (1972). The Economic Theory of Price Indices, Academic Press, New York.

MARTINI M. (1992). A General Function of Axiomatic Index Numbers, Journal of the Italian Statistics Society, 1 (3), 359-376.

KÖVES P. (1983). Index Theory and Economic Reality, Budapest: Akad. Kiad.

OLT B. (1996). Axiom und Struktur in der statistischen Preisindextheorie, Frankfurt, Peter Lang.

MOUTLON B., SESKIN E. (1999). A preview of the 1999 comprehensive revision of the national income an product accounts , Survey of Current Business No. 79.

SHELL K. (1998). Economics Analysis of Production Price Indexes, Cambridge University Press, UK.

TÖRNQVIST L. (1936). The Bank of Finland's consumption price index, Bank of Finland Monthly Bulletin 10, 1-8.

VON DER LIPPE P. (2007). Index Theory and Price Statistics, Peter Lang, Frankfurt, Germany.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 157—178

MULTIPLE – EQUATION MODELS OF ORDERED DEPENDENT VARIABLES IN EXPLORATION OF THE RESULTS OF REHABILITATION OF LOCOMOTIVE ORGAN DISORDERS

Jerzy Grosman[1] , Mieczysław Kowerski[2]

ABSTRACT

In the present paper concerning the analysis of the factors determining patient's self-service during the admission and release from hospital a two-equation model of ordered dependent variables is proposed. These types of models are especially useful when the results of rehabilitation of locomotive organ disorders are not described by means of exact values obtained by mechanical measurements but they are described by means of qualitative valuation (ranking) made by a therapist when the distances between neighbouring ranks are not known. The advantages of the proposed model were presented on the basis of the results of estimation based on data of 4063 patients of hospitals from Mazowieckie and Warmińsko-Mazurskie provinces.

Key words: Rehabilitation of locomotive organ disorders, Weiss test, multiple-equation model of ordered dependent variable, maximum likelihood estimation, McFadden determination coefficient - pseudo R^2, count R-squared, probability of the norm in the self-service test.

1. Preface

The results of rehabilitation of locomotive organ disorders are often described not by exact values of measurements with specific devices but by values, ranked from the lowest to the highest, for which the distances between consecutive ranks are not known (e.g. in the Weiss test the patients are classified using the following ranks: 1 – patient without pain; 2 – periodic pain, low severity, not reported by the patient; 3 – postexercise pain after a long walk, etc.). In this situation classic regression method should not be used to model the relationships between the

[1] Akademia Wychowania Fizycznego w Warszawie.
[2] Wyższa Szkoła Zarządzania i Administracji w Zamościu, ul. Akademicka 4, 22-400 Zamość, tel. 84 6776711, mowerski@wszia.edu.pl.

results of rehabilitation and their determinants. Models of ordered dependent variable seem to be much more relevant.

In the present paper the models of ordered dependent variable have been used to describe the self-service status of a patient. Since the self-service status is measured both at the hospital admission and at the hospital release of a patient, a two-equation recursive model is proposed, in which the first equation describes the patient status at the admission and the second equation – at the release from the hospital.

The source of information for all values of dependent and explanatory variables in this paper are the Weiss test results for 4063 patients.

2. Weiss test

The Weiss test is used to assess the locomotive ability and the social and occupational situation of a patient [Weiss, Zembaty 1983, Grossman 1987, Górski et al. 1985, Grossman 1988]. It is used not only for patients with stroke but also for geriatric patients.

As a result of numerous clinical trials nine test were selected as representative for the whole research:

- pain test (variable P_1 at the admission and variable K_1 at the release),
- motion test (variable P_2 at the admission and variable K_2 at the release),
- muscular strength test (variable P_3 at the admission and variable K_3 at the release),
- fitness test (variable P_4 at the admission and variable K_4 at the release),
- locomotion test (variable P_5 at the admission and variable K_5 at the release),
- wheelchair locomotion test (variable P_6 at the admission and variable K_6 at the release),
- self-service test (variable P_7 at the admission and variable K_7 at the release),
- hand-grip test (variable P_8 at the admission and variable K_8 at the release),
- occupational adaptation test (variable P_9 at the admission and variable K_9 at the release)

Each of these test has a six-rank assessment scale, where 1 is the norm. Higher values mean an unfavourable (negative) intensification of the variable.

3. Data

The data used in the present paper has been collected from hospitals and health centres in Warsaw and the provinces of Mazowieckie and Warminsko-Mazurskie during last 20 years. The information about patients (age, sex, length of hospitalization in months) and the Weiss test results at the admission and at the release was recorded for 4063 persons[1].

Table 1. Basic statistics of the population

Variable	N	Mean	Std. Dev.	Median	Mode	Min	Max.	Q1	Q3	Skewness	Kurtosis
Sex	4063	1.44	0.5914	1	1	1	22	1	2	10.51	357.95
Age	4063	43.97	17.9941	45	45	10	92	27	59	0.06	-1.09
Length of hospitalization	4063	47.41	39.0778	39	35	1	589	22	60	2.86	17.90
P_1	4063	2.65	1.4984	3	1	1	6	1	4	0.49	-0.77
P_2	4063	3.37	1.5667	3	2	1	6	2	5	0.13	-1.06
P_3	4063	3.08	1.4697	3	3	1	6	2	4	0.40	-0.61
P_4	4063	2.25	1.3641	2	1	1	6	1	3	1.10	0.52
P_5	4063	3.57	1.8683	3	6	1	6	2	6	0.05	-1.43
P_6	4063	2.57	1.7985	2	1	1	6	1	4	0.77	-0.86
P_7	4063	3.15	1.9434	3	1	1	6	1	5	0.27	-1.48
P_8	4063	1.59	1.2580	1	1	1	6	1	1	2.27	4.24
P_9	4063	4.24	2.1084	6	6	1	6	2	6	-0.60	-1.39
K_1	4063	1.55	0.8174	1	1	1	6	1	2	1.56	2.41
K_2	4063	2.59	1.4309	2	2	1	6	1	4	0.75	-0.31
K_3	4063	2.42	1.3541	2	1	1	6	1	3	0.90	0.19
K_4	4063	1.67	0.9481	1	1	1	6	1	2	1.74	3.47
K_5	4063	2.58	1.5056	2	1	1	6	1	3	0.87	-0.09
K_6	4063	1.65	1.1276	1	1	1	6	1	2	1.87	2.95
K_7	4063	2.19	1.4574	2	1	1	6	1	3	1.06	0.06
K_8	4063	1.45	1.0559	1	1	1	6	1	1	2.67	6.71
K_9	4063	3.42	2.0999	3	1	1	6	1	6	0.08	-1.69

Note: Variables: P_1, P_2,...,P_9 represent the Weiss test results at the admission to the hospital while the K_1, K_2,...,K_9 variables represent the Weiss test results at the release from the hospital.

Source: calculated with Statistica 6.0

[1] Patients who died during the rehabilitation process were excluded from the panel.

It was assumed in this paper that the dependent variables are: the self-service test results at the admission (P_7) and at the release (K_7) from the hospital.

Table 2. Correlation coefficients between the dependent variables and potential explanatory variables

Variable	P_7	K_7
Sex	-0.0633*	-0.0441*
Age	0.3573*	0.2046*
Length of hospitalization	x	0.1359*
P_1	0.1926*	0.0992*
P_2	0.5070*	0.4633*
P_3	0.4572*	0.4709*
P_4	0.5231*	0.4683*
P_5	0.7073*	0.5173*
P_6	0.6732*	0.4783*
P_7	1.0000	0.6509*
P_8	0.3931*	0.3717*
P_9	0.6448*	0.4578*
K_1	x	0.0815*
K_2	x	0.4937*
K_3	x	0.5221*
K_4	x	0.5140*
K_5	x	0.6739*
K_6	x	0.6025*
K_7	x	1.0000
K_8	x	0.3892*
K_9	x	0.5956*

Note: (*) means that the correlation coefficient is significant at the level of at least .05.
Source: calculated with Statistica 6.0.

The values of Pearson's linear correlation coefficients[1] for P_7 show that there is a statistically significant relationship between the dependent variable and all potential explanatory variables. The older the patient is, the poorer (higher value

[1] The qualitative (rank) nature of the test results makes the Pearson's linear correlation coefficient not the best tool to assess dependencies between the variables. However, it was chosen to present the results in the preliminary analysis because it is very popular. The more so because many research showed that the assessment of relationships based on the Pearson's correlation coefficients gives approximate results to assessments based on qualitative measures of associations.

of test) self-service results at the admission are achieved. At the same time, the poorer the results of other tests are, the poorer self-service results at the admission are recorded. The K_7 variable also shows a statistically significant positive relationship with age and the test results both at the admission and at the release. To summarise: the older the person is and the higher test results at the admission are, the higher self-service test score is at the release of a patient. The statistically significant positive value of the correlation coefficient between the self-service test at the release and the length of stay in the hospital may be surprising. This can be explained as follows: the patients who stayed longer in the hospital were released in a worse self-service condition. However, this conclusion is based just on single coefficient of correlation, which does not reflect the relation between the length of hospitalization and the patient's age or self-service status at the admission. This issue will be subject to further analysis.

The relationships between qualitative discreet variables (ranks) should be analysed with the Pearson χ^2 test [Stanisz 2001, s. 286-287].

Table 3. Results of the Pearson χ^2 test for the self-service test and the other tests

Variable	P_7		K_7	
	χ^2	p	χ^2	p
P_1	327.616	<0.000001	142.840	<0.000001
P_2	1283.420	<0.000001	1190.780	<0.000001
P_3	1117.670	<0.000001	1142.210	<0.000001
P_4	1456.920	<0.000001	1244.480	<0.000001
P_5	2782.570	<0.000001	1423.680	<0.000001
P_6	2462.490	<0.000001	1146.290	<0.000001
P_7	x	x	3132.180	<0.000001
P_8	930.575	<0.000001	897.134	<0.000001
P_9	2379.180	<0.000001	1227.830	<0.000001
K_1	x	x	106.511	<0.000001
K_2	x	x	1275.980	<0.000001
K_3	x	x	1358.270	<0.000001
K_4	x	x	1689.340	<0.000001
K_5	x	x	2638.060	<0.000001
K_6	x	x	2304.370	<0.000001
K_8	x	x	1009.920	<0.000001
K_9	x	x	2094.300	<0.000001

Source: calculated with Statistica 6.0.

The results of the χ^2 – Pearson test for a relationship of a self-service test at the admission and the other Weiss tests show that these relationships are very close and statistically significant at .000001. Similar situation is between the self-service tests at the admission and the other Weiss tests – these relationships are statistically significant at least at .000001. The self-service tests at the release can also be an example of such dependence.

Table 4. Self-service tests at the admission vs. at the release from a hospital

Value	At the admission		At the release		Change of (%)
	No. of patients	(%)	No. of patients	(%)	
1	1308	32.2	1935	47.6	15.4
2	551	13.6	748	18.4	4.8
3	537	13.2	602	14.8	1.6
4	341	8.4	363	8.9	0.5
5	524	12.9	264	6.5	-6.4
6	802	19.7	151	3.7	-16.0
Total	4063	100.0	4063	100.0	0.0

Source: authors' calculation.

The treatment applied caused an improvement of patient self-service. At the admission to a hospital only one third of patients showed a normal status of self-service, whereas at the release this number grew to almost a half. On the other hand, the number of patients with rank 6 (who require a permanent care) dropped from 19.7% at the admission to 3.7% at the release (16 points).

4. Two-equation model of patient's self-service

Taking into account the presented relationships between dependent and potential explanatory variables it is proposed to use a two-equation recursive model where the self-service statuses at the admission and at the release play a role of dependent variables.

These variables may adopt one out of six possible states (values) of patient's self-service:

1. Full ability to perform all daily indoor and outdoor activities
2. Ability to perform basic daily indoor activities without special equipment
3. Partial ability to perform daily activities, which does not require assistance from other persons
4. Independence in daily activities, but special equipment is needed
5. Partial independence in daily activities, which requires other person's assistance and special equipment in some activities

6. Fully dependent, permanent care by other person required

The recursive model is composed of two equations.

First equation describes patient's self-service status at the admission to a hospital

$$P_7 = f\left(\text{Sex}, \text{Age}, P_1, P_2, P_3, P_4, P_5, P_6, P_8, P_9, \varepsilon_1\right) \tag{1}$$

The second equation describes patient's self-service at the release (K_7) and considers the length of stay in the hospital and the self-service status at the admission, which is represented – typically for recursive models – by the theoretical values of the first dependent variable, estimated from the first equation [Syczewska 2003, pp. 209-210] $\hat{P}_7$. In this way the estimated value of P_7 is representative for the test results at the admission and the patient's age.

$$K_7 = f\left(\text{Sex, Age, Length of hosp}, \hat{P}_7, K_1, K_2, K_3, K_4, K_5, K_6, K_8, K_9, \varepsilon_2\right) \tag{2}$$

Each dependent value adopts six values ranked from the lowest to the highest and the distances between consecutive ranks are unknown. In this situation the most appropriate approach to explain the variation of the dependent variable (reasons for different values of the self-service test) is the use of ordered dependent variable models [Greene 2003, pp. 736-740]. In case of both tests (descriptions of daily activities) the assumption of equal distances between neighbouring categories, e.g. between rank 2 and rank 3 and between rank 3 and rank 4, can hardly be maintained. It is assumed, however, that there is a model of self-service – latent variable[1] y^*:

$$y_i^* = \mathbf{X}_i^T \beta + \varepsilon_i \tag{3}$$

where:

y_i^* – value of latent dependent variable for an *i*-th patient,

$\mathbf{X}_i^T$ – a vector of values of independent variables for an *i*-th patient,

β – a vector of structural parameters of the model,

ε_i – an error term,

on the basis of which the assessment of patient's self-service is made both at the admission and at the release.

The value of the ordered dependent value y_i depends then on the value of the latent variable y_i^*, according to the rule[2]

---

[1] *Y* means any dependent variable, P_7 and K_7 in our case. The **X** vector is a vector of independent variables, which in our case includes the results of other tests and other characteristic of the patients (sex, age, length of stay in a hospital).

[2] EViews 5.1 User's Guide Quantitative Micro Software, LLC, Irvine CA, s. 639.

$$y_i = \begin{cases} 1 \text{ dla } y_i^* \le \gamma_1 \\ 2 \text{ dla } \gamma_1 < y_i^* \le \gamma_2 \\ 3 \text{ dla } \gamma_2 < y_i^* \le \gamma_3 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ M \text{ dla } \gamma_M < y_i^* \end{cases} \tag{4}$$

It does not matter that the ranks were chosen arbitrarily and may adopt different values in other models. The only requirement for ordered variables is to meet the following rule:

$$\text{If } y_i^* < y_j^* \text{ then } y_i < y_j \tag{5}$$

A big advantage of this model, especially in medical research, is that the results may be interpreted as probabilities. The probability of each rank of the Y variable is as follows:

$$\begin{gathered} \Pr(y_i = 0 | x_i, \beta, \gamma) = F(\gamma_1 - x_i^T \beta) \\ \Pr(y_i = 1 | x_i, \beta, \gamma) = F(\gamma_2 - x_i^T \beta) - F(\gamma_1 - x_i^T \beta) \\ \Pr(y_i = 2 | x_i, \beta, \gamma) = F(\gamma_3 - x_i^T \beta) - F(\gamma_2 - x_i^T \beta) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \Pr(y_i = M | x_i, \beta, \gamma) = 1 - F(\gamma_M - x_i^T \beta) \end{gathered} \tag{6}$$

where F is a distribution function of the random variable ε.

The γ values are the boundary points and are estimated together with the β parameters by means of the maximum likelihood method through maximizing the log likelihood.

The maximization of the likelihood function may be done using a few methods. The EViews software offers Newton-Raphson and Goldfeld-Quandt (quadratic hill climbing) methods. In most cases it is assumed that the models of ordered dependent variable are logit or probit models.

If we assume that the ordered dependent variable adopts a probit model, then the F function is a distribution function of a standardized normal distribution, which value may be read directly from the standardized normal distribution tables. If we assume that the ordered dependent variable adopts a logit model, then the F function is a logistic distribution function, which values may be calculated using the following formula:

$$F(y) = \frac{e^y}{1 + e^y} \tag{7}$$

While choosing the most appropriate model the criterion of maximisation of McFadden determination coefficient was applied [McFadden 2002, p. 65],

together with the rule that the model should consist of only such variables, which parameters are significant and coincident [Hellwig 1976].

5. Results of the estimation of the two-equation patient's self-service model

In the present paper logit models of the ordered dependent variable were adopted.

5.1. Estimation of the first equation[1]

When the criterion of significance and coincidence of all parameters was applied, the optimum model describing the P_7 variable includes Age, Sex, and the results of 7 tests at the admission.

[1] An example of an effective application of the ordered dependent variable models in economic research is the work of Bielak J., M. Kowerski, An approach to identify the factors determining the assessment of the regional labour market in the Lubelskie province, Prace i Materiały Instytutu Rozwoju Gospodarczego nr 80, SGH, Warsaw 2008, pp. 233–258 (in polish).

Table 5. The results of estimation of the first equation of the patient's self-service model at the admission

Dependent Variable: P_7

Method: ML – Ordered Logit (Quadratic hill climbing)

Sample: 4063

Included observations: 4063

Number of ordered indicator values: 6

Estimation settings: tol= 0.0001

Convergence achieved after 6 iterations

Covariance matrix computed using second derivatives

Estimations of structural parameters:

	Coefficient	Std. Error	z-Statistic	Prob,
Sex	0.1191	0.0604	1.9738	0.048400
Age	0.0231	0.0021	11.1381	<0.000001
P_2	0.2155	0.0289	7.4508	<0.000001
P_3	0.1407	0.0313	4.4893	<0.000001
P_4	0.1406	0.0301	4.6687	<0.000001
P_5	0.5073	0.0259	19.5766	<0.000001
P_6	0.4138	0.0258	16.0272	<0.000001
P_8	0.5137	0.0340	15.1179	<0.000001
P_9	0.3209	0.0213	15.0702	<0.000001
Limit Points				
LIMIT_2:C(1)	5.8970	0.1737	33.9478	<0.000001
LIMIT_3:C(2)	7.1503	0.1859	38.4608	<0.000001
LIMIT_4:C(3)	8.3850	0.1995	42.0382	<0.000001
LIMIT_5:C(4)	9.2419	0.2092	44.1779	<0.000001
LIMIT_6:C(5)	10.7596	0.2271	47.3848	<0.000001
Goodness-of-fit				
Akaike info criterion	2.2881	Schwarz criterion		2.3098
Log likelihood	-4634.227	Hannan-Quinn criter.		2.2958
Restr. log likelihood	-6889.575	Avg. log likelihood		-1.1406
LR statistic (9 df)	4510.696	LR index (Pseudo-R^2)		0.3274
Probability(LR stat)	0.000000			

Source: calculations with the eViews software.

The estimated model shows that all variables are stimulants, i.e. the higher value they adopt, the higher theoretical value of self-service test is. In other words, since the lowest rank means a norm, the worse the self-service test results are. In terms of likelihood it means that the higher values of independent variables, the higher probability of increased value of the self-service test at the admission. The McFadden coefficient is .3274, which is not a high value, but the significance of the likelihood ratio, measuring the significance of the whole set of variables in the model, proves high explanatory potential of the model. On the other hand, numerous research confirmed that the comparatively low value McFadden determination coefficients (for example exceeding .25) are usually sufficient to accept the goodness-of-fit of the model [Gruszczyński 2002, p. 55]. Additionally, the count R-squared, a percentage of correct indications of the model out of all indications [Maddala 1992, p. 334] is 65.1%.

The model allows analysis of relationship between age and the self-service status at the admission. In simulations it was assumed that a hypothetical patient has the following values of the independent variables:

1. 1^{st} quartiles of variables for the whole population
2. Medians of variables for the whole population
3. Means of variables for the whole population
4. 3^{rd} quartiles of variables for the whole population.

Table 6. Values of independent variables adopted for simulation

Variable	Q_1	Median	Mean	Q_3
P_2	2	3	3.37	5
P_3	2	3	3.08	4
P_4	1	2	2.25	3
P_5	2	3	3.57	6
P_6	1	2	2.57	4
P_8	1	1	1.59	1
P_9	2	6	4.24	6

Source: authors' calculation.

Table 7. Change of self-service probabilities depending on other tests, age and sex

Age	Men						Women					
	P_{71}	P_{72}	P_{73}	P_{74}	P_{75}	P_{76}	P_{71}	P_{72}	P_{73}	P_{74}	P_{75}	P_{76}
Independent variables (tests) at the level of the 1st quartile												
20	**0.868**	0.091	0.029	0.007	0.004	0.001	**0.853**	0.100	0.033	0.008	0.005	0.001
35	**0.823**	0.119	0.040	0.010	0.006	0.002	**0.805**	0.131	0.045	0.011	0.007	0.002
50	**0.766**	0.154	0.055	0.014	0.008	0.002	**0.744**	0.166	0.062	0.016	0.009	0.003
65	**0.699**	0.192	0.075	0.020	0.012	0.003	**0.673**	0.205	0.083	0.022	0.013	0.004
80	**0.622**	0.230	0.100	0.027	0.016	0.005	**0.593**	0.243	0.110	0.030	0.018	0.005
Independent variables (tests) at the level of the median												
20	**0.305**	0.301	0.235	0.085	0.057	0.017	0.281	**0.297**	0.247	0.093	0.063	0.019
35	0.237	**0.284**	0.268	0.109	0.078	0.024	0.216	0.275	**0.277**	0.118	0.086	0.027
50	0.180	0.255	**0.291**	0.136	0.104	0.034	0.163	0.243	**0.295**	0.145	0.115	0.038
65	0.135	0.218	**0.299***	0.163	0.137	0.047	0.121	0.205	**0.298***	0.172	0.150	0.053
80	0.099	0.179	**0.292**	0.187	0.177	0.066	0.089	0.166	**0.286**	0.194	0.192	0.073
Independent variables (tests) at the level of the mean												
20	0.230	**0.281**	0.271	0.112	0.081	0.025	0.209	0.272	**0.280**	0.121	0.089	0.028
35	0.174	0.251	**0.293**	0.139	0.108	0.035	0.158	0.238	**0.297**	0.149	0.119	0.040
50	0.130	0.213	**0.299***	0.167	0.142	0.049	0.117	0.200	**0.298***	0.175	0.155	0.055
65	0.095	0.174	**0.290**	0.190	0.182	0.068	0.086	0.161	**0.283**	0.197	0.197	0.076
80	0.070	0.138	**0.266**	0.206	0.227	0.094	0.062	0.126	**0.255**	0.209	0.243	0.104
Independent variables (tests) at the level of the 3rd quartile												
20	0.020	0.047	0.131	0.170	**0.358**	0.273	0.018	0.042	0.120	0.161	**0.361***	0.297
35	0.014	0.034	0.101	0.143	**0.361***	0.347	0.013	0.031	0.091	0.134	0.358	**0.374**
50	0.010	0.025	0.076	0.116	0.345	**0.429**	0.009	0.022	0.068	0.107	0.336	**0.458**
65	0.007	0.018	0.056	0.091	0.314	**0.515**	0.006	0.016	0.050	0.083	0.301	**0.544**
80	0.005	0.013	0.041	0.069	0.273	**0.600**	0.005	0.011	0.036	0.063	0.257	**0.628**

Note: 1 P_{71} means a probability of adopting a value of 1 in the self-service test, P_{72} means a probability of adopting a value of 2 in the self-service test, etc. 2 (*) inflection point. 3 The maximum probabilities for a given age are bolded, this means that the variable P_7 adopts a values from a column with bolded probability.

Source: authors' calculation

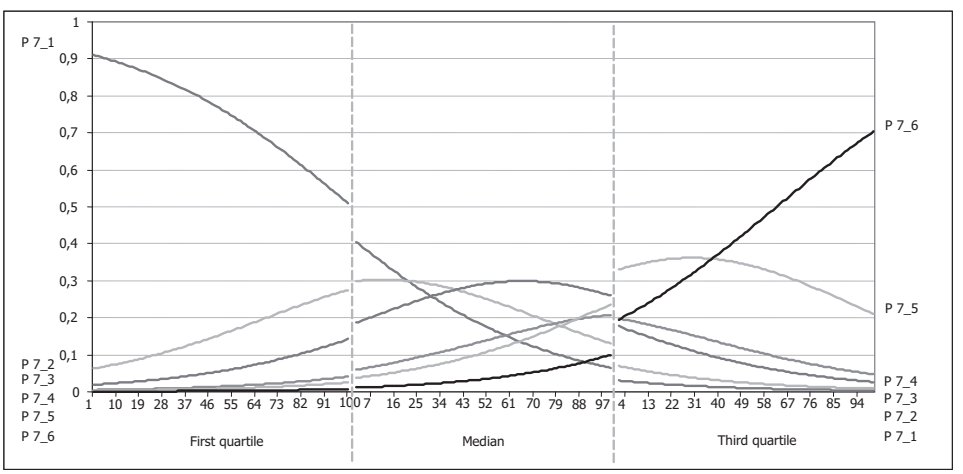
The simulation research shows that regardless of the results of other tests included in the model, the probability of good self-service declines and the probability of poorer self-service grows with ageing process. When the other tests reach values of their 1st quartiles, which means the norm for P_4, P_6 and P_8, or the value of 2 (at least good) for P_2, P_3, P_5 and P_9, the probability of normal self-service is .868 for a 20-year old man and .853 for a 20-year old woman. The

probability of norm for a 60-year old man is .699, and .673 for a 60-year old woman. In case of 80-year old persons these probabilities are .622 for a man and .593 for a woman. When the results of other tests worsen, the probability of norm is lower and the probability of poorer self-service increases. For example, in case of 35-year old man, for whom the other tests reached their median value, the probability of norm is .237 and is lower than the probability of that the self-service test adopts the value of 2 (probability of .284) [1].

In case of a 80-year old woman, for whom the other tests adopted values of the 3rd quartile (poor or very poor) the probability of norm is only .005, which means that only 5 women out of 1000 would have a self-service test result equal to the norm. For 63 it would be 4, for 257 the value would be 5, but the majority of women (628) would obtain the worst value of the self-service test.

It is also worth mentioning that for all groups the probability of norm is lower for women than for men at the same age.

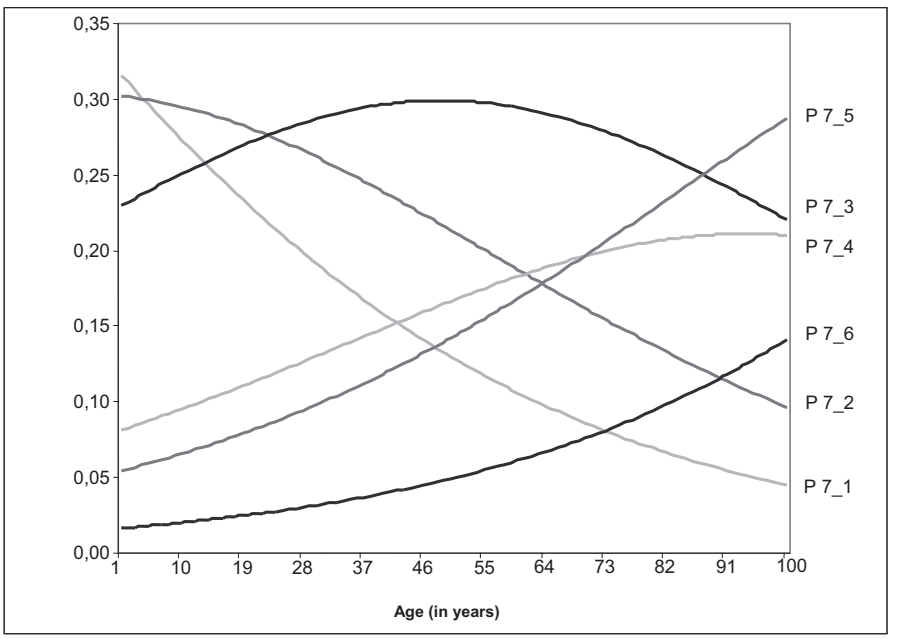
Figure 1. Probabilities for a given self-service status at the admission (P_7) at the assumed test results for P_2, P_3, P_4, P_5, P_6, P_8, P_9 at the admission versus age. Men.



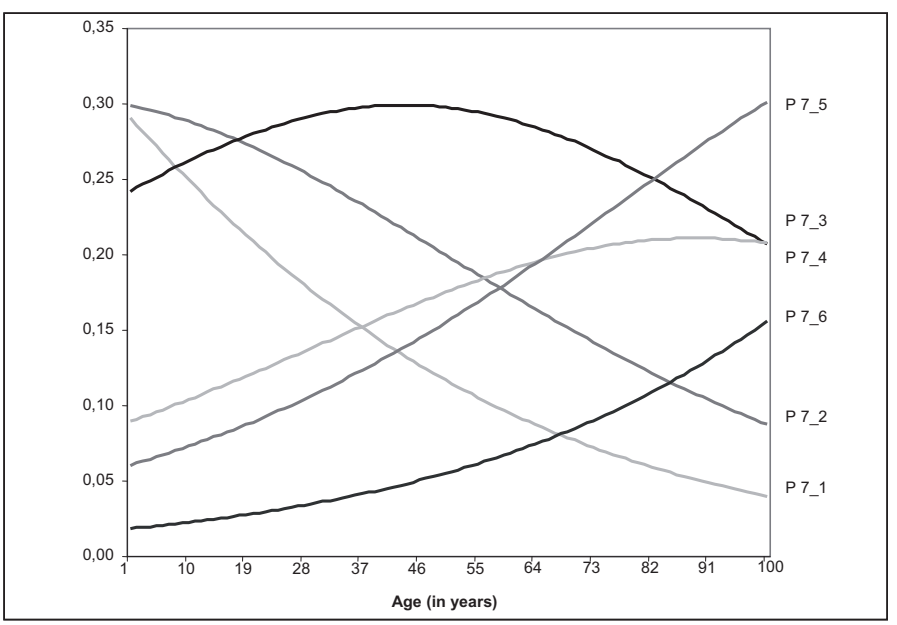
[1] If we change again from a theoretical probability to a theoretical rank, the latter for the variable P_7 will be 2 because the probability for 2 is the highest.

Figure 2. Probabilities for a given self-service status at the admission (P_7) assuming that the test results for P_2, P_3, P_4, P_5, P_6, P_8, P_9 at the admission reach the mean value, versus age.

Men



Women



The detailed analysis of changes in probability of particular self-service ranks versus age (incl. Fig. 1 and Fig. 2) leads to following conclusions:

1. The probability of the norm (1) decreases with age along a „reversed" logistic curve
2. The probability of the worst self-service rank (6) increases with age along a logistic curve
3. The probabilities of the other self-service ranks (from 2 to 5) increase up to a certain maximum and then decrease[1]; the "earliest" maximum is reached by the curve representing the probability of that the self-service rank is 2, then by the curve representing the probability of that the self-service rank is 3 and finally by the curve representing the probability of that the self-service rank is 5.

5.2. Estimation results for the second equation

According to the accepted methodology of building a recursive model the second equation describes changes of the self-service assessments at the release (K_7) but the patient's status at the admission is described by theoretical values of $\hat{P}_7$. There are two possible ways to calculate the theoretical value of the dependent variable in the first equation:

- the $\hat{P}_7$ variable adopts the rank, for which the probability estimated by the model is the highest
- the $\hat{P}_7$ variable is a mean rank weighted by the probabilities of the ranks

$$\hat{P}_7 = 1 \circ P(P_7 = 1) + 2 \circ P(P_2 = 2) + 3 \circ P(P_7 = 3) + 4 \circ P(P_7 = 4) + 5 \circ P(P_5 = 5) + 6 \circ P(P_7 = 6)$$

The second option was used in this paper.

[1] The analysis of probability functions for three-rank answers was presented in Kowerski M., On the Informative Value of „Remain Unchanged", Barometr Regionalny nr 4 (14), Wyższa Szkoła Zarządzania i Administracji w Zamościu, Zamość 2008, pp. 47–62 (in polish).

Table 8. Estimation results for the patient's self-service at the release

Dependent Variable: K_7

Method: ML - Ordered Logit (Quadratic hill climbing)

Sample: 1 4063

Included observations: 4063

Number of ordered indicator values: 6

Estimation settings: tol= 0.00010

Convergence achieved after 8 iterations

Covariance matrix computed using second derivatives

Structural parameters

	Coefficient	Std. Error	z-Statistic	Prob.
Sex	0.2046	0.0668	3.0652	0.002175
Age	0.0063	0.0023	2.6907	0.007131
P_{7teor}	0.2408	0.0328	7.3456	<0.000001
K_2	0.1910	0.0326	5.8588	<0.000001
K_3	0.2339	0.0352	6.6487	<0.000001
K_4	0.2510	0.0409	6.1349	<0.000001
K_5	0.5055	0.0330	15.3153	<0.000001
K_6	0.3833	0.0368	10.4107	<0.000001
K_8	0.3064	0.0326	9.3887	<0.000001
K_9	0.3474	0.0225	15.4546	<0.000001
Limit Points				
LIMIT_2:C(1)	5.9213	0.1789	33.1022	<0.000001
LIMIT_3:C(2)	7.5226	0.1956	38.4567	<0.000001
LIMIT_4:C(3)	9.1167	0.2143	42.5391	<0.000001
LIMIT_5:C(4)	10.4956	0.2318	45.2824	<0.000001
LIMIT_6:C(5)	12.4096	0.2654	46.7615	<0.000001
Goodness-of-fit				
Akaike info criterion	1.9866	Schwarz criterion		2.0099
Log likelihood	-4020.713	Hannan-Quinn criter.		1.9948
Restr. log likelihood	-5946.303	Avg. log likelihood		-0.9896
LR statistic (10 df)	3851.178	LR index (Pseudo-R^2)		0.3238
Probability(LR stat)	0.000000			

Source: calculated with eViews software.

In the model estimated here all parameters are significant and their signs are as expected. The probability that the self-service test is closer to the norm at the release is higher, if the patient is younger, the self-service test result at the admission was better and the results of other tests at the release were better. The McFadden determination coefficient is at the level of that for the first equation, but the accuracy of estimation for the second equation, measured by the count R-squared, is much higher and reaches 80.0%. A simulation of self-service status at the release was run depending on the theoretical values of the self-service test at the admission, age and sex, assuming that the results of other tests at the release are at the median level.

Table 9. Medians for the tests at the release

Variable	Median
K_2	2
K_3	2
K_4	1
K_5	2
K_6	1
K_8	1
K_9	3

Table 10. Changes of the self-service probabilities at the release depending on the theoretical values of the self-service test at the admission, age and sex, assuming that the results of other tests at the release are at the median level

Age	Men						Women					
	K_{71}	K_{72}	K_{73}	K_{74}	K_{75}	K_{76}	K_{71}	K_{72}	K_{73}	K_{74}	K_{75}	K_{76}
The theoretical value of the self-service test at the admission is 1												
20	**0.819**	0.139	0.034	0.007	0.002	0.000	**0.786**	0.162	0.041	0.008	0.002	0.000
35	**0.804**	0.149	0.037	0.007	0.002	0.000	**0.770**	0.173	0.045	0.009	0.003	0.000
50	**0.789**	0.160	0.040	0.008	0.002	0.000	**0.753**	0.185	0.049	0.010	0.003	0.000
65	**0.773**	0.171	0.044	0.009	0.003	0.000	**0.735**	0.197	0.053	0.011	0.003	0.001
80	**0.756**	0.183	0.048	0.010	0.003	0.000	**0.716**	0.210	0.058	0.012	0.003	0.001
The theoretical value of the self-service test at the admission is 2												
20	**0.780**	0.166	0.042	0.009	0.002	0.000	**0.743**	0.192	0.051	0.010	0.003	0.001
35	**0.763**	0.178	0.046	0.009	0.003	0.000	**0.725**	0.204	0.056	0.011	0.003	0.001
50	**0.746**	0.190	0.050	0.010	0.003	0.001	**0.705**	0.217	0.061	0.013	0.004	0.001
65	**0.728**	0.202	0.055	0.011	0.003	0.001	**0.686**	0.230	0.066	0.014	0.004	0.001
80	**0.709**	0.215	0.060	0.012	0.004	0.001	**0.665**	0.243	0.072	0.015	0.004	0.001
The theoretical value of the self-service test at the admission is 3												
20	**0.736**	0.197	0.053	0.011	0.003	0.001	**0.694**	0.224	0.064	0.013	0.004	0.001
35	**0.717**	0.209	0.058	0.012	0.003	0.001	**0.674**	0.237	0.069	0.014	0.004	0.001
50	**0.698**	0.222	0.063	0.013	0.004	0.001	**0.653**	0.250	0.075	0.016	0.005	0.001
65	**0.678**	0.235	0.068	0.014	0.004	0.001	**0.631**	0.263	0.082	0.017	0.005	0.001
80	**0.657**	0.248	0.074	0.016	0.005	0.001	**0.609**	0.276	0.089	0.019	0.006	0.001
The theoretical value of the self-service test at the admission is 4												
20	**0.687**	0.229	0.066	0.014	0.004	0.001	**0.641**	0.258	0.079	0.017	0.005	0.001
35	**0.666**	0.242	0.072	0.015	0.004	0.001	**0.619**	0.271	0.086	0.018	0.005	0.001
50	**0.645**	0.255	0.078	0.016	0.005	0.001	**0.597**	0.283	0.093	0.020	0.006	0.001
65	**0.623**	0.268	0.085	0.018	0.005	0.001	**0.574**	0.296	0.101	0.022	0.006	0.001
80	**0.604**	0.279	0.091	0.019	0.006	0.001	**0.551**	0.308	0.109	0.024	0.007	0.001
The theoretical value of the self-service test at the admission is 5												
20	**0.633**	0.263	0.082	0.017	0.005	0.001	**0.584**	0.290	0.097	0.021	0.006	0.001
35	**0.611**	0.275	0.089	0.019	0.006	0.001	**0.561**	0.303	0.105	0.023	0.007	0.001
50	**0.588**	0.288	0.096	0.021	0.006	0.001	**0.538**	0.315	0.114	0.025	0.007	0.001
65	**0.565**	0.301	0.104	0.023	0.007	0.001	**0.514**	0.326	0.123	0.028	0.008	0.001
80	**0.542**	0.313	0.112	0.025	0.007	0.001	**0.491**	0.336	0.132	0.030	0.009	0.002
The theoretical value of the self-service test at the admission is 6												
20	**0.575**	0.295	0.100	0.022	0.006	0.001	**0.525**	0.321	0.119	0.027	0.008	0.001
35	**0.552**	0.307	0.108	0.024	0.007	0.001	**0.501**	0.332	0.128	0.029	0.009	0.002
50	**0.529**	0.319	0.117	0.026	0.008	0.001	**0.478**	0.342	0.138	0.032	0.009	0.002
65	**0.505**	0.330	0.126	0.029	0.009	0.001	**0.454**	0.351	0.148	0.035	0.010	0.002
80	**0.482**	0.340	0.136	0.031	0.009	0.002	**0.431**	0.359	0.159	0.038	0.011	0.002

Source: authors' calculation.

Probability of the norm at the release of a 20-year old man, whose self-service test rank at the admission was at the level of the norm and the other tests at the release reached the median, was .819. If a 20-year man's self-service status at the admission was 2 and the other tests at the release were equal to their medians, the probability of the norm at the release is .780. In case of a 20-year old woman the results are a bit lower and reach .786 and .743, respectively. The probability of the norm at the release decreases with patient's age and the self-service level at the admission. For example, the probability of the norm for a 80-year old man, whose self-service status at the admission was 6 and the other tests at the release reached the median, was .482. For a 80-year old woman it is even lower and reaches .431.

Such high probabilities result from good results of the other tests at the release (median levels).

Table 11. Test means at the release

Variable	Mean
K_2	2.59
K_3	2.42
K_4	1.67
K_5	2.58
K_6	1.65
K_8	1.45
K_9	3.42

Source: authors' calculation.

Table 12. Changes of the self-service probabilities at the release depending on the theoretical values of the self-service test at the admission, age and sex, assuming that the results of other tests at the release are at the mean level

Age	Men						Women					
	K_{71}	K_{72}	K_{73}	K_{74}	K_{75}	K_{76}	K_{71}	K_{72}	K_{73}	K_{74}	K_{75}	K_{76}
The theoretical value of the self-service test at the admission is 1												
20	**0.574**	0.296	0.101	0.022	0.006	0.001	**0.523**	0.321	0.119	0.027	0.008	0.001
35	**0.551**	0.308	0.109	0.024	0.007	0.001	**0.500**	0.332	0.129	0.029	0.009	0.002
50	**0.527**	0.320	0.118	0.026	0.008	0.001	**0.476**	0.342	0.138	0.032	0.010	0.002
65	**0.504**	0.330	0.127	0.029	0.009	0.001	**0.453**	0.351	0.149	0.035	0.010	0.002
80	**0.481**	0.341	0.137	0.031	0.009	0.002	**0.430**	0.359	0.160	0.038	0.011	0.002
The theoretical value of the self-service test at the admission is 2												
20	**0.514**	0.326	0.123	0.028	0.008	0.001	**0.463**	0.347	0.144	0.033	0.010	0.002
35	**0.491**	0.336	0.132	0.030	0.009	0.002	**0.440**	0.356	0.155	0.037	0.011	0.002
50	**0.467**	0.346	0.142	0.033	0.010	0.002	**0.417**	0.363	0.166	0.040	0.012	0.002
65	**0.444**	0.354	0.153	0.036	0.011	0.002	**0.394**	0.369	0.177	0.044	0.013	0.002
80	**0.421**	0.362	0.164	0.039	0.012	0.002	0.372	**0.374**	0.189	0.048	0.015	0.003
The theoretical value of the self-service test at the admission is 3												
20	**0.454**	0.351	0.148	0.035	0.010	0.002	**0.404**	0.367	0.172	0.042	0.013	0.002
35	**0.431**	0.359	0.159	0.038	0.011	0.002	**0.382**	0.372	0.184	0.046	0.014	0.002
50	**0.408**	0.366	0.170	0.041	0.013	0.002	0.360	**0.376**	0.196	0.050	0.015	0.003
65	**0.393**	0.369	0.178	0.044	0.013	0.002	0.338	**0.379**	0.209	0.054	0.017	0.003
80	0.364	**0.376**	0.194	0.049	0.015	0.003	0.318	**0.380**	0.221	0.059	0.018	0.003
The theoretical value of the self-service test at the admission is 4												
20	**0.395**	0.369	0.177	0.043	0.013	0.002	0.348	**0.378**	0.203	0.052	0.016	0.003
35	0.373	**0.374**	0.189	0.047	0.014	0.003	0.327	**0.380**	0.216	0.057	0.018	0.003
50	0.352	**0.377**	0.201	0.052	0.016	0.003	0.306	**0.380***	0.229	0.062	0.019	0.003
65	0.330	**0.379**	0.213	0.056	0.017	0.003	0.287	**0.379**	0.242	0.067	0.021	0.004
80	0.310	**0.380**	0.226	0.061	0.019	0.003	0.268	**0.377**	0.255	0.073	0.023	0.004
The theoretical value of the self-service test at the admission is 5												
20	0.340	**0.379**	0.208	0.054	0.017	0.003	0.295	**0.380**	0.236	0.065	0.020	0.004
35	0.319	**0.380***	0.221	0.059	0.018	0.003	0.276	**0.378**	0.249	0.071	0.022	0.004
50	0.299	**0.380**	0.234	0.064	0.020	0.004	0.258	**0.375**	0.262	0.077	0.024	0.004
65	0.279	**0.378**	0.247	0.070	0.022	0.004	0.240	**0.370**	0.275	0.083	0.027	0.005
80	0.261	**0.376**	0.260	0.076	0.024	0.004	0.223	**0.365**	0.287	0.090	0.029	0.005
The theoretical value of the self-service test at the admission is 6												
20	0.288	**0.379**	0.241	0.067	0.021	0.004	0.248	**0.373**	0.269	0.080	0.026	0.005
35	0.269	**0.377**	0.254	0.073	0.023	0.004	0.231	**0.367**	0.282	0.087	0.028	0.005
50	0.251	**0.373**	0.267	0.079	0.025	0.005	0.214	**0.361**	0.294	0.094	0.031	0.006
65	0.234	**0.368**	0.280	0.086	0.028	0.005	0.199	**0.353**	0.307	0.102	0.034	0.006
80	0.217	**0.362**	0.292	0.093	0.030	0.005	0.184	**0.344**	0.318	0.110	0.037	0.007

Source: authors' calculation.

When the results of other tests at the release worsen, the probabilities of the norm decrease. For example, the probability of the norm for a 80-year old man, whose self-service status at the admission was 6 and the other tests at the release reached the mean, was .362

Table 13. Changes of the probability of K_7 reaching a particular value when theoretical P_7 changes from 1 to 6 (based on data in the Table 12)

Age	Men						Women					
	K_{71}	K_{72}	K_{73}	K_{74}	K_{75}	K_{76}	K_{71}	K_{72}	K_{73}	K_{74}	K_{75}	K_{76}
20	-0.286	0.083	0.140	0.045	0.015	0.003	-0.276	0.051	0.150	0.054	0.018	0.003
35	-0.282	0.069	0.145	0.049	0.016	0.003	-0.269	0.035	0.153	0.058	0.020	0.004
50	-0.277	0.054	0.149	0.053	0.018	0.003	-0.262	0.019	0.156	0.062	0.021	0.004
65	-0.270	0.038	0.153	0.057	0.019	0.003	-0.254	0.002	0.158	0.067	0.023	0.004
80	**-0.263**	0.021	0.156	0.061	0.021	0.004	-0.245	-0.015	0.159	0.072	0.025	0.005

Source: authors' calculation.

Obviously, the worse self-service status at the admission, the lower probability of the norm at the release (assuming that the other test results are the same). For example, for 20-year old men, if the self-service status at the admission worsens from the norm (1) to total disability (6), the probability of the norm at the release decreases by .263, assuming that the other tests remain at the mean level.

6. Conclusions

The two-equation model estimated suggests that the models of ordered dependent variable are suitable to determine the factors affecting locomotive abilities of the patients. This model allows simulating the probabilities of the patient's self-service status both at the admission and at the release from a hospital depending on various factors describing a patient.

REFERENCES

BIELAK J., M. KOWERSKI 2008: Próba określenia czynników determinujących oceny regionalnego rynku pracy przez mieszkańców województwa lubelskiego. Prace i Materiały Instytutu Rozwoju Gospodarczego nr 80, SGH, Warszawa, s. 233–258.

EViews 5.1 User's Guide. Quantitative Micro Software, LLC, Irvine CA.

GÓRSKI W., J. GROSSMAN, B. NEJMAN 1985: Problemy rehabilitacji ruchowej w geriatrii. Wydawnictwo Akademii Wychowania Fizycznego w Warszawie, Warszawa.

GREENE W. H. 2003: Econometric Analysis. Fifth Edition, Prentice Hall, New Jersey.

GROSSMAN J. 1987: Niektóre uwarunkowania efektywności rehabilitacji dzieci z wadami rozwojowymi. Wydawnictwo Akademii Wychowania Fizycznego w Warszawie, Warszawa.

GROSSMAN J., G. RAWICZ-MAŃKOWSKI 1988: Analiza czynników determinujących wyniki rehabilitacji narządów ruchu u dzieci. Wydawnictwo Akademii Wychowania Fizycznego w Warszawie, Warszawa.

GRUSZCZYŃSKI M. 2002: Modele i prognozy zmiennych jakościowych w finansach i bankowości. Oficyna Wydawnicza Szkoły Głównej Handlowej w Warszawie, Warszawa.

HELLWIG Z. 1976: Przechodniość relacji skorelowania zmiennych losowych I płynące stąd wnioski ekonometryczne. Przegląd Statystyczny.

KOWERSKI M. 2008: Wartość informacyjna odpowiedzi „bez zmian" w badaniach nastrojów gospodarczych. Barometr Regionalny. Analizy I prognozy nr 4(14), Wyższa Szkoła Zarządzania i Administracji w Zamościu, Zamość, s. 47–62.

MADDALA G. S. 1992: Introduction to econometrics, 2nd ed., Macmillan, New York.

MCFADDEN D. 1974: Conditional logit analysis of qualitative choice behaviour. [w:] P. Zarembka (ed.): Frontiers in econometrics. Academic Press, New York.

STANISZ A. 2001: Przystępny kurs statystyki w oparciu o program STATICTICA PL na przykładach medycznych. StatSoft Polska sp. z o.o., Kraków.

SYCZEWSKA E. M. 2001: Wielorównaniowe modele ekonometryczne. [w:] Gruszczyński M., M. Podgórska (red.): Ekonometria. Wydanie szóste, Oficyna Wydawnicza Szkoły Głównej Handlowej w Warszawie, Warszawa, s. 209–210.

WEISS M., A. ZEMBATY (red.) 1983: Fizjoterapia, Państwowy Zakład Wydawnictw Lekarskich, Warszawa, 1983.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 179—192

IDENTIFYING AN APPROPRIATE FORECASTING MODEL FOR FORECASTING TOTAL IMPORT OF BANGLADESH

TANVIR KHAN[1]

ABSTRACT

Forecasting future values of economic variables are some of the most critical tasks of a country. Especially the values related to foreign trade are to be forecasted efficiently as the need for planning is great in this sector. The main objective of this research paper is to select an appropriate model for time series forecasting of total import (in taka crore) of Bangladesh. The decision throughout this study is mainly concerned with seasonal autoregressive integrated moving average (SARIMA) model, Holt-Winters' trend and seasonal model with seasonality modeled additively and vector autoregressive model with some other relevant variables. An attempt was made to derive a unique and suitable forecasting model of total import of Bangladesh that will help us to find forecasts with minimum forecasting error.

Key words*:* ARIMA model, Holt Winters' trend and seasonality method, VAR model, Forecasting accuracy, Out-of-sample accuracy measurement.

1. Introduction

An important economic concept involves international trade and finance. International trade in goods and services allows nations to raise their standards of living by exporting and importing goods and services. In a modern economy the economic condition is highly affected by the amount of its foreign trade and its balance of trade. Imports, along with exports, form the basis of trade. Bangladesh has had a negative trade balance since its independence, and the gap between

[1] MS Student. Institute of Statistical Research and Training, University of Dhaka, Dhaka-1000, Bangladesh. E-mail: tkhan1@isrt.ac.bd.

export and import are still widening. The country is importing a lot of goods from foreign countries and the aim of this paper is to find a forecasting model that will help us to get ideas about the future values of total import of Bangladesh.

1.1. Data and variables

This study is conducted on total import of Bangladesh. The data set has 133 observations, during the time period from July 1998 to July 2009, in the initialization set and 14 observations, during the time period from August 2009 to September 2010, in the test set. For vector autoregressive model two more variables - total export and net foreign assets, are also used. The data were obtained from the following sources for validation:

1. Statistical Department of Bangladesh Bank
2. 'Economic Trends', a monthly report published by Bangladesh Bank.

2. Methodology

In this study three methods are used, i.e. seasonal ARIMA model, Holt Winters' trend and seasonality method and vector autoregressive model. The goal is to find an appropriate model that has both in-sample and out-of-sample forecasting errors as small as possible.

2.1. ARIMA model

A model containing p autoregressive terms and q moving average terms is classified as ARMA(p,q) model. If the series is differenced d times to achieve stationary, the model is classified as ARIMA(p,d,q), where the symbol 'I' signifies 'integrated'. The equation for the ARIMA (p,d,q) model is as follows:

$$Y_t = c + a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots . + a_p Y_{t-p} + e_t - b_1 e_{t-1} - b_2 e_{t-2} - \ldots . - b_q e_{t-q} \quad (1)$$

Or, in backshift notation:

$$\left(1 - a_1 L - a_2 L^2 - \ldots . a_p L^p\right)(1 - L) Y_t = c + (1 - b_1 L - b_2 L^2 - \ldots . - b_q L^q) e_t \quad (2)$$

The ARIMA notation can be extended readily to handle seasonal aspects, and a Seasonal ARIMA (p,d,q)(P,D,Q) or SARIMA model can be represented as

$$\left(1-L-L^2-L^3-\ \ldots\ L^d\right)\left(1-L^S-L^{2(S)}-\ \ldots\ldots\ L^{D(S)}\right)\left(1-a_1L-a_2L^2-\ \ldots\ a_pL^p\right)\left(1-A_1L-A_2L^2\ldots A_PL^P\right)Y_t=\left(1-b_1L-b_2L^2-\ \ldots.-b_qL^q\right)\left(1-B_1L-B_2L^2-\ \ldots-B_QL^Q\right) \quad (3)$$

2. 2. Holt Winters trend and seasonality method

The Holt-Winters' method is based on three smoothing equations-one for the level, one for trend, and one for seasonality. In fact, there are two different Holt-Winters' methods, depending on whether seasonality is modeled in an additive or multiplicative way. The basic equations for Holt-Winters' multiplicative method is as follows:

Level: $L_t=\alpha\frac{Y_t}{S_{t-s}}+(1-\alpha)(L_{t-1}+b_{t-1})$ (4)

Trend: $b_t=\beta(L_t-L_{t-1})+(1-\beta)b_{t-1}$ (5)

Seasonal: $S_t=\gamma\frac{Y_t}{L_t}+(1-\gamma)S_{t-s}$ (6)

Forecast: $F_{t+m}=(L_t+b_tm)S_{t-s+m}$ (7)

Where s is the length of seasonality (e.g. number of month or quarter in a year), L_t represents the level of the series, b_t represents the trend, S_t is the seasonal component, and F_{t+m} is the forecast for m period ahead. The first two equations for additive model are identical to the first two equations of the multiplicative method. The only difference is in the third equation, that is the seasonal indices are now added and subtracted, i.e.:

Seasonal: $S_t=\gamma(Y_t-L_t)+(1-\gamma)S_{t-s}$ (8)

2.3. VAR model

In a VAR model the variables which have bilateral causality with each other have been included. A VAR model consists of a set of variables $Y_t=(Y_{1t},Y_{2t},\ldots,Y_{Kt})$ which can be represented as

$$Y_t=\alpha+A_1Y_{t-1}+A_2Y_{t-2}+\ \ldots+A_pY_{t-p}+u_t \quad (9)$$

With A_i are (K×K) coefficient matrix for i=1,2,….p and u_t is a p dimensional process with E$(u_t)=0$ and covariance matrix $E(u_tu_t^T)=\sum_u$.

3. COMPARISON AMONG THE FORECASTING METHODS

To make comparison among the methods some well known measures of forecast error are used. The model that gives the minimum measures of these errors will be the expected model for further forecasting. The measures used are cited below.

Mean Error (ME): The mean error gives the average forecast error, i.e.:

$$ME = \frac{1}{n}\sum_{i=1}^{n} e_t$$

Where

$$e_t = Y_t - F_t$$

Y_t= The observation at time t, F_t=Forecasted value at time t, n= the number of observation.

Mean Absolute Error (MAE): The MAE is first defined by making each error positive by taking its absolute value, and then averaging the result, i.e.:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_t|$$

Mean Squared Error (MSE): The MSE is defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n} {e_t}^2$$

Mean Percentage Error (MPE): The MPE is the mean of the relative or percentage error and is given by

$MPE = \frac{1}{n}\sum_{i=1}^{n} PE_t$; Where $PE_t = \frac{Y_t - F_t}{Y_t} \times 100$ is the relative or percentage error at time t.

Mean Absolute Percentage Error (MAPE): The MAPE is defined as

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} |PE_t|$$

3.1. Out-of-sample accuracy measurement

The summary statistics described so far measures the goodness of fit of the model to historical data. Such fitting does not necessarily imply good forecasting. An MSE or MAPE of zero can always be obtained in the fitting phase by using a polynomial of sufficiently high order. These problems can be overcome by measuring true out-of-sample forecast accuracy. That is, the total data are divided into an 'initialization' set and a 'test' set or 'holdout' set. Then the initialization set is used to estimate any parameters and to initialize the method. Forecasts are made for the test set. The accuracy measures are computed for the errors in the test set only.

4. Analysis of data

The initialization set of the data has 133 data points, starting from July 1998 to July 2009. The first step of the analysis is to plot the whole dataset to visualize the nature of it. The time plot for total import (in taka crore) of Bangladesh is shown in the figure given below.

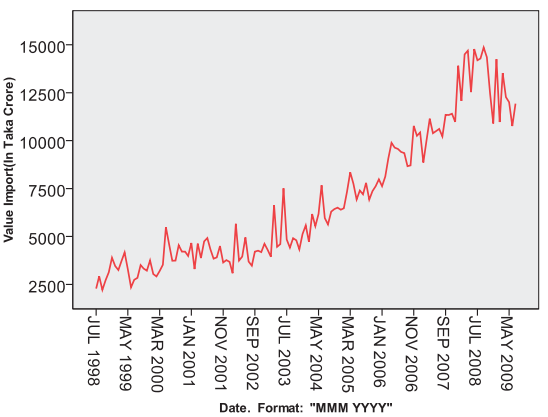


Figure 1. Time series plot for actual data

4.1. ARIMA model

It can be seen from Figure 1 that the data had an upward trend but from October 2008 it started to show a downward trend. From simple view of the plot we can have an idea about non-stationarity in mean but stationarity in variance. To become more certain about stationarity of the data, Augmented Dickey Fuller test has been conducted. The calculated value of Dickey Fuller is -2.0622 with P

value .5506 at suggested lag 5. So, we fail to reject the null hypothesis of non-stationarity.

The PACF of the data and ACF of the first differenced data are:

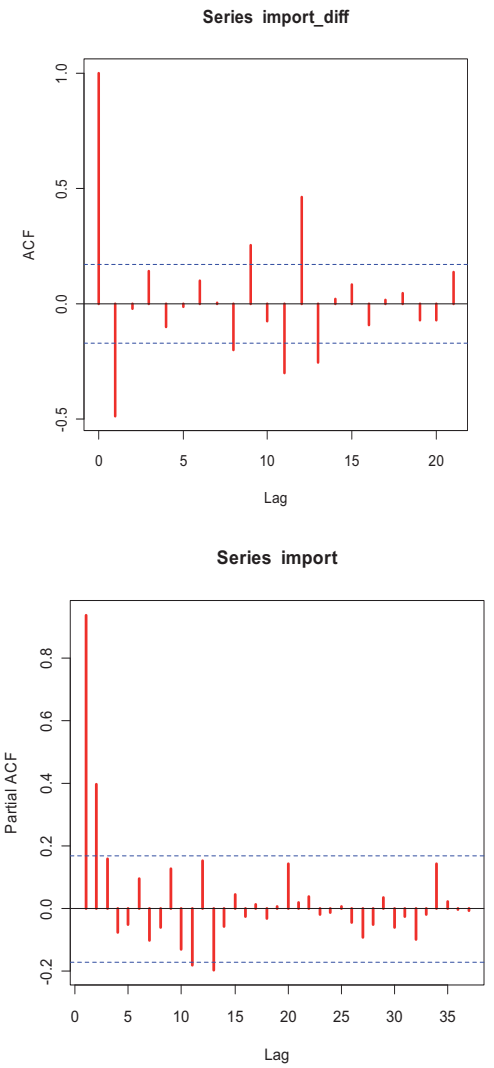


Figure 2. PACF of original data and ACF of first differenced data

From the PACF plot it can be seen that the partial autocorrelation at lag 12 is insignificant and so are at further seasonal lags (i.e. 24, 36 etc.). Thus, it can be assumed that the data do not have strong seasonality.

To obtain stationarity the original data are differenced at order one. As the data set do not have strong seasonality, a first order non-seasonal difference is taken rather than a seasonal difference and ADF test is conducted again. The value of Dickey Fuller for the first differenced data is -5.3819 with P value less than .01 at suggested lag 5. This clearly indicates that our data are now stationary. The ACF plot of the first differenced data shows significant spikes at lag 12, which gives a clear indication that a seasonal term must be included in the model. To be certain about the form of the appropriate model, the AIC and BIC values are checked for all the probable models and it is found that ARIMA (0,1,1,)(1,0,0)[12] model has the smallest AIC and BIC values, so, this model should be the desired ARIMA model.

The estimated ARIMA (0,1,1) (1,0,0)12 model is given below:

$$\widehat{M}_t = 0.4733M_{t-12} - M_{t-1} + 0.4733M_{t-13} + .5088e_{t-1} \quad (10)$$

The value of Box-Pierece Q statistic 30.6526 is with degrees of freedom 42 with P value .9026 at lag 44. So, the null hypothesis that residuals of ARIMA (0,1,1)(1,0,0)12 model are white noise fails to be rejected. Again, the value of Ljung-Box Q statistic 36.4945 is with degrees of freedom 42 with P value 0.7107 at lag 44. So, it can be said that the residuals are white noise for ARIMA(0,1,1)(1,0,0)12 model.

The plot of forecasted value along with the original time series is

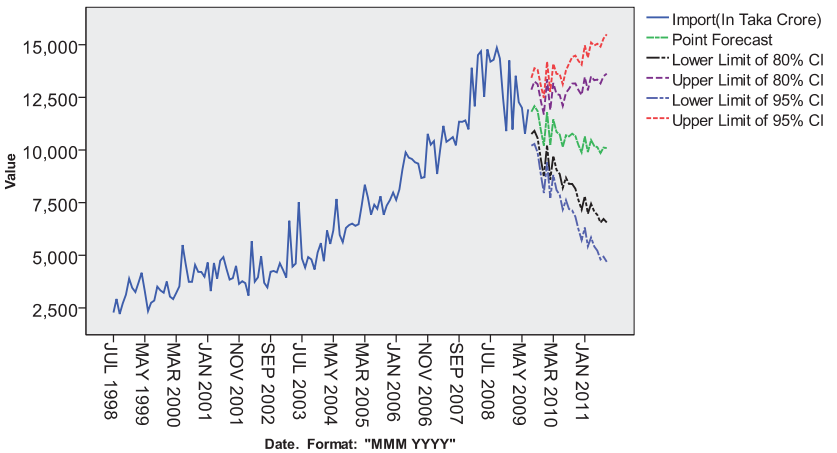


Figure 3. Point and interval forecasting with ARIMA (0,1,1)(1,0,0)12 model

4.2. Holt Winters trend and seasonality method

To initialize the Holt-Winters' forecasting method, initial values of the level L_t, the trend b_t, and the seasonal indices S_t are needed. To determine initial estimates of the seasonal indices 12 data points were used as the data set is monthly. The estimated model is

$$L_t = .3925046(M_t - S_{t-12}) + (1 - .3925046)(L_{t-1} + b_{t-1}) \quad (11)$$

$$b_t = b_{t-1} \quad (12)$$

$$S_t = 0.8301905(M_t - L_t) + (1 - 0.8301905)S_{t-12} \quad (13)$$

The plot of forecasted value obtained from Holt Winters model is given below.

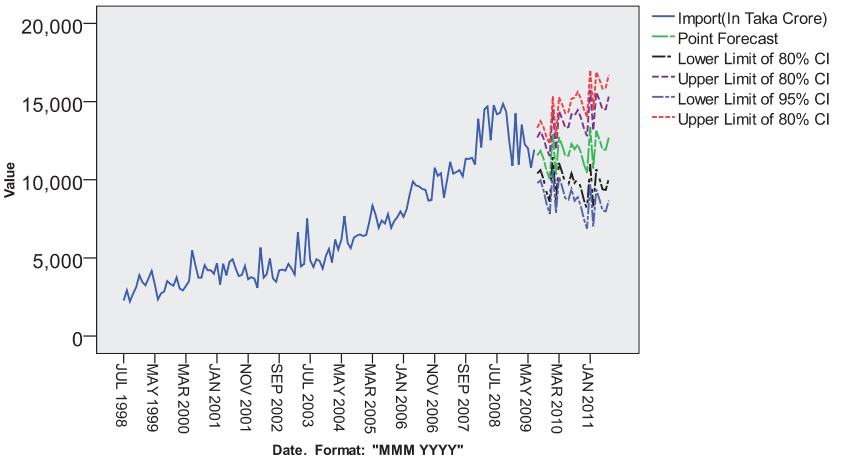


Figure 4. Point and Interval Forecasts obtained from Holt-Winters Model

4.3. VAR model

To fit a VAR model, at first an attempt was made to find relevant variables that have bilateral causality with total import of Bangladesh as well as with each other. Primarily 5 variables were chosen. They are: total export, net foreign assets, domestic credit, exchange rate and inflation rate. Among these variables, it was found that only total export and net foreign assets have bilateral relationship with each other as well as with total import. The results or Granger causality tests are given below.

Table 1. Result of Granger causality test

relationship	Lag	F-statistic	P-value
Export→Import	3	3.1070343	2.901318e-02
Import→Export	3	7.5056727	1.180776e-04
Net Foreign Assets→Import	5	2.582494	2.966103e-02
Import→Net Foreign Assets	5	2.460508	3.696343e-02
Export→Import	9	2.758777	0.0061726486
Import→Export	9	2.954010	0.0036209259
Net Foreign Assets→Export	9	3.881317	0.0002814135
Export→Net Foreign Assets	9	2.346345	0.0187481360

Thus, it can be assumed that total export and net foreign assets can be used as endogenous variables with total import in the desired VAR model. The natural log forms of all the three variables are taken and for dealing with seasonality all the three variables are seasonally adjusted by the classical

multiplicative decomposing method. The modified data set are then analyzed. As at lag 2 the model gives the smallest AIC, FP and HQ values, the model is fitted with lag 2. The estimated model has the form

$$\widehat{M}_t = 0.52533 + 0.37965M_{t-1} - 0.02456M_{t-2} - 0.15031X_{t-1} + 0.5257X_{t-2} + 0.02215F_{t-1} + 0.18856F_{t-2} \quad (14)$$

$$\hat{X}_t = 0.0342 - 0.1211M_{t-1} + 0.0234M_{t-2} + 0.14806X_{t-1} - 0.0225X_{t-2} + 1.07017F_{t-1} - 0.09139F_{t-2} \quad (15)$$

$$\hat{F}_t = 0.0342 - 0.12107M_{t-1} + 0.02343M_{t-2} + 0.14806X_{t-1} - 0.0225X_{t-2} + 1.07017F_{t-1} - 0.09139F_{t-2} \quad (16)$$

For checking whether the required type of causality exists in our model or not, Granger causality test was done again.

Table 2. Results of Multivariate Granger Causality test

Null Hypothesis	F value	P value	Decision
Total export does not Granger cause total import and net foreign assets	6.6997	3.244e-05	Null hypothesis is rejected with more than 99% confidence
total import does not Granger cause total export and net foreign assets	3.7346	0.0054	Null hypothesis is rejected with more than 99% confidence
net foreign assets does not Granger cause total import and total export	2.9141	0.02141	Null hypothesis is rejected with more than 95% confidence

At lag 31, asymptotic Portmanteau test has the Chi-Squared value 250.7905 with P value 0.664 at lag 31 and the adjusted Portmanteau test has Chi-Squared value 286.9564 with P value 0.1294. Thus, the model has white noise residuals. The forecasted values of the model are then back-transformed. The plot of original data and forecasted values is

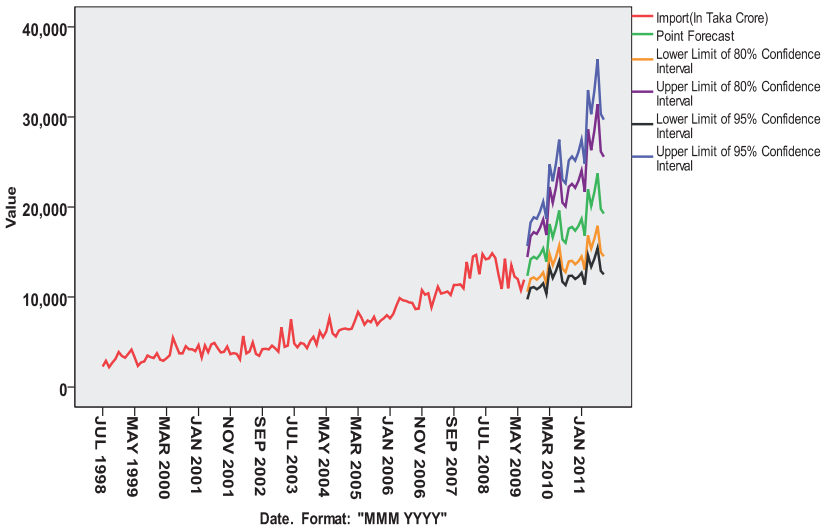


Figure 5. Point and Interval Forecasts obtained from VAR model

5. Comparison among arima, holt winters trend and seasonality method and VAR model

The forecasting performance of these three models have been compared with each other with respect to different measures of error and the summary measures are listed in the table below.

Table 3. In-sample error measures of the methods

Measures of Error	ARIMA (0,1,1)(1,0,0)12	Holt-Winters' Method	VAR Model
Mean Error	58.61606930	80.57279	32.63461
Mean Absolute Error	588.55960438	676.039	616.8842
Mean Squared Error	671123.9	808836.5	780904.8
Mean Percentage Error	0.01089789	0.08561744	-0.7211356
Mean Absolute Percentage Error	9.48140336	10.56153	9.086229

In-sample error measures do not necessarily imply good forecasting model. To find which forecasting method is better, true out-of-sample forecast accuracy was measured.

The plot of the test set observations with the forecasted values using all three methods is given below.

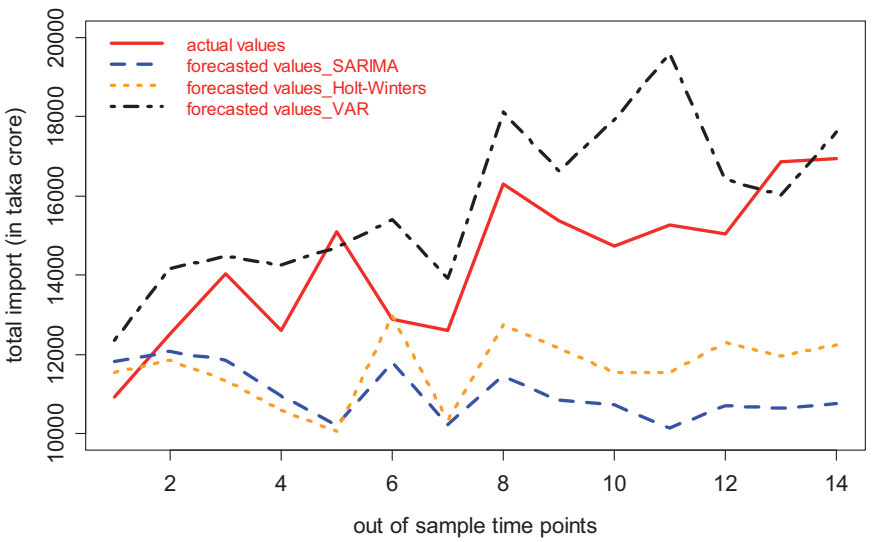


Figure 6. Plot of test set of the data and forecasted values obtained from the three methods

The out-of-sample accuracy measures are given below.

Table 4. Out-of-sample accuracy measurements

Measures of Error	ARIMA(0,1,1)(1,0,0)12	Holt-Winters' Method	VAR Model
Mean Error	3350.570	2712.007	-1460.056
Mean Absolute Error	3477.229	2815.95	1636.428
Mean Squared Error	15747374	10227508	3775589
Mean Percentage Error	21.81878	17.79637	-10.5399
Mean Absolute Percentage Error	22.97802	18.72796	11.62293

It is clear from the above table that VAR model is giving minimum values for all the measures of forecast error. Therefore, one may take VAR model of total import, total export and net foreign assets as an appropriate model for forecasting total import of Bangladesh.

6. Conclusion

The basic aim of this paper is to select an appropriate model that will be helpful for understanding future behavior of total import of Bangladesh. Three methods are used - seasonal ARIMA, Holt-Winters' trend and seasonality method as well as VAR model. Various measures of forecasting accuracy were also measured for all the three models. The comparison shows that VAR model of total import with total export and net foreign assets as other endogenous variables is better than the other two methods as it is producing both in-sample and out-of-sample forecasting errors. But this is not the end. More researches have to be done to handle seasonality in a better way and to find more relevant variables that are useful for forecasting total import of Bangladesh.

REFFERENCES:

AMISANO, G. AND GIANNINI, C. (1997). Topics in Structural VAR Econometrics, Springer-Verlag, Berlin, 2nd edition.

GUJRATI, D. N. AND SANGEETHA (2007). Basic Econometrics, McGraw-Hill Book Co, New York.

HAMILTON,J.D. (1994). Time Series Analysis, Princeton University Press, Princeton.

MAKRIDAKIS, S., WHEELWRITGHT, S. C. AND HYNDMAN, R. J. (1998). Forecasting Methods and Applications, John Wiley and Sons, Ink., New York.

PFAFF, B. (2008). VAR, SVAR and SVEC Models: Implementation within R Package vars, New York. URL: http://CRAN.R-project.org/package=vars.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 193—206

DOES A CHANGE OF OCCUPATION LEAD TO HIGHER EARNINGS?

Barbara Liberda[1], Marek Pęczkowski[2]

ABSTRACT

The aim of this paper is to identify how the mobility between different types of broadly defined occupation (hired work, self-employment in industry, services and agriculture or social security beneficiaries) changes personal income of individuals. We apply the Markov matrices to the panel data on 30,540 individuals for 2007-2008 from the Polish Household Budget Surveys. Our hypothesis is that a change of occupation affects individual capability to earn income, controlling for the occupation a person quits and the occupation a person starts, as well as age, education level and a permanent or temporary character of work. We test our hypothesis using the regression analysis. Our results show that the inter-occupational mobility matters mostly for those quitting hired work for self-employment, for the better educated, as well as for respondents above 60 years of age.

Key words: income, earning, mobility, occupation, hired work, self-employment
JEL: D11, D12, D14

1. Introduction

Earning capability of an individual depends on the level and quality of human capital and the occupation a person performs as well as social and cultural relations in which the process of earning takes place. The hidden individual and social capabilities of earning can be measured by their influence on manifest variables like education or occupation.

In the neoclassical economics the consumer is rational and maximizing utility. The standard consumption theory of life cycle and the permanent income theory (Modigliani 1986; Friedman 1957) assume that income is earned to smooth consumption during the life cycle or in conceivably long periods when consumers can assume that their income will be relatively stable. The type and place of work

[1] University of Warsaw, Faculty of Economic Sciences. E-mail: barbara.liberda@uw.edu.pl
[2] University of Warsaw, Faculty of Economic Sciences. E-mail: mpeczkowski@wne.uw.edu.pl

or occupation affect income uncertainty, e.g. income earned in agriculture is treated as less stable than income earned in industry or services.

The human capital theories (Becker 1994; Arrow 1962; Mincer 1974; Schultz 1971; Ben Porath 1967) view earnings as the outcome of utilization of skills and capabilities accumulated in individuals during the period of education, training and learning by doing in the process of production. The human capital earning function of Mincer (1974) shows the positive correlation of earnings and the level of human capital expressed in years of schooling and years of experience or age (if training and experience cannot be fully observed in age groups). Earnings are also dependent on the signaling value of a school or university diploma which is treated at the labor market as a pure signal independent from skills that it should prove (Spence, 1973).

Gary Becker (1994, third edition) points to the importance of on-the-job training for increasing human capital of employees and hence their productivity and earnings. Training can be of general character or specific to a particular firm, but in both cases it is strongly connected to the workplace.

Particular skills gained from education and training may be utilized differently in occupations that are performed by individuals. This may lead to earning differences of workers (with similar human capital and age) who are economically active in different sectors and occupations. Thus, earnings are a function not only of education, experience and age, but also of the type of economic activity in which the person acts as employee, self-employed (in industry, services, agriculture), unemployed or retiree.

The earning capability depends also on individual expectations or desires of a preferred income (Liberda 2007). If a person perceives personal income as insufficient for individual and family needs she or he may make a decision to increase earning by a change of occupation or a place of work within a country or between countries (Milanovic 2008).

In this paper we analyze the inter-occupational mobility of individuals and the change of personal income when they quit one type of broadly defined occupation and enter another occupation. The analysis is based on the panel data for 30,540 respondents from the 2007-2008 Polish Household Budget Surveys. We apply the Markov mobility matrices and the regression analysis to account for the increase or decrease of personal income when the occupation is changed by individuals grouped by age, education level, as well as the permanent or temporary character of work.

2. Data

The data concern 30,540 individuals who were surveyed in the same month during two consecutive years (2007-2008) in the Polish Household Budget Surveys and thus constitute a panel. Since 1993 household budget surveys in Poland have been based on the monthly rotation method. The monthly rotation

assumes that one household participates in the survey for one month in a year, and then, part of households participate in the survey in the corresponding month of the following year. Until 2000, 50% of randomly selected households participated in surveys for 4 consecutive years (constant sample) and 50% of households participated in surveys only once (variable sample). It enabled to create four-year panels of households.

Since 2001 the same households have been surveyed only for two consecutive years. In 2007, the following two independent subsamples selected in 2005 were surveyed:

subsample 1 for the surveys in 2006-2007,

subsample 2 for the surveys in 2007-2008.

It means that during one year half of surveyed households are exchanged. Currently, during one year, over 30,000 households are surveyed, which is about 0.25% of the population. In selected households all persons included in the household participate in the survey. The person at the age of 16 and over who gains the highest income of all the household members is a reference person.

Household's net disposable income is defined as a sum of all household members' current incomes from various sources reduced by prepayments on personal income tax and by social security and health insurance contributions. The household disposable income covers also an estimated value of income in kind (e.g. natural consumption) as well as goods and services received free of charge, e.g. social security benefits and gifts received from other households. The main sources of income are:

income from hired work,

income from a private farm in agriculture,

income from self-employment (outside a private farm in agriculture or from free profession),

income from property or from rental of a property or land,

social security benefits (old age pension, disability pension or family pension),

other social benefits,

other income (including gifts and alimonies).

The main source of maintenance is defined as the only or prevailing source of maintenance. A person who does not have her/his own source of maintenance (either earned or unearned) and is maintained by other members of the household is a dependent.

The panel of households created in 2007-2008 consists of 15,853 households (unweighted sample). To analyze incomes of household members, we have created a panel of persons included in the surveyed households. The panel of individuals included the persons who were members of the households in 2007 as well as in 2008, and were at the age of 19 or more in 2007.

The 2007-2008 Panel of individuals (members of households) consists of 30,540 persons.

For this research we used the net personal income of individuals measured in nominal terms in the same month in 2007 and 2008. Occupation of a person was defined by her/his main source of income which was the main source of maintenance for this person.

3. Markov mobility matrices

Table 1 shows the structure of a surveyed panel by persons' main source of maintenance in 2007 and 2008, respectively. We see that among persons aged 19 years or more the dominating source of maintenance is hired work (40% of the total number of respondents in the sample) and old age or disability pension (one third of the sample). The rate of persons being dependent is about 10% of the 2007-2008 sample. Principally, they are pupils or students.

Socio-economic changes in the last years in Poland have created the necessity for people to change the main source of maintenance. Some of these changes are caused by retirement or entrance into the labour market. Others changes of occupation are unplanned and result from the climate on the labour market or random events, e.g. loss of work and transition to unemployment benefits or dependency on other members of the household, transition to a disability pension by a working person, etc. Very interesting is the analysis of a change of a form of employment from hired work to self-employment or vice versa. In the 2008 household survey a category "hired work" (permanent and occasional) was divided into two subcategories: work in the country of origin and work abroad.

Table 1. Structure of respondents by a source of income in a panel of individuals 2007-2008

Source of income	2007	2008
Employment		
hired permanently	38,3	40,3
hired occasionally	1,4	1,1
Farm	5,8	5,6
Self-employment	4,8	4,8
Pensions	32,0	33,0
Other social benefits	5,1	4,2
Other income	1,2	1,1
Dependents	11,5	9,9
Total	100,0	100,0

Source: Household Budget Surveys, 2007-2008, Panel of individuals, Poland, Central Statistical Office, Warsaw.

In 2008, 1.5% out of permanent workers worked abroad. In a much smaller group of part-time workers, one fourth of them worked abroad (Table 2).

Table 2. Mobility of respondents between different occupations (source of income) in a panel of individuals 2007-2008

2008 Source of income in 2007	Employment hired permanently at home country	Employment hired permanently abroad	Hired work occasionally at home country	Hired work occasionally abroad	Farm	Self-employment	Pensions	Other benefits	Other income	Dependents	Total
Employment hired perma-nently	10564	166	41	15	39	119	240	225	43	248	11700
Employment hired occasion-nally	172	12	131	7	14	10	12	9	6	49	423
Farm	101	9	17	5	1544	17	36	18	1	35	1782
Self-employ-ment	123	6	12	1	12	1240	27	8	4	22	1454
Pensions	130	1	4	1	22	14	9491	57	17	27	9763
Other benefits	289	4	26	1	14	12	185	810	26	180	1546
Other income	70	4	10	2	5	12	14	19	192	34	361
Depen-dents	640	16	68	6	59	51	59	126	54	2431	3510
Total	12090	218	309	37	1708	1474	10064	1272	342	3027	30540

Source: Household Budget Surveys, 2007-2008 Panel of individuals, Poland, Central Statistical Office, Warsaw.

Most of the respondents are maintained from hired work and old-age/disability pensions. For persons for whom hired work was the main source of income in 2007, this source of maintenance remains unchanged in 90% of cases in 2008.

Only 1% of the hired workers in 2007 started acting as self-employed in 2008, namely 129 persons out of 12,123 of the total number of hired workers in 2007 (Tables 2 and 3). Similar number of the self-employed moved to hired work in 2008, namely 142 persons out of 1,454 of the total number of self-employed in 2007. They constituted 10% of the total number of self-employed in 2008 (Tables 2 and 4).

Table 3. Mobility of respondents from hired work to other occupations in a panel of individuals 2007-2008 (in %)

2008 Source 2007	Hired work	Farm	Self-employ-ment	Pensions	Other benefits	Other income	Dependents	Total
Hired work	91,6	0,4	1,1	2,1	1,9	0,4	2,4	100,0

Source: Household Budget Surveys, 2007-2008 Panel of individuals, Poland, Central Statistical Office, Warsaw.

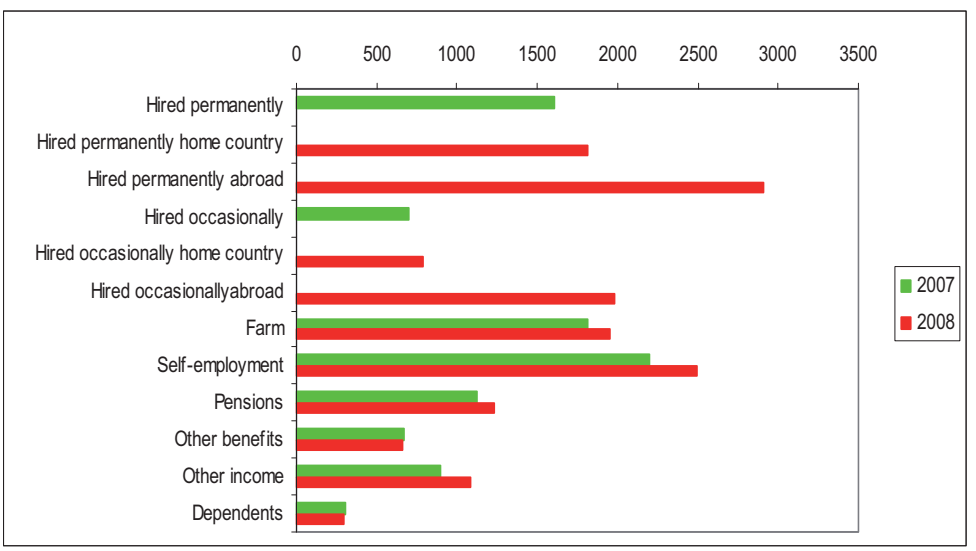
Table 4. Mobility of respondents from other occupations to hired work in a panel of individuals 2007-2008 (in %)

2007 Source 2008	Hired work	Farm	Self-employ-ment	Pensions	Other benefits	Other income	Dependents	Total
Hired work	87,8	1,0	1,1	1,1	2,5	0,7	5,8	100,0

Source: Household Budget Surveys, 2007-2008 Panel of individuals, Poland, Central Statistical Office, Warsaw.

Working abroad gives permanent workers 50% more income than income earned in their home country. Part-time workers earn more than twice as much abroad as at home. However, moving to another country is related to a higher risk of becoming unemployed and other risks related to unknown labour market institutions and rules, including language requirements (Figure 1).

Figure 1. Average monthly personal income by occupation in a panel of individuals 2007-2008 (in zlotys)



Source: Household Budget Surveys, 2007-2008 Panel of individuals, Poland, Central Statistical Office, Warsaw.

The highest increase in personal income was observed when respondents switched from hired work in 2007 to a self-employment activity in 2008. It says that a risky decision to start one's own business proved to be profitable, although only for a very small group of respondents (Table 5).

Young people in the age group of 30-39 are those mainly involved in switching from hired work to self-employment. Older respondents (50-54) more often benefit from the social security system than from self-employment. When hired employment is discontinued, due to either individual reasons or due to structural unemployment, it is more difficult to start one's own activity at an older age than at a young age (Tables 6 and 7).

Table 5. Increase in personal income by a change of occupation in a panel of individuals 2007-2008 (in zlotys)

	Increase in personal income					
	Source of income in 2008 (to)					
Source of income in 2007 (from)	Hired work	Farm	Self-employment	Pensions	Other benefits	Dependents
Hired work	290	-139	594	126	-481	-414
Farm	486	145	9	-527	-96	-352
Self-employment	270	762	255	-396	-491	-298
Pensions	695	627	1053	93	-238	-508
Other benefits	497	-38	552	355	19	-111
Dependents	794	521	594	657	237	71

Source: Household Budget Surveys, 2007-2008 Panel of individuals, Poland, Central Statistical Office, Warsaw.

Table 6. Mobility of respondents from hired work to other occupations and from other occupations to hired work, by age groups, in a panel of individuals 2007-2008

	Age group										
	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65+	Total
From hired work											
to self-employment	8	15	30	21	18	18	14	6	1		129
to other benefits	24	42	36	26	25	26	40	14	1		234
To hired work											
from self-employment	6	19	28	16	24	18	19	8	2		141
from pensions	26	12	5	13	2	14	18	25	17	4	135
from other benefits	47	58	45	38	38	27	41	23	2		319
dependents	268	142	77	56	61	49	48	18	4		724
Total	380	288	221	168	168	151	179	93	28	4	1683

Source: Household Budget Surveys, 2007-2008 Panel of individuals, Poland, Central Statistical Office, Warsaw.

When mobility from other occupations (types of economic activity) to hired work is concerned, the picture is more diversified. When available, hired work is attractive for young respondents at the age of 30-34 and 40-44, who tried a self-employment activity but decided to switch to hired work due to either lack of profits or a will to reduce risk – a disadvantage of such an activity.

There is also a relatively large group of respondents who became employed in 2008 being at the pre-retirement age (55-59 years) and whose main income in 2007 was a pension. In 2008, they earned mainly from wage employment. The pre-retirement and retirement age in Poland are relatively low (55/60 for women and 60/65 for men), and the effective retirement age is even lower by 2-3 years below the limits. Many people benefit from the pension system and work extra, either legally or illegally.

Table 7. Increase in personal income by mobility from hired work to other occupations and from other occupations to hired work, by age groups, in a panel of individuals 2007-2008 (in zlotys)

	Age group										
	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65+	Total
From hired work											
to self-employment	804	522	1240	1358	30	261	-468	36			594
to other benefits	-387	-364	-1305	-320	-402	-206	-299	-386			-481
To hired work											
from self-employment	269	381	359	406	49	236	189	250	558		270
from pensions	781	472	516	789	-109	585	934	327	1090	1038	695
from other benefits	552	518	444	490	480	353	492	702	19		497
dependents	651	938	877	709	585	1263	554				794
Total	444	330	208	491	164	300	195	279	847	1038	315

Source: Household Budget Surveys, 2007-2008 Panel of individuals, Poland, Central Statistical Office, Warsaw.

Legal employment of early retired persons is allowed in Poland up to some limits of earned income. The data show that for many respondents such a solution was profitable, and they earned more from wage employment than from a pension or they could postpone their pensions for some time.

Looking at the mobility by respondents' educational levels, we see that two groups of respondents: those with primary and with vocational education were prone to moving from hired work to self-employment. They took risks and earned mainly from their own business in 2008, as opposed to wage employment in 2007. Respondents with secondary education are not mobile (Table 8).

Table 8. Mobility of respondents from hired work to other occupations and from other occupations to hired work, by educational levels, in a panel of individuals 2007-2008 (% of persons working)

	Level of education in 2008				
	Primary	Vocational	Secondary	Tertiary	Total
From hired work					
to self-employment	1.3	1.1	0.	1.2	1.1
to other benefits	4.5	2.7	1.7	0.5	1.9
To hired work					
from self-employment	13.8	12.9	1.8	4.7	3.6
from pensions	0.4	1.7	2.3	2..0	1.4
from other benefits	12.9	19.9	21.9	49.5	20.6
from dependents	14.3	22.0	19.7	32.9	20.6

Source: Household Budget Surveys, 2007-2008 Panel of individuals, Poland, Central Statistical Office, Warsaw.

Mobility in the opposite direction, from other activities in 2007 to hired work in 2008, concerned also mainly the primary and vocational education groups. The group of persons with tertiary education is relatively smaller than other educational groups, but it is mobile at a comparable scale (Table 8).

The highest increase in personal income by mobility from hired work in 2007 to self-employment in 2008 was gained by young people (30-39 years of age). Persons at the age of 50-54 years suffered a loss in 2008 moving in the same direction. It seems that starting one's own business may be more risky for older than for younger persons (Table 9).

Table 9. Increase in personal income by mobility from hired work to other occupations and from other occupations to hired work, by educational levels, in a panel of individuals 2007-2008 (in zlotys)

	Level of education in 2008				
	Primary	Vocational	Secondary	Tertiary	Total
From hired work					
to self-employment	422	68	652	1291	594
to other benefits	-54	-329	-930	-326	-481
To hired work					
from self-employment	89	217	305	339	270
from pensions	553	657	703	904	695
from other benefits	384	397	569	678	497
dependents	681	941	641	904	794
Total	283	215	268	688	315

Source: Household Budget Surveys, 2007-2008 Panel of individuals, Poland, Central Statistical Office, Warsaw.

A move in the opposite direction to hired work seems to be the most beneficial for persons at the age of more than 60 or even more than 65, when they undertook the wage employment in 2008, having lived previously (in 2007) on a pension. Here, the experience and competence would matter.

However, the increase in income caused by mobility from wage employment to self-employment was the highest in a group of persons with tertiary education. It proves that highly educated persons start activities that pay the most. Respondents with secondary education moving from hired work to own business increased their income by the half of the increase of those with tertiary education. And those with vocational education earned only slightly by switching from wage to self-employment.

When shifting to hired work from other activities similar trends were observed with tertiary-educated respondents, earning the most when they switched from pension or from self-employment to wage employment. But the persons with secondary and vocational education also benefited relatively well by moving in the same direction (Table 9). Again, it proves that it is profitable to work even when one has the right to obtain a pension. And when one's own business is not profitable enough, it is better to move to wage employment than experience losses. Mobility matters. And mobility of the more educated matters more.

4. Regression analysis

For a calculation of a model we have chosen a panel of 3095 individuals who changed their occupation (and their main source of maintenance) between 2007 and 2008. We have omitted persons who became dependents because of insignificant number of persons in some subgroups by age and education.

In the linear regression model, a dependent variable *dincome* is an increase in net personal income of individuals who changed their occupation. Incomes are counted nominally in the same month of the survey in 2007 as well as in 2008. An increase in income is a difference of personal net income between 2007 and 2008. Independent variables consist of indicator (binary) variables describing:

- change of the main source of maintenance,
- respondents' education groups,
- respondents' age groups.

Apart from that, one continuous variable – *income,* enters the set of independent variables – a net personal income of individuals (members of the households) in 2007.

Variables *agr_hw, soc_hw*, *pen_hw, mt_hw, hw_sem*, *hw_pen*, *hw_soc*, *sem_agr*, *sem_hw* are equal to 1 when the respondent has changed her/his main source of maintenance:

agr_hw from work in a rural farm to hired work,
soc_hw from social security to hired work,

pen_hw from pension to hired work,
mt_hw from maintenance as dependent to hired work,
hw_ sem from hired work to self-employment,
hw_ pen from hired work to a pension,
hw_ soc from hired work to another type of social security, respectively.
sem_agr from self-employment to work in a rural farm,
sem_hw from self-employment to hired work.
Otherwise these above mentioned variables are equal to 0.

Variables *edu_ter*, *edu_sec*, *edu_voc* are equal to 1 when a person reports higher, secondary or basic vocational education, respectively. Primary education is a reference category. In this case, for a person with primary education, variables *edu_ter*, *edu_sec*, *edu_voc* are set to 0.

Variables concerning age are equal to 1, when (respectively)
age35_49, for age 35-49
age50_59, for age 50-59
age60plus, for age 60 and more

The age group 20-34 is a reference category.

There were 3095 observations taken into account in the estimated model. The determination coefficient is equal to $R^2 = 0.533$. The variable *edu_voc*, is not significant at the significance level $\alpha = 0.1$. The rest of explanatory variables are significant at the level $\alpha = 0.02$.

Table 10. Results of the regression of an increase in personal income on a change of occupation of individuals between 2007 and 2008

Variable	B	Std. Error	p-value
agr_hw	648.58	137.95	< 0,001
soc_hw	305.04	85.79	< 0,001
pen_hw	811.39	113.37	< 0,001
mt_hw	310.16	130,18	0.017
hw_sem	1241.28	124.20	< 0,001
hw_soc	-221.26	94.27	< 0,001
hw_pen	*439.96*	93.55	0.019
sem_agr	990.98	414.28	0.017
sem_hw	517.66	117.62	< 0,001
edu_ter	862.74	105.57	< 0,001
edu_sec	310.65	82.84	< 0,001
edu_voc	117.35	81.96	0,152
age35-49	279.71	70.12	< 0,001
age50_59	254.88	74.72	0.001
age60plus	838.75	106.30	< 0,001
income	-0,78	0,02	< 0,001
Const	267.86	91.59	0.003

Dependent variable: *dincome* is an increase in net personal income in a panel of 3095 persons who changed their occupation between 2007 and 2008.

Source: Household Budget Surveys, 2007-2008 Panel of individuals, Poland, Central Statistical Office, Warsaw.

We interpret regression coefficients as changes of binary variables of the increase in net personal income of given categories in comparison to a reference category.

Mobility in all directions, e.g. to different occupations, increases income except for quitting a job for the social security benefit which is lower than former wage. The highest increase in income is gained by those who changed the hired work for self-employment in industry or services. The second highest gain in income was observed by a move from self-employment to an activity in agriculture. In this case mobility concerns a move between similar types of activity.

A passage from pension to hired work and from an activity in agriculture to hired work as well as from self-employment in industry or services to hired work brought the relatively lower but still significant increases in earnings.

As far as education is concerned, respondents with the education level above primary education experienced a higher increase in income than those with primary education. The highest increase in income was observed by respondents with tertiary education. Persons in the age group of 60 and more achieved the biggest increase in income in comparison with the age group of 20-34 years. Persons who earned a higher income in 2007 experienced a relatively lower increase in income in 2008, so the regression coefficient is negative at the variable *income*.

5. Conclusions

In this paper we have analyzed how a change of occupation affects individual capability to earn income. The analysis concerns the mobility of individuals between broadly defined types of occupation such as: hired work, self-employment in industry, services or agriculture and social security. The change of occupation was examined in groups of individuals according to: age, level of education and a permanent or temporary character of work.

The Markov matrices were applied to the panel data on 30,540 individuals for 2007-2008 from the Polish Household Budget Surveys. Then the regression analysis was conducted for a group of 3095 persons who changed their occupation and their main source of maintenance between 2007 and 2008.

Our results show that in most cases a change of broadly defined occupation increases income, except for quitting a job for the social security benefit. The mobility matters mostly for those who changed hired work for self-employment in industry or services. A shift from self-employment to an activity in agriculture brought the second highest gain in personal income. Other changes of occupation,

e.g. from self-employment in agriculture, industry or services to hired work lead to relatively lower but still significant increases in earnings.

Respondents with tertiary education who were mobile in different directions reported the highest increase in income and those with secondary education benefited from changing occupation as well. The inter-occupational mobility matters mostly for persons in the age group of 60 and more. Persons at the age of 35-59 also noted a sizable increase in income in comparison with the age group of 20-34 years.

The increase in income in 2008 was relatively lower for persons who earned a higher income in 2007, which is partially a statistical effect of a higher income base in 2007. However, it shows that mobility between different occupations can be beneficial for individuals at all levels of initial income.

Acknowledgments

The research was conducted within the project COMPETE*, PL0104,* funded by the EEA Grants and Norway Grants (85%) as well as by the Ministry of Science and Higher Education of Poland (15%).

REFERENCES

ARROW KENNETH J., 1962, The Economic Implications of Learning by Doing, *The Review of Economic Studies*, Vol. 29, No. 3, 155-173.

BECKER GARY S., 1994, *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*, Third Edition, New York, National Bureau of Economic Research.

BEN PORATH, YORAM, 1967, The Production of Human Capital and the Life Cycle of Earnings, *Journal of Political Economy*, 75(4), part 1, 352-365.

FRIEDMAN MILTON, 1957, A Theory of the Consumption Function, Princeton: Princeton University Press.

HOUSEHOLD BUDGET SURVEYS, 2007-2008 Panel of individuals, Poland, Central Statistical Office, Warsaw.

LIBERDA BARBARA, 2007, Income Preferences and Household Savings, *Gospodarka Narodowa,* No 9, pp. 19-30.

MILANOVIC BRANKO, 2008, Where in the World Are You? Assessing the Importance of Circumstance and Effort in a World of Different Mean Country Incomes and (almost) No Migration, *Policy Research Working Paper Series* 4493, The World Bank.

MINCER JACOB, 1974, Schooling, Experience and Earnings, New York: National Bureau of Economic Research.

MODIGLIANI FRANCO, 1986, Life Cycle, Individual Thrift, and the Wealth of Nations, *American Economic Review,* Vol. 76, No. 3, pp. 297-313.

SPENCE ANDREW MICHAEL, 1973, Job Market Signaling, *Quarterly Journal of Economics,* 87 (3), pp. 355–374.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 207—222

NEONATAL MORTALITY BY GESTATIONAL AGE AND BIRTH WEIGHT IN ITALY, 1998-2003: A RECORD LINKAGE STUDY

Cristiano Marini[1], Alessandra Nuccitelli[2]

ABSTRACT

Neonatal mortality rates by gestational age and birth weight category are important indicators of maternal and child health and care quality. However, due to recent laws on administrative simplification and privacy, these specific rates have not been calculated in Italy since 1999. The main aim of this work is to assess the possibility of retrieving information on neonatal mortality by the linkage between records related to live births and records related to infant deaths within the first month of life, with reference to 2003 and 2004 birth cohorts. From a strict methodological point of view, some critical aspects of the most used record linkage approach are highlighted: specific problems may arise from the choice of records to be linked if there are consistency constraints between pairs (in this context, one death record can be linked to at most one birth record). In the light of considerations on the quality of the starting data, the retrieval of information on neonatal mortality by gestational age and birth weight is restricted to Northern Italy. Specific neonatal mortality rates are provided with reference to 2003 and discussed with particular emphasis on quality issues in the data collection processes.

Key words: administrative data, Boolean linear programming, data quality, greedy algorithm

1. Introduction

Neonatal mortality rates by gestational age and birth weight category are important indicators of maternal and child health (Zeitlin *et al.*, 2003) and care quality (Horbar, 1999)[3]. However, due to recent laws on administrative

[1] University of Rome "La Sapienza", Italy, e-mail: cristiano.marini@uniroma1.it.
[2] National Statistical Institute, Italy, e-mail: nuccitel@istat.it.
[3] A neonatal death is the loss of a live-born infant within the first month of life. The neonatal mortality rate is the number of deaths within the first month of life per 1,000 live births.

simplification and privacy, these specific rates have not been calculated in Italy since 1999.

The time-honoured system of birth registration managed by Istat, the National Statistical Institute, was dismantled in 1998 and later rebuilt entrusting it to the Ministry of Health. Istat has remained in charge of the register of deaths; the transfer of birth certificates from the Ministry to Istat has been only permitted after deletion of personal identifiers. Thus, the linkage of birth certificates - containing crucial information, such as birth weight and gestational age - to the corresponding infant death records (if any) has turned more difficult.

The main aim of this work is to assess the possibility of retrieving information on neonatal mortality by a linkage between records related to live births and records related to neonatal deaths, with reference to 2003 and 2004 birth cohorts.

The paper is organized as follows. Some quality issues regarding the data used for the linkage are discussed in the next section, with special emphasis on the coverage for the new medical birth register entrusted to the Ministry of Health. The coverage of neonatal deaths can be assumed to be total. An overview of record linkage problems is provided in section 3; some critical aspects of the most used approach to choose records to be linked, when there are consistency constraints between record pairs, are highlighted. The strategy adopted for the matching is described in section 4. Some results from a provisional linkage at national level are reported in section 5; in the light of such preliminary results, the retrieval of information on neonatal mortality is restricted to Northern Italy. In an attempt to give an idea of the trend of the phenomenon at least for this area, specific neonatal mortality rates are provided with reference to 2003, the first year from which the time series calculation can be resumed (section 6). Finally, some conclusions are drawn.

2. Data quality issues

2.1. Coverage of births

Data on births are collected through a medical register entrusted in 2002 to the Ministry of Health, rather than to Istat, as it had been up to 1998.

This change has caused some organizational problems, and in 2004 the coverage[4] for the new registration system was still about 86 percent at national level (Ministero della Salute, 2007). Over the period 2002-2004 the coverage showed a strong heterogeneity across regions, with a tendency to decrease from north to south.

[4] The coverage is measured by the percent ratio of number of birth certificates to number of deliveries registered by the hospital discharge system.

2.2. Quality of information used for the linkage

For the sake of simplicity, let *A* and *B* be two data files containing n_A and n_B records, respectively. Each record corresponds to an entity (or individual) and consists of several fields (or variables).

The aim of record linkage is to partition the set of all record pairs $A \times B = \{(a,b); a \in A, b \in B\}$ into two disjoint subsets: *M*, the set of matches, and *U*, the set of non-matches. Generally, a record linkage procedure classifies the record pairs on the basis of results from the comparison between corresponding fields (or matching variables) common to both files.

Matching variables may be subject to errors and omissions and usually have a different identifying power. For example, a field such as *sex* only has two value states and consequently could not impart enough information to identify the records related to the same individual. Conversely, a field such as *municipality of birth* imparts much more information, but it may frequently be reported incorrectly. Besides, some variables may change their values during the course of a person's lifetime, making more difficult to recognize records related to the same individual (*e.g. municipality of residence*).

Rather than considering all pairs $A \times B$, comparisons are sometimes restricted to those records that agree on some matching variables (blocking variables). Clearly, in this case the reliability and efficiency of a record linkage procedure is highly dependent upon the way in which blocking is carried out. Errors in blocking variables can result in a failure to compare records corresponding to the same entity.

The matching variables used in retrieving information on neonatal mortality are:

a. *infant's sex*;
b. *plurality of pregnancy*;
c. *infant's day of birth*;
d. *infant's month of birth*;
e. *infant's municipality of birth*;
f. *infant's province of birth*;
g. *mother's day of birth*;
h. *mother's month of birth*;
i. *mother's year of birth*;
j. *mother's municipality of residence*;
k. *mother's province of residence*;
l. *mother's nationality*.

In such a framework, the variables concerning mother's residence and nationality can be considered virtually unchanging over the lifetime of the infants of interest (at most, one month). The quality of the data to be linked can be

assessed in terms of incidence of invalid or missing values for each matching variable, with reference to 2003 and 2004 live birth cohorts[5] (Tables 1 and 2).

As displayed in Table 1, the accuracy of the neonatal death data is not high. The quality is poor for the matching variables concerning mother's date of birth and *plurality of pregnancy*. In particular, at national level unacceptable values for these variables increase respectively from 37 and 0 percent, for 2003 cohort, to 42 and 30 percent, for 2004 cohort. The incidence of invalid or missing values for Northern Italy is lower overall than for the whole country.

Table 1. Invalid or missing values (absolute and percent frequencies) in the neonatal death data for Italy and Northern Italy, by matching variable and infant's year of birth

matching variable	*invalid or missing values*							
	2003				*2004*			
	Italy		*Northern Italy*		*Italy*		*Northern Italy*	
	abs. fr.	*%*	*abs. fr.*	*%*	*abs. fr.*	*%*	*abs. fr.*	*%*
a	-	-	-	-	-	-	-	-
b	2	0.13	1	0.18	465	30.45	129	24.57
c	-	-	-	-	-	-	-	-
d	-	-	-	-	-	-	-	-
e	23	1.51	5	0.88	-	-	-	-
f	14	0.92	3	0.53	-	-	-	-
g	563	36.85	171	30.00	640	41.91	163	31.04
h	563	36.85	171	30.00	640	41.91	163	31.04
i	558	36.52	169	29.65	640	41.91	163	31.04
j	92	6.02	34	5.96	2	0.13	-	-
k	84	5.50	34	5.96	2	0.13	-	-
l	60	3.93	18	3.16	14	0.92	6	1.14

The birth data appear more accurate than the neonatal death ones (Table 2). In particular, with regard to *plurality of pregnancy* and *mother's nationality*, invalid or missing values decrease respectively from 12 and 3 percent, for 2003 cohort, to 1 and 2 percent, for 2004 cohort, while unacceptable values for the variables concerning the place of birth tend to disappear. However, the quality of *infant's month of birth* and of almost all other information on the infant's mother declines for the 2004 cohort. Compared to the whole country, for Northern Italy the infant data are more accurate, while the incidence of unacceptable values for the variables regarding the infant's mother is slightly higher.

[5] Other types of mistake in the data are not considered, since no subset of records for which the matching variables are observed without errors is available.

Table 2. Invalid or missing values (absolute and percent frequencies) in the birth data for Italy and Northern Italy, by matching variable and infant's year of birth

matching variable	*invalid or missing values*							
	2003				*2004*			
	Italy		*Northern Italy*		*Italy*		*Northern Italy*	
	abs. fr.	*%*	*abs. fr.*	*%*	*abs. fr.*	*%*	*abs. fr.*	*%*
a	598	0.13	263	0.12	530	0.11	143	0.06
b	55,138	12.17	39	0.02	4,447	0.94	324	0.14
c	558	0.12	153	0.07	313	0.07	147	0.07
d	738	0.16	157	0.07	12,464	2.62	147	0.07
e	1,019	0.22	36	0.02	44	0.01	21	0.01
f	1,022	0.23	39	0.02	44	0.01	20	0.01
g	2,466	0.54	1,450	0.66	2,663	0.56	1,997	0.89
h	2,743	0.61	1,484	0.67	2,892	0.61	2,015	0.90
i	3,157	0.70	1,629	0.74	6,602	1.39	2,177	0.97
j	4,925	1.09	1,586	0.72	7,077	1.49	3,565	1.59
k	6,203	1.37	2,227	1.01	8,128	1.71	4,040	1.80
l	14,710	3.25	2,812	1.27	8,002	1.69	4,790	2.14

3. Constrained record linkage problems

Specifying a record linkage procedure requires:

i. a measure of similarity between records;
ii. a decision rule using this measure for deciding when to classify a record pair as a link.

If k matching variables $X_1, X_2, \ldots, X_j, \ldots, X_k$ are compared for the record pair (a,b), the results can be expressed by a k-dimensional comparison vector representing the level of agreement between the records $a \in A$ and $b \in B$:

$$\mathbf{y}_{ab} = (y_{ab,1}, y_{ab,2}, \ldots, y_{ab,j}, \ldots, y_{ab,k}).$$

The easiest and most used way of defining the component j ($j = 1, 2, \ldots, k$) of $\mathbf{y}_{ab}$ is to assume that $y_{ab,j} = 1$, if a and b agree on the field j, and $y_{ab,j} = 0$, if a and b disagree on j or at least one value for j is missing.

Once the comparison vector is defined, the second step is to decide how $\mathbf{y}_{ab}$ can be used to classify the record pairs into M or U. According to the most usual approach, $\mathbf{y}_{ab}$ can be given a weight w_{ab} on which to base the decision rule for

(a,b) :

$$\begin{aligned} &\text{if } w_{ab} \geq t_u && (a,b) \text{ is classified as a link;} \\ &\text{if } t_l \leq w_{ab} < t_u && (a,b) \text{ is classified as a potential link and the final decision will be only taken after manual review} \\ &\text{if } w_{ab} < t_l && (a,b) \text{ is classified as a non-link.} \end{aligned} \tag{1}$$

The estimation of w_{ab}, $(a,b) \in A \times B$, and the selection of the threshold values t_l and t_u $(t_l \leq t_u)$ are crucial steps in defining a record linkage procedure. In a very straightforward way, weights and thresholds can be specified according to a deterministic criterion, with a view to the specific goals of the matching[6].

In the decision rule proposed by Fellegi e Sunter (1969) the weights are defined in a probabilistic way by means of the log-likelihood ratio:

$$w_{ab} = log\left(\frac{P(\mathbf{y}_{ab} | (a,b) \in M)}{P(\mathbf{y}_{ab} | (a,b) \in U)} \right) \qquad (a,b) \in A \times B \tag{2}$$

Most software for probabilistic record linkage uses the Expectation-Maximization (EM) algorithm to estimate (2), while t_l and t_u are usually selected in an empirical way, by inspection of the distribution of the estimated weights (Jaro, 1989; Winkler, 2000). Besides, some software enables to adjust such estimates for relative frequencies of specific values observed for the matching variables, in order to better take into account of their different identifying power[7].

In most applications, a one-to-many (or one-to-one) matching is required, depending on whether each entity is represented at most once in one file (or in each file) to be matched. When the decision rule is defined separately for each pair, as in (1), this requirement leads to one of the following methodological questions. First, there is the issue of how to modify the weight estimation method to incorporate the requirement. The second issue is how to solve the assignment problem arising in a one-to-many (or one-to-one) matching when a record in one file (or in each file) is included in more than one record pair classified as a link or a potential link by the decision rule.

---

[6] For instance, a very simple deterministic criterion may consist in setting the weights equal to the number of agreements on the matching variables and both t_l and t_u equal to $k-1$ (*i.e.* only the record pairs with at least $k-1$ agreements are classified as links).

[7] For example, with regard to *infant's municipality of birth*, an agreement observed on a small municipality (in terms of population) enables to identify the records related to the same individual more easily than an agreement observed on a larger municipality.

Since current implementations of the EM algorithm usually do not incorporate any explicit mechanism to force the matching requirement, additional procedures based on operational research techniques are needed to solve the assignment problem.

Relatively early software for record linkage introduced greedy algorithms to solve the assignment problem (Hill and Pring-Mill, 1985). With reference to the one-to-one case, a greedy algorithm usually works by ordering the record pairs by matching weight. At each step, the record pair with the highest weight is retained and included in the final assignment scheme. All other record pairs that conflict with this best pair, meaning that they include records that are also in the best pair, are discarded. This is repeated for all pairs whose weights are not less than t_l or t_u.

Some recent computer software uses the core of the algorithm proposed by Jaro (1989) to force the matching requirement. Specifically, the optimal assignment scheme for the one-to-one case can be obtained as the solution of the following Boolean linear programming problem[8]:

$$\begin{aligned} &\text{maximize} && \textstyle\sum_{a=1}^{n_A}\sum_{b=1}^{n_B} w_{ab} z_{ab} \\ &\text{subject to:} && \textstyle\sum_{a=1}^{n_A} z_{ab} \le 1 \\ &\quad b = 1,\ \ 2,\ \ \ldots,\ \ n_B && \\ &\qquad \textstyle\sum_{a=1}^{n_B} z_{ab} \le 1 && \quad a = 1,\ \ 2,\ \ \ldots,\ \ n_A\,, \end{aligned} \tag{3}$$

where $z_{ab} = 1$, if a is assigned to b, and $z_{ab} = 0$, if a is not assigned to b, for $(a,b) \in A \times B$. From now on, this approach will be also called global. According to Jaro, once the optimal assignment scheme is obtained, an assigned pair can be classified as a link if its weight is not less than t_u.

Armstrong e Saleh (2000) found the algorithm introduced by Jaro to produce incorrect links in practice. In order to give here a general explanation for that, consider the records $a, a' \in A$ and $b, b' \in B$ and all the possible record pairs (a,b), (a,b'), (a',b), (a',b') whose weights are related as follows:

$$\begin{aligned} &w_{ab} + w_{a'b'} > w_{ab'} + w_{a'b}\,, \\ &\quad w_{ab'} > w_{ab}\,, w_{a'b'}\,, w_{a'b}\,. \end{aligned}$$

[8] With reference to the one-to-many case (one record in *A* can be linked to several records in *B*, but one record in *B* can be linked to at most one record in *A*), the number of constraints decreases:

$$\begin{aligned} &\text{maximize} && \textstyle\sum_{a=1}^{n_A}\sum_{b=1}^{n_B} w_{ab} z_{ab} \\ &\text{subject to:} && \textstyle\sum_{a=1}^{n_A} z_{ab} \le 1 \qquad b = 1,\ \ 2,\ \ \ldots,\ \ n_B\,. \end{aligned}$$

Suppose that all the weights are not less than t_u and each entity can be represented at most once in each file to be matched. According to the global approach, (a, b') is not labelled as a link, despite its highest weight. Such a final decision may turn out to be unreliable, especially if a and b' agree perfectly on all matching variables.

It is worth noting that the assignment scheme obtained by maximizing the sum of weights depends upon the numerical values of the weights involved and not only on their order.

As a consequence, this scheme proves to be even less robust to possible biases in estimates of weights[9] than that resulting from a greedy strategy. Besides, the widespread use of logarithms in defining weights (2) can affect final results in a global approach.

This approach may perform well in record linkage applications characterized by very reliable matching variables with a high identifying power (*e.g. surname* or *address*), as in the case study reported by Jaro (1989). However, in other situations in which most of matching variables are subject to frequent errors or missing values and have a low identifying power, it is advisable to ensure the consistency between links by a more suitable approach to the problem.

4. The matching strategy

The possibility of retrieving information on neonatal mortality by gestational age and birth weight is explored by a deterministic linkage. The files to be linked consist of the following records:

file *A1*. 1,528 records related to neonatal deaths in Italy[10], with reference to 2003 birth cohort;

file *A2*. 1,527 records related to neonatal deaths in Italy, with reference to 2004 birth cohort;

file *B1*. 452,984 records related to live births in Italy in 2003;

file *B2*. 474,893 records related to live births in Italy in 2004.

---

[9] For instance, consider the records $a, a' \in A$ and $b, b' \in B$ and all the possible record pairs with the following weights: $w_{ab} = 15 > w_{ab'} = 12 > w_{a'b} = 10 > w_{a'b'} = 8$. Suppose that all the weights are not less than t_u and each entity can be represented at most once in each file to be matched. By adopting either a global or a greedy approach, both (a, b) and (a', b') are classified as links. If the weights are modified, for example, by setting $w_{ab} = 13 > w_{ab'} = 12 > w_{a'b} = 10 > w_{a'b'} = 8$, the global solution changes as well - (a, b') and (a', b) are now classified as links - while the greedy one is unchanged.

[10] The death records from the province of Bolzano are excluded, as the birth certificates for 2003 and 2004 are not available.

The matching is carried out between files related to the same birth cohort (*A1* and *B1*, *A2* and *B2*); therefore, *infant's year of birth* - which can be considered error-free - represents an implicit blocking variable. As only a low percentage of infants die within the first month of life, the size of the files to be linked is very different and the identification of records corresponding to the same individual is particularly difficult.

The matching weights are defined as a measure of similarity between records; this measure is based on the frequencies of the values observed for each matching variable in the larger file to be linked, which is related to live births. Let *A* and *B* represent two files related to the same birth cohort, *A1* and *B1* or *A2* and *B2*. The weight for (a,b) is:

$$w_{ab} = \frac{\sum_{j=1}^{k}[y_{ab,j} \times (n_B - fr_{md(x_{b,j}^B)})]}{k \times n_B - \sum_{j=1}^{k} fr_{md(x_{r,j}^B)}}, \tag{4}$$

where:

- n_B is the number of records in *B*;
- $y_{ab,j} = 1$, if *a* and *b* agree on *j*, and $y_{ab,j} = 0$, if *a* and *b* disagree on *j* or at least one value for *j* is missing;
- $fr_{md(x_{b,j}^B)}$ is the frequency (in *B*) of the value observed for *j* on the record $b \in B$;
- *r* is the record in *B* corresponding to the rarest perfect link, *i.e.* the pair whose records agree on all matching variables and for which $\sum_{j=1}^{k} fr_{md(x_{b,j}^B)}$ is minimum.

The weights (4) usually take values between 0 and 1. In particular, the pairs whose records disagree on all matching variables have a weight equal to 0; the maximum weight is usually attained at the rarest perfect link. On the basis of some tests on data from Italian regions characterized by different levels of coverage of births, such weights lead to the same final assignment scheme attainable by adopting the weights (2).

The threshold values t_l and t_u are selected by inspection of the distribution of the estimated weights. In this context, t_l and t_u are set to take values in the ranges 0.45-0.55 and 0.55-0.65, respectively (with $t_u - t_l \leq 0.1$).

As one death record can be linked to at most one birth record, a greedy algorithm is used to obtain the final assignment scheme. In case of one-to-many assignments[11] with the same matching weight, only the pair related to the infant with the lowest birth weight is labelled as a link, exploiting the findings from recent studies on neonatal mortality (Branum and Schoendorf, 2003).

[11] These pairs are related to cases in which one of the infants born from the same pregnancy died.

5. Results of linkage

5.1. Results from provisional linkage at national level

In this subsection the results from a provisional linkage of national data are reported. The weights are set equal to the number of agreements on the matching variables and $t_l = t_u = k$, *i.e.* only the record pairs with k agreements are classified as links. In case of one-to-many assignments with the same matching weight, only the pair related to the infant with the lowest birth weight is labelled as a link.

As displayed in Table 3, this provisional linkage does not produce a high number of links. The ratio of number of links to number of deaths is 29.2 and 23.6 percent for 2003 and 2004 cohorts, respectively. The largest decrease in the percent ratio can be observed for Central Italy (from 25.0 to 10.3 percent). With regard to Southern Italy and Islands, about one out of five death records is matched. Northern Italy shows the highest percentage of links, in the face of a decrease from 42.5 to 37.5 over the two-year period. The links coming from one-to-many assignments with equal matching weight - which are related to infants born from the same pregnancy - are 22 and 20 for 2003 and 2004 cohorts, respectively (for the sake of brevity, these data are not reported in Table 3).

The links achieved by perfect agreement on the matching variables are not many; so, a more sophisticated strategy for linkage is required. Besides, in the light of the considerations on coverage of births and quality of data used for the linkage (section 2), the retrieval of information on neonatal mortality by gestational age and birth weight is restricted to Northern Italy.

Table 3. Results from provisional linkage between records related to neonatal deaths and records related to live births, by geographical area of death and infant's year of birth

geographical area of death	*number of deaths*		*number of links*		*n. of links / n. of deaths (%)*	
	2003	*2004*	*2003*	*2004*	*2003*	*2004*
Northern Italy	570	525	242	197	42.5	37.5
Central Italy	288	302	72	31	25.0	10.3
Southern Italy and Islands	670	700	135	135	20.1	19.3
Italy	1,528	1,527	449	363	29.2	23.6

5.2. Results from linkage for Northern Italy

With regard to Northern Italy, the proposed strategy for linkage leads to a high number of final links (Table 4). The links coming from one-to-many assignments with equal matching weight are related to infants of the same gender, born from the same pregnancy (for the sake of brevity, such data are not reported in Table 4), and are about 4.8 and 5.0 percent of the total links for 2003 and 2004 cohorts, respectively. Besides, 91.4 and 95.7 percent of the links (for 2003 and 2004 cohorts, respectively) are labelled automatically, as their weights are above t_u, while 8.6 and 4.3 percent are designated manually, as their weights are between t_l and t_u.

Table 4. Results from final linkage between records related to neonatal deaths and records related to live births, by infant's year of birth - Northern Italy

infant's year of birth	*n. of deaths*	*n. of links*	*n. of links with weight* $\geq t_u$	*n. of links with weight between* t_l *and* t_u	*n. of links / n. of deaths (%)*
2003	570	526	481	45	92.3
2004	525	460	440	20	87.6

The high number of links that agree on all or most of the matching variables - at least 8 agreements occur for over 96 percent of links (Table 5) - does not rule out, however, the presence of errors in the final results (false matches or missing matches), mainly due to some critical issues concerning the data: the coverage of births is heterogeneous across regions; the matching variables do not have a great identifying power; the information on mother's date of birth in the death records is not very accurate. On the other hand, matching errors cannot be easily estimated, as no subset of pairs for which the linkage status is known is available; besides, model-based estimates of matching errors, according to other approaches proposed in literature (*e.g.* Armstrong and Mayda, 1993), could result inaccurate, due to the above mentioned critical aspects.

Table 5. Percentage distribution of links by number of agreements on the matching variables and infant's year of birth - Northern Italy

number of agreements	*links (%)*	
	2003	*2004*
12	46.01	42.83
11	12.17	18.26
10	8.17	6.30
9	19.96	15.22
8	9.70	14.57
7	2.47	2.61
below 7	1.52	0.22
total	100.00	100.00

6. Specific neonatal mortality rates

Neonatal mortality rates by gestational age and birth weight category have not been calculated in Italy since 1999. On the basis of the results from record linkage, estimates of these specific rates are provided for Northern Italy with reference to 2003, the first year from which the calculation of the time series can be resumed and for which the quality of the starting data appears to be slightly higher.

The neonatal mortality rates presented here, related to 2003 birth cohort, are calculated by using:

live births in Northern Italy by gestational age and birth weight category, at the denominator;

deaths within the first month of life, in Northern Italy, by gestational age and birth weight category, at the numerator[12].

The 570 neonatal deaths are assumed to have the same distribution by gestational age and birth weight category as the 481 links with weight not less than t_u [13].

The results in Figures 1 and 2 appear to be in line with those reported in recent studies on the survival of preterm and low birth weight infants (Demissie *et al*., 2001; Horbar *et al*., 2002) and with the progressive downward trend in neonatal mortality observed in the nineties - especially for preterm and low birth weight infants. Compared to 1998, in 2003 the neonatal mortality rate declined:

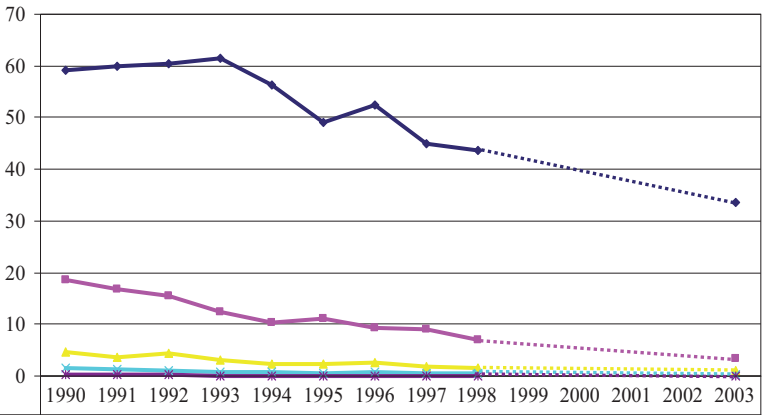
from 43.7 to 33.6 and from 36.7 to 30.8, respectively for the lowest categories of gestational age (below 28 weeks) and birth weight (below 1,000 grams);

by half, for gestations between 28 and 31 weeks (from 7.0 to 3.4) and for infants with birth weight between 1,000 and 1,499 grams (from 6.1 to 3.1).

[12] For 2003, neonatal mortality rates are calculated by cohort, while in the past (up to 1998) they were cross-sectional. However, if neonatal deaths are assumed to have a uniform distribution over the year, the differences between these types of rate can be negligible.

[13] Of the total 526 links, only those with weight not less than t_u (column 4 of table 4) are considered, in order to contain the introduction of bias due to possible false matches.

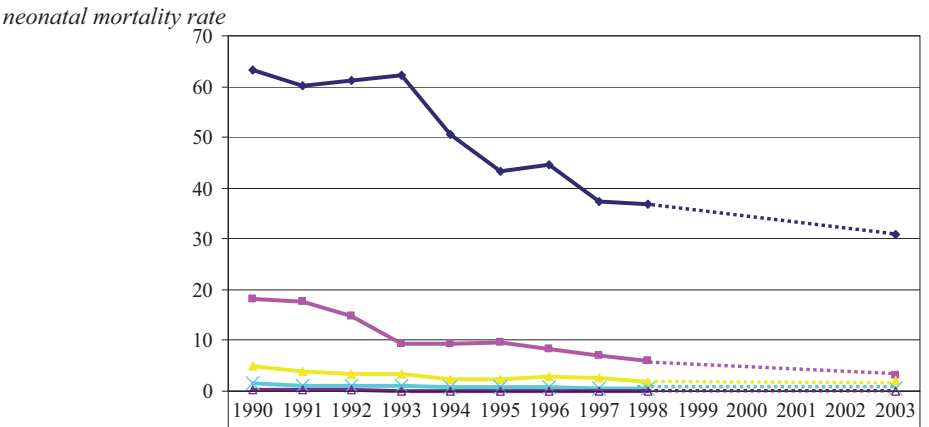
Figure 1. Neonatal mortality rates by gestational age category and year - Northern Italy, years 1990-98 and 2003 (values per 100 live births in the same gestational age category)



birth weight category (in grams)	1990	1991	1992	1993	1994	1995	1996	1997	1998	2003
<28	59.1	59.8	60.4	61.5	56.3	49.1	52.3	44.9	43.7	33.6
28-31	18.6	16.7	15.5	12.4	10.3	11.2	9.3	9.0	7.0	3.4
32-34	4.8	3.7	4.3	3.2	2.4	2.3	2.5	1.8	1.6	1.1
35-36	1.5	1.2	1.0	0.8	0.8	0.6	0.8	0.6	0.5	0.3
>=37	0.2	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Besides, compared to the past, in 2003 there was a higher concentration of neonatal deaths in the lowest categories of gestational age and birth weight: 42 and 55 percent of neonatal deaths were related to gestations below 28 and 32 weeks, while 40 and 48 percent concerned infants with birth weight below 1,000 and 1,500 grams.

Figure 2. Neonatal mortality rates by birth weight category and year - Northern Italy, years 1990-98 and 2003 (values per 100 live births in the same birth weight category)



birth weight category (in grams)	1990	1991	1992	1993	1994	1995	1996	1997	1998	2003
<1000	63.1	60.2	61.3	62.3	50.5	43.3	44.7	37.2	36.7	30.8
1000-1499	18.2	17.6	14.7	9.4	9.3	9.7	8.3	7.0	6.1	3.1
1500-1999	4.8	3.8	3.3	3.5	2.3	2.4	2.8	2.6	1.8	2.2
2000-2499	1.6	1.2	1.1	1.1	0.8	0.7	0.9	0.6	0.6	0.4
>=2500	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

7. Concluding remarks

In the light of the considerations on coverage of births and quality of data used for the linkage, the information on neonatal mortality by gestational age and birth weight is only retrieved for Northern Italy. The percentage of cases for which the retrieval is achieved by perfect agreement on the matching variables is low; so, a more sophisticated strategy for linkage is used. In this work, the possibility of retrieving crucial information on neonatal mortality is explored by using a deterministic record linkage. The adopted weights are based on the frequencies of the values observed for each matching variable in the larger file to be linked, which is related to live births. On the basis of some tests on data from Italian regions characterized by different levels of coverage of births, this choice - even though not very sophisticated - leads to the same links attainable by adopting a probabilistic strategy.

In this framework, one death record can be linked to at most one birth record. This requirement is the starting point for a critical review of the most used approach to choose records to be linked, when the decision rule is defined separately for each pair. Maximizing the sum of matching weights, incorrect links can be produced, as the solution depends upon the numerical values of the

weights involved and not only on their order. A greedy approach is believed to be more suitable, apart from the specific context.

However, the high number of final links does not rule out the presence of errors in the results and, as a consequence, the mentioned findings for neonatal mortality should be treated with caution. The reliability of neonatal mortality estimates depends on the accuracy and completeness of reporting and recording of births and deaths. In the retrieval of the information needed to compute specific rates, further efforts should be made to control matching errors, and this cannot be accomplished without improving the data collection processes over the whole country.

REFERENCES

ARMSTRONG J. B., MAYDA J. E. (1993), Model-based estimation of record linkage error rates, *Survey Methodology*, 19, 137-147.

ARMSTRONG J., SALEH M. (2000), Weight estimation for large scale record linkage applications, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1-10.

BRANUM A. M., SCHOENDORF K. C. (2003), The effect of birth weight discordance on twin neonatal mortality, *Obstetrics & Gynecology*, 101, 570-574.

DEMISSIE K., RHOADS G. G., ANANTH C. V., ALEXANDER G. R., KRAMER M. S., KOGAN M. D., JOSEPH K. S. (2001), Trends in preterm birth and neonatal mortality among blacks and whites in the United States from 1989 to 1997, *American Journal of Epidemiology*, 154, 307-315.

DEMPSTER A. P., LAIRD N. M., RUBIN D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, B, 39, 1-38.

FELLEGI I. P., SUNTER A. B. (1969), A theory for record linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

HERZOG T. N., SCHEUREN F. J., WINKLER W. E. (2007), *Data quality and record linkage techniques*, Springer, New York.

HILL T., PRING-MILL F. (1985), Generalized iterative record linkage system, in: KILSS. B., ALVEY W. (eds.), *Record linkage techniques - 1985, Proceedings of the Workshop in Exact Matching Methodologies*, Arlington, Virginia, May 9-10, 1985, 327-333.

HORBAR J. D. (1999), The Vermont Oxford network: evidence-based quality improvement for neonatology, *Pediatrics*, 103, 350-359.

HORBAR J. D., BADGERS G. J., CARPENTER J. H., FANAROFF A. A., KILPATRICK S., LACORTE M., PHIBBS R., SOLL R. F. (2002), Trends in mortality and morbidity for very low birth weight infants 1991-1999, *Pediatrics*, 110, 143-151.

JARO M. A. (1989), Advances in record-linkage methodology as applied to matching the Census of Tampa, Florida, *Journal of the American Statistical Association*, 84, 414-420.

MINISTERO DELLA SALUTE (2007), *Certificato di assistenza al parto (CeDAP). Analisi dell'evento nascita - Anno 2004*, Dipartimento della Qualità, Direzione Generale del Sistema Informativo, Ufficio di Direzione Statistica.

WINKLER W. E. (2000), Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage, *Statistical Research Report Series* RR2000/05, U. S. Bureau of the Census, Washington.

ZEITLIN J., WILDMAN K., BRÉART G., ALEXANDER S., BARROS H., BLONDEL B., BUITENDIJK S., GISSLER M., MACFARLANE A. (2003), Selecting an indicator set for monitoring and evaluating perinatal health in Europe: criteria, methods and results from the PERISTAT project, *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 111, 5-14.

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 223—230

BOOK REVIEW

Sampling: Design and Analysis. Sharon L. Lohr. 2nd Edition, International Publication, CB, ©2010, 608 Pages ISBN-10: 0495105279; ISBN-13: 9780495105275

Professor Sharon L. Lohr has already published a number of papers and books devoted to sampling surveys. I have had an opportunity to study some of them, and especially the first edition of her book on *Sampling: Design and Analysis*, published in 1999.

Now I have studied the 2nd edition of the book *Sampling: Design and Analysis*, published in 2010. The book covers many topics not found in other textbooks at this level. The book provides a modern introduction to the field of survey sampling intended for a wide audience of statistics students. The author concentrates on the statistical aspects of taking and analyzing a sample, incorporating a multitude of applications from a variety of disciplines. The text gives guidance on how to tell when a sample is valid or not, and how to design and analyze many different forms of sample surveys. Recent research on theoretical and applied aspects of sampling is included, as well as optional technology instructions for using statistical software with survey data.

Six main features distinguish this book from other texts about sampling methods.

1) The book is *accessible to students with a wide range of statistical backgrounds*, and is flexible for content and level. By approximate choice of sections, this book can be used for a first-year graduate course of statistics students or for class with students for business, sociology, psychology or biology who wants to learn about designing and analyzing data from sample surveys. It is also *useful for a person doing research who wants to learn more about statistical aspects of surveys and recent developments*.

2) The author used *real data as much as possible*. The examples and exercises come from *social sciences, engineering, agriculture, ecology, medicine* and a variety of other disciplines, and are selected *to illustrate the wide applicability of sampling methods*. A number of data sets have extra variables not specifically referred to the text; an instructor can use these for additional exercises or variations.

3) The author has incorporated *model-based* as well *as randomization theory* into the text, with the goal of placing sampling methods within the framework used in other areas of statistics. *Many of the important results in the last years of sampling research* have involved models, and *an understanding of both approaches is essential for the survey practitioners.*

4) As I stressed above, *the book covers many topics not found in other textbooks at this level.* Chapters 7 through 15 discuss how to *analyze complex surveys* such as those conducted by different countries[1], computer-intensive methods for estimating variances in complex surveys, *what to do if there is nonresponse*, and how to perform chi-squared tests and regression analyses using data from complex surveys.

5) This book emphasized *the importance of graphing the data*. Graphical analysis of survey data is challenging because of the large sizes and complexity of survey datasets but graphs can provide insight into the data structure.

6) Design of surveys is emphasized throughout, and is related to methods for analyzing the data from a survey. The book *presents the philosophy that the design* is by far the most important aspect of any survey: *no amount of statistical analysis can be compensated for a badly designed survey.* Models are used to motivate designs, and graphs are presented to check the sensitivity of the design to model assumptions.

Chapters 1 through 6 cover the building blocks of simple random, stratified, and cluster sampling, as well as ratio and regression estimation. To *read them requires* familiarity with *basic ideas of expectation, sampling distributions, confidence intervals, and linear regression* – material covered in most introductory statistics classes. Optional sections on the statistical theory for designs are marked with asterisks – these require to be familiar with calculus and mathematical statistics.

Chapter 7 is devoted to ***complex surveys*** and deals with assembling *design components, sampling weights, estimating a distribution function, plotting data from a complex survey, design effects*. Chapter 8 deals with ***nonresponse*** and begins with effects of ignoring nonresponse; *designing surveys to reduce nonsampling errors*; callbacks and two-phase sampling; *mechanisms for nonresponse*; *weighting methods for nonresponse*; *imputation;* parametric models for nonresponse; *what is an acceptable response rate*.

[1] Comp. e.g.: J. Kordos, A. Zięba: Development of Standard Error Estimation Methods in Complex Household Sample Surveys in Poland, *Statistics in Transition - new series, Vol. 11, No 2, , 2010, pp. 231-252.*

The material *in Chapter 9 through 15 can* be covered in almost any order, *with topics chosen to fit the needs of the students*. Appendix *A* reviews probability concepts.

The second edition introduces some organizational changes to the chapters. The central concept of sampling weights is now introduced in Chapter 2. Stratified sampling has been moved earlier to Chapter 3, preceding ratio and regression estimation. This allows students to become more familiar with the use of weights to account for inclusion probabilities before they are exposed to adjusting weight for calibration. Chapter 9 – ***variance estimation in complex surveys*** – has expanded treatment of computer-intensive methods such as *linearization (Taylor series) methods, jackknife and bootstrap*; generalized variance functions, and confidence intervals.

Material in Chapter 12 of the first edition has been expanded in Chapters 12 to 14 of the second edition. Chapter 15 – ***Quality survey*** – on total survey design is *completely new,* and ties together much of the material in the earlier chapters. I will devote a special attention to this chapter later.

Each chapter now concludes with *a chapter summary*, including key terms and references for further exploration.

The exercises in the second edition *have been reordered into four categories* in each chapter, with many new exercises added to the book's already extensive problem sets.

- *Introductory Exercises* give more routine problems intended to develop skills at the basic ideas in the book. Many of these would be suitable for hand calculations.
- *Working with Survey Data* exercises ask students to analyze data from real surveys.
- *Working with Theory* exercises are intended for a more mathematically oriented class, allowing students to work though proofs of results in a step-by-step manner and explore the theory of sampling in more depth.

As I have mentioned above, my attention was particularly drawn to chapter 15 on *Survey quality.* As the author has stressed *in this chapter*, "In many surveys, *the margin of error reported* is *based entirely on the sampling error*; *nonsampling errors* are sometimes acknowledged in the text, but *generally are not included in the reported measures of uncertainty"*. This is common practice in many countries and the author here concentrates on *survey quality.* First, the author quotes Dalenius (1977, p. 21) who referred to the practice of reporting only sampling error and ignoring other sources of error as "*'strain at a gnat and swallow a camel'; this characterization applies especially to the practice with respect to the accuracy: the sampling error plays the role of the gnat; sometimes*

malformed, while the nonsampling error plays the role of the camel, often of unknown size and always of unwieldy shape".

In this chapter, the author explores approaches to survey design and analysis taking into account ***total survey error*** which is the sum of five components*: coverage error, nonresponse error, measurement error, processing error and sampling error*. Such approach to *total survey error* does not exist in any other book on survey sampling. Sampling error produces variability in the estimates, and measures only precision of the estimates. It means that we may have very precise and not accurate estimates.

Additionally, I would like to cite Eurostat publication on data quality assessment: methods and tools[2].This publication underlines that "*production of high quality statistics depends on the assessment of data quality. Without a systematic assessment of data quality, the statistical office will risk to lose control of the various statistical processes such as data collection, editing or weighting. Doing without data quality assessment would result in assuming that the processes cannot be further improved and that problems will always be detected without systematic analysis. At the same time, data quality assessment is a precondition for informing the users about the possible uses of the data, or which results could be published with or without a warning. Certainly, without good approaches for data quality assessment statistical institutes are working in the blind and can make no justified claim of being professional and of delivering quality in the first place. Assessing data quality is therefore one of the core aspects of a statistical institute's work".*

For details see contents.

Contents

[2] Eurostat (2007), *Handbook on Data Quality Assessment: Methods and Tools*.

Appendices: Probability Concepts Used in Sampling.
Probability.
Random Variables and Expected Value.
Conditional Probability.
Conditional Expectation.

REFERENCES

Summing up, I would like to stress that the book is very comprehensive, covering both fundamental sampling theory and analysis, and real survey examples. The book also covers many of the current survey sampling topics that are not usually covered in standard sampling books, such as sampling weights, nonresponse, complex data analysis, and survey quality.

Information for the Polish statisticians: this book is available at the Central Statistical Library of the Central Statistical Office of Poland.

Prepared by: Jan Kordos, Warsaw School of Economics, e-mail: jan1kor2@aster.pl

STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 231—236

ISI Satellite Conference on Improving Statistical Systems Worldwide - Building Capacity

Being one of the oldest scientific associations running in the contemporary world, International Statistical Institute hosts its biennial ISI World Statistics Congresses (WSC) since 1853. Each time the congress is organized in a different country. In 1929 and 1975 Central Statistical Office of Poland had the honour to host WSC in Warsaw. Forthcoming WSC will be the 58th one, organized in Ireland, Dublin from the 21-26, August 2011.Each WSC provides an excellent opportunity for participants and observers to share ideas, develop plans, and experience a new vision of a future statistics. As well as standard activities of the WSC, there are Satellite Meetings organized all around the world and taking place before and after the ISI World Statistics Congresses.

ISI Satellite Conference on Improving Statistical Systems Worldwide - Building Capacity will be held in Krakow, Poland from 18 to 19 August 2011. The huge world of statistics holds remarkable potential for in depth discussions that highlight the new trends and provide an opportunity to anticipate the future developments and challenges that Statistical Systems will have to face.The aim of the Satellite Conference is to focus on the recent changes in policy regarding the work on improving statistics in developing countries. The new trend is to advocate more broad changes in national statistical system, trying to address a range of shortcomings which are interlinked and therefore often cripple national statistical systems. Taking place at the institutional, organizational and management levels as well as in the financing, the process of change should include not only the national statistical office but also the statistical units in the line ministries. Whether it's the development and strengthening the capacity of statistical systems, financing statistical development, cooperation in public statistics among transition and post- transition countries, in- country partnerships for statistical development or technical assistance – our conference has it in one place. It will address these real dilemmas and look at some of the possible solutions, and will go beyond the simple exchange of ideas and information.

The conference will be organised in seven plenary sessions, i.e:

Session 1. Statistical development strategies and national development plans.

National development plans demand the improvement of the ability of countries, agencies, and institutions to manage the developmental process in order to achieve for results. This process has equally emphasized the importance of better statistical data for highlighting issues, making policy choices, allocating

resources, monitoring outcomes and evaluating impact. The Paris Declaration on Aid Effectiveness, endorsed on March 2, 2005 to harmonize, align, and manage their aid programs to achieve measurable results using a set of measurable actions and indicators of progress. This is how a focus was put on statistics and statistical development plans.

Session 2. In-country partnerships for statistical development

The development of a national partnership for statistics is an essential aspect of the development of the national statistical system. Main producers, main users and policy makers should come together to agree on the direction of the national strategy and what mechanisms will be used to implement the strategy. In order to promote improved in-country partnerships for development the partners need to rely on government and partners working together to provide coordinated support to implement national statistical strategies.

Session 3.Technical assistance, training and tools.

Technical assistance is about the design of the most effective way to address the needs of the specific countries. In general there are two approaches. The first one is the country by country approach. The other approach is called the regional approach. In that case a package of assistance is delivered to a group of countries. In 1999 the UN Economic and Social council adapted a document called: *Some guiding principles for good practices in technical assistance for statistics.* These guidelines mention that regional approaches should be considered where appropriate. This aspect is also relevant for the south-south cooperation topic.

Session 4. North - south and south -south cooperation.

Countries have decided to make more efforts to support the development in the lesser developed countries and the least developed countries. This has led to new approaches, like the system-wide approach, and new tools, like the Virtual Statistical System. These new approaches also demand new working relations between the partners. and aim at empowering the partners countries to take more ownership of these activities and to be strongly involved in all phases of this work, from design till implementation.

Session 5. Financing statistical development / donor cooperation and coordination

While countries have not yet been able to mobilize the resources needed to implement *National Strategies for the Development of Statistics* for any number of reasons, a key requirement is new finance. For many low-income countries new finance inevitably means donor finance because countries are unable to allocate either the recurrent or development resources needed from their national budgets. Official statistics are a typical public good and their production and dissemination is therefore generally financed from tax revenue.

Session 6. Cooperation in public statistics among transition and post-transition countries.

Due to globalization processes the economy becomes a blurred system in many cases, with supranational and regional labour, production, trade and services markets being created. On the one hand, one can observe integration

processes, i.e. common market, free trade zone, duty-free and non-visa traffic, on the other, disintegration ones are visible, for instance tightening of rules while crossing the border. Integration (currently prevailing) and disintegration processes cause the need for information on cross-border areas to grow. It is mainly because of the changes of the function of borders and increased dynamics of processes in cross-border areas. Under conditions of dynamic socio-economic changes, public statistics could act as one of the most important factors in the integration process, with special focus on institutional transformations.

Session 7. Agriculture Statistics

During this session the progress of the work on the Global Strategy to Improve Agriculture Statistics and Rural will be discussed, especially issues of agricultural policy and quality statistics.

This meeting is unique in bringing together professionals from across the disciplinary spectrum - statisticians, researchers, academics, practitioners and related experts from various institutions to share their experience, expertise and challenges. This conference will provide a global forum for national, regional and international experts to appreciate the current state of Statistical System. The Scientific Programme Committee has created a schedule that promises to stick to the cutting edge of our field. It is a place to gain insight from world leaders and an opportunity to share best practices by meeting colleagues from around the globe.

Information on the Congress of Polish Statistics to Celebrate the 100th Anniversary of The Polish Statistical Association

The 100th anniversary of the Polish Statistical Association (PSA will be celebrated in 2012. The Association was created in Cracow 1912 to integrate specialists involved in public statistics, including members of the academic community, decision-makers and practitioners at the government and local administration agencies, and all professionals interested in statistical theory and research. The Association contributes to the advancement of theoretical, methodological and practical aspects of statistical research, to promotion of statistical information use, and to popularization of statistical knowledge in society. It maintains cooperation with statistical associations in several countries, and with such organizations as Bernoulli Society for Mathematical Statistics and Probability, International Society for Quality of Life Research, International Society for Quality-of-Life Studies or International Federation of Classification Societies. The PAS is an affiliated member of the International Statistical Institute ISI)..

Currently the Polish Statistical Association has about 750 members organized in 17 regional branches. The PSA publishes jointly with the Central Statistical Office the monthly journal „Statistical News" („Wiadomości Statystyczne") and an international journal „Statistics in Transition – New series".

To celebrate the one hundredth anniversary of The Polish Statistical Association the Congress of Polish Statistics will be held on 18 – 20 April 2012 in Poznan, combining this event with the celebration of the Polish Statistics Day in 2012.

The preliminary programme of the Congress comprises of a number of thematic sessions, including the anniversary (historical) session, as well as other ones devoted to the methodology of statistical research, regional statistics,

population statistics, socio-economic statistics, the problems of statistical data and the statistics of health, sport and tourism. The Congress will also host two panel discussions on: fundamental problems of statistics in the modern world, and the future of statistics

PRELIMINARY PROGRAMME:

- **ANNIVERSARY SESSION**
- **DEVELOPMENT OF POLISH STATISTICAL RESEARCH**
- **EMINENT POLISH STATISTICIANS**
  - Jerzy Spława Neyman
  - Jan Czekanowski
- **METHODOLOGY OF STATISTICAL RESEARCH**
  - Development of Statistical Research Methodology
  - Small Area Statistics
  - Methods of Data Analysis and Classification
  - Survey Methodology
- **REGIONAL STATISTICS**
- **POPULATION STATISTICS**
  - New demography of Europe and Poland: challenges for demographic analysis
  - Demographic modelling and projecting: factors affecting the demographic reality and future
  - Changes in demographic structures: facts and consequences for economy and society
  - New standards in population statistics (incl. census, panel studies)
  - Historical demography
- **SOCIAL STATISTICS**
  - Poverty, social exclusion and inequality
  - Standard and quality of life
  - Human capital
  - Labour market
- **ECONOMIC STATISTICS**
- **STATISTICAL DATA**
- **HEALTH STATISTICS**
- **STATISTICS OF SPORTS AND TOURISM**
- **DISCUSSION PANELS**
  - Fundamental problems of statistics in the modern world
  - The future of statistics

Papers presented during the Congress (both those delivered and presented during the poster session), on receiving a positive review, will be published in special issues of the following journals: „Biblioteka Wiadomości Statystycznych", „Ruch Prawniczy, Ekonomiczny i Socjologiczny", „Statistics in

Transition", „Studia Demograficzne", „Wiadomości Statystyczne" and Zeszyty Naukowe Uniwersytetu Ekonomicznego w Poznaniu.

The Organizing Committee would like to invite everyone to join in the unique celebration of Polish statistics and take part in the Congress of Polish Statistics. Updated information about the conference are published on the Congress website: http://www.stat.gov.pl/pts/kongres2012/english/index.htm

We welcome paper proposals on various aspects of statistical research, both theoretical and applied. Paper abstracts should be attached with the participation form and submitted on-line or sent by email at the address of the Congress secretariat by 10.10.2011.

Chairperson of the Organizing Committee of the Congress:

Dr. Elżbieta Gołata, Univ. Prof of UEP:
e-mail: elzbieta.golata@ue.poznan.pl

Secretariat of the Organizing Committee of the Congress:

Statistical Office in Poznan,
ul. J. H. Dąbrowskiego 79, 60-959 Poznań:
e-mail: kongres2012@stat.gov.pl
tel. + 48 61 27 98 325 (Mo-Fr: 8-15),
+48 61 27 98 343 (Mo: 9-12, Tu: 7-10,We: 7-10),
fax. +48 61 27 98 101