STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 14, Number 1, Spring 2013

Editorial Office **ISSN 1234-7655**
Marek Cierpiał-Wolan, Ph.D.: Scientific Secretary
m.wolan@stat.gov.pl
Beata Witek: Secretary
b.witek@stat.gov.pl. Phone number 00 48 22 — 608 33 66
Rajmund Litkowiec: Technical Assistant
Address for correspondence
GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 1—4

FROM THE EDITOR

A set of eleven articles in this issue is arranged in three parts, containing respectively papers devoted to sampling and estimation, a research paper, and papers addressing diverse statistical problems.

Relatively largest, the first part starts with a paper by **M. Liberts** (of Latvia) on ***The Cost Efficiency of Sampling Designs*** that is aimed at developing a mathematical framework to compare various sample designs with respect to the expected precision of estimates and the data collection cost. A framework is proposed which employs artificial population data generation, survey sampling techniques, survey cost modelling, Monte Carlo simulation experiments and other techniques. This framework is applied next to analyze the cost efficiency of the sample design used for the Latvian Labour Force Survey (LFS). The advantage of the framework is that no extra data collection is required as it utilizes data already available to a statistical agency (administrative records, population census data or sample survey data). The only requirement is that it must be possible to describe the sampling process of a design as an R function. It is proven that the two-stage sampling design used currently for the LFS provides more precise parameter estimates under the condition of equal fieldwork cost when compared to two other simpler sampling designs.

In their paper ***Synthetic Estimators Using Auxiliary Information in Small Domains,*** **P. K. Rai, K. K. Pandey** discuss the generalized class of synthetic estimators for estimating the population mean of small domains under the information of two auxiliary variables. They describe the special cases under the different values of the constant beta involved in the proposed generalized class of synthetic estimator. A numerical illustration for the two auxiliary variables and compared results for the synthetic ratio estimator under single and two auxiliary variables are being given. In conclusion, it shows that at least two auxiliary variables will be the better choice over a single one when the sample size decreases and that it is useful to make use of information on the auxiliary variable to increase the precision of the estimators.

R. Shanker and **A. Mishra** propose ***A Two-parameter Lindley Distribution***, focusing on the case in which one-parameter Lindley Distribution (LD) is a particular one. Its moments, failure rate function, mean residual life function and stochastic orderings are discussed. The maximum likelihood method and the method of moments have been discussed for estimating its parameters. The distribution has been fitted to some data sets to test its goodness of fit. Finally, the proposed distribution has been fitted to a number of data sets relating to waiting and survival times to test its goodness of fit to which the one-parameter LD was

fitted earlier. However, it is found that two-parameter LD provides better fits than those by the one-parameter LD.

Best Linear Unbiased Estimators of Population Mean on Current Occasion in Two-Occasion Rotation Patterns are studied by **G. N. Singh** and **S. Prasad**. Behavior of the proposed estimators and their respective optimum replacement policies are discussed. Empirical studies are carried out to examine the performance of the proposed estimators and consequently the suitable recommendations are made. In conclusion, the proposed estimators are proved to be the best linear unbiased estimators of population mean $\overline{Y}$ with their respective minimum variance. They may be seen as new innovative ideas in the survey literature as they nicely utilized the information on an auxiliary variable in order to improve the precision of the estimates. According to the analyzed results, the proposed estimators enhance the precision of estimates as well as reduces the cost of the survey. Therefore, they may be recommended to survey practitioners for use in real life problems.

J. Subramani and **G. Kumarapandiyan** in ***Estimation of Finite Population Mean Using Deciles of an Auxiliary Variable*** discuss the problem of a class of modified ratio estimators for estimation of population mean of the study variable when the population deciles of the auxiliary variable are known. The bias and the mean squared error of the proposed estimators are derived and compared with that of existing modified ratio estimators for certain known populations. Also, authors derive the conditions for which the proposed estimators perform better than the existing modified ratio estimators. Based on the numerical study they conclude that the proposed modified ratio estimators perform better than the existing modified ratio estimators and they recommend that these estimators be employed in practical situations.

This part is concluded by **N. Bokun's** paper on ***Sample Surveys of Households in Belarus: State and Perspectives.*** The main principles, characteristics and problems of three sample surveys of households conducted by the State Statistics of Belarus are discussed: (1) The Household Sample Surveys (on expenses and incomes); (2) the Private Subsidiary Plots in rural areas (PSP); and (3) the Labour Force Survey (LFS). For each of them the purpose, sample design, data collection, methods of estimation and possible ways to improve the surveys are briefly presented. The sample units are households and some target population groups (for example, persons aged 15-74). The surveys cover the whole country: the regions and the city of Minsk (the sample fraction is at the level of 0.2-0.6% of HHs; sample frames are Census and additional databases; face-to-face interview is a mode of data collection). The main challenges faced so far relate to sample localization, the construction of regional (district) samples, non-sampling errors, non-response (20-30%), presence of atypical units, not appropriate extrapolation, the use of different weighting schemes, the assessment of structural employment and unemployment indicators (for LFS), improving the representativeness of the quarterly data.

There is one research paper: ***Maternal Nutritional Status and Lactational Amenorrhea in India: A Simulation Analysis*** by **L. K. Dwivedi**. Its main objective is to examine the linkages between maternal nutritional status (measured by body mass index-BMI) and postpartum amenorrhea among currently breast-feeding women in India and its region. The probability to remain amenorrheic through simulative approach has been estimated to get better understanding of the impact of maternal nutritional status on postpartum amenorrhea. Using National Family Health Survey-2 data, women who were not pregnant, who were breast-feeding and who were not using any hormonal contraceptives at the time of the survey were included in the analysis. There was no significant difference existing between mean BMI of each region of India before and after imputation of missing cases. The interaction term between maternal nutritional status and duration of breast-feeding (child's age) was significantly associated with the likelihood of having resumed menstruation after controlling for breast-feeding practices, child nutritional status and socio-economic and demographic covariates. The effect of maternal nutritional status on lactational amenorrhea was not found to be significant when women were breast-feeding since last 12 months except in the northern region of India. However, after 12 months of breast-feeding, the probability of undernourished women to remain amenorrheic was likely to be greater and this trend was highly consistent across all the six regions included in the analysis.

The 'other articles' part is opened by **R. Calabrese's** paper ***A probabilistic scheme with uniform correlation structure***. The probabilistic schemes with independence between the trials show different dispersion characteristics depending on the behaviour of the probabilities of the binary event in the trials. The author proposes a probabilistic scheme with uniform correlation structure that leads to different dispersion characteristics depending on the sign of the linear correlation. A hypothesis test is suggested to identify the type of the dispersion of the probabilistic scheme.

In his article ***Bias Reduction of Finite Population Imputation by Kernel Methods***, **N. Pettersson** discusses missing data problem and proposes real donor imputation for item nonresponse. A pool of donor units with similar values on auxiliary variables is matched to each unit with missing values. The missing value is then replaced by a copy of the corresponding observed value from a randomly drawn donor. Although such methods can to some extent protect against nonresponse bias, the estimator and the nature of the data also matter. Techniques adopted from kernel estimation are used to deal with this problem. Using Pólya urn sampling the set of potential donors with units already imputed was sequentially updated; multiple imputations via Bayesian bootstrap was used to account for imputation uncertainty. Simulations with a single auxiliary variable show that such imputation method performs almost as well as competing methods with linear data, but better when data is nonlinear, especially with large samples.

A. K. Shukla, **S. K. Yadav**, and **G. C. Misra**, in their paper ***A Linear Model for Uniformity Trial Experiments*** propose such a type of model along showing

that it yields better results than other existing models. Uniformity trial experiments are required to assess fertility variation in agricultural land. Several models have appeared in the literature, of which Fairfield Smith's Variance Law assuming a nonlinear relationship between the coefficient of variation (C.V.) and a plot size has been extensively used in uniformity trial studies. The expression for point of maximum curvature for the proposed model is much simpler as compared to the model of Fairfield Smith. The appropriateness of the proposed model has also been verified with the help of a data set. In conclusions, authors recommend that the discussed linear model should be preferred in uniformity trial experiments.

In the last paper, ***Model of Latent Profile Factor Analysis for Ordered Categorical Data*, P. Tarka** discusses some problems associated with application of the factor analysis starting with observation that a common factor analysis solution is typically being used based on continuous data. This paper addresses the issue of latent variable models where discrete variables are used. One of them, called Latent Profile Factor Analysis (LPFA) is of particular interest. In order to prove the model's functionality in practice of market research, a brief example of LPFA model for ordered categorical data (based on one-factorial solution) in reference to hedonic consumption data is given in the paper. The proposed solution clarifies, simplifies and reduces ordered data (categorical responses) into more simple form than the previous model based on classic factor analysis.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 5

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, Spring 2013
Vol. 14, No. 1, pp. 7–30

THE COST EFFICIENCY OF SAMPLING DESIGNS

MĀRTIŅŠ LIBERTS[1]

ABSTRACT

The aim of a sample survey is to obtain high quality estimates of population parameters with low cost. The expected precision of estimates and the expected data collection cost are usually unknown making the choice of sampling design a complicated task. Analytical methods can not be used often because of the complexity of the sampling design or data collection process. The aim of this paper is to develop a mathematical framework to compare chosen sampling designs with respect to the expected precision of estimates and the data collection cost. As a result a framework is developed which employs artificial population data generation, survey sampling techniques, survey cost modelling, Monte Carlo simulation experiments and other techniques. The framework is applied to analyse the cost efficiency of the sampling design currently used for the Latvian Labour Force Survey.

Key words: cost efficiency, simulation study, survey cost estimation, survey methodology, variance of estimators.

1. Introduction

The inspiration for this paper comes from pure practical necessity. National Statistical Institutes (NSIs) are the main providers of official statistics in most countries. A large proportion of official statistics produced by NSIs are done so using data collected via sample surveys, with the main customer of official statistics being the general public (or tax payers, in other words). These days, cost efficiency is an essential consideration in all government spending; the question is, *are NSI sample surveys cost efficient?*

There is not a simple answer to the question posed. A sample survey can possess one of many different sampling designs. The simplest sampling designs do not necessarily provide the lowest data collection cost. More complex sampling designs are considered in theory and applied in practice to obtain statistical information with an acceptable precision at a lower cost. In designing a sample survey, the following considerations should be decided upon: *What is the expected precision of the estimates of population parameters? What is the expected data collection cost? Which sampling design should be chosen in order to minimise sampling errors under a*

[1]University of Latvia, Raiņa bulvāris 19, Rīga, LV-1586, Latvia, *martins.liberts@gmail.com.*

fixed data collection cost? These are commonly asked questions during the planning stage of a sample survey. In most cases, the answers to the questions posed cannot be gained through analytical means and NSIs are usually reliant on expert judgement to some extent.

The relation between the precision of estimates and survey cost has been discussed in literature for at least 70 years, though the topic has not been comprehensively addressed. Different aspects of the relationship have been analysed and different goals of analysis have been set by authors but it is possible to observe the lack of common foundations for the topic. One of the first papers devoted to the topic are by Mahalanobis (1940) and Jessen (1942). The topic is extensively discussed by Hansen, Hurwitz, and Madow (1953) and Kish (1965). Significant book regarding the topic is by Groves (1989). The author advocates simulation studies to be the best-suited for a sample design analysis because of usual complexity of cost and precision functions.

Several events have been organised recently, in the United States of America, devoted to the topics of survey cost estimation and simulation models for survey fieldwork operations. For example "Survey Cost Workshop" (2006, Washington, D.C.) and "Workshop on Microsimulation Models for Surveys" (2011, Washington, D.C.). The research of survey field operations is a brand new topic in the scope of statistical research. Several research activities have been devoted to the topic only recently (Chen, 2008; Cox, 2012).

The Latvian Labour Force Survey (LFS) is the main object of the study in the paper. It was organised for the first time in November 1995 (Lapiņš, 1997) and ran biannually. The first redesign of the LFS sampling design was done after the 2000 Latvian Population Census with the new sampling design launched in 2002 (Lapiņš, Vaskis, Priede, & Bāliņa, 2002). It become a continuous survey after the redesign. The second redesign of the survey occurred in 2006. The re-launch of the LFS with the new sampling design and a much larger sample size took place in 2007. Finally, the latest redesign of the LFS sampling design was done by the author in 2009 (Liberts, 2010). The main reason for redesigning the LFS sampling design for the third time was the necessity to update the population frame used for the first-stage sampling units. The redesign resulted in a new sample drawn which was used to run the LFS since 2010. More information regarding the history of the LFS is given by Central Statistical Bureau of Latvia (2012) and European Commission (2012a, 2012b).

The target population and the parameters of interest in the case of the LFS are described in the second section. Artificial population data reflecting the target population of the LFS are necessary to do simulation experiments. A methodology to develop artificial population data is presented in the third section. Artificial population data with characteristics similar to the target population of the LFS has been produced with this methodology. The fourth section of the paper is devoted to the

development and the application of the framework for the cost efficiency analysis of sampling designs.

2. Target population and parameters of interest

The target population of the LFS is defined as all residents permanently living in private households. Residents at working-age (15–74 years) compose the main domain of interest. The target population is continuously changing over time, for example some individuals are losing or gaining employment every day. The target population is observed on a weekly basis by the methodology of the LFS (European Commission, 2012b, p. 5).

An individual is called **unit** and denoted by v_i (there are cases when households are used as units). The set of all units is denoted by V. The size of V is M. The units are labelled with an index i where $i \in \overline{1, M}$, $V = \{v_1, v_2, \ldots, v_M\}$. The observation of unit v_i in week w is called **element** and denoted by $u_{i,w}$. The set of all elements in week w is denoted by U_w. There are M elements in U_w. The elements of U_w are labelled with a double index (i, w) where i refers to a unit and w refers to a week, $U_w = \{u_{1,w}, u_{2,w}, \ldots, u_{M,w}\}$. Values $y_{i,w}$ are associated to elements $u_{i,w}$ from U_w. The total of a variable y in week w is defined as

$$Y_w = \sum_{i=1}^{M} y_{i,w}.$$

The total number of weeks observed is denoted by W and w is the week index, $w \in \overline{1, W}$. The set of elements over W weeks is denoted by U, $U = \cup_{w=1}^{W} U_w$. Each U_w consists of the observation of units from V observed in different weeks. The size of U_w is constant over time, $|U_w| = M$ for all w. The size of U is denoted by N, $|U| = \sum_{w=1}^{W} M = MW = N$. An index k is used to label elements over W weeks, $k \in \overline{1, N}$. The elements of each U_w are ordered according to the order of the units of V. The indices

$$\{k : ((k-1) \bmod M) + 1 = i\}$$

correspond to the unit v_i. The example of the set U is given in Table 1. The M rows of the table represent units. The W columns of the table represent weeks observed. The cells of the table represent elements. The dimension of the table is $M \times W$.

The total of the variable y over W weeks is defined as

$$Y = \sum_{w=1}^{W} Y_w = \sum_{w=1}^{W} \sum_{i=1}^{M} y_{i,w} = \sum_{k=1}^{N} y_k.$$

Two types of parameter are considered in the further analysis – the average of weekly totals and the quarterly ratio of two totals. The average of weekly totals is defined

Table 1. Example of set U

i	$w=1$	$w=2$	$w=3$	$w=4$	$w=5$	$\cdots$	$w=W$
1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$\cdots$	$y_{1,W}$
2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$\cdots$	$y_{2,W}$
3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	$y_{3,5}$	$\cdots$	$y_{3,W}$
				$\cdots$			
M	$y_{M,1}$	$y_{M,2}$	$y_{M,3}$	$y_{M,4}$	$y_{M,5}$	$\cdots$	$y_{M,W}$

by

$$Y_q = \frac{1}{13}\sum_{w=1}^{13} Y_w = \frac{1}{13}\sum_{w=1}^{13}\sum_{i=1}^{M} y_{i,w} = \frac{1}{13}\sum_{k=1}^{N} y_k = \frac{1}{13}Y,$$

and the quarterly ratio of two totals is defined by

$$R_q = \frac{Y_q}{Z_q} = \frac{\sum_{w=1}^{13} Y_w}{\sum_{w=1}^{13} Z_w} = \frac{\sum_{w=1}^{13}\sum_{i=1}^{M} y_{i,w}}{\sum_{w=1}^{13}\sum_{i=1}^{M} z_{i,w}} = \frac{\sum_{k=1}^{N} y_k}{\sum_{k=1}^{N} z_k}.$$

The estimators of Y_q and R_q are constructed using the π estimator (Särndal, Swensson, & Wretman, 1992, p.42, 176) as

$$\hat{Y}_q = \frac{1}{13}\sum_{(i,w)\in s}\frac{y_{i,w}}{\pi_{i,w}} = \frac{1}{13}\sum_{k\in s}\frac{y_k}{\pi_k}, \tag{1}$$

$$\hat{R}_q = \frac{\sum_{(i,w)\in s}\frac{y_{i,w}}{\pi_{i,w}}}{\sum_{(i,w)\in s}\frac{z_{i,w}}{\pi_{i,w}}} = \frac{\sum_{k\in s}\frac{y_k}{\pi_k}}{\sum_{k\in s}\frac{z_k}{\pi_k}} \tag{2}$$

where s is a probability sample of elements and $\pi_{i,w}$ is an inclusion probability of element $u_{i,w}$ in a sample.

3. Artificial population data

Artificial population data are necessary to carry out simulation experiments. Artificial population data are created from the data of the Statistical Household Register (a statistical register owned and maintained by the Central Statistical Bureau of Latvia) and the survey data of the LFS. The artificial population data are represented by two files – one for a static population (the population of units) and other for a dynamic population (the population of elements). There are several assumptions incorporated in the artificial population model:

- the set of units V is fixed over W weeks,
- background variables such as age and place of residence are fixed during W weeks, while study variables (for example, employment status) can change from week to week,

- the membership of individuals to households is fixed over W weeks.

3.1. Static population data

Two data sources are used to construct the static population data. The list of individuals aged 15–74 on 30th January 2011 is extracted from the Statistical Household Register. There are 1 705 048 records (individuals) in the list. The list of individuals forms the frame for the static population. Demographic information (age and gender) and residence information (region, dwelling ID and geographical coordinates) is attached to the list. Dwelling ID allows individuals to be grouped by households (assume a single household per dwelling).

The LFS data are used to create study variables for the static population. The LFS data from 2007–2010 are used. The variables describing demographic information (age and gender), residence information (region and dwelling ID) and economic activity status are extracted from the survey data.

The data from both sources are merged using an imputation technique where recipients are the units in the register data and donors are the units in the survey data. Random donor imputation within classes is used (United Nations, 2010, p.162). However, this is not the classical application of random donor imputation because non-response is not the cause of data missingness here. The cause of data missingness is the fact that the register data do not contain the variable describing economic activity. Imputation classes are built in both data sets according to the same specification using demographic and residence information as auxiliary information.

The imputation is done at seven levels where imputation units are households at the first five levels and imputation units are individuals at the last two levels. Different specification of classes is used at each level. Donors and recipients are grouped in very detailed classes at the first level. As it is not possible to impute all households at the first level (there are not enough donors in each class at the first level), the imputation process is repeated for the not-imputed households at the succeeding levels by merging the imputation classes. There are 26 variables used to define household classes at the first level, 16 at the second level, 12 at the third level, 11 at the forth level and 10 at the fifth level (see Table 2 for more details, where strata is a variable with four values: “Riga”, “Cities”, “Towns”, and “Rural areas”; region is a variable with six values). Strata, region, gender and age are variables used to create imputation classes at the sixth level, and strata, region, gender and age group (12 age groups) are variables used to create imputation classes at the seventh level when imputation units are individuals.

The description of imputation procedure done at each level is given here. The imputation is done in each class c independently. A donor $d_k \in D_c$ is assigned to a recipient $r_i \in R_c$ with a probability $\frac{1}{|D_c|}$ if $|D_c| \geq 10$ where D_c is the set of donors in a class c, R_c is the set of recipients in a class c, and $|D_c|$ is the total number of donors in a class c. A donor $d_k \in D_c$ can be assigned to several recipients from R_c.

Table 2. Household imputation classes at the first five levels

Variable	Level 1	Level 2	Level 3	Level 4	Level 5
Males 15–19	1	1	1	1	1
Males 20–24	2	2	2	2	2
Males 25–29	3	3	3	3	3
Males 30–34	4	3	3	3	3
Males 35–39	5	4	3	3	3
Males 40–44	6	4	3	3	3
Males 45–49	7	5	4	4	4
Males 50–54	8	5	4	4	4
Males 55–59	9	6	4	4	4
Males 60–64	10	6	4	4	4
Males 65–69	11	7	5	5	5
Males 70–74	12	7	5	5	5
Females 15–19	13	8	6	6	6
Females 20–24	14	9	7	7	7
Females 25–29	15	10	8	8	8
Females 30–34	16	10	8	8	8
Females 35–39	17	11	8	8	8
Females 40–44	18	11	8	8	8
Females 45–49	19	12	9	9	9
Females 50–54	20	12	9	9	9
Females 55–59	21	13	9	9	9
Females 60–64	22	13	9	9	9
Females 65–69	23	14	10	10	10
Females 70–74	24	14	10	10	10
Strata	25	15	11	11	.
Region	26	16	12	.	.

The imputation is not done in a class c if $0 \leq |D_c| < 10$. The units imputed at one level are not re-imputed any more at the succeeding imputation levels. The units not imputed at one level will be imputed at one of succeeding imputation levels.

The imputation of households as units at the first five levels allows one to keep demographic and economic composition of households the same as observed in the survey data. The specification of the classes at the first five levels is hierarchical. The classification of the classes is the most detailed at the first level. The classes are merged by each succeeding level. Economic activity status is imputed for 82.2% of all individuals from the register data at the first five levels. The imputation for all individuals can not be done in this manner because there are classes of households in the register data which have not been observed in the survey data or have been observed only in few cases (less than 10).

Economic activity status is imputed for the rest of individuals at the last two levels with the same imputation technique except imputation units are individuals and other specification of classes is used. The classification of the classes here is based on the same auxiliary information as used at the first five levels, though it is used at the individual level rather than at the household level. The specification of the classes is hierarchical here as well. It is possible to impute economic activity status for all remaining individuals at the last two imputation levels.

3.2. Dynamic population data

A dynamic population according to the description in Section 2 is generated. A variable – economic activity status is extrapolated from the static population to the dynamic population. Let y_i be the economic activity status of an individual v_i from the static population. A Markov chain model is used to generate the dynamic population. The economic activity status y_i can take any of three different values, $y_i \in \{1, 2, 3\}$:

- $y_i = 1$ if an individual v_i is employed,
- $y_i = 2$ if an individual v_i is unemployed,
- $y_i = 3$ if an individual v_i is economically inactive.

The value of y_i is defined once in a week by the LFS methodology. Let $y_{i,w}$ be the economic activity status for an individual v_i on week $w \in \{0, 1, 2, \ldots\}$. Let $y_{i,w}$ be random variables and sequence $y_{i,0}, y_{i,1}, y_{i,2}, \ldots$ be a time-inhomogeneous Markov chain for an individual v_i. The state space of the Markov chain is $\{1, 2, 3\}$. The probability of going from a state k to a state l after a week for an individual v_i is

$$p_{i,w,w+1,k,l} = P\left(y_{i,w+1} = l \mid y_{i,w} = k\right).$$

Constant transition probabilities for all v_i are assumed

$$p_{i,w,w+1,k,l} = p_{w,w+1,k,l},$$

and a time-dependent transition matrix the same for every individual v_i is

$$\boldsymbol{P}_{w,w+1} = \begin{pmatrix} p_{w,w+1,1,1} & p_{w,w+1,1,2} & p_{w,w+1,1,3} \\ p_{w,w+1,2,1} & p_{w,w+1,2,2} & p_{w,w+1,2,3} \\ p_{w,w+1,3,1} & p_{w,w+1,3,2} & p_{w,w+1,3,3} \end{pmatrix}.$$

The estimate of $\boldsymbol{P}_{w,w+1}$ is necessary to generate artificial dynamic population data. It is assumed there are 52 weeks in each year, and 52 weeks are split in four seasonal quarters by 13 weeks in each.

The first quarter is shown as an example here. It is assumed that all 13 weekly transition matrices are equal for the first quarter. Thus, the following equivalence holds for the 13 weekly transition matrices:

$$\boldsymbol{P}_{0,1} = \boldsymbol{P}_{1,2} = \ldots = \boldsymbol{P}_{12,13}.$$

In general the transition matrix after 13 weeks is equal to the product of 13 weekly transition matrices: $\boldsymbol{P}_{0,13} = \prod_{w=0}^{12} \boldsymbol{P}_{w,w+1}$. Because of the previous equivalence we can write

$$\boldsymbol{P}_{0,13} = \boldsymbol{P}_{w,w+1}^{13} \quad \text{for all} \quad w \in \overline{0,12}.$$

It follows from the previous equation

$$\boldsymbol{P}_{w,w+1} = \sqrt[13]{\boldsymbol{P}_{0,13}} \quad \text{for all} \quad w \in \overline{0,12}.$$

The LFS is a rotating panel survey. There is a 50% overlap between the succeeding quarterly samples. The individuals are interviewed with 13 weeks shift between the succeeding quarterly samples. Theoretically it is possible to estimate $\boldsymbol{P}_{0,13}$ from the LFS data, because there are respondents who are observed both at week $w = 0$ and week $w = 13$. Practically the estimation of $\boldsymbol{P}_{0,13}$ will not be precise if only data from overlapping respondents of weeks $w = 0$ and $w = 13$ are used. It is because the number of such respondents is small.

Thus, the decision was made to estimate $\boldsymbol{P}_{0,13}$ using the LFS data from overlapping respondents of the first and the second quarter:

$$\hat{\boldsymbol{P}}_{0,13} = \hat{\boldsymbol{p}}_{1,2}$$

where $\hat{\boldsymbol{p}}_{1,2}$ is the estimate of transition matrix from the first quarter to the second quarter using the LFS data. This estimation is introducing some bias to the estimate of $\boldsymbol{P}_{0,13}$, but it is more stable estimate.

Thus, the estimate of the weekly transition matrix for the first quarter is estimated as

$$\hat{\boldsymbol{P}}_{w,w+1} = \sqrt[13]{\hat{\boldsymbol{p}}_{1,2}} \quad \text{for all} \quad w \in \overline{0,12}.$$

Similarly, the weekly transition matrices for the second quarter are estimated as

$$\hat{\boldsymbol{P}}_{w,w+1} = \sqrt[13]{\hat{\boldsymbol{p}}_{2,3}} \quad \text{for all} \quad w \in \overline{13,25},$$

where $\hat{\boldsymbol{p}}_{2,3}$ is the estimate of a quarterly transition matrix from the second quarter to the third quarter and so on.

A time-inhomogeneous Markov chain is used to introduce a seasonal component in dynamic population data as it is observed in the survey data with respect to the changes of economic activity status of individuals. The estimates of the quarterly transition matrices and the weekly transition matrices are available in Table 3. The estimated weekly transition matrices are used to generate the dynamic population data by weeks. The variable of economic status from the static population is used as the initial state ($w = 0$) for each individual.

Table 3. Estimates of Transition Matrices

q	w	$\hat{\boldsymbol{p}}_{q,q+1}$	$\hat{\boldsymbol{P}}_{w,w+1}$
1	$\overline{0,12}$	$\begin{pmatrix} 0.950 & 0.021 & 0.029 \\ 0.251 & 0.541 & 0.209 \\ 0.058 & 0.052 & 0.890 \end{pmatrix}$	$\begin{pmatrix} 0.996 & 0.002 & 0.002 \\ 0.025 & 0.952 & 0.022 \\ 0.004 & 0.006 & 0.990 \end{pmatrix}$
2	$\overline{13,25}$	$\begin{pmatrix} 0.944 & 0.021 & 0.035 \\ 0.253 & 0.540 & 0.206 \\ 0.055 & 0.055 & 0.891 \end{pmatrix}$	$\begin{pmatrix} 0.995 & 0.002 & 0.003 \\ 0.026 & 0.952 & 0.022 \\ 0.004 & 0.006 & 0.990 \end{pmatrix}$
3	$\overline{26,38}$	$\begin{pmatrix} 0.937 & 0.028 & 0.035 \\ 0.199 & 0.609 & 0.192 \\ 0.048 & 0.042 & 0.910 \end{pmatrix}$	$\begin{pmatrix} 0.995 & 0.003 & 0.003 \\ 0.019 & 0.962 & 0.019 \\ 0.004 & 0.004 & 0.992 \end{pmatrix}$
4	$\overline{39,51}$	$\begin{pmatrix} 0.930 & 0.033 & 0.037 \\ 0.183 & 0.596 & 0.221 \\ 0.042 & 0.043 & 0.915 \end{pmatrix}$	$\begin{pmatrix} 0.994 & 0.003 & 0.003 \\ 0.018 & 0.960 & 0.022 \\ 0.003 & 0.004 & 0.993 \end{pmatrix}$

4. Cost efficiency

Assume an arbitrary population parameter θ. There is a probability sample s_p drawn by a sampling design $p(s)$. The parameter θ is estimated by an estimator $\hat{\theta}_p$. The variance of $\hat{\theta}_p$ is denoted by $\mathrm{Var}_p\left(\hat{\theta}_p\right)$. There is a cost function $c(s_p)$. The operational cost of a sample s_p is computed by the cost function $c_p = c(s_p)$. The result of the cost function is a random variable because s_p is a random sample. The expectation of c_p under a sampling design $p(s)$ is denoted as $\mathrm{E}(c_p) = C_p$. Definition 1 is used to compare two sampling designs with respect to cost efficiency where γ is a survey budget available.

Definition 1. A sampling design $p(s)$ is more cost efficient than a sampling design $q(s)$ for estimation of a population parameter θ with a survey budget γ if

$$\mathrm{Var}_p\left(\hat{\theta}_p \middle| C_p \approx \gamma\right) < \mathrm{Var}_q\left(\hat{\theta}_q \middle| C_q \approx \gamma\right).$$

The parameter γ can be replaced by a parameter vector $\boldsymbol{\gamma}$ denoting budget allocation by operational domains in Definition 1. Specifying the budget as a vector is useful in practice if the allocation of a budget by operational domains is important. The practical application of Definition 1 to analyse cost efficiency of sampling designs is achieved by the following steps:

- selection of sampling designs to be analysed with respect to the cost efficiency,
- definition of a cost function $c(s)$,

- setting the total budget γ or a budget allocation $\boldsymbol{\gamma}$,
- setting specific sample design parameters for each chosen sample design to achieve the expected total cost or cost allocation for all designs approximately equal to γ or $\boldsymbol{\gamma}$ accordingly,
- selection of population parameters for analysis,
- calculation of variance for the estimators of parameters selected,
- determination of the most cost efficient sample design using Definition 1.

4.1. Sampling designs

A modified simple random sampling design (mSRS) is introduced as an alternative to the current LFS sampling design. The notation of Section 2 is used here. The set of sampled units is denoted by $\tilde{s} \subseteq V$. The set of sampled elements in week w is denoted by $s_w \subseteq U_w$. The set of sampled elements over W weeks is denoted by $s = \cup_{w=1}^{W} s_w \subseteq U$. The weekly sample size is denoted by m. The total sample size n is computed as mW. The value of m has to be chosen so that $n = mW \leq M$ because each unit can be sampled only once during W weeks. The goals of the mSRS are:

- all elements of U have sampling probabilities equal to $\pi_k = \frac{n}{N} = \frac{m}{M}$,
- weekly samples for W weeks are drawn,
- all weekly samples are drawn with equal sample size, $|s_w| = m$ for all w, making the total sample size equal to $n = mW$,
- all n sampled elements refer to n different units, one and only one element $u_{i,w}$ may be sampled for a unit v_i.

There are several techniques to achieve the sample by the mSRS. An example is presented here. The sample is selected in two steps. The first step is to select n units by simple random sampling without replacement from M units. The sampled units are sorted in a random order. The ordered sample of units is systemically split into W blocks with length m. The units of the first block determine the sampled elements for the first week, the units of the second block determine the sampled elements for the second week and so on until the units of the last block determine the sampled elements for the week W.

A probability to sample a unit v_i at the first step is equal to $\frac{n}{M}$. The probability of a unit v_i to be located in a block w after the random ordering is equal to $\frac{1}{W}$. A sampled element is determined by the index i of a sampled unit v_i and the index w of a block containing the unit v_i. Therefore, the sampling probability of an element is equal to $\pi_{i,w} = \pi_k = \frac{n}{M}\frac{1}{W} = \frac{n}{N} = \frac{m}{M}$.

A stratified mSRS sampling design is realised if units are stratified in H strata and mSRS is applied independently in each stratum with sample size n_h. The stratified mSRS is denoted as mSSRS.

Three sampling designs are chosen for the cost efficiency study. The first design is mSSRS with individuals as sampling units (denoted as mSSRSi). Each sampled individual is interviewed by a household questionnaire and an individual questionnaire. This is a similar sampling design used for LFS in Sweden and Denmark – stratified random sampling of individuals, and only sampled individuals take part in a survey (European Commission, 2012a).

The second sampling design is mSSRS with households as sampling units (denoted as mSSRSh). Each sampled household is interviewed by a household questionnaire and all household members are interviewed by an individual questionnaire. This is a similar sampling design used for LFS in Malta, Austria and United Kingdom – stratified random sampling of dwellings or households and all members of a sampled dwelling or household take part in a survey (European Commission, 2012a).

The third sampling design is two-stage sampling design (denoted as TSSh) used in practice for the Latvian LFS. The primary sampling units (PSUs) are census counting areas at the first stage. Census counting areas are geographically compact areas with low variation by size (here and afterwards the size of PSU is measured as the number of dwellings in PSU) making them useful for sampling purposes. The average PSU size is 238 in Riga (capital city), 219 in other cities (excluding Riga), 190 in towns and 141 in rural areas.

PSUs are stratified in four strata by the level of urbanisation (Riga – the capital of Latvia, other cities, towns and rural areas). PSUs are sampled by systematic πps sampling with random starting point and sampling probabilities proportional to PSU size. PSUs are ordered in "serpentine" order in each stratum allowing for implicit stratification by administrative territories. The systematic sampling of PSUs allows the implementation of the chosen rotation scheme `2-(2)-2` (European Commission, 2012a, p.7).

Dwellings are the secondary sampling units sampled by simple random sampling with fixed sample size in each stratum. Usually there is only one household in each dwelling. Each sampled dwelling is interviewed by a household questionnaire and all household members are interviewed by an individual questionnaire. More details about the TSSh design are available at Liberts (2010).

The two-stage sampling design using census counting areas as PSUs has been used for the Latvian LFS since 2002. Several questions about the chosen sampling design have been raised quite often: *Why should Central Statistical Bureau of Latvia (CSB) use such complex (two-stage) sampling design? Why CSB are not switching to more simpler (one-stage) sampling design?* One of the main reasons for these questions was the fact that design is using census counting areas as PSUs. The frame of census counting areas (PSUs) has to be updated using the resources of the CSB (it is because the census counting areas are not available in any administrative register). Thus, the question regarding the most appropriate sampling design for

LFS has been open for quite a long time. This explains the choice of the alternative sampling designs for this study.

It is not obvious which of the selected sampling designs is the most cost efficient in the case of the LFS. The mSSRSi and the mSSRSh could provide more precise estimates with smaller sample sizes because of lower cluster effect (in the case of the LFS). However, the TSSh requires lower fieldwork cost per unit because of shorter travelling distances for interviewers allowing to select larger sample size.

Other sampling designs can be analysed as well, for example, mixed designs where one-stage sampling is used for high density areas and two-stage sampling for low density areas (to reduce travelling cost). This kind of sampling design was used for the Latvian LFS in 1995–2001 (Lapiņš et al., 2002, p.628). On the one hand this kind of sampling design could have good cost efficiency properties.

On the other hand, the complexity of the design is higher making the estimators of population parameters and estimators of precision more complex. This could be an obstacle for the external users of survey micro-data or for automatic precision estimation systems assuming unified sampling design used throughout the survey. It will be possible to observe further in the paper that mixed sampling design (with chosen stratification) would not be more cost effective compared to the three chosen sampling designs.

4.2. Cost function

Assume a survey done by face to face personal interviews where interviewers are travelling to respondents. Two components of fieldwork cost are assumed – travel cost and interview cost. Travel cost is approximated by a function $c_1(s) = dK_f C_f k_d$ where d is the total travelling distance done by interviewers expressed in kilometres, K_f is the average fuel consumption expressed in litres per kilometre, C_f is the average price of fuel expressed in lats per litre (lats is the national currency of the Republic of Latvia, 1 lats = 0.702804 euro), and k_d is an adjustment coefficient specified by a statistician.

There are G interviewers available and there is an interviewer assigned to each unit in population. Sampled units for week w are split by interviewers according to the predefined interviewer assignment in population. Geographical coordinates are known for the sampled units and also for the residence places of interviewers. Distances between sampled units and interviewers residence are computed as the Euclidean distance.

The shortest path connecting the residence of an interviewer g and the sampled units assigned to an interviewer g is found by solving a travelling salesperson problem (TSP). The TSP is solved by the nearest insertion algorithm (Rosenkrantz, Ste-

arns, & Lewis, 1977, p.572). The total travel distance d is computed by

$$\sum_{g=1}^{G}\sum_{w=1}^{W} d_{g,w}$$

where W is the total number of weeks observed and $d_{g,w}$ is the length of the path found by solving a TSP for an interviewer g in week w. The constants K_f, C_f and k_d are set.

The interview cost is computed by a function $c_2(s) = aC_a + bC_b$ where a is the total number of individuals in a sample s, b is the total number of households in a sample s, C_a is the interview cost for an individual questionnaire, and C_b is the interview cost for a household questionnaire. A cost function $c(s) = c_1(s) + c_2(s) = K_f C_f k_d \sum_{g=1}^{G} d_g + aC_a + bC_b$ is used further in the study.

4.3. Fieldwork budget allocation

The field work budget $\boldsymbol{\gamma}$ is set equal to the survey budget necessary to run the LFS by the current sampling design (TSSh) for a quarter allocated by three operational domains: "Riga", "Cities" and "Towns and rural areas". The estimation of $\boldsymbol{\gamma}$ is done by a Monte Carlo simulation experiment.

The expected values of d_l, a_l and b_l are estimated by a Monte Carlo simulation experiment where l is the operational domain index. A sample is selected by the TSSh and the values of d_l, a_l and b_l are computed in each iteration. The total number of the iterations of the simulation is 6000. The values of K_f, C_f, C_a, C_b and k_d are set according to the available information about the LFS fieldwork organisation.

The resulting total survey cost for a quarter with TSSh design is 36 004.8. The allocation of the survey cost by operational domains and resulting field work budget is set as $\boldsymbol{\gamma} = \{5395.1, 7719.5, 22\,890.1\}$ (for "Riga", "Cities" and "Towns and rural areas" accordingly).

4.4. Design parameters of alternative sampling designs

The mSSRSi and mSSRSh are chosen as alternative sampling designs. The stratification of both designs is set equal to the operational domains of TSSh. Therefore, three strata ("Riga", "Cities", "Towns and rural areas") are created for each design. Units are individuals for the mSSRSi and units are households for the mSSRSh. A sample size is estimated independently for each design and each stratum (six cases). A stratum sample size n_h is the only parameter for the designs. The valid values of n_h are

$$\{n_h : (0 < n_h \leq M_h \;\&\; n_h \bmod 13 = 0)\}$$

where M_h is the total number of units stratum h. The aim is to find n_h so that $C(n_h) \approx \gamma_h$ where $C(n_h)$ is the expected survey cost with sample size n_h, and γ_h

is the survey budget for a stratum h. The solution is defined as

$$n_h^* = \underset{\{n_h : C(n_h) > \gamma_h\}}{\arg\min} C\left(n_h\right).$$

The solution is found by a stepwise procedure for each design and each stratum independently:

- Eight values of n_h widely spread in the interval of valid sample sizes are selected and $C\left(n_h\right)$ is estimated for each selected n_h with a Monte Carlo simulation.
- The relation between the expected cost and sample size is approximated by a non-linear regression $C\left(n_h\right) \sim \beta_0 + \beta_1 n_h + \beta_2 \sqrt{n_h}$. The regression coefficients β_0, β_1 and β_2 are estimated from the eight pairs of $\left\{n_h, \hat{C}\left(n_h\right)\right\}$.
- An approximate solution $\hat{n}_h^*$ is computed from the regression equation by

$$\hat{n}_h^* = \frac{\left(\sqrt{\hat{\beta_2}^2 - 4\hat{\beta_1}\left(\hat{\beta_0} - \gamma_h\right)} - \hat{\beta_2}\right)^2}{4\hat{\beta_1}^2}.$$

- It has been observed that the exact solution n_h^* is close to $\hat{n}_h^*$. The exact solution is found by another Monte Carlo simulation experiment estimating the cost for a sampling design with seven different sample sizes close to $\hat{n}_h^*$. The sample sizes chosen for the simulation are $\hat{n}_h^* - 39$, $\hat{n}_h^* - 26$, $\hat{n}_h^* - 13$, $\hat{n}_h^*$, $\hat{n}_h^* + 13$, $\hat{n}_h^* + 26$, $\hat{n}_h^* + 39$.

The resulting sample size and survey cost for each stratum and sampling design are available in Table 4, where table columns are: `n.PSU` – number of PSUs, `n.h` – number of households in sample, `n.i` expected number of individuals in sample, `c.travel` – expected travel cost, `c.interview` – expected interview cost, `c.total` – expected total survey cost (the total survey cost is slightly higher than the budget available for mSSRSi and mSSRSh sampling designs to preserve a conservative position with respect to the TSSh).

4.5. Parameters of interest

There are six parameters considered:

- `a.empl` – the average of weekly totals of employed individuals,
- `a.unem` – the average of weekly totals of unemployed individuals,
- `a.inact` – the average of weekly totals of economically inactive individuals,
- `r.act` – the activity rate (the total number of employed and unemployed individuals by the total number of working-age individuals),

Table 4. Sample size and survey cost by stratum and sampling design

stratum	design	n.PSU	n.h	n.i	c.travel	c.interview	c.total
Riga	mSSRSi	.	.	1 261	403.2	5 036.6	5 439.8
Riga	mSSRSh	.	1 001	2 105	351.7	5 107.6	5 459.4
Riga	TSSh	104	1 040	2 185	90.6	5 304.6	5 395.1
Cities	mSSRSi	.	.	1 781	660.1	7 099.0	7 759.2
Cities	mSSRSh	.	1 404	2 963	581.9	7 174.6	7 756.5
Cities	TSSh	208	1 456	3 073	278.8	7 440.8	7 719.5
Other	mSSRSi	.	.	2 834	11 631.4	11 301.7	22 933.1
Other	mSSRSh	.	2 340	5 554	10 356.7	12 573.7	22 930.4
Other	TSSh	416	3 536	8 318	3 964.2	18 925.9	22 890.1

- `r.empl` – the employment rate (the total number of employed individuals by the total number of working-age individuals),
- `r.unem` – the unemployment rate (the total number of unemployed individuals by the total number of employed and unemployed individuals).

Six parameters are estimated for the whole target population and also in breakdowns by domains. Three sets of domains are considered:

- geographical domain (4) – Riga, cities (excluding Riga), towns, and rural areas,
- age group (2) – individuals aged 15–24 and 25–74 years,
- geographical domain (4) $\times$ age group (2).

It makes 90 parameters (45 averages of weekly totals and 45 ratios of two totals) selected for the cost efficiency analysis.

4.6. Variance of parameter estimators

The variance of $\hat{Y}_q$ (1) by the mSSRSi and the mSSRSh is computed by

$$\mathrm{Var}\left(\hat{Y}_q\right) = \frac{1}{169}\sum_{h=1}^{H}\left(\frac{M_h^2}{m_h}\sum_{w} S_{w,h}^2\left(y\right) - M_h\sum_{w}\sum_{v} S_{w,v,h}\left(y\right)\right)$$

where h is a stratum index, H is the total number of strata, M_h is the total number of units in the unit population of a stratum h, m_h is the total number of units in the sample of a stratum h, $S_{w,h}^2\left(y\right)$ is the variance of a variable y in week w and a stratum h, and $S_{w,v,h}\left(y\right)$ is the covariance of a variable y between weeks w and v in a stratum h. The approximate variance of $\hat{R}_q$ (2) by the mSSRSi and the mSSRSh is computed by

$$\mathrm{AVar}\left(\hat{R}_q\right) = \frac{1}{Z_q^2}\sum_{h=1}^{H}\left(\frac{M_h^2}{m_h}\sum_{w} S_{w,h}^2\left(u\right) - M_h\sum_{w}\sum_{v} S_{w,v,h}\left(u\right)\right)$$

where Z_q is the denominator of R_q, and u is the so called linearised variable for the ratio of two totals (Särndal et al., 1992, p.178). The variance of $\hat{Y}_q$ and $\hat{R}_q$ by the TSSh is estimated by a Monte Carlo simulation experiment.

4.7. Cost efficiency analysis

The three selected designs are compared by their cost efficiency using Definition 1 for the estimation of each selected parameter. A hypothesis testing is used in the case when the estimate of the variance by the TSSh is compared to the variance by the mSSRSi or the mSSRSh. An assumption is made that the estimates of the parameters by the TSSh are normally distributed:

$$\hat{\theta} \sim N\left(\mu, \sigma^2\right)$$

where σ^2 is unknown and is estimated by $s^2 = s^2(\boldsymbol{x})$ from the data $\boldsymbol{x}$ of the simulation experiment. The length of $\boldsymbol{x}$ is equal to the total number of iterations in the simulation, $|\boldsymbol{x}| = J = 20\,000$ in this case. The aim is to compare σ^2 by the TSSh with the known σ_0^2 under alternative design. A one-sided hypothesis testing (Wasserman, 2004) is done:

$$\begin{aligned} H_0 &: \sigma^2 \geq \sigma_0^2, \\ H_1 &: \sigma^2 < \sigma_0^2. \end{aligned} \tag{3}$$

A test statistic is computed as

$$T(\boldsymbol{x}) = \frac{(J-1)\, s^2}{\sigma_0^2},$$

and a rejection region R is defined as

$$R = \{\boldsymbol{x} : T(\boldsymbol{x}) \leq c\}$$

where $c = F_{J-1}^{-1}(\alpha)$ is the value of the inverse cumulative distribution function of χ^2_{J-1} at α. The following statements with respect to H_0 are set:

$$\begin{aligned} &T(\boldsymbol{x}) \leq c \Rightarrow \text{reject } H_0, \\ &T(\boldsymbol{x}) > c \Rightarrow \text{retain (do not reject) } H_0. \end{aligned}$$

The smallest α which rejects H_0 is called p-value, and p-value is equal to the value of the cumulative distribution function of χ^2_{J-1} at the point $\frac{(J-1)s^2}{\sigma_0^2}$.

The most cost efficient sampling design for the estimation of a parameter is determined by the following procedure:

1. The value of σ_0^2 is computed as min $\left(\sigma^2_{mSSRSi}, \sigma^2_{mSSRSh}\right)$.
2. The hypothesis testing (3) is done by computing p-value.

3. The TSSh is chosen as the most cost efficient sampling design for a parameter and the procedure stops here if p-value is less than 0.01. The procedure is continued to the step 4 if p-value is equal or greater than 0.01.
4. The mSSRSi is chosen as the most cost efficient sampling design for a parameter if $\sigma^2_{mSSRSi} < \sigma^2_{mSSRSh}$, and the mSSRSh is chosen as the most cost efficient sampling design for a parameter otherwise.

The expected precision of parameter estimates by the three sampling designs and the most efficient sampling design determined is given in Tables 5, 6, 7, and 8. The columns of the tables are:

- `param`: the name of parameter,
- `dom`: five geographical domains – "Latvia", "Riga", "Cities" (excluding city Riga), "Towns" or "Rural" (rural areas),
- `age`: three age groups – "15–74", "15–24" or "25–74",
- `value`: the true value of a population parameter computed from the artificial population data,
- σ_1: the expected standard error of an estimate by the mSSRSi,
- σ_2: the expected standard error of an estimate by the mSSRSh,
- σ_3: the estimated standard error of an estimate by the TSSh,
- p-`val`: p-value of the hypothesis testing (3),
- `des.eff`: the most cost efficient sampling design determined by the framework – "mSSRSi", "mSSRSh" or "TSSh".

Table 5. Precision of the estimates for the average of weekly totals in Latvia

param	dom	age	value	σ_1	σ_2	σ_3	p-val	des.eff
a.empl	Latvia	15–74	972 327	11 034	12 061	11 437	1.000	mSSRSi
a.unem	Latvia	15–74	133 746	6 173	4 958	4 654	0.000	TSSh
a.inact	Latvia	15–74	545 052	10 513	9 109	8 605	0.000	TSSh
a.empl	Latvia	15–24	102 838	5 410	4 344	4 097	0.000	TSSh
a.unem	Latvia	15–24	27 693	2 868	2 191	2 034	0.000	TSSh
a.inact	Latvia	15–24	157 176	6 487	5 373	5 078	0.000	TSSh
a.empl	Latvia	25–74	869 489	11 204	10 802	10 150	0.000	TSSh
a.unem	Latvia	25–74	106 054	5 565	4 393	4 121	0.000	TSSh
a.inact	Latvia	25–74	387 876	9 499	7 800	7 282	0.000	TSSh

The efficiency of the sampling designs strongly depends on a domain and the type of a parameter. The mSSRSi is selected as the most efficient design only for three parameters – "the average of weekly totals of employed individuals" in the domains "Latvia", "Riga" and "Cities". The mSSRSh is reasonably efficient for the estimation of the averages of totals in the domain "Riga" – it has been selected as the most efficient design in five out of nine cases. There are five other parameters

Table 6. Precision of the estimates for the ratio of two totals in Latvia

param	dom	age	value	σ_1	σ_2	σ_3	p-val	des.eff
r.act	Latvia	15–74	0.670	0.0064	0.0049	0.0045	0.000	TSSh
r.empl	Latvia	15–74	0.589	0.0067	0.0051	0.0048	0.000	TSSh
r.unem	Latvia	15–74	0.121	0.0055	0.0043	0.0040	0.000	TSSh
r.act	Latvia	15–24	0.454	0.0161	0.0122	0.0113	0.000	TSSh
r.empl	Latvia	15–24	0.357	0.0155	0.0118	0.0109	0.000	TSSh
r.unem	Latvia	15–24	0.212	0.0197	0.0148	0.0138	0.000	TSSh
r.act	Latvia	25–74	0.716	0.0067	0.0053	0.0050	0.000	TSSh
r.empl	Latvia	25–74	0.638	0.0072	0.0057	0.0053	0.000	TSSh
r.unem	Latvia	25–74	0.109	0.0056	0.0044	0.0040	0.000	TSSh

in the domains “Riga” and “Cities” which are the most efficiently estimated by the mSSRSh.

The TSSh is the most efficient design for the estimation of ratios in the domain “Riga” and also for the estimation of totals and ratios in the domain “Cities”. The TSSh dominates in the domains “Towns” and “Rural areas” – all parameters in these domains are the most efficiently estimated by the TSSh. It is because travelling distances are longer in these domains compared to the domains “Riga” and “Cities”. The TSSh is the most efficient also for the estimation of the parameters representing the domain “Latvia” (only one parameter for the domain “Latvia” is more efficiently estimated by the mSSRSi).

The cost efficiency analysis is done from a conservative position with respect to the TSSh. Firstly, the total sample size of each stratum for the mSSRSi and the mSSRSh is chosen slightly larger compared to the TSSh (Section 4.4).

Secondly, the TSSh is chosen as the most efficient design only in the cases when it is supported by strong evidence (p-value of the hypothesis testing is less than 0.01). The mSSRSi and the mSSRSh are preferred in the cases when there is uncertainty in the determination of the most efficiency design. For example, there are several cases when the precision of estimates achieved by the mSSRSh and the TSSh is quite similar.

The TSSh sampling design can be used reasonably well in some of these cases even if the mSSRSh has been chosen as the most efficient design, for example, in cases for the estimation of the average of weekly totals of inactive individuals in the domain “Riga” and the average of weekly totals of employed individuals aged 25–74 in the domain “Riga” (these are the cases when p-value is slightly higher than 0.01).

The TSSh has achieved the highest precision of estimates in most cases despite the conservative position with respect to it. Therefore, it is recommended to use the currently used two-stage sampling design for the Latvian LFS to achieve the

Table 7. Precision of the estimates for the average of weekly totals

param	dom	age	value	σ_1	σ_2	σ_3	p-val	des.eff
a.empl	Riga	15–74	330 855	7 381	8 272	8 329	1.000	mSSRSi
a.unem	Riga	15–74	47 160	4 284	3 569	3 504	0.000	TSSh
a.inact	Riga	15–74	160 949	6 938	6 062	6 009	0.040	mSSRSh
a.empl	Riga	15–24	31 245	3 543	2 903	2 960	1.000	mSSRSh
a.unem	Riga	15–24	8 152	1 851	1 452	1 435	0.011	mSSRSh
a.inact	Riga	15–24	40 138	3 980	3 300	3 301	0.508	mSSRSh
a.empl	Riga	25–74	299 610	7 533	7 509	7 430	0.017	mSSRSh
a.unem	Riga	25–74	39 007	3 928	3 222	3 184	0.008	TSSh
a.inact	Riga	25–74	120 810	6 322	5 329	5 250	0.001	TSSh
a.empl	Cities	15–74	196 200	3 870	4 304	4 126	1.000	mSSRSi
a.unem	Cities	15–74	26 352	2 125	1 746	1 713	0.000	TSSh
a.inact	Cities	15–74	110 307	3 703	3 250	3 261	0.754	mSSRSh
a.empl	Cities	15–24	19 779	1 860	1 532	1 500	0.000	TSSh
a.unem	Cities	15–24	5 362	991	782	764	0.000	TSSh
a.inact	Cities	15–24	30 430	2 267	1 903	1 846	0.000	TSSh
a.empl	Cities	25–74	176 421	3 926	3 878	3 736	0.000	TSSh
a.unem	Cities	25–74	20 990	1 913	1 536	1 510	0.000	TSSh
a.inact	Cities	25–74	79 877	3 360	2 839	2 850	0.784	mSSRSh
a.empl	Towns	15–74	166 623	5 991	6 139	3 325	0.000	TSSh
a.unem	Towns	15–74	23 376	2 493	1 935	1 395	0.000	TSSh
a.inact	Towns	15–74	96 256	4 808	4 206	2 549	0.000	TSSh
a.empl	Towns	15–24	17 418	2 160	1 687	1 203	0.000	TSSh
a.unem	Towns	15–24	5 101	1 179	873	639	0.000	TSSh
a.inact	Towns	15–24	29 682	2 797	2 284	1 593	0.000	TSSh
a.empl	Towns	25–74	149 205	5 749	5 487	2 967	0.000	TSSh
a.unem	Towns	25–74	18 275	2 212	1 676	1 224	0.000	TSSh
a.inact	Towns	25–74	66 574	4 085	3 361	2 167	0.000	TSSh
a.empl	Rural	15–74	278 650	7 004	7 761	5 583	0.000	TSSh
a.unem	Rural	15–74	36 859	3 103	2 405	2 085	0.000	TSSh
a.inact	Rural	15–74	177 540	6 129	5 698	4 516	0.000	TSSh
a.empl	Rural	15–24	34 396	3 001	2 401	2 043	0.000	TSSh
a.unem	Rural	15–24	9 078	1 568	1 165	1 023	0.000	TSSh
a.inact	Rural	15–24	56 926	3 802	3 252	3 013	0.000	TSSh
a.empl	Rural	25–74	244 254	6 779	6 787	4 821	0.000	TSSh
a.unem	Rural	25–74	27 781	2 710	2 043	1 754	0.000	TSSh
a.inact	Rural	25–74	120 615	5 285	4 461	3 473	0.000	TSSh

Table 8. Precision of the estimates for the ratio of two totals

param	dom	age	value	σ_1	σ_2	σ_3	p-val	des.eff
r.act	Riga	15–74	0.701	0.0129	0.0101	0.0099	0.000	TSSh
r.empl	Riga	15–74	0.614	0.0137	0.0109	0.0106	0.000	TSSh
r.unem	Riga	15–74	0.125	0.0111	0.0090	0.0088	0.000	TSSh
r.act	Riga	15–24	0.495	0.0366	0.0287	0.0281	0.000	TSSh
r.empl	Riga	15–24	0.393	0.0358	0.0281	0.0277	0.001	TSSh
r.unem	Riga	15–24	0.207	0.0422	0.0329	0.0329	0.542	mSSRSh
r.act	Riga	25–74	0.737	0.0134	0.0109	0.0107	0.000	TSSh
r.empl	Riga	25–74	0.652	0.0145	0.0119	0.0116	0.000	TSSh
r.unem	Riga	25–74	0.115	0.0113	0.0092	0.0090	0.000	TSSh
r.act	Cities	15–74	0.669	0.0111	0.0088	0.0086	0.000	TSSh
r.empl	Cities	15–74	0.589	0.0116	0.0092	0.0091	0.001	TSSh
r.unem	Cities	15–74	0.118	0.0093	0.0075	0.0073	0.000	TSSh
r.act	Cities	15–24	0.452	0.0288	0.0227	0.0221	0.000	TSSh
r.empl	Cities	15–24	0.356	0.0277	0.0218	0.0213	0.000	TSSh
r.unem	Cities	15–24	0.213	0.0352	0.0275	0.0269	0.000	TSSh
r.act	Cities	25–74	0.712	0.0117	0.0097	0.0095	0.013	mSSRSh
r.empl	Cities	25–74	0.636	0.0125	0.0103	0.0102	0.069	mSSRSh
r.unem	Cities	25–74	0.106	0.0095	0.0076	0.0074	0.000	TSSh
r.act	Towns	15–74	0.664	0.0146	0.0105	0.0079	0.000	TSSh
r.empl	Towns	15–74	0.582	0.0153	0.0111	0.0082	0.000	TSSh
r.unem	Towns	15–74	0.123	0.0125	0.0092	0.0069	0.000	TSSh
r.act	Towns	15–24	0.431	0.0359	0.0263	0.0195	0.000	TSSh
r.empl	Towns	15–24	0.334	0.0342	0.0249	0.0184	0.000	TSSh
r.unem	Towns	15–24	0.227	0.0462	0.0334	0.0246	0.000	TSSh
r.act	Towns	25–74	0.716	0.0154	0.0116	0.0087	0.000	TSSh
r.empl	Towns	25–74	0.637	0.0165	0.0125	0.0092	0.000	TSSh
r.unem	Towns	25–74	0.109	0.0126	0.0093	0.0070	0.000	TSSh
r.act	Rural	15–74	0.640	0.0113	0.0083	0.0072	0.000	TSSh
r.empl	Rural	15–74	0.565	0.0117	0.0086	0.0074	0.000	TSSh
r.unem	Rural	15–74	0.117	0.0095	0.0070	0.0061	0.000	TSSh
r.act	Rural	15–24	0.433	0.0259	0.0188	0.0166	0.000	TSSh
r.empl	Rural	15–24	0.343	0.0248	0.0180	0.0156	0.000	TSSh
r.unem	Rural	15–24	0.209	0.0323	0.0234	0.0204	0.000	TSSh
r.act	Rural	25–74	0.693	0.0122	0.0093	0.0081	0.000	TSSh
r.empl	Rural	25–74	0.622	0.0128	0.0098	0.0084	0.000	TSSh
r.unem	Rural	25–74	0.102	0.0096	0.0071	0.0061	0.000	TSSh

highest overall precision under the current budget constrains. Switching to a simpler sampling design will result in one of two negative effects. The first possible negative effect is the loss of overall precision if the survey cost is kept in the current budget level. The second possible negative effect is the increase in the survey cost if overall precision level is kept equal to the current level.

5. Conclusions

The aim of this paper was to develop a mathematical framework to compare chosen sampling designs with respect to the expected precision of estimates and the data collection cost. The framework has been developed and its application in case of Latvian Labour Force Survey has been demonstrated. The framework presented in the paper utilises Monte Carlo simulation experiment techniques when analytical methods can not be applied.

The framework allows the user to gain information about the sampling design properties (for example, the expected fieldwork cost or the expected precision of estimates) in a relatively short time and with relatively low cost. This information is very valuable for survey planning and the decision making processes. The advantage of the framework is that no extra data collection is required. The framework utilises data already available to a statistical agency (administrative records, population census data or sample survey data).

A set of procedures is developed to support the implementation of the framework in practice. The aim of the procedures is to run Monte Carlo simulations of sampling designs. The procedures are developed in R which is a free software environment for statistical computing and graphics (R Core Team, 2013). The code of the procedures is available online at the "GitHub" repository (Liberts, 2013). The procedures are developed as modular functions. It allows for the extension of the procedures with additional functions if necessary. There is no limitation on the types of design that can be analysed by the procedures. The only requirement is that it must be possible to describe the sampling process of a design as an R function.

The cost efficiency of three sampling designs is analysed using the framework. The properties of the chosen sampling designs are explored and recommendations with respect to an appropriate sampling design for the Latvian LFS are given. It is proven that the two-stage sampling design used currently for the LFS provides more precise parameter estimates under the condition of equal fieldwork cost when compared to two other simpler sampling designs.

The developed framework for cost efficiency analysis is flexible. It can be applied for different surveys and arbitrary sampling designs. There are broad possibilities of tuning the framework to specific aspects under analysis, for example, the survey cost estimation can be extended to take into account other processes from

real fieldwork operations. The developed framework can be used both by national statistical agencies and private companies organising sample surveys.

The research can be continued by extending the framework with non-response modelling. The set of the developed R procedures has to be extended with additional procedures. The additional procedure is necessary to simulate the process of the non-response of sampled units. The cost function has to be adjusted to take into account the actions done by interviewers in the case of non-response. The procedure for estimation of the population parameters in the case of non-response is necessary.

Acknowledgements

This work has been supported by the European Social Fund within the project **«Support for Doctoral Studies at University of Latvia – 2»**.

I would like to thank the anonymous referees for their constructive comments and suggestions, helping to improve the manuscript.

REFERENCES

Central Statistical Bureau of Latvia., (2012). *Employment and unemployment* [Metadata]. Riga. Retrieved 15.12.2012, from http://ej.uz/CSB-LFS

CHEN, B.-C., (2008). *Stochastic simulation of field operations in surveys* (Research report). Washington: U. S. Census Bureau. Retrieved from https://www.census.gov/srd/www/byyear.html

COX, L., (2012). The case for simulation models of federal surveys. In *Research conference papers of federal committee on statistical methodology research conference 2012.* Washington.
Retrieved from http://www.fcsm.gov/events/papers2012.html

European Commission. (2012a). *Labour force survey in the EU, candidate and EFTA countries – Main characteristics of national surveys, 2011* (Tech. Rep.). Luxembourg: Eurostat. Retrieved from http://epp.eurostat.ec.europa.eu/

European Commission. (2012b). *Quality report of the European Union Labour Force Survey – 2010* (Tech. Rep.). Luxembourg: Eurostat. Retrieved from http://epp.eurostat.ec.europa.eu/

GROVES, R. M., (1989). *Survey errors and survey costs*. New Jersey: Wiley.

HANSEN, M. H., HURWITZ, W. N., & MADOW, W. G., (1953). *Sample survey methods and theory* (Vol. I). New-York: Wiley.

JESSEN, R. J., (1942). *Statistical investigation of a sample survey for obtaining farm facts* (Research Bulletin No. 304). Iowa State College of Agriculture and Mechanic Arts. KISH, L. (1965). *Survey sampling*. New-York: John Wiley & Sons.

LAPIŅŠ, J., (1997). Sampling surveys in Latvia: Current situation, problems and future development. *Statistics in Transition*, *3*(2), 281–292.

LAPIŅŠ, J., VASKIS, E., PRIEDE, Z., & BĀLIŅA, S., (2002). Household surveys in Latvia. *Statistics in Transition*, *5*(4), 617–641. Retrieved from http://www.stat.gov.pl/pts/15_ENG_HTML.htm

LIBERTS, M., (2010). The redesign of Latvian Labour Force Survey. In M. Carlson, H. Nyquist, & M. Villani (Eds.), *Official statistics – methodology and applications in honour of Daniel Thorburn* (pp. 193–203). Stockholm, Sweden: Stockholm University. Retrieved from http://officialstatistics.wordpress.com/

LIBERTS, M., (2013). *Survey-design-simulation* [Online code repository]. Retrieved from https://github.com/djhurio/Survey-Design-Simulation

MAHALANOBIS, P. C., (1940). A sample survey of the acreage under jute in Bengal. *Sankhyā: The Indian Journal of Statistics*, *4*(4), 511–530.

R CORE TEAM., (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria. Retrieved from http://www.r-project.org

ROSENKRANTZ, D., STEARNS, R., & LEWIS, P., II., (1977). An analysis of several heuristics for the traveling salesman problem. *SIAM Journal on Computing*, *6*(3), 563–581.

SÄRNDAL, C.-E., SWENSSON, B., & WRETMAN, J., (1992). *Model assisted survey sampling*. New-York: Springer.

United Nations. (2010). *Handbook on population and housing census editing: Revision 1*. New York: United Nations.

WASSERMAN, L., (2004). *All of statistics*. New-York: Springer.

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 31–44

SYNTHETIC ESTIMATORS USING AUXILIARY INFORMATION IN SMALL DOMAINS

P. K. Rai[1], **K. K. Pandey**[2]

ABSTRACT

In the present article we discuss the generalized class of synthetic estimators for estimating the population mean of small domains under the information of two auxiliary variables, and describe the special cases under the different values of the constant beta involved in the proposed generalized class of synthetic estimator. In addition we have taken a numerical illustration for the two auxiliary variables and compared the result for the synthetic ratio estimator under single and two auxiliary variables.

Key words: auxiliary information, small area (domain) estimation, synthetic estimation, optimum weights.

1. Introduction

An estimator is called a synthetic estimator if a reliable direct estimator for a larger area, covering several small areas, is used to derive an indirect estimator for a small area under the assumption that the small areas have the same characteristics as the large area (Gonzalez, 1973). Such estimators have been studied by Gonzalez (1973), Gonzalez and Waksberg (1973). It is a fact that if small domain sample sizes are relatively small the synthetic estimator performs better than the simple direct estimators, whereas when sample sizes are large the direct estimators perform better than the synthetic estimators (Schaible, Brock, Casady and Schnack, 1977). The classes of synthetic estimators proposed by the above authors give consistent estimators if the corresponding synthetic assumptions are satisfied. These authors, further, discuss the generalized class of synthetic estimators under simple random sampling and stratified random

[1] Department of Mathematics and Statistics, Banasthali University, Rajasthan India-304022. E-mail: raipiyush5@gmail.com.
[2] College of management & Economic Studies, University of Petroleum & Energy Studies (UPES), Energy Acres, P.O. Bidholi, Dehradun-248007. E-mail: krishan.pandey@gmail.com, kkpandey@ddn.upes.ac.in.

sampling schemes. In sample surveys usually auxiliary variables are used to increase the precision of the estimators. A ratio estimator is one of the most commonly used estimators among others for the population mean or population total with the help of an auxiliary character. It was shown by Tikkiwal, G.C. and Ghiya, A. (2004), Tikkiwal, G.C. and Pandey, K.K. (2007), Pandey Krishan K. and Tikkiwal, G.C. (2010), Pandey, Krishan K. (2010), that when an auxiliary variable is closely related with the variable under study, the small area estimators based on auxiliary information perform better than those which do not use auxiliary information. Further, Tikkiwal, G.C. and Pandey, K.K. (2007) discuss the generalized class of synthetic and composite estimators under Lahiri-Midzuno and systematic sampling schemes. The relative performances of these estimators are empirically assessed for the problem of crop acreage estimation for small domains.

It is rather difficult to assess the performance of these estimators theoretically. Here we have discussed the different aspect of the generalized class of synthetic estimators for small area estimation problems when more than one auxiliary information is available.

2. Generalized class of synthetic estimators in sample surveys

We define a generalized class of synthetic estimators for estimating the population mean $\overline{Y}$ under 'k' auxiliary variables $x_1, x_2, ...x_k$, as follows

$$\overline{y}_{syn} = \sum_{i=1}^{k} W_i \overline{y} \left(\frac{\overline{x}_i}{\overline{X}_i} \right)^{\beta_i} \tag{1}$$

where β_i's are equal to the $-\rho_{0i} \frac{C_0}{C_i}$ and W_i are the weights to be obtained by minimizing the variance of (1) subject to the condition that $\sum_{i=1}^{k} W_i = 1$. Here, $\overline{x}_i$ and $\overline{X}_i$ denote the sample mean and population mean of $x_i (i = 1, 2, ..., p)$ respectively, $\rho_{ij} (i \neq j = 0, 1, ..., p)$ denotes the correlation coefficient between x_i and x_j, and $C_i (i = 0, 1, ..., p)$ denotes the coefficient of variation of x_i; the suffix 0 stands for the variable y and $\overline{y}$ is the sample mean of the variable under study.

3. Notations and formulation under small domains

Let us represent the important notations which are to be used in this paper. Suppose that a finite population U= (1,…, i,…, N) is divided into 'A' non-overlapping domains U_a of size N_a (a=1,…, A) for which estimates are required. The domains may be numerous and represent small geographical areas of a sampled population, which may be a state or a sub-division of the state as the case may be. Let the characteristic under study be denoted by 'y'. Further, assume that the auxiliary information is also available and denoted by 'x'. A simple random sample (without replacement) s= (1,…, i, …, n) of size n is selected such that n_a (a=1,…, A) units in the sample 's' come from small area 'a'. Consequently,

$$\sum_{a=1}^{A} N_a = N \text{ and } \sum_{a=1}^{A} n_a = n \tag{2}$$

Let us consider the case of generalized synthetic estimator for estimating the population mean $\overline{Y}_a$ for domain 'a' under two auxiliary variables x_1 and x_2;

$$\overline{y}_{syn,a} = W_1 \overline{y} \left(\frac{\overline{x}_1}{\overline{X}_{1a}} \right)^{\beta_1} + W_2 \overline{y} \left(\frac{\overline{x}_2}{\overline{X}_{2a}} \right)^{\beta_2} \tag{3}$$

Here, W_1 and W_2 are the weights such that $W_1 + W_2 = 1$ and β_1 , β_1 are suitably chosen constants. To find the expectation and mean square error of the estimator $\overline{y}_{syn,a}$ define

$$\varepsilon_0 = \frac{\overline{y} - \overline{Y}}{\overline{Y}}, \qquad \varepsilon_1 = \frac{\overline{x}_1 - \overline{X}_1}{\overline{X}_1}, \qquad \varepsilon_2 = \frac{\overline{x}_2 - \overline{X}_2}{\overline{X}_2} \tag{4}$$

Then, clearly

$$E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_2) = 0 \tag{5}$$

$$E(\varepsilon_0^2) = \frac{f}{n} C_0^2, \quad E(\varepsilon_1^2) = \frac{f}{n} C_1^2, \quad E(\varepsilon_2^2) = \frac{f}{n} C_2^2 \tag{6}$$

and $$E(\varepsilon_o \varepsilon_1) = \frac{f}{n} C_{01}, \qquad E(\varepsilon_o \varepsilon_2) = \frac{f}{n} C_{02}, \qquad E(\varepsilon_1 \varepsilon_2) = \frac{f}{n} C_{12} \tag{7}$$

where

$$f = \frac{N-n}{N}, \qquad C_0^2 = \frac{S_y^2}{\overline{Y}^2}, \qquad C_1^2 = \frac{S_{x_1}^2}{\overline{X}_1^{\,2}}, \qquad C_2^2 = \frac{S_{x_2}^2}{\overline{X}_1^{\,2}},$$
$$C_{01} = \frac{S_{yx_1}}{\overline{Y}\,\overline{X}_1}, \quad C_{02} = \frac{S_{yx_2}}{\overline{Y}\,\overline{X}_2}, \quad C_{12} = \frac{S_{x_1x_2}}{\overline{X}_1\overline{X}_2}, \tag{8}$$

and

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2 ; \qquad S_{x_1}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{1i} - \overline{X}_1)^2$$

$$S_{x_2}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{2i} - \overline{X}_2)^2 ;$$

$$S_{yx_1} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})(x_{1i} - \overline{X}_1)$$

$$S_{yx_2} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})(x_{2i} - \overline{X}_2)$$

$$S_{x_1x_2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_{1i} - \overline{X}_1)(x_{2i} - \overline{X}_2) \tag{9}$$

4. Bias and mean square error

In this section bias and mean square error expressions are considered up to the terms of order (1/n) only. The $\overline{y}_{syn,a}$ can be expressed as

$$\overline{y}_{syn,a} = W_1\overline{Y}(1+\varepsilon_0)\left(\frac{\overline{X}_1(1+\varepsilon_1)}{\overline{X}_{1a}}\right)^{\beta_1} + W_2\overline{Y}(1+\varepsilon_0)\left(\frac{\overline{X}_2(1+\varepsilon_2)}{\overline{X}_{2a}}\right)^{\beta_2} \tag{10}$$

assuming that the contribution of terms involving powers in $\varepsilon_0, \varepsilon_1$ and ε_2 higher than the second order is negligible. The design bias of $\overline{y}_{syn,a}$ and $MSE(\overline{y}_{syn,a})$ is given below as

$$B(\overline{y}_{syn,a}) = W_1\overline{Y}\left(\frac{\overline{X}_1}{\overline{X}_{1a}}\right)^{\beta_1}\left(1+\frac{f}{n}\left(\frac{\beta_1(\beta_1-1)}{2!}C_1^2+\beta_1 C_{01}\right)\right) + W_2\overline{Y}\left(\frac{\overline{X}_2}{\overline{X}_{2a}}\right)^{\beta_2}\left(1+\frac{f}{n}\left(\frac{\beta_2(\beta_2-1)}{2!}C_2^2+\beta_2 C_{02}\right)\right)-\overline{Y}_a \tag{11}$$

$$\begin{aligned} MSE(\overline{y}_{syn,a}) = {} & \overline{Y}^2\left[W_1\left(\frac{\overline{X}_1}{\overline{X}_{1a}}\right)^{\beta_1}+W_2\left(\frac{\overline{X}_2}{\overline{X}_{2a}}\right)^{\beta_2}\right]^2 \\ & +\overline{Y}^2W_1^2\left(\frac{\overline{X}_1}{\overline{X}_{1a}}\right)^{2\beta_1}\left[\frac{f}{n}\left\{C_0^2+\beta_1(2\beta_1C_1^2-C_1^2+4C_{01})\right\}\right] \\ & +\overline{Y}^2W_2^2\left(\frac{\overline{X}_2}{\overline{X}_{2a}}\right)^{2\beta_2}\left[\frac{f}{n}\left\{C_0^2+\beta_2(2\beta_2C_2^2-C_2^2+4C_{02})\right\}\right] \\ & +2W_1W_2\overline{Y}^2\left(\frac{\overline{X}_1}{\overline{X}_{1a}}\right)^{\beta_1}\left(\frac{\overline{X}_2}{\overline{X}_{2a}}\right)^{\beta_2}\left(\frac{f}{n}\left\{\begin{array}{l} C_0^2+\beta_1\left(2C_{01}+\frac{(\beta_1-1)}{2!}C_1^2\right) \\ +\beta_2\left(2C_{02}+\frac{(\beta_2-1)}{2!}C_2^2\right)+\beta_1\beta_2C_{12}\end{array}\right\}\right) \\ & -2\overline{Y}_a\left[\begin{array}{l} W_1\overline{Y}\left(\frac{\overline{X}_1}{\overline{X}_{1a}}\right)^{\beta_1}\left(1+\frac{f}{n}\left(\frac{\beta_1(\beta_1-1)}{2!}C_1^2+\beta_1C_{01}\right)\right) \\ +W_2\overline{Y}\left(\frac{\overline{X}_2}{\overline{X}_{2a}}\right)^{\beta_2}\left(1+\frac{f}{n}\left(\frac{\beta_2(\beta_2-1)}{2!}C_2^2+\beta_2C_{02}\right)\right)\end{array}\right]+\overline{Y}_a^2 \end{aligned} \tag{12}$$

Also, the optimum value for the weights W_1^{opt} and W_2^{opt} can be obtained by minimizing mean square error term of (12).

5. Special cases: various synthetic estimators

The generalized synthetic estimator $\overline{y}_{syn,a}$ reduces to the simple synthetic estimator if β_1 and β_2 equal to zero, i.e. $\beta_1 = \beta_2 = 0$

$$\overline{y}_{syn,a} = \overline{y} = \overline{y}_{syn,s,a} \tag{13}$$

and synthetic assumption $\overline{Y}_a \left(\overline{X}_a\right)^{\beta} \cong \overline{Y}(\overline{X})^{\beta}$ reduces to $\overline{Y}_a \cong \overline{Y}$. Substituting $\beta_1 = \beta_2 = 0$ in the expression (11) we get

$$B(\overline{y}_{syn,a}) = W_1\overline{Y} + W_2\overline{Y} - \overline{Y}_a = \overline{Y} - \overline{Y}_a = B(\overline{y}_{syn,s,a}) \tag{14}$$

This is the expression for design bias of the simple synthetic estimator. The design bias of the synthetic estimator vanishes if the synthetic assumption, i.e. $\overline{Y}_a \cong \overline{Y}$ is satisfied. Now $\beta_1 = \beta_2 = 0$ in the expression (12) gives

$$MSE(\overline{y}_{syn,s,a}) = \overline{Y}^2\left(W_1 + W_2\right)^2 + \overline{Y}^2\frac{f}{n}C_0^2\left(W_1 + W_2\right)^2 - 2\overline{Y}_a\overline{Y}(W_1 + W_2) + \overline{Y}_a^2$$

$$= \overline{Y}^2\frac{f}{n}C_0^2 = \frac{N-n}{Nn}S_y^2 \tag{15}$$

This is the mean square error of simple synthetic estimator under said synthetic assumption.

The generalized Synthetic estimator $\overline{y}_{syn,a}$ reduces to ratio synthetic estimator under two auxiliary variables, if β_1 and β_2 equal to -1, i.e. $\beta_1 = \beta_2 = -1$

$$\overline{y}_{syn,r,a} = W_1\left(\frac{\overline{y}}{\overline{x}_1}\right)\overline{X}_{1a} + W_2\left(\frac{\overline{y}}{\overline{x}_2}\right)\overline{X}_{2a} \tag{16}$$

Substituting $\beta_1 = \beta_2 = -1$ in the expression (11) and (12) we get the expressions for the bias and mse for the ratio synthetic estimator.

The generalized synthetic estimator $\overline{y}_{syn,a}$ reduces to the product synthetic estimator under two auxiliary variables, if β_1 and β_2 equal to +1, i.e. $\beta_1 = \beta_2 = +1$

$$\overline{y}_{syn,p,a} = W_1\overline{y}\left(\frac{\overline{x}_1}{\overline{X}_{1a}}\right) + W_2\overline{y}\left(\frac{\overline{x}_2}{\overline{X}_{2a}}\right) \tag{17}$$

Substituting $\beta_1 = \beta_2 = +1$ in the expression (11) and (12) we get the expressions for the bias and mse for the product synthetic estimator.

6. Numerical illustration

We consider the study variable as REV84, the real estate values according to 1984 assessment and use the two auxiliary variables as population under municipalities of 1975 and 1985 of the different geographic region indicator of Swedish municipalities. Just draw the sample of different sizes using SRSWOR scheme and analyze the cases for regions 1, 2 and 3 as small domains.

We have considered the cases under single and double auxiliary variables and computed the biases and mse's for the different sample sizes. Using the expressions of optimum weights we have computed the value of weights for the generalized synthetic estimator under $\beta_1 = \beta_2 = -1$ which reduces to synthetic ratio estimator, thus W_1^{opt} = 0.978828466 and W_2^{opt} =0.021171534. And $\overline{Y}_a$ = 3011.683, $\overline{X}_{1a}$ =28.92308, $\overline{X}_{2a}$ = 255.0192, $\overline{Y}$ =3133.862676, $\overline{X}_1$ = 28.80986, $\overline{X}_2$ = 111.9471831

Under single auxiliary variable the bias and mse for the synthetic ratio estimator is given by

$$B_2 = B\left(\overline{y}_{syn,r,a}\right) = \frac{\overline{Y}}{\overline{X}}\overline{X}_a\left[1+\frac{N-n}{Nn}\left(C_x^2 - C_{xy}\right)\right] - \overline{Y}_a \tag{18}$$

and

$$MSE\left(\overline{y}_{syn,r,a}\right) = \left(\frac{\overline{Y}}{\overline{X}}\overline{X}_a\right)^2\left[1+\frac{N-n}{Nn}\left\{3C_x^2 + C_y^2 - 4C_{xy}\right\}\right] - 2\overline{Y}_a\left(\frac{\overline{Y}}{\overline{X}}\overline{X}_a\right)\left[1+\frac{N-n}{Nn}\left(C_x^2 - C_{xy}\right)\right] + \overline{Y}_a^2 \tag{19}$$

Using the equations above we show the results for bias and mse of synthetic ratio estimator under two scenarios for the different sample sizes in the given tables in appendix. The results can be also explored by the following graphical presentation.

7. Conclusions

At least two auxiliary variables will be the better choice over a single one when the sample size decreases. In sample surveys it is useful to make use of information on the auxiliary variable to increase the precision of the estimators. The above study will provide the motivation towards the use of generalized class of synthetic estimators in the small area estimation, when the information on two auxiliary variables is available.

Acknowledgements

The authors are grateful to the University Grant Commission (UGC), F.No.36-341/2008 (SR), New Delhi for providing support and facilitation to this research and development work.

REFERENCES

AGRAWAL, M. C. and ROY, D. C., (1997). Efficient Estimators for Small Domains. Jour. Ind. Soc. Ag. Statistics 52(3), 327-337.

GHOSH, M. and RAO, J. N. K., (1994). Small Area Estimation: An Appraisal. Statistical Science, 9, 55-93.

GONZALEZ, M. E., (1973). Use and Evaluation of Synthetic Estimates. Proceedings of the Social Statistical Section of American Statistical Association, 33-36.

GONZALEZ, M. E. and WAKSBERG, J., (1973). Estimation of the Error of Synthetic Estimates. Paper presented at first meeting of the International Association of Survey Statisticians, Vienna, Austria, 18-25, August 1973.

HEDAYAT, A. S. and SINHA, B. K., (1991). Design and Inference in Finite Population Sampling. John Wiley and Sons, New York.

PANDEY, KRISHAN K. and TIKKIWAL, G. C., (2010). „Generalized class of synthetic estimators for small area under systematic sampling design" Statistics in Transition-new series, Poland, Vol.11 No.1, pp. 75-89.

PANDEY, KRISHAN, K., (2010). „Aspects of small area estimation using auxiliary information". Book published by VDM Verlag Dr. Muller GmbH & Co. KG, Germany. (ISBN: 978-3-639-31569-1).

PLATEK, R., RAO, J. N. K., SARNDAL, C. E. and SINGH, M. P., (1987). Small Area Statistics: An International Symposium. John Wiley and Sons, New York.

PURCELL, N. J. and KISH, L., (1979). Estimation for Small Domain. Biometrics, 35, 365-384.

SARNDAL, C. E., SWENSSON, B. and WRETMAN, J., (1992). Model Assisted Survey Sampling. Springer-Verlag, New York.

SCHAIBLE, W. L., BROCK, D. B., CASADY, R. J. and SCHNACK, G. A., (1977). An Empirical Comparison of the Simple Inflation, Synthetic and Composite Estimators for Small Area Statistics. Proceedings of the American Statistical Association, Social Statistics Section, 1017-1021.

SINGH, M. P., GAMBINO, J. and MANTEL, H., (1993). Issues and Options in the Provision of Small Area Data. Proceedings of International Scientific Conference on Small Area Statistics and survey Design (Held in September, 1992 in Warsaw, Poland), 37-75.

SRIVASTAVA, S. K., (1967). An estimator using auxiliary information in sample surveys, Calcutta Statistical Association Bulletin, 16, 121-132.

TIKKIWAL, B. D. and TIKKIWAL, G. C., (1991). Sampling Strategies in Surveys. The Role of the Theory of T-Classes and Computers. Symposium Ind. Agri. Stat. Research Institute, New Delhi, 287-296.

TIKKIWAL, G. C. and GUPTA, A. K., (1991). Estimation of Population Mean under Successive Sampling When Various Weights and Regression Coefficient are Unknown. Biometrical Journal, 33, 529-538.

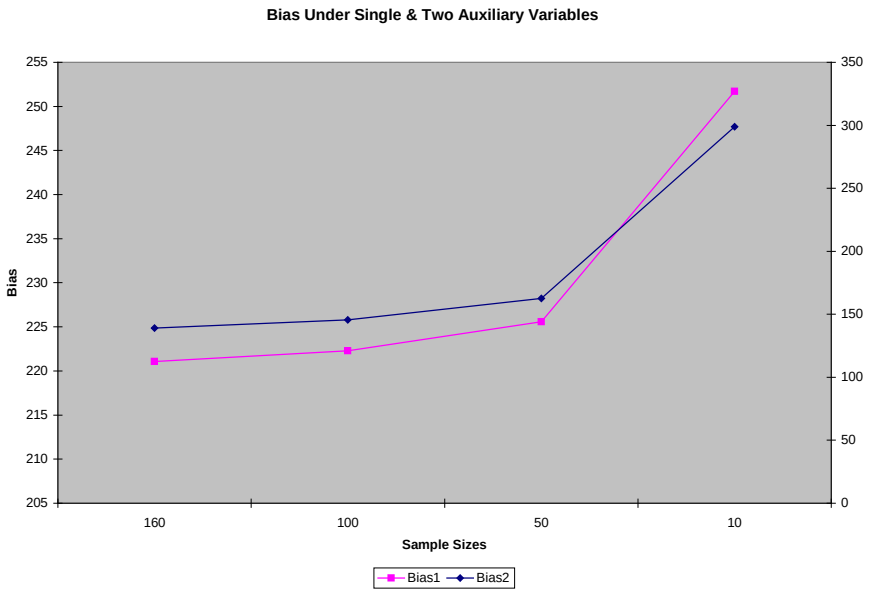
TIKKIWAL, G. C. and GHIYA, ALKA., (2000). A Generalized Class of Synthetic Estimators with Application to Crop Acreage Estimation for Small Domains. Biometrical Journal, 42 (7), 865-876.

TIKKIWAL, G. C. and PANDEY, K. K., (2007). On Some Aspects of Small Area Estimation Using Auxiliary Information. Ph.D. Thesis under supervision of Prof. G. C. Tikkiwal, Head of Department of Mathematics and Statistics J. N. V. University Jodhpur Rajasthan.

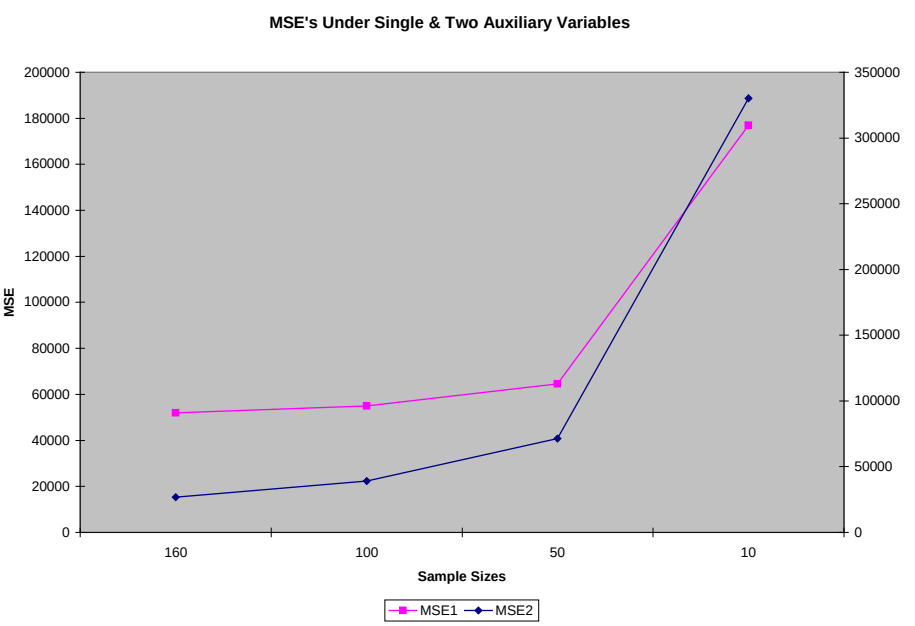
APPENDICES

FIGURES

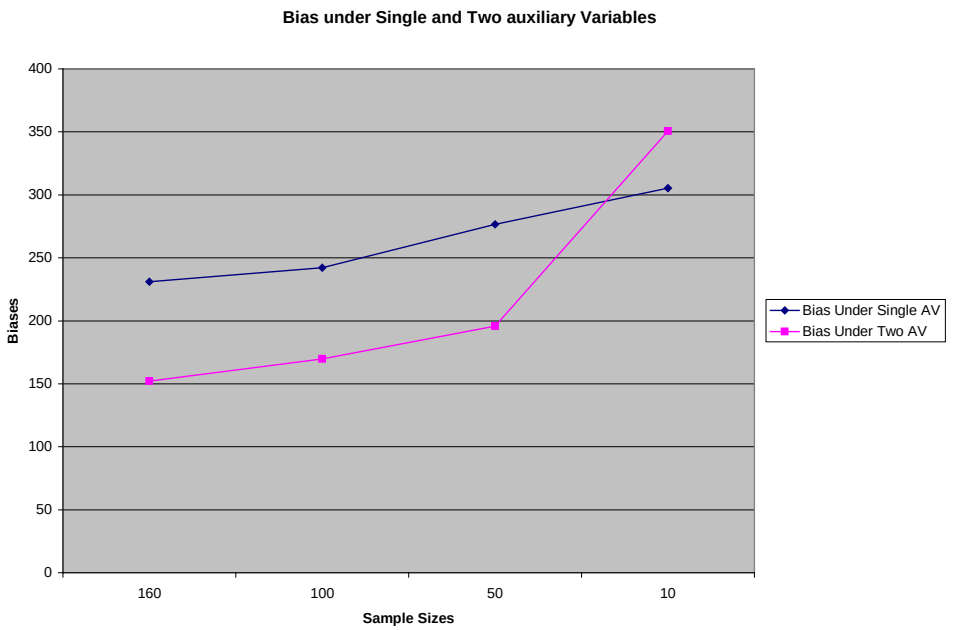
Graph 1.



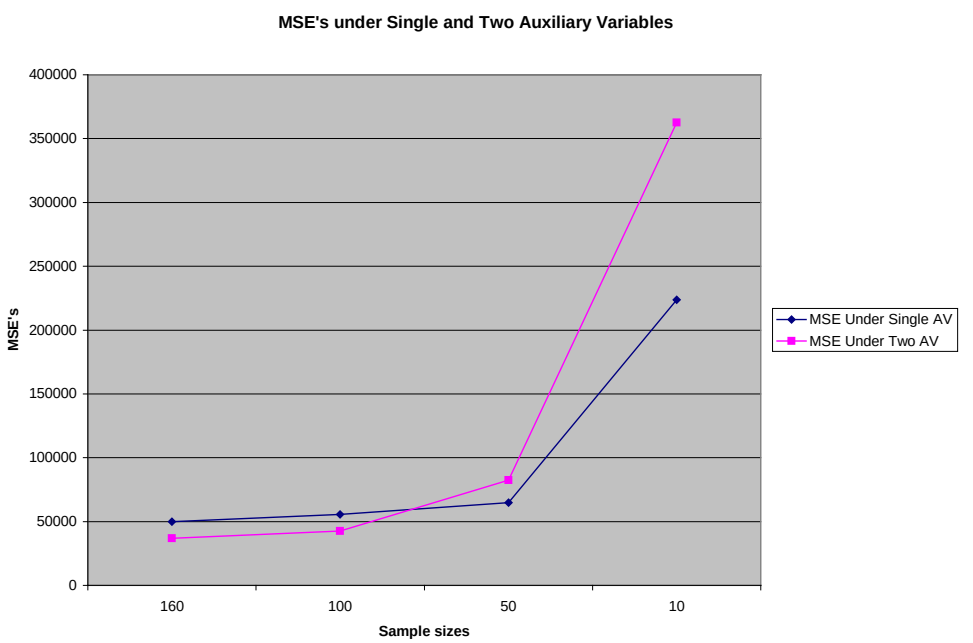
Graph 2.



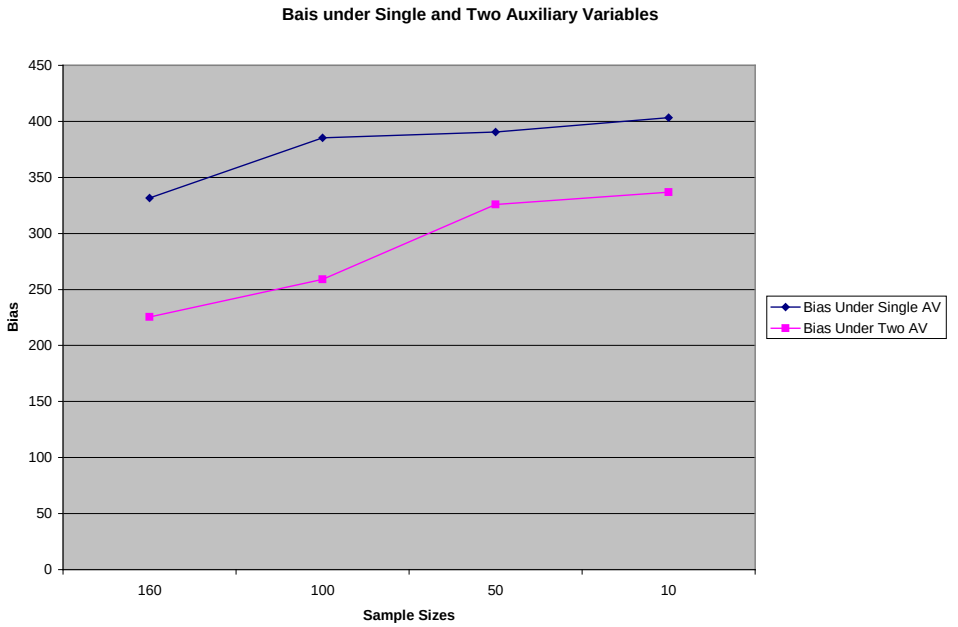
Graph 3.



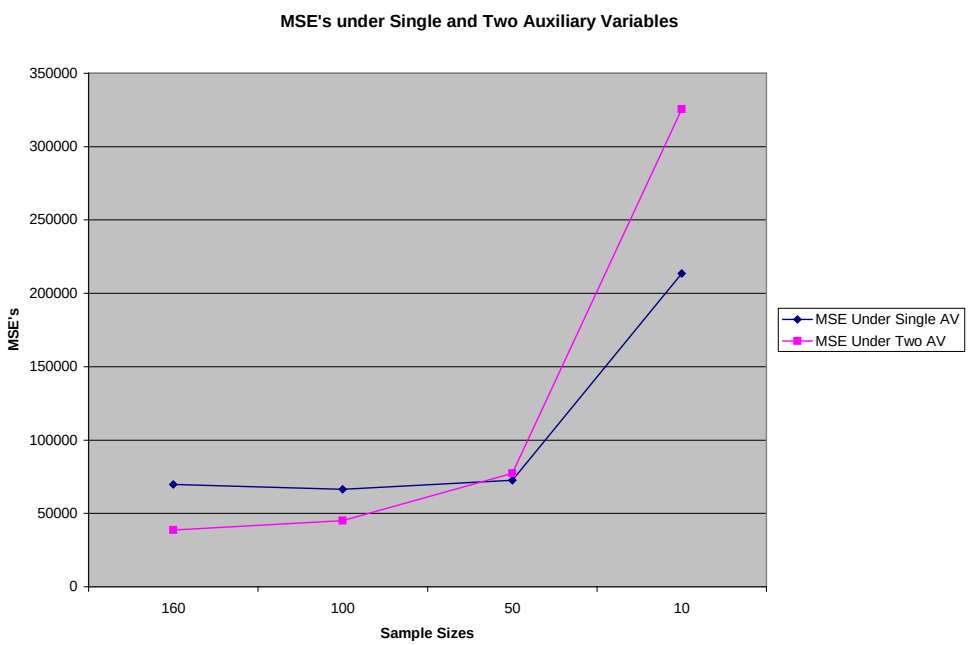
Graph 4.



Graph 5.



Graph 6.



TABLES

Table No. 6.1
Coefficient of Variation involved in Computation

C_0^2	**2.54**
C_1^2	**3.34**
C_2^2	**2.81**
C_{01}	**2.81**
C_{02}	**1.56**
C_{12}	**1.87**

Table No. 6.2
Bias of Synthetic Ratio Estimator under Single and Two Auxiliary Variables

Sample Sizes	Bias Under Single AV	Bias Under Two AV
160	221.07	139.14
100	222.30	145.53
50	225.57	162.56
10	251.71	298.79

Table No. 6.3
MSE of Synthetic Ratio Estimator under Single and Two Auxiliary Variables

Sample Sizes	MSE Under Single AV	MSE Under Two AV
160	51908.99	26914.89
100	54960.31	39043.23
50	64560.66	71385.74
10	176930.48	330125.22

Table No. 6.4
Bias of Synthetic Ratio Estimator under Single and Two Auxiliary Variables

Sample Sizes	Bias Under Single AV	Bias Under Two AV
160	231.13	152.10
100	242.29	169.76
50	276.56	195.83
10	305.31	350.77

Table No. 6.5
MSE of Synthetic Ratio Estimator under Single and Two Auxiliary Variables

Sample Sizes	MSE Under Single AV	MSE Under Two AV
160	50025.81	36980.87
100	55635.45	42565.40
50	64860.68	82344.91
10	223692.88	362512.46

Table No. 6.6
Bias of Synthetic Ratio Estimator under Single and Two Auxiliary Variables

Sample Sizes	Bias Under Single AV	Bias Under Two AV
160	331.74	225.37
100	385.35	258.88
50	390.56	325.83
10	403.30	336.73

Table No. 6.7
MSE of Synthetic Ratio Estimator under Single and Two Auxiliary Variables

Sample Sizes	MSE Under Single AV	MSE Under Two AV
160	69871.32	38565.61
100	66523.33	45021.33
50	72563.24	77452.26
10	213564.13	325613.65

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 45–56

A TWO-PARAMETER LINDLEY DISTRIBUTION

R. Shanker[1], **A. Mishra**[2]

ABSTRACT

A two-parameter Lindley distribution, of which the Lindley distribution (LD) is a particular case, has been introduced. Its moments, failure rate function, mean residual life function and stochastic orderings have been discussed. The maximum likelihood method and the method of moments have been discussed for estimating its parameters. The distribution has been fitted to some data-sets to test its goodness of fit.

Key words: Lindley distribution, moments, failure rate function, mean residual life function, stochastic ordering, estimation of parameters, goodness of fit.

1. Introduction

Lindley (1958) introduced a one-parameter distribution, known as Lindley distribution, given by its probability density function

$$f(x;\theta) = \frac{\theta^2}{\theta+1}(1+x)\,e^{-\theta x}; \qquad x>0, \quad \theta>0 \tag{1.1}$$

It can be seen that this distribution is a mixture of exponential (θ) and gamma $(2,\theta)$ distributions. Its cumulative distribution function has been obtained as

$$F(x) = 1 - \frac{\theta+1+\theta x}{\theta+1}\,e^{-\theta x};\ x>0,\ \theta>0 \tag{1.2}$$

Ghitany et al (2008) have discussed various properties of this distribution and showed that in many ways (1.1) it provides a better model for some applications than the exponential distribution. The first four moments about origin of the Lindley distribution have been obtained as

[1] College of Business and Economics, Asmara, Eritrea (N-E Africa). E-mail: shankerrama2009@gmail.com.
[2] Department of Statistics, Patna University, Patna, (India). E-mail: mishraamar@rediffmail.com.

$$\mu_1' = \frac{\theta+2}{\theta(\theta+1)}, \quad \mu_2' = \frac{2(\theta+3)}{\theta^2(\theta+1)}, \quad \mu_3' = \frac{6(\theta+4)}{\theta^3(\theta+1)}, \quad \mu_4' = \frac{24(\theta+5)}{\theta^4(\theta+1)} \tag{1.3}$$

and its central moments have been obtained as

$$\mu_2 = \frac{\theta^2+4\theta+2}{\theta^2(\theta+1)^2}, \quad \mu_3 = \frac{2(\theta^3+6\theta^2+6\theta+2)}{\theta^3(\theta+1)^3}, \quad \mu_4 = \frac{3(3\theta^4+24\theta^3+44\theta^2+32\theta+8)}{\theta^4(\theta+1)^4} \tag{1.4}$$

Ghitany et al (2008) studied various properties of this distribution. A discrete version of this distribution has been suggested by Deniz and Ojeda (2011) having its applications in count data related to insurance. Sankaran (1970) obtained the Lindley mixture of Poisson distribution. Mazucheli and Achcar (2011), Ghitany et al (2009, 2011) and Bakouchi et al (2012) are some among others who discussed its various applications. Zakerzadah and Dolati (2009) obtained a generalized Lindley distribution and discussed its various properties and applications.

In this paper, a two parameter Lindley distribution, of which the Lindley distribution (1.1) is a particular case, has been suggested. Its first four moments and some of the related measures have been obtained. Its failure rate, mean residual rate and stochastic ordering have also been studied. Estimation of its parameters has been discussed and the distribution has been fitted to some of those data sets where the Lindley distribution has earlier been fitted by others, to test its goodness of fit.

2. A Two-parameter Lindley distribution

A two-parameter Lindley distribution with parameters α and θ is defined by its probability density function (p.d.f)

$$f(x;\alpha,\theta) = \frac{\theta^2}{\alpha\theta+1}(\alpha+x)e^{-\theta x}; \; x>0, \; \theta>0, \; \alpha\theta>-1 \tag{2.1}$$

It can easily be seen that at $\alpha=1$, the distribution (2.1) reduces to the Lindley distribution (1.1) and at $\alpha=0$, it reduces to the gamma distribution with parameters $(2,\theta)$. The p.d.f. (2.1) can be shown as a mixture of exponential (θ) and gamma $(2,\theta)$ distributions as follows:

$$f(x;\alpha,\theta) = pf_1(x) + (1-p) f_2(x) \tag{2.2}$$

where $p = \dfrac{\alpha\theta}{\alpha\theta+1}$, $f_1(x) = \theta e^{-\theta x}$ and $f_2(x) = \theta^2 x e^{-\theta x}$.

The first derivative of (2.1) is obtained as

$$f'(x) = \frac{\theta^2}{\alpha\theta+1}(1-\alpha\theta-x\theta)\, e^{-\theta x}$$

and so $f'(x) = 0$ gives $x = \dfrac{1-\alpha\theta}{\theta}$. From this it follows that

(i) for $|\alpha\theta| < 1$, $x_0 = \dfrac{1-\alpha\theta}{\theta}$ is the unique critical point at which $f(x)$ is maximum.

(ii) for $\alpha \geq 1$, $f'(x) \leq 0$ i.e. $f(x)$ is decreasing in x.

Therefore, the mode of the distribution is given by

$$\text{Mode} = \begin{cases} \dfrac{1-\alpha\theta}{\theta}, & |\alpha\theta| < 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.3}$$

The cumulative distribution function of the distribution is given by

$$F(x) = 1 - \frac{1+\alpha\theta+\theta x}{\alpha\theta+1} e^{-\theta x};\ x > 0, \theta > 0, \alpha\theta > -1 \tag{2.4}$$

3. Moments and related measures

The *r*th moment about origin of the two-parameter Lindley distribution has been obtained as

$$\mu_r' = \frac{\Gamma(r+1)(\alpha\theta+r+1)}{\theta^r(\alpha\theta+1)};\ r = 1,2,\ldots \tag{3.1}$$

Taking $r = 1,2,3$ and 4 in (3.1), the first four moments about origin are obtained as

$$\mu_1' = \frac{\alpha\theta+2}{\theta(\alpha\theta+1)},\ \mu_2' = \frac{2(\alpha\theta+3)}{\theta^2(\alpha\theta+1)},\ \mu_3' = \frac{6(\alpha\theta+4)}{\theta^3(\alpha\theta+1)},\ \mu_4' = \frac{24(\alpha\theta+5)}{\theta^4(\alpha\theta+1)} \tag{3.2}$$

It can be easily verified that for $\alpha = 1$, the moments about origin of the distribution reduce to the respective moments of the Lindley distribution. Further, the mean of the distribution is always greater than the mode, the distribution is positively skewed. The central moments of this distribution have thus been obtained as

$$\mu_2 = \frac{\alpha^2\theta^2 + 4\alpha\theta + 2}{\theta^2(\alpha\theta+1)^2}, \tag{3.3}$$

$$\mu_3 = \frac{2\left(\alpha^3\theta^3 + 6\alpha^2\theta^2 + 6\alpha\theta + 2\right)}{\theta^3(\alpha\theta+1)^3}, \tag{3.4}$$

$$\mu_4 = \frac{3\left(3\alpha^4\theta^4 + 24\alpha^3\theta^3 + 44\alpha^2\theta^2 + 32\alpha\theta + 8\right)}{\theta^4(\alpha\theta+1)^4} \tag{3.5}$$

It can be easily verified that for $\alpha = 1$, the central moments of the distribution reduce to the respective moments of the Lindley distribution.

The coefficients of variation (γ), skewness $\left(\sqrt{\beta_1}\right)$ and the kurtosis (β_2) of the two-parameter LD are given by

$$\gamma = \frac{\sigma}{\mu_1'} = \frac{\sqrt{\alpha^2\theta^2 + 4\alpha\theta + 2}}{\alpha\theta + 2} \tag{3.6}$$

$$\sqrt{\beta_1} = \frac{2\left(\alpha^3\theta^3 + 6\alpha^2\theta^2 + 6\alpha\theta + 2\right)}{\left(\alpha^2\theta^2 + 4\alpha\theta + 2\right)^{3/2}} \tag{3.7}$$

$$\beta_2 = \frac{3\left(3\alpha^4\theta^4 + 24\alpha^3\theta^3 + 44\alpha^2\theta^2 + 32\alpha\theta + 8\right)}{\left(\alpha^2\theta^2 + 4\alpha\theta + 2\right)^2} \tag{3.8}$$

4. Failure rate and mean residual life

For a continuous distribution with p.d.f. $f(x)$ and c.d.f. $F(x)$, the failure rate function (also known as the hazard rate function) and the mean residual life function are respectively defined as

$$h(x) = \lim_{\Delta x \to 0} \frac{P\left(X < x + \Delta x \middle| X > x\right)}{\Delta x} = \frac{f(x)}{1 - F(x)} \tag{4.1}$$

$$\text{and} \quad m(x) = E\left[X - x \middle| X > x\right] = \frac{1}{1-F(x)} \int_x^\infty \left[1 - F(t)\right] dt \tag{4.2}$$

The corresponding failure rate function $h(x)$ and the mean residual life function $m(x)$ of the distribution are thus given by

$$h(x) = \frac{\theta^2(\alpha + x)}{1 + \alpha\theta + x\theta} \tag{4.3}$$

$$\text{and } m(x) = \frac{1}{(1+\alpha\theta+\theta x)e^{-\theta x}} \int_x^\infty (1+\alpha\theta+\theta t) e^{-\theta t} dt = \frac{2+\alpha\theta+\theta x}{\theta(1+\alpha\theta+\theta x)} \tag{4.4}$$

It can be easily verified that $h(0) = \frac{\theta^2\alpha}{\alpha\theta+1} = f(0)$ and $m(0) = \frac{\alpha\theta+2}{\theta(\alpha\theta+1)} = \mu_1'$ .It is also obvious that $h(x)$ is an increasing function of x, α and θ, whereas $m(x)$ is a decreasing function of x, α and increasing function of θ. For $\alpha = 1$, (4.3) and (4.4) reduce to the corresponding measures of the Lindley distribution. The failure rate function and the mean residual life function of the distribution show its flexibility over Lindley distribution and exponential distribution.

5. Stochastic orderings

Stochastic ordering of positive continuous random variables is an important tool for judging the comparative behaviour. A random variable X is said to be smaller than a random variable Y in the

(i) stochastic order $(X \leq_{st} Y)$ if $F_X(x) \geq F_Y(x)$ for all x

(ii) hazard rate order $(X \leq_{hr} Y)$ if $h_X(x) \geq h_Y(x)$ for all x

(iii) mean residual life order $(X \leq_{mrl} Y)$ if $m_X(x) \leq m_Y(x)$ for all x

(iv) likelihood ratio order $(X \leq_{lr} Y)$ if $\frac{f_X(x)}{f_Y(x)}$ decreases in x.

The following results due to Shaked and Shanthikumar (1994) are well known for establishing stochastic ordering of distributions

$$X \leq_{lr} Y \Rightarrow X \leq_{hr} Y \Rightarrow X \leq_{mrl} Y \tag{5.1}$$
$$\Downarrow$$
$$X \leq_{st} Y$$

The two-parameter LD is ordered with respect to the strongest 'likelihood ratio' ordering as shown in the following theorem:

Theorem: Let $X \sim$ two-parameter LD(α_1, θ_1) and $Y \sim$ two-parameter LD(α_2, θ_2). If $\alpha_1 = \alpha_2$ and $\theta_1 \geq \theta_2$ (or if $\theta_1 = \theta_2$ and $\alpha_1 \geq \alpha_2$), then $X \leq_{lr} Y$ and hence $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$.

Proof: We have

$$\frac{f_X(x)}{f_Y(x)} = \left(\frac{\theta_1}{\theta_2}\right)^2 \left(\frac{\alpha_2\theta_2 + 1}{\alpha_1\theta_1 + 1}\right)\left(\frac{\alpha_1 + x}{\alpha_2 + x}\right) e^{-(\theta_1 - \theta_2)x} \quad ; \; x > 0$$

Now

$$\log \frac{f_X(x)}{f_Y(x)} = 2\log\left(\frac{\theta_1}{\theta_2}\right) + \log\left(\frac{\alpha_2\theta_2 + 1}{\alpha_1\theta_1 + 1}\right) + \log(\alpha_1 + x) - \log(\alpha_2 + x) - (\theta_1 - \theta_2)x.$$

Thus

$$\frac{d}{dx}\log \frac{f_X(x)}{f_Y(x)} = \frac{1}{\alpha_1 + x} - \frac{1}{\alpha_2 + x} + (\theta_2 - \theta_1)$$

$$= \frac{\alpha_2 - \alpha_1}{(\alpha_1 + x)(\alpha_2 + x)} + (\theta_2 - \theta_1) \tag{5.2}$$

Case (i). If $\alpha_1 = \alpha_2$ and $\theta_1 \geq \theta_2$, then $\frac{d}{dx}\log \frac{f_X(x)}{f_Y(x)} < 0$. This means that $X \leq_{lr} Y$ and hence $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$.

Case (ii). If $\theta_1 = \theta_2$ and $\alpha_1 \geq \alpha_2$, then $\frac{d}{dx}\log \frac{f_X(x)}{f_Y(x)} < 0$. This means that $X \leq_{lr} Y$ and hence $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$.

This theorem shows the flexibility of two-parameter LD over Lindley and exponential distributions.

6. Estimation of parameters

6.1. Maximum likelihood estimates

Let $x_1, x_2, \ldots, x_n$ be a random sample of size *n* from a two-parameter Lindley distribution (2.1) and let f_x be the observed frequency in the sample corresponding to $X = x$ $(x = 1, 2, ..., k)$ such that $\sum_{x=1}^{k} f_x = n$, where k is the

largest observed value having non-zero frequency. The likelihood function, L of the two-parameter Lindley distribution (2.1) is given by

$$L = \left(\frac{\theta^2}{\alpha\theta + 1} \right)^n \prod_{x=1}^{k} (\alpha + x)^{f_x} e^{-n\theta\bar{X}} \tag{6.1.1}$$

and so the log likelihood function is obtained as

$$\log L = n \log \theta^2 - n \log(\alpha\theta + 1) + \sum_{i=1}^{k} f_x \log(\alpha + x) - n\theta\bar{X} \tag{6.1.2}$$

The two log likelihood equations are thus obtained as

$$\frac{\partial \log L}{\partial \theta} = \frac{2n}{\theta} - \frac{n\alpha}{\alpha\theta + 1} - n\bar{X} = 0 \tag{6.1.3}$$

$$\frac{\partial \log L}{\partial \alpha} = -\frac{n\theta}{\alpha\theta + 1} + \sum_{x=1}^{k} \frac{f_x}{\alpha + x} = 0 \tag{6.1.4}$$

Equation (6.1.3) gives $\bar{X} = \dfrac{\alpha\theta + 2}{\theta(\alpha\theta + 1)}$, which is the mean of the two-parameter Lindley distribution. The two equations (6.1.3) and (6.1.4) do not seem to be solved directly. However, the Fisher's scoring method can be applied to solve these equations. We have

$$\frac{\partial^2 \log L}{\partial \theta^2} = -\frac{2n}{\theta^2} + \frac{n\alpha^2}{(\alpha\theta + 1)^2} \tag{6.1.5}$$

$$\frac{\partial^2 \log L}{\partial \theta \, \partial \alpha} = -\frac{n}{(\alpha\theta + 1)^2} \tag{6.1.6}$$

$$\frac{\partial^2 \log L}{\partial \alpha^2} = \frac{n\theta^2}{(\alpha\theta + 1)^2} - \sum_{x=1}^{k} \frac{f_x}{(\alpha + x)^2} \tag{6.1.7}$$

The following equations for $\hat{\theta}$ and $\hat{\alpha}$ can be solved

$$\begin{bmatrix} \dfrac{\partial^2 \log L}{\partial \theta^2} & \dfrac{\partial^2 \log L}{\partial \theta \partial \alpha} \\ \dfrac{\partial^2 \log L}{\partial \theta \partial \alpha} & \dfrac{\partial^2 \log L}{\partial \alpha^2} \end{bmatrix}_{\substack{\hat{\theta}=\theta_0 \\ \hat{\alpha}=\alpha_0}} \begin{bmatrix} \hat{\theta} - \theta_0 \\ \hat{\alpha} - \alpha_0 \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \log L}{\partial \theta} \\ \dfrac{\partial \log L}{\partial \alpha} \end{bmatrix}_{\substack{\hat{\theta}=\theta_0 \\ \hat{\alpha}=\alpha_0}} \tag{6.1.8}$$

where θ_0 and α_0 are the initial values of θ and α respectively. These equations are solved iteratively till sufficiently close estimates of $\hat{\theta}$ and $\hat{\alpha}$ are obtained.

6.2. Estimates from moments

Using the first two moments about origin, we have

$$\frac{\mu_2'}{\mu_1'^2} = k \text{ (say)} = \frac{2(\alpha\theta+3)(\alpha\theta+1)}{(\alpha\theta+2)^2} \tag{6.2.1}$$

Taking $b = \alpha\theta$, we get

$$\frac{\mu_2'}{\mu_1'^2} = \frac{2(b+3)(b+1)}{(b+2)^2} = \frac{2b^2+8b+6}{b^2+4b+4} = k \ .$$

$$\text{This gives} \quad (2-k)b^2 + 4(2-k)b + 2(3-2k) = 0 \tag{6.2.2}$$

which is a quadratic equation in b .Replacing the first and the second moments μ_1' and μ_2' by the respective sample moments $\bar{X}$ and m_2' an estimate of k can be obtained, using which , the equation (6.2.2) can be solved and an estimate of b obtained. Substituting this estimate of b in the expression for the mean of the two-parameter LD, an estimate of θ can be obtained as

$$\hat{\theta} = \left(\frac{b+2}{b+1}\right)\frac{1}{\bar{X}} \tag{6.2.3}$$

Finally to get an estimate of α , we substitute the value b and estimate of θ in the expression $b = \alpha\theta$, which gives an estimate of α as

$$\hat{\alpha} = \frac{b}{\hat{\theta}} \tag{6.2.4}$$

7. Goodness of fit

The two-parameter Lindley distribution has been fitted to a number of data-sets to which earlier the Lindley distribution has been fitted by others and to almost all these data-sets the two-parameter Lindley distribution provides closer fits than the one parameter Lindley distribution.

The fittings of the two-parameter Lindley distribution to three such data-sets have been presented in the following tables. The data sets given in tables-1, 2 and 3 are the data sets reported by Ghitany et al (2008), Bzerkedal (1960) and Paranjpe and Rajarshi (1986) respectively. The expected frequencies according to the one parameter Lindley distribution have also been given for ready comparison with those obtained by the two-parameter Lindley distribution. The estimates of the parameters have been obtained by the method of moments.

Table 1. Waiting times (in minutes) of 100 bank customers

Waiting Time (In minutes)	Observed frequency	Expected frequency	
		One-parameter LD	Two-parameter LD
0 – 5	30	30.4	30.2
5 – 10	32	30.7	30.9
10 – 15	19	19.2	19.3
15 – 20	10	10.3	10.3
20 – 25	5	5.1	5.0
25 – 30	1	2.4	2.4
30 – 35	2	1.1	1.1
35 – 40	1	0.8	0.8
Total	100	100.0	100.0
Estimates of parameters		$\hat{\theta} = 0.187$	$\hat{\theta} = 0.191139$ $\alpha = 0.894052$
	χ^2	0.09402	0.07481
	d.f.	4	3

Table 2. Data of survival times (in days) of 72 guinea pigs infected with virulent tubercle bacilli

Survival Time (In days)	Observed frequency	Expected frequency	
		One-parameter LD	Two-parameter LD
0 – 80	8	16.1	10.7
80 – 160	30	21.9	26.9
160 – 240	18	15.4	17.7
240 – 320	8	9.0	9.2
320 – 400	4	5.5	4.3
400 – 480	3	1.8	1.9
480 – 560	1	2.3	1.3
Total	72	72.0	72.0
Estimates of parameters		$\hat{\theta} = 0.011$	$\hat{\theta} = 0.012992$ $\hat{\alpha} = -20.08359$
	χ^2	7.7712	1.2335
	d.f.	3	2

Table 3. Mortality grouped data for blackbird species

Survival Time (In days)	Observed frequency	Expected frequency	
		One-parameter LD	Two-parameter LD
0 – 1	192	173.5	168.0
1 – 2	60	98.6	88.4
2 – 3	50	46.5	46.2
3 – 4	20	20.1	24.0
4 – 5	12	8.1	12.4
5 – 6	7	3.2	6.4
6 – 7	6	1.4	3.3
7 – 8	3	0.3	1.7
≥ 8	2	0.3	1.6
Total	352	352.0	352.0
Estimates of parameters		$\hat{\theta} = 0.984$	$\hat{\theta} = 0.731104$ $\hat{\alpha} = 10.266582$
χ^2		49.846	16.5342
d.f.		4	4

It can be seen that the two-parameter LD gives much closer fits than the one parameter Lindley distribution and thus provides a better alternative to the Lindley distribution.

8. Conclusions

In this paper, a two-parameter Lindley distribution (LD), of which the one-parameter LD is a particular case, has been proposed. Several properties of the two-parameter LD such as moments, failure rate function, mean residual life function, stochastic orderings, estimation of parameters by the method of maximum likelihood and the method of moments have been discussed. Finally, the proposed distribution has been fitted to a number of data sets relating to waiting and survival times to test its goodness of fit to which earlier the one-parameter LD has been fitted, and it is found that two-parameter LD provides better fits than those by the one-parameter LD.

Acknowledgments

The authors express their gratitude to the referee for his valuable comments and suggestions which improved the quality of the paper.

REFERENCES

BAKOUCH, H. S., AL-ZAHRANI, B. M., AL-SHOMRANI, A. A., MARCHI, V. A. A., LOUZADA, F., (2012). An extended Lindley distribution, Journal of the Korean Statistical Society, Vol. 41 (1), 75-85.

BJERKEDAL, T., (1960). Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli, American Journal of Epidemiol, Vol. 72 (1), 130-148.

DENIZ, E. G., and OJEDA, E. C., (2011). The discrete Lindley distribution-Properties and Applications, Journal of Statistical Computation and Simulation, Vol. 81 (11), 1405-1416.

GHITANY, M. E., ATIEH, B., NADARAJAH, S., (2008). Lindley distribution and its Applications, Mathematics and Computers in Simulation, Vol. 78(4), 493-506.

GHITANY, M. E., AL-MUTAIRI, D. K., NADARAJAH, S., (2008). Zero-truncated Poisson-Lindley distribution and its Applications, Mathematics and Computers in Simulation, Vol. 79(3), 279-287.

GHITANY, M. E. and AL-MUTAIRI, D. K., (2008). Size-biased Poisson-Lindley distribution and its Applications, Metron-International Journal of Statistics, Vol. LXVI, no. 3, 299-311.

GHITANY, M. E. and AL-MUTAIRI, D. K., (2009). Estimation methods for the discrete Poisson- Lindley distribution, Journal of Statistical Computation and Simulation, Vol. 79 (1), 1-9.

GHITANY, M. E. AL-QALLAF, F., AL-MUTAIRI, D. K., HUSSAIN, H. A., (2011). A two parameter weighted Lindley distribution and its applications to survival data, Mathematics and Computers in Simulation, Vol. 81 (6), 1190-1201.

LINDLEY, D. V., (1958). Fiducial distributions and Bayes' theorem, Journal of the Royal Statistical Society, Series B, 20, 102-107.

LINDLEY, D. V., (1965). Introduction to Probability and Statistics from Bayesian viewpoint, part II, Inference, Cambridge university press, New York.

MAHMOUDI, E. and ZAKERZADAH, H., (2010). Generalized Poisson Lindley distribution, Communications in Statistics-Theory and Methods, 39, 1785-1798.

MAZUCHELI, J. and ACHCAR, J. A., (2011). The Lindley distribution applied to competing risks lifetime data, Computer Methods and Programs in Biomedicine, Vol. 104 (2), 188-192.

PARANJPE, S. and RAJARSHI, M. B., (1986). Modeling non-monotonic survivorship data with bath tube distributions, Ecology, Vol. 67 (6), 1693-1695.

SANKARAN, M., (1970). The discrete Poisson-Lindley distribution, Biometrics, 26, 145-149.

SHAKED, M. and SHANTHIKUMAR, J. G., (1994). Stochastic Orders and Their Applications, Academic Press, New York.

ZAKERZADAH, H. and DOLATI, A., (2009). Generalized Lindley distribution, Journal of Mathematical Extension, Vol. 3(2), 13-25.

ZAMANI, H. and ISMAIL, N., (2010). Negative Binomial- Lindley distribution and its Applications, Journal of Mathematics and Statistics, Vol. 6(1), 4-9.

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 57–74

BEST LINEAR UNBIASED ESTIMATORS OF POPULATION MEAN ON CURRENT OCCASION IN TWO-OCCASION ROTATION PATTERNS

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

ABSTRACT

Best linear unbiased estimators have been proposed to estimate the population mean on current occasion in two-occasion successive (rotation) sampling. Behavior of the proposed estimators have been studied and their respective optimum replacement policies are discussed. Empirical studies are carried out to examine the performance of the proposed estimators and consequently the suitable recommendations are made.

Key words: successive sampling, auxiliary information, unbiased, variance, optimum replacement policy.

1. Introduction

Often in sample surveys on successive occasions for the same population, the current or most recent estimates are of the primary interest if the characteristics of the population are likely to change rapidly over time. For example, monthly surveys are carried out to collect data on prices of goods to determine the consumer price index, labor force surveys are conducted on monthly basis to estimate the numbers of people in employment and industries, collect information at regular intervals to know popularity of their products, etc. In such studies, successive (rotation) sampling may be an impressive statistical tool to generate reliable and cost effective estimates of different population parameters on successive points of time (occasions) in chronological order. It also provides effective estimates of changing patterns over a period of time.

The problem of successive (rotation) sampling with a partial replacement of sampling units was initiated by Jessen (1942) in the analysis of agricultural survey data. He pioneered using the entire information collected during the previous investigations. The theory of successive (rotation) sampling was further

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

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 strategies for estimating the population mean on the current occasion using information on two auxiliary variables readily available on the 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 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-occasion successive sampling.

In many situations, information on an 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 is known in environmental surveys, nature of employment status, educational status, food availability and medical aids of a locality is well known in advance for estimating various demographic parameters in demographic surveys. Utilizing auxiliary information on both the occasions, Feng and Zou (1997), Biradar and Singh (2001), 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. Recently Singh and Vishwakarma (2009) have suggested a general estimation procedure for population mean in successive (rotation) sampling. Motivated with the above works and utilizing the information on an auxiliary variable, readily available on both the occasions, we have proposed best linear unbiased estimators for estimating the current population mean in two-occasion successive (rotation) sampling. Behaviors of the proposed estimators are examined through empirical means of comparison and subsequently the 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 (large) units which is assumed to remain unchanged over two occasions. Let x (y) be the character under study on the first (second) occasion respectively. It is assumed that the information on an auxiliary variable z (stable over occasion), is readily available for both the occasions, whose population mean is known and it is highly positively correlated to x and y on the first and second occasions respectively. A simple random sample (without replacement) of size n units is drawn on the first occasion and a random sub-sample of size $m = n\lambda$ units from the sample on the first occasion is retained (matched) for its use on the current (second) occasion. A fresh (un-matched) sample of size $u = (n-m) = n\mu$ units is drawn on the current occasion from the entire population by simple random sampling (without replacement) method so that the sample size on the current occasion is

also n. λ and μ ($\lambda+\mu=1$) are the fractions of the matched and fresh samples, respectively, on the current occasion. We consider the following notations for further use:

$\bar{X}$, $\bar{Y}$, $\bar{Z}$: Population means of the variables x, y and z respectively.

$\bar{x}_m$, $\bar{x}_n$, $\bar{y}_u$, $\bar{y}_m$, $\bar{z}_u$, $\bar{z}_m$, $\bar{z}_n$: Sample means of the respective variables based on the sample sizes shown in suffices.

ρ_{yx}, ρ_{yz}, ρ_{xz}: Correlation coefficients between the variables shown in suffices.

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

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

3. Formulation of the estimator

To estimate the population mean $\bar{Y}$ on the current (second) occasion, we consider the following minimum variance linear unbiased estimator of $\bar{Y}$, which is as follows:

$$T_1 = \{a_1\bar{y}_u + a_2\bar{y}_m\} + \{a_3\bar{x}_m + a_4\bar{x}_n\} + \{a_5\bar{z}_u + a_6\bar{z}_m + a_7\bar{z}_n + a_8\bar{Z}\} \qquad (1)$$

where a_1, a_2, a_3, a_4, a_5, a_6, a_7 and a_8 are constants to be determined so that

(i) T_1 becomes an unbiased estimator of $\bar{Y}$ and

(ii) the variance of T_1 attains a minimum value.

For unbiasedness condition, we must have

$(a_1+a_2) = 1$, $(a_3+a_4) = 0$ and $(a_5+a_6+a_7+a_8) = 0$.

Substituting $a_1 = \varphi_1$, $a_3 = \beta_1$ and $a_8 = -(a_5+a_6+a_7)$, the estimator T_1 defined in equation (1) reduces to the following form

$$\begin{aligned} T_1 &= \{\varphi_1\bar{y}_u + (1-\varphi_1)\bar{y}_m\} + \beta_1\{\bar{x}_m-\bar{x}_n\} + \{a_5(\bar{z}_u-\bar{Z}) + a_6(\bar{z}_m-\bar{Z}) + a_7(\bar{z}_n-\bar{Z})\} \\ &= \varphi_1[\bar{y}_u + k_1(\bar{z}_u-\bar{Z})] + (1-\varphi_1)[\bar{y}_m + k_2\{\bar{x}_m-\bar{x}_n\} + k_3(\bar{z}_m-\bar{Z}) + k_4(\bar{z}_n-\bar{Z})] \\ &= \varphi_1 T_{1u} + (1-\varphi_1)T_{1m} \end{aligned} \qquad (2)$$

where $T_{1u} = \bar{y}_u + k_1(\bar{z}_u-\bar{Z})$; an estimator based on the fresh sample of size u

and $T_{1m}=\bar{y}_m+k_2\left(\bar{x}_m-\bar{x}_n\right)+k_3\left(\bar{z}_m-\bar{Z}\right)+k_4\left(\bar{z}_n-\bar{Z}\right)$; an estimator based on the matched sample of size m, $k_1=\frac{a_5}{\varphi_1}$, $k_2=\frac{\beta_1}{1-\varphi_1}$, $k_3=\frac{a_6}{1-\varphi_1}$, $k_4=\frac{a_7}{1-\varphi_1}$ and φ_1 are the unknown constants to be determined under certain criterions.

Remark 3.1. For estimating the population mean on each occasion the estimator T_{1u} is suitable, which implies that more belief on T_{1u} could be shown by choosing φ_1 as 1 (or close to 1), while for estimating the change over the occasions, the estimator T_{1m} could be more useful and hence φ_1 might be chosen as 0 (or close to 0). For asserting both the problems simultaneously, the suitable (optimum) choice of φ_1 is desired.

4. Properties of the estimator T_1

T_1 is an unbiased estimator of $\bar{Y}$ whose variance, ignoring finite population corrections, is derived in the following theorem.

Theorem 4.1. Variance of the estimator T_1 is obtained as

$$V(T_1)=\varphi_1^2\,V\left(T_{1u}\right)+\left(1-\varphi_1\right)^2V\left(T_{1m}\right) \tag{3}$$

where $$V(T_{1u})=\frac{1}{u}\eta_1 S_y^2 \tag{4}$$

$$V(T_{1m})=\left[\frac{1}{m}\eta_2+\left(\frac{1}{m}-\frac{1}{n}\right)\eta_3+\frac{1}{n}\eta_4\right]S_y^2 \tag{5}$$

$$\eta_1=\left(1+k_1^2+2k_1\rho_{yz}\right),\ \eta_2=\left(1+k_3^2+2k_3\rho_{yz}\right),$$

$$\eta_3=\left(k_2^2+2k_2\rho_{yx}+2k_2k_3\rho_{xz}\right) \text{ and } \eta_4=\left(k_4^2+2k_4\rho_{yz}+2k_3k_4\right).$$

Proof: It is obvious that the variance of the estimator T_1 is given by

$$V(T_1)=E\left[T_1-\bar{Y}\right]^2=E\left[\varphi_1\left(T_{1u}-\bar{Y}\right)+\left(1-\varphi_1\right)\left(T_{1m}-\bar{Y}\right)\right]^2$$

$$=\varphi_1^2\,V(T_{1u})+\left(1-\varphi_1\right)^2V\left(T_{1m}\right)+2\varphi_1\left(1-\varphi_1\right)C_{11} \tag{6}$$

where

$$V\left(T_{1u}\right)=E\left[T_{1u}-\bar{Y}\right]^2,\ V(T_{1m})=E\left[T_{1m}-\bar{Y}\right]^2 \text{ and } C_{11}=E\left[\left(T_{1u}-\bar{Y}\right)\left(T_{1m}-\bar{Y}\right)\right]$$

Substituting the expressions of T_{1u} and T_{1m} in equation (6), taking expectations and ignoring finite population corrections, we have the expression of the variance of the estimator T_1 as given in equation (3).

Remark 4.1. Results in equation (3) are derived under the assumption that the population mean squares of the variables x, y and z are almost equal.

Remark 4.2. T_{1u} and T_{1m} are based on two independent samples of sizes u and m respectively and they are unbiased estimators of $\overline{Y}$, hence the covariance term C_{11} between T_{1u} and T_{1m} vanishes.

5. Minimum variance of the estimator T_1

Since the variance of the estimator T_1 in equation (3) is the function of the unknown constants k_1, k_2, k_3, k_4 and φ_1, therefore it is minimized with respect to these constants, and subsequently the optimum values of k_1, k_2, k_3, k_4 and φ_1 are obtained as

$$k_1^* = -\rho_{yz} \tag{7}$$

$$k_2^* = \frac{\rho_{yz}\rho_{xz} - \rho_{yx}}{1-\rho_{xz}^2} \tag{8}$$

$$k_3^* = \frac{\rho_{yx}\rho_{xz} - \rho_{yz}}{1-\rho_{xz}^2} \tag{9}$$

$$k_4^* = \frac{\rho_{xz}\left(\rho_{yz}\rho_{xz} - \rho_{yx}\right)}{1-\rho_{xz}^2} \tag{10}$$

$$\varphi_{1_{opt}} = \frac{V(T_{1m})}{V(T_{1u}) + V(T_{1m})} \tag{11}$$

Substituting the values of k_1^*, k_2^*, k_3^* and k_4^* in equations (4) and (5), we get the optimum variances of T_{1u} and T_{1m} as

$$V(T_{1u})_{opt} = \frac{1}{u} A_1 S_y^2 \tag{12}$$

$$V(T_{1m})_{opt} = \left[\frac{1}{m}A_2 + \left(\frac{1}{m} - \frac{1}{n}\right)A_3 + \frac{1}{n}A_4\right] S_y^2 \tag{13}$$

where $A_1 = 1-\rho_{yz}^2$, $A_2 = \dfrac{1+\rho_{xz}^2\left(\rho_{xz}^2 - 2 + \rho_{yx}^2 + 2\rho_{yz}^2 - 2\rho_{yz}\rho_{yx}\rho_{xz}\right) - \rho_{yz}^2}{\left(1-\rho_{xz}^2\right)^2}$, $A_3 = -k_2^{*2}$

and $A_4 = -k_4^{*2}$.

Further, substituting the values of $V(T_{1u})_{opt}$ and $V(T_{1m})_{opt}$ from equations (12) and (13) in equation (11), we get the optimum value φ_{1opt} with respect to k_1^*, k_2^*, k_3^* and k_4^* as

$$\varphi_{1opt}^* = \frac{V(T_{1m})_{opt}}{V(T_{1u})_{opt}+V(T_{1m})_{opt}} \tag{14}$$

Again from equation (14) substituting the value of φ_{1opt}^* in equation (3), we get the optimum variance of T_1 as

$$V(T_1)_{opt} = \frac{V(T_{1m})_{opt}.V(T_{1u})_{opt}}{V(T_{1u})_{opt}+V(T_{1m})_{opt}} \tag{15}$$

Further, substituting the values from equations (12) and (13) in equations (14) and (15), the simplified values of φ_{1opt}^* and $V(T_1)_{opt}$ are obtained as

$$\varphi_{1opt}^* = \left[\frac{\mu_1(A_5+\mu_1 A_6)}{A_1+\mu_1 A_7+\mu_1^2 A_6}\right] \tag{16}$$

$$V(T_1)_{opt} = \frac{1}{n}\left[\frac{A_1(A_5+\mu_1 A_6)}{A_1+\mu_1 A_7+\mu_1^2 A_6}\right]S_y^2 \tag{17}$$

where $A_5 = A_2+A_4$, $A_6 = A_3-A_4$, $A_7 = A_5-A_1$ and μ_1 is the fraction of fresh sample for the estimator T_1.

6. Optimum replacement policy

To determine the optimum value of μ_1 (fraction of a sample to be drawn afresh on the current occasion) so that the population mean $\bar{Y}$ may be estimated with the maximum precision, we minimize the $V(T_1)_{opt}$ given in equation (17) with respect to μ_1, which result in a quadratic equation in μ_1 and respective solutions of μ_1 say μ_1^0 is given below:

$$Q_1\mu_1^2+2Q_2\mu_1+Q_3=0 \tag{18}$$

$$\mu_1^0=\frac{-Q_2\pm\sqrt{Q_2^2-Q_1Q_3}}{Q_1} \tag{19}$$

where $Q_1 = A_6^2$, $Q_2 = A_5A_6$ and $Q_3 = A_7A_5-A_1A_6$.

From equation (19), it is obvious that the real values of μ_1^0 exist if the quantity under square root is greater than or equal to zero. Two real values of μ_1^0 are possible. Hence, while choosing the value of μ_1^0, it should be remembered that $0 \le \mu_1^0 \le 1$. All other values of μ_1^0 are inadmissible. Substituting the admissible value of μ_1^0 say $\hat{\mu}_1$ from equation (19) into equation (17), we have the optimum value of $V(T_1)_{opt}$ as

$$V(T_1^0)_{opt} = \frac{1}{n}\left[\frac{A_1(A_5+\hat{\mu}_1 A_6)}{A_1+\hat{\mu}_1 A_7+\hat{\mu}_1^2 A_6}\right] S_y^2 \tag{20}$$

7. Efficiency comparison

To study the performance of the estimator T_1 the percent relative efficiencies of the estimator T_1 with respect to (i) $\overline{y}_n$, when there is no matching, and (ii) the estimator T_2, when no auxiliary information is used at any occasion, have been computed for different choices of correlations. The estimator T_2 is defined under the same circumstances as the estimator T_1, but in the absence of the auxiliary variable z on both the occasions and proposed as

$$T_2 = \{b_1\overline{y}_u + b_2\overline{y}_m\} + \{b_3\overline{x}_m + b_4\overline{x}_n\} \tag{21}$$

where b_1, b_2, b_3 and b_4 are constants to be determined so that

(i) T_2 becomes an unbiased estimator of $\overline{Y}$ and

(ii) The variance of T_2 attains a minimum value.

For unbiasedness condition, we must have $(b_1+b_2) = 1$ and $(b_3+b_4) = 0$.

Substituting $b_1 = \varphi_2$ and $b_3 = \beta_2$, the estimator T_2 defined in equation (21) reduces to the following form

$$\begin{aligned} T_2 &= \{\varphi_2\overline{y}_u + (1-\varphi_2)\overline{y}_m\} + \beta_2\{\overline{x}_m - \overline{x}_n\} \\ &= \varphi_2\overline{y}_u + (1-\varphi_2)[\overline{y}_m + k_5\{\overline{x}_m - \overline{x}_n\}] \\ &= \varphi_2 T_{2u} + (1-\varphi_2)T_{2m} \end{aligned} \tag{22}$$

where $T_{2u} = \overline{y}_u$; an estimator based on the fresh sample of size u

and $T_{2m} = \bar{y}_m + k_5(\bar{x}_m - \bar{x}_n)$; an estimator based on the matched sample of size m, $k_5 = \dfrac{\beta_2}{1-\varphi_2}$ and φ_2 are the unknown constants to be determined in such a way that they minimize the variance of the estimator T_2. Following the methods discussed in Sections 4, 5 and 6, the optimum values of k_5, μ_2 (fraction of fresh sample for the estimator T_2), variance of $\bar{y}_n$ and optimum variance of T_2 for large N are given by

$$k_5^* = -\rho_{yx} \tag{23}$$

$$\hat{\mu}_2 = \frac{1 \pm \sqrt{1-\rho_{yx}^2}}{\rho_{yx}^2} \tag{24}$$

$$V(\bar{y}_n) = \frac{S_y^2}{n} \tag{25}$$

$$V(T_2^0)_{opt} = \frac{1}{n}\left[\frac{1 - \hat{\mu}_2\rho_{yx}^2}{1 - \hat{\mu}_2^2\rho_{yx}^2}\right]S_y^2 \tag{26}$$

For different choices of ρ_{yx}, ρ_{xz} and ρ_{yz}, Table 1 shows the optimum values of μ_1 and percent relative efficiencies E_1 and E_2 of the estimator T_1 with respect to the estimators $\bar{y}_n$ and T_2 respectively, where

$$E_1 = \frac{V(\bar{y}_n)}{V(T_1^0)_{opt}} \times 100 \quad \text{and} \quad E_2 = \frac{V(T_2^0)_{opt}}{V(T_1^0)_{opt}} \times 100.$$

8. Analysis of results for estimator T_1

The following conclusions can be read out from Table 1.

(a) For fixed values of ρ_{xz} and $\rho_{yz,}$ the values of μ_1 and E_1 are increasing with the increasing values of ρ_{yx}. The values of E_2 are decreasing for the lower values of ρ_{yx} while increasing pattern may be seen for the higher values of $\rho_{yx.}$

(b) For fixed values of ρ_{xz} and $\rho_{yx,}$ the values of μ_1 are decreasing with the increasing values of $\rho_{yz..}$ Values of E_1 and E_2 are increasing with the increasing values of $\rho_{yz..}$This behavior is highly desirable, since it concludes that if highly

correlated auxiliary variable is available, it pays in terms of enhance precision of the estimates as well as it reduces the cost of the survey.

(c) For fixed values of ρ_{yz} and $\rho_{yx,}$ the values of μ_1 are decreasing with the increasing values of ρ_{xz} . Similar patterns are visible for the efficiencies E_1 and E_2.

(d) Minimum value of μ_1 is 0.4329, which indicates that only 43 percent of the total sample size is to be replaced on the current occasion for the corresponding choices of the correlations.

9. Use of auxiliary variable only at the current occasion

In section 3 we have formulated the estimator T_1 on the assumption that information on a stable auxiliary variable z was readily available on both the occasions. If the duration between two successive occasions is small then one may expect the stability of the auxiliary variable but the stability character of the auxiliary variable may not sustain if the duration between two successive occasions is appreciably large. In such situation it may not be wise to use the auxiliary information from the previous occasion. Motivated with the above argument, we formulate the estimator T_3 when the information on an auxiliary variable z is available only on the current (second) occasion. The estimator T_3 is formulated as

$$T_3 = \left\{c_1\bar{y}_u + c_2\bar{y}_m\right\} + \left\{c_3\bar{x}_m + c_4\bar{x}_n\right\} + \left\{c_5\bar{z}_u + c_6\bar{z}_m + c_7\bar{Z}\right\} \qquad (27)$$

where c_1, c_2, c_3, c_4, c_5, c_6 and c_7 are constants to be determined so that

(i) T_3 becomes an unbiased estimator of $\bar{Y}$ and

(ii) The variance of T_3 attains a minimum value.

For unbiasedness condition, we must have

$$\left(c_1+c_2\right) = 1,\ \left(c_3+c_4\right) = 0 \text{ and } \left(c_5+c_6+c_7\right) = 0.$$

Substituting $c_1 = \varphi_3$, $c_3 = \beta_3$ and $c_7 = -\left(c_5 + c_6\right)$, the estimator T_3 defined in equation (27) reduces to the following form

$$\begin{aligned} T_3 &= \left\{\varphi_3\bar{y}_u + \left(1-\varphi_3\right)\bar{y}_m\right\} + \beta_3\left\{\bar{x}_m - \bar{x}_n\right\} + \left\{c_5\left(\bar{z}_u - \bar{Z}\right) + c_6\left(\bar{z}_m - \bar{Z}\right)\right\} \\ &= \varphi_3\left[\bar{y}_u + l_1\left(\bar{z}_u - \bar{Z}\right)\right] + \left(1-\varphi_3\right)\left[\bar{y}_m + l_2\left\{\bar{x}_m - \bar{x}_n\right\} + l_3\left(\bar{z}_m - \bar{Z}\right)\right] \\ &= \varphi_3 T_{3u} + \left(1-\varphi_3\right)T_{3m} \end{aligned} \qquad (28)$$

where $T_{3u} = \bar{y}_u + l_1\left(\bar{z}_u - \bar{Z}\right)$; an estimator based on the fresh sample of size u and $T_{3m} = \bar{y}_m + l_2\left(\bar{x}_m - \bar{x}_n\right) + l_3\left(\bar{z}_m - \bar{Z}\right)$; an estimator based on the matched sample of size m, $l_1 = \dfrac{c_5}{\varphi_3}$, $l_2 = \dfrac{\beta_3}{1-\varphi_3}$, $l_3 = \dfrac{c_6}{1-\varphi_3}$ and φ_3 are the unknown constants to be determined under certain criterions.

9.1. Properties of the estimator T_3

T_3 is an unbiased estimator of $\bar{Y}$ whose variance is given in the following theorem.

Theorem 9.1. Variance of the estimator T_3 is obtained as

$$V\left(T_3\right) = \varphi_3^2\, V\left(T_{3u}\right) + \left(1-\varphi_3\right)^2 V\left(T_{3m}\right) \tag{29}$$

where $$V\left(T_{3u}\right) = \frac{1}{u}\left(1 + l_1^2 + 2l_1\rho_{yz}\right)S_y^2 \tag{30}$$

$$V\left(T_{3m}\right) = \left[\frac{1}{m}\left(1 + l_3^2 + 2l_3\rho_{yz}\right) + \left(\frac{1}{m} - \frac{1}{n}\right)\left(l_2^2 + 2l_2\rho_{yx} + 2l_2l_3\rho_{xz}\right)\right]S_y^2 \tag{31}$$

Proof: It is obvious that the variance of the estimator T_3 is given by

$$\begin{aligned} V\left(T_3\right) &= E\left[T_3 - \bar{Y}\right]^2 = E\left[\varphi_3\left(T_{3u} - \bar{Y}\right) + \left(1-\varphi_3\right)\left(T_{3m} - \bar{Y}\right)\right]^2 \\ &= \varphi_3^2\, V\left(T_{3u}\right) + \left(1-\varphi_3\right)^2 V\left(T_{3m}\right) + 2\varphi_3\left(1-\varphi_3\right)R_{11} \end{aligned} \tag{32}$$

where $V\left(T_{3u}\right) = E\left[T_{3u} - \bar{Y}\right]^2$, $V\left(T_{3m}\right) = E\left[T_{3m} - \bar{Y}\right]^2$ and

$R_{11} = E\left[\left(T_{3u} - \bar{Y}\right)\left(T_{3m} - \bar{Y}\right)\right]$

Substituting the expressions of T_{3u} and T_{3m} in equation (32), taking expectations and ignoring finite population corrections, we have the expression of the variance of T_3 as given in equation (29).

Remark 9.1. Results in theorem 9.1 is derived similar to the results obtained in theorem 4.1.

9.2. Minimum variance of the estimator T_3

Since the variance of the estimator T_3 in equation (29) is the function of the unknown constants l_1, l_2, l_3 and φ_3, therefore it is minimized with respect to these constants and subsequently the optimum values of l_1, l_2, l_3 and φ_3 are obtained as

$$l_1^* = -\rho_{yz} \tag{33}$$

$$l_2^* = \frac{\rho_{yz}\,\rho_{xz} - \rho_{yx}}{1-\mu_3\rho_{xz}^2} \tag{34}$$

$$l_3^* = \frac{\mu_3\rho_{yx}\rho_{xz} - \rho_{yz}}{1-\mu_3\rho_{xz}^2} \tag{35}$$

$$\varphi_{3_{opt}} = \frac{V\left(T_{3m}\right)}{V\left(T_{3u}\right)+V\left(T_{3m}\right)} \tag{36}$$

Now, substituting the values of l_1^*, l_2^* and l_3^* in equations (30) and (31), we get the optimum variances of T_{3u}, T_{3m} as

$$V\left(T_{3u}\right)_{opt} = \frac{1}{u}B_1S_y^2 \tag{37}$$

$$V\left(T_{3m}\right)_{opt} = \frac{1}{m}\left[\frac{B_1+\mu_3B_5+\mu_3^2B_2}{\left(1-\mu_3\rho_{xz}^2\right)^2}\right]S_y^2 \tag{38}$$

where $B_1 = 1-\rho_{yz}^2$, $B_2 = \rho_{xz}^2\left(\rho_{xz}^2+\rho_{yx}^2-2\rho_{yz}\rho_{yx}\rho_{xz}\right)$, $B_3 = -2\rho_{xz}^2B_1$, $B_4 = -\left(\rho_{yz}\rho_{xz}-\rho_{yx}\right)^2$ and $B_5 = B_3+B_4$.

Further, substituting the values of $V(T_{3u})_{opt}$ and $V(T_{3m})_{opt}$ from equations (37) and (38) in equation (36), we get the optimum value $\varphi_{1_{opt}}$ with respect to l_1^*, l_2^* and l_3^* as

$$\varphi_{3opt}^* = \frac{V\left(T_{3m}\right)_{opt}}{V\left(T_{3u}\right)_{opt}+V\left(T_{3m}\right)_{opt}} \tag{39}$$

Again, from equation (39) substituting the value of φ_{3opt}^* in equation (29), we get the optimum variance of T_3 as

$$V\left(T_3\right)_{opt} = \frac{V\left(T_{3m}\right)_{opt}\ V\left(T_{3u}\right)_{opt}}{V\left(T_{3m}\right)_{opt} + V\left(T_{3u}\right)_{opt}} \tag{40}$$

Further, substituting the values from equations (37) and (38) in equations (39) and (40), the simplified values of φ^{*}_{3opt} and $V\left(T_3\right)_{opt}$ are obtained as

$$\varphi^{*}_{3opt} = \left[\frac{\mu_3\left(B_1+\mu_3B_5+\mu_3^2B_2\right)}{\mu_3^3B_6+\mu_3^2B_7+\mu_3B_3+B_1}\right] \tag{41}$$

$$V\left(T_3\right)_{opt} = \frac{1}{n}\left[\frac{B_8+\mu_3B_9+\mu_3^2B_{10}}{\mu_3^3B_6+\mu_3^2B_7+\mu_3B_3+B_1}\right]S_y^2 \tag{42}$$

where $B_6 = B_2\text{-}B_1\rho_{xz}^4$, $B_7 = B_1\rho_{xz}^4+2B_1\rho_{xz}^2+B_5$, $B_8 = B_1^2$, $B_9 = B_1B_5$, $B_{10} = B_1B_2$ and μ_3 is the fraction of fresh sample for the estimator T_3.

9.3. Optimum replacement policy

To determine the optimum value of μ_3 (fraction of a sample to be drawn afresh on the current occasion) so that population mean $\overline{Y}$ may be estimated with the maximum precision, we minimize the $V\left(T_3\right)_{opt}$ given in equation (42) with respect to μ_3, which result in fourth degree equation in μ_3 and respective solutions of μ_3 is discussed below:

$$P_1\mu_3^4+P_2\mu_3^3+P_3\mu_3^2+P_4\mu_3+P_5 = 0 \tag{43}$$

where $P_1 = \text{-}B_6B_{10}$, $P_2 = \text{-}2B_6B_9$,

$P_3 = B_3B_{10}\text{-}B_7B_9\text{-}3B_6B_8$, $P_4 = 2\left(B_1B_{10}\text{-}B_7B_8\right)$,

$P_5 = B_1B_9\text{-}B_3B_8$

From equations (43) it is obvious that the four real values of μ_3 are possible. Hence, while choosing the values of μ_3, it should be remembered that $0 \le \mu_3 \le 1$. All the other values of μ_3 are inadmissible. If more than one admissible values are obtained, the lowest admissible value is the best choice as it reduces the cost of the survey. From equation (43), substituting the admissible value of μ_3 say $\hat{\mu}_3$ into equation (42), we have the optimum value of $V\left(T_3\right)_{opt}$ as

$$V\left(T_3^0\right)_{opt} = \frac{1}{n}\left[\frac{B_1\left(B_1 - \hat{\mu}_3 B_4 + \hat{\mu}_3^2 B_5\right)}{\hat{\mu}_3^3 B_6 + \hat{\mu}_3^2 B_7 + \hat{\mu}_3 B_8 + B_1}\right] S_y^2 \qquad (44)$$

9.4. Efficiency comparison

To study the performance of the estimator T_3, the percent relative efficiencies of the estimator T_3 with respect to (i) $\bar{y}_n$, when there is no matching, and (ii) the estimator T_2, when no auxiliary information is used at any occasion, have been obtained for different choices of correlations. For different choices of ρ_{yx}, ρ_{xz} and ρ_{yz}, Table 2 shows the optimum values of μ_3 and percent relative efficiencies E_3 and E_4 of the estimator T_3 with respect to the estimators $\bar{y}_n$ and T_2 respectively, where

$$E_3 = \frac{V\left(\bar{y}_n\right)}{V\left(T_3^0\right)_{opt}} \times 100 \text{ and } E_4 = \frac{V\left(T_2^0\right)_{opt}}{V\left(T_3^0\right)_{opt}} \times 100.$$

9.5. Analysis of results for estimator T_3

The following conclusions can be read out from Table 2:

(a) For fixed values of ρ_{xz} and ρ_{yz}, the values of μ_3 and E_3 are increasing with the increasing values of ρ_{yx}. Efficiencies E_4 are decreasing for the increasing values of ρ_{yx}.

(b) For fixed values of ρ_{xz} and ρ_{yx}, the values of μ_3 increase for the lower values of ρ_{yz} and decrease for the higher values of ρ_{yz}. Efficiencies E_3 and E_4 are increasing with the increasing values of ρ_{yz}.

(c) For fixed values of ρ_{yz} and ρ_{yx}, the values of μ_3 are increasing with the increasing values of ρ_{xz}. Efficiencies E_3 and E_4 increase for the lower values of ρ_{xz} while decreasing patteren may also be seen for the higher values of ρ_{xz}.

(d) Minimum value of μ_3 is 0.5365, which indicates that only 54 percent of the total sample size is to be replaced at the current occasion for the corresponding choices of the correlations.

Table 1. Optimum values of μ_1 and percent relative efficiencies of T_1 with respect to $\bar{y}_n$ and T_2

$\rho_{xz}\downarrow$	$\rho_{yz}\downarrow$	$\rho_{yx}\rightarrow$	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.5	0.5	$\hat{\mu}_1$	0.5006	0.5051	0.5147	0.5307	0.5556	0.5953	0.6672
		E_1	133.48	134.69	137.25	141.51	148.14	158.74	177.91
		E_2	130.41	129.07	128.06	127.35	126.97	126.99	127.72
	0.7	$\hat{\mu}_1$	0.4956	0.4965	0.5039	0.5190	0.5450	0.5899	0.6818
		E_1	194.36	194.70	197.60	203.53	213.74	231.35	267.36
		E_2	189.88	186.58	184.36	183.17	183.18	185.08	191.95
	0.9	$\hat{\mu}_1$	0.4352	0.4329	0.4404	0.4597	0.4982	0.5823	*
		E_1	458.11	455.69	463.55	483.92	524.45	612.93	-
		E_2	447.55	436.67	432.49	435.53	449.49	490.35	-
0.7	0.5	$\hat{\mu}_1$	0.4987	0.4996	0.5070	0.5221	0.5481	0.5929	0.6844
		E_1	132.97	133.22	135.19	139.22	146.16	158.11	182.51
		E_2	129.91	127.65	126.13	125.30	125.27	126.49	131.03
	0.7	$\hat{\mu}_1$	0.5187	0.5040	0.5000	0.5060	0.5232	0.5574	0.6271
		E_1	203.39	197.62	196.09	198.41	205.17	218.58	245.91
		E_2	198.71	189.37	182.96	178.57	175.85	174.86	176.55
	0.9	$\hat{\mu}_1$	0.6616	0.5334	0.4915	0.4796	0.4897	0.5286	0.6436
		E_1	696.46	561.47	517.36	504.86	515.52	556.42	677.46
		E_2	680.42	538.03	482.70	454.37	441.84	445.13	486.37
0.9	0.5	$\hat{\mu}_1$	0.4820	0.4883	0.5003	0.5201	0.5515	0.6050	0.7230
		E_1	128.54	130.20	133.43	138.68	147.07	161.33	192.78
		E_2	125.58	124.77	124.48	124.81	126.05	129.06	138.41
	0.7	$\hat{\mu}_1$	0.6548	0.5435	0.5049	0.4943	0.5051	0.5440	0.6564
		E_1	256.79	213.15	198.01	193.85	198.07	213.33	257.41
		E_2	250.87	204.25	184.74	174.47	169.76	170.66	184.81
	0.9	$\hat{\mu}_1$	*	*	*	*	0.5509	0.5003	0.5317
		E_1	-	-	-	-	579.85	526.68	559.70
		E_2	-	-	-	-	496.96	421.35	401.83

Note: "*" indicates $\hat{\mu}_1$ do not exist.

Table 2. Optimum values of μ_3 and percent relative efficiencies of T_3 with respect to $\bar{y}_n$ and T_2

$\rho_{xz}\downarrow$	$\rho_{yz}\downarrow$	$\rho_{yx}\rightarrow$	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.5	0.5	$\hat{\mu}_3$	0.5365	0.5410	0.5505	0.5663	0.5907	0.6294	0.6983
		E_3	133.46	134.50	136.70	140.32	145.89	154.64	169.95
		E_4	130.38	128.88	127.54	126.28	125.04	123.71	122.02
	0.7	$\hat{\mu}_3$	0.5367	0.5367	0.5434	0.5580	0.5834	0.6273	0.7163
		E_3	196.35	196.35	198.63	203.56	212.11	226.71	255.75
		E_4	191.83	188.15	185.32	183.20	181.79	181.37	183.61
	0.9	$\hat{\mu}_3$	0.5572	0.5381	0.5381	0.5572	0.6065	0.7550	*
		E_3	545.67	528.32	528.32	545.66	589.99	719.76	-
		E_4	533.09	506.26	492.93	491.09	505.66	575.81	-
0.7	0.5	$\hat{\mu}_3$	0.5842	0.5842	0.5907	0.6049	0.6294	0.6712	0.7538
		E_3	133.48	133.48	134.72	137.40	141.96	149.55	163.89
		E_4	130.41	127.91	125.70	123.66	121.67	119.64	117.66
	0.7	$\hat{\mu}_3$	0.6014	0.5872	0.5834	0.5892	0.6058	0.6381	0.7019
		E_3	201.09	197.15	196.09	197.69	202.30	211.13	227.92
		E_4	196.46	188.92	182.95	177.92	173.38	168.89	163.63
	0.9	$\hat{\mu}_3$	*	0.6751	0.6065	0.5845	0.5897	0.6257	0.7378
		E_3	-	593.11	543.53	527.17	531.03	557.65	636.23
		E_4	-	568.34	507.12	474.45	455.13	446.12	456.78
0.9	0.5	$\hat{\mu}_3$	0.7143	0.6983	0.6983	0.7143	0.7538	0.8596	*
		E_3	135.37	133.55	133.55	135.37	139.64	149.42	-
		E_4	132.25	127.97	124.60	121.83	119.68	119.54	-
	0.7	$\hat{\mu}_3$	*	0.7730	0.7163	0.6974	0.7019	0.7325	0.8217
		E_3	-	208.23	199.41	196.24	197.001	202.04	215.00
		E_4	-	199.53	186.05	176.62	68.84	161.63	154.36
	0.9	$\hat{\mu}_3$	*	*	*	*	0.7378	0.6967	0.7226
		E_3	-	-	-	-	544.564	526.45	538.03
		E_4	-	-	-	-	66.72	421.16	386.27

Note: "*" indicates $\hat{\mu}_3$ do not exist.

10. General conclusions

The estimators T_1 and T_3 proposed in this work are proved to be the best linear unbiased estimators of population mean $\overline{Y}$ with their respective minimum variance. These estimators may be seen as new innovative ideas in survey literature as they nicely utilized the information on an auxiliary variable in order to improve the precision of the estimates. From the analysis of the results shown in Tables 1-2, the propositions of the estimators T_1 and T_3 are vindicated because it enhances the precision of estimates as well as reduces the cost of the survey. Therefore, the proposed estimators may be recommended to survey practitioners for use in real life problems.

Acknowledgements

Authors are thankful to the referee for his valuable suggestions. Authors are also thankful to the UGC, New Delhi and Indian School of Mines, Dhanbad for providing financial assistance and necessary infrastructures to carry out the present research 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.

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. Agr. 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, 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., (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, *J. Indian 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, 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.

SINGH, H. P. and VISHWAKARMA, G. K., (2009). A general procedure for estimating population mean in successive sampling, *Communications in Statistics - Theory and Methods,* 38(2), 293-308.

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 75–88

ESTIMATION OF FINITE POPULATION MEAN USING DECILES OF AN AUXILIARY VARIABLE

J. Subramani, G. Kumarapandiyan[1]

ABSTRACT

The present paper deals with a class of modified ratio estimators for estimation of population mean of the study variable when the population deciles of the auxiliary variable are known. The biases and the mean squared errors of the proposed estimators are derived and compared with that of existing modified ratio estimators for certain known populations. Further, we have also derived the conditions for which the proposed estimators perform better than the existing modified ratio estimators. From the numerical study it is also observed that the proposed modified ratio estimators perform better than the existing modified ratio estimators.

Key words: mean squared error, natural populations, simple random sampling.

1. Introduction

Consider a finite population $U = \{U_1, U_2, \dots, U_N\}$ of N distinct and identifiable units. Let Y be a real variable with value Y_i measured on $U_i, i = 1,2,3, \dots, N$ giving a vector $Y = \{Y_1, Y_2, \dots, Y_N\}$. The problem is to estimate the population mean $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ on the basis of a random sample selected from the population U. The simple random sample mean is the simplest estimator of population mean. If an auxiliary variable X closely related to the study variable Y is available then one can improve the performance of the estimator of the study variable by using the known values of the population parameters of the auxiliary variable. That is, when the population parameters of the auxiliary variable X such as Population Mean, Co-efficient of Variation, Co-efficient of Kurtosis, Co-efficient of Skewness, etc., are known, a number of estimators such as ratio, product and linear regression estimators and their modifications are proposed in the literature. Among these estimators the ratio estimator and its modifications are widely used for the

[1] Department of Statistics, Ramanujan School of Mathematical Sciences, Pondicherry University. R V Nagar, Kalapet, Puducherry – 605014.
E-mail: drjsubramani@yahoo.co.in; kumarstat88@gmail.com.

estimation of the mean of the study variable. Before discussing further about the modified ratio estimators and the proposed modified ratio estimators the notations to be used in this paper are described below:

- N – Population size
- n – Sample size
- $f = n/N$, Sampling fraction
- Y – Study variable
- X – Auxiliary variable
- $\overline{X}, \overline{Y}$ – Population means
- x, y - Sample totals
- $\bar{x}, \bar{y}$ – Sample means
- S_X, S_y – Population standard deviations
- C_X, C_y – Coefficient of variations
- ρ – Coefficient of correlation
- $\beta_1 = \frac{N\sum_{i=1}^{N}(X_i-\overline{X})^3}{(N-1)(N-2)S^3}$, Coefficient of skewness of the auxiliary variable
- $\beta_2 = \frac{N(N+1)\sum_{i=1}^{N}(X_i-\overline{X})^4}{(N-1)(N-2)(N-3)S^4} - \frac{3(N-1)^2}{(N-2)(N-3)}$, Coefficient of kurtosis of the auxiliary variable
- $B(.)$ – Bias of the estimator
- $MSE(.)$ – Mean squared error of the estimator
- $\widehat{\overline{Y}}_i(\widehat{\overline{Y}}_{pi})$ – Existing (proposed) modified ratio estimator of $\overline{Y}$

The Ratio estimator for estimating the population mean $\overline{Y}$ of the study variable Y is defined as

$$\widehat{\overline{Y}}_R = \frac{\overline{y}}{\overline{x}}\overline{X} = \widehat{R}\,\overline{X} \tag{1.1}$$

where $\widehat{R} = \frac{\overline{y}}{\overline{x}} = \frac{y}{x}$ is the estimate of $R = \frac{\overline{Y}}{\overline{X}} = \frac{Y}{X}$. The Ratio estimator given in (1.1) is used for improving the precision of estimate of the population mean compared to simple random sampling when there is a positive correlation between X and Y. Further improvements are achieved on the classical ratio estimator by introducing a large number of modified ratio estimators with the use of known Co-efficient of Variation, Co-efficient of Kurtosis, Co-efficient of Skewness, etc. The lists of modified ratio estimators, which are to be compared with that of the proposed estimators, are divided into two classes namely Class 1 and Class 2, and are given respectively in Table 1.1. and Table 1.2. As stated above, some of the existing

modified ratio estimators together with their biases, mean squared errors and constants available in the literature are presented in the following tables:

Table 1.1. Existing modified ratio type estimators (Class 1) with their biases, mean squared errors and their constants

Estimator	**Bias - B**(.)	**Mean squared error MSE**(.)	**Constant** θ_i
$\hat{\bar{Y}}_1 = \bar{y}\left[\frac{\bar{X}+C_x}{\bar{x}+C_x}\right]$ Sisodia and Dwivedi [13]	$\frac{(1-f)}{n}\bar{Y}(\theta_1^2C_x^2-\theta_1C_xC_y\rho)$	$\frac{(1-f)}{n}\bar{Y}^2(C_y^2+\theta_1^2C_x^2-2\theta_1C_xC_y\rho)$	$\theta_1=\frac{\bar{X}}{\bar{X}+C_x}$
$\hat{\bar{Y}}_2 = \bar{y}\left[\frac{\bar{X}+\beta_2}{\bar{x}+\beta_2}\right]$ Singh et.al [11]	$\frac{(1-f)}{n}\bar{Y}(\theta_2^2C_x^2-\theta_2C_xC_y\rho)$	$\frac{(1-f)}{n}\bar{Y}^2(C_y^2+\theta_2^2C_x^2-2\theta_2C_xC_y\rho)$	$\theta_2=\frac{\bar{X}}{\bar{X}+\beta_2}$
$\hat{\bar{Y}}_3 = \bar{y}\left[\frac{\bar{X}+\beta_1}{\bar{x}+\beta_1}\right]$ Yan and Tian [15]	$\frac{(1-f)}{n}\bar{Y}(\theta_3^2C_x^2-\theta_3C_xC_y\rho)$	$\frac{(1-f)}{n}\bar{Y}^2(C_y^2+\theta_3^2C_x^2-2\theta_3C_xC_y\rho)$	$\theta_3=\frac{\bar{X}}{\bar{X}+\beta_1}$
$\hat{\bar{Y}}_4 = \bar{y}\left[\frac{\bar{X}+\rho}{\bar{x}+\rho}\right]$ Singh and Tailor [10]	$\frac{(1-f)}{n}\bar{Y}(\theta_4^2C_x^2-\theta_4C_xC_y\rho)$	$\frac{(1-f)}{n}\bar{Y}^2(C_y^2+\theta_4^2C_x^2-2\theta_4C_xC_y\rho)$	$\theta_4=\frac{\bar{X}}{\bar{X}+\rho}$
$\hat{\bar{Y}}_5 = \bar{y}\left[\frac{\bar{X}C_x+\beta_2}{\bar{x}C_x+\beta_2}\right]$ Upadhyaya and Singh [14]	$\frac{(1-f)}{n}\bar{Y}(\theta_5^2C_x^2-\theta_5C_xC_y\rho)$	$\frac{(1-f)}{n}\bar{Y}^2(C_y^2+\theta_5^2C_x^2-2\theta_5C_xC_y\rho)$	$\theta_5=\frac{\bar{X}C_x}{\bar{X}C_x+\beta_2}$
$\hat{\bar{Y}}_6 = \bar{y}\left[\frac{\bar{X}\beta_2+C_x}{\bar{x}\beta_2+C_x}\right]$ Upadhyaya and Singh [14]	$\frac{(1-f)}{n}\bar{Y}(\theta_6^2C_x^2-\theta_6C_xC_y\rho)$	$\frac{(1-f)}{n}\bar{Y}^2(C_y^2+\theta_6^2C_x^2-2\theta_6C_xC_y\rho)$	$\theta_6=\frac{\bar{X}\beta_2}{\bar{X}\beta_2+C_x}$
$\hat{\bar{Y}}_7 = \bar{y}\left[\frac{\bar{X}\beta_2+\beta_1}{\bar{x}\beta_2+\beta_1}\right]$ Yan and Tian [15]	$\frac{(1-f)}{n}\bar{Y}(\theta_7^2C_x^2-\theta_7C_xC_y\rho)$	$\frac{(1-f)}{n}\bar{Y}^2(C_y^2+\theta_7^2C_x^2-2\theta_7C_xC_y\rho)$	$\theta_7=\frac{\bar{X}\beta_2}{\bar{X}\beta_2+\beta_1}$
$\hat{\bar{Y}}_8 = \bar{y}\left[\frac{\bar{X}\beta_1+\beta_2}{\bar{x}\beta_1+\beta_2}\right]$ Yan and Tian [15]	$\frac{(1-f)}{n}\bar{Y}(\theta_8^2C_x^2-\theta_8C_xC_y\rho)$	$\frac{(1-f)}{n}\bar{Y}^2(C_y^2+\theta_8^2C_x^2-2\theta_8C_xC_y\rho)$	$\theta_8=\frac{\bar{X}\beta_1}{\bar{X}\beta_1+\beta_2}$
$\hat{\bar{Y}}_9 = \bar{y}\left[\frac{\bar{X}C_x+\beta_1}{\bar{x}C_x+\beta_1}\right]$ Yan and Tian [15]	$\frac{(1-f)}{n}\bar{Y}(\theta_9^2C_x^2-\theta_9C_xC_y\rho)$	$\frac{(1-f)}{n}\bar{Y}^2(C_y^2+\theta_9^2C_x^2-2\theta_9C_xC_y\rho)$	$\theta_9=\frac{\bar{X}C_x}{\bar{X}C_x+\beta_1}$

Table 1.2. Existing modified ratio type estimators (Class 2) with their biases, mean squared errors and their constants

Estimator	Bias-B(.)	Mean squared error MSE(.)	Constant R_i
$\hat{\bar{Y}}_{10} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{\bar{x}}\bar{X}$ Kadilar and Cingi [2]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{10}^2$	$\frac{(1-f)}{n}\left(R_{10}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{10} = \frac{\bar{Y}}{\bar{X}}$
$\hat{\bar{Y}}_{11} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + C_x)}(\bar{X} + C_x)$ Kadilar and Cingi [2]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{11}^2$	$\frac{(1-f)}{n}\left(R_{11}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{11} = \frac{\bar{Y}}{\bar{X} + C_x}$
$\hat{\bar{Y}}_{12} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + \beta_2)}(\bar{X} + \beta_2)$ Kadilar and Cingi [2]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{12}^2$	$\frac{(1-f)}{n}\left(R_{12}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{12} = \frac{\bar{Y}}{\bar{X} + \beta_2}$
$\hat{\bar{Y}}_{13} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\beta_2 + C_x)}(\bar{X}\beta_2 + C_x)$ Kadilar and Cingi [2]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{13}^2$	$\frac{(1-f)}{n}\left(R_{13}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{13} = \frac{\bar{Y}\beta_2}{\bar{X}\beta_2 + C_x}$
$\hat{\bar{Y}}_{14} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + \beta_2)}(\bar{X}C_x + \beta_2)$ Kadilar and Cingi [2]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{14}^2$	$\frac{(1-f)}{n}\left(R_{14}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{14} = \frac{\bar{Y}C_x}{\bar{X}C_x + \beta_2}$
$\hat{\bar{Y}}_{15} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + \beta_1)}(\bar{X} + \beta_1)$ Yan and Tian [15]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{15}^2$	$\frac{(1-f)}{n}\left(R_{15}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{15} = \frac{\bar{Y}}{\bar{X} + \beta_1}$
$\hat{\bar{Y}}_{16} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\beta_1 + \beta_2)}(\bar{X}\beta_1 + \beta_2)$ Yan and Tian [15]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{16}^2$	$\frac{(1-f)}{n}\left(R_{16}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{16} = \frac{\bar{Y}\beta_1}{\bar{X}\beta_1 + \beta_2}$
$\hat{\bar{Y}}_{17} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + \rho)}(\bar{X} + \rho)$ Kadilar and Cingi [3]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{17}^2$	$\frac{(1-f)}{n}\left(R_{17}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{17} = \frac{\bar{Y}}{\bar{X} + \rho}$
$\hat{\bar{Y}}_{18} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + \rho)}(\bar{X}C_x + \rho)$ Kadilar and Cingi [3]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{18}^2$	$\frac{(1-f)}{n}\left(R_{18}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{18} = \frac{\bar{Y}C_x}{\bar{X}C_x + \rho}$
$\hat{\bar{Y}}_{19} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + C_x)}(\bar{X}\rho + C_x)$ Kadilar and Cingi [3]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{19}^2$	$\frac{(1-f)}{n}\left(R_{19}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{19} = \frac{\bar{Y}\rho}{\bar{X}\rho + C_x}$
$\hat{\bar{Y}}_{20} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\beta_2 + \rho)}(\bar{X}\beta_2 + \rho)$ Kadilar and Cingi [3]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{20}^2$	$\frac{(1-f)}{n}\left(R_{20}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{20} = \frac{\bar{Y}\beta_2}{\bar{X}\beta_2 + \rho}$
$\hat{\bar{Y}}_{21} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + \beta_2)}(\bar{X}\rho + \beta_2)$ Kadilar and Cingi [3]	$\frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}}R_{21}^2$	$\frac{(1-f)}{n}\left(R_{21}^2 S_x^2 + S_y^2(1-\rho^2)\right)$	$R_{21} = \frac{\bar{Y}\rho}{\bar{X}\rho + \beta_2}$

It is to be noted that "the existing modified ratio estimators" mean the list of modified ratio estimators to be considered in this paper unless otherwise stated. It does not mean the entire list of modified ratio type estimators available in the literature. For a more detailed discussion on the ratio estimator and its modifications one may refer to Cochran [1], Kadilar and Cingi [2, 3], Koyuncu and Kadilar [4], Murthy [5], Prasad [6], Rao [7], Singh [9], Singh and Tailor [10,12], Singh et.al [11], Sisodia and Dwivedi [13], Upadhyaya and Singh [14], Yan and Tian [15] and the references cited therein.

The modified ratio type estimators discussed above are biased but have minimum mean squared errors compared to the classical ratio estimator. The list of estimators given in Table 1.1. and Table 1.2. uses the known values of the parameters like $\bar{X}$, C_x, β_1, β_2, ρ and their linear combinations. However, it seems no attempt is made to use the known values of the population deciles of the auxiliary variable to improve the ratio estimator. Further, we know that the value of deciles is unaffected and robustness by the extreme values or the presence of outliers in the population values unlike the other parameters like the mean, coefficient of variation, coefficient of skewness and coefficient of kurtosis, etc. The points discussed above have motivated us to introduce modified ratio estimators using the known value of the population deciles of the auxiliary variable. It is observed that the proposed estimators perform better than the existing modified ratio type estimators listed in Table 1.1. and Table 1.2. The materials of this paper are arranged as follows: The proposed modified ratio estimators with known population deciles of an auxiliary variable are presented in section 2 whereas the conditions in which the proposed estimators perform better than the existing modified ratio estimators are derived in section 3. The performances of the proposed modified ratio estimators compared to the existing modified ratio estimators are assessed for certain natural populations in section 4 and the conclusion is presented in section 5.

2. Proposed modified ratio type estimators using deciles of the auxiliary variable

As we stated earlier one can always improve the performance of the estimator of the study variable by using the known population parameters of the auxiliary variable, which are positively correlated with that of study variable. In this section, we have suggested a class of modified ratio type estimators using the population deciles, denoted by $D_j ; j = 1, 2, 3, \ldots, 10$ of the auxiliary variable. The proposed modified ratio type estimators $\widehat{\bar{Y}}_{pj}$, $j = 1,2, \ldots ,10$ for estimating the

population mean $\overline{Y}$ together with the first degree of approximation, the biases, mean squared errors and the constants are given below:

Table 2.1. Proposed modified ratio type estimators (Class 3) with their biases, mean squared errors and their constants

Estimator	Bias - B(.)	Mean squared error MSE(.)	Constant θ_i
$\widehat{\overline{Y}}_{p1} = \bar{y}\left[\frac{\overline{X} + D_1}{\bar{x} + D_1}\right]$	$\frac{(1-f)}{n}\overline{Y}\,(\theta_{p1}^2 C_x^2 - \theta_{p1} C_x C_y \rho)$	$\frac{(1-f)}{n}\overline{Y}^2(C_y^2 + \theta_{p1}^2 C_x^2 - 2\theta_{p1} C_x C_y \rho)$	$\theta_{p1} = \frac{\overline{X}}{\overline{X} + D_1}$
$\widehat{\overline{Y}}_{p2} = \bar{y}\left[\frac{\overline{X} + D_2}{\bar{x} + D_2}\right]$	$\frac{(1-f)}{n}\overline{Y}\,(\theta_{p2}^2 C_x^2 - \theta_{p2} C_x C_y \rho)$	$\frac{(1-f)}{n}\overline{Y}^2(C_y^2 + \theta_{p2}^2 C_x^2 - 2\theta_{p2} C_x C_y \rho)$	$\theta_{p2} = \frac{\overline{X}}{\overline{X} + D_2}$
$\widehat{\overline{Y}}_{p3} = \bar{y}\left[\frac{\overline{X} + D_3}{\bar{x} + D_3}\right]$	$\frac{(1-f)}{n}\overline{Y}\,(\theta_{p3}^2 C_x^2 - \theta_{p3} C_x C_y \rho)$	$\frac{(1-f)}{n}\overline{Y}^2(C_y^2 + \theta_{p3}^2 C_x^2 - 2\theta_{p3} C_x C_y \rho)$	$\theta_{p3} = \frac{\overline{X}}{\overline{X} + D_3}$
$\widehat{\overline{Y}}_{p4} = \bar{y}\left[\frac{\overline{X} + D_4}{\bar{x} + D_4}\right]$	$\frac{(1-f)}{n}\overline{Y}\,(\theta_{p4}^2 C_x^2 - \theta_{p4} C_x C_y \rho)$	$\frac{(1-f)}{n}\overline{Y}^2(C_y^2 + \theta_{p4}^2 C_x^2 - 2\theta_{p4} C_x C_y \rho)$	$\theta_{p4} = \frac{\overline{X}}{\overline{X} + D_4}$
$\widehat{\overline{Y}}_{p5} = \bar{y}\left[\frac{\overline{X} + D_5}{\bar{x} + D_5}\right]$	$\frac{(1-f)}{n}\overline{Y}\,(\theta_{p5}^2 C_x^2 - \theta_5 C_x C_y \rho)$	$\frac{(1-f)}{n}\overline{Y}^2(C_y^2 + \theta_{p5}^2 C_x^2 - 2\theta_{p5} C_x C_y \rho)$	$\theta_{p5} = \frac{\overline{X}}{\overline{X} + D_5}$
$\widehat{\overline{Y}}_{p6} = \bar{y}\left[\frac{\overline{X} + D_6}{\bar{x} + D_6}\right]$	$\frac{(1-f)}{n}\overline{Y}\,(\theta_{p6}^2 C_x^2 - \theta_{p6} C_x C_y \rho)$	$\frac{(1-f)}{n}\overline{Y}^2(C_y^2 + \theta_{p6}^2 C_x^2 - 2\theta_{p6} C_x C_y \rho)$	$\theta_{p6} = \frac{\overline{X}}{\overline{X} + D_6}$
$\widehat{\overline{Y}}_{p7} = \bar{y}\left[\frac{\overline{X} + D_7}{\bar{x} + D_7}\right]$	$\frac{(1-f)}{n}\overline{Y}\,(\theta_{p7}^2 C_x^2 - \theta_{p7} C_x C_y \rho)$	$\frac{(1-f)}{n}\overline{Y}^2(C_y^2 + \theta_{p7}^2 C_x^2 - 2\theta_{p7} C_x C_y \rho)$	$\theta_{p7} = \frac{\overline{X}}{\overline{X} + D_7}$
$\widehat{\overline{Y}}_{p8} = \bar{y}\left[\frac{\overline{X} + D_8}{\bar{x} + D_8}\right]$	$\frac{(1-f)}{n}\overline{Y}\,(\theta_{p8}^2 C_x^2 - \theta_{p8} C_x C_y \rho)$	$\frac{(1-f)}{n}\overline{Y}^2(C_y^2 + \theta_{p8}^2 C_x^2 - 2\theta_{p8} C_x C_y \rho)$	$\theta_{p8} = \frac{\overline{X}}{\overline{X} + D_8}$
$\widehat{\overline{Y}}_{p9} = \bar{y}\left[\frac{\overline{X} + D_9}{\bar{x} + D_9}\right]$	$\frac{(1-f)}{n}\overline{Y}\,(\theta_{p9}^2 C_x^2 - \theta_{p9} C_x C_y \rho)$	$\frac{(1-f)}{n}\overline{Y}^2(C_y^2 + \theta_{p9}^2 C_x^2 - 2\theta_{p9} C_x C_y \rho)$	$\theta_{p9} = \frac{\overline{X}}{\overline{X} + D_9}$
$\widehat{\overline{Y}}_{p10} = \bar{y}\left[\frac{\overline{X} + D_{10}}{\bar{x} + D_{10}}\right]$	$\frac{(1-f)}{n}\overline{Y}\,(\theta_{p10}^2 C_x^2 - \theta_{p10} C_x C_y \rho)$	$\frac{(1-f)}{n}\overline{Y}^2(C_y^2 + \theta_{p10}^2 C_x^2 - 2\theta_{p10} C_x C_y \rho)$	$\theta_{p10} = \frac{\overline{X}}{\overline{X} + D_{10}}$

3. Efficiency of the proposed estimators

For want of space, for the sake of convenience to the readers and for the ease of comparisons, the modified ratio type estimators given in Table 1.1, Table 1.2 and the proposed modified ratio estimators given in Table 2.1 are represented into three classes as given below:

Class 1: The biases, the mean squared errors and the constants of the modified ratio type estimators $\hat{\bar{Y}}_1$ to $\hat{\bar{Y}}_9$ listed in the Table 1.1 are represented in a single class (say Class 1), which will be very much useful for comparing with that of proposed modified ratio estimators, and are given below:

$$B\left(\hat{\bar{Y}}_i\right) = \frac{(1-f)}{n}\,\bar{Y}\,(\theta_i^2 C_x^2 - \theta_i C_x C_y\,\rho)$$

$$MSE\left(\hat{\bar{Y}}_i\right) = \frac{(1-f)}{n}\,\bar{Y}^2\left(C_y^2 + \theta_i^2 C_x^2 - 2\theta_i C_x C_y\,\rho\right)\ i = 1,2,3,\ldots,9 \qquad (3.1)$$

where $\theta_1 = \frac{\bar{X}}{\bar{X}+C_x}, \theta_2 = \frac{\bar{X}}{\bar{X}+\beta_2}, \theta_3 = \frac{\bar{X}}{\bar{X}+\beta_1}, \theta_4 = \frac{\bar{X}}{\bar{X}+\rho}, \theta_5 = \frac{\bar{X}\,C_x}{\bar{X}\,C_x+\beta_2}, \theta_6 = \frac{\bar{X}\,\beta_2}{\bar{X}\,\beta_2+C_x}$,

$\theta_7 = \frac{\bar{X}\,\beta_2}{\bar{X}\,\beta_2+\beta_1}, \theta_8 = \frac{\bar{X}\,\beta_1}{\bar{X}\,\beta_1+\beta_2}$ and $\theta_9 = \frac{\bar{X}\,C_x}{\bar{X}\,C_x+\beta_1}$

Class 2: The biases, the mean squared errors and the constants of the 12 modified ratio estimators $\hat{\bar{Y}}_{10}$ to $\hat{\bar{Y}}_{21}$ listed in the Table 1.2. are represented in a single class (say Class 2), which will be very much useful for comparing with that of proposed modified ratio estimators, and are given below:

$$B\left(\hat{\bar{Y}}_i\right) = \frac{(1-f)}{n}\,\frac{S_x^2}{\bar{Y}}\,R_i^2$$

$$MSE\left(\hat{\bar{Y}}_i\right) = \frac{(1-f)}{n}\left(R_i^2 S_x^2 + S_y^2(1-\rho^2)\right)\ i = 10,11,12,\ldots,21 \qquad (3.2)$$

where

$R_{10} = \frac{\bar{Y}}{\bar{X}}, R_{11} = \frac{\bar{Y}}{\bar{X}+C_x}, R_{12} = \frac{\bar{Y}}{\bar{X}+\beta_2}, R_{13} = \frac{\bar{Y}\beta_2}{\bar{X}\beta_2+C_x}, R_{14} = \frac{\bar{Y}C_x}{\bar{X}C_x+\beta_2}, R_{15} = \frac{\bar{Y}}{\bar{X}+\beta_1}, R_{16} = \frac{\bar{Y}\beta_1}{\bar{X}\beta_1+\beta_2}$,

$R_{17} = \frac{\bar{Y}}{\bar{X}+\rho}, R_{18} = \frac{\bar{Y}C_x}{\bar{X}C_x+\rho}, R_{19} = \frac{\bar{Y}\rho}{\bar{X}\rho+C_x}, R_{20} = \frac{\bar{Y}\beta_2}{\bar{X}\beta_2+\rho}$ and $R_{21} = \frac{\bar{Y}\rho}{\bar{X}\rho+\beta_2}$

Class 3: The biases, the mean squared errors and the constants of the 10 proposed modified ratio estimators $\hat{\bar{Y}}_{p1}\ to\ \hat{\bar{Y}}_{p10}$ listed in the Table 2.1. are represented in a single class (say Class 3), which will be very much useful for comparing with that of existing modified ratio estimators given in Class1 and Class 2, and are given below:

$$B\left(\hat{\bar{Y}}_{pj}\right) = \frac{(1-f)}{n}\,\bar{Y}\,(\theta_{pj}^2 C_x^2 - \theta_{pj} C_x C_y\,\rho)$$

$$MSE\left(\hat{\bar{Y}}_{pj}\right) = \frac{(1-f)}{n}\,\bar{Y}^2\left(C_y^2 + \theta_{pj}^2 C_x^2 - 2\theta_{pj} C_x C_y\,\rho\right),\quad j = 1,2,3,\ldots,10 \qquad (3.3)$$

where $\theta_{p1} = \frac{\overline{X}}{\overline{X}+D_1}, \theta_{p2} = \frac{\overline{X}}{\overline{X}+D_2}, \theta_{p3} = \frac{\overline{X}}{\overline{X}+D_3}, \theta_{p4} = \frac{\overline{X}}{\overline{X}+D_4}, \theta_{p5} = \frac{\overline{X}}{\overline{X}+D_5}, \theta_{p6} = \frac{\overline{X}}{\overline{X}+D_6}$,

$$\theta_{p7} = \frac{\overline{X}}{\overline{X}+D_7}, \theta_{p8} = \frac{\overline{X}}{\overline{X}+D_8}, \theta_{p9} = \frac{\overline{X}}{\overline{X}+D_9} \text{ and } \theta_{p10} = \frac{\overline{X}}{\overline{X}+D_{10}}$$

From the expressions given in (3.1) and (3.3) we have derived the conditions for which the proposed estimator $\hat{\bar{Y}}_{pj}$ is more efficient than the existing modified ratio type estimators given in Class 1, $\hat{\bar{Y}}_i ; i = 1, 2, 3, \ldots, 9$, and which are given below.

$$MSE\left(\hat{\bar{Y}}_{pj}\right) < MSE\left(\hat{\bar{Y}}_i\right) \text{ if } \rho < \frac{(\theta_{pj}+\theta_i)}{2}\frac{C_x}{C_y} ; i = 1, 2, 3, \ldots, 9, j = 1, 2, 3, \ldots, 10 \tag{3.4}$$

From the expressions given in (3.2) and (3.3) we have derived the conditions for which the proposed estimator $\hat{\bar{Y}}_{pj}$ is more efficient than the existing modified ratio estimators given in Class 2, $\hat{\bar{Y}}_i ; i = 10, 11, 12, \ldots, 21$, and which are given below:

$$MSE\left(\hat{\bar{Y}}_{pj}\right) < MSE\left(\hat{\bar{Y}}_i\right) \text{ if } \frac{\theta_{pj}C_x - R_i^* S_x}{C_y} < \rho < \frac{R_i^* S_x + \theta_{pj} C_x}{C_y} \text{ or } \frac{R_i^* S_x + \theta_{pj} C_x}{C_y} < \rho < \frac{\theta_{pj} C_x - R_i^* S_x}{C_y} \tag{3.5}$$

$i = 10, 11, 12, \ldots, 21$, $j = 1, 2, 3, \ldots, 10$

where $R_i^* = \frac{R_i}{\overline{Y}}$.

4. Empirical study

The performances of the proposed modified ratio estimators listed in Table 2.1. are assessed with that of existing modified ratio estimators listed in Table 1.1. and Table 1.2. for certain natural populations. In this connection, we have considered three natural populations for the assessment of the performances of the proposed modified ratio estimators with that of existing modified ratio estimators. They are: Population 1 is the closing price of the industry ACC in the National Stock Exchange from 2, January 2012 to 27, February 2012 [16]; Population 2 and Population 3 are taken from Singh and Chaudhary [8] given in page 177. The population parameters and the constants computed from the above populations are given below:

Parameters	Population 1	Population 2	Population 3
N	40	34	34
n	20	20	20
$\overline{Y}$	5141.5363	856.4117	856.4117

Parameters	Population 1	Population 2	Population 3
$\bar{X}$	1221.6463	208.8823	199.4412
ρ	0.9244	0.4491	0.4453
S_y	256.1464	733.1407	733.1407
C_y	0.0557	0.8561	0.8561
S_x	102.5494	150.5059	150.2150
C_x	0.0839	0.7205	0.7531
β_2	-1.5154	0.0978	1.0445
β_1	0.3761	0.9782	1.1823
D_1	1111.8150	70.3000	60.6000
D_2	1119.4800	76.8000	83.0000
D_3	1139.2000	108.2000	102.7000
D_4	1159.8400	129.4000	111.2000
D_5	1184.2250	150.0000	142.5000
D_6	1252.5500	227.2000	210.2000
D_7	1307.1000	250.4000	264.5000
D_8	1345.7200	335.6000	304.4000
D_9	1366.7850	436.1000	373.2000
D_{10}	1389.3000	564.0000	634.0000

The constants, the biases and the mean squared errors of the existing and proposed modified ratio estimators for the above populations are respectively given in the Tables 4.1. to 4.3.

Table 4.1. The constants of the existing and proposed modified ratio type estimators

Estimator	Constants θ_i or R_i		
	Population 1	Population 2	Population 3
$\hat{\bar{Y}}_1$ Sisodia and Dwivedi [13]	0.9999	0.9966	0.9962
$\hat{\bar{Y}}_2$ Singh et.al [11]	1.0012	0.9995	0.9948
$\hat{\bar{Y}}_3$ Yan and Tian [15]	0.9997	0.9953	0.9941
$\hat{\bar{Y}}_4$ Singh and Tailor [10]	0.9992	0.9979	0.9978
$\hat{\bar{Y}}_5$ Upadhyaya and Singh [14]	1.0150	0.9994	0.9931
$\hat{\bar{Y}}_6$ Upadhyaya and Singh [14]	1.0000	0.9658	0.9964
$\hat{\bar{Y}}_7$ Yan and Tian [15]	1.0002	0.9542	0.9944
$\hat{\bar{Y}}_8$ Yan and Tian [15]	1.0033	0.9995	0.9956
$\hat{\bar{Y}}_9$ Yan and Tian [15]	0.9963	0.9935	0.9922
$\hat{\bar{Y}}_{10}$ Kadilar and Cingi [2]	4.2087	4.1000	4.2941
$\hat{\bar{Y}}_{11}$ Kadilar and Cingi [2]	4.2084	4.0859	4.2779
$\hat{\bar{Y}}_{12}$ Kadilar and Cingi [2]	4.2139	4.0981	4.2717
$\hat{\bar{Y}}_{13}$ Kadilar and Cingi [2]	4.2089	3.9598	4.2786
$\hat{\bar{Y}}_{14}$ Kadilar and Cingi [2]	4.2718	4.0973	4.2644

Table 4.1. The constants of the existing and proposed modified ratio type estimators (cont.)

Estimator	Constants θ_i or R_i		
	Population 1	**Population 2**	**Population 3**
$\widehat{\bar{Y}}_{15}$ Yan and Tian [15]	4.2074	4.0809	4.2688
$\widehat{\bar{Y}}_{16}$ Yan and Tian [15]	4.2226	4.0980	4.2751
$\widehat{\bar{Y}}_{17}$ Kadilar and Cingi [3]	4.2055	4.0912	4.2845
$\widehat{\bar{Y}}_{18}$ Kadilar and Cingi [3]	4.1711	4.0878	4.2814
$\widehat{\bar{Y}}_{19}$ Kadilar and Cingi [3]	4.2084	4.0687	4.2579
$\widehat{\bar{Y}}_{20}$ Kadilar and Cingi [3]	4.2108	4.0115	4.2849
$\widehat{\bar{Y}}_{21}$ Kadilar and Cingi [3]	4.2143	4.0957	4.2441
$\widehat{\bar{Y}}_{p1}$(Proposed estimator)	0.5235	0.7482	0.7670
$\widehat{\bar{Y}}_{p2}$(Proposed estimator)	0.5218	0.7312	0.7061
$\widehat{\bar{Y}}_{p3}$(Proposed estimator)	0.5175	0.6588	0.6601
$\widehat{\bar{Y}}_{p4}$(Proposed estimator)	0.5130	0.6175	0.6420
$\widehat{\bar{Y}}_{p5}$(Proposed estimator)	0.5078	0.5820	0.5833
$\widehat{\bar{Y}}_{p6}$(Proposed estimator)	0.4938	0.4790	0.4869
$\widehat{\bar{Y}}_{p7}$(Proposed estimator)	0.4831	0.4548	0.4299
$\widehat{\bar{Y}}_{p8}$(Proposed estimator)	0.4758	0.3836	0.3958
$\widehat{\bar{Y}}_{p9}$(Proposed estimator)	0.4720	0.3239	0.3483
$\widehat{\bar{Y}}_{p10}$(Proposed estimator)	0.4679	0.2703	0.2393

Table 4.2. The biases of the existing and proposed modified ratio type estimators

Estimator	Bias **B**(.)		
	Population 1	**Population 2**	**Population 3**
$\widehat{\bar{Y}}_1$Sisodia and Dwivedi [13]	0.3505	4.2233	4.8836
$\widehat{\bar{Y}}_2$ Singh et.al [11]	0.3522	4.2631	4.8621
$\widehat{\bar{Y}}_3$ Yan and Tian [15]	0.3502	4.2070	4.8519
$\widehat{\bar{Y}}_4$ Singh and Tailor [10]	0.3497	4.2406	4.9064
$\widehat{\bar{Y}}_5$ Upadhyaya and Singh [14]	0.3697	4.2607	4.8369
$\widehat{\bar{Y}}_6$ Upadhyaya and Singh [14]	0.3507	3.8212	4.8860
$\widehat{\bar{Y}}_7$ Yan and Tian [15]	0.3509	3.6732	4.8556
$\widehat{\bar{Y}}_8$ Yan and Tian [15]	0.3548	4.2630	4.8739
$\widehat{\bar{Y}}_9$Yan and Tian [15]	0.3460	4.1831	4.8236
$\widehat{\bar{Y}}_{10}$ Kadilar and Cingi [2]	0.9058	9.1539	10.0023
$\widehat{\bar{Y}}_{11}$ Kadilar and Cingi [2]	0.9056	9.0911	9.9272
$\widehat{\bar{Y}}_{12}$ Kadilar and Cingi [2]	0.9080	9.1454	9.8983

Table 4.2. The biases of the existing and proposed modified ratio type estimators (cont.)

Estimator	Bias B(.)		
	Population 1	**Population 2**	**Population 3**
$\widehat{\bar{Y}}_{13}$ Kadilar and Cingi [2]	0.9058	8.5387	9.9303
$\widehat{\bar{Y}}_{14}$ Kadilar and Cingi [2]	0.9331	9.1420	9.8646
$\widehat{\bar{Y}}_{15}$ Yan and Tian [15]	0.9052	9.0688	9.8847
$\widehat{\bar{Y}}_{16}$ Yan and Tian [15]	0.9118	9.1452	9.9143
$\widehat{\bar{Y}}_{17}$ Kadilar and Cingi [3]	0.9044	9.1147	9.9578
$\widehat{\bar{Y}}_{18}$ Kadilar and Cingi [3]	0.8896	9.0995	9.9432
$\widehat{\bar{Y}}_{19}$ Kadilar and Cingi [3]	0.9056	9.0149	9.8348
$\widehat{\bar{Y}}_{20}$ Kadilar and Cingi [3]	0.9066	8.7630	9.9597
$\widehat{\bar{Y}}_{21}$ Kadilar and Cingi [3]	0.9081	9.1349	9.7711
$\widehat{\bar{Y}}_{p1}$(Proposed estimator)	0.0424	1.4697	2.0008
$\widehat{\bar{Y}}_{p2}$(Proposed estimator)	0.0430	1.3223	1.4125
$\widehat{\bar{Y}}_{p3}$(Proposed estimator)	0.0447	0.7548	1.0164
$\widehat{\bar{Y}}_{p4}$(Proposed estimator)	0.0464	0.4741	0.8726
$\widehat{\bar{Y}}_{p5}$(Proposed estimator)	0.0483	0.2581	0.4499
$\widehat{\bar{Y}}_{p6}$(Proposed estimator)	0.0533	0.2394	0.0939
$\widehat{\bar{Y}}_{p7}$(Proposed estimator)	0.0568	0.3281	0.3279
$\widehat{\bar{Y}}_{p8}$(Proposed estimator)	0.0591	0.5266	0.4367
$\widehat{\bar{Y}}_{p9}$(Proposed estimator)	0.0602	0.6218	0.5499
$\widehat{\bar{Y}}_{p10}$(Proposed estimator)	0.0614	0.6515	0.6387

Table 4.3. The mean squared errors of the existing and proposed modified ratio type estimators

Estimator	Mean Squared Error MSE(.)		
	Population 1	**Population 2**	**Population 3**
$\widehat{\bar{Y}}_1$Sisodia and Dwivedi [13]	995.2787	10514.2250	10929.0458
$\widehat{\bar{Y}}_2$ Singh et.al [11]	1000.0116	10535.8620	10916.9080
$\widehat{\bar{Y}}_3$ Yan and Tian [15]	994.4171	10505.3563	10911.1914
$\widehat{\bar{Y}}_4$ Singh and Tailor [10]	992.8028	10523.6171	10941.9491
$\widehat{\bar{Y}}_5$ Upadhyaya and Singh [14]	1050.6525	10534.5417	10902.7384
$\widehat{\bar{Y}}_6$ Upadhyaya and Singh [14]	995.6899	10298.4432	10930.3879
$\widehat{\bar{Y}}_7$ Yan and Tian [15]	996.2592	10220.4736	10913.2804
$\widehat{\bar{Y}}_8$ Yan and Tian [15]	1007.5083	10535.7860	10923.6103
$\widehat{\bar{Y}}_9$Yan and Tian [15]	982.4136	10492.3779	10895.2039

Table 4.3. The mean squared errors of the existing and proposed modified ratio type estimators (cont.)

Estimator	Mean Squared Error MSE(.)		
	Population 1	**Population 2**	**Population 3**
$\widehat{\bar{Y}}_{10}$ Kadilar and Cingi [2]	4954.6195	16673.4489	17437.6451
$\widehat{\bar{Y}}_{11}$ Kadilar and Cingi [2]	4953.9796	16619.6435	17373.3111
$\widehat{\bar{Y}}_{12}$ Kadilar and Cingi [2]	4966.1946	16666.1389	17348.6192
$\widehat{\bar{Y}}_{13}$ Kadilar and Cingi [2]	4955.0419	16146.6142	17376.0389
$\widehat{\bar{Y}}_{14}$ Kadilar and Cingi [2]	5095.3661	16663.3064	17319.7468
$\widehat{\bar{Y}}_{15}$ Yan and Tian [15]	4951.7534	16600.5393	17336.9770
$\widehat{\bar{Y}}_{16}$ Yan and Tian [15]	4985.4911	16665.9758	17362.2582
$\widehat{\bar{Y}}_{17}$ Kadilar and Cingi [3]	4947.5796	16639.8457	17399.5196
$\widehat{\bar{Y}}_{18}$ Kadilar and Cingi [3]	4871.7809	16626.8702	17387.0811
$\widehat{\bar{Y}}_{19}$ Kadilar and Cingi [3]	4953.9273	16554.4002	17294.1864
$\widehat{\bar{Y}}_{20}$ Kadilar and Cingi [3]	4959.2739	16338.6465	17401.1397
$\widehat{\bar{Y}}_{21}$ Kadilar and Cingi [3]	4967.1427	16657.1867	17239.6579
$\widehat{\bar{Y}}_{p1}$(Proposed estimator)	334.8577	9194.9620	9454.2668
$\widehat{\bar{Y}}_{p2}$(Proposed estimator)	336.2980	9139.9570	9214.1709
$\widehat{\bar{Y}}_{p3}$(Proposed estimator)	340.0837	8956.7638	9074.5845
$\widehat{\bar{Y}}_{p4}$(Proposed estimator)	344.1636	8889.1069	9029.7423
$\widehat{\bar{Y}}_{p5}$(Proposed estimator)	349.1280	8852.3417	8922.5155
$\widehat{\bar{Y}}_{p6}$(Proposed estimator)	363.7720	8857.3224	8874.7609
$\widehat{\bar{Y}}_{p7}$(Proposed estimator)	376.1193	8882.6263	8921.3976
$\widehat{\bar{Y}}_{p8}$(Proposed estimator)	385.1501	9010.2560	8975.8044
$\widehat{\bar{Y}}_{p9}$(Proposed estimator)	390.1632	9178.8233	9085.0541
$\widehat{\bar{Y}}_{p10}$(Proposed estimator)	395.5824	9377.5847	9481.5539

From the values of Table 4.2 it is observed that the biases of the proposed modified ratio estimators are lower than the biases of all the 21 existing modified ratio estimators. Similarly, from the values of Table 4.3, it is observed that the mean squared errors of the proposed modified ratio estimators are lower than the mean squared errors of all the 21 existing modified ratio estimators.

5. Conclusion

In this paper we have proposed a class of modified ratio type estimators using known values of population deciles of the auxiliary variable. The biases and mean squared errors of the proposed estimators are obtained and compared with that of existing modified ratio estimators. Further, we have derived the conditions for

which the proposed estimators are more efficient than the existing modified ratio estimators. We have also assessed the performances of the proposed estimators for some known populations. It is observed that the biases and mean squared errors of the proposed estimators are lower than the biases and mean squared errors of the existing modified ratio estimators for certain known populations. Hence, we strongly recommend that the proposed modified estimators may be preferred over the existing modified ratio estimators for the use of practical applications.

Acknowledgements

The second author wishes to record his gratitude and thanks to the Vice Chancellor, Pondicherry University and other University authorities for having given the financial assistance to carry out this research work through the University Fellowship.

REFERENCES

COCHRAN, W. G., (1977). Sampling Techniques, Third Edition, Wiley Eastern Limited.

KADILAR, C. and CINGI, H., (2004). Ratio estimators in simple random sampling, Applied Mathematics and Computation 151, 893-902.

KADILAR, C. and CINGI, H., (2006). An Improvement in Estimating the Population mean by using the Correlation Coefficient, Hacettepe Journal of Mathematics and Statistics Volume 35 (1), 103-109.

KOYUNCU, N. and KADILAR, C., (2009). Efficient Estimators for the Population mean, Hacettepe Journal of Mathematics and Statistics, Volume 38(2), 217-225.

MURTHY, M. N., (1967). Sampling theory and methods, Statistical Publishing Society, Calcutta, India.

PRASAD, B., (1989). Some improved ratio type estimators of population mean and ratio in finite population sample surveys, Communications in Statistics: Theory and Methods 18, 379-392.

RAO, T. J., (1991). On certain methods of improving ratio and regression estimators, Communications in Statistics: Theory and Methods 20 (10), 3325-3340.

SINGH, D. and CHAUDHARY, F. S., (1986). Theory and Analysis of Sample Survey Designs, New Age International Publisher.

SINGH, G. N., (2003). On the improvement of product method of estimation in sample surveys, Journal of the Indian Society of Agricultural Statistics 56 (3), 267-265.

SINGH, H. P. and TAILOR, R., (2003). Use of known correlation coefficient in estimating the finite population means, Statistics in Transition 6 (4), 555-560.

SINGH, H. P., TAILOR, R., TAILOR, R. and KAKRAN, M. S., (2004). An Improved Estimator of population mean using Power transformation, Journal of the Indian Society of Agricultural Statistics 58(2), 223-230.

SINGH, H. P. and TAILOR, R., (2005). Estimation of finite population mean with known coefficient of variation of an auxiliary, STATISTICA, anno LXV, n.3, 301-313.

SISODIA, B. V. S. and DWIVEDI, V. K., (1981). A modified ratio estimator using coefficient of variation of auxiliary variable, Journal of the Indian Society of Agricultural Statistics 33(1), 13-18.

UPADHYAYA, L. N. and SINGH, H. P., (1999). Use of transformed auxiliary variable in estimating the finite population mean, Biometrical Journal 41 (5), 627-636.

YAN, Z. and TIAN, B., (2010). Ratio Method to the Mean Estimation Using Coefficient of Skewness of Auxiliary Variable, ICICA 2010, Part II, CCIS 106, pp. 103-110.

www.nseindia.com/index.htm. Historical Security-wise Price Volume Data- Data for ACC - EQ from 02-01-2012 to 27-02-2012.

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 89–106

SAMPLE SURVEYS OF HOUSEHOLDS IN BELARUS: STATE AND PERSPECTIVES

Natalia Bokun[1]

ABSTRACT

The main principles, characteristics and problems of three sample surveys of households (HH), conducted by the State Statistics of Belarus are considered: 1) The Household Sample Surveys (on expenses and incomes), 2) Private Subsidiary Plots in rural areas (PSP) and 3) Labour Force Survey (LFS). For each of them the purpose, sampling plan, sample design, data collection mode, the methods of estimation and possible ways to improve the surveys are discussed.

Key words: sample fraction, territorial probabilistic multistage sampling, weighting, non-responses, private subsidiary plots, Labour Force Survey.

1. Introduction

Over 70% of Belarus's population of 9.49 million resides in urban areas. According to the Census (2009) there were 2.5 million households in rural areas and 1.1 million in urban areas. There is a big income inequality. About 20% of the population have incomes below the minimum consumer budget which is set to 1171.6 thousands of Belarusian rubles or 144.6$ for a single person. The biggest part of the household expenditures is spent on purchasing foodstuffs (37–40%). Expenditures on clothing, footwear, textiles, furniture, and household goods make up 17–18%, housing and utility are about 7–8%, and costs for education, health, culture, recreation and sport amount to 7–9%. Almost all rural residents have personal subsidiary plots. Thus, households produce about 30–35% of all agricultural products, about 89–90% of all potatoes, more than 80% of vegetables, 32–33% of eggs and 13–19% of all milk. The main information source about the household status is the census but it is complemented by three nation-wide sample surveys: the Household Sample Survey, the sampling of subsidiary plots and the Labour Force Survey. They will be described in three separate sections below, which are followed by a discussion of the future development of sample surveys and statistics in Belarus.

[1] Belarus State Economic University, Minsk. E-mail: nataliabokun@rambler.ru.

In addition, in Belarus a number of experimental surveys of health care are held (Institute of Statistics, 2005), the living standards of certain categories of workers (Institute of Statistics, 2006), consumption of alcoholic beverages (Institute of Statistics, 1999, Belarus National Academy of Sciences, 2009-2011) and public opinion polls. They are of small size and they are held irregularly. In 2005 and 2012 Multiple Indicator Cluster Surveys were held (MICS 3 and MICS 4). These surveys were conducted under the auspices of UNICEF. Despite the extensive program, the questions about illness and health are not detailed enough. For the information in the field of small businesses development, retail trade, wages in the context of professions and positions can only be obtained on the basis of industry enterprises sample surveys.

The implementation process of sampling methods in practical statistics is extremely slow. The survey of reproductive health and marketing surveys are not conducted; sample surveys of enterprises cover a limited range of issues. The priority is given to the continuous reporting.

2. Household Sample Survey

Until 1995 a survey of family budgets of working people was conducted in Belarus. The sample size was 3.5 thousand persons. Two-stage sample design was used: at the first stage the enterprises were selected within branches and then employees were selected. This principle of selection ensured representativeness of data about employees' incomes, but due to development of market relations and liberalization of labor activity the statistics of family budgets has ceased to provide objective information about amounts and sources of income. In this regard a new model for Household Surveys was developed and implemented in the statistical practice. It was based on the international standards in sample design, development tools, data processing (Metodicheskie ukazania, 1997). In accordance with the proposed methodology Household Sample Survey (HSS) has been conducted since January, 1995.

HSS is the only information basis for studying living standards. Its main purpose is to get information about the welfare of all population and particular demographic groups, detailed income and expenditure data.

The information obtained is actively used by the government, research institutes and other users. The data are used for analysis and publication to assess living standards, development of the social policy, billing the household sector SNA, in the CPI and other economical and statistical calculations.

The survey is carried out in all regions and separately in Minsk. Private households are sampled. The participation in the survey is voluntary.

The household (HH) is a group of people living together and maintaining a joint unit. Persons not belonging to any HH and living and managing a household are considered as single person HHs.

Sampling plan. The sample size is approximately 0.2% or 6000 HHs. The survey covers 49 cities and 53 rural soviets.

The sampling frame is based on the Census data. The sample design is multistage sampling. Territorial three-stage probability sampling is used:

1) at the first stage sample units are cities and rural soviets (village councils);
2) at the second stage sample units are local polling districts in cities and settlements (villages and hamlets) listed in the registers of the rural soviets (village councils);
3) at the third stage sample units are households.

At the first stage large cities are fully observed (over 72 thousands of people); small cities are selected through the sampling step, which is proportional to the population of each region. At the second and third stage systematic sampling is also used. The first unit is determined randomly.

The procedure of cities and rural soviets selection is repeated once in ten years, selection of polling districts and HHs is carried out annually.

Weighting procedure. The methodology of weighing and extrapolation data on a general population is based on assignment to each unit (HH) the corresponding weight (B_i):

$$B_i = \frac{1}{p_1 \cdot p_2 \cdot p_3}, \tag{2.1}$$

where p_1 - the probability of selecting a city or a rural soviet; p_2 - the probability of a polling district in a city, zone or rural soviet; p_3 - the probability of selecting a household within a polling district or zone.

Base HH weights are corrected *for uninhabited apartments* and *non-responses* by using overweighting procedures. Weighted cells are constructed with the usage of the following characteristics: region, type of a settlement, type of housing, size of HH. Each cell includes at least 20 HHs. On the basis of the modified cells new weights are calculated:

$$V_{ki}^{*} = \frac{\sum_{j=1}^{M_k} V_{kj} + \sum_{j=1}^{N_k} V_{kj}}{\sum_{j=1}^{N_k} V_{kj}} \cdot V_{ki}, \tag{2.2}$$

$$V_{li} = \frac{\sum_{j=1}^{L} B_{lj} + \sum_{j=1}^{K} B_{lj}}{\sum_{j=1}^{K} B_{lj}} \cdot B_{li}\,, \tag{2.3}$$

where V_{ki} *- weight of HH, which fell in the *k*th cell, corrected for non-responses; V_{ki} - weight of HH, which fell in the *k*th cell, corrected for non-residential apartments; M_k - number of non-responses in the *k*th cell; N_k - number of responses in the *k*th cell; B_{li} - base weight for the *i*th HH in the *l*th region; V_{li} - base weight for the *i*th HH in the *l*th region, corrected for non-residential apartments; $l = 1,7$ - number of the region; L – number of non-residential apartments in *l*th region; M - number of HHs left in the *l*th region.

Data collection. Data are collected with the use of face-to-face interviews using paper and pencil (PAPI). The field staff comprises 150 interviewers. Before visiting the HH the interviewer sends a copy of a letter-appeal to each selected address with the request to take part in the survey. The letter briefly describes the procedure of examination and the date of the first visit.

The sample program assumes filling in some questionnaires (living conditions, personal subsidiary plots, education, health, and employment), daily and quarterly questionnaires: expenditures on foodstuffs and nonfood products, payment of services, etc.

The main components of the survey are: the main interview (a questionnaire, which is to be filled at the beginning of the survey); four quarterly interviews (conducted in April, July, October and January); four two-weeks diaries, which the households get once every quarter and in which they have to indicate their expenditures on foodstuffs and non-food products separately for each day. More than 10 000 variables are observed in the survey.

The average response rate is 70-80%. Refusals are 1-2% of the total non-response; remaining 98% are “impossible to contact”, “not at home”, “unable to answer”, “incapacity”, “unreturned questionnaire”. State statistics bodies ensure the confidentiality of the information provided by households. This information is used exclusively for the compilation of summary statistics.

Dissemination of the HH data to users is carried out by the publication of statistical books, bulletins and with the help of the Belstat website. The main publications are: Incomes and expenditures of the population in the Republic of Belarus; Social status and living standards of the population of the Republic of Belarus; Statistical Yearbook of the Republic of Belarus.

Problems. The mechanism of quarterly HH survey samples is sufficiently worked out. Nevertheless, the survey process has such problems as: high non-response rate (up to 30%), the need of building regional and demographic sub-samples, usage of more sophisticated models of imputation, in addition the replacement of non-response data by neighboring units etc. The solutions can be: to increase the number of weighted variables, to increase the number of interviewers, and to use ratio imputation and regression imputation.

3. Sampling of Subsidiary Plots

A Personal Subsidiary Plot (PSP) is a small plot of land around the house that is worked by the holder. In Belarus the sample survey of Personal Subsidiary Plots (PSP) has been conducted since June, 2010. Its main purposes are:

- to obtain data on the output of plant growing and livestock products, the number of livestock and poultry, the size of sown areas, the amount of feed consumption of livestock and poultry, and the amount of sales;
- to calculate the gross output in agriculture;
- to develop food balance sheets and funds for personal food consumption.

Survey objects are HHs, personal subsidiary households of citizens in rural areas. These households are examined for each region. Participation in the survey is voluntary.

PSP data and results are used extensively by Belstat and other government bodies to estimate total agricultural output in Belarus and in each region, to develop regional economic policy taking into account trends in agricultural production, output of PSP.

Sampling plan. The sample size is approximately 3600 PSPs, the sample fraction is 0.32%, and the maximum relative error is 5–10%.

The sampling frame is based on the last Census data and household register. The sampling frames consist of:

- a set of districts in each region;
- a set of village councils (rural soviets) in each selected area;
- villages (settlements) in each selected village council;
- the totality of the households in each village (data household register).

Two indicators are recorded for each unit: the size of the total land area and the number of conventional livestock.

Territorial four-stage probability sampling is used (Bokun, N., (2010); Nauchno-obosnovannoe metodologicheskoe obespechenie, 2010).

At the first stage sample units are districts within the region; at the second stage – village councils within selected districts; at the third – villages within selected rural councils; at the fourth – private household farms in the selected villages.

At each stage the selection of units is based on the probability that is proportional to the sample indicators (total land area, number of conventional livestock).

The first stage. Districts are selected. In Belarus a district is an administrative unit in rural areas, providing statistical data. Therefore, in the frames of HH survey sample their maximal representativeness is desirable. The maximum sampling rate that exceeds the normal or mean can range from 50% to 90%. For the area selection it is reasonable to use the middle of the interval, i.e. sampling rate 70%. 80 districts are selected from the 118 districts ($118 \cdot 0.7 \approx 80$), which are distributed over the regions (Table 1). A systematic sampling algorithm is used: districts are ranked by the number of households. For each district the values of indicators "total area of land" and "amount of conditional livestock" are calculated for the private plots.

Table 1. The composition of district sample

Regions	Number of districts		District sampling rate, d_1	Number of households		
	in region, N_1	in sample, n_1		total, N	in selected districts, N_{11}	selected, (n), distribution by regions
Brest	16	11	0.687	210606	147614	495
Vitebsk	21	14	0.667	161100	117648	630
Gomel	21	14	0.667	174159	140204	630
Grodno	17	12	0.706	166640	127045	540
Minsk	22	15	0.682	273061	209715	675
Mogilev	21	14	0.667	128692	100381	630
TOTAL	**118**	**80**	**0.678**	**1114258**	**842607**	**3600**

The second stage. One village council is randomly chosen in each of the selected districts, which leads to a certain degree of uncertainty in the representativeness of the district. But since summary information is presented only by regions, and not by districts, it is neglected. In addition, every interviewer is assigned to one village council. The interviewer does not need to make long journeys to conduct surveys in other areas. Averages, totals, figures for the oscillation of the analyzed characteristics are estimated. Then village council of medium size with minimal values of deviations from the mean values in a district is selected.

The third stage. The list of settlements ranked by the number of households is composed for each selected village council. The settlements consisting of a small number of households (1, 2, 3 etc. HHs) are excluded. Unit selection is determined by a random number generator or by a table of random numbers.

The fourth stage. The number of subsidiary plots is determined. The disproportionate approach is used. This means that 45 households are sampled for each selected locality settlement. Household selection is done in a mechanical way using the accumulated amount of the parameters "total area of land" and "conditional livestock".

Weighting procedure. The extrapolating (here and in the following you should replace extrapolation/extrapolate by estimation/estimate) of mean and total values of sowing and harvesting areas, and all kinds of cultures, the total land area, the number of cattle, the gross collection of crops, livestock production, feed consumption of livestock and poultry and others are provided.

Extrapolation is carried out by the following methods:

1) by simplified method;
2) by the probability of selection for each of the four stages of selection;
3) by ratio estimation.

Simplified method. The methodology of weighing and estimations is based on assigning to each unit (PSP) the corresponding weight (B_e):

$$B_e = \frac{1}{p_p \cdot p_c \cdot p_n \cdot p_q}, \tag{3.1}$$

where p_p – probability of selection of region in a district; p_c – probability of selection of village council in the selected district; p_n – probability of selecting a point in the selected village council; p_q – the probability of selection of each household within a sampled settlement.

The investigator has to take into account variability of the studied parameters. Therefore, for each HH several basic weights are calculated (for crop, livestock, etc.).

Taking into account the variability of the studied indicators some basic weights are calculated for each HH (for the extrapolation of indicators for crops, livestock, etc.). For the calculation of the selection probability the size of PSP is estimated by land area and livestock. For example, the estimation of the crop base weight is determined by the formula:

$$B_{e'p} = \frac{1}{p'_p \cdot p'_c \cdot p'_n \cdot p'_q}; \tag{3.2}$$

$$p'_p = \frac{S_{pj} \cdot n_{1i}}{S_i};\ p'_c = \frac{S_{cj}}{S_{pj}};\ p'_n = \frac{S_{nj}}{S_{cj}};\ p'_q = \frac{S_{ej} \cdot n_{4j}}{S_{nj}}, \tag{3.3}$$

where p'_p, p'_c, p'_n, p'_q are selection probabilities of district, village council, village, HH, calculated taking into account the land area of PSP in the region, district, village council, village, separate HH respectively. S_{Pj} is land area of PSP in the selected *j*-th district (1st stage); S_i - total area of PSP land in the *i*-th region;

n_{1i} - the number of districts selected in the *i*-th region; n_{4j} - the number of households selected in the *j*-th village council; S_{ej} - land area in the selected *e*-th household (4 stage); S_{cj} - total land area of PSP in the selected village hall *j*-th district (2nd stage); S_{nj} - the total land area of PSP in the selected village *j*-th district (3rd stage).

Data extrapolation on the probability of selection for each of the four stages of sampling. At each stage of the selection the value of the average and total values of a characteristic are extrapolated. The calculation is made separately for livestock and crop.

IV stage. Estimation of each characteristic for the crop is carried out by the following weight:

$$B_{ep_4} = \frac{S_n}{S_e}, \tag{3.4}$$

where Bep_4 is the reciprocal of the probability of selecting PSP from the totality of human settlements on indicators of crop at stage 4; S_n – the area of PSP land in all selected locations; S_e – the land area in the *e*-th HH included in the sample.

III stage. For estimation of crop characteristics the weight of a settlement is calculated as follows:

$$B_{ep_3} = \frac{S_c}{S_{nj}}, \tag{3.5}$$

where Bep_3 – is the reciprocal of the probability of selecting a settlement in a village council selected at the stage 2; S_c – the area of PSP land in selected village council; S_{nj} – the area of PSP village council land in the selected village *j*-th region.

II stage. Weighting of the village council (for plants):

$$B_{ep_2} = \frac{S_p}{S_{cj}}, \tag{3.6}$$

where S_p – the area of household land in all selected districts of a region; S_{cj} – the area of land in the selected council in the *j*-th district.

I stage. The area weights are calculated as follows:

- crop $$B_{ep_1} = \frac{S_i}{S_{pj}}; \tag{3.7}$$

- livestock $$B_{el_1} = \frac{(S_i + Y_i)}{(S_{pj} + Y_{pj})}, \quad (3.8)$$

where S_i – the area of PSP land in the *i*-th region; Y_i – conventional livestock in PSP of *i*-th region; S_{pj} and Y_{pj} – the area of land and conventional livestock in the selected *j*-th district in the *i*-th region respectively.

Extrapolated total value of a characteristic is defined as a product of the average value of the trait and the number of households in the region, or as a sum of weighted values of a variable at the first stage.

Ratio estimation. The sample population for each region is formed. Average and total values are extrapolated using of the raising coefficients (*Kp*):

- crop $$K_{pp_1} = S : \sum_{1}^{n_4} S_e ; \quad (3.9)$$

- livestock $$K_{pl} = \frac{S + Y}{\sum_{1}^{n_4} (S_e + Y_e)}. \quad (3.10)$$

Selection of the optimal extrapolation method depends on the initial data and is determined by the minimal standard sample error. We use the classical formula for calculating the variance for multi-stage sample, as well as the variance of ratio estimators (Bokun, N., Chernysheva, T (1997); Cochran, W (1997)).

Non-response adjustment is based on the donor imputation: selection of values with replacement from the set of respondents.

The results of subsidiary plots sample survey held in Belarus in 2010 are shown in Table 2, where: X_1 is gross harvest of grains and legumes (quintals); X_2 is gross harvest of potatoes (quintals); X_3 is gross harvest of vegetables (quintals); X_4 is the number of cows; X_5 is the number of pigs.

Table 2. Sample survey of subsidiary plots in Belarus, 2010

Indicators	Total value of parameter			Sample error, %
	sample	general	estimated value	
1. Simplified extrapolation method				
X1	409	27209	23858.5	9.8
X2	1020	80306	91629.1	14.1
X3	6231.34	799409.3	945701.2	18.3
X4	33426.31	4604302.6	5460922.4	3.1
X5	6124.32	1104521.2	1129925.1	2.3

Table 2. Sample survey of subsidiary plots in Belarus, 2010 (cont.)

Indicators	Total value of parameter			Sample error, %
	sample	general	estimated value	
2. Data extrapolating on the probability of selection for each of the four stages				
X3	6231.34	799409.3	815397.48	2.0
X4	33426.31	4604302.6	4765453.1	3.5
X5	6124.32	1104521.2	1158893.6	3.9
3. Ratio estimated				
X1	409	27209	32786.8	20.5
X2	1020	80306	81831.8	1.9
X3	6231.34	799409.3	945701.2	18.3
X4	33426.31	4604302.6	4415526.2	4.1
X5	6124.32	1104521.2	976396.8	11.6

Data presented in Table 2 are examples of different estimation methods used in subsidiary plots sample surveys held in Belarus. The most preferred extrapolation methods are based on using base weights, which take into account the sizes of cultivated areas and livestock. In some cases ratio estimators are better. Additional usage of extrapolation over probabilities at each of the four selection stages is also possible (in the case of high error in the first two methods, for example, when evaluating the total yield of vegetables).

Preliminary assessment of the acceptable degree of accuracy shows that the standard relative error for the whole Belarus is 1-2%; for the regions it is 5-6%; for small-size areas it is 8-15%. The standard relative error of the sample for sown area is 5-6%; for land area – 0.1-0.5%; for the number of livestock – 5-10%; for the planted area with potatoes and vegetables – 5-5.6%.

Data collection. Face-to-face interviews are used to survey the items of interest in the questionnaire. According to the national specificities the optimal interviewer load is nearly 45 households. The data are collected by 80 field workers using paper questionnaires. Respondents maintain their accounting records of the volume of crop production, livestock, provide information about the presence and movement of poultry livestock, acreage size of family members, etc.

Five questionnaires are used: basic questionnaire (as of 1 January), questionnaire on the crop area (as of 1 June), on the presence and movement of livestock and poultry (quarterly), diary registration of crop production (5 times a year, June-October), diary of livestock products registration and feeds accounting

(monthly). The collected information is confidential and it is used for the aggregate indicators calculation.

The average response rate is 85-90%. Refusals make 60-65% of the total non-response. Main publications are: Agriculture in the Republic of Belarus; Statistical Yearbook of the Republic of Belarus. Survey results are also presented on the website www.belstat.gov.by.

Problems. The results of surveys in 2010-2011 have shown: 1) real response rate was higher than the planned one (85-90% versus 80%). This fact indicates a positive attitude of respondents to the survey; 2) at the regional level for the investigated variables (land area, crop area, number of pigs, etc.), the discrepancy between the estimates and the data of households recording are within an acceptable range (10-15%); an exception is the number of indicators of cattle, for which estimates are much lower than the continuous data records; this may be due to errors in the sampling frame: in some areas the number of livestock is overestimated, and it needs updating; 3) relative standard errors for most indicators of questionnaires did not exceed the permissible level; 4) it is quite difficult to select any option of extrapolation for various indicators of the questionnaire. Further improvement of the survey methodology may be related, firstly, to updating household recording, secondly, to the development of algorithm of choosing the optimal method for extrapolating the individual indicators (sections) of the questionnaire, and, thirdly, to study the possible application of the iterative weighting.

4. Labour force survey

Nowadays, the National Statistical Committee of the Republic of Belarus together with some foreign and national experts makes the preparatory work on implementation of the Labour Force Survey (LFS). In November 2011 a test sample survey was conducted. Since 2012 LFS has been provided on a regular basis.

The purposes are:

- to obtain empirical statistics on the labour force, economically active population, employed, unemployed;
- to obtain empirical statistics on labour force, employed, unemployed by sex, regions, rural, urban;
- to determine real labour force demand and supply.

Frequency of the results: quarterly and annual.

LFS data will be widely used for the labour market analysis, assess the actual level of unemployment, making optimal management decisions in the field of employment.

The survey covers the whole country: urban and rural areas in each region. Private households are surveyed. Participation in the survey is voluntary.

The target population comprises all residents aged 15-74.

Sampling plan. The size of the sample is perhaps the most important parameter of the sample design, as it affects the precision, cost and duration of the survey more than any other factors.

To calculate the ***sample size***, with the usage of the appropriate formula, recommended strategy for calculating the sample size is to take into account several factors, connected with sample precision, design-effect (*deff*), household size and non-responses. These factors are:

- the precision, needed relative sample error;
- desired confidence level;
- estimated (or known) proportion of the population in the specified target group;
- predicted coverage rate, or prevalence for the specified indicator;
- sample *deff*;
- average household size;
- adjustment for potential loss of sampled households due to non-response.

Design-effect *(deff)* is a ratio of sample variances of the actual stratified cluster sample (σ_a^2) and of a simple random sample of the same overall sample size (σ^2):

$$deff = \sigma_a^2 / \sigma^2 . \qquad (4.1)$$

Two sets of problems arise at this stage. First, the value of *deff* can be easily calculated after the survey, it is not often known before the survey. Second, the value of *deff* is different for each indicator and each target group. Consequently, it is necessary to choose one more important key indicator. International statistical practice has shown that the optimal value of *deff* is 1.5 (Multiple Indicator Cluster Survey Manual (2009), p. 4.3-4.8) (which may be sometimes high). Therefore, the sample size will be large enough to measure all main indicators.

Key indicator is the most important indicator that will yield the largest sample size.

Selection of the target group and key indicator includes the following stages:

1. Selection of two or three target populations that comprise small percentages of the total population (1-year, 2-year, 5-year age groups) (Multiple Indicator Cluster Survey Manual (2009), p. 4.8).
2. Review of important indicator based on these groups, ignoring indicators that have very low (less than 5%) or very high (more than 50%) prevalence.
3. Maximal indicator value, calculated for target group (10-15% of the population) is 15-20% [6; 7].
4. Do not pick from desirable low coverage indicators an indicator that is already acceptably low.

Key indicator, used in Belorussian LFS, is the real unemployment rate (by the Census results). Target groups are economically active populations (rural, urban, by regions, 5-year groups).

The sample size formula is used (Bokun, N., Chernysheva, T (1997), p. 44-53; Multiple Indicator Cluster Survey Manual (2009), p. 4.5-4.8, 4.11):

$$n = \frac{4r(1-r)\cdot f \cdot 1.2}{(0.12r)^2 \cdot p \cdot n_h}, \tag{4.2}$$

where n – required size for the key indicator; 4 – the factor to achieve 95% level of confidence, t-criteria; r – predicted prevalence for the key indicator; 1.2 – essential factor in order to raise the sample size by 20% for non-response; f – the symbol for deff (1.5); 0.12 – recommended relative sample error (95% level of confidence); p – proportion of the total population upon which the indicator (r) is based; n_h – average household size.

Several types of the sample size calculations were executed:
1) random selection for rural and urban population for each region;
2) random selection for Belarus (for target groups);
3) random selection for each region;
4) stratified sampling for each region.

The examples of sample size determination are given in Tables 3 and 4.

Table 3. Sample size for LFS. Variant 2

Target group	Real unemployment rate		Target group size		Average household size, n_h	Number of persons aged 15-74 on average, falling to one HH, n'_h	Predicted sample size	
	persons	%, r	to total population, p	to 15-74 years age group, p'			$n_1 = \frac{4r(1-r)\cdot f \cdot 1.2}{(0.12r)^2 \cdot p \cdot n_h}$	$n_2 = \frac{4r(1-r)\cdot 1.5 \cdot 1.2}{(0.12r)^2 \cdot p' \cdot n'_h}$
Economically active population aged 20-24 (565833 persons)	60627	10.7	5.95	7.5	2.43	1.94	28860	28860
Economically active population aged 15-74 in rural area (1051627 persons)	69346	6.6	11.06	14.0	2.43	1.94	26328	26052

Table 4. Sample size for LFS. Variant 3

Regions	Population aged 15-74, *N*, persons	Number of unemployment, persons	Proportion unemployed in the population aged 15-74, *w*	Number of persons aged 15-74 on average, falling to one HH, n'_h	Sample size, *n* , number of households	
					Relative standard error μ =0,06, relative limited error Δ =0,12, (without *deff*)	Relative standard error μ =0,075, relative limited error Δ =0,15, (with *deff*)
Brest region	1073227	50065	0.047	1.92	3502	3380
Vitebsk region	979845	37108	0.038	1.87	4480	4312
Gomel region	1132928	46840	0.041	1.89	4102	3946
Grodno region	829263	31757	0.038	1.87	4474	4308
Minsk	1513844	56293	0.037	2.06	4191	4043
Minsk region	1113871	37345	0.033	1.94	4997	4811
Mogilev region	868907	38511	0.044	1.97	3651	3513
Total	7511885	297919	0.040	1.94	29397	28313

Calculation results by different variants have shown that required annual sample size is 26-29 thousand of households, or in average – 28 thousand. Without taking into account non-responses the sample size is 22 thousand. Therefore, predicted sample fraction is 0.6%, or 22 000 HHs. It is planned to examine 7 000 HHs on a quarterly basis.

Sample frame is based on the 2009 Census and includes:

- set of cities in each region;
- set of village councils in each region;
- census enumeration districts in each selected city;
- villages (settlements) in each selected village council;
- the household totality in each census enumeration district and village.

Annual updating of the lists of enumeration areas and HHs is assumed.

Sample design. The territorial three-stage sample is used: primary unit – city or village council; secondary unit – census enumeration district or village (zone); final sampling unit – household.

There are 25 census enumeration districts in cities and 16 village councils (zones).

At each stage units are selected with systematic sampling with the probability that is proportional to population size or to the number of households. Variables used for the stratification are: administrative districts, urban/rural.

Weighting procedure is connected with HH weights and individual's weights. HH weights are calculated as reciprocal of overall sample probabilities:

$$B_i = \frac{1}{p_1 \cdot p_2 \cdot p_3}, \tag{4.3}$$

where p_1 - the probability of selecting a city or a rural soviet; p_2 - the probability of selecting each polling district in cities, zones and rural soviets; p_3 - the probability of selecting each household within the Census enumerated district or zone.

For the case of *non-response* an additional array of HHs is reserved within not less than 20% of the total sample ($28000 \cdot 0{,}2 \approx 6000$).

Individual's weights are based on iterative weighting (Multiple Indicator Cluster Survey Manual (2009); Metodika provedenia bazovyh obsledovanij naselenija (1997)):

Iteration I:

a) weights are calculated separately by sex within 5-year age groups;
b) the first correction coefficient (k_1) is calculated; weighted variables are: region, sex, rural/urban;
c) the second correction coefficient (k_2) is calculated; variables are: region, sex, eleven 5-year age groups .

Individual weights are equal within each region, 5-year age groups, one kind of a settlement.

Iteration II:

At the second iteration the operations are implemented on the subsequent adjustment of the basic weight and intermediate extrapolated data on the same criteria as for the first iteration.

Final individual weights for each 5-year age group:

$$K_i = B_b \cdot k_1 \cdot k_2 \cdot k_3, \tag{4.4}$$

where: $B_b = \frac{S_j}{s_j}$; $k_1 = \frac{S_t}{S_E}$; $k_2 = \frac{S_{jt}}{S_{E2}}$; S_j, s_j – population size in *j*-th sex-age group based on the result of the Census and survey; S_t – population size in *t*-th group by rural (urban), sex (on the Census data); S_E – extrapolated population size in *t*-th group (by B_b); S_{jt} – population size in *jt*-th sex-age rural (urban) group; S_{E2} – extrapolated population size in *jt*-th group (by B_b and k_1); k_3 – generic correction coefficient, calculated in the second iteration ($k_3 = k_{31} \cdot k_{32} \cdot \ldots \cdot k_{3n}$).

Preliminary results of iterative weighting for unemployment rate and employment rate, calculated for Mogilev region (Table 5) have shown that received sample population is representative. Relative errors for the region do not

exceed 7-8%: for the number of unemployed – 6%, the number of employed – 1.8%, the unemployment rate – 6.6%.

Table 5. Indicators of sample representativeness. Mogilev region. Iterative weighting

Indicators	Characteristic value		Error	
	extrapolated, $Э_x$	in the general population, x	in absolute terms, $\Delta a = \lvert x - Э_x \rvert$	in % $\Delta = \frac{\lvert x - Э_x \rvert}{x}$
Number of employed, persons	506231.11	515876	9644.89	1.87
Urban area	402333.2	412962	10628.8	2.57
- Male	194657.81	205508	10850.2	5.28
- Female	207675.39	207454	221.39	0.11
Rural area	103897.91	102914	983.91	0.96
- Male	55227.66	55228	0.34	0.0006
- Female	48670.25	47686	984.25	2.06
Total number of employed, persons				
- Male	249885.05	260736	10851	4.16
- Female	256345.64	255140	1205.64	0.47
Number of unemployed, persons	40510.33	38511	1899.33	4.19
Urban area	32094.01	29332	2762.01	9.42
- Male	20045.51	18381	1664.51	9.06
- Female	12048.50	10951	997.5	9.10
Rural area	8416.32	9179	762.68	8.31
- Male	5931.53	6572	640.47	9.75
- Female	2484.79	2607	122.21	4.69
Number of unemployed (persons) among				
- Male	25977.04	24953	1024.04	4.10
- Female	14533.29	13558	975.29	7.19
Unemployment rate, %	7.41	6.95	0.46	6.62
Urban area	7.39	6.63	0.76	10.46
- Male	9.34	8.21	1.13	13.76
- Female	5.48	5.01	0.47	9.38
Rural area	7.49	8.19	0.7	8.55
- Male	9.70	10.63	0.93	8.75
- Female	4.86	5.18	0.32	6.18
Unemployment rate (%) among:				
- Male	9.42	8.73	0.69	7.90
- Female	5.37	5.05	0.32	6.34

The results of trial calculations and testing of the first version of methodological and software sampling have shown that the main difficulties are associated with the use of different weighting schemes, determining the number of iterations steps, evaluation of structural indicators of employment and unemployment, the presence of atypical employment on the level of primary units (cities, districts).

Data collection. The data are collected by 200 field workers with face-to-face interviews using paper questionnaires. The optimal interviewer's load in the cities is 40 HHs, in rural areas – 30 HHs. The predicted response rate is 80%. The reference week is the week before the interview.

The main component of the survey is "The questionnaire on studying employment for the surveyed week". It includes 57 questions, which are combined into seven sections, and includes the details about the respondent, basic and additional paid work, self-employment, unemployment, employment in the PSPs.

Preliminary results of the survey are to be presented on the website of Belstat.

Problems. Under a given load and a limited number of interviewers (200), it is not possible to question the estimated number of HHs (28 000) on a quarterly basis. On the basis of the selected annual array of HHs (28 000), built by regions, for each quarter, randomly generated four sub-samples are formed (each includes 7 000 HHs). If the annual array of information makes it possible to obtain sufficiently representative data at the level of the republic and regions on most indicators (number of employed, unemployed, the economically active population, employment, unemployment, and in the context of all sex-age groups, urban and rural areas), the quarterly array makes it possible to design and evaluate the indicators with an acceptable degree of accuracy (10-12%) only at the level of the country. To improve the representativeness by region the indicators of the survey can be formed on the basis of the three samples – the average for three consecutive quarters. In addition, improving the quality of sample data is possible due to testing and using various schemes of the iterative weighting.

5. Concluding remarks

The household surveys make it possible to get the information on living standards of the population, actual employment and unemployment and products produced in PSPs.

The sample units are HHs and target population groups (for example, persons aged 15-74), Personal Subsidiary Plots of citizens in rural areas. The surveys cover the whole country: the regions and Minsk city. The sample fraction is at the level of 0.2-0.6% of HHs, sample frames are Census and additional databases (household survey for the PSPs). Face-to-face paper assisted interview is used.

The experience of household sample survey construction in Belarus has shown that the most applicable form of HH selection is multi-stage territorial probability sampling. The population can be stratified by the group of indicators: the administrative center, the type of housing, the size and composition of the HH. For the survey of PSPs the additional stratification variables are: the area of land, conventional livestock, and for LFS - gender and age groups of those aged 15-74. Weighting and extrapolation are carried out both on the basis of individual weights that are calculated with the usage of linear functions (e. g., the reciprocal product of the probability of selection units at various stages of the sample), and with the usage of sophisticated estimates (ratio estimators are applied for estimation of some parameters of the PSP population).

The main problems for researchers and practitioners of statistics are: the issues of sample localization, the construction of regional (district) samples, non-sampling errors, non-response (20-30%), presence of atypical units, not appropriate extrapolation, the use of different weighting schemes, the assessment of structural employment and unemployment indicators (for LFS), improving the representativeness of the quarterly data.

Possible directions for improvement of the surveys are connected with using ratio and regression imputation, demographic and territorial sub-samples, usage of combined estimation methods for each indicator, presented in questionnaire (PSP), clarifying the steps and subsequent realization of iterative weighting scheme (LFS). It would be interesting to evaluate the goodness of sample strategy by means of Monte Carlo simulation from the census data (LFS) and household register data (PSP).

REFERENCES

BOKUN, N., (2010). Sampling of Subsidiary Plots in Belarus: methodological problems of population formation and data estimation. *Workshop on Survey sampling theory and methodology*. August, 23-27. Vilnius, Lithuania.

BOKUN, N., CHERNYSHEVA, T., (1997). Metody vyborochnyh obsledovanij, Minsk.

COCHRAN, W., (1997). Sampling techniques. *John, Willey and sons, inc.* New-York.

Metodicheskie ukazania po vyborochnomu obsledovaniu domashnih hoziajstv v Respublike Belarus. (1997). Minsk.

Metodika provedenia bazovyh obsledovanij naselenija. (2008). Kiev.

Metodologichni osnovi formuvannia viborkovyh sukupnostej dlia provedennia organami derzhavnoj statistiki Ukraini bazovyh derzhavnyh viborkovyh obstezhen naselenia (domogospodarstv). (2005). – 156 p., Kiev.

Money incomes and expenditures of population of the Republic of Belarus: statistical book. (2011). National Statistical Committee of Republic of Belarus, Minsk.

Multiple Indicator Cluster Survey Manual. (2009). Eurostat.

Nauchno-obosnovannoe metodologicheskoe obespechenie po formirovaniu vyborochnoj sovokupnosti lichnyh podsobnyh hoziajstv grazhdan, postojanno prozhivajuschih v selskoj mestnosti: otchet o NIR № GR. (2010). Nauchnyj rukovoditel – N. Bokun, BSEU, Minsk.

Social conditions and living standards of population in the Republic of Belarus: statistical book. (2012). National Statistical Committee of Republic of Belarus, Minsk.

Statistical Yearbook: statistical book. (2012). National Statistical Committee of Republic of Belarus, Minsk.

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 107–128

MATERNAL NUTRITIONAL STATUS AND LACTATIONAL AMENORRHEA IN INDIA: A SIMULATION ANALYSIS

Laxmi K. Dwivedi[1]

ABSTRACT

Apart from breast-feeding, socio-economic and biological factors, maternal health also influences the length or distribution of waiting time to conception. The main objective of this paper is to examine the linkages between maternal nutritional status (measured by body mass index-BMI) and postpartum amenorrhea among currently breast-feeding women in India and its region. Further, the probability to remain amenorrheic through simulative approach has been estimated to get better understanding of the impact of maternal nutritional status on postpartum amenorrhea. Using National Family Health Survey-2 data, women who were not pregnant, who were breast-feeding and who were not using any hormonal contraceptives at the time of the survey were included in the analysis. Missing cases for body mass index and child nutritional status were imputed by fitting the linear regression equation. There was no significant difference existing between mean BMI of each region of India before and after imputation of missing cases. The interaction term between maternal nutritional status and duration of breast-feeding (child's age) was significantly associated with the likelihood of having resumed menstruation after controlling for breast-feeding practices, child nutritional status and socio-economic and demographic covariates. The effect of maternal nutritional status on lactational amenorrhea was not found to be significant when women were breast-feeding since last 12 months except in the northern region of India. However, after 12 months of breast-feeding, the probability of undernourished women to remain amenorrheic was likely to be greater and this trend was highly consistent across all the six regions included in the analysis.

Key words: simulative approach, maternal nutritional status, body mass index, postpartum amenorrhea, India.

[1] Assistant Professor, Department of Mathematical Demography & Statistics, International Institute for Population Sciences, Govandi Station Road, Deonar, Mumbai-400088, India.
E-mail: laxmikdwivedi@gmail.com, laxmikant@iips.net.

1. Introduction

The interval between birth of a child and subsequent return of menstrual cycle is known as lactational amenorrhea. There are number of important factors which affect lactational amenorrhea either directly or through breast-feeding and its complete understanding is somewhat a complex phenomenon. Apart from breast-feeding, socio-economic and biological factors, maternal health also influences the length or distribution of waiting time to conception. It is easily visualized that maternal health affects the duration/frequency of breast-feeding. For instance, an undernourished woman might think that her milk is not sufficient and/or nutritious or she will not be competent enough to breast-feed for a longer duration and ultimately it adversely affects the duration of amenorrhea. The other possibility is that an undernourished woman may prefer to increase the frequency as well as duration of breast-feeding because she might be not capable to produce sufficient nutritious milk for her child.

However, nutritionists argued that the nutritional status of woman is also directly linked with the quality and duration of breast-feeding. Frisch (1983) found that nutritional intake influences fecundity. Further, Frisch *et al.* (1973) and Frisch and McArthur (1974) investigated the effect of nutrition on ovarian function and they have formulated the "critical body composition hypothesis." This hypothesis suggests that a minimal amount of fat as percentage of body weight is necessary for attaining menarche and for maintaining ovarian cycles. However, it still remains controversial and some researchers have suggested that nutritional status of women has a strong impact on postpartum amenorrhea (PPA).

The arguments that shorter duration of breast-feeding results in short duration of birth interval may deteriorate the nutritional status of mother. The term 'maternal depletion syndrome' in the literature refers to "the effect of a rapid succession of pregnancies and periods of lactation which erode the nutritional status of the mother" (Cleland and Sathar, 1984). There have been a few studies dealing with the effects of birth interval on maternal mortality (measurement of maternal health) due to non-availability of data. It is very difficult to measure the health effect of high fertility or short birth intervals on mothers. However, it is also argued that longer duration of breast-feeding has a negative impact on health. It is not easy to measure maternal health due to intense and longer breast-feeding in the analyses of reproductive performance and health from currently available data.

India and the central region in particular are well known for high fertility leading to the burden on women who already have poor nutritional status. Therefore, it is felt that the duration of amenorrhea might change for well-nourished and undernourished women. Huffman *et al.* (1987) suggested maternal nutrition is not likely to shorten the length of PPA significantly. Moreover, the relationship between maternal nutritional status and lactational amenorrhea is not clearly understood. Some researchers have argued that undernourished women have less chance of maintaining the ovarian cycle (Frisch *et al.*, 1973; Frisch and

McArthur, 1974). But some researchers have also argued that this relationship is biologically insignificant (Diaz *et al*., 1988). Therefore, the specific objective of this chapter is to examine the independent impact of maternal nutritional status on lactational amenorrhea among breast-feeding women.

2. Method and materials

This study uses National Family Health Survey (NFHS) data conducted in the years 1998-99. The analysis was carried out for India and its six regions, namely - the northern region includes Delhi, Haryana, Himachal Pradesh, Jammu and Kashmir, Punjab and Rajasthan; the central region consists of Chhattisgarh, Madhya Pradesh, Uttaranchal and Uttar Pradesh; the eastern region comprises of Bihar, Jharkhand, Orissa and West Bengal; the northeastern region consists of Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim and Tripura; the western region includes Goa, Gujarat and Maharashtra; and finally the southern region includes Andhra Pradesh, Karnataka, Kerala and Tamil Nadu.

Regions follow the classification scheme contained in the NFHS published report (IIPS and ORC Macro, 2000). Region specific analysis was carried out after assigning a proper weight to adjust for the differences in sample size across states. Sample weights were calculated to provide region-wide estimates, for example, for the northeastern region (which contains eight states):

$1/[8*(n_s/ n_p)]$ where n_s is the sample size for each state and n_p is the sample size for pooled data. Whenever it is required, values for missing cases have been imputed using linear regression equation and results have compared before and after imputing the missing values.

Women who were not pregnant, not using any hormonal contraceptives and were currently breast-feeding at the time of the survey were selected for the study. The NFHS-2 data obtained information from 90,303 ever married women in the age group of 15-49 years. There were 22,597 women who were currently breast-feeding at the time of the survey, of whom 959 women who were currently pregnant and 737 women who were using modern contraceptive pills were dropped from the analysis. The analysis was carried out for 20,901 currently breast-feeding women. At the time of the survey, women were asked if their menstruation had returned since the birth of their youngest child. The lactational amenorrhea is defined as a dichotomous variable. The dependent variable was the women who reported not to have resumed menstruation after the delivery of the last child and were coded as amenorrheic or as non-amenorrheic if women resumed menstruation.

An anthropometric measurement, body mass index (BMI) has been used as an indicator for measuring nutritional status of women. Chronic energy deficiency in women is usually indicated by BMI of less than 18.5 kg/m^2. BMI is a valid indicator for assessment of nutritional status of women as literature suggests that

BMI is consistently highly correlated with body weight and is relatively independent of the stature or height of the individual. Such type of measurement is very highly reliable than the measurement which is solely based on reporting. However, child weight-for-age is identified as an indicator for child's nutritional status.

The combined variables, namely maternal nutrition and duration of breast-feeding (equivalent to the age of the child) is considered as an independent variable. The four categories of this variable are:

undernourished (BMI<18.5 kg/m^2) women and child aged <=12 months;
undernourished (BMI<18.5 kg/m^2) women and child aged 13-35 months;
well-nourished (BMI>=18.5 kg/m^2) women and child aged <=12 months;
and well-nourished (BMI>=18.5 kg/m^2) women and child aged 13-35 months.

The other independent variables are: region (north/central/east/northeast/west/south); place of residence (rural/urban); respondent's education (illiterate/middle school complete/high school complete and above); standard of living (low/medium/high); sex of index child (female/male); maternal age (in years) (15-24/25-34/35-49); parity (1/2/>=3); child's weight-for-age Z-score (>=-2/<-2); and breast-feeding status (breast-feeding with supplements/exclusive breast-feeding/breast-feeding with plain water only). Analyses were also carried out in India and for six regions, separately.

The information on maternal body mass index (BMI) for 1795 women and child's weight-for-age for 4130 cases was found to be missing. The missing values of maternal BMI and child's weight-for-age were imputed with the help of multiple linear regression analysis. The significance level of coefficients in the multivariate framework was compared before and after imputing the missing values. The independent variables used for imputing the missing values of maternal BMI are region (north/central/east/northeast/west/south), place of residence (rural/urban), respondent's education (illiterate/middle school complete/high school complete and above), standard of living (low/medium/high), parity (continuous), current age of woman (continuous), and breast-feeding status (breast-feeding with supplements/exclusive breast-feeding/breast-feeding with plain water only). In addition to the above mentioned covariates, current age of child was also included for imputing the missing values of child weight-for-age in the regression analysis.

The mean value of maternal nutritional status (BMI) was computed by selected characteristics of women. Further, the survival probability of the pattern of PPA was estimated using the non-parametric method of Kaplan-Meier (K-M). Log-rank test has been applied to determine whether there were significant differences in the median duration of PPA between undernourished and well-nourished women.

Since we have controlled the duration of breast-feeding by creating the combined variable of maternal nutritional status and duration of breast-feeding, we have preferred multiple logistic regression analysis over Cox hazards model.

A simulative approach has been adopted in this paper to find out the adjusted effects of maternal nutritional status on lactational amenorrhea. It is decided to compute the adjusted proportion of amenorrheic women by maternal nutritional status on the assumption that all women in the sample were having undernourished child and also by considering all women in the sample were having well-nourished child, separately. From this approach, an attempt was made to obtain the important information of proportion of remaining amenorrheic for a particular variable by keeping other variables at the mean level. Similarly, these probabilities for a particular combination of variables may also be computed by holding the remaining variables at the mean level (Dwivedi, 2006; Dwivedi *et al.*, 2007).

3. Results

3.1. Kaplan-Meier (K-M) survival analysis

Kaplan-Meier (K-M) survival analysis was carried out to find the median duration of PPA according to maternal nutritional status. The technique also helps in proper categorization of duration of breast-feeding (child's age) as a categorical predictor in the multivariate analysis.

The results of K-M survival probability of PPA in relation to BMI for those women who were currently breast-feeding, not pregnant and were not using any hormonal contraceptives at the time of the survey for India and its regions are presented in the Figure 1. It is evident that at the first month of PPA period, 88 percent of better-nourished women and 90 percent of malnourished women in case of India were still amenorrheic, whereas at the end of six months, 57 percent of better-nourished and 63 percent of malnourished women in case of all India were still amenorrheic. However, at the end of 12 months, these rates for two different categories of women declined to about half of the previous values, i.e., to 26 percent and 32 percent, respectively. Later, at the end of 32 months, the corresponding figures were eight percent and six percent, respectively. It is clear from the present analysis that undernourished women were more likely to be found in amenorrheic state than better-nourished women (Figure 1).

In the case of better-nourished women, the percentage of women who were still amenorrheic at the end of first month was highest in the northeastern region and lowest in the northern region. But for undernourished women, the percentage of women who were still amenorrheic at the end of first month was highest in the southern region and lowest in the northern region. The percentage of women who remained amenorrheic at the end of first month was found to be relatively higher among malnourished women in the entire region except in the northeastern region where the corresponding percentage was equal for both better-nourished and malnourished women. But, the percentage of women who were amenorrheic at the end of 12 months was considerably higher among malnourished in the entire

region of India. At the end of 20 months also, a similar pattern was observed. But, the difference in percentages to remain amenorrheic between better-nourished and malnourished women was remarkably higher in the western and southern regions. For example, at the end of 12 months, the percentages of women remaining amenorrheic among undernourished was 31 percent compared to 18 percent who were well-nourished in the western region of India. Kaplan-Meier estimates show that there has been a sharp decline in the percentages of amenorrheic women during 11 to 12 month of the postpartum period in all the regions of India, irrespective of maternal nutritional status.

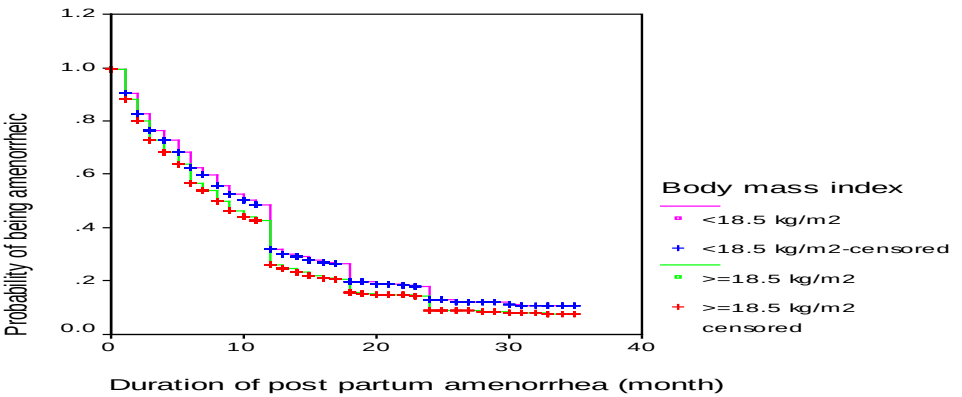
The median duration of PPA and its 95 percent confidence interval estimates for women who were currently breast-feeding, not pregnant and were not using any hormonal contraceptives at the time of the survey with respect to their body mass index in case of India and its regions have also been calculated and are presented in Table 1. Log-rank test showed that there was a significant difference in the duration of amenorrhea between the two groups under study. Undernourished women had a significantly longer duration of PPA than well-nourished women in India ($p<0.00001$).

Table 1. Median duration of postpartum amenorrhea and its 95% confidence interval (CI) estimates for women who were currently breast-feeding, not pregnant and were not using any hormonal contraceptives at the time of the survey with respect to body mass index of women in India and its regions-1998-99.

Country/Regions	BMI>=18.5 kg/m^2			BMI<18.5 kg/m^2			Log-rank test Test-statistic
	Median	95% CI		Median	95% CI		
		L	U		L	U	
India	**9.00**	**8.68**	**9.32**	**11.00**	**10.79**	**11.21**	**76.90***
North	7.00	6.49	7.51	8.00	7.14	8.86	5.38***
Central	12.00	11.73	12.27	12.00	11.85	12.15	2.86
East	10.00	9.56	10.44	12.00	11.59	12.41	14.18**
Northeast	8.00	7.48	8.52	9.00	8.01	9.99	3.64
West	6.00	5.36	6.64	11.00	10.39	11.61	45.53*
South	7.00	6.54	7.46	9.00	8.16	9.84	35.62*

Note: *p<=0.00001; ** p<=0.0002; ***p<=0.0204

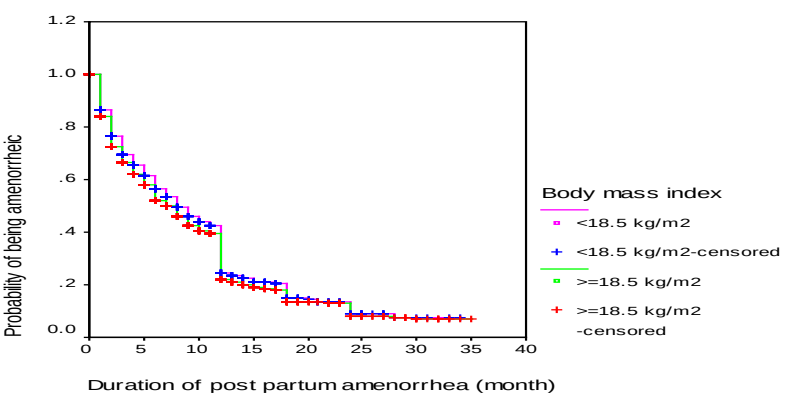
Figure 1. Survival curve based on Kaplan-Meier method for India



The median duration of PPA was significantly longer for undernourished women in the northern, eastern, western and southern regions and difference was significantly more apparent in the western and southern regions. However, there was no significant difference in the median duration of PPA between better-nourished and malnourished women in the central and northeastern regions. The survival curves based on Kaplan-Meier method for all the regions of India presented in the figures, also indicates more clearly the same findings (Figure 1).

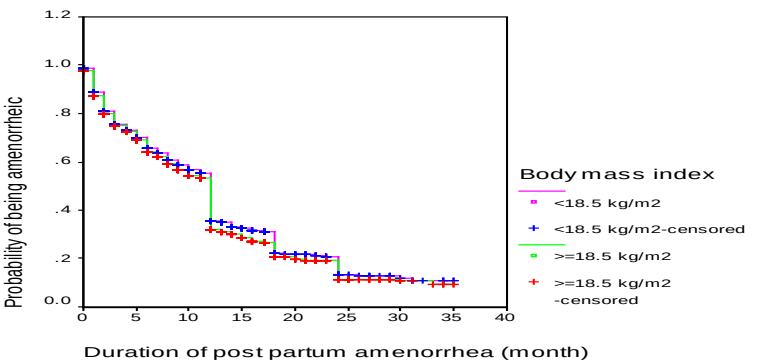
Cont…

Figure 1.1. Survival curve based on Kaplan-Meier method for Northern region of India



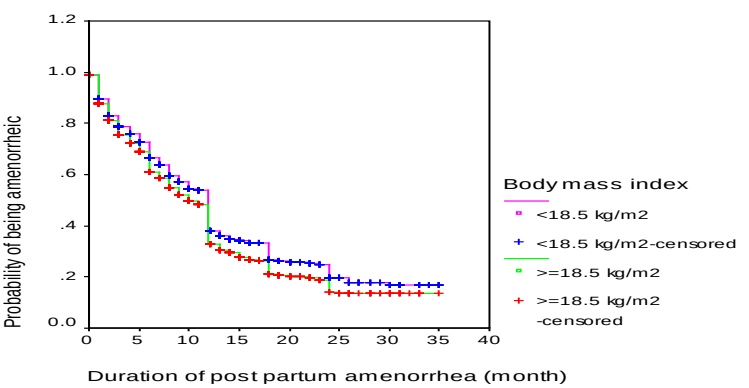
Cont…

Figure 1.2. Survival curve based on Kaplan-Meier method for Central region of India



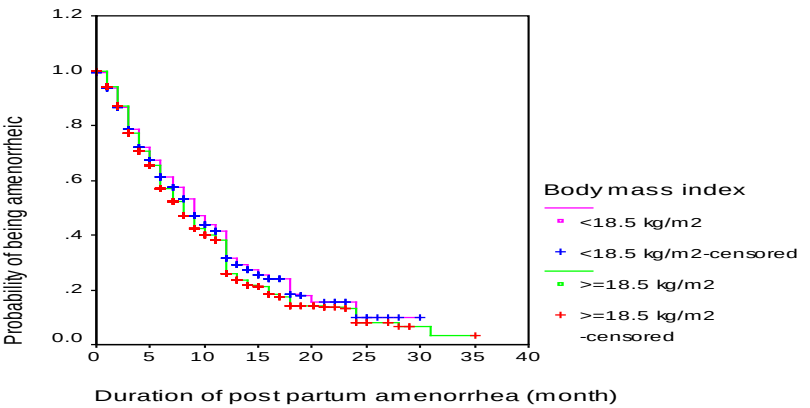
Cont…

Figure 1.3. Survival curve based on Kaplan-Meier method for Eastern region of India



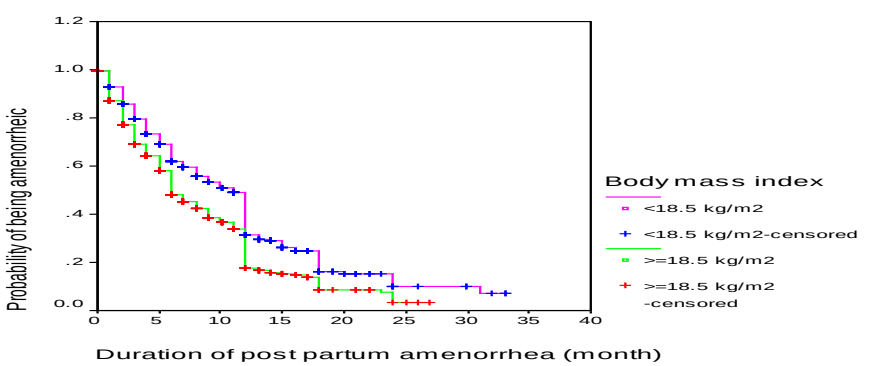
Cont…

Figure 1.4. Survival curve based on Kaplan-Meier method for Northeastern region of India



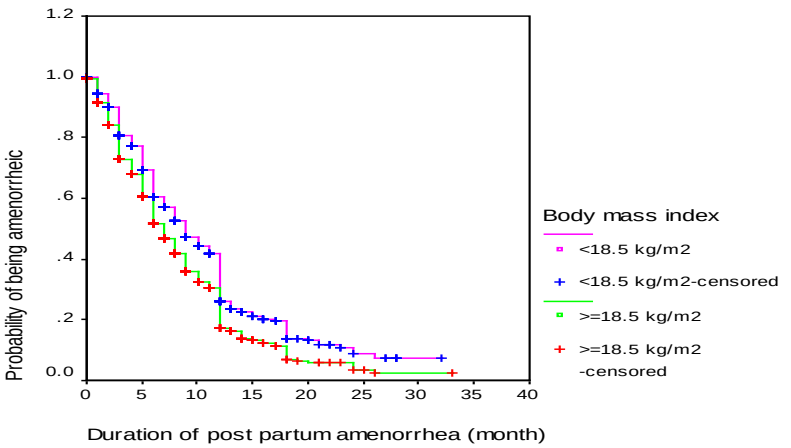
Cont…

Figure 1.5. Survival curve based on Kaplan-Meier method for Western region of India



Cont…

Figure 1.6. Survival curve based on Kaplan-Meier method for Southern region of India



3.2. Multivariate analysis

To identify the independent effect of maternal nutritional status on lactational amenorrhea among breast-feeding women, all the important determinants were considered in the multivariate logistic regression analysis. The choice of the explanatory variables included in the logistic model was governed by two considerations: first, its relation with the dependent variable should be statistically significant in the bivariate analysis; and second, the inclusion of that variable could also be theoretically justified.

The adjusted odds ratio and its 95 percent confidence interval estimates of amenorrheic versus non-amenorrheic women who were currently breast-feeding, not pregnant and were not using any hormonal contraceptives at the time of the survey for India by selected characteristics are presented in Table 2. The results presented in the table clearly show that the adjusted chance to remain amenorrheic was more evident among women of all regions of India except in the southern region, although the odds ratio was found to be significant only in the central and eastern regions. There was no variation in the magnitude as well as significance level of odds ratio after imputing the missing values. However, the magnitude of the adjusted odds ratio increased slightly compared to the unadjusted odds ratio in the eastern region. This indicates that after controlling the other important factors, women from the eastern region had strong positive tendency to remain amenorrheic than those in the northern region of India.

Women who were from urban areas belonged to educated category and had medium or high standard of living, were significantly less likely to be

amenorrheic. This pattern was almost similar after imputing the missing values. Women with a male child were more likely to be amenorrheic, but the relationship was not statistically significant. Maternal age was also inversely associated with lactational amenorrhea. However, parity was positively associated with the chance that the woman will be in the state of PPA. The adjusted chance to remain amenorrheic increases significantly with an increase in parity. For example, women who were at parity three or above had greater likelihood of being found in amenorrheic state at the time of the survey; the chance was double than that for women of parity one. The result was statistically significant. After imputing the missing values of maternal BMI and child weight-for-age, the magnitude of odds ratio for parity and maternal age remained unchanged for the same level of significance. Women who had undernourished children were significantly less likely to remain amenorrheic than women with well nourished children. The likelihood was almost similar but the level of significance was changed after imputing the missing values. Among women who did not report their child's weight-for-age the chances to remain amenorrheic increased significantly. Further, the direction of the regression coefficient was the same as it was in the case of unadjusted coefficient. Women who were exclusively breast-feeding or breast-feeding as well as giving plain water were significantly more likely to remain amenorrheic, but the odds ratio was relatively higher among those women who were exclusively breast-feeding.

Table 2. Adjusted odds ratio and 95% confidence interval (CI) estimates of amenorrheic versus non-amenorrheic women who were currently breast-feeding, not pregnant and were not using any hormonal contraceptives at the time of the survey, India by selected characteristics-1998-99.

Variables	India			India*		
	Exp (β)	95% CI for Exp (β)		Exp (β)	95% CI for Exp (β)	
		Lower	Upper		L	U
Region of residence						
North®	1.00	-	-	1.00	-	-
Central	1.15	1.03	1.27	1.15	1.04	1.28
East	1.28	1.15	1.43	1.30	1.16	1.45
Northeast	1.06	0.94	1.20	1.04	0.92	1.17
West	1.05	0.93	1.20	1.05	0.92	1.20
South	0.94	0.83	1.06	0.94	0.83	1.06
Place of residence						
Rural®	1.00	-	-	1.00	-	-
Urban	0.81	0.75	0.89	0.79	0.73	0.87
Respondent's education						
Illiterate®	1.00	-	-	1.00	-	-
Middle school complete	0.79	0.73	0.86	0.77	0.71	0.84
High school complete and above	0.74	0.66	0.84	0.72	0.64	0.82
Standard of living						
Low®	1.00	-	-	1.00	-	-
Medium	0.74	0.69	0.80	0.72	0.67	0.78
High	0.54	0.48	0.61	0.52	0.46	0.59
Sex of child						
Female®	1.00	-	-	1.00	-	-
Male	1.06	1.00	1.14	1.07	1.00	1.14
Maternal age (in years)						
15-24®	1.00	-	-	1.00	-	-
25-34	0.88	0.81	0.96	0.89	0.82	0.97
35-49	0.77	0.66	0.88	0.78	0.67	0.90
Parity						
1 Child	1.00	-	-	1.00	-	-
2 Children	1.44	1.31	1.58	1.44	1.31	1.59
>=3 Children	2.11	1.91	2.34	2.16	1.95	2.39
Child wt-for-age						
Z-score >=-2®	1.00	-	-	1.00	-	-
Z-score < -2	0.99	0.91	1.08	0.91	0.84	0.99
Missing	1.25	1.12	1.40	NA	NA	NA
Breast-feeding status						
Breast-feeding+supplements®	1.00	-	-	1.00	-	-
Exclusive breast-feeding	5.32	4.74	5.98	4.51	4.01	5.08
Breast-feeding+ plain water only	2.45	2.22	2.69	2.27	2.06	2.51
Maternal BMI & Child age						
<18.5Kg/m^2 and 13-35 months®	1.00	-	-	1.00	-	-
<18.5Kg/m^2 and <=12 months	6.10	5.43	6.86	6.14	5.46	6.90
>=18.5Kg/m^2 and <=12 months	6.71	6.01	7.49	6.78	6.08	7.57
>=18.5Kg/m^2 and 13-35 months	0.76	0.68	0.86	0.77	0.70	0.86
Missing BMI and 0-35 months	1.91	1.63	2.25	NA	NA	NA

Note: *Odds ratio includes imputed values for missing cases of body mass index and child wt-for-age.

Table 3. Adjusted odds ratio and 95% confidence interval (CI) estimates of amenorrheic versus non-amenorrheic women who were currently breast-feeding, not pregnant and were not using any hormonal contraceptives at the time of the survey for India and its regions by maternal body mass index (BMI) & child age-1998-99.

Variables	India			India*			North*			Central*		
	Exp (β)	95% CI for Exp (β)		Exp (β)	95% CI for Exp (β)		Exp (β)	95% CI for Exp (β)		Exp (β)	95% CI for Exp (β)	
		Lower	Upper		Lower	Upper		Lower	Upper		Lower	Upper
Maternal BMI & Child age												
<18.5Kg/m^2 and <=12 months®	1.00	-	-	1.00	-	-	1.00	-	-	1.00	-	-
<18.5Kg/m^2 and 13-35 months	0.16	0.15	0.18	0.16	0.15	0.18	0.17	0.13	0.22	0.15	0.12	0.20
>=18.5Kg/m^2 and <=12 months	1.10	1.00	1.21	1.11**	1.00	1.22	1.34	1.10	1.64	1.02	0.83	1.26
>=18.5Kg/m^2 and 13-35 months	0.13	0.11	0.14	0.13	0.11	0.14	0.14	0.11	0.18	0.14	0.11	0.18
Missing BMI and 0-35 months	0.31	0.27	0.37	NA	NA	NA	NA	NA	NA	NA	NA	NA

Note: *Odds ratio includes imputed values for missing cases of body mass index and child wt-for-age.
** Results significant at P<=0.05. It includes 1 in 95% confidence interval because of rounding.
All other considered variables in the Table 2 have been controlled.

Cont…

Table 3.1. Adjusted odds ratio and 95% confidence interval (CI) estimates of amenorrheic versus non-amenorrheic women who were currently breast-feeding, not pregnant and were not using any hormonal contraceptives at the time of the survey for different regions of India by maternal body mass index (BMI) & child age-1998-99.

Variables	East			Northeast			West			South		
	Exp (β)	95% CI for Exp (β)		Exp (β)	95% CI for Exp (β)		Exp (β)	95% CI for Exp (β)		Exp (β)	95% CI for Exp (β)	
		Lower	Upper		Lower	Upper		Lower	Upper		Lower	Upper
Maternal BMI & Child age												
<18.5Kg/m^2 and <=12 months®	1.00	-	-	1.00	-	-	1.00	-	-	1.00	-	-
<18.5Kg/m^2 and 13-35 months	0.20	0.16	0.26	0.18	0.12	0.26	0.15	0.11	0.22	0.11	0.08	0.16
>=18.5Kg/m^2 and <=12 months	1.16	0.92	1.46	1.23	0.93	1.62	0.81	0.60	1.09	0.96	0.74	1.24
>=18.5Kg/m^2 and 13-35 months	0.15	0.12	0.19	0.12	0.09	0.17	0.09	0.06	0.13	0.07	0.05	0.10

Note: Odds ratio includes imputed values for missing cases of body mass index and child wt-for-age.
All other considered variables in the Table 3.5.1 have been controlled.

Women who had a child of age less than 13 months were significantly more likely to remain amenorrheic, irrespective of their nutritional status, than undernourished women who had a child of age 13 to 35 months and the odds ratio was found to be similar after imputing the missing values. On the other hand, the adjusted chance to remain amenorrheic was found to be lower among women who were better-nourished and had a child of age 13 to 35 months compared to their undernourished counterparts who had a child of the same age. The result was consistent after imputation of missing values of maternal BMI and child weight-for-age.

A possible reason for this finding may be that the duration of breast-feeding (child's age) is one of the important predictors in determining the lactational amenorrhea among breast-feeding women. Therefore, for comparison purpose, we have taken undernourished women with the age of the child less than 13 months as a reference category in the multivariate analysis (Table 3). Results clearly revealed that there was no significant difference in postponing the return of ovulation after birth of a child between undernourished and better-nourished women with the age of the child less than 13 months in all the regions of India except the northern region. In the northern region, the probability to remain amenorrheic was high among better-nourished women with the age of the child less than 13 months as opposed to undernourished women with the same age of the child.

3.3. Simulation analysis

The predicted probabilities to remain amenorrheic were calculated for a particular variable by holding all remaining variables at their average level in the model. These results are presented in the Table 4.

Some selected variables and a combination of variables considered in the present prediction analysis are: (i) breast-feeding status, (ii) place of residence, (iii) respondent's education, (iv) standard of living, (v) sex of child, (vi) child nutritional status, (vii) maternal BMI and child age (duration of breast-feeding), (viii) maternal BMI and child age with breast-feeding status, (ix) maternal BMI and child age with child nutritional status, (x) maternal BMI and child's age with breast-feeding status and child's nutritional status.

The results indicate that the probability to remain amenorrheic was comparatively higher among those women who were exclusively breast-feeding. By assuming that all women in India were exclusively breast-feeding, the chance to remain amenorrheic has increased to around 13 percent than among those women who were breast-feeding and giving only plain water. The probability to remain amenorrheic was more than average among those women who were breast-feeding and giving only plain water, whereas the corresponding probability for women who were breast-feeding as well as giving supplements was lower in comparison to the average value (0.418).

A similar pattern was found in all the regions of India where the percentage of amenorrheic women who were breast-feeding and giving some supplements was lower than the average value of the respective regions. The gain in terms of percentage of women who continued PPA period due to practice of exclusive breast-feeding was high in the western region followed by the eastern and the southern parts of the country and it was lowest in the northeastern region.

Living in urban areas was inversely associated with amenorrhea compared with living in rural areas and the rural-urban differentials were more pronounced in the northern and the eastern parts of the country. The probability to remain amenorrheic among rural women was comparatively high in the western regions followed by those in the eastern and central regions and was lowest in the southern region. As education and standard of living increases, the percentage of amenorrheic women decreases. However, the pattern was not consistent with respect to education of women in the western and the eastern regions of India. By assuming that all women in the sample were educated up to high school and above, the chance to remain amenorrheic has reduced considerably in all the regions of India except the northeastern, the southern and the western regions of the country. This change in the probability was more apparent in the western region followed by the north and the northeast regions of India. Similarly, the chance to remain amenorrheic has reduced in the entire region on the assumption that all women have a higher standard of living, and this change was more manifested in the central and eastern regions. Women with male child were slightly more likely to be amenorrheic than women with girl child except in the western region. However, if it is assumed that all women in India had only male child, the percentage of women with lactational amenorrhea has reduced by around five percent from the mean value.

The percentage of amenorrheic women was higher among those with malnourished children than their counterparts with well-nourished children in all the regions of India except the south India. However, this difference was almost negligible in the southern region. The probability to remain amenorrheic was highest among those women who were better-nourished and had a child of age less than 13 months in the central, east, north and northeast regions of India, whereas in the southern and western regions, the chance was found to be highest among those women who were undernourished and had a child of age less than 13 months. Further, more undernourished women with a child of age 13-35 months were amenorrheic than better-nourished women with a child of the same age in all the regions of India. If it is assumed that all women in the sample had a child of age less than 13 months, the probability to remain amenorrheic was higher than the average value, irrespective of their maternal nutritional status. This result is true for all the regions of India. In addition, there was a substantial reduction in the percentage of amenorrheic women among better-nourished women with a child of age 13-35 months in all the regions of India from their respective mean values. The reduction in probability was more apparent in the central parts of the country followed by the southern region.

The percentage of amenorrheic women was found to be highest among those who were exclusively breast-feeding regardless of their BMI and child's age. Further, the probability to remain amenorrheic was found to be highest in the western region among those undernourished women who had a child of age less than 13 months. If it is assumed that all women in the sample were exclusively breast-feeding a child of age 13-35 months then the probability to remain amenorrheic increases from the average value among undernourished women in all the regions of India. Whereas, this probability was lower than the average value among better-nourished women with a child at the same age of 13-35 months in all the regions of India except the central, eastern and western parts of the country. Moreover, after keeping the age of the child as 13-35 months, the probability to remain amenorrheic was high among undernourished women than their well-nourished counterparts in all the regions of India.

If it is assumed that all women were exclusively breast-feeding a child of age less than 13 months, the probability to remain amenorrheic was higher among those women who were malnourished than well-nourished women with a child of the same age in the southern and western regions. The result has become inverted in other regions of India. However, after making unvarying the age of a child as less than 13 months, the chance to remain amenorrheic was not consistent across the different regions of India by maternal nutritional status.

The percentage of amenorrheic women was found to be relatively higher among those who were breast-feeding and giving plain water compared to those women who were breast-feeding and giving supplements, irrespective of maternal BMI and child's age, in all the regions of India. If it is considered that all women were breast-feeding with plain water only, the probability to remain amenorrheic increases from the average value when women had a child of age less than 13 months regardless of their BMI in the entire region. However, it decreases from the average value when women had a child of age 13-35 months in all the regions of India. On the other hand, the level of corresponding probability decreases but the pattern was the same if it is considered that all women were breast-feeding and giving supplements. With regard to women who were breast-feeding with plain water or any supplements, the probability was relatively higher for well-nourished women who had a child of age less than 13 months in the entire region except in the southern and western regions where the corresponding highest figure was for undernourished women with the age of the child less than 13 months.

When child's weight-for-age and maternal BMI with child age are taken into consideration, it is evident that the chance of remaining amenorrheic was comparatively higher among women whose children were better-nourished, regardless of maternal BMI and child's age except in the southern region. On assuming the age of the child as 13-35 months, the percentage of amenorrheic women was found highest among those undernourished who had a well-nourished child in all the regions of India. If it is considered that all women in India had malnourished children of age less than 12 months, the probability of remaining amenorrheic increases from the average value among all women regardless of

their nutritional status. This pattern was consistent in all the regions of India (table not shown).

Once child's weight-for-age, maternal BMI with child's age and breast-feeding status are considered together, the probability was found to be highest among women who were exclusively breast-feeding and had better-nourished children regardless of their BMI and child's age. The pattern was found to be the same in all the regions except in the southern region where the chance was comparatively higher among women who were exclusively breast-feeding and had undernourished children regardless of their BMI and child's age. The chance of remaining amenorrheic was highest among better-nourished women who were exclusively breast-feeding and had better-nourished children of less than 13 months in all the regions of India except the southern region, and lowest among better-nourished women who were breast-feeding and giving supplements and had better-nourished children of age 13-35 months in all the regions of India. When it is assumed that all women in India were undernourished with malnourished children of age 13-35 months and were exclusively breast-feeding, the probability of remaining amenorrheic increases around four points from the average value. This increase in probability was found to be highest (13 points) in the eastern region and lowest in the northern region (one point). However, for the central and northeastern regions, the figure was lower than the average value of the respective regions (table not shown).

Table 4. Estimated probabilities of remaining amenorrheic by selected combinations of characteristics for India and its regions-1998-99.

Variable	Probability ± Standard Deviation						
	India	North	Central	East	Northeast	West	South
Average	0.418±0.282	0.388±0.284	0.457±0.283	0.445±0.275	0.392±0.266	0.406±0.301	0.394±0.296
Breast-feeding status							
Breast-feeding+supplements	0.359±0.222	0.329±0.220	0.399±0.232	0.390±0.216	0.357±0.228	0.320±0.213	0.321±0.217
Exclusive breast-feeding	0.642±0.222	0.620±0.235	0.619±0.226	0.688±0.189	0.590±0.246	0.702±0.219	0.652±0.245
Breast-feeding+ plain water only	0.514±0.240	0.486±0.248	0.536±0.239	0.533±0.221	0.495±0.253	0.555±0.246	0.471±0.257
Place of residence							
Rural	0.524±0.239	0.505±0.246	0.541±0.238	0.542±0.219	0.503±0.253	0.562±0.245	0.476±0.258
Urban	0.479±0.239	0.431±0.241	0.512±0.240	0.474±0.222	0.459±0.250	0.548±0.246	0.458±0.255
Respondent's education							
Illiterate	0.502±0.238	0.450±0.242	0.529±0.237	0.494±0.219	0.476±0.253	0.582±0.242	0.499±0.261
Middle school complete	0.452±0.236	0.409±0.236	0.480±0.238	0.427±0.216	0.449±0.250	0.511±0.245	0.442±0.252
High school complete and above	0.439±0.235	0.340±0.233	0.436±0.234	0.437±0.217	0.439±0.248	0.540±0.245	0.410±0.244
Standard of living							
Low	0.490±0.237	0.466±0.242	0.490±0.236	0.460±0.218	0.473±0.253	0.607±0.238	0.459±0.256
Medium	0.427±0.231	0.407±0.233	0.419±0.229	0.416±0.214	0.417±0.244	0.533±0.244	0.393±0.238
High	0.364±0.217	0.336±0.213	0.347±0.211	0.363±0.205	0.409±0.242	0.455±0.239	0.351±0.223
Sex of child							
Female	0.357±0.215	0.326±0.209	0.338±0.208	0.358±0.203	0.397±0.240	0.476±0.241	0.339±0.217
Male	0.369±0.218	0.344±0.216	0.354±0.213	0.369±0.206	0.420±0.245	0.437±0.236	0.362±0.227
Child wt-for-age							
Z-score >=-2	0.375±0.217	0.351±0.214	0.368±0.210	0.370±0.206	0.425±0.243	0.454±0.231	0.362±0.227
Z-score < -2	0.358±0.212	0.326±0.205	0.334±0.199	0.367±0.205	0.398±0.237	0.407±0.222	0.364±0.228
Maternal BMI & Child age							
<18.5Kg/m^2 and 13-35months	0.169±0.043	0.132±0.033	0.148±0.034	0.215±0.048	0.204±0.049	0.216±0.068	0.134±0.028
<18.5Kg/m^2 and <=12 months	0.546±0.078	0.472±0.076	0.525±0.072	0.566±0.071	0.580±0.075	0.627±0.100	0.569±0.060
>=18.5Kg/m^2 and <=12 months	0.570±0.077	0.543±0.077	0.530±0.072	0.601±0.070	0.628±0.073	0.578±0.103	0.560±0.060
>=18.5Kg/m^2 and 13-35 months	0.136±0.036	0.112±0.028	0.140±0.033	0.170±0.040	0.150±0.038	0.134±0.046	0.088±0.020
Maternal BMI & Child age with Breast-feeding status							
Breast-feeding with supplements							
+ <18.5Kg/m^2 and 13-35months	0.176±0.072	0.140±0.062	0.202±0.073	0.240±0.080	0.173±0.058	0.154±0.080	0.116±0.050
+ <18.5Kg/m^2 and <=12 months	0.543±0.129	0.472±0.133	0.600±0.119	0.589±0.115	0.525±0.102	0.509±0.150	0.510±0.122
+ >=18.5Kg/m^2 and <=12 months	0.566±0.128	0.540±0.134	0.605±0.118	0.623±0.113	0.573±0.100	0.461±0.149	0.500±0.122
+ >=18.5Kg/m^2 and 13-35 months	0.143±0.061	0.120±0.054	0.191±0.070	0.192±0.068	0.127±0.045	0.094±0.052	0.076±0.034
Exclusive breast-feeding							
+ <18.5Kg/m^2 and 13-35months	0.471±0.128	0.422±0.129	0.433±0.116	0.576±0.116	0.416±0.099	0.561±0.149	0.438±0.120
+ <18.5Kg/m^2 and <=12 months	0.831±0.079	0.796±0.095	0.818±0.080	0.861±0.064	0.789±0.070	0.880±0.067	0.859±0.062
+ >=18.5Kg/m^2 and <=12 months	0.844±0.074	0.837±0.081	0.821±0.079	0.877±0.059	0.820±0.063	0.857±0.077	0.854±0.064

Variable	Probability ± Standard Deviation						
	India	North	Central	East	Northeast	West	South
+ >=18.5Kg/m^2 and 13-35 months	0.412±0.123	0.380±0.124	0.417±0.114	0.508±0.116	0.333±0.090	0.428±0.147	0.332±0.108
Breast-feeding with plain water only							
+ <18.5Kg/m^2 and 13-35months	0.319±0.109	0.271±0.102	0.335±0.103	0.387±0.107	0.302±0.086	0.374±0.141	0.227±0.085
+ <18.5Kg/m^2 and <=12 months	0.719±0.109	0.665±0.124	0.748±0.099	0.740±0.097	0.694±0.088	0.770±0.110	0.696±0.105
+ >=18.5Kg/m^2 and <=12 months	0.738±0.105	0.723±0.114	0.752±0.098	0.766±0.092	0.734±0.082	0.733±0.120	0.688±0.107
+ >=18.5Kg/m^2 and 13-35 months	0.269±0.099	0.237±0.093	0.321±0.101	0.324±0.098	0.232±0.072	0.258±0.116	0.157±0.065

4. Discussion and conclusions

Analyses are very much in line with the previous finding that breast-feeding practices affect the likelihood of resumption of menstrual cycle after birth. Mothers who breast-feed their child exclusively were less likely to have resumed menstruation. Socio-economic status and education of women were found to be inversely associated with duration of lactational amenorrhea. Living in urban areas of the north and the east regions and higher level of maternal education except the east, northeast and west regions were all associated with a lower likelihood of remaining amenorrheic. It is clear that mother needs a balanced diet based on a variety of nutritious food items especially during pregnancy and lactation. Women from low socio-economic status will not be able to manage the nutritious food. Thus, they may breast-feed their child more frequently or spend more time per day doing it, and as a result have a higher likelihood of remaining amenorrheic. Parity was found to be positively associated with the likelihood of remaining amenorrheic. The reason may be that higher parities women preferred longer duration of breast-feeding as compared to those women who were at lower parities. Mothers who had underweight children were poorly ($p>0.05$) associated with lactational amenorrhea except in the central region. Moreover, mothers who had underweight children were more likely to remain amenorrheic in the central region. There is a possibility that poorly nourished children are those who are not receiving adequate weaning foods after six months of age and are more likely to breast-feeding more intensively (Kurz *et al.,* 1993; Dewey *et al.,* 1997). Child gender did not have significant impact on the return of ovulation period. It shows that there is no change in the breast-feeding behaviour of mothers by sex of the child.

The interaction term between maternal nutritional status and duration of breast-feeding (child's age) was significantly associated with the likelihood of having resumed menstruation after controlling for breast-feeding practices, child nutritional status and socio-economic and demographic covariates. The effect of maternal nutritional status on lactational amenorrhea was not found to be

significant when women were breast-feeding since last 12 months except in the northern region of India. However, after 12 months of breast-feeding, the probability of undernourished women remaining amenorrheic was likely to be greater and this trend was highly consistent across all the six regions included in the analysis. The possibility is that undernourished women have less body fat to support the return of menses after giving birth. As Kurz *et al.* (1993) came out with the findings that poorly nourished women may experience greater inhibition of the ovulatory hormones than better-nourished women, given the same amount of suckling, and so they were found to be more amenorrheic. At the same time, other researchers have also argued that undernourished women produce less milk per nourishing episode (Delgado *et al.,* 1982 and Lunn *et al.,* 1984), and their children need to suck longer or more intensely than children of better-nourished mothers to obtain the amount of milk that they require. This increase in sucking frequency or intensity might be associated with an increase in plasma prolactin level and thus increase the likelihood of being amenorrheic (Loudon *et al.,* 1983).

Thus, the result clearly shows that maternal nutritional status has not had an independent impact on lactational amenorrhea. Gournis *et al.* (1997) have tried to explain the specific biological mechanisms that explain such type of findings. They found that unrestricted access to food (well-nourished) to ovariectomized rats during lactation was associated with higher levels of luteinizing hormone and follicle stimulating hormone. Therefore, these rats had shorter postpartum anestrus period. However, in this seminal study, it was not possible to separate the influence of maternal body composition from behaviours leading to less sucking behaviour on the metabolic/physiologic changes determining the duration of the anestrus period.

In addition to human epidemiologic studies, there is accumulating evidence strongly suggesting that maternal nutritional status does exercise an independent role in the return of menstruation. Leptin, a protein hormone released from adipocytes, appears to play an important role in reproductive performance (Frübeck, 1997). Studies show that there is a crucial link in leptin, maternal nutritional status and postpartum amenorrhea (Kopp *et al.*, 1997). Kurz *et al.* (1993) reported a significant negative relation between maternal nutritional status and postpartum amenorrhea. Further, they have stated that after controlling the infant supplementation, the association became only marginally significant and this study has little biological importance.

However, Loudon *et al.* (1983) has suggested that changes in sucking behaviour are more likely than maternal nutrition per se to influence the duration of postpartum amenorrhea. This study also supports the argument; otherwise the results should be very much consistent across all the six regions of India. Moreover, NFHS-3 data shows that the mean number of day and night time feeds was found to be low in the north and the central regions as compared to rest of the regions of the country. Further, the biologically significant role of maternal nutritional status on postpartum amenorrhea has never been contested, but it is argued that when undernourished women had certain level of duration and

intensity of breast-feeding then they were found to be more likely to remain amenorrheic.

Moreover, one could not fully breakdown the effect of socio-economic status on all intermediate and proximate determinants explaining lactational amenorrhea. The fact that socio-economic status remained significantly associated with lactational amenorrhea in India and in countries` all six regions has been considered. It is important for future studies to include the factors such as breast-feeding duration, that is, minutes breast-feeding per day and intensity in the analysis of postpartum amenorrhea, which is not available in NFHS. An attempt should also be made to collect the duration of PPA from those women who were not interested to breast-feed their child but they could not do so because of child loss.

REFERENCES

CLELAND, J. G., ATHAR, Z. A., (1984). The Effect of Birth Spacing on Childhood Mortality in Pakistan, *Population Studies*, 38 (3): 401–418.

DELGADO, H. L., MARTORELL, R., KLEIN, R. E., (1982). Nutrition, Lactation, and Birth Interval Components in Rural Guatemala, *American Journal of Clinical Nutrition*, 35: 1468–1476.

DEWEY, K. G., COHEN, R. J., RIVERA, L. L., CANAHUATI, J., BROWN, K. H., (1997). Effect of Age at Introduction of Complementary Foods to Breastfed Infants on Duration of Lactational Amenorrhea in Honduran Women, *American Journal of Clinical Nutrition*, 65: 1403–1409.

DIAZ, S., RODRIQUEZ, G., MARSHALL, G., DEL, P. G., CASADO, M.E., MIRANDA, P., SCHIAPPACASSE V., CROXATTO, H. B., (1988). Breastfeeding Pattern and the Duration of Lactational Amenorrhea in Urban Chilean Women, *Contraception*, 38 (1): 37–51.

DWIVEDI, L. K., (2006). Contraceptive Use in India: A Multivariate Decomposition and Related Simulation Analysis, *Demography India*, 35(2): 291–302.

DWIVEDI, L. K., RAM, F., RESHMI, R. S., (2007). An Approach to Understanding Change in Contraceptive Behaviour in India, *GENUS*, LXIII (3-4): 19–54.

FRISCH, R. E., (1983). Population, Nutrition and Fecundity, In Malthus: Past and Present: Dupaquier, J (Eds.), A Fauve-Chamoux, and Grebinik, E, Academic Press, London: 393–404.

FRISCH, R. E., MCARTHUR, J. W., (1974). Menstrual Cycles: Fatness as a Determinant of Minimum Weight for Height Necessary for their Maintenance or Onset, *Science (Wash. DC),* 185(4155): 949–951.

FRISCH, R. E., REVELLE, R., COOK, S., (1973). Components of the 'Critical' Weight at Menarche and at Initiation of the Adolescent Spurt: Estimated Total Water, Lean Body Mass and Fat, *Human Biology*, 45: 469–483.

FRÜBECK, G., (1997). Leptin involvement in reproductive performance, *Journal of Nutrition*, 127: 1533.

GOURNIS, E., MCGUIRE, M. K., RASMUSSEN, K. M., (1997). Food supplementation during lactation shortens anestrus and elevates gonadotropins in rats, *Journal of Nutrition,* 127: 785–790.

HUFFMAN, S. L., FORD, K., ALLEN, H. A., STREBLE, P., (1987). Nutrition and Fertility in Bangladesh: Breastfeeding and Postpartum Amenorrhoea, *Population Studies,* 41(3): 447–462.

INTERNATIONAL INSTITUTE FOR POPULATION SCIENCES (IIPS), ORC MACRO., (2000). National Family Health Survey (NFHS-2), 1998–1999. India. Mumbai: IIPS.

KOPP, W., BLUM, W. F., VON PRITTWITZ, S., ZIEGLER, A., LUBBERT, H., EMONS, G., HERZOG, W., HERPERTZ, S., DETER, H.C., REMSCMIDT, H, HEBERBRAND, J., (1997). Low Leptin Levels Predict Amenorrhea in Underweight and Eating Disordered Females. *Mol. Psychiatry* 2: 335–340.

KURZ, K. M., HABICHT, J. P., RASMUSSEN, K. M., SCHWAGER, S. J., (1993). Effects of Maternal Nutritional Status and Maternal Energy Supplementation on Length of Postpartum Amenorrhea among Guatemalan Women, *American Journal of Clinical Nutrition*, 58: 636–642.

LOUDON, A. S. I., MCNEILLY, A. S., MILNE, J. A., (1983). Nutrition and Lactational Control of Fertility in Red Deer, *Nature (London),* 302: 145–147.

LUNN, P. G., AUSTIN, S., PRENTICE, A. M., WHITEHEAD, R. G., (1984). The Effect of Improved Nutrition on Plasma Prolactin Concentrations and Postpartum Infertility in Lactating Gambian Women, *American Journal of Clinical Nutrition*, 39: 227–235.

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 129–138

A PROBABILISTIC SCHEME WITH UNIFORM CORRELATION STRUCTURE

Raffaella Calabrese[1]

ABSTRACT

The probabilistic schemes with independence between the trials show different dispersion characteristics depending on the behaviour of the probabilities of the binary event in the trials. This work proposes a probabilistic scheme with uniform correlation structure that leads to different dispersion characteristics depending on the sign of the linear correlation. Finally, a hypothesis test is proposed to identify the type of the dispersion of the probabilistic scheme.

Key words: probabilistic scheme, uniform correlation, binary event.

1. Introduction

Binary events clustered into groups are analysed by the probabilistic schemes (Feller, 1968, p.146). Under the assumption of the independence between the trials, by changing the characteristics of the probabilities of the binary events the Bernoulli, Poisson, Lexis and Coolidge probabilistic schemes (Kendall, 1994, p.164) are defined. In this paper the above-mentioned schemes are analysed by highlighting how the different characteristic of the probabilities of the binary events lead to different dispersion properties. By removing the assumption of the independence of the trials, a probabilistic scheme with uniform correlation structure is proposed in this paper.

Analogously to the previous schemes, the dispersion of the proposed scheme can be normal, subnormal and supernormal, depending on whether the correlation is zero, negative or positive, respectively. Finally, a hypothesis test is proposed to verify the assumption of binomial dispersion.

The present paper is organized as follows. In the next section the probabilistic schemes with independence between the trials is analysed. In the following section a probabilistic scheme with unform correlation is proposed. Section

[1] Dynamic Labs Geary Institute University College Dublin.
E-mail: raffaella.calabrese@ucd.ie.

4 suggests a hypothesis test to identify the kind of dispersion of a probabilistic scheme. Finally, the last section contains some concluding remarks.

2. The probabilistic schemes with independence between the trials

Let us assume to be interested in attaining an event A (success) in k series of n_j trials each with $j = 1, 2, \ldots, k$. For the subsequent results the assumption of independence between both the k series and the n_j trials of each series will be essential. Thus, let A_{ji} be the Bernoulli random variable associated to the i-th trial of the j-th series, with $i = 1, 2, ..., n_j$ and $j = 1, 2, ..., k$

$$A_{ji} = \begin{cases} 1 & \text{the event } A \text{ occurs in the } i\text{-th trial of the } j\text{-th series} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

having the following success and failure probabilities

$$P\{A_{ji} = 1\} = p_{ji} \qquad P\{A_{ji} = 0\} = 1 - p_{ji} = q_{ji}.$$

In addition, let us define the random variables $X_j = \sum_{i=1}^{n_j} A_{ji}$ which indicates the number of times the event A occurs in the n_j trials of the j-th series and $X = \sum_{j=1}^{k} \sum_{i=1}^{n_j} A_{ji}$ which represents the number of times the event A occurs in the $n = \sum_{j=1}^{k} n_j$ trials. For the previous assumptions the n indicator random variables A_{ji} are thus mutually independent.

The relative frequency of the event A in the n_j trials of the j-th series can be represented through the random variable $\hat{p}_j = \frac{X_j}{n_j}$; while the relative frequency of the event A on the total of the n trials is $\hat{p} = \frac{X}{n} = \frac{1}{n} \sum_{j=1}^{k} \hat{p}_j n_j$; which coincides with the weighted arithmetic mean of the relative frequencies of the k series with weights equal to n_j .

The variables defined in this way show therefore the following expectations and variances:

$$\mathbb{E}(\hat{p}_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} p_{ji} \tag{2.2}$$

$$\mathbb{V}(\hat{p}_j) = \frac{1}{n_j^2} \sum_{i=1}^{n_j} p_{ji}(1 - p_{ji}) \tag{2.3}$$

$$\mathbb{E}(\hat{p}) = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} p_{ji} \tag{2.4}$$

$$\mathbb{V}(\hat{p}) = \frac{1}{n^2} \sum_{j=1}^{k} \sum_{i=1}^{n_j} p_{ji}(1 - p_{ji}) \tag{2.5}$$

Thus, probabilistic schemes with independence between both the trials and the series require carrying out k series of n_j trials each.

These schemes can be classified according to the conditions under which these trials are performed, which influence the probability of success p_{ji}.

2.1. The Bernoulli probabilistic scheme

In the Bernoulli probabilistic scheme the assumption is made that the probability of success is constant from trial to trial and from series to series

$$p_{ji} = p \qquad \text{with} \quad i = 1, 2, ..., n_j \quad \textit{and} \quad j = 1, 2, ..., k.$$

Under such conditions the indicator random variables A_{ji} are independent and identically distributed with common parameter p.

The expectation and the variance of the relative frequency $\hat{p}$, for the calculation of which it is advisable to determine the mathematical expectation and the variance of $\hat{p}_j$, in a Bernoulli scheme are

(2.6)

$$\mathbb{E}(\hat{p}_j) = p \qquad \mathbb{V}(\hat{p}_j) = \frac{p(1-p)}{n_j}$$

(2.7)

$$\mathbb{E}(\hat{p}) = p \qquad \mathbb{V}(\hat{p}) = \frac{p(1-p)}{n^2} \sum_{j=1}^{k} n_j = \frac{pq}{n}$$

To analyse the dispersion of the Bernoulli scheme, the following quantity is computed

$$\mathbb{E}\left[\sum_{j=1}^{k} (\hat{p}_j - p)^2 n_j\right] = kpq. \tag{2.8}$$

The Bernoulli scheme is defined as *normal dispersion* scheme (Feller, 1968, p.146).

In this probabilistic scheme the relative frequency $\hat{p}_j$ of the j-th series can be approximated with a normal having mean and variance given by the (2.6). This means that the random quantity

$$\sum_{j=1}^{k} \frac{(\hat{p}_j - p)^2}{pq} n_j, \tag{2.9}$$

approximates, as n_j diverges, to a chi-square with k degree of freedom. Similar considerations applied to the relative frequency $\hat{p}$, having expectancy and

variance given by the equations (2.7), enable one to state that the following random variable

$$\frac{(\widehat{p}-p)^2}{pq}n \tag{2.10}$$

can be approximated, as *n* diverges, to a chi-square with one degree of freedom. In the random quantities defined by the expressions (2.9) and (2.10) the probability of success *p*, whose value is usually unknown, is included. For this reason, it is advisable to modify the above said random quantities so that they become functions of known parameters.

The following relation is deduced from the decomposition of the deviance

$$\sum_{j=1}^{k}(\widehat{p}_j-\widehat{\overline{p}})^2 n_j = \sum_{j=1}^{k}(\widehat{p}_j-p)^2 n_j - n(\widehat{p}-p)^2.$$

Dividing both members of the previous equation by the factor *pq* we obtain

$$\sum_{j=1}^{k}\frac{(\widehat{p}_j-\widehat{\overline{p}})^2}{pq}n_j = \sum_{j=1}^{k}\frac{(\widehat{p}_j-p)^2}{pq}n_j - \frac{(\widehat{p}-p)^2}{pq}n.$$

Because of the associative property of the random variable chi-square, we can deduce that the following expression

$$\sum_{j=1}^{k}\frac{(\widehat{p}_j-\widehat{\overline{p}})^2}{pq}n_j \tag{2.11}$$

can be approximated, as the number of occurrences *nj* diverges, to a chi-square with (k − 1) degrees of freedom. From the convergence in probability of the relative frequency $\hat{p}$ to the unknown parameter *p* and by applying Slutsky's theorem (Cramer, 1996, pp. 254-255) we observe that the random quantity

$$\sum_{j=1}^{k}\frac{(\widehat{p}_j-\widehat{\overline{p}})^2}{pq}n_j\frac{pq}{\widehat{\overline{p}}\widehat{\overline{q}}} = \sum_{j=1}^{k}\frac{(\widehat{p}_j-\widehat{\overline{p}})^2}{\widehat{\overline{p}}\widehat{\overline{q}}}n_j, \tag{2.12}$$

converges in distribution, as the number of occurrences n_j diverges, to a chi-square random variable with (k − 1) degrees of freedom.

2.2. The Poisson probabilistic scheme

In 1830 Poisson formalized the scheme of repeated trials in conditions of independence with probabilities of success p_{ji} varying from trial to trial within the same series. The probabilistic scheme called after this author considers constant from series to series both the partial means

$$\overline{p}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}p_{ji} = \overline{p},$$

with $j = 1, 2, \ldots, k$, and the variances

$$\mathbb{V}_j(p_{ji}) = \frac{1}{n_j}\sum_{i=1}^{n_j}(p_{ji} - \overline{p}_j)^2 = \sigma_j^2(p) = \sigma^2(p)$$

between the probabilities of the trials of each series[1], with $j = 1, 2, \ldots, k$.

By considering the deviation λ_{ji} between the probability of success of the i-th trial of the j-th series and the overall mean

$$\lambda_{ji} = p_{ji} - \overline{p} \qquad i = 1, 2, ..., n_j \quad \text{and} \quad j = 1, 2, ..., k \tag{2.13}$$

we obtain

$$\begin{aligned}\mathbb{V}(X_j) &= \sum_{i=1}^{n_j} p_{ji} - \sum_{i=1}^{n_j} p_{ji}^2 \\ &= n_j\overline{p} - n_j\overline{p}^2 - n_j\sigma^2(p) - 2\overline{p}\sum_{i=1}^{n_j}\lambda_{ji} \\ &= n_j\overline{p}(1-\overline{p}) - n_j\sigma^2(p).\end{aligned}$$

Like in the case of the Bernoulli scheme, to analyse the dispersion we calculate the expectation of the weighted sum of the deviations squared between the relative frequencies $\hat{p}_j$ and the overall average probability $\overline{p}$, with weight equal to the number of occurrences n_j

$$\mathbb{E}\left[\sum_{j=1}^{k}(\hat{p}_j - \overline{p})^2 n_j\right] = \sum_{j=1}^{k}\frac{\mathbb{V}(X_j)}{n_j} = k\overline{p}(1-\overline{p}) - k\sigma^2(p). \tag{2.14}$$

Comparing this result with the outcome obtained (2.8) in the Bernoulli scheme with constant success probability equal to $\overline{p}$, we understand why the Poisson scheme is defined as *subnormal dispersion scheme* (Kendall, 1996, p.166). The dispersion of the Poisson scheme, therefore, depends on the variability $\sigma^2(p)$ among the probabilities of a series, in particular, the higher it is, the lower will the expectation of the 'deviation' of the relative frequencies $\hat{p}_j$ be.

2.3. The Lexis probabilistic scheme

In 1876 Lexis proposed the following probabilistic scheme that was named after him, in which the probabilities of success p_{ji} stay constant within the same series $p_{ji} = p_j$, with $i = 1, 2, ..., n_j$ and $j = 1, 2, \ldots, k$, but vary from series to series[2].

[1] The features of Poisson's probabilistic scheme coincide with those of a stratified sampling scheme in which a single sample unit is extracted from each population layer (Cochran, 1953, chapter 5).

[2] A. Lexis probabilistic scheme represents a particular sampling scheme in two stages (Cochran, 1953, p. 274).

Thus, let

$$\overline{p} = \sum_{j=1}^{k} p_j \frac{n_j}{n}$$

be the average probability of success, obtained as the arithmetic mean of the probabilities of success of the individual series with weights equal to the number of occurrences n_j , and let

$$\sigma^2(p_j) = \sum_{j=1}^{k} (p_j - \overline{p})^2 \frac{n_j}{n}$$

be the variance among the probabilities of the different series.

In order to compare the dispersion of the Bernoulli scheme with constant probability of success equal to $\bar{p}$ with the one of the Lexis scheme, the following deviation is considered. We obtain

$$\mathbb{E}\left[\sum_{j=1}^{k} (\hat{p}_j - \overline{p})^2 n_j\right] = k\overline{p}(1-\overline{p}) + (1-2\overline{p}) \sum_{j=1}^{k} \lambda_j + \sum_{j=1}^{k} (p_j - \overline{p})^2 (n_j - 1).$$

Defining $\widetilde{p} = \frac{1}{k} \sum_{j=1}^{k} p_j$ as the simple (not weighted) mean of the probabilities of success of the various series, we can rewrite the previous equation

$$\mathbb{E}\left[\sum_{j=1}^{k} \left(\frac{X_j}{n_j} - \overline{p}\right)^2 n_j\right] = k\overline{p}(1-\overline{p}) + k(1-2\overline{p})(\widetilde{p} - \overline{p}) + \sum_{j=1}^{k} (p_j - \overline{p})^2 (n_j - 1). \tag{2.15}$$

As the number of occurrences n_j diverges, the last summation tends to infinity. Now, it is possible to compare the result obtained with that (2.8) obtained previously in the Bernoulli scheme with constant probability of success equal to $\bar{p}$. Therefore, the Lexis scheme shows a *supernormal* dispersion (Kendall, 1996, p. 166).

2.4. The Coolidge probabilistic scheme

Finally, let us consider the probabilistic scheme proposed by Coolidge in 1921, which represents a generalization of the schemes of repeated trials examined before, since the probabilities of success p_{ji} are free to vary both from trial to trial and from series to series.

To determine the properties of the random variable X associated to the Coolidge scheme we associate to each series the random variable X_j of the Poisson probabilistic scheme and then go ahead with mixing the k variables

determined with weights n_j. This method enables to use some of the results obtained previously.

Following the same method used for the Poisson probabilistic scheme, the deviation (2.13) is used to obtain the following result

$$\mathbb{V}(X_j) = n_j\overline{p}_j - n_j\overline{p}^2 - \sum_{i=1}^{n_j}(p_{ji} - \overline{p}_j)^2 - n_j(\overline{p}_j - \overline{p})^2 - 2\overline{p}n_j(\overline{p}_j - \overline{p}).$$

At this point we calculate the expectation of the mixture of the k random variables $\hat{p}_j$ with weights n_j by obtaining

$$\mathbb{E}\left[\sum_{j=1}^{k}(\hat{p}_j - \overline{p})^2 n_j\right] =$$
$$= \sum_{j=1}^{k}\overline{p}_j - k\overline{p}^2 - \sum_{j=1}^{k}\sigma_j^2(p) + \sum_{j=1}^{k}(\overline{p}_j - \overline{p})^2(n_j - 1) - 2\overline{p}\sum_{j=1}^{k}(\overline{p}_j - \overline{p}). \tag{2.16}$$

Depending on various assumptions of different probabilistic schemes, the previous expression includes as special cases the results obtained for the Bernoulli, Poisson and Lexis schemes.

To be able to make some considerations on the above result we should consider a Coolidge scheme made up of k series, all having a constant number of occurrences equal to m; under such circumstances the quantity (2.16) becomes

$$\mathbb{E}\left[\sum_{j=1}^{k}\left(\frac{X_j}{m} - \overline{p}\right)^2 m\right] = k\overline{p}(1-\overline{p}) + (m-1)\sum_{j=1}^{k}(\overline{p}_j - \overline{p})^2 - \sum_{j=1}^{k}\sigma_j^2(p).$$

In the Coolidge scheme, the last two addenda of the previous equation are both not equal to zero, but since the two summations $\sum_{j=1}^{k}(\overline{p}_j - \overline{p})^2$ and $\sum_{j=1}^{k}\sigma_j^2(p)$ have the same magnitude, as m diverges, the positive component prevails on the negative one, thus obtaining a supernormal dispersion scheme, also in the case in which the assumptions of the Lexis probabilistic scheme are not met. Therefore, in order for a phenomenon with subnormal dispersion to manifest itself, both the assumptions of the Poisson scheme need to be met, i.e. the probabilities must vary within the same series, while the average probabilities $\bar{p}_j$ and the variances $\sigma_j^2(p)$ have to remain constant from series to series. To find supernormal dispersion instead, it is not necessary that the probabilities of success remain constant from trial to trial in each series, as long as they vary from series to series.

Since, in empirical terms, the average probabilities $\bar{p}_j$ and the variances $\sigma_j^2(p)$ are seldom constant from series to series, it is obvious why a minor number of phenomena displays hypo-binomial dispersion, a property of the Poisson scheme, if compared to those with hyperbinomial dispersion, which mostly follow the Coolidge probabilistic scheme and only to a small extent the Lexis scheme.

3. A probabilistic scheme with uniform correlation between the trials

Into a probabilistic scheme, in which the goal is always that of obtaining an event A (success) in k series of n_j trials each with $j = 1, 2, \ldots, k$, we introduce at this point the assumption of dependence between the n_j trials of each series, maintaining the assumption of independence between the k series, though.

Since the following analysis focuses on the relationships of dependence between the variables, we assume to simplify matters that the probability of success p is constant from trial to trial and from series to series. Let us consider the case in which the (linear) dependence between each pair of random variables A_{ji} and A_{jl}, with $i \neq l$, of the j-th series, manifests itself in a uniform way

$$r(A_{ji}, A_{jl}) = \rho \quad i \neq l; \quad i, l = 1, 2, \ldots, n_j \quad \text{and} \quad j = 1, 2, \ldots, k,$$

by obtaining

$$Cov(A_{jl}, A_{ji}) = \rho(1 - p)p.$$

From the assumption of independence between the series we deduce that

$$r(A_{ji}, A_{sl}) = 0 \quad j \neq s; \quad j, s = 1, 2, \ldots, k \text{ and } i = 1, 2, \ldots, n_j ; \; l = 1, 2, \ldots, n_s.$$

As for the case of independence between the trials, the following variance is computed

$$V(X_j) = n_j p(1 - p) + n_j(n_j - 1)\,\rho p(1 - p) \qquad j = 1, 2, \ldots, k.$$

It follows that

$$\mathbb{E}\left[\sum_{j=1}^{k} (\hat{p}_j - p)^2 n_j\right] = kpq + \rho pq(n - k). \tag{3.1}$$

Comparing this result to those determined previously in the various probabilistic schemes with independence between trials, we obtain the following relation between the linear correlation coefficient ρ and the dispersion of the probabilistic scheme considered:

- if $\rho > 0$ the dispersion is supernormal, same behaviour as for the Lexis scheme;
- if $\rho = 0$ the dispersion is normal, same behaviour as for the Bernoulli scheme;
- if $\rho < 0$ the dispersion is subnormal, same behaviour as for the Poisson scheme.

An estimator of the linear correlation coefficient ρ is

$$\widehat{\rho} = \frac{\sum_{j=1}^{k}\left[(\hat{p}_j - \widehat{p})^2 \frac{n_j}{n}\right] - \frac{k\widehat{p}\widehat{q}}{n}}{\widehat{p}\widehat{q}\left(1 - \frac{k}{n}\right)}.$$

If the number of trials n is very high in comparison to k, we can approximate the previous equation as follows

$$\widehat{\rho} \simeq \frac{\sum_{j=1}^{k}\left[\left(\hat{p}_j - \widehat{p}\right)^2 \frac{n_j}{n}\right]}{\widehat{pq}}.$$

It should be noted that the numerator of this proportion represents the variability of the relative frequencies $\hat{p}_j$, whereas the denominator consists of the variability of the indicator random variable A_{ji} in the Bernoulli probabilistic scheme with constant probability of success equal to p.

4. A hypothesis test for the dispersion of a probabilistic scheme

To find out whether a test meets the assumptions of the Bernoulli scheme we propose to consider the following ratio

$$\frac{\sum_{j=1}^{k}\left[\left(\hat{p}_j - \widehat{p}\right)^2 \frac{n_j}{n}\right]}{(k\widehat{pq})n^{-1}} = \frac{\sum_{j=1}^{k}\left[\left(\hat{p}_j - \widehat{p}\right)^2 n_j\right]}{k\widehat{pq}}. \tag{4.1}$$

As already previously mentioned, the numerator of the ratio (4.1) represents the variability of the relative frequencies $\hat{p}_j$, and, as we can see from the result (2.8), in Bernoulli's probabilistic scheme the expectation of the numerator and the denominator coincide. This means that if the ratio (4.1) is close to unity, then the test taken into consideration meets the assumptions of the Bernoulli probabilistic scheme[1]. Since from the equation (3.1) the expectation of the numerator of the ratio (4.1) appears to be smaller than the denominator, we deduce that if this proportion is sizeably smaller than one, we should, instead, be inclined to use a probabilistic scheme with subnormal dispersion, that is the Poisson scheme, or a scheme with uniform negative correlation between the indicator random variables of each series. If, finally, this proportion is sizeably bigger than one, for the relations (2.15) and (3.1), a scheme with supernormal dispersion is preferred, namely the Lexis scheme or a scheme with uniform positive correlation between the trials of each series. It has to be pointed out that in the latter case, in which the value of the Lexis divergence quotient obtained is considerably higher than one, we might also take into consideration the Coolidge scheme, since it approximately displays supernormal dispersion as the number of tests of each series diverges.

Defining a significance level equal to α, we observe that if the value obtained by the test statistics

$$\sum_{j=1}^{k} \frac{\left(\frac{X_j}{n_j} - \widehat{p}\right)^2}{\widehat{pq}} n_j$$

[1] It has to be pointed out that in this case we might also consider a probabilistic scheme with dependence and uncorrelation ($\rho = 0$) between the trials, which means that between the indicator random variables of each series there is a tie of dependence of the non-linear kind, but given the rarity of the case we prefer to disregard this possibility.

is included within the values assumed by the quantiles of the $\alpha/2$ and $(1 - \alpha/2)$ orders of a random variable chi-square with $(k - 1)$ degrees of freedom, then we accept the null hypothesis that the test considered fits the Bernoulli scheme. If instead the value assumed by the test statistics is higher than the quantile of the $(1 - \alpha/2)$ order of a random variable chi-square with $(k - 1)$ degrees of freedom, then we accept the alternative hypothesis and choose either a Lexis scheme or a positive (uniform) correlation scheme between the trials of each series. In the latter case, in which the value assumed by the aforementioned test statistics is lower than the quantile of the $\alpha/2$ order of a random variable chi-square with $(k - 1)$ degrees of freedom, we always accept the alternative hypothesis which, however, consists in the Poisson scheme or in a scheme with negative (uniform) correlation between the trials of each series.

5. Conclusion remarks

The probabilistic schemes (Bernoulli, Poisson, Lexis and Coolidge) with independence between the trials show different dispersion properties. By introducing a uniform correlation structure between the trials, a new probabilistic scheme is proposed. By changing the type of correlation, the suggested scheme shows the same dispersion characteristics of the probabilistic schemes analysed in the literature. To identify the type of the dispersion of the probabilistic scheme, a hypothesis test is proposed.

REFERENCES

COCHRAN, W. G., (1953). Sampling techniques, Wiley.

CRAMER, H., (1996). Mathematical Methods of Statistics. Princeton University Press, Princeton.

FELLER, W., (1968). An introduction to Probability Theory and Its Applications. Vol. I John Wiley & Sons, New York.

JOHNSON, N. L., KEMP A. W. and KOTZ S., (2005). Univariate Discrete Distributions. Wiley, New York.

JOHNSON, N. L., KEMP A. W. and KOTZ S., (1969). Discrete Distributions. Houghton Mifflin, Boston.

KENDALL, S., (1994). The Advanced Theory of Statistics. Vol. I. Hafner Publishing Company, New York.

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 139–160

BIAS REDUCTION OF FINITE POPULATION IMPUTATION BY KERNEL METHODS

Nicklas Pettersson[1]

ABSTRACT

Missing data is a nuisance in statistics. Real donor imputation can be used with item nonresponse. A pool of donor units with similar values on auxiliary variables is matched to each unit with missing values. The missing value is then replaced by a copy of the corresponding observed value from a randomly drawn donor. Such methods can to some extent protect against nonresponse bias. But bias also depends on the estimator and the nature of the data. We adopt techniques from kernel estimation to combat this bias. Motivated by Pólya urn sampling, we sequentially update the set of potential donors with units already imputed, and use multiple imputations via Bayesian bootstrap to account for imputation uncertainty. Simulations with a single auxiliary variable show that our imputation method performs almost as well as competing methods with linear data, but better when data is nonlinear, especially with large samples.

Key words: bayesian bootstrap, boundary and nonresponse bias, missing data, multiple imputation, Pólya urn models, real donor imputation.

1. Introduction

In sample surveys missing data often has to be dealt with. Imputation is a standard treatment for sporadically missing values in the sample data due to item nonresponse. Given observed auxiliary variable(s) X related to the incomplete study variable Y, an imputation model is usually estimated from units where both x and y values are observed, modelled by the missing at random (MAR) mechanism which assumes that the probability of missingness only depends on observed values. The missing y values are then replaced by imputed values, and multiple imputation can account for the fact that imputed values differs from the true ones, so that standard methods can be used (Rubin, 1987). Imputed values may be non-observable values derived from a model, or real-donor values derived from observed values (Laaksonen, 2000). Donors to each donee (or recipient) are

[1] Department of Statistics, Stockholm University. E-mail: nicklas.pettersson@stat.su.se.

usually found by selecting units close to the donee according to some distance measure on X.

Imputation methods employing parametric models may be effective (Schafer, 1997), but their benefits diminish with sample size and can lead to severe bias if the underlying assumptions are violated. Methods based on nonparametric models can then provide robustness to nonresponse bias at the cost of some efficiency. Bias of methods based on nonparametric models also depends on the derivation of the imputed values, and the nature of the bounded data. The bias of a mean estimate of y is related to the individual unit bias of x, the expectation over donor x's minus the actual x, through individual unit bias of y. When X is continuous, the asymptotic bias of x for an interior donee can easily be set to zero. This is more difficult for donees that lie on the boundary of the data. By viewing imputation as pointwise kernel smoothing, and adopting bias reduction techniques from that area, we propose a real donor method which aims at mitigating such bias of individual x as to implicitly reduce bias of the mean estimator of y.

Our method starts out from the popular hot deck imputation; see Little and Andridge (2010) for a review. For each donee unit where y is missing, a pool consisting of k potential donor units with observed y-values is identified. The missing y value of the donee is then filled in by a copy of the observed y value from a unit in the donor pool. Adjustment cells imputation bring together all zero distance donors and donees, having the same categorized x, creating an illusion that individual x's are unbiased. Cells may therefore only contain donees. This is avoided by non-categorizing distance measures, which produce donor pools that can be better matched to the donee, but the number of k nearest neighbour (kNN) donors has to be decided. Justified by Bayesian exchangeability through Pólya sampling (Feller, 1971), we extend the set of potential donors to include previously imputed donees, and handle imputation uncertainty through multiple imputation.

Individual bias in x is first addressed by relating distances between the donee and the donors to the donor selection probabilities, giving closer donors higher donation probability. Siddique and Belin (2008) set selection probabilities inversely proportional to the distance between predictive means of donor and donee units, while Conti, Marella and Scanu (2008) let a Gaussian kernel decide the selection probabilities. We propose to use an Epanechnikov (1969) kernel, which asymptotically can minimize mean squared error of an estimate. We expect reduction of variance in general and boundary donee bias of x.

Boundary bias can also be reduced by letting the selection probabilities be found from local linearization (Simonoff, 1996). Aerts, Claeskens, Hens and Molenberghs (2002) use non-negative constrained weights asymptotically equivalent to kernel weights as selection probabilities. We calibrate our selection probabilities by a Lagrange function, similar to calibration of design weights (Deville and Särndal, 1992), but on a pointwise level.

Our third bias reduction method is inspired by Rice (1984), who tightened the kernel at the boundary. By reducing k for boundary donees, on average closer but

fewer donors are obtained compared to interior donees, which contribute to the bias reduction of x.

The paper is structured as follows: Section 2 presents real donor imputation with Pólya urn sampling and multiple imputation. Our proposed methods are described in Section 3, and further studied by simulations in Section 4. The paper is then concluded in Section 5.

2. Background on real donor and multiple imputation

A simple random sample (SRS) of i=1, …, n units from a population of N units is drawn with the aim to estimate the mean $\bar{y} = \sum y_i / N$ of the study variable Y, and the value y_i is observed in the sample. The indicator R_i=1 for the r units where y_i is observed, while R_i=0 for nonresponding units. In real donor imputation, each donee i should have a donor pool of k_i units. Denote by q_i the number of units that possibly could enter pool i. Given our SRS design, we simply set k_i=q_i=r for all i, and use all respondents as potential donors. Later we allow k_i, q_i, and r to differ, and may omit index i when it is dispensable.

For each donee i, a donor j is selected with probability λ_{ij}, and the imputed value $\hat{y}_i$ is a copy of y_j. When all n-r missing values have been imputed, an estimate of $\bar{y}$ is

$$\hat{\bar{y}} = \frac{1}{n}\left(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \hat{y}_i \right). \tag{1}$$

Since the expectation of an imputed value is

$$E(\hat{y}_i) = \sum_{j=1}^{q} \lambda_{ij} y_j \,, \tag{2}$$

the individual bias of $\hat{y}_i$ is

$$B(\hat{y}_i) = E(\hat{y}_i) - y_i \,. \tag{3}$$

Due to the SRS design it follows that $E(y_i) = \bar{y}$. The bias of (1) is therefore

$$B\left(\hat{\bar{y}}\right) = E\left(\hat{\bar{y}}\right) - \bar{y} = \frac{1}{n}\left\{ \sum_{i=1}^{r} E(y_i) + \sum_{i=r+1}^{n} E(\hat{y}_i) \right\} - \bar{y} = \frac{1}{n} \sum_{i=r+1}^{n} B(\hat{y}_i). \tag{4}$$

Now, assume a known auxiliary variable X and a MAR mechanism, so that the response probability does not depend on y; $P(R=1|Y,X)=P(R=1|X)$. We further assume that the expected value of Y does not depend on R, $E(Y|X)=g(X)$, which is another consequence of MAR. Denote the x-value of the donor selected for donee

i by $\hat{x}_i$. Its expectation is $E(\hat{x}_i) = \sum_{j=1}^{q} \lambda_{ij} x_j$. We may expect to reduce (3), and thereby (4), by reducing the bias of x_i

$$B(\hat{x}_i) = E(\hat{x}_i) - x_i = \sum_{j=1}^{q} \lambda_{ij} x_j - x_i \,. \tag{5}$$

2.1. Adjustment cells and k-nearest neighbour imputation

As a background we first describe two common methods for imputation, adjustment cells and nearest neighbour imputation. Our suggested method in Subsection 3.3 is based on the latter. All methods are illustrated on the simple dataset in Table 1, where x is observed on all n=7 units, while y is only observed on r=5 units. Table 1 is ordered after x. The cut off between the two adjustment cells is set to x=0. Let $\lambda_{ij} = 1/k_i$ for donee i=3, 6 and donor j. Since donor pools are determined from x, we usually have that $k_i<q_i$.

Example 1. Imputation within adjustment cells. Only units within the same adjustment cell may be used as donors. Thus, although $q_3=r=5$, the $k_3=4$ potential donors for Unit 3 are Units 1, 2, 4 and 5, and $B(\hat{x}_3) \approx -0.013$. We randomly draw one of them, say Unit 4, and impute the missing y-value as $\hat{y}_3 = 0.022$. Unit 7 is the only ($k_6=1$) potential donor to Unit 6, so $B(\hat{x}_6) = 0.231$ and we impute $\hat{y}_6 = -0.099$. If single donor situations are not allowed, a common solution is to collapse adjustment cells. Units 1, 2, 4, 5 and 7 are then the ($k_3=q_3=5$) potential donors to Unit 3, and $B(\hat{x}_3) \approx -0.107$. Assume again we draw Unit 4. Unit 6 has the same donors, so $B(\hat{x}_6) \approx -0.247$. If the imputed Unit 3 also had been allowed to act as a donor (so that $k_6=q_6=r+1=6$) we would have had $B(\hat{x}_6) = -0.265$.

Table 1. Data in Examples 1-5, with x and y generated by model *NO* in Subsection 4.1.

Unit no.	1	2	3	4	5	6	7
x-cat.	1	1	1	1	1	2	2
x	-0.413	-0.381	-0.255	-0.152	-0.125	0.099	0.330
y	-0.555	-0.476	Missing (-0.136)	0.022	0.349	Missing (0.335)	-0.099

Note: (the true but unknown value in parenthesis is given here for illustrative purposes.)

Example 2. Imputation by kNN. We now discard the categorization of x, and use 4NN imputation (i.e. $k_3=k_6=4$). Since Units 1, 2, 4 and 5 are the closest (among the $q_3=5$) units to donee Unit 3, $B(\hat{x}_3) \approx -0.013$ as in Example 1. Assume unit 4 was drawn. Unit 6 then has Units 2, 4, 5 and 7 as donors with $B(\hat{x}_6) = -0.181$. By allowing the imputed Unit 3 as a donor (so that $q_6=r+1=6$) we get $B(\hat{x}_6) \approx -0.150$ based on Units 3, 4, 5, and 7.

Adjustment cells imputation effectively matches donors to a donee and is widely used. But having a single donor can severely affect variances, as explained in Subsection 2.3. Collapsing cells is a simple solution, but *k*NN can provide better matching. Since Unit 3 has half of its donors on each side (as k_3=4) we call it an interior unit, while Unit 6 with only a single donor on the right is called a boundary unit. We will make use of this distinction in Subsection 3.3, where we suggest how to further improve *k*NN matching and try to reduce bias. Allowing imputed donees to act as donors for subsequent donees differs from usual donor imputation, but a Bayesian justification based on Exchangeability and Pólya urns is given in Subsection 2.2.

2.2. Imputation by Pólya urn sampling and Bayesian bootstrap

Descriptions of imputation methods which use previously imputed values in subsequent imputations can be found in Rubin (1987) and Kong, Liu and Wong (1994). These methods attempt to impute the missing values by draws from their posterior predictive distributions, and rely on a Bayesian motivation going back to de Finetti's (1931) theorem on exchangeable sequences. If the probability distribution for any finite sequence of *n* random variables drawn from an infinite series of random variables is the same, then any such infinite series is exchangeable. A sequence of independent and identically distributed (iid) random variables is always exchangeable, but the opposite is not true. But under some assumptions any exchangeable sequence is distributed as a sequence that is iid, given some parameters which in turn have a prior distribution. Hewitt and Savage (1955) generalized de Finetti's theorem to non-binary variables, and Diaconis and Freedman (1980) showed that it is approximately true for long but finite sequences of variables, implying finite exchangeability.

Pólya urn sampling produces an exchangeable but non-iid series, see Feller (1971). Assume a sample of *n* units where we have observed either the value 0 or 1 on variable *Y*. Then, 1) draw a single unit at random from the sample, 2) duplicate the drawn unit, and 3) replace both the drawn and the duplicated unit into the sample. The procedure is then repeated, but now with the updated sample of size *n*+1. By repeating the procedure ad infinitum, the generated sequence of values on the units is then an infinite exchangeable sequence. Blackwell and MacQueen (1973) generalized Pólya urn sampling to allow for more than two categories, and Ferguson (1973) extended to continuous variables.

Finite population Bayesian bootstrap (FPBB) (Lo, 1988) is based on Pólya urn sampling from a sample (of size *n*) to a large finite population (of size *N*). If a sample is drawn by SRS and the observed units are randomly drawn from the sample itself by SRS, then the observed units may be treated as a part of an exchangeable series of variables. In our example (Table 1) we may treat the sample as the population, and the five observed units as our sample. Pólya sampling may then be applied to reconstruct the remaining *n*-*r* units from the *r* observed ones, corresponding to imputation within the collapsed adjustment cells

using Unit 3 as potential donor to Unit 6 in Example 1 (where $k_6=q_6= r+1=6$). Knowing the full population size, Pólya sampling can be done to the whole population, starting from the r observed units, and sequentially impute all N-r units. An estimate of $\bar{y}$ is then simply the mean of the bootstrap population.

As $N \to \infty$, FPBB approaches the model based Bayesian bootstrap by Rubin (1981). They raise two objections to bootstrap methods in connection to the exchangeability assumption. First, they ask whether it is reasonable to assume that all possible distinct values of a variable have been observed in a sample. The objection is definitely valid with the continuous and very small sample in Table 1. Assuming unlimited precision all realized values of a continuous variable are unique, so we will not observe all values until we have observed the whole population. But our ability to grasp the data distribution should improve with the sample size, unless data is censored or if missingness in other ways is concentrated to certain regions of the data. This (strong) dependence on sample size is a characteristic common to nonparametric methods, simply because they refrain from parametric assumptions.

Assuming all possible distinct values are observed, Rubin's second objection is that the probabilities of occurrences for similar values might be dependent. This calls for smoothing of probabilities, but bootstrapping assumes strict independence. If the distribution of realized or bootstrap samples differs much from the true population, some estimators might perform poorly. As for the first objection, the larger the sample, the more likely we are to observe the distribution of the true data, so benefits from smoothing should, in general, diminish.

2.3. Bayesian bootstrap and multiple imputation

Imputation by FPBB basically corresponds to multiple imputation (Rubin, 1987). A general overview of variance estimation with single imputation is given in Little and Rubin (2002), and an overview for hot deck imputation in Andridge and Little (2010).

Assume a sample from a finite population of exchangeable units with n-r missing values on variable Y imputed d=1, …, D times. The distribution of the estimates

$$\hat{\bar{y}}_{d,n} = \frac{1}{n}\left(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \hat{y}_{di} \right), \; d=1,\ldots,D, \tag{6}$$

then reflects the imputation uncertainty due to that imputed values for the same unit differs between the imputed datasets. A point estimate of $\bar{y}$ is given by

$$\hat{\bar{\bar{y}}}_n = \frac{1}{D}\sum_{d=1}^{D} \hat{\bar{y}}_{d,n}, \tag{7}$$

and the variance of $\hat{\bar{\bar{y}}}_n$ is estimated as

$$\hat{V}\left(\hat{\bar{\bar{y}}}_n\right) = \frac{D+1}{D} B_n + \overline{W_n}. \tag{8}$$

Component $B_n = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\bar{y}}_{d,n} - \hat{\bar{\bar{y}}}_n\right)^2$ accounts for imputation uncertainty, and sampling uncertainty is covered by the variance component $\overline{W_n} = \frac{1}{D}\sum_{d=1}^{D} W_{d,n}$, where

$$W_{d,n} = \left(\frac{N-n}{N-1}\right)\left(\frac{1}{n-1}\right)\left[\left(\sum_{i=1}^{r} y_{di} - \hat{\bar{y}}_{d,n}\right)^2 + \left(\sum_{i=r+1}^{n} \hat{y}_{di} - \hat{\bar{y}}_{d,n}\right)^2\right], \; d=1,\ldots,D, \tag{9}$$

is the estimated variance within a bootstrap set. The term $\frac{N-n}{N-1}$ is the finite population correction. If both the *n-r* non-responding and the *N-n* non-sampled units in each bootstrap set had been imputed, then a population estimate similar to (7) would have been

$$\hat{\bar{\bar{y}}}_N = \frac{1}{D}\sum_{d=1}^{D} \hat{\bar{y}}_{d,N} = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{N}\left(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{N} \hat{y}_{di}\right). \tag{10}$$

Sampling uncertainty vanishes with a completely imputed population, so (8) simplifies to

$$\hat{V}\left(\hat{\bar{\bar{y}}}_N\right) = \frac{D+1}{D} B_N = \left(\frac{D+1}{D}\right)\left(\frac{1}{D-1}\right)\sum_{d=1}^{D}\left(\hat{\bar{y}}_{d,N} - \hat{\bar{\bar{y}}}_N\right)^2. \tag{11}$$

With missing values deterministically imputed, as in the uncollapsed cell in Example 1 with a single donor (k_6=1), all imputed bootstrap sets will have the same value imputed, so B_N (or B_n) will be underestimated. In particular, if all values are deterministically imputed, then $\hat{\bar{y}}_{1,N} = \ldots = \hat{\bar{y}}_{D,N} = \hat{\bar{\bar{y}}}_N$, implying that $B_N = 0$, so that $\hat{V}\left(\hat{\bar{\bar{y}}}_N\right) = 0$ in (11).

3. Kernel estimation and kernel imputation

One may look at donor imputation from the view of kernel estimation. We give a brief introduction to the area, describe the connections to imputation, and suggest how to improve estimation and achieve bias reduction of (7) or (10) using auxiliary variable *X*.

3.1. Short background on kernel estimation

Kernel estimation is a method to estimate density. Assume that q values are observed on x and the density $f(x)$ at a point x_i is to be estimated. Denote the distance x_i-x_j by $\tilde{x}_{ij}$. Given a kernel function K, the pointwise kernel estimate of f at x_i is then

$$\hat{f}(x_i) = \frac{1}{qh}\sum_{j=1}^{q} K\left(\frac{\tilde{x}_{ij}}{h}\right) = \frac{1}{q}\sum_{j=1}^{q} K_h(\tilde{x}_{ij}),$$

where K is typically symmetric, unimodal and integrates to 1. We confine our considerations to situations where K is proportional to the indicator function $I(|\tilde{x}_{ij}| < h)$, which is 1 if the statement is true. Function K_h is K scaled by the bandwidth (or smoothing) parameter h, which determines that K is positive if $|\tilde{x}_{ij}| < h$, and zero if $|\tilde{x}_{ij}| \geq h$. The choice of h is usually more important than K. If h is fixed for all i, the number of units $k_i \geq 0$ within the range $x_i \pm h$ is random. Instead, if the number of units k_i is fixed at k, the bandwidth h_i will be random. Methods to select a fixed h or k range from subjective judgement of plots and simple automatic rules of thumb, to more sophisticated methods based on cross-validation and plug-in estimates (Wand and Jones, 1995). Fixing h is more frequent, and a fixed k is best used when the exact size is noncritical, typically with $k \approx q^{1/2}$ (Silverman, 1986).

A commonly used measure of accuracy is the mean integrated squared error (MISE)

$$MISE(\hat{f}) = \int E(\hat{f}(x) - f(x))^2 dx = \int (B\{\hat{f}(x)\})^2 dx + \int V\{\hat{f}(x)\} dx, \qquad (12)$$

where a pointwise approximation of the bias component is given by

$$B\{\hat{f}(x_i)\} = E\{\hat{f}(x_i)\} - f(x_i) = \frac{1}{q}\sum_{j=1}^{q} E\{K_h(\tilde{x}_{ij})\} - f(x_i), \qquad (13)$$

and an approximation of the variance with independent x_j is given by

$$V\{\hat{f}(x_i)\} = \frac{1}{k_i} V\{K_h(\tilde{x}_{ij})\}. \qquad (14)$$

Given that K is symmetric and h (or k) is reduced, bias in (13) will decrease while variance in (14) will increase. The variance goes to zero as $qh \to \infty$ (or $k \to \infty$), while bias depends on the curvature of f and is asymptotically unrelated to q, unless $h \to 0$ (or $k/q \to 0$) as $q \to \infty$. Bias then converge to zero if x_i lies in the interior (unbounded) part of x, while if x_i lies within a bandwidth h from the boundary of x, the bias will not vanish. Given an optimal choice of h, MISE in

(12) is approximately minimized if K is set to the unimodal Epanechnikov (1969) function

$$K_h^{Ep}\left(\tilde{x}_{ij}\right)=\frac{3}{4}\left\{1-\left(\tilde{x}_{ij}\right)^2\right\}I\left(\left|\tilde{x}_{ij}\right|<h\right). \tag{15}$$

3.2. Kernel imputation

Assume K_h is a positive function scaled so that $\sum_{j=1}^{q}K_h\left(\tilde{x}_{ij}\right)=1$, where $\tilde{x}_{ij}=x_i-x_j$ and the sum is over the donor pool described in Subsection 2.1. When the selection probabilities are given by $\lambda_{ij}=K_h\left(\tilde{x}_{ij}\right)$ we call the technique kernel imputation. The expectation of $\hat{y}_i$ in (2) thus becomes the Nadaraya-Watson (1964) estimator

$$E\left(\hat{y}_i\right)=\sum_{j=1}^{q}\lambda_{ij}y_j=\sum_{j=1}^{q}K_h\left(\tilde{x}_{ij}\right)y_j\ .$$

With a uniform kernel $K_h^{Un}\left(\tilde{x}_{ij}\right)\propto I\left(\left|\tilde{x}_{ij}\right|<h\right)$, the donee i has k potential donor units within the range $x_i\pm h$ with selection probabilities $\lambda_{ij}=K_h^{Un}\left(\tilde{x}_{ij}\right)=1/k$, and q-k units outside the range with $\lambda_{ij}=0$. When donor data at x_i is sparse, fixing k instead of h will cover more distant donors, which avoids situations with no or few donors. With donors densely located in a vicinity of x_i, using an adaptable parameter h_i (caused by the fixed k) will in general result in donor pools that are better matched to donee i.

3.3. Kernel imputation with bias reduction

We suggest the use of multiple kernel imputation but also add three special devices, mainly to decrease imputation bias, but also to decrease the random errors. The bias $B\left(\hat{\bar{y}}_i\right)$ in (4) is related to $B\left(\hat{x}_i\right)$ in (5) and $B\left\{\hat{f}\left(x_i\right)\right\}$ in (13) through $B\left(\hat{y}_i\right)$ in (3) and $\lambda_{ij}\propto K_h\left(\tilde{x}_{ij}\right)$. Given a model $E(Y|X)=g(X)$ and a response mechanism $P(R=1|X)$, we will probably reduce $B\left(\hat{\bar{y}}_i\right)$ by reducing $B\left\{\hat{f}\left(x_i\right)\right\}$ or $B\left(\hat{x}_i\right)$. Examples 3 to 5 are in line with this, and each presents one of our three proposed devices.

Example 3. Imputation with Epanechnikov selection probabilities. It is easy to believe that giving donors close to the donee higher probabilities is better than using a uniform kernel function. This is the idea behind this example. Due to the optimality properties shown by the non-negative Epanechnikov function in Kernel estimation, we suggest to use it here. In Example 2, donee 3 had k=4 donors, with

$B(\hat{x}_3) \approx -0.013$ and $E(\hat{y}_3) = -0.176$. With Epanechnikov probabilities $\lambda_{3j}^{Ep} = K_h^{Ep}(\tilde{x}_{3j})$ from (15), the closer (furthest) donor is more (less) likely to donate. With h_3=0.3715, Units 1, 2, 4 and 5 are assigned probabilities 0.238, 0.252, 0.260 and 0.250, so $B(\hat{x}_3) \approx -0.010$ and $E(\hat{y}_3) = -0.170$. Suppose that we draw Unit 4. If h_6=0.417 units 3, 4, 5 and 7 will get the probabilities λ_{6j}^{Ep} at 0.125, 0.274, 0.304 and 0.297, so that $B(\hat{x}_6) = -0.113$ and $E(\hat{y}_6) = 0.068$, compared to $B(\hat{x}_6) \approx -0.150$ and $E(\hat{y}_6) \approx 0.052$ in Example 2.

Given a symmetric kernel function the expected bias of interior donees is zero, so we only expect a reduction of variance by the change from K^{Un} to K^{Ep}. But given the same bandwidth *h* (or *k*), we do expect some reduction of bias for boundary donees since we switch from K^{Un} to the parabolic shaped K^{Ep}.

Example 4. Imputation with adjusted selection probabilities. A technique which fully eliminates $B(\hat{x}_i)$ is to adjust the probabilities given by the kernel so that the expectation over the *x*-values equals the donee x_i. More technically we propose to replace λ_{ij} by λ_{ij}' as close as possible but such that $E(\hat{x}_i) = x_i$ holds. λ_{ij}' is easily found by Lagrange minimisation as the solution to

$$\min_{\Lambda_1,\Lambda_2,\lambda_{ij},\, j=1,\ldots,k} \sum_{j=1}^{k} L(\lambda_{ij} - \lambda_{ij}') + \Lambda_1 \left\{ \sum_{j=1}^{k} \lambda_{ij}' (\tilde{x}_{ij}) \right\} + \Lambda_2 \left(\sum_{j=1}^{k} \lambda_{ij} - 1 \right), \tag{16}$$

where $L(\lambda_{ij} - \lambda_{ij}')$ is a distance function and Λ_1 and Λ_2 are Lagrange multipliers.

For the data in Table 1 and using Euclidean distances we get $\lambda_{31}^{Ep}' \approx 0.217$, $\lambda_{32}^{Ep}' \approx 0.235$, $\lambda_{34}^{Ep}' \approx 0.277$ and $\lambda_{35}^{Ep}' \approx 0.272$, with $E(\hat{y}_3) \approx -0.143$. Assuming Unit 4 is drawn, we get $\lambda_{6j}^{Ep}' \approx 0.011$, $\lambda_{6j}^{Ep}' \approx 0.217$, $\lambda_{6j}^{Ep}' \approx 0.263$ and $\lambda_{6j}^{Ep}' \approx 0.508$, with $E(\hat{y}_6) \approx 0.036$. Both $B(\hat{x}_i)$ are zero.

By solving (16) it is possible to obtain λ_{ij}' that results in $B(\hat{x}_i) = 0$ for both interior and boundary donees, as long as there are possible donors at both sides of x_i. (Other restrictions, for example, deterministic situations, may also prohibit unbiased solutions). The proposed adjustment of selection probabilities resembles the use of approximate kernel regression weights in imputation (Aerts, Claeskens, Hens, and Molenberghs, 2002), or calibration of design weights (Deville and Särndal, 1992) but on a pointwise level.

Example 5. Imputation with fewer donors at the boundary. Problems occur at the boundaries since there may be none or only few possible donor *x*-values at one side of x_i. We suggest that the width of the kernel then should be decreased. With

multidimensional x one could also use an oblong donor pool instead of a spherical (quadratic) one.

Consider only boundary Unit 6. Setting k=2 shrinks the bandwidth from h_6=0.417 to h_6=0.241, which results in selection probabilities $\lambda_{65}^{Ep} \approx 0.624$ and $\lambda_{67}^{Ep} \approx 0.376$ for donors 5 and 7, with $B(\hat{x}_6) \approx -0.053$ and $E(\hat{y}_6) \approx 0.181$, compared to $B(\hat{x}_6) \approx -0.150$ and $E(\hat{y}_6) \approx 0.052$ from Example 3. Applying the Lagrange adjustment in (16) results in $\lambda_{65}^{Ep\prime} \approx 0.508$ and $\lambda_{67}^{Ep\prime} \approx 0.492$, with $B(\hat{x}_6) = 0$ and $E(\hat{y}_6) \approx 0.128$.

The expected bias of boundary units is directly related to the bandwidth and the reduction of $|B(\hat{x}_6)|$ from shrinking k is in line with this. But this bias reduction is expected to come at the cost of higher $V(\hat{x}_6)$ since we use fewer possible donors.

4. Simulation study

Here we use our suggested bias reduction methods from Subsection 3.3 in a design-based simulation study with simulated data, and compare with other imputation methods.

4.1. Setup of simulation study

We construct two related populations. First *N=1 600* values are simulated from a *Un(0,1)* distribution (*u*) and a standard normal distribution (*e*) using R (R Development Core Team, 2009). The populations are then constructed, one with a linear (*LI*) relationship (x^{LI}=*u-1/2;* y^{LI}=*u+e/7-1/2*) and one with a nonlinear (*NO*) relationship (x^{LI}=*u-1/2;* y^{NO}=*sin(uπ)+e/7-2/π*). From each population we draw *1 000* samples of size n=*100*, *400* and *900*. In each sample we create 50% nonresponse on $y_{\_}$, using the MAR mechanism $P(y_{\_} \text{ is observed}) \propto 1\text{-}u^{1/4}$.

Table 2. Bias correction in kernel imputation

ID for kernel imputation methods	*U*	*E*	*L*	*S*	*EL*	*ES*	*LS*	*ELS*
Epanechnikov selection probabilities	No	Yes	No	No	Yes	Yes	No	Yes
Lagrange adjustment of biased units	No	No	Yes	No	Yes	No	Yes	Yes
Shrinkage to $k=k^{5/6}$ at boundary	No	No	No	Yes	No	Yes	Yes	Yes

The missing data in the sample or the population were imputed by all combinations of the three bias correction methods: Epanechnikov (*E*) selection probabilities, Lagrange (*L*) adjustment, and shrinkage (*S*) of the donor pool for boundary biased units. The methods' initial letters are used for notation as displayed in Table 2. The k potential donors were found using Euclidian distance

and a square root rule $k = q^{½}$, where q is the number of eligible (observed and imputed) donor units.

Mean estimates of $\bar{y}^{LI}$ and $\bar{y}^{NO}$ from our methods are compared to estimates based on complete data (*CD*) and complete cases (*CC*). Estimates $\hat{\bar{\bar{y}}}_n^-$ based on imputed samples are also compared to estimates from ten single imputation methods, SI_i i=1,…,10, and thirteen multiple imputation methods, MI_i i=1,…,13. Estimates $\hat{\bar{\bar{y}}}_N^-$ based on fully imputed populations are only compared to the MI_i methods. All MI_i and SI_i methods are derived from the R-packages described in Appendix 1. Appendix 2 and 3 contain results for estimates of $\bar{y}^{LI}$ and $\bar{y}^{NO}$ with the comparison methods.

The SI_i point and variance estimates $\hat{\bar{y}}_n^-$ and $\hat{V}\left(\hat{\bar{y}}_n^-\right)$ are calculated as in (6) and (9), while all multiple imputation estimates $\hat{\bar{\bar{y}}}_n^-$ and $\hat{\bar{\bar{y}}}_N^-$ are calculated as in (7) and (10), with variance estimates $\hat{V}\left(\hat{\bar{\bar{y}}}_-^-\right)$ given by (8) and (11). We used either D=5 or D=20 replicates for all multiple imputation methods. To simplify the description, we henceforth replace $\hat{\bar{\bar{y}}}_-^-$ by $\hat{\bar{y}}_-^-$. Empirical averages from simulations, with M representing n or N, are calculated as $G_M^- = \frac{1}{1000}\sum_{g=1}^{1000} G_{g,M}^-$, where $G_{g,M}^-$ is a function based on the g:th data, such as a point estimate $\hat{\bar{y}}_{g,M}^-$, the empirical mean squared error $MSE\left(\hat{\bar{y}}_{g,M}^-\right) = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\bar{y}}_{gi,M}^- - \bar{y}^-\right)^2$, bias $B\left(\hat{\bar{y}}_{g,\mathrm{M}}^-\right) = \hat{\bar{y}}_{g,\mathrm{M}}^- - \bar{y}^-$ or variance $V\left(\hat{\bar{y}}_{g,M}^-\right) = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\bar{y}}_{gi,M}^- - \hat{\bar{y}}_{g,M}^-\right)^2$, the average estimated variance $\hat{V}\left(\hat{\bar{y}}_{g,\mathrm{M}}^-\right)$, or the average double sided confidence interval length $CIL = 2t_{(1-\alpha,df)}\left\{\hat{V}\left(\hat{\bar{y}}_{g,\mathrm{M}}^-\right)\right\}^{1/2}$ and coverage $CIC = I\left\{\left|\hat{\bar{y}}_{g,M}^-\right| \le CIL/2\right\}$. The significance level of the t-statistic is always set to α=0.05, and the degrees of freedom $v = (D+1)\left(1 + \frac{1}{D+1}\frac{\overline{W}_M}{B_M}\right)^2$ where $\overline{W}$ and B are the variances components of (8) as described in Subsection 2.3 (Rubin, 1987). We always multiply G_M^- by 100 (100^2) if $G_{g,M}^-$ is a first (second) moment function.

4.2. Results from simulation study

Results for $\hat{\bar{y}}_-^{LI}$ ($\hat{\bar{\bar{y}}}_-^{NO}$) are presented in Table 3 (4), and for comparison methods in Appendix 2 (3). We only show results for sample sizes 100 and 900,

and 20 imputed datasets for multiple (including kernel) imputation. Results using n=400 ended up in between n=100 and n=900 with kernel imputation. This was mostly the case for multiple imputation comparison methods as well, except for bias (and sometimes for MSE dominated by bias) which tended to be highest with n=400. Comparing D=20 and D=5, most simulation results were up to 15% lower for kernel imputation with D=20 compared to D=5. Confidence coverage was only slightly smaller, but interval lengths were down to 30% shorter. Bias was rather unaffected by D, with $B\left(\hat{\bar{y}}_n^{LI}\right)$ as an exception which almost halved but from a low level. Results for multiple imputation comparison methods had the same tendencies, but were more mixed.

Table 3. Simulation results for estimates of $\bar{y}^{LI}$, including 95% confidence intervals

		Sample size n=100, nonresponse r=50						*Sample size n=900, nonresponse r=450*					
M	**ID**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**
n; sample imputed	*U*	14.4	0.69	13.9	11.9	93.2	14.1	0.85	0.17	0.82	0.79	95.5	3.7
	E	13.8	0.47	13.6	11.8	92.9	14.0	0.84	0.13	0.82	0.80	95.8	3.7
	L	14.0	0.60	13.6	12.0	92.9	14.1	0.84	0.15	0.81	0.85	96.3	3.8
	S	14.2	0.58	13.8	11.9	93.5	14.1	0.84	0.14	0.81	0.80	95.2	3.7
	EL	13.7	0.40	13.5	11.8	93.0	14.0	0.85	0.12	0.83	0.84	95.7	3.8
	ES	13.6	0.41	13.5	11.7	92.9	13.9	0.84	0.12	0.83	0.80	95.8	3.7
	LS	13.9	0.52	13.7	12.0	93.6	14.1	0.85	0.13	0.83	0.85	95.6	3.8
	ELS	13.6	0.36	13.5	11.8	93.6	14.0	0.84	0.11	0.83	0.84	95.6	3.8
N; population imptued	*U*	6.2	0.78	5.6	4.4	91.0	8.6	0.44	0.15	0.42	0.40	93.6	2.6
	E	6.0	0.58	5.7	4.0	88.9	8.2	0.45	0.13	0.43	0.40	93.7	2.6
	L	6.1	0.66	5.7	4.4	90.0	8.6	0.48	0.17	0.45	0.45	93.7	2.8
	S	6.1	0.70	5.6	4.3	90.0	8.5	0.44	0.13	0.42	0.39	94.9	2.6
	EL	6.1	0.52	5.8	4.0	89.5	8.2	0.47	0.13	0.45	0.43	93.7	2.7
	ES	5.9	0.51	5.7	3.7	88.6	7.9	0.44	0.12	0.43	0.41	94.4	2.6
	LS	6.1	0.59	5.8	4.2	88.6	8.4	0.48	0.15	0.46	0.45	94.0	2.8
	ELS	6.0	0.45	5.8	3.9	89.4	8.0	0.47	0.12	0.45	0.44	94.1	2.7

With the sample imputed in Table 3, bias decreased with increased sample size and added bias corrections (*E, S* or *L*). Variance dominated mean squared error, and seemed to decrease slightly with bias corrections and n=100. Average estimated variance was below the true value for n=100 and 400, but the underestimation was ameliorated by the added bias correction and it almost disappeared for n=900. Confidence interval coverage (CIC) was slightly below the stated 95% for n=100 and 400, but slightly above for n=900. Confidence interval lengths (CIL) decreased with sample size. Patterns were similar for the

whole population imputed but all figures were lower. An exception is $B\left(\hat{\bar{y}}_{\mathrm{n}}^{LI}\right)$, which was smaller than $B\left(\hat{\bar{y}}_{\mathrm{N}}^{LI}\right)$, but became more alike with increased sample size.

Single imputation methods (in Appendix 2) had similar or slightly better MSE compared to ELS, except SI3- SI6 which also had large bias. They always underestimated variance, and interval coverage decreased with sample size. Many multiple imputation methods behaved as well or somewhat better than ELS. Exceptions were MI5 and MI13 (and mostly MI12) with underestimated variance and poor coverage. MI13 also had huge bias. With the whole sample imputed MI6 also underestimated variance severely, and MI9 and MI10 had extremely large bias for n=100.

Table 4. Simulation results for estimates of $\bar{y}^{NO}$, including 95% confidence intervals

		Sample size n=100, nonresponse r=50						*Sample size n=900, nonresponse r=450*					
M	**ID**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**
n; sample imputed	*U*	20.9	2.29	15.6	13.6	89.2	15.1	1.30	0.68	0.84	0.86	91.3	3.8
	E	17.6	1.65	14.9	13.0	92.1	14.8	1.12	0.53	0.84	0.87	93.4	3.8
	L	18.1	1.83	14.7	14.0	91.6	15.3	1.14	0.55	0.84	0.94	94.4	4.0
	S	19.0	1.90	15.3	13.4	90.8	15.0	1.18	0.58	0.84	0.86	92.8	3.8
	EL	16.3	1.37	14.4	13.3	92.3	14.9	1.05	0.45	0.85	0.92	94.8	4.0
	ES	16.8	1.40	14.8	12.9	91.5	14.7	1.05	0.46	0.84	0.87	93.6	3.8
	LS	17.2	1.60	14.6	13.6	92.1	15.1	1.07	0.49	0.84	0.93	94.3	4.0
	ELS	15.9	1.23	14.4	13.1	92.9	14.8	1.00	0.40	0.83	0.91	95.2	3.9
N; population imptued	*U*	14.0	2.45	8.0	6.5	83.1	10.4	0.88	0.65	0.46	0.43	82.6	2.7
	E	10.4	1.81	7.1	5.4	84.8	9.5	0.73	0.51	0.46	0.43	86.6	2.7
	L	11.2	1.96	7.4	6.6	87.3	10.5	0.77	0.54	0.48	0.49	87.0	2.9
	S	11.5	1.97	7.6	6.0	86.2	10.0	0.76	0.55	0.46	0.42	86.7	2.7
	EL	9.4	1.54	7.0	5.4	86.6	9.5	0.66	0.43	0.47	0.46	90.3	2.8
	ES	9.0	1.48	6.8	4.9	86.3	9.0	0.66	0.46	0.45	0.42	88.8	2.7
	LS	10.1	1.66	7.4	6.0	87.4	10.0	0.70	0.47	0.48	0.48	89.5	2.9
	ELS	8.5	1.30	6.8	5.0	87.2	9.1	0.62	0.38	0.47	0.46	90.9	2.8

In Table 4, both $MSE\left(\hat{\bar{y}}_{n}^{NO}\right)$ and $B\left(\hat{\bar{y}}_{n}^{NO}\right)$ decreased in all cases with added bias correction and increasing sample size when the sample was imputed. Variance fell with sample size and somewhat with bias corrections for n=100. The underestimation of variance lessened with sample size, and $\hat{V}\left(\hat{\bar{y}}_{\mathrm{n}}^{NO}\right)$ was even somewhat higher then $V\left(\hat{\bar{y}}_{n}^{NO}\right)$ with n=900. Confidence interval coverage increased with sample size and added bias corrections, but was always below the stated 95% except for *ELS* with n=900. Confidence interval lengths decreased

with sample size. The patterns were similar when the whole population was imputed, but all figures were lower except for bias, which was somewhat higher with n=100, about the same with n=400, and slightly lower with n=900.

With only the sample imputed, nearest neighbour methods SI_7- SI_{10} and predictive mean matching methods MI_5-MI_6 in Appendix 3 had MSE similar to *ELS*, but with lower bias and higher variance. Their underestimation of variance also increased with sample size, with worsening confidence interval coverage. With the whole population imputed, MI_5-MI_6 gave small or zero estimates of variance. Method MI_{12} gave better coverage rate than *ELS*, both with the sample and population imputed, but overestimated the high variance severely and gave very wide confidence intervals. All other methods had much larger MSE than *ELS*, due to larger bias or variance. Several methods that rely on regression models had MSE similar to complete cases, with bias dominating the MSE.

5. Conclusions

Our proposed imputation method for missing value of a study variable assumes a relationship to a fully observed continuous auxiliary variable. Common to other methods based on nonparametric models, our method relies on having observed the data dispersion, which is more probable with larger samples. The non-informative Bayesian approach with Pólya urn sampling only using the sample as a prior and with multiple imputation can effectively address uncertainty with minimal assumptions. Given a missing at random mechanism, the real donor approach with imputed values selected among already observed (and thus presumably realistic) values, can also effectively remove nonresponse bias even with nonlinearities in the data. The use of kernel methods addresses the bias caused by having sparse and bounded finite sample data.

As expected, the simulation study with linear data demonstrated a small loss of efficiency compared to methods utilizing parametric assumptions, but with the nonlinear data the improvement by bias corrections was relatively larger, and comparison methods were generally outperformed. In both cases, our three suggested devices (Epanechnikov kernel, Lagrange adjustment, and shrinkage at the boundary) always reduced bias. Properties seemed to improve with increasing the sample size, which agrees with the nonparametric reliance on the sample size. Many of the multiple imputation comparison methods managed to give at least 95% coverage with linear data, which kernel imputation only did for the largest sample imputed. However, except for one extremely inefficient comparison method, kernel imputation with all bias corrections and the largest sample was the only method which reached 95% coverage with the nonlinear data. Since the response probabilities were strongly related to the study variable through the auxiliary, imputation methods with linear parametric assumptions displayed bias (and hence MSE) sometimes even larger than for complete cases when imputing the nonlinear data.

Variance (and hence MSE) went down when the whole population was imputed instead of just the sample. The effect is similar to what would have been expected from applying (post-) stratification weights based on the auxiliary. Since the bias share of MSE increased when the sample was imputed the confidence interval coverage rates fell. A similar but weaker effect was seen when the number of imputed datasets was increased.

Several extensions of the proposed method could be explored, including multivariate auxiliary and study variables, use of more or other prior information, estimators other than means, alternative distance metrics, more elaborate ways of choosing the number of donors, including the degree of shrinkage, or other aspects related to boundary donees.

REFERENCES

AERTS, M. CLAESKENS, G. HENS, N. and MOLENBERGHS G., (2002). Local multiple imputation. *Biometrika*, 89 (2), pp. 375–388.

ANDRIDGE, R. R. and LITTLE, R. J. A., (2010). A review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, 78 (1), pp. 40–64.

CONTI, P. L. MARELLA, D. and SCANU, M., (2008). Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. *Computational Statistics and Data Analysis*, 53 (2), pp. 354–365.

DE FINETTI, B., (1931). Funzione caratteristica di un fenomeno aleatorio, *Atti della R. Academia Nazionale dei Lincei, Classe di Scienze Fisiche, Mathematice e Naturale*, 6(4), pp. 251–299.

DEVILLE, J-C. and SÄRNDAL, C-E., (1992), Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87 (418), pp. 376–382.

DIACONIS, P. and FREEDMAN, D., (1980). Finite exchangeable sequences. *Annals of Probability*, 8(4), pp. 745–764.

EPANECHNIKOV, V. A., (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications*, 14 (1), pp. 153–158.

FELLER, W., (1971). An Introduction to Probability Theory and Its Applications, 2nd ed. Wiley, New York.

FERGUSON, T. S., (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2), pp. 209–230.

GELMAN, A. HILL, J. SU, Y-S. YAJIMA, M. and PITTAU, M. G., (2010). mi: Missing Data Imputation and Model Checking. R package version 0.09–11.

GRAMACY, R. B., (2010). monomvn: Estimation for multivariate normal and Student-t data with monotone missingness. R package version 1.8–3.

GROSS, K. and BATES, D., (2008). mvnmle: ML estimation for multivariate normal data with missing values. R package version 0.1–8.

HARRELL, F. E., (2010). Hmisc: Harrell Miscellaneous. R package version 3.8–3.

HEWITT, E. and SAVAGE, L. J., (1955). Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80 (2), pp. 470–501.

HOFF, P., (2010). sbgcop: Semiparametric Bayesian Gaussian copula estimation and imputation. R package version 0.975.

HONAKER, J. KING, G. and BLACKWELL, M., (2011) Amelia: Amelia II: A Program for Missing Data. R package version 1.5–4.

KIM, K-Y. and YI, G-S., (2008). SeqKnn: Sequential KNN imputation method. R package version 1.0.1.

KONG, A. LIU, J. S. and WONG, W. H., (1994) Sequential Imputations and Bayesian Missing Data Problems. *Journal of the American statistical association*, 89(425), pp. 278–288.

LAAKSONEN, S., (2000). Regression-based nearest neighbour hot decking, *Computational Statistics*, 15(1), pp. 65–71.

LITTLE R. J. A. and RUBIN, D. B., (2002). *Statistical analysis with missing data*. Hoboken: Wiley.

LO, A. Y., (1988). A Bayesian bootstrap for a finite population. *The Annals of Statistics*, 16 (4), pp. 1684–1695.

NADARAYA, E. A., (1964). On estimating regression. *Theory of Probability and its Applications*, 9 (1), pp. 141–142.

R DEVELOPMENT CORE TEAM. (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RICE, J., (1984). Boundary modification for kernel regression. *Communications in statistics - Theory and methods*, 13 (7), pp. 893–900.

RUBIN, D. B., (1981). The Bayesian Bootstrap. *The Annals of Statistics*, 9 (1), pp. 130–134.

RUBIN, D. B., (1987). *Multiple imputation for nonresponse in surveys*. Hoboken; Wiley.

SCHAFER, J. L., (1997). *Analysis of incomplete multivariate data*. London; Chapman and Hall.

SIDDIQUE, J. and BELIN, T. R., (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine*, 27 (1), pp. 83–102.

SILVERMAN, B. W., (1986). *Density estimation for statistics and data analysis*. London; Chapman and Hall.

SIMONOFF, J. S., (1996). *Smoothing methods in statistics*. New York; Springer-Verlag.

STACKLIES, W. REDESTIG, H. and WRIGHT, K., (2011). pcaMethods: A collection of PCA methods. R package version 1.24.0.

TEMPL, M. HRON, K. and FILZMOSER, P., (2010). robCompositions: Robust Estimation for Compositional Data. R package version 1.4.3.

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K., (2010). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, (in press).

WAND, M. P. and JONES, M. C., (1995). *Kernel smoothing*. London; Chapman and Hall.

WATSON, G. S., (1964). Smooth regression analysis. *Sankhya Series A*, 26 (4), pp. 359–372.

APPENDICES

Appendix 1. R-packages and code for alternative estimators

R-Package	ID	R-code
monomvn. Gramacy (2010)	SI_1	monomvn(data)
mvnmle. Gross (2008)	SI_2	mlest(data)
*pcaMethods.** Stacklies, Redestig and Wright (2011)	SI_3	llsImpute(data,k=1,center=T,correlation="pearson",verbose=F,allVariables=T)
	SI_4	pca(data,method="nipals")
	SI_5	pca(data,method="ppca")
	SI_6	pca(data,method="svdImpute")
robCompositions. Templ, Hron and Filzmoser (2010)	SI_7	impKNNa(data,k=1,metric="Euclidean",agg="median",primitive=T)
	SI_8	impKNNa(data,k=5,metric="Euclidean",agg="median",primitive=T)
SeqKnn. Kim and Yi (2008)	SI_9	SeqKNN(data,k=1)
	SI_{10}	SeqKNN(data,k=5)
Amelia. Honaker, King and Blackwell (2011)	MI_1	amelia(data,m = D)
Hmisc. Harrell (2010)	MI_2	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='regression',match='closest',nk=0,curtail=T,boot.method="approximate bayesian")
	MI_3	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='regression',match='closest',nk=0,curtail=F,boot.method="approximate bayesian")
	MI_4	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='regression',match='weighted',nk=0,curtail=T,boot.method="approximate bayesian")
	MI_5	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='pmm',match='closest',nk=0,curtail=T,boot.method="approximate bayesian")
	MI_6	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='pmm',match='weighted',nk=0,curtail=T,boot.method="approximate bayesian")
	MI_7	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='regression',match='closest',nk=c(0,3:5),B=10,curtail=T,boot.method= "approximate bayesian")
	MI_8	aregImpute(as.formula(~I(x)+I(y)),n.impute=D,type='regression',match='closest',nk=c(0,3:5),B=10,tlinear=F,curtail=T,boot.method="approximate bayesian")
mi. Gelman (2010)	MI_9	mi(data.frame(data),n.imp=D,add.noise=noise.control(method="reshuffling",K=1,post.run.iter=20),n.iter=30)
	MI_{10}	mi(data.frame(data),n.imp=D,add.noise=noise.control(method="fading",pct.aug=10,post.run.iter=20),n.iter=30)
mice. van Buuren and Groothuis-Oudshoorn (2010)	MI_{11}	mice(data,m=D,method="norm")
	MI_{12}	mice(data,m=D,method="pmm")
sbgcop. Hoff (2010)	MI_{13}	sbgcop.mcmc(data,nsamp=D)

R-packages for single (SI) and multiple (MI) imputation methods are available at *http://cran.r-project.org/web/packages/* and (*) *http://www.biocondoctor.org/biocLite.R.*

The object 'data' is created as 'data <- cbind(x,y)' in R, where 'x' is the fully observed auxiliary variable vector, and 'y' is the partly observed study variable vector. Object 'D' is the number of imputed datasets.

Appendix 2. Simulation results, alternative $\bar{y}^{LI}$ - estimators

		Sample size n=100, nonresponse r=50						*Sample size n=900, nonresponse r=450*					
M	**ID**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**	**MSE**	**B**	**V**	**V^**	**CIC**	**CIL**
n	*CD*	9.7	-.03	9.7	10.0	94.2	12.4	0.49	-.01	0.49	0.51	95.1	2.8
	CC	32.7	3.5	20.4	10.7	70.2	12.8	13.4	3.43	1.67	0.87	11.6	3.7
n; sample imputed	SI_1	12.1	-.15	12.1	10.1	92.3	12.4	0.73	-.02	0.73	0.51	90.8	2.8
	SI_2	12.1	-.15	12.1	9.9	92.3	12.3	0.73	-.02	0.73	0.51	90.8	2.8
	SI_3	20.9	2.77	13.2	8.7	79.6	11.6	11.3	3.24	0.85	0.45	1.8	2.6
	SI_4	39.5	4.81	16.4	5.0	42.6	8.7	31.6	5.51	1.18	0.25	0.0	1.9
	SI_5	23.7	3.09	14.2	8.1	73.2	11.1	37.0	5.98	1.30	0.23	0.0	1.9
	SI_6	51.2	5.72	18.4	4.6	32.4	8.3	44.3	6.55	1.40	0.23	0.0	1.9
	SI_7	14.1	-.06	14.1	9.9	89.1	12.3	1.21	-.02	1.21	0.51	78.4	2.8
	SI_8	13.3	0.08	13.3	9.2	88.7	11.8	0.92	-.06	0.92	0.48	83.8	2.7
	SI_9	14.3	-.03	14.3	9.9	88.6	12.3	1.24	-.03	1.24	0.51	78.5	2.8
	SI_{10}	13.3	0.01	13.3	9.4	89.4	12.0	0.95	-.01	0.95	0.49	83.5	2.7
	MI_1	12.5	-.54	12.2	12.4	95.4	14.4	0.75	0.11	0.73	0.90	97.9	3.9
	MI_2	12.3	0.04	12.3	12.2	95.7	14.3	0.74	0.03	0.73	0.82	96.8	3.7
	MI_3	12.2	-.14	12.1	12.8	95.5	14.7	0.75	-.03	0.75	0.85	97.4	3.8
	MI_4	12.2	0.05	12.2	12.2	95.2	14.3	0.73	0.03	0.73	0.82	97.3	3.7
	MI_5	14.1	-.06	14.1	9.9	89.3	12.3	1.20	-.02	1.20	0.51	80.3	2.9
	MI_6	12.9	0.11	12.9	10.6	93.0	13.3	0.80	0.25	0.73	0.63	93.1	3.2
	MI_7	12.3	0.04	12.3	12.1	94.9	14.3	0.74	0.03	0.74	0.82	96.7	3.7
	MI_8	12.2	0.06	12.2	12.2	95.3	14.3	0.74	0.03	0.74	0.82	97.0	3.7
	MI_9	12.1	-.03	12.1	13.2	96.1	14.9	0.74	-.10	0.73	0.87	97.4	3.8
	MI_{10}	12.4	-.25	12.3	12.6	95.0	14.5	0.75	-.03	0.75	0.86	97.3	3.8
	MI_{11}	12.2	-.26	12.2	12.9	96.0	14.7	0.73	0.07	0.73	0.92	98.1	4.0
	MI_{12}	12.7	0.19	12.7	12.2	94.9	14.3	1.12	0.09	1.11	0.76	90.4	3.6
	MI_{13}	28.5	3.71	14.7	12.1	81.7	14.3	14.5	3.67	0.99	0.70	2.4	3.4
N; population imptued	MI_1	5.3	0.12	5.3	4.9	94.1	9.0	0.39	-.05	0.39	0.36	93.8	2.5
	MI_2	5.4	0.44	5.2	8.7	98.6	12.1	0.41	0.03	0.41	0.68	98.9	3.4
	MI_3	5.2	0.11	5.2	10.7	98.8	13.4	0.40	-.01	0.39	0.71	98.9	3.5
	MI_4	5.7	0.46	5.5	9.5	98.2	12.7	0.41	0.03	0.41	0.69	98.2	3.4
	MI_5	7.5	0.29	7.4	0	0	0	0.85	-.06	0.84	0.00	6.5	0.2
	MI_6	6.9	0.62	6.5	0.1	21.8	1.3	0.45	0.24	0.39	0.10	65.8	1.3
	MI_7	5.6	0.46	5.4	9.0	98.4	12.3	0.39	0.04	0.39	0.69	99.5	3.4
	MI_8	5.3	0.45	5.1	8.6	97.9	12.0	0.40	0.03	0.40	0.69	99.0	3.4
	MI_9	52.8	6.01	16.7	14.4	68.5	15.9	0.39	0.12	0.38	0.49	96.8	2.9
	MI_{10}	65.9	7.03	16.5	8.3	38.1	12.0	0.39	0.08	0.38	0.46	97.1	2.8
	MI_{11}	5.5	0.82	4.9	4.2	93.0	8.5	0.37	0.00	0.37	0.60	99.3	3.2
	MI_{12}	6.3	0.56	6.0	4.3	90.0	8.6	0.76	0.27	0.68	0.22	71.0	1.9
	MI_{13}	59.0	6.29	19.5	0.5	9.5	3.1	26.3	5.02	1.10	0.25	0.0	2.1

Estimators are based on complete data (*CD*), complete cases (*CC*), multiply imputed (*MI*) and singly imputed (*SI*) datasets. Confidence interval coverage (CIC) and length (CIL) are from double-sided intervals with 5% significance level.

Appendix 3. Simulation results, alternative $\bar{y}^{NO}$ - estimators

		Sample size n=100. nonresponse r=50						*Sample size n=900. nonresponse r=450*					
M	**ID**	***MSE***	***B***	***V***	***V^***	***CIC***	***CIL***	***MSE***	***B***	***V***	***V^***	***CIC***	***CIL***
n	*CD*	10.6	0.05	10.6	11.2	95.4	13.1	0.6	0.02	0.55	0.57	95.8	3.0
	CC	63.9	6.6	19.9	9.2	42.8	11.8	44.0	6.51	1.55	0.75	0.0	3.4
n; sample imputed	SI_1	40.3	4.4	21.1	10.7	65.8	12.8	40.2	6.20	1.75	0.54	0.0	2.9
	SI_2	40.3	4.4	21.1	10.5	65.3	12.6	40.2	6.20	1.75	0.54	0.0	2.9
	SI_3	24.7	-.5	24.5	9.4	76.1	12.0	2.1	0.35	2.00	0.48	65.0	2.7
	SI_4	29.1	3.1	19.3	5.5	57.4	9.1	22.8	4.61	1.49	0.28	0.4	2.1
	SI_5	26.0	2.4	20.0	5.2	60.6	8.9	25.0	4.83	1.58	0.29	0.2	2.1
	SI_6	25.9	2.4	20.1	5.2	60.7	8.9	15.0	3.66	1.56	0.26	2.1	2.0
	SI_7	15.2	0.4	15.1	10.7	89.3	12.8	1.4	0.16	1.39	0.57	79.2	2.9
	SI_8	14.9	0.8	14.2	9.7	87.5	12.2	1.1	0.19	1.09	0.53	83.1	2.9
	SI_9	15.5	0.4	15.4	10.7	88.2	12.8	1.5	0.15	1.43	0.57	78.3	2.9
	SI_{10}	14.8	0.6	14.4	10.1	88.8	12.4	1.2	0.20	1.13	0.54	83.4	2.9
	MI_1	36.3	3.8	22.1	23.0	89.4	19.7	44.3	6.52	1.77	2.38	1.4	6.4
	MI_2	44.1	4.8	21.1	24.7	85.7	20.5	40.8	6.24	1.88	2.28	2.0	6.2
	MI_3	46.9	5.0	21.7	27.6	87.0	21.5	41.1	6.26	1.84	2.31	1.6	6.3
	MI_4	43.8	4.7	21.5	24.7	86.5	20.4	40.7	6.23	1.82	2.27	1.9	6.2
	MI_5	15.2	0.4	15.1	10.7	89.3	12.8	1.4	0.16	1.39	0.57	79.6	3.0
	MI_6	13.9	0.7	13.4	11.5	93.1	13.8	1.7	0.95	0.84	0.68	79.9	3.4
	MI_7	43.7	4.8	21.0	25.3	86.7	20.7	40.4	6.21	1.84	2.27	1.6	6.2
	MI_8	43.5	4.8	20.9	25.0	86.3	20.6	40.5	6.22	1.83	2.29	2.0	6.3
	MI_9	42.7	4.6	21.7	22.3	85.4	19.5	39.9	6.17	1.84	1.90	1.1	5.7
	MI_{10}	38.8	4.0	22.8	21.8	86.6	19.2	40.4	6.22	1.76	2.03	1.5	5.9
	MI_{11}	38.3	4.1	21.1	23.4	88.1	20.0	42.8	6.40	1.75	2.31	1.3	6.3
	MI_{12}	59.5	-2.2	54.6	74.6	93.7	35.1	2.1	0.85	1.37	18.6	100	17.6
	MI_{13}	36.6	4.2	19.0	14.5	79.4	15.7	28.2	5.17	1.49	1.00	1.0	4.1
N; population imputed	MI_1	97.9	8.5	26.3	24.2	61.0	20.1	36.9	5.91	1.95	1.93	1.7	5.7
	MI_2	92.6	8.3	23.1	42.0	80.1	26.6	37.5	5.95	2.03	3.57	9.2	7.8
	MI_3	105	8.9	25.9	53.9	84.4	30.1	38.1	6.00	2.05	3.65	8.8	7.9
	MI_4	98.6	8.7	23.1	45.5	81.0	27.7	37.6	5.97	2.00	3.58	8.4	7.8
	MI_5	9.1	1.0	8.0	0.0	0	0	0.8	0.06	0.84	0.00	7.8	0.2
	MI_6	9.8	1.6	7.2	0.1	19.6	1.4	1.3	0.93	0.41	0.12	36.5	1.4
	MI_7	94.2	8.4	23.4	44.5	79.8	27.4	37.5	5.96	2.01	3.55	8.1	7.8
	MI_8	95.3	8.5	22.9	42.2	79.8	26.6	37.8	5.97	2.07	3.61	9.7	7.9
	MI_9	45.3	5.1	19.2	18.0	79.6	17.7	35.2	5.77	1.89	2.20	3.5	6.1
	MI_{10}	58.7	6.3	18.9	10.7	55.3	13.7	34.2	5.68	2.01	1.83	2.9	5.6
	MI_{11}	110	9.3	23.5	17.3	46.3	17.4	37.3	5.95	1.88	2.66	3.6	6.8
	MI_{12}	75.4	-1.3	73.6	146	96.9	47.9	2.3	-.90	1.49	24.7	100	20.3
	MI_{13}	38.6	4.3	19.9	0.7	20.9	3.4	18.4	4.10	1.54	0.53	2.3	3.0

Estimators are based on complete data (*CD*), complete cases (*CC*), multiply imputed (*MI*) and singly imputed (*SI*) datasets. Confidence interval coverage (CIC) and length (CIL) are from double-sided intervals with 5% significance level.

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 161–170

A LINEAR MODEL FOR UNIFORMITY TRIAL EXPERIMENTS

Alok K. Shukla[1], **Subhash K. Yadav**[2], **Govind Ch. Misra**[3]

ABSTRACT

Uniformity trial experiments are required to assess fertility variation in agricultural land. Several models have appeared in literature, of which Fairfield Smith's Variance Law assuming a nonlinear relationship between the coefficient of variation (C.V.) and a plot size has been extensively used in uniformity trial studies. A linear model has been proposed for uniformity trial experiments and it has shown better results as compared to existing models. The expression for point of maximum curvature for the proposed model is much simpler as compared to the model of Fairfield Smith. The appropriateness of the proposed model has also been verified with the help of a data set.

Key words: Fairfield Smith's Variance Law, linear model, uniformity trial experiments.

1. Introduction

Uniformity trials are needed to determine suitable shape and size of the plot for knowing the nature and extent of fertility variation in land, so that if some treatment has given good result, one should be confirmed that it is true and is not due to some other unknown reason. In these trials, a particular variety of crop is sown on the entire experimental field and throughout the growing season it is managed uniformly. All sources of variation except that are due to natural soil differences, and are held constant to the maximum extent. At the time of harvest a substantial border is removed from all sides of the field. The rest of the field is divided into number of small plots which are termed as basic units, with the same dimensions. The production from these basic units is harvested and recorded separately for each basic unit. Then the yields in these basic units are collected separately. The usefulness of a uniformity trial lies in the fact that neighboring units may be amalgamated to form larger plots of various sizes and shapes. The

[1] Department of Statistics, D. A-V. College, Kanpur, India.
[2] Department of Mathematics & Statistics, Dr R M L Avadh University, Faizabad-224001, India.
[3] Department of Statistics, D. A-V. College, Kanpur, India.

variation in yield over the field due to soil heterogeneity and other manual errors are generally summed up in the term "Experimental Error" and may be calculated for each type of plot thus formed. Hence, all efforts in designing field experiments are directed to measure and control this source of variation.

The coefficient of variation (C.V.), the ratio of standard deviation to arithmetic mean is a normalized measure of dispersion of a probability distribution. It tells us about the size of variation relative to the size of the observation, and is independent of the units of observation. It is an index of the precision of the experiment. The coefficient of variation and the plot size relationship has been investigated by several researchers including Mahalanobis (1940) and Panse (1941), etc. Panse and Sukhatme (1954) gave detailed description of uniformity trial experiments. The determination of the optimum plot size is an important step in field experimentation as it takes into account variability, both due to crop species and soil heterogeneity.

Smith (1938) gave an empirical model for describing relationship between the variance and the plot size for his field experiments. His model can be reduced to the following simple form as

$$Y = a X^{b} \tag{1}$$

where Y is Coefficient of Variation and X is size of the plot, a and b being parameters of the model to be estimated.

Haque *et al.* (1988) considered the following two models along with the model (1) for describing relationship between the plot size (X) and the Coefficient of Variation (Y) as,

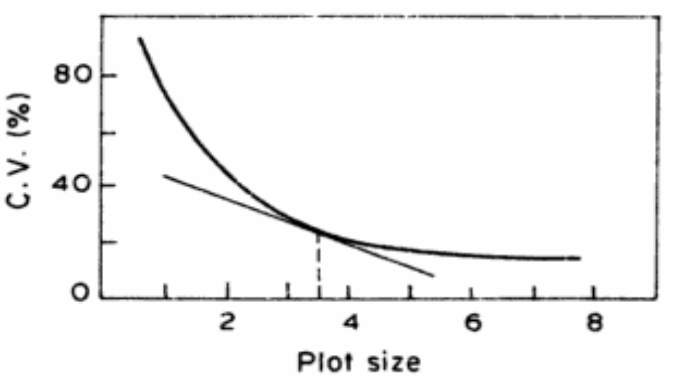
$$Y = a b^{X} \tag{2}$$

$$Y = a + \frac{b}{X} \tag{3}$$

Haque *et al.* (1988) arrived at the conclusion that the relationship (1) is the best among relationships (1) to (3) to describe the coefficient of variation and the plot size relationship. They calculated the point of maximum curvature for determining the optimum plot size and found the optimum plot size which corresponds to coefficient of variation (C.V.) of magnitude 25%. But this C.V. is quite high. They mentioned that in field experiments, generally the C.V. should not be more than 10-15%. If the C.V. is very high the reliability of the experimental results becomes doubtful. Therefore they suggested that instead of maximum curvature, it would be more logical to consider C.V. as the criterion for deciding the optimum plot size. In reference to the shape of the plots they showed that in all cases when x_1 (length) is measured along the fertility gradient and x_2 (width) across the fertility gradient rectangular plots are always optimum. They also suggested that if the experimenter has no idea of fertility gradient of the field, it is safer to use square shaped plots.

Draper and Smith (1998) classified the models (1) and (2) as intrinsically linear models, as they can be transformed into a form in which parameters appear linearly. The estimation of parameters a and b of these models can be done only after transforming them into a form in which parameters appear linearly by the well known method of least squares. The models (1) and (2) can be brought into linear form by using log transformation. However, it presupposes a multiplicative error term, a condition not so easy to justify. The direct application of least square method is not possible to estimate parameters of the models (1) and (2). Non-linear least squares estimation involves complicated iterative procedures. Convergence of solution is a serious problem in non-linear least squares estimation. Obtaining prior guess values of parameters in non-linear least squares estimation poses a serious problem before an investigator. The relation (3) is, however, a linear model and its parameter estimates can be obtained by direct application of classical least squares method of estimation.

The curvature is the amount by which a curve deviates from being flat. It is defined in different ways depending on the context. In uniformity trial experiments, the basic units of uniformity trials are combined to form new units. The new units are formed by combining columns, rows or both. Combination of columns and rows should be done in such a way that no column or row is left out. For each set of units, the coefficient of variation (C.V.) is computed. A curve is plotted by taking the plot size (in terms of basic units) on the X-axis and the C.V. values on the Y-axis of a graph sheet. The point at which the curve takes a turn that is the point of maximum curvature is located by inspection. The value corresponding to the point of maximum curvature will be the optimum plot size (Sundarraj, 1977). The following figure shows the point of maximum curvature expressed by dotted line.



This is only an approximate method of fixing the optimum plot size. Another method to obtain the point of maximum curvature is the calculus method. Fairfield Smith (1938) derived expression for maximum curvature for his model described by the relation (1) as

$$C = \frac{\left[1+(y_1)^2\right]^{3/2}}{y_2} \tag{4}$$

$$y_1 = \frac{dY}{dX} \quad \& \quad y_2 = \frac{d^2Y}{dX^2}$$

and therefore $$C = -\frac{1}{bX}(X^2+b^2)^{3/2} \tag{5}$$

On putting, $\frac{dC}{dX} = 0$, the solution of X will define the point of maximum curvature. For Fairfield Smith's model the value of C is

$$C = \frac{[1+\{abX^{(b-1)}\}^2]^{\frac{3}{2}}}{ab(b-1)X^{(b-2)}} \tag{6}$$

Putting $\frac{dC}{dX} = 0$ and substituting estimated values of the parameters a and b in it, the point of maximum curvature can be obtained.

In the present study we propose a linear model which relates the coefficient of variation to the plot size in a better way as compared to existing models. The expression for calculating the point of maximum curvature is also simple as compared to that of Fairfield Smith's model.

2. Proposed model

A linear model with its deterministic component is proposed to relate the plot size represented by X and Coefficient of Variation represented by Y as

$$Y = a + b\log X \tag{7}$$

The proposed model describes the relationship between the plot size and C.V. in a better way as compared to existing empirical models. a and b are parameters of the model which appear linearly in it and can be estimated by least squares method of estimation. The proposed model (7) was used by Shukla (2011) for his studies on uniformity trial experiments.

The model (7) admits an additive error term and can be written as

$$Y_i = a + b\log(X_i) + U_i, \qquad i = 1, 2....., n \tag{8}$$

where Y_i and X_i are i[th] observations of Y and X respectively. *U*s are independently and identically distributed random variables with mean zero and fixed variance σ^2. If *U*s follow normal distribution, i.e. $U \sim N(0, \sigma^2)$, the maximum likelihood estimates of a and b can also be obtained.

Following standard procedures as described in Draper and Smith (1998), the classical least squares estimators of parameters can be easily obtained. Let $\hat{a}$ and $\hat{b}$ are the least square estimates of a and b, respectively. The least squares estimate of Y_i that is $\hat{Y}_i$ will be

$$\hat{Y}_i = \hat{a} + \hat{b} \log X_i \tag{9}$$

The residual e_i is

$$e_i = Y_i - \hat{Y}_i \tag{10}$$

The appropriateness of the proposed model has been verified by examining the values of coefficient of determination- R^2, mean residual sum of square- s^2, mean absolute error (MAE), Akaike Information Criterion (AIC) and standardized residuals. Adopting the procedures as described in Montgomery *et al.* (2003), the analysis of residuals have been performed to verify the assumptions of zero mean, normal distribution and fixed variance of residuals. The point of maximum curvature can be obtained for the proposed model (7) as below,

$$C = -\frac{1}{bX}(X^2 + b^2)^{3/2} \tag{11}$$

On putting, $\frac{dC}{dX} = 0$, the point of maximum curvature can be obtained. It leads to the solution of $X = \pm\sqrt{b^2/2}$. As X will assume only positive values, the point of maximum curvature will be at $X = \sqrt{b^2/2}$. It is observed that expression for obtaining point of maximum curvature is much simpler for the proposed model as compare to that of Fairfield Smith's model.

3. Empirical study

The appropriateness and model adequacy of the proposed linear model (7) has been verified with the help of primary data given in Haque *et al.* (1988). Haque *et*

al. (1988) worked on field experiments for wheat and taken a piece of land measuring 45 x 39 m^2 at Rajendra Agricultural University, Bihar, India. At the time of harvest, the land was subdivided into 45 x 39 = 1755 basic units, of size 1 x 1 m^2 , and grain yield was recorded in gram for each unit separately. We have computed the values of Coefficient of determination R^2 , residual mean square s^2 , Mean absolute error (MAE) and Akaike Information Criterion (AIC) for the models (1) to (3) & (7) and these values are listed in table 1 along with parameter estimates of the model (7). An analysis of residuals has also been performed for the model (7) by plotting normal probability plot and residual versus explanatory variable plot. The normal probability plot (Fig.1) is almost a straight line which conforms the assumption of normal distribution of residuals. The plot of residuals versus explanatory variables (Fig.2) for the model (7) does not show any systematic pattern. It conforms the assumption of homoscedasticity for residuals. The MAE values are also negligible. Thus, we infer that residuals of the model (7) admit the assumption of zero mean, normal distribution and fixed variance. We can conclude that the proposed linear model (7) adequately explains the relationship between the plot size and the C.V.

Figure 1.

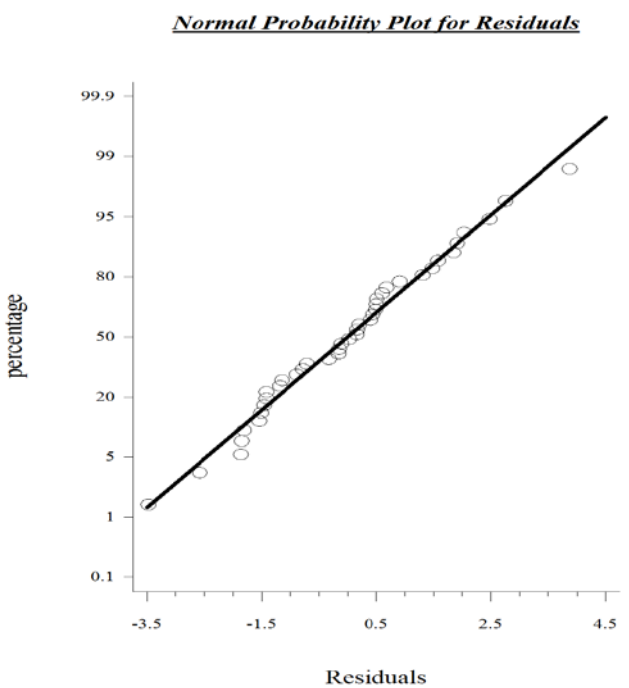
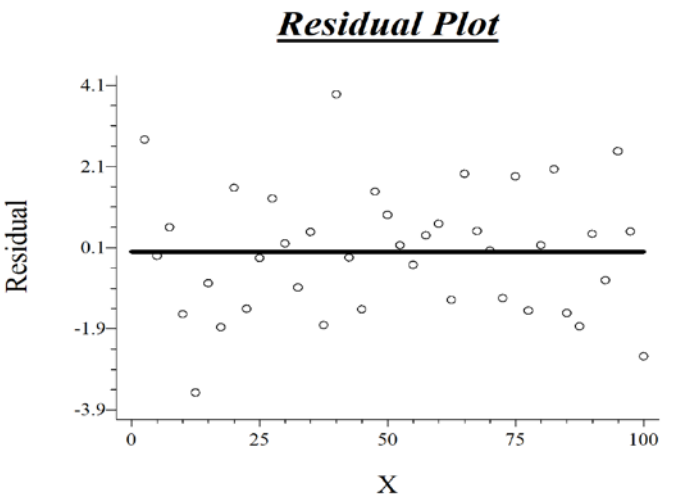


Figure 2.



On comparing values of R^2, s^2, MAE and AIC for the models (1) to (3) and (7) we have observed that the proposed linear model (7) has highest R^2 values and lowest s^2, MAE and AIC values. Thus, the model (7) better fits data sets as compared to the models (1) to (3). The model (7) is more appropriate to be used in uniformity trial experiments.

Table 1.

Parameters Estimates of Model (7)			$\hat{a}$ = 32.9928	$\hat{b}$ = -4.6674
	R^2	s^2	MAE	AIC
Model (7)	**0.9186**	**2.4638**	**1.2472**	**2.5868**
Model (1)	0.9110	2.6980	1.2547	2.8325
Model (2)	0.7810	6.6170	1.8190	6.9475
Model (3)	0.6750	9.7969	2.4658	10.2859

Table-2 gives the values of C.V. for different plot areas. Estimated values of C.V. using the model (1) and (7) are also given.

Table 2.

S. No.	Area(m^2) x	C.V. y	C.V. $\hat{y}_{(1)}$	C.V. $\hat{y}_{(7)}$	S. No.	Area(m^2) x	C.V. y	C.V. $\hat{y}_{(1)}$	C.V. $\hat{y}_{(7)}$
1	1	35.75	37.39	32.99	21	30	17.28	16.57	17.12
2	2	29.65	31.67	29.76	22	32	16.49	16.31	16.82
3	3	28.47	28.74	27.87	23	35	16.80	15.97	16.40
4	4	24.98	26.83	26.52	24	36	16.95	15.86	16.27
5	5	22.01	25.44	25.48	25	40	14.59	15.46	15.78
6	6	23.85	24.35	24.63	26	42	17.46	15.28	15.55
7	7	22.05	22.73	23.91	27	45	15.73	15.03	15.23
8	9	24.31	22.10	22.74	28	48	14.95	14.80	14.92
9	10	20.83	21.55	22.25	29	50	13.59	14.66	14.73
10	12	21.24	20.63	21.39	30	54	16.23	14.39	14.37
11	14	21.98	19.88	20.68	31	56	12.75	14.27	14.20
12	15	20.55	19.56	20.35	32	60	14.04	14.03	13.88
13	16	19.17	19.26	20.05	33	63	15.69	13.87	13.66
14	18	19.99	18.72	19.50	34	64	12.07	13.82	13.58
15	20	17.20	18.25	19.01	35	70	11.32	13.53	13.16
16	21	22.66	18.04	18.78	36	72	13.47	13.43	13.03
17	24	18.01	17.47	18.16	37	80	11.83	13.10	12.54
18	25	16.55	17.30	17.97	38	81	14.96	13.06	12.48
19	27	19.09	16.99	17.61	39	90	12.49	12.74	11.99
20	28	18.35	16.84	17.44	40	100	08.92	12.42	12.50

The point of maximum curvature for the proposed model (7) is $x = 3.30$, hence the optimum plot size which falls just near to this point of maximum curvature is $3m^2$ corresponding to which the C.V. is 27.86%, which is quite high. Therefore, as suggested by Haque *et al.* (1988), it would be more logical to consider C.V. as the criterion for deciding the optimum plot size.

4. Conclusions

It is submitted that the linear model (7) is a better alternative to describe the relationship between the plot size and the coefficient of variation in uniformity trial experiment. The proposed model (7) has highest R^2 values as compared to the models (1) to (3) which include Fairfield Smith's model also. Apart from it the model (7) has smallest values of s^2, MAE and AIC, as compared to all other models (1) to (3). The analysis of residuals also conforms the assumptions of zero mean, normal distribution and fixed variance for residuals. The expression for obtaining the point of maximum curvature is also easy to use for the model (7). The parameter estimates of the proposed model posses good statistical properties. Another advantage with this model is that it admits additive error term. The predictions and inferences as well as test of significance procedures for the model (7) can be easily carried out. It is therefore recommended that the linear model (7) should preferably be used in uniformity trial experiments.

Acknowledgements

The authors are very much thankful to the editor in chief and the unknown learned referee for critically examining the manuscript and giving valuable suggestions to improve it.

REFERENCES

DRAPER, N. R. AND SMITH, H., (1998). Applied Regression Analysis, Wiley, New York.

HAQUE, H. N. AZAD, N. K. JHA, R. N. and SINGH, S. N., (1988). Optimum Size and Shape of Plots for Wheat, Annals of Agricultural Research, 9, 2, 165-170.

MAHALANLBIS, P. C., (1940). A sample survey of the acreage under jute in Bengal, Sankhya, 4, 511-530.

MONTGOMERY, D. C. PECK, E. A. and VINING, G. C., (2003). Introduction to Linear Regression Analysis, Wiley.

PANSE, V. G., (1941). studies of the technique of field experiments size and shape of blocks and arrangements of plots in cotton trials, Indian J. agric. Sci, 11, 6, 850-865.

PANSE, V. G. and SUKHATME, P.V., (1954). Statistical Methods for Agricultural Workers, I C A R New Delhi.

SHUKLA, A. K., (2011). Some Contributions in Estimation of Asymptotic Regression Statistical Models, the Unpublished Thesis of Ph. D., CSJM University, Kanpur.

SMITH, H. F., (1938). An empirical law describing heterogeneity in the yield of Agricultural crops, Journal of Agricultural Science, 28, 1-23.

SUNDARARAJ, N., (1977). Technique for estimating optimum plot size and shape from fertilizer trial data, a modified approach. J. Ind. Soc. Agric. Stat., 29, 2, 80-4.

STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 171–182

MODEL OF LATENT PROFILE FACTOR ANALYSIS FOR ORDERED CATEGORICAL DATA

Piotr Tarka[1]

ABSTRACT

In the literature factor analysis is admittedly a well-known and effective multivariate method in the reduction of extensive and broad data, e.g., in the analysis of too many variables. It is also known for the process of unidimensional or multidimensional scale/s construction. Typically, in many studies (especially those pertaining to market research area) a common factor analysis solution is used (based on continuous data). However, there are rarely ever undertaken studies pertaining to latent variable models where other type of data is used based on discrete variables. One of these models might be called Latent Profile Factor Analysis - LPFA. In this article author's main objective is to propose and discuss its (LPFA) main assumptions. In order to prove the model's functionality in practice of market research, a brief example of LPFA model for ordered categorical data (based on one-factorial solution) in reference to hedonic consumption data is given at the end of the paper.

Key words: latent profile factor analysis model, ordered categorical data.

1. Introduction

Most of professional researchers in the socio-economic field, when analyzing market and people-customers' traits, often conduct projects in statistical research based on qualitative data. Most of them are thus forced to describe customers by simply asking questions (including prepared earlier set of items) about their hidden and unknown structure concerning for example personal attitudes, feelings or values. In consequence, in order to examine internal structure of customers, they need to implement an appropriate model for the purposes of data reduction, facing the problems of broad data, e.g. including too many variables. Researchers struggle also with the selection of appropriate method in order to increase precision level in the analysis according to the type of collected data. Solutions, as

[1] Poznan University of Economics, Department of Marketing Research, al. Niepodległości 10, Poland, e-mail: piotr.tarka@ue.poznan.pl.

usual, come with latent variable models based on multivariate complexity. And because in social sciences and in many surveys (undertaken within market research) collected data is mainly of categorical nature, and categorical variables are definitely more used than continuous variables, hence they need a more sophisticated latent variable model to examine this type of data as compared to classical solution based on common factor analysis (Vasdekis et al., 2008). By term "ordered categorical" we will refer to type of data being measured on ordinal variables. For instance in market survey, respondents are often asked to characterize their opinions or attitudes (e.g. about products, etc.) on measurement scales where answers are ranging from "strongly disagree" to "strongly agree". This is a common example of such data. This type of data is also known as the *ordinal logit, ordered polytomous logit, constrained cumulative logit, proportional odds* (Borooah, 2002; Cohen et al., 2003; DeMaris, 2004; Hoffmann, 2004; Long and Freese, 2006). The most natural way to view structure in ordinal data is to postulate the existence of an underlying latent (unobserved) variable associated with each respondent's response – observed variable. Unfortunately as it often happens in research practice, the analysis of such data is performed without regard to their ordinal nature (Agresti 2007). For this reason in this article the author investigates the most important characteristics and specificity of Latent Profile Factor Analysis (LPFA). LPFA model is designed for data, that is originating strictly from ordered categorical responses. At the end of paper a practical example of this model is given.

2. Generalized Linear Latent Variable Model

Generalized Linear Latent Variable Model (GLLVM) could be approximately a framework or some kind of a background for construction of Latent Profile Factor Analysis Model (LPFA) for ordered categorical responses. As far as the GLLVM model is concerned, it includes (Moustaki and Knott 2000; Moustaki 2003):

- the random component in which each of the p random response variables, $\left(x_1,...,x_p\right)$ has a distribution from the exponential family such as Bernoulli, Poisson, Multinomial, Normal, Gamma,
- the systematic component in which latent variables vector and covariates vector $\mathrm{z}'=\left(\mathrm{z}_1,...,\mathrm{z}_q\right)$, $\mathrm{x}'=\left(x_1,...,x_r\right)$ produce a linear predictor η_i corresponding to each category of x:

$$\eta_i=\alpha_{i0}+\sum_{j=1}^{q}\alpha_{ij}\mathrm{z}_j+\sum_{l=1}^{r}\beta_{il}x_l,\qquad i=1,...,p. \tag{1}$$

- the links between the systematic component and the conditional means of the random component distributions:

$$\eta_i=\upsilon_i\left(\mu_i\left(\mathrm{z},\mathrm{x}\right)\right) \tag{2}$$

where:

$$\mu_i\left(\mathrm{z},\mathrm{x}\right)=E\left(x_i\middle|\mathrm{z},\mathrm{x}\right) \tag{3}$$

and υ_i is called the link function which can be any monotonic differentiable function and may be different for different manifest variables $x_i \;\; i=1,...,p.$

We shall also assume that $\left(x_1,x_2...,x_p\right)$ denotes a vector of p manifest variables where each variable has a distribution in the exponential family taking the form:

$$g_i\left(x_i;\theta_i,\phi_i\right)=\exp\left\{\frac{x_i\theta_i-b_i\left(\theta_i\right)}{\phi_i}+d_i\left(x_i,\phi_i\right)\right\},\quad i=1,...,p, \tag{4}$$

where $b_i\left(\theta_i\right)$ and $d_i\left(x_i,\phi_i\right)$ are specific functions taking a different form depending on the distribution of the response variable x_i.

Because of the existence of different types of collected responses (depending on type of used measurement scale) there will be different distribution forms, which we rearrange respectively to their specific transformation functions (Table 1).

Table 1. Distributions and transformation functions from Generalized Linear Model approach

Scale type x_i	**Distribution** $f\left(x_i\middle\|\theta\right)$	**Transformation** $g\left[E\left(x_i\middle\|\theta\right)\right]$
Dichotomous	Binomial	Logit
Nominal	Multinomial	Logit
Ordinal	Multinomial	Restricted logit
Count	Poisson	Log
Continuous	Normal	Identity

Source: Vermunt and Magidson 2005.

And from perspective of GLLVM approach we may further assume a four-fold classification including sub-models of latent variables. They are: *Factor Analysis (FA), Latent Trait Factor Analysis (LTFC), Latent Profile Factor Analysis (LPFA), and Latent Class Analysis (LCA),* as shown in Table 2. The fundamental distinction in this classification is the one between continuous and discrete latent variables, so that a researcher has to decide whether to treat the underlying latent variable(s) as continuous or discrete. In case of LPFA model, the latent variable is assumed to be discrete and to come from a multinomial distribution.

Table 2. Classification of Latent Variable Models

Manifest Variables	**Latent variables**	
	Continuous	Categorical
Continuous	**Factor Analysis (FA)**	**Latent Profile Factor Analysis (LPFA)**
Categorical	**Latent Trait Factor Analysis (LTFA)**	**Latent Class Analysis (LCFA)**

Source: own construction based on Bartholomew and Knott 1999.

3. Latent Profile Factor Analysis (LPFA) against a background of other useful models

Classical Factor Analysis (FA) is a popular used tool in market research where in a given set of manifest variables one wants to find a set of latent variables $\xi_1,...,\xi_k$, fewer in number than the manifest variables, which contain essentially the same information. Although FA is meant in general for continuous observed indicators, it is often used by researchers with ordinal models which are based on other types of discrete variables. This mistake yields in the end results that might be incorrect. Not only parameter estimates may be biased, but also goodness-of-fit indices cannot be trusted (Moustaki and Jöreskog, 2006).

Latent Profile Factor Analysis (LPFA) differs from standard Factor Analysis mainly in the sense that the observed variables are either ordered categorical variables (e.g.: "very much", "a little", "not very much") or measured on attitudinal statements (such as: "strongly disagree", "disagree", "agree", "strongly agree"). These answers collected from survey fall into only one category. Such categorization makes the data of ordinal nature. However, as already mentioned, assumptions of ordinality of data in practice of market research is often ignored and numbers such as 1, 2, 3, 4, 5 representing ordered categories are treated as numbers having metric properties 1-2-3-4-5 which yields incorrect results. In consequence, ordered categorical data (which has for example number of five or seven categories) is by mistake of many analysts treated as if there were some kinds of interval level variables in it. Indeed, proceeding in that way with standard factor analysis allows them to compute correlations on the basis of so-called *pseudo-continuous variables*. Moreover, this uncritical approach to application of factor analysis associated with ordered categorical data is likely to give biased estimates of the factor loadings. Hence, the better solution in finding relationships between ordered categorical data comes with minor modifications of Item Response Theory Models where one assumes that the responses to the ordinal items are independent conditional on the latent variables (conditional independence) (Bartholomew 2002). For ordered-response categories (which

appear in LPFA) IRT models[1] are definitely more informative and reliable than simply scored items by FA.

However, IRT solution is not yet enough. In order to construct a good model of LPFA we need to focus additionally on Latent Trait Factor Analysis for binary data, where we usually analyze the probability of a randomly selected individual giving a positive response to an item as a function of the latent variables. In case of ordinal data, where more than two categories exist, we simply need to specify probabilities for each category. As a result the observed ordinal variables are denoted by $x_1,\ldots,x_p$. Let us suppose that there are m_i categories for variable i labelled $(1,\ldots,m_i)$. For binary items $m_i = 2$ (for each i) the category labels are usually denoted as 0 and 1 but they could equally well have been marked as 1 and 2. In LPFA we need to redefine a response probability for each category. Let now $\pi_{i(s)}(\mathrm{F})$ be the probability so that given F a response falls in category s for i-th variable. The position with two categories can be then compared with the general case as follows:

Categories	0	1
Response probability	$1-\pi_i(\mathrm{f})$	$\pi_i(\mathrm{f})$

Categories	1	2	…	s	…	m_i
Response probability	$\pi_{i(1)}(\mathrm{f})$	$\pi_{i(2)}(\mathrm{f})$	…	$\pi_{i(s)}(\mathrm{f})$	…	$\pi_{i(m_i)}(\mathrm{f})$

In both cases, the response probabilities sum to one. The question is now on how to use logit model (expressing the logit of probability of response in category as a linear function of *f*) for more than just two categories. Suppose we divided categories into two groups with categories $(1,2,\ldots,s)$ - into one group and $(s+1,s+2,\ldots,m_i)$ - into other group and were merely to report into which of these two groups the response fall. We would thereby need to reduce the polytomous variables to a binary variable. Therefore, it seems reasonable to infer that any model we choose for polytomous case should be consistent with the one

[1] IRT model(s) may be characterized by a few options such as (Embretson and Reise, 2000): *Graded Response Model* (Samejima, 1969), *Modified Graded Response Model* (Muraki, 1992), (which is used with questionnaires that have a common rating scale format (e.g., all item responses scored on a five-point scale). These two models are considered as "indirect" models because a two-step process is needed to determine the conditional probability of the response in particular category. The other remaining models are considered as "direct" IRT models because only a single equation is needed to describe the relationship between respondent response level and the probability of responding in particular category. Specifically there are two polytomous models that are extensions of the Rasch model, e.g. *Partial Credit Model* (Masters, 1982) and *Rating Scale Model* (Andrich, 1978 a-b).

which we use also for the binary case. As a result, in order to make ordered categorical model (in LPFA) more effective, we need to apply binary logit model. To do so we must split the binary model where the probabilities of a response fall into the first and second group, which may be written as follows (Bartholomew, 2002):

$$\gamma_{i(s)}(\mathrm{f}) = \Pr(y_i \le s) = \pi_{i(1)}(\mathrm{f}) + \pi_{i(2)}(\mathrm{f}) + \ldots + \pi_{i(s)}(\mathrm{f}), \tag{5}$$

and

$$1 - \gamma_{i(s)}(\mathrm{f}) = \Pr(y_i > s) = \pi_{i(s+1)}(\mathrm{f}) + \pi_{i(s+2)}(\mathrm{f}) + ,\ldots, + \pi_{i(m_i)}(\mathrm{f}), \tag{6}$$

where: s denotes the category into which the i-th variable falls.

$\gamma_{i(s)}(\mathrm{f})$ - the probabilities are referred to as cumulative response probabilities .

Next we need to define the model, supposing that binary logit model holds for all possible divisions of the m_i categories into two groups. We can do this by specifying the model in terms of logit $\gamma_{i(s)}(\mathrm{f})$ or logit $\left(1 - \gamma_{i(s)}(\mathrm{f})\right)$.

The model is thus expressed as follows:

$$\log\left[\frac{\gamma_{i(s)}(\mathrm{f})}{1 - \gamma_{i(s)}(\mathrm{f})}\right] = \alpha_{i(s)} - \sum_{j=1}^{k} \alpha_{ij} f_j, \tag{7}$$

where: $(s = 1,\ldots, m_i - 1;\ i = 1,\ldots, p)$.

For a positive factor loading α_{ij} the higher the value of an individual on the latent variable f_j, the higher the probability of that individual responding in the higher categories of item *i*. The intercept parameter $\alpha_{i(s)}$ is one for each category. The ordering of the categories implies that the intercept parameters are also ordered:

$$\alpha_{i(1)} \le \alpha_{i(2)} \le \ldots \le \alpha_{i(m_i)}. \tag{8}$$

In consequence the factor loadings remain the same across categories of the same variable. Otherwise, the discriminating power of the item does not depend on where the split into two groups was made. The π's are obtained from the γ's by:

$$\pi_{i(s)}(\mathrm{f}) = \gamma_{i(s)}(\mathrm{f}) - \gamma_{i(s-1)}(\mathrm{f}) \qquad (s = 2,\ldots, m_i), \tag{9}$$

where $\gamma_{i(1)}(\mathrm{f}) = \pi_{i(1)}(\mathrm{f})$ and $\gamma_{i(m_i)}(\mathrm{f}) = 1$. We refer to $\gamma_{i(s)}(\mathrm{f})$ as cumulative response function and to $\pi_{i(s)}(\mathrm{f})$ as category response function.

4. Goodness-of-fit in Latent Profile Factor Analysis – (LPFA)

The LPFA model should be fitted in the same way as the binary latent trait model using the method of maximum likelihood. Goodness-of-fit can likewise be judged using the same criteria based on the likelihood ratio G^2 and the Pearson chi-squared χ^2 statistics calculated from the whole response patterns as follows:

$$G^2 = 2\sum_{i=1} O_i \log \frac{O_i}{E_i} \tag{10}$$

$$\chi^2 = \sum_{i=1} \frac{\left(O_i - E_i\right)^2}{E_i} \tag{11}$$

where O_i and E_i are the observed and expected frequency of response pattern i. When the sample size n is large and p small, the statistics under the hypothesis that the model fits follow a chi-square distribution with degrees of freedom the number of response patterns minus the number of independent parameters minus one. As the number of items increases, the chi-square approximation to the distribution of either goodness-of-fit statistic ceases to be valid. Parameter estimates are still valid but it is difficult to assess the model.

The goodness-of-fit in Latent Profile Factor Analysis can be also assessed by looking at the two or three-way margins. The pairwise distribution of any two variables is then displayed as a two or three-way contingency table, and chi-squared residuals are constructed by comparing the observed and expected frequencies. The differences are computed using G^2 and χ^2 statistic. If there are small differences, it means the associations between all pairs of responses are well predicted by the model.

5. Example on construction *LPFA one-factorial model* in reference to hedonic consumption data

For demonstration purposes the data set that was extracted from the earlier study conducted by the author (Tarka, 2010) was prepared. The data included the responses given by 232 individuals to four below listed items concerning attitude to hedonic consumption-oriented lifestyle. For each item (statement), respondents were asked the following response alternatives based on four-point scale: [1] = strongly disagree, [2] = disagree to some extent, [3] = agree to some extent, [4] = strongly agree. And the chosen questions were given:

1: *I'm money-oriented person and looking for wealth in my life* [money]
2: *I'm striving to be free in my private life with no family frontiers* [freedom]
3: *I'm having a good time and enjoying only things I like and prefer* [party]
4: *I'm looking for adventurous and risky life* [full of life]

The output of the one-factor analysis for above ordered categorical data is given below. For calculations the author used LAMI software which contains an interface that allows users of the GENLAT and LATCLASS programs to run their analyses conveniently than using the original DOS programs directly. The program fits a latent trait model for ordinal observed variables with up to two latent variables. The program computes parameter estimates, standard errors, chi-squared residuals, and scoring methods.

In order to start program input file parameters were specified as follows:

One-Factorial Model = 1
Number of Observed Variables = 4
Number of Ordinal Variables = 4
Number of Cases Sampled = 232
Proportion of Response Patterns with at Least One Missing Observation = 0,00
Number of Quadrature Points Used = 8
Maximum Number of Iterations Permitted = 2000
Convergence Tolerance For The Relative Likelihood Value = 0.00000000
…….
NFAC: Number of factors (1)
INIT: 0 if the initial parameter values are set in the program or 1 if the initial parameter estimates are to be read from file
ITER: Number of iterations (maximum is 2000)
PREC: Precision for maximization (e.g. 0.0000001, convergence tolerance of the EM algorithm)
SCOR: 1 if scoring results to be printed, 0 otherwise.

Source: Own construction based on LAMI software.

Finally, we obtained the following estimated scores (according to printed version in LAMI software. These results are shown in Tables 3-8. From Table 3 we can observe that percentage of individuals *agreeing to some extent* or *agreeing strongly* (categories 3 and 4) is larger as compared to other two response categories denoted by 1 and 2. Having inspected the results (in case of binary data), we looked at pairwise associations between four mentioned above items which suggest there exists some real common underlying factor including all four items. They can be considered as indicators for measuring respondents' attitude to lifestyle based on hedonic consumption.

Table 3. Items – category frequencies

Item 1		Item 2	
1	0,0043103	1	0,0732759
2	0,0689655	2	0,2413793
3	**0,6810345**	**3**	**0,5689655**
4	**0,2456897**	**4**	**0,1163793**

Table 3. Items – category frequencies (cont.)

Item 3		Item 4	
1	0,0431034	1	0,0387931
2	0,1810345	2	0,2413793
3	**0,5474138**	**3**	**0,5172414**
4	**0,2284483**	**4**	**0,2025862**

Source: Own calculations based on LAMI software.

Also the parameters and standard errors based on maximum likelihood estimates for particular categories associated with respective items (Tab. 4) indicate that the strongest relationships appear mainly in the third category containing positive values and smaller standard errors (S.E.). This result simply means that third category of the respective item (I) will compose to a greater part our considered one factorial-model.

Table 4. Maximum likelihood estimates of item parameters and standard errors (S.E.)

Item	Category	(I - Item, J - Factor)	S.E
1	1	-5,888	6,498
1	2	-2,920	2,523
1	**3**	**1,328**	**0,328**
1	4	-3,420	3,613
2	1	-3,035	1,059
2	2	-1,003	0,743
2	**3**	**2,481**	**0,371**
2	4	-2,820	1,583
3	1	-4,587	1,876
3	2	-2,019	1,304
3	**3**	**1,964**	**0,369**
3	4	-4,423	3,523
4	1	-3,616	2,470
4	2	-1,112	0,803
4	**3**	**1,624**	**0,314**
4	4	-1,920	0,423

Source: Own calculations based on LAMI software.

Now, if we decide to fit this type of one-factor model based on hedonic consumption data, we need to obtain the estimates given in Table 5. The Alpha's are simultaneously representing factor loadings. They are defined in the literature as discriminating parameters. If the values of factor loadings are large and they all are positive but the standard errors are small, then there is an underlying factor which is common to all items. And this is purely visible in our case. The high values of standardized loadings (Table 6) also suggest that the single factor model provides a good explanation for all four ordinal (ordered categorical) items,

especially for item number 3. However, before putting too much weight on this conclusion, we need to look at how well the model fits.

Given the sparsity of the data (at total frequency of 232 spread over multiple response categories), it is not feasible to carry out global tests. Instead, we need to look at the fits to the margins. Therefore, for each pair of items, (see Table 7) we need to calculate the sum of the chi-squared residuals over each pair of item categories. Sixteen chi-squared residuals for each pair of items were generated, since each variable had four response categories.

Table 5. Alpha as factor loadings and standard errors (S.E.) for items

Items	Alpha(1,I)	S.E
1	0,982	0,241
2	1,167	0,313
3	**2,009**	0,398
4	1,000	0,212

Source: Own calculations based on LAMI software.

Table 6. Standardized Loadings for items

Items	St. Alpha(1,I)
1	0,7008
2	0,7592
3	**0,8952**
4	0,7072

Source: Own calculations based on LAMI software.

Table 7 shows how the entry 20,47 (due to calculations based on two-way margins of selected items "Money" and "Full of Life" of Table 8) is computed. The sum of the entries of Table 8 is 20,47. In similar way we computed the sums of chi-squared residuals for other two-way tables including another pairs of items. In order to confirm if the model is correct, we need to check the chi-squared residuals. Values greater than about 4 would indicate a poor fit. For instance, as observed from Table 8, values larger than 4 do not appear. For the best part of cells they are considerably below 4. In other words these associations make up a good configuration for our items in the model.

Table 7. Sum of chi-squared residuals for all pairs of items derived from the two-way margins for one-factorial model

Items	Money	Freedom	Party
F.o.life	**20,47**	10,58	23,32
Money		17,21	12,45
Freedom			9,56

Source: Own calculations based on LAMI software.

Table 8. Chi-squared residuals for the two-way margins of selected pair of items "Money" and "Full of Life"

Category		**Money** 1	2	3	4
F.o.Life	1	0,87	1,65	0,34	0,90
	2	2,40	3,40	2,10	1,09
	3	2,34	1,13	1,09	0,56
	4	1,02	1,30	0,21	0,94

Source: Own calculations based on LAMI software.

Since the sum of these residuals over all the cells in a two-way marginal table is analogous to Pearson's chi-squared statistic for goodness-of-fit, but because the model has been fitted to the full multi-way table, the standard chi-squared test does not apply. We may still use this sum, as a diagnostic, e.g. D. Larger value of D would then suggest that the associations in two-way table are not well explained. As a rule of thumb that D is too large we need to take into account value that is greater than upper 1% point of a chi-square distribution with $\left[\left(m_i \times m_j\right)-1\right]$ degrees of freedom. And as observed from the results (Table 7), each pair of all analyzed six entries has values of D less than 28,58 (the upper 1% point of chi-square with 15 degrees of freedom). Therefore, the fit to each two-way marginal table (pair of item) appears satisfactory. Overall, the one-factor model appears to give an adequate description of data. Therefore, we can use this factor as a summary measure of attitude to hedonic consumption issues.

6. Conclusions

Latent Profile Factor Analysis (LPFA), being a part of four latent variable models, is a powerful and useful tool for researchers. However, this model has been languishing too long on the borders of statistics and most importantly in research practice. It is slowly and surely taking its right place in the main stream, stimulated in part by the recognition of its greater value and sound foundations which have been given to it within a statistical framework. Assuredly, this new solution clarifies, simplifies and reduces broad data as far as the ordered categorical responses are concerned into more simple form than the previous model based on classical factor analysis. LPFA model would not be for sure possible without earlier progress of Item Response Theory which supported to a greater extent the development of Latent Profile Factor Analysis.

REFERENCES

AGRESTI, A., (2007). An introduction to categorical data analysis, 2nd ed. New Jersey: John Wiley and Sons.

ANDRICH, D., (1978a). Application of psychometric model to ordered categories which are scored with successive integers, "Applied Psychological Measurement", 2, pp. 581–594.

ANDRICH, D., (1978b). A rating formulation for ordered response categories, "Psychometrika", 43, pp. 561–573.

BARTHOLOMEW, D. J., KNOTT, M., (1999). Latent variable models and factor analysis. London: Arnold.

BARTHOLOMEW, D. J., (2002). Old and new approaches to latent variable modelling, [in:] MARCOULIDES, G. A., MOUSTAKI, I. (Eds.). Latent variable and latent structure models, New Jersey: Lawrence Erlbaum Associates, pp. 1–15.

COHEN, J., COHEN, P., WEST, S. G., AIKEN, L. S., (2003). Applied multiple regression/correlation analysis for the behavioral sciences, (3rd ed.). Mahwah, New York: Lawrence Erlbaum.

DEMARIS, A., (2004). Modeling continuous and limited response variables. Hoboken, New York: John Wiley and Sons.

EMBRETSON, S. E., REISE, S., (2000). Item response theory for psychologists, New Jersey: Lawrence Erlbaum Associates.

JÖRESKOG, K. G., MOUSTAKI, I., (2006). Factor analysis of ordinal variables with full information maximum likelihood, unpublished report.

LONG, J. S., FREESE, J., (2006). Regression models for categorical dependent variables using Stata, (2nd ed.). College Station, TX: Stata Press.

MAGIDSON, J., VERMUNT, J. K., (2005). Factor analysis with categorical indicators – a comparison between traditional and latent class approaches, [In:], VAN DER ARK, L. A., CROON, M. A., SIJTSMA, K., (Eds.), New developments in categorical data analysis for the social and behavioral sciences, New Jersey: Lawrence Erlbaum Associates, pp. 41–62.

MASTERS, G. N., (1982). A Rasch model for partial credit scoring, "Psychometrika", 47, pp. 149–174.

MOUSTAKI, I., KNOTT, M., (2000). Generalized latent trait models, "Psychometrika", 65, pp. 391–411.

MOUSTAKI, I., (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables, "British Journal of Mathematical and Statistical Psychology", 56, pp. 337–357.

SAMEJIMA, F., (1969). Estimation of latent ability using a response pattern of graded scores, "Psychometrika Monograph Supplement", 17, pp. 1–97.

TARKA, P., (2010). Latent variable models - issues on measurement and finding exact constructs – "Przegląd Statystyczny", 4, pp. 142–167.