STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

From the Editor 187

Submission information for authors 191

Sampling methods and estimation

Dihidar K., Bhattacharya M., Estimating sensitive population proportion using a combination of binomial and hypergeometric randomized responses by direct and inverse mechanism 193

Titt E. M., Residency Testing. Estimating the true population size of Estonia 211

Beevi N. T., Chandran C., Efficient family of ratio-type estimators for mean estimation in successive sampling using auxiliary information on both occasions 227

Alkaya A., Ayhan Ö. H., Esin A., Sequential data weighting procedures for combined ratio estimators in complex sample surveys 247

Research articles

Łukaszonek W., A multidimensional and dynamised classification of Polish provinces based on selected features of higher education 271

Shanker R., Shukla K. K., Mishra A., A three-parameter weighted Lindley distribution and its applications to model survival time 291

Karadeniz P. G., Ercan I., Examining tests for comparing survival curves with right censored data 311

Rai P. K., Pareek S., Joshi H., Met and unmet need for contraception: Small Area Estimation for rajasthan state of India 329

Conference reports

The XXXV International Conference on Multivariate Statistical Analysis MSA 2016 (7-9 November, 2016), Łódź, Poland (Szczepocki P., Białek, J.) 361

About the Authors 365

Volume 18, Number 2, June 2017

ISSN 1234-7655

Editorial Office
Marek Cierpiał-Wolan, Ph.D., Scientific Secretary
m.wolan@stat.gov.pl
Secretary:
Patryk Barszcz, P.Barszcz@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, June 2017
Vol. 18, No. 2, pp. 187−189, DOI 10. 21307

FROM THE EDITOR

In the current issue of *Statistics in Transition new series* four articles are devoted to the problems of estimation and sampling while in the next four ('research articles') a variety of different topics are discussed.

In the first paper, entitled ***Estimating sensitive population proportion using a combination of binomial and hypergeometric randomized responses by direct and inverse mechanism,*** **Kajal Dihidar** and **Manjima Bhattacharya** follow the approach discussed in some earlier papers (Chaudhuri and Dihidar 2014, and Dihidar 2016) towards estimating the sensitive population proportion using a combination of binomial and hypergeometric randomized responses by direct and inverse mechanism. Along with the traditional simple random sampling, with and without replacement, here they consider sampling of respondents by unequal probabilities. They start with an observation that individuals in a sample survey may (for various reasons) prefer not to confide to the interviewer the correct answers to certain potentially sensitive questions such as the illegal use of drugs, illegal earning, or incidence of acts of domestic violence, etc. The resulting evasive answer bias is ordinarily difficult to assess. The use of a randomized response method for estimating the proportion of individuals possessing those sensitive attributes can potentially eliminate the bias, as the authors demonstrate with numerical illustration, comparing a study of the relative efficiencies of the direct and an inverse mechanism.

Ene-Margit Titt's paper, ***Residency Testing. Estimating the true population size of Estonia***, discusses the idea of using the residency index as a tool created for estimating the under- and over-coverage of population census and calculation of proper population size. For this aim the concept of a sign of life – a binary variable depending on register i, person j and year k has been introduced showing if the person was active in the register in a given year. The weighted sum of signs of life indicates the probability that the person belongs to the set of residents in a given year. To improve the stability of the index a linear combination of the previous value of the index and the sum of signs of life is used. Compared with other types of statistical models, the advantage of this methodology is the possibility of using a large number of different registers, which may have both positive and negative impact on the residency status. Some obstacles to using such an index are also mentioned – in some situations, the residency indexes cannot be used, for instance, for describing interior migration and population in small areas.

In the next article, ***Efficient Family of Ratio-Type Estimators for Mean Estimation in Successive Sampling Using Auxiliary Information on Both Occasions,*** **Beevi Nazeema T., Chandran C.,** propose an efficient family of ratio-type estimators using one auxiliary variable for the estimation of the current population mean under successive sampling scheme following methodology originally studied by Ray and Sahai (1980) under simple random sampling using one auxiliary variable for estimation of the population mean. They employ these estimators in successive sampling, with usual ratio estimator being identified as a particular case of the suggested estimators. The proposed family of estimators at optimum condition is compared with the simple mean per unit estimator, Cochran (1977) estimator and existing other members of the family. Among the results, the authors indicate that a smaller fresh sample at current occasion is required if a highly positively correlated auxiliary character is available, what also may reduce the cost of the survey.

Paper by **Alkaya Aylin, Ayhan Öztes H., Esin Alptekin**, ***Sequential Data Weighting Procedures for Combined Ratio Estimators in Complex Sample Surveys***, is devoted to the issue of sample surveys weighting procedure to increase the quality of estimates. While there are many types of situations calling for such procedure, the authors point to unequal probability of selection, compensation for nonresponse, and post-stratification as to the most important reasons of weighting. The authors propose a sequential data weighting procedure for the estimators of combined ratio mean in complex sample surveys and general variance estimation for the population ratio mean. They illustrate the utility of the proposed estimator using data from Turkish Demographic and Health Survey 2003, and showing that that auxiliary information on weights can considerably improve the efficiency of the estimates of post-stratification.

The research articles set starts with the paper by **Wojciech Łukaszonek's** paper, ***A Multidimensional and Dynamised Classification of Polish Provinces Based on Selected Features of Higher Education***, which addresses the issue of distribution of the unprecedented, fivefold increase in the number of students and the number of higher educational (HE) institutions in Poland over the so-called, post-1989, transition period. Given that the distribution was not uniform in any respect (space or time), the regional differentiation between country provinces is analysed with special attention being paid to the years 2002–2013. The applied procedure uses new statistical methods applicable to a space of double multivariate data. The covariance matrix used to construct principal components is structured as a Kronecker product. The results led to the identification of six groups of provinces, including two consisting of a single province – Masovian and Lesser Poland – which contain the biggest and the highest-ranked HE institutions in Poland (the University of Warsaw and Jagiellonian University).

The next paper, by **Shanker Rama, Shukla Kamlesh Kumar, Mishra Amarendra,** ***A Three-Parameter Weighted Lindley Distribution and its Applications to Model Survival Time*** treats about a semiparametric additive risks regression model for analysing middle-censored lifetime data arising from an unknown population. The authors propose a dual approach to estimating the regression parameters and the unknown baseline survival function: the first method uses the martingale-based theory, and the other one uses an iterative procedure. Results of simulation studies are reported to assess the finite sample behaviour of the estimators, followed by illustration of the utility of the model with a real life data set.

Pinar Gunar Karadeniz's and Ilker Ercan's paper ***Examining Tests for Comparing of Survival Curves with Right Censored Data*** addresses the problem faced in survival analysis – in estimating the survival probability of a population and comparing the survival experiences of different groups – given that data obtained from survival studies contains frequently censored observations. The authors examine several tests (Logrank, Gehan-Wilcoxon, Tarone-Ware, Peto-Peto, Modified Peto-Peto and tests belonging to Fleming-Harrington test family with (p, q) values; (1, 0), (0.5, 0.5), (1, 1), (0, 1) ve (0.5, 2) are examined by means of Type I error rate obtained from a simulation study). As a result of the simulation study, Type I error rate of Logrank test is equal or close to the nominal value. The authors conclude that in the situation when survival data were generated from lognormal and inverse Gaussian distribution, Type I error rate of Gehan-Wilcoxon, Tarone-Ware, Peto-Peto, Modified Peto-Peto and Fleming-Harrington (1,0) tests were close to the nominal value.

Piyush Kant Rai, Pareek Sarla, and **Joshi Hemlata,** employ SAE methodology in their article ***Met and Unmet Need for Contraception: Small Area Estimation for Rajasthan State of India,*** focused on a policy important, yet sensitive issue of access to contraception. Using data for 187 towns (of Rajasthan state of India) and the data from the District Level Household Survey (DLHS) 2002-04, and of the Census 2001 of India they estimate the proportion of women having met and unmet need (spacing and limiting fertility) of family planning. They employ Generalized Linear Mixed Model with logit-link function given the binomial nature of variables. The authors believe that the results of their analysis is of relevance to designing and implementing better policies and programmes in this area of possible state intervention.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 1, pp. 191, DOI 10. 21307

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.,
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://stat.gov.pl/en/sit-en/guidelines-for-authors/

STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 193–210, DOI 10. 21307

ESTIMATING SENSITIVE POPULATION PROPORTION USING A COMBINATION OF BINOMIAL AND HYPERGEOMETRIC RANDOMIZED RESPONSES BY DIRECT AND INVERSE MECHANISM

Kajal Dihidar [1], **Manjima Bhattacharya** [2]

ABSTRACT

For various reasons individuals in a sample survey may prefer not to confide to the interviewer the correct answers to certain potentially sensitive questions such as the illegal use of drugs, illegal earning, or incidence of acts of domestic violence, etc. In such cases the individuals may elect not to reply at all or to reply with incorrect answers. The resulting evasive answer bias is ordinarily difficult to assess. The use of a randomized response method for estimating the proportion of individuals possessing those sensitive attributes can potentially eliminate the bias. Following Chaudhuri and Dihidar (2014) and Dihidar (2016), here, as a possible variant, we have made an attempt to estimate the sensitive population proportion using a combination of binomial and hypergeometric randomized responses by direct and inverse mechanism. Along with the traditional simple random sampling, with and without replacement, we consider here sampling of respondents by unequal probabilities. Essential theoretical derivations for unbiased estimator, variance and variance estimators are presented for several sampling schemes. A numerical illustration is performed to make a comparative study of the relative efficiencies of the direct and inverse mechanism..

1. Introduction

Surveys for eliciting information on sensitive or stigmatizing attributes are plagued by the problem of untruthful responses or non-cooperation by respondents, both of which lead to biased estimates. To avoid this evasive answer bias and to preserve the privacy of the respondent, Warner (1965) introduced an innovative technique commonly referred to as randomized response (RR) technique. In his model, a respondent answers 'Yes' or 'No' to either the sensitive question of interest or the complementary question. For example, suppose that we are interested in whether

[1] Corresponding Author. Sampling and Official Statistics Unit, Indian Statistical Institute, Kolkata. E-mail: kajaldihidar@gmail.com

[2] Credit Swiss Company, Mumbai, Maharastra, India. E-mail: bhattacharya.manjima@gmail.com

a person belongs to the sensitive group A. The respondent uses a chance device to select Question 1:'Do you belong to Group A?,' or Question 2: 'Do you belong to Group A^c?,' where A^c is the complement of A, with probabilities, say, $p : (1 - p)$, where $p \neq 0.5$. Thus, if the respondent says 'Yes', the interviewer does not know whether the 'Yes' refers to Question 1 or Question 2. These RRs gathered from a sample of persons chosen by SRSWR provide an unbiased estimator for the sensitive population proportion, say, θ. The variance of this estimator and an unbiased estimator for that variance are also given by Warner (1965).

Later significant developments to Warner's model are made by many researchers. For example, to expect the greater participation rate of the respondents, Horvitz et al. (1967), Greenberg et al. (1969) developed the unrelated question model, where in place of both questions being about a sensitive characteristic, one question is about sensitive, and the other is completely unrelated to the sensitive characteristic, e.g. 'Do you prefer football to cricket?' or 'Is red your favourite colour?'. Boruch (1971) introduced the forced response model where the randomization determines whether a respondent truthfully answers the sensitive question or simply replies with a forced answer, 'yes' or 'no'. The idea behind the forced response design is that a certain proportion of respondents are expected to respond 'yes' or 'no' regardless of their truthful response to the sensitive question, and the design protects the anonymity of respondents' answers. That is, interviewers and researchers can never tell whether observed responses are in reply to the sensitive question. Kuk (1990) proposed a method, where each person selected by simple random sampling with replacement (SRSWR) is given two boxes, say, Box-1 and Box-2. Each of the two boxes are filled with cards of two types, say, red and blue with their mixing proportions being $p_1 : (1 - p_1), 0 < p_1 < 1$ in one box and $p_2 : (1 - p_2), 0 < p_2 < 1$ in the other; $p_1 \neq p_2$ and $p_1 + p_2 \neq 1$. Every selected person is requested to draw cards for a fixed number of times, say, K times independently, either from the first box or from the second, according as whether this person bears characteristic A or not. The respondent is requested to report the number of red cards obtained out of K cards drawn. Based on these RRs an unbiased estimator for θ, variance and variance estimator are obtained.

Likewise, many contributors of this area have enriched the randomized response literature, for instance, Moors (1971), Raghavarao (1978), Eichhorn and Hayre (1983), Chaudhuri and Mukerjee (1987), Mangat and Singh (1990), Mangat (1994), Haung (2004), Kim and Warde (2004), Gjestvang and Singh (2006), Chaudhuri, Bose and Dihidar (2011a, 2011b), Singh and Grewal (2013), Singh and Sedory (2013) among others. We refer to Hedayat and Sinha (1991) as an example of an early text book on sampling which covers this area as a separate chapter (see Chap-

ter 11). For a comprehensive review of the literature on these techniques, we refer to the books by Chaudhuri and Mukerjee (1988) and Chaudhuri (2011) and the various articles in Chaudhuri et al. (2016).

In general, all the approaches of RR techniques assume that the respondents answer truthfully according to the randomized response device. However, the methods are often criticized as being susceptible to cheaters, that is, respondents who do not answer truthfully as directed by the randomizing device. Clark and Desharnais (1998) has shown that by splitting the sample into two groups and assigning each group a different randomization probability, it is possible to detect whether significant cheating is occurring and to estimate its extent while simultaneously protecting the identity of cheaters and those who may have engaged in sensitive activities. In Feth et al. (2015) different forms of cheating is described and it has been shown in detail how to obtain general solution for detecting the extent to which various forms of cheating occurs and extends these analyses with practical hints for the flexible use of these methods. However, although there may be some possibility of cheating occurrences, for the present research work we assume that the respondents are tried to be well convinced to answer truthfully according to the randomizing device and therefore, based on this assumption, below we make an attempt to develop alternative RR techniques for estimating the sensitive proportion.

As stated above, in Kuk's (1990) approach the cards are drawn from either of the two boxes with replacement. A natural question arises - what will happen if the cards are drawn without replacement? In this paper we look into this matter. We know that while drawing n cards with replacement from a box containing two types of cards, the number of the first type of cards obtained follows binomial distribution whereas the number of trials to obtain a fixed number of the first type of cards follows a negative binomial distribution, and drawing the cards without replacement instead of with replacement will result in the hypergeometric and negative hypergeometric distributions respectively for the same. Here, we consider estimating the sensitive population proportion by generating randomized responses using a combination of binomial and hypergeometric distributions in the direct approach as well by using a combination of negative binomial and negative hypergeometric distributions in the inverse approach. Also, keeping in mind that many large scale sample surveys consist of sampling of respondents by unequal probability sampling even without replacement, in this paper we develop unbiased estimators for sensitive population proportion by general sampling schemes instead of only simple random sampling with replacement scheme of respondents. We organize our findings of this research work in the following sections.

In Section 2 below, we present the necessary derivations for generating random-

ized responses using binomial distribution for Box-1 and hypergeometric distribution for Box-2. In Section 3, we present the same by negative binomial and negative hypergeometric distributions respectively. In Section 4, we present the unbiased estimators for θ, variance and variance estimators based on some sampling methods, namely simple random sampling (SRS) both with and without replacement (WR or WOR), and some unequal probability sampling methods, namely probability proportional to size with replacement (PPSWR), Rao, Hartley and Cochran's (1962) and Midzuno's (1952) sampling schemes. We present the numerical illustration in Section 5 for comparison purpose. Finally, we give some concluding remarks in Section 6.

2. Generating RR by direct approach

Let $U = (1,2,\ldots,N)$ denote a finite, identifiable population of N persons labeled 1 to N. Let

$$\begin{aligned} y_i &= 1, \text{ if } i^{th} \text{ person bears the sensitive character, say, } A \\ &= 0, \text{ otherwise.} \end{aligned}$$

We want to estimate the population proportion $\theta = \frac{1}{N}\sum_{i=1}^{N} y_i$, proportion of individuals bearing the sensitive character A.

In our proposed methodology, two randomized response boxes, say Box-1 and Box-2 are used, and each of the two boxes are filled with two types of cards, say 'Red' and 'Blue'; in proportion $p_1 : (1-p_1)$ in Box-1; and in proportion $p_2 : (1-p_2)$ in Box-2, $0 < p_1 \neq p_2 < 1$. Suppose Box-1 contains N_1 total number of balls out of which r_1 are red and the rest are blue, and Box-2 contains N_2 total number of balls out of which r_2 are red and the rest are blue. Hence, $p_1 = r_1/N_1$ and $p_2 = r_2/N_2$. Each respondent in sample s of units, collected with a given probability $p(s) > 0$ according to a given sampling design p, is given two boxes. Every selected person is instructed to use the first box if he bears A, otherwise to use the second box, unnoticed by the interviewer, thus protecting the privacy of the respondent. Additional instruction is also given to the selected respondent to draw cards at random independently for a specified number of times, say, K times, with replacement if he chooses the Box-1 and without replacement if he chooses the Box-2. Every selected person is requested to report finally how many times a 'Red' marked cards are actually drawn out of K trials. Let us denote f_i as the number out of K trials, a 'Red' card happened to be obtained as reported by the person labeled i. Additionally, let E_R, V_R, C_R denote the expectation, variance and covariance operators with

respect to the randomized response generation. Then

$$E_R(f_i) = K[y_i p_1 + (1-y_i)p_2]$$

and

$$V_R(f_i) = K\left[y_i p_1(1-p_1) + (1-y_i)\frac{N_2-K}{N_2-1}p_2(1-p_2)\right]$$

leading to

$$E_R\left[\frac{f_i}{K}\right] = y_i(p_1-p_2)+p_2$$

$$\Rightarrow E_R\left[\frac{\frac{f_i}{K}-p_2}{p_1-p_2}\right] = y_i, \quad \text{on noting that} \quad p_1 \neq p_2.$$

Let $r_i = \dfrac{\frac{f_i}{K}-p_2}{p_1-p_2}$ with $E_R[r_i] = y_i$ and

$$V_R(r_i) = \frac{1}{(p_1-p_2)^2}V_R(\frac{f_i}{K})$$

$$= \frac{1}{(p_1-p_2)^2}\frac{1}{K^2}V_R(f_i)$$

$$= \frac{1}{K(p_1-p_2)^2}\left[y_i p_1(1-p_1) - y_i\left(\frac{N_2-K}{N_2-1}\right)p_2(1-p_2) + \left(\frac{N_2-K}{N_2-1}\right)p_2(1-p_2)\right]$$

$$= ay_i + b$$

$$= V_i, \quad \text{say,}$$

where

$$a = \frac{1}{K(p_1-p_2)^2}\left[p_1(1-p_1) - \left(\frac{N_2-K}{N_2-1}\right)p_2(1-p_2)\right]$$

and

$$b = \frac{p_2(1-p_2)(N_2-K)}{K(p_1-p_2)^2(N_2-1)}.$$

Then, an unbiased estimator of V_i is

$$\hat{V}_R(r_i) = v_i = ar_i + b, \;\; i \in s$$

because $E_R(v_i) = ay_i + b = V_i$.

3. Generating RR by inverse mechanism

Here, every selected respondent is given instruction to use the first box if he bears A, otherwise to use the second box. Additional instruction is given to each respondent that, if he uses Box-1, he should draw cards WR until he gets a specified number, say, t_1 'Red' cards, then he should report the number of required draws to obtain t_1 'Red' cards, say, $G = g$, where G is the random variable denoting the number of draws obtained from Box-1; similarly, if he uses Box-2, he should draw cards WOR until he gets a specified number, say, $t_2 (t_2 < r_2)$ 'Red' cards, then he should report the number of required draws to obtain t_2 'Red' cards, say, $H = h$, where H is the random variable denoting the number of draws obtained from Box-2.

Then G follows a negative binomial distribution with parameters t_1 and p_1 and its probability mass function is given by:

$$P(G = g|t_1, p_1) = \binom{g-1}{t_1-1} p_1^{t_1} (1-p_1)^{g-t_1}; g = t_1, t_1+1, \ldots$$

Similarly, the random variable H follows the negative hypergeometric distribution with parameters N_2, r_2, t_2 and its probability mass function is given by:

$$P(H = h|N_2, r_2, t_2) = \frac{\binom{r_2}{t_2-1}\binom{N_2-r_2}{h-t_2}}{\binom{N_2}{h-1}} \times \frac{r_2 - t_2 + 1}{N_2 - h + 1}; h = t_2, t_2+1, \ldots, (N_2 - r_2 + t_2)$$

At this stage, we may note that it may be possible that the response of an individual with A could be $g < t_2$ or $g > N_2 - r_2 + t_2$, in which case it would be known that the individual has characteristic A, compromising the privacy of the respondent. So, in order to protect the privacy of the respondent, we consider $t_1 = t_2 = t$, say, and ask the respondent to stop drawing when he reaches at the number of draws at $N_2 - r_2 + t$, so that after getting the number of draws from respondent it will not be possible to find out from which box the draws are made. Hence, instead of usual negative binomial distribution, we consider the following truncated negative binomial distribution. We also note that as the number of successes is fixed at t, the number of failures is the random variable, and following Mir (2008) and Shonkwiler (2016) we utilize below the properties of the un-truncated and the truncated random variable. So, if X is the random variable denoting the number of failures preceding t successes, then the probability mass function of the usual negative binomial distribution is given by :

$$P(X = x|t, p_1) = \frac{t}{t+x}\binom{t+x}{x} p_1^t (1-p_1)^x; x = 0, 1, 2, \ldots,$$

for which the expectation and variance are

$$E(X) = t\frac{1-p_1}{p_1} \quad \text{and } V(X) = t\frac{1-p_1}{p_1^2}.$$

Following Shonkwiler (2016), we obtain the expectation and variance of the right truncated negative binomial distributed variable as

$$E(X|X \leq N_2 - r_2) = t\frac{1-p_1}{p_1} - \frac{\frac{t}{p_1}(N_2 - r_2 + 1)h(N_2 - r_2 + 1)}{tP(X \leq N_2 - r_2)} = \mu_0, \quad \text{say,}$$

where $h(N_2 - r_2 + 1)$ is the un-truncated negative binomial probability mass function $P(X = x|t, p_1)$ evaluated at $N_2 - r_2 + 1$, and

$$V(X|X \leq N_2 - r_2)$$
$$= \mu_0 + (N_2 - r_2)\left(\mu_0 - t\frac{1-p_1}{p_1}\right) + \mu_0 t\frac{1-p_1}{p_1}\left(1 + \frac{1}{t}\right) - \mu_0^2 = V_0, \quad \text{say.}$$

Hence,

$$E(G = t + X|G \leq N_2 - r_2 + t) = t + E(X|X \leq N_2 - r_2)$$
$$= t + t\frac{1-p_1}{p_1} - \frac{\frac{t}{p_1}(N_2 - r_2 + 1)h(N_2 - r_2 + 1)}{tP(X \leq N_2 - r_2)} = t + \mu_0 = \mu_1, \quad \text{say.}$$

And

$$V(G = t + X|G \leq N_2 - r_2 + t) = V(X|X \leq N_2 - r_2) = V_0.$$

So, if Z_i denotes the randomized response obtained from i^{th} chosen person, and if G_T denotes the above defined truncated negative binomial distribution, then

$$\begin{aligned} Z_i &= G_T \quad \text{if } i^{th} \text{ person bears } A \\ &= H \quad \text{if } i^{th} \text{ person bears } A^c \,. \end{aligned}$$

On noting the expectation and variance of G_T as derived above and that for the negative hypergeometric distribution $H(N_2, r_2, t)$ as

$$E(H(N_2, r_2, t)) = t\frac{N_2 + 1}{r_2 + 1}, \quad V(H(N_2, r_2, t)) = t\frac{(N_2 - r_2)(N_2 + 1)(r_2 + 1 - t)}{(r_2 + 1)^2(r_2 + 2)},$$

we have

$$\begin{aligned} E_R(Z_i) &= y_i E_R(G_T) + (1-y_i)E_R(H) \\ &= y_i\mu_1 + (1-y_i)t\frac{N_2+1}{r_2+1} \\ &= y_i\left(\mu_1 - \frac{t(N_2+1)}{r_2+1}\right) + t\frac{N_2+1}{r_2+1} \end{aligned}$$

This implies that if

$$\mu_1 - \frac{t(N_2+1)}{(r_2+1)} \neq 0 \quad \text{and} \quad r_i' = \frac{Z_i - t\frac{N_2+1}{r_2+1}}{\mu_1 - \frac{t(N_2+1)}{(r_2+1)}}$$

then

$$E_R(r_i') = y_i.$$

We now note that

$$\begin{aligned} V_i' = V_R(r_i') &= \frac{V_R(Z_i)}{\left[\mu_1 - \frac{(N_2+1)t}{r_2+1}\right]^2} \\ &= cy_i + d, say \end{aligned}$$

where, on writing

$$\phi = \left[\mu_1 - \frac{(N_2+1)t}{r_2+1}\right]^2,$$

$$c = \frac{V_0 - \frac{t(N_2-r_2)(N_2+1)(r_2+1-t)}{(r_2+1)^2(r_2+2)}}{\phi},$$

$$d = \frac{t(N_2-r_2)(N_2+1)(r_2+1-t)}{(r_2+1)^2(r_2+2)\phi}.$$

An unbiased estimator for $V_i' = V_R(r_i')$ is

$$\hat{V}_R(r_i') = v_i' = cr_i' + d, \;\; i \in s,$$

because

$$E_R(v_i') = E_R(cr_i' + d) = cE_R(r_i') + d = cy_i + d = V_i'.$$

4. Comparative efficiencies of the inverse method versus direct one under different sampling schemes

We now present a study of the relative efficiencies of the direct versus inverse RRT as $e = 100\frac{V}{V'}$, where V is the variance of the usual estimator of θ for direct method and V' as the variance of the estimator of θ for the inverse method in different situations. We consider (1) Simple Random Sampling With Replacement (SRSWR) by n draws and (2) Simple Random Sampling Without Replacement (SRSWOR) in n draws, in these two cases the sample means of the transformed randomized responses are used to estimate θ. Also some unequal probability schemes, for example, (3) probability proportional to size with replacement (PPSWR), (4) Rao, Hartley and Cochran's (RHC,1962) sampling scheme and (5) Midzuno's (1952) scheme are used for the estimation of θ and variance of that estimator.

Let us denote E_p, V_p as the expectation and variance operators for design p, then the overall expectation, variance operators denoted by E and V are given as $E = E_p E_R$ and $V = E_p V_R + V_p E_R$. We present below the essential formulation for the estimator considering the direct method for generating RR (as described earlier) and variance and variance estimators for θ based on the various sampling schemes considered in this paper. For the inverse counterpart, r_i, V_i and v_i will of course change in the manners described already, replacing them by r_i', V_i' and v_i' respectively.

4.1. SRSWR in n draws

Let us denote y_k as the y-value for a person chosen on the k^{th} draw $(k = 1, \ldots, n)$ and r_k as the transformed RR generated by the direct method from that person. Then an unbiased estimator for $\theta = \frac{1}{N}\sum_{i=1}^{N} Y_i = \overline{Y}$ is given by $\bar{r} = \frac{1}{n}\sum_{k=1}^{n} r_k$ with $V(\bar{r}) = V_p E_R(\bar{r}) + E_p V_R(\bar{r}) = V_p(\bar{y}) + E_p(\frac{1}{n^2}\sum_{k=1}^{n} V_R(r_k)) = \frac{1}{n}[\theta(1-\theta)] + \frac{1}{Nn}\sum_{i=1}^{N} V_i$, where $V_i = V_R(r_i)$. $V(\bar{r})$ can be unbiasedly estimated by

$$\hat{V}(\bar{r}) = v(\bar{r}) = \frac{1}{n(n-1)}\sum_{k=1}^{n}(r_k - \bar{r})^2.$$

4.2. SRSWOR in n draws

In this case also an unbiased estimator for θ is $\bar{r} = \frac{1}{n}\sum_{i\in s} r_i$ because $E(\bar{r}) = E_p E_R(\bar{r}) = E_p(\bar{y}) = \overline{Y} = \theta$ and $V(\bar{r}) = \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2 + \frac{1}{Nn}\sum_{i=1}^{N} V_i$. $V(\bar{r})$ is unbiasedly estimated by

$$\hat{V}(\bar{r}) = v(\bar{r}) = \frac{N-n}{Nn}\frac{1}{(n-1)}\sum_{i\in s}(r_i - \bar{r})^2 + \frac{1}{Nn}\sum_{i\in s} v_i.$$

4.3. PPSWR in n draws

Let us consider that for unequal probability sample drawing the normed size measures $p_i s$ are from an auxiliary variable z with known $z_i > 0$ for all i having $Z = \sum_{i=1}^{N} z_i$ such that $p_i = \frac{z_i}{Z}$, where $0 < p_i < 1, i = 1,2,\ldots,N$ and $\sum_{i=1}^{N} p_i = 1$. Let us denote p_k as the normed size measure, y_k as the y-value for a person chosen at the k^{th} draw $(k = 1,2,\ldots,n)$. And also let us denote r_k as the transformed RR generated by the direct method for generating randomized response for a person chosen at the k^{th} draw, for $k = 1,2,\ldots,n$. Then, following Hansen and Hurwitz (1943) an unbiased estimator for θ is given by $e_{PPSWR} = \frac{1}{Nn}\sum_{k=1}^{n} \frac{r_k}{p_k}$ with $V(e_{PPSWR}) = V_p E_R(e_{PPSWR}) + E_p V_R(e_{PPSWR}) = V_p(\frac{1}{Nn}\sum_{k=1}^{n} \frac{y_k}{p_k}) + E_p(\frac{1}{N^2 n^2}\sum_{k=1}^{n} \frac{V_R(r_k)}{p_k^2}) = \frac{1}{N^2}\left[\frac{V}{n} + \frac{1}{n}\sum_{i=1}^{N} \frac{V_i}{p_i}\right]$, where

$$V = \sum_{i=1}^{N-1}\sum_{j>i}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$$

and $V_i = V_R(r_i)$. $V(e_{PPSWR})$ can be unbiasedly estimated by

$$\hat{V}(e_{PPWSR}) = v(e_{PPSWR}) = \frac{1}{N^2}\left[\frac{1}{2n^2(n-1)}\sum_{k=1}^{n}\sum_{k'\neq k,k=1}^{n}\left(\frac{r_k}{p_k} - \frac{r_{k'}}{p_{k'}}\right)^2\right].$$

4.4. Rao, Hartley and Cochran's sampling scheme of size n

Rao, Hartley and Cochran's (RHC, 1962) sampling of n persons from N population units consists of making n non-overlapping random groups of the population units of group sizes being N_i, $i = 1,\ldots,n$ such that $\sum_{i=1}^{n} N_i = N$. Let Q_i denote the sum of the normed size measures of the N_i units falling in the i^{th} group. Then, independently from every group only one unit is selected with probability proportional to the normed size measures, thus yielding a sample of required size n by RHC method. For simplicity in notation, we denote the value obtained from the unit selected from i^{th} group as y_i and its normed size measure as p_i. With this notation, the unbiased estimator for θ is

$$e_{RHC} = \frac{1}{N}\sum_{n} r_i \frac{Q_i}{p_i}.$$

Here $\sum_n$ means the sum over the n disjoint groups into which the population U is divided into random groups. Following Rao et al. (1962), the optimal choices of group sizes N_is are given by $N_i = [N/n]$ for $i = 1,2,\ldots,k$ and $N_i = [N/n] + 1$ for $i = k+1, k+2,\ldots,n$, k being determined by solving $\sum_{i=1}^{n} N_i = N$. Following

Chaudhuri and Dihidar (2014) we have

$$V(e_{RHC}) = \frac{1}{N^2}\left[C\sum_{i=1}^{N}\frac{V_i}{p_i} + (1-C)\sum_{i=1}^{N}V_i + C\left(\sum_{i=1}^{N}\frac{y_i^2}{p_i} - Y^2\right)\right], \quad \text{with } C = \frac{\sum_n N_i^2 - N}{N(N-1)}.$$

$V(e_{RHC})$ is unbiasedly estimated by

$$\hat{V}(e_{RHC}) = v(e_{RHC}) = \frac{1}{N^2}\left[D\sum_{n}\sum_{n'}Q_iQ_{i'}\left(\frac{r_i}{p_i} - \frac{r_{i'}}{p_{i'}}\right)^2 + \sum_n v_i\frac{Q_i}{p_i}\right],$$

where

$$D = \frac{\sum_n N_i^2 - N}{N^2 - \sum_n N_i^2}.$$

Here $\sum_n \sum_{n'}$ denotes the sum over non-overlapping pairs of n groups.

4.5. Midzuno's (1952) sampling scheme of n persons

For our illustrative purpose we consider the fifth scheme as Midzuno's (1952) scheme of unequal probability sampling of n units. Sampling by this scheme is done first by drawing one unit by probability proportional to size measure of the auxiliary variable, say, z with $Z = \sum_{i=1}^{N} z_i$. Then, keeping the selected unit aside, the remaining $(n-1)$ units are chosen by simple random sampling without replacement (SRSWOR) out of the remaining $(N-1)$ population units. Under this scheme, the first and second order inclusion probabilities, π_i and $\pi_{ij}, i \neq j$ are as follows.

$$\pi_i = \frac{z_i}{Z} + \frac{Z - z_i}{Z}\frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} = \frac{z_i}{Z}\frac{N-n}{N-1} + \frac{n-1}{N-1} \quad \forall i = 1,2,\ldots,N, \tag{1}$$

and

$$\pi_{ij} = \frac{z_i}{Z}\frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} + \frac{z_j}{Z}\frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} + \frac{Z - z_i - z_j}{Z}\frac{\binom{N-3}{n-3}}{\binom{N-1}{n-1}}$$

$$= \frac{z_i + z_j}{Z}\frac{(N-n)(n-1)}{(N-1)(N-2)} + \frac{(n-1)(n-2)}{(N-1)(N-2)}, \quad \forall i \neq j \in U. \tag{2}$$

For this scheme, $\pi_i\pi_j > \pi_{ij}, \forall i \neq j \in U$. An unbiased estimator for the sensitive population proportion θ is given by Horvitz and Thompson(1952)'s estimator as

$$e_{HT} = \frac{1}{N}\sum_{i\in s}\frac{r_i}{\pi_i}.$$

Utilizing Yates and Grundy (1953)'s form of variance of the HT estimator the variance of e_{HT} is given by

$$V(e_{HT}) = \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1,j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{V_i}{\pi_i}\right].$$

It is unbiasedly estimated by

$$\hat{V}(e_{HT}) = v(e_{HT}) = \frac{1}{N^2}\left[\sum_{i\in s}\sum_{j\in s,j>i}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{v_i}{\pi_i}\right].$$

4.6. Comparison of the efficiencies

It is clear from the variance formulae of the unbiased estimators for θ as given in the above section, that in each case since $E_R(r_i) = y_i = E_R(r_i')$, the V_pE_R term will remain same for both the direct and inverse RRT and the difference will be only in the E_pV_R term. So, to compare the efficiencies of the two methods, we need to examine the relative magnitudes of $V_R(r_i')$ versus $V_R(r_i)$. So, the inverse method will be superior to the direct one if

$$V_R(r_i') \le V_R(r_i), \quad \text{that is if } cy_i + d \le ay_i + b \quad \text{or } y_i \le \frac{b-d}{c-a}, \quad \text{provided } c-a > 0.$$

Maintaining the constraint $c-a>0$, this condition may be equivalently stated by

$$0 \le \theta \le \frac{b-d}{c-a} \quad \text{or } 0 \le \theta \le \frac{b-d}{(b-d) + \left[\frac{V_0}{\phi} - \frac{r_1(N_1-r_1)}{KN_1^2(\frac{r_1}{N_1} - \frac{r_2}{N_2})^2}\right]}.$$

Because of the complicated form of the above inequality, it seems excessively difficult to have any insightful idea about the superiority or otherwise of the estimators of θ realized by survey data obtained through the revised RRT approach rather than the procedure following direct method of randomized response generation. However, we present below a simulation-based numerical evaluation for efficiency comparison.

5. Numerical illustration

For numerical illustration, the values of y_i s and size measures z_i s for $i = 1,2,\ldots,N = 117$ are taken from Chaudhuri and Dihidar (2014) and n is taken throughout as 24. For those data we have $\theta = 0.188$. For illustration of the simulation purpose, we

have used the device parameters as $N_1 = 30, r_1 = 17, t = 9, N_2 = 33, r_2 = 13$ and $K = 12$. We have checked that all the conditions evolved in earlier sections are satisfied with these chosen device parameters. Below we present the relative efficiencies of the inverse mechanism for RR generation versus the direct one based on the various sampling schemes considered in this paper for comparative illustration purpose. We also show below few instances of our findings for the estimated standard error (se), which is the positive square root of $\hat{V}(\hat{\theta})$ and estimated coefficient of variation (cv), which is $cv = 100\frac{se}{\hat{\theta}}$ for various situations.

Table 1. Relative performances of the direct RRT versus the inverse RRT based on SRSWR

	Method 1:Direct RRT			Method 2:Indirect RRT		
serial number	est	se	cv	est	se	cv
1	0.227	0.192	84.582	0.183	0.118	64.481
2	0.289	0.187	64.706	0.372	0.180	48.387
3	0.229	0.127	55.459	0.201	0.105	52.239
4	0.209	0.132	63.158	0.212	0.115	54.245
5	0.278	0.150	53.957	0.292	0.144	49.315

Efficiency = 100(V(Method 1)/V(Method 2)) = 119.36.
Out of 100 cases estimated cv(Method 2) $<$ estimated cv(Method 1) in 60 cases.

Table 2. Relative performances of the direct RRT versus the inverse RRT based on SRSWOR

	Method 1:Direct RRT			Method 2:Indirect RRT		
serial number	est	se	cv	est	se	cv
1	0.248	0.151	60.887	0.252	0.131	51.984
2	0.208	0.151	72.596	0.223	0.118	52.915
3	0.294	0.181	61.565	0.251	0.127	50.598
4	0.234	0.148	63.248	0.241	0.127	52.697
5	0.224	0.143	63.839	0.264	0.121	45.833

Efficiency = 100(V(Method 1)/V(Method 2)) = 120.52.
Out of 100 cases estimated cv(Method 2) $<$ estimated cv(Method 1) in 65 cases.

We observe from Tables 1-5 that the randomized response model considered in this paper can be profitably modified by generating randomized responses by the inverse method having greater efficiencies in comparison to the direct one. Also, from the results obtained from the simulation exercise, it reveals that the inverse RRT has relatively lower values of the estimated coefficient of variations than the ones for

Table 3. Relative performances of the direct RRT versus the inverse RRT based on PPSWR

	Method 1:Direct RRT			Method 2:Indirect RRT		
serial number	est	se	cv	est	se	cv
1	0.214	0.193	90.187	0.174	0.146	83.908
2	0.149	0.128	85.906	0.162	0.131	80.864
3	0.139	0.108	77.698	0.192	0.108	56.250
4	0.289	0.202	69.896	0.164	0.102	62.195
5	0.214	0.131	61.215	0.194	0.106	54.639

Efficiency = 100(V(Method 1)/V(Method 2)) = 103.92.
Out of 100 cases estimated cv(Method 2) $<$ estimated cv(Method 1) in 54 cases.

Table 4. Relative performances of the direct RRT versus the inverse RRT based on Rao, Hartley and Cochran's sampling

	Method 1:Direct RRT			Method 2:Indirect RRT		
serial number	est	se	cv	est	se	cv
1	0.401	0.269	67.082	0.179	0.108	60.335
2	0.328	0.260	79.268	0.249	0.111	44.578
3	0.421	0.266	63.183	0.223	0.121	54.260
4	0.321	0.254	79.128	0.253	0.145	57.312
5	0.317	0.249	78.549	0.173	0.125	72.254

Efficiency = 100(V(Method 1)/V(Method 2)) = 104.89.
Out of 100 cases estimated cv(Method 2) $<$ estimated cv(Method 1) in 56 cases.

Table 5. Relative performances of the direct RRT versus the inverse RRT based on Midzuno's Scheme of sampling

	Method 1:Direct RRT			Method 2:Indirect RRT		
serial number	est	se	cv	est	se	cv
1	0.157	0.119	75.796	0.273	0.157	57.509
2	0.147	0.118	80.272	0.263	0.139	52.852
3	0.217	0.125	57.604	0.203	0.109	53.695
4	0.272	0.124	45.588	0.293	0.119	40.614
5	0.162	0.134	82.716	0.190	0.131	68.947

Efficiency = 100(V(Method 1)/V(Method 2)) = 120.20.
Out of 100 cases estimated cv(Method 2) $<$ estimated cv(Method 1) in 62 cases.

the direct RRT. Therefore, one may use the inverse method profitably in practical survey situation in place of the direct counterpart, in any general sampling scheme, some of which are considered here including the unequal probability sampling of respondents, and hence this is the justification of this research.

6. Concluding Remarks

The study in this paper is relevant for socio-economic surveys where the underlying variable is stigmatizing and qualitative and the objective is to estimate the population proportion of the variable. Here we propose a randomization device which generates the randomized response data from a combination of the binomial and hypergeometric distribution. We also present the alternative procedure of generating randomized responses by combining the inverse of these two distributions. While preparing the randomized response devices, we take care about the privacy of the respondents. In both cases we present the related estimation procedures considering the sample of respondents as chosen by simple random sampling and various unequal probability sampling schemes as well. At the same time, we concentrate on comparing these two approaches. Our numerical simulation-based comparison shows that the inverse approach may be used in practical survey situation in place of the direct approach not only in simple random sampling of respondents, but also profitably in general unequal probability sampling of respondents.

Acknowledgements

The authors are grateful to two anonymous reviewers for giving many insightful and constructive comments and suggestions which led to improvement of the earlier manuscript.

REFERENCES

BORUCH, R. F., (1971). Assuring confidentiality of responses in social research: A note on strategies, The American Sociologist, 6, 308–311.

CHAUDHURI, A., (2011). Randomized response and indirect questioning techniques in surveys, CRC Press, Boca Raton, FL.

CHAUDHURI, A., BOSE, M., DIHIDAR, K., (2011a). Estimation of a sensitive proportion by Warner's randomized response data through inverse sampling. Statistical Papers. 52, 343–354.

CHAUDHURI, A., BOSE, M., DIHIDAR, K., (2011b). Estimating sensitive proportions by Warner's randomized response technique using multiple randomized responses from distinct persons sampled, Statistical Papers, 52, 111–124.

CHAUDHURI, A., CHRISTOFIDES, T. C., RAO, C. R., (2016). Handbook of Statistics 34, Data gathering, analysis and protection of privacy through randomized response techniques. Elsevier, Amsterdam.

CHAUDHURI, A., DIHIDAR, K., (2014). Generating randomized response by inverse mechanism. Model Assisted Statistics and Applications, 9, 343–351.

CHAUDHURI, A., MUKERJEE, R., (1987). Randomized response techniques: a review. Statistica Neerlandica, 41, 27–44.

CHAUDHURI, A., MUKERJEE, R., (1988). Randomized responses: Theory and Techniques, Marcel Dekker, New York, NY.

CLARK, S. J., DESHARNAIS, R. A., (1998). Honest answers to embarrassing: Detecting cheating in the randomized response model. Psychological Methods, 3 (2), 160–168.

DIHIDAR, K., (2016). Estimating sensitive population proportion by generating randomized response following direct and inverse hypergeometric distribution. Handbook of Statistics, 34 : Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits. Edited by Arijit Chaudhuri, Tasos C. Christofides and C.R. Rao. Elsevier, North Holland, Amsterdam, The Netherlands. 427–441.

EICHHORN, B. H., HAYRE, L. S., (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. Journal of Statistical Planning and Inference, 7, 307–316.

FETH, S., FRENGER, M., PITSCH, W., SCHMELZEISEN, P., (2015). Cheater Detection in the Randomized Response technique: derivation, analysis and application. Monsenstein & Vannerdat.
URL: http://universaar.uni-saarland.de/monographien/volltexte/2015/134/

GJESTVANG, C. R., SINGH, S., (2006). A new randomized response model. Journal of the Royal Statistical Society, Series B, 68, 523–530.

GREENBERG, B. G., ABUL-ELA, ABDEL-LATIF, A., SIMMONS, W. R., HORVITZ, D. G., (1969). The unrelated question RR model : theoretical framework. Journal of the American Statistical Association. 64, 520–539.

HANSEN, M. H., HURWITZ, W. N., (1943). On the theory of sampling from finite populations. Annals of Mathematical Statistics, 14 (4), 333–362.

HAUNG, K. (2004). A survey technique for estimating the proportion and sensitivity in a dichotomous finite population. Statistica Neerlandica, 58, 75-82.

HEDAYAT, A.S., SINHA, BIKAS. K., (1991). Design and Inference in Finite Population Sampling. Wiley.

HORVITZ, D. G., THOMPSON, D. J., (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663–685.

HORVITZ, D. G., SHAH, B. V., SIMMONS, W. R., (1967). The unrelated question randomized response model. Social Statistics Section, Proceedings of the American Statistical Association, 65–72.

KIM J., WARDE, W. D., (2004). A mixed randomized response model. Journal of Statistical Planning and Inference, 110, 1–11.

KUK, A. Y. C., (1990). Asking sensitive questions indirectly. Biometrika, 77 (2), 436–438.

MANGAT, N.S., (1994). An improved randomized response strategy. Journal of the Royal Statistical Society, Series B, 56, 93–95.

MANGAT, N. S., SINGH, R., (1990). An alternative randomized response procedure. Biometrika, 77 (2), 439–442.

MIDZUNO, H., (1952). On the sampling system with probabilities proportionate to sum of sizes. Annals of the Institute of Statistical Mathematics, 3, 99–107.

MIR, K. A., (2008). Size-biased generalized negative binomial distribution. Journal of Modern Applied Statistical Methods, 7 (2), 446–453.

MOORS, J. J. A., (1971). Optimization of the Unrelated Question Randomized Response Model. Journal of the American Statistical Association, 66, 627–629.

RAGHAVARAO, D., (1978). On an estimation problem in Warner's randomized response technique. Biometrics, 34, 87–90.

RAO, J. N. K., HARTLEY, H. O., COCHRAN, W. G., (1962). On a simple procedure of unequal probability sampling without replacement. Journal of the Royal Statistical Society, Series B, 24, 482–491.

SHONKWILER, J. S., (2016). Variance of the truncated negative binomial distribution. Journal of Econometrics, 195, 209–210.

SINGH, S., GREWAL, I. S., (2013). Geometric distribution as a randomization device: Implemented to the Kuks model. International Journal of Contemporary Mathematical Sciences, 8 (5), 243–248.

SINGH, S., SEDORY, S.A., (2013). A new randomized response device for sensitive characteristics: An application of negative hypergeometric distribution. Metron, 71, 3–8.

WARNER, S. L., (1965). Randomized response: a survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60, 63–69.

YATES, F. & GRUNDY, P. M., (1953). Selection without replacement from within strata with probability proportional to size. Journal of the American Statistical Association, 75, 206–211.

STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 211–226, DOI 10. 21307

RESIDENCY TESTING. ESTIMATING THE TRUE POPULATION SIZE OF ESTONIA

Ene-Margit Tiit[1]

ABSTRACT

The number of residents or population size is important for all countries. Nowadays in many countries a series of registers have been created, which can be used for assessing the population size. The residency index is a tool created for estimating the under- and over-coverage of population census and calculation of proper population size. For this aim the concept of a sign of life – a binary variable depending on register i, person j and year k has been introduced showing if the person was active in the register in a given year. The weighted sum of signs of life indicates the probability that the person belongs to the set of residents in a given year. To improve the stability of the index a linear combination of the previous value of the index and the sum of signs of life is used. Necessary parameters were estimated using empirical data.

Key words: population size, under-coverage of census, sign of life

Residency and population size. The case of Estonia

The number of residents or population size is important for all countries, but also cities, towns and municipalities. For a long time, the census has been the only way to get information about the number of residents.

From the time when different registers were created and implemented, the situation has changed, as the number of residents can also be counted from registers. Therefore, it seems that in the countries that have a population register or some other good (administrative) registers the population size can be calculated at any time without interviewing the people [1].

In reality, however, the situation is not so simple. Multiple sources of information sometimes complicate the situation because the results may be inconsistent. For instance, in Estonia after the population and household census of 2011 (PHC2011), we had three different numbers of population size:

- The size of census population – 1 294 455.

[1] University of Tartu, Faculty of Science and Technology. E-mail: ene.tiit@ut.ee.

- Population size calculated using registered population events and the population size of census PHC2000 – 1 320 000.
- The number of Estonian residents in the Estonian Population Register (EPR) – 1 365 000.

In some age groups, the difference between various estimates was almost 10%.

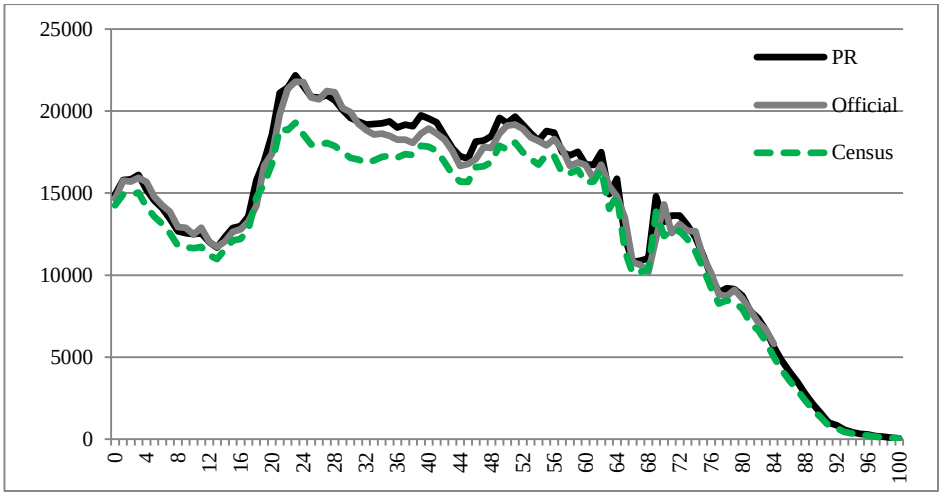


Figure 1. Age-sex distribution of the Estonian population by three different sources

Under-coverage of PHC2011 and estimating the true population size

After PHC2011, it became evident that census population was under-covered. This situation is very common nowadays when the people are very mobile and migration between the countries belonging to the EU and/or the Schengen group is free. It also seemed that probably the population size fixed in EPR was over-covered. In 2012, immediately after PHC2011, the true size of the Estonian population was estimated, see [2—4].

For this aim, the set of people belonging to EPR, but not enumerated in PCH2011 (60 000 persons, about 4.6% of the population) was investigated using the existing system of administrative registers, which includes 12 registers. The activities of these 60 000 problematic persons during the year 2011 were checked in all registers. Thus, 12 binary variables demonstrating their activity in every register were created for each person. Residency was estimated statistically, using these binary variables as explanatory variables for logistical and linear regression. For completing training groups needed in statistical procedures the census data were used. For different age-sex groups different models were created, as the activity in registers depends on age, see Figure 2, where for each person from the training group the sum of all binary variables is presented.

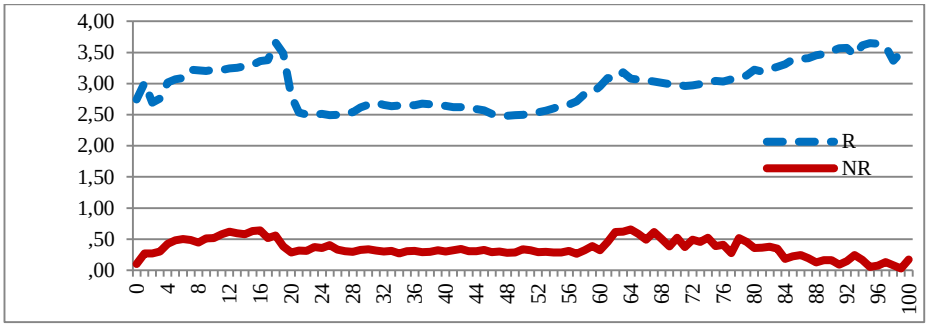


Figure 2. Total activity of residents and non-residents depending on age

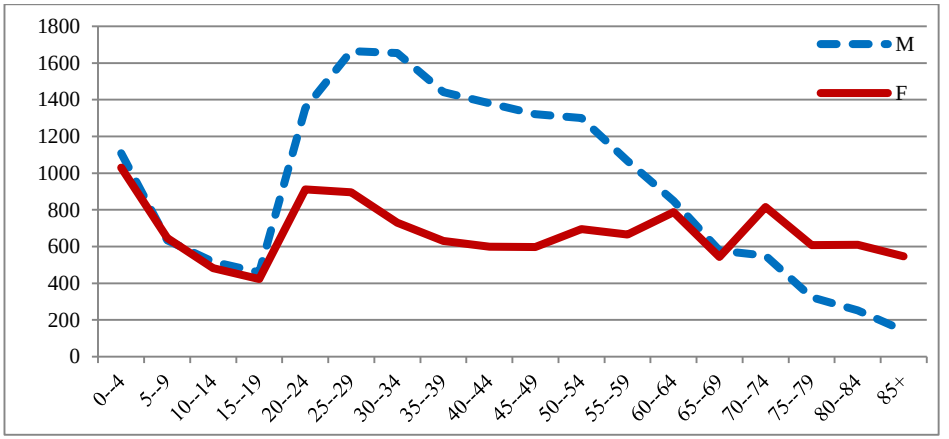


Figure 3. The sex-age distribution of estimated residents (under-coverage of census) added to the census population

About 30 000 persons (2.3% of population) were added to the census population to get the "official" population which Statistics Estonia was using for demographic calculations. In all sex-age groups, the inclusion and exclusion errors of the model were less than 5%. Each added person was identified by his/her recoded ID-code. Such codes allow combining the person's all data from different registers without identifying the person.

There were two main reasons why the census population and the population of Estonian residents in EPR differed. The population of Estonian residents in EPR included non-registered emigrants who had left Estonia during more than 10 years, and hence it was over-covered. The same situation is common in many other transition countries. The census population was under-covered as people are very mobile nowadays, and they also appreciate their privacy very highly, and therefore, are not very keen on participating in censuses. This problem is common in most European countries [5].

Preparation for PHC2020. Estimation of census population

As Estonia has a quite well-functioning system of registers, it has been decided that the following population and household census in 2020/2021 will be organised without personal enumeration and interviewing, but based on registers, as this has already been done in the Nordic countries, Austria, Slovenia and the Netherlands [6]. That means it is necessary to know the census population – the identified set of residents – beforehand. All the census variables about these people will be collected and/or calculated by the data gained from the existing registers.

It is reasonable that the task of estimating the (future) census population relies on current calculation of annual population: every year the population of the previous year is corrected via adding the immigrants and the children born that year and subtracting the emigrants and the people who died that year. While the data of natural increase (births and deaths) is exact nowadays, then migration data might be quite inaccurate due to defective registration that has lasted for decades. Due to errors made in the past, it is complicated to include into the list of immigrants people who have left without registering and returned after some years.

One possibility is to create the model (similar to the model of estimating under-coverage) for residency testing using all the existing registers as explanatory variables. In this case the following problems arise:

- Who are the people to be checked?
- How to get reliable training groups?
- Are all registers equally important, reliable and also independent?

One attempt to solve this problem was made in 2015, three years after the first correction of the "true" population [7, 8]. The solution was the following:

Training groups were formed using information of PHC2011, the current population events and EPR. The population to be checked consisted of the total EPR population (residents and non-residents). All the functioning registers (21 registers and sub-registers) were used to create explanatory variables. Decisions were made using logistic regression. The results were also checked with the help of linear regression and discriminant analysis.

The results of the attempt were positive in the sense that the ideology worked. It was possible to create explanatory variables by registers and check the residency of people using the statistical model. But some problems also occurred.

1. The estimated number of residents (population size) was about 5% lower than the currently calculated official "true" population size;
2. There was a number of indeterminable people who had a different pattern of registers' activities compared with typical residents and typical non-residents; see Figure 4, where on the right, there is the cluster of residents, on the left, the

column of non-residents, and between them, the series of indeterminable subjects.

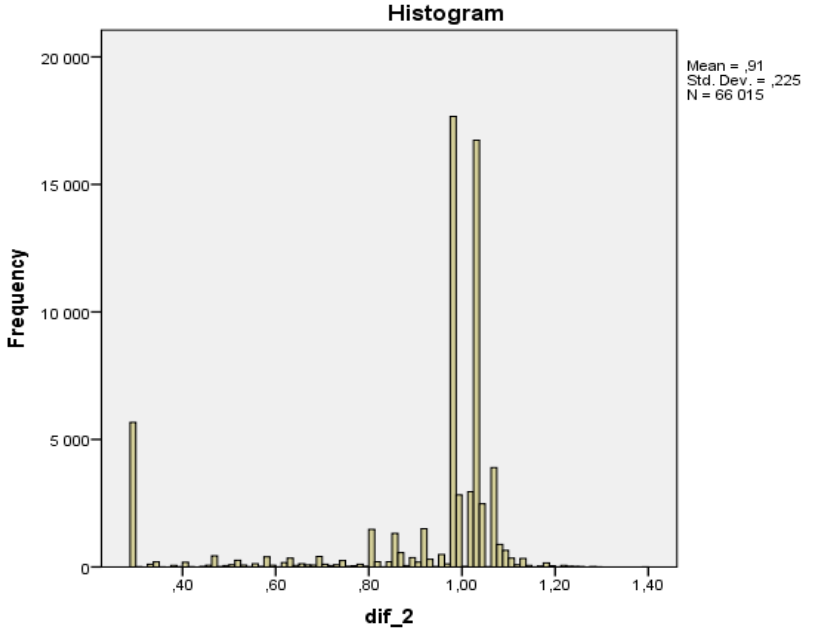


Figure 4. The histogram of distribution of values ascribed to people by the discrimination model

3. There exist a number of persons who are residents but do not show any activity in registers. In the population of PHC2011, there were about 3% of people who had not been active in 2011 in any of the 12 registers used for testing under-coverage. Such people were mainly working-aged men who did not study, did not visit doctors, did not get any social support and, probably, did some non-registered (the so-called black or grey) work.
4. If the strategy is to make a decision separately for each year using new models every time, then it is difficult to gain stability of population (on the personal level). Conversely, when the same model has been used for several years, it will lose optimality.

To avoid the problems listed, a new approach – using the residency index – was invented [9, 10].

Residency index

Principal concepts for formulating the task of residency testing. Fuzzy sets

Time. The whole process of checking residency is connected with one fixed year. This fact follows from the tradition of assessing the population number at the

beginning of the year. The common residency rule used in census statistics also has the lag-time of one year – a person attains (and also loses) the residency of a country during a year. Hence, the residency status of a person in year k is defined by his activities in year k—1.

Persons. Let us have maximum population M, that is a set of persons j, j=1, 2, …, J about whom we have to make the decision if they are residents or non-residents. The content of maximum population changes every year – people will be added to M if they immigrate or are born. The only feasible reason for dropping off from population M is death.

Registers. Let us have a set of registers/sub-registers i, i = 1, 2, …, I. We assume that they are independent. For each person j, register i and year k, a binary variable $B(i,j,k)$ is defined in the following way:

$$B(i,j,k) = \begin{cases} 1, \text{ if person } j \text{ has at least once been active in register } i \text{ in year } k \\ 0 \text{ else.} \end{cases} \quad (1)$$

We say that $B(i,j,k)$ is the ***sign of life*** (SL) of person j in year k. This term has been introduced by Li-Chun Zhang and Dunne [10].

Summarised SLs

Let us form, for every subject j of population M, a linear combination of all binary variables SL reflecting his/her activity in registers in year k,

$$X_j(k) = \sum_{i=1}^{I} a_i B(i,j,k), \quad (2)$$

where a_i are fixed coefficients. The value $X_j(k)$, the generalised sum of SLs, may have a different content depending on the concrete task and values of weights.

1. When k is fixed and all the parameters are equal to 1, then $X_j(k)$ is the ***simple sum of*** SLs.
2. When k is fixed and parameters a_i are calculated as coefficients of the discriminating model (e.g. linear or logarithmic regression), then the value X_j is the prognosis of the residency status of subject j, see [7].

Residency index

To assure the stability of the estimated resident population, the idea of the ***residency index*** has been launched. The main essence of the idea is to predict the residency status for all potential residents every year, using the whole information about them collected during the preceding years.

Assume that for all persons from population M their residency status for year k has been fixed and define the **residency index $R_j(k)$** for them in the following way:

- $R_j(k) = 1$ if person j is resident in year k;
- $R_j(k) = 0$ if person j is not a resident in year k;
- $0 < R_j(k) < 1$ if person's j residency status is not clear.

By definition, the inequalities (3) always hold:

$$0 \leq R_j(k) \leq 1. \tag{3}$$

The residency index can also be treated using the concept of fuzzy sets introduced by Zadeh and Klaua [11, 12]. In this framework, R is for a given year k a membership function from the population M, $R: M \rightarrow [0.1]$ and for each person j the $R_j(k)$ is the grade of membership in year k. $R_j(k)$ can also be interpreted as (subjective) probability that subject j is a resident in year k. To ensure the condition (3), the value of indicator $R_j(k)$ must always be truncated.

In the practical decision process, not only the people having the residency index equal to 1, but also some others belong to the set of residents. That means there exists a ***threshold*** c $(0 < c < 1)$ so that

If $R_j(k) \geq c$, then person j has been considered as resident in year k. (4)

For calculation/assigning the value c, there are some traditional rules in the case when $R_j(k)$ has been defined using statistical models. In general, the value of threshold c must be derived considering rational calculations, and their consonance with empirical data should be tested statistically. In the following, we say that residents having $R_j(k) = 1$ are ***confident residents*** (CR) and non-residents having $R_j(k) = 0$ are ***confident non-residents*** (CNR).

Recalculation of the residency index

The key question in defining the residency index is – how to calculate the residency index for all members of population M for consecutive years? We assume at the beginning of year $k+1$ that most people from population M have the index $R_j(k)$ from the previous year that should be recalculated. The only people who do not have the index are newcomers. All people j who were added to population M during year k will have

$$R_j(k+1) = 1. \tag{5}$$

In the case of immigrants, it is not important if they enter for the first time or have also been residents earlier.

For other persons from *M*, the most logical and simple way is to use the linear combination of two indicators from the previous year – the residency index $R_j(k)$ and GSL $X_j(k)$:

$$R_j\,(k+1) = d\,R_j(k) + gX_j(k). \tag{6}$$

Both the ***stability parameter*** d and the ***SL parameter*** g must satisfy the conditions $0 \le d,g \le 1$. As term $X_j(k)$ is not restricted by 1, there is no need to use the convexity condition $d + g = 1$. To ensure the condition (3) the value $R_j(k+1)$ is truncated:

$$\text{If } R_j(k+1) > 1, \text{ then } R_j(k+1) := 1. \tag{6a}$$

Estimation of parameters

The three parameters – c, d and g defining the residency of persons are connected with the following decisions:

- How long a CR can stay in the status of a resident without any SL? This is the ***exclusion time*** q_1.
- How long does it take for a CNR to obtain residency status on the basis of SLs? This is the ***inclusion time*** q_2.

Parameters *c* and *d* and exclusion time

As regards the first question, we can see that the bigger the value d, the more ***stable*** the process, and the more likely the persons are to retain their residency status for a longer time. Exclusion probability and exclusion time also depend on the value of c: the higher the value c, the more probable it is that a person j will be excluded from the set of residents. We can also say that the higher c, the more ***conservative*** the decision.

Hence, it depends on the combination of values of d and c how long a person will retain the status of a resident not having any SL. If the condition (7) holds, then the CR having no SL retains the status of a resident for q—1 years, and loses it after that.

$$d^q < c \le d^{q-1}. \tag{7}$$

From (7) it follows that the change of residency happens at the moment

$$q = \frac{\ln c}{\ln d}. \tag{8}$$

As the recalculation of index happens at the beginning of the year, the exclusion time is the smallest integer q_1 satisfying the condition

$$q_1 \ge q.$$

Parameter g and inclusion time

To analyse the second question, we have to pay attention to CNRs who are obtaining residency status using SLs. Here, we assume that a_i =1. Then, the condition that the person obtains residency exactly in q years having every year f SLs is the following:

$$fg(1 + d + d^2 + \cdots + d^q) \geq c > fg(1 + d + d^2 + \cdots + d^{q-1}).$$

As the brackets contain the sum of geometric progression, we get the inequalities for parameter g, see (9):

$$\frac{c(1-d)}{f(1-d^{q+1})} \leq g < \frac{c(1-d)}{f(1-d^{q})}. \tag{9}$$

The necessary conditions for positive probability of getting residency status by SLs follow from the sum of geometric progression:

$$g \geq \frac{c(1-d)}{f}. \tag{10}$$

From inequality (9), we get the expression for inclusion time:

$$q = \ln\left\{\frac{fg-c+cd}{fg}\right\}/\ln d - 1 \tag{11}$$

q_2 is the smallest integer fulfilling the condition $q_2 \geq q$, defined by (11).

The special case when $fg \geq c$, then q_2=1 means the person gets the status of R in the first year.

Table 1. Exclusion and inclusion time in the case of a selected set of parameters

Case No	c	d	g	f	q_1	q_2
1	0.7	0.75	0.2	1	2	7
2	0.7	0.75	0.2	2	2	1
3	0.7	0.75	0.25	1	2	4
4	0.7	0.75	0.25	2	2	1
5	0.7	0.8	0.2	1	2	5
6	0.7	0.8	0.2	2	2	1
7	0.7	0.8	0.25	1	2	3
8	0.7	0.8	0.25	2	2	1
9	0.75	0.75	0.2	1	2	9
10	0.75	0.75	0.2	2	2	2
11	0.75	0.75	0.25	1	2	4
12	0.75	0.75	0.25	2	2	1
13	0.75	0.8	0.2	1	2	6
14	0.75	0.8	0.2	2	2	2
15	0.75	0.8	0.25	1	2	4
16	0.75	0.8	0.25	2	2	1

From the table we can see that the exclusion time is quite stable (and does not depend on *f* and *g*), while the number of SLs has a big influence on the inclusion time. For the following example we will choose the parameters c=0.7, d = 0.8 and g= 0.2. That means the exclusion time is 2 years (the CR will be excluded from the set of Rs after two years without SLs), and inclusion time in the case of f= 1 and 2 is correspondingly 5 and 1. Hence, a CNR will gain the residency status in one year if s/he gets 2 SLs and in five years getting one SL each year. The last assertion is true in the case of the simple sum of SLs. When the weighted sum SLs is used, then this calculation is true in average. That means if the SLs have weights that differ from the mean weight, the inclusion time might be somewhat different: using a SL having low weight, the inclusion time might be longer, and in the case of SLs having high weight, the inclusion might be faster.

Example. Using the residency index for the estimation of Estonian population

Maximal population model parameters and initial residency index

The first step in defining the set of residents is fixing the initial maximum population *M*. This population should contain all people who, in principle, might belong to the set of residents. In Estonia this set is the population of (living) people fixed in EPR, being either residents or not but having an Estonian ID-code. Population *M* also includes people who were enumerated in PHC2011 but were not Estonian residents in EPR (the number of such persons was very small). In the future, the size of population *M* may somewhat increase when people also fixed in other registers but not in EPR will be included in *M*.

The model parameters will be defined in the following way: d = 0.8, c = 0.7 and g = 0.2. Then, the model is rather conservative, and it takes several years to obtain residency by SL. For instance, a CNR can get the status of a resident by having one SL only during ten years. To get the status of a resident in one year, the person must have 5 SLs.

We will define the initial residency index R0 in the following way: using the fact that the critical moment of PHC2011 in Estonia was 31.12.2011, it almost coincides with the beginning of the year 2012.

- R0 = 0 for persons who were not Estonian residents in EPR on 1.01.2015 and were not enumerated in PHC2011;
- R0 = 1 for persons who were Estonian residents by EPR on 1.01.2015 and either were enumerated in PHC2011 or were born in 2012–2014.
- R0 = 0.8 for persons who were Estonian residents by EPR on 1.01.2015 but were not enumerated in PHC2011 and were added to the Estonian population using residency criteria in 2012 (the so-called under-coverage).
- R0 = 0.7 for all persons who officially immigrated in 2012–2014.
- R0 = 0.5 for all other persons from population *M*.

Table 2. Distribution of index R0

Population group	CNR	unclear	immigrants	under-coverage of PHC2012	CR	Total
Index	0.0	0.5	0.7	0.8	1.0	
Frequency	78 387	69 111	20106	25 094	1 270 161	1 462 859

Using the threshold 0.7, we have the number of residents 1 315 361, of which 1 270 161 are CRs.

Weighting SL

There are different ways to define coefficients a_i in expression (1). The activity in registers depends on the sex and age of person, see Figure 5. This fact was taken into consideration when preparing the models [3. 7], but it would be too troublesome to use different SLs for different age groups in calculating indexes, and it would cause instability of the processes.

It is common for all age-sex groups that the average activity of CRs and CNRs in registers is quite different, see Figure 8. But there are still differences between registers. Non-residents are more active in registers connected with health services (which are considerably cheap in Estonia) and also in the pension register.

From here it follows that it is reasonable to weight the register-based SLs taking into account the popularity of registers among residents and non-residents. Figure 7 depicts the ratio of activity of residents and non-residents in all the registers.

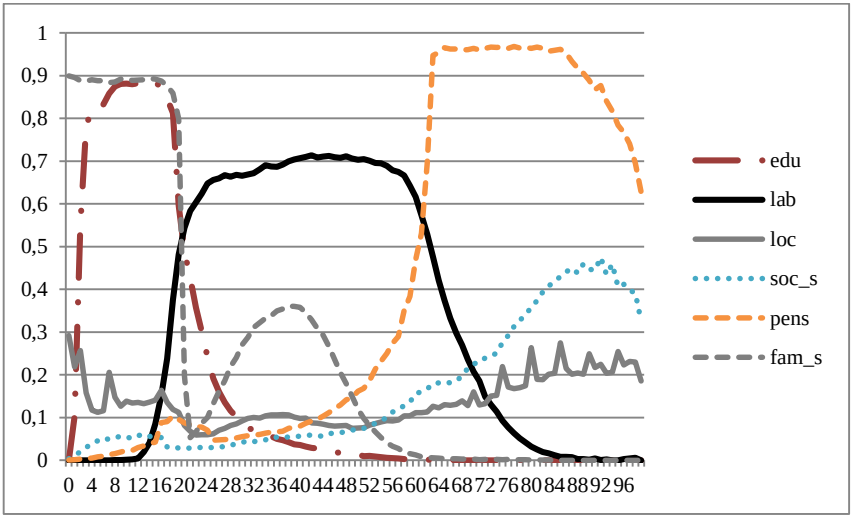


Figure 7. Average activity of all persons from *M* in registers depending on age.

Explanation of the names of variables: edu – learning in an Estonian school; lab – working in an organisation situated in Estonia; loc – getting any support from local administration; soc_s – getting social support or stipend from government; pens – getting pension; fam_s – family support, children benefit.

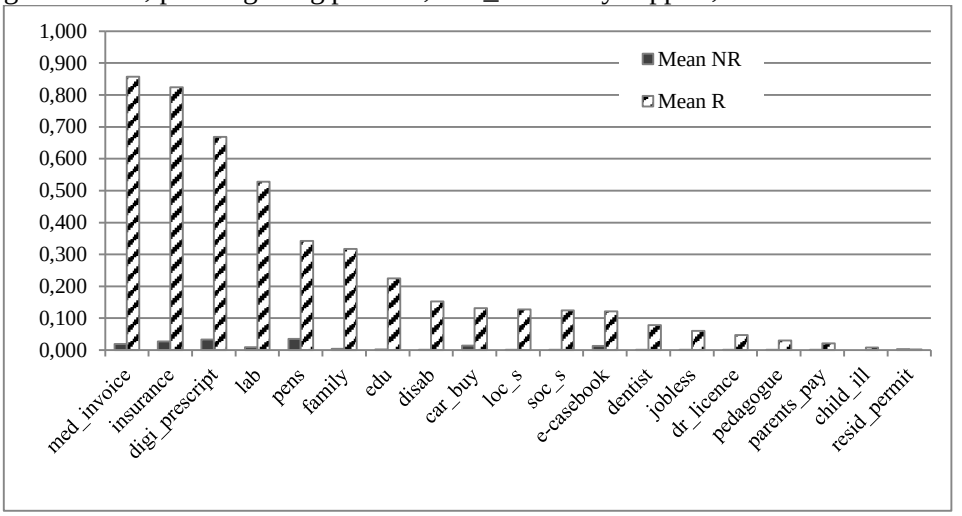


Figure 8. Average activity of CRs and CNRs in different registers.

Explanation of the names of variables: med_invoice – invoice for medical procedure; insurance – health insurance; digi_prescription – digital medical prescription; family – family support; disab – disability fixed by doctor; car_buy – buying or selling a car; e-casebook – fixed event in e-casebook; dentist – visiting dentist; joblessness – active in register of unemployed; de_licence – having doctot's licence; pedagogue – working as pedagogue; parents_pay – getting parents' support; child_ill – document of child's illness; resid_permit – residency permission;

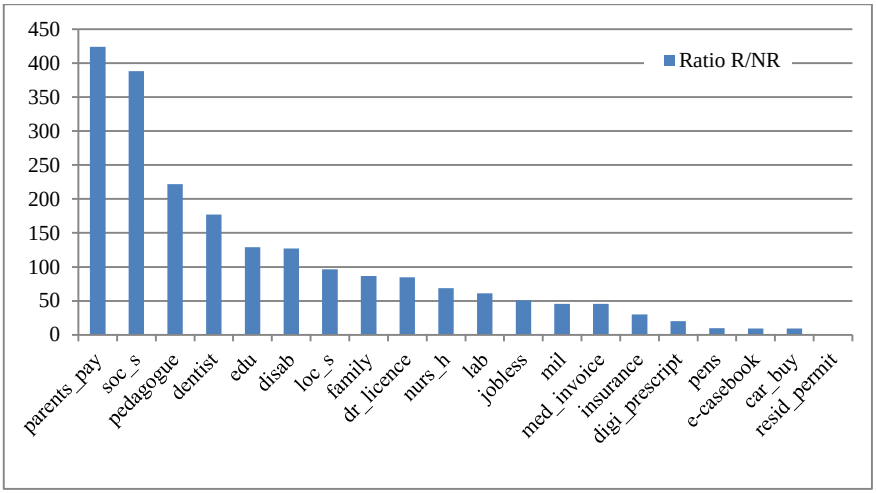


Figure 9. Ratio of average activities of CRs and CNRs in all registers

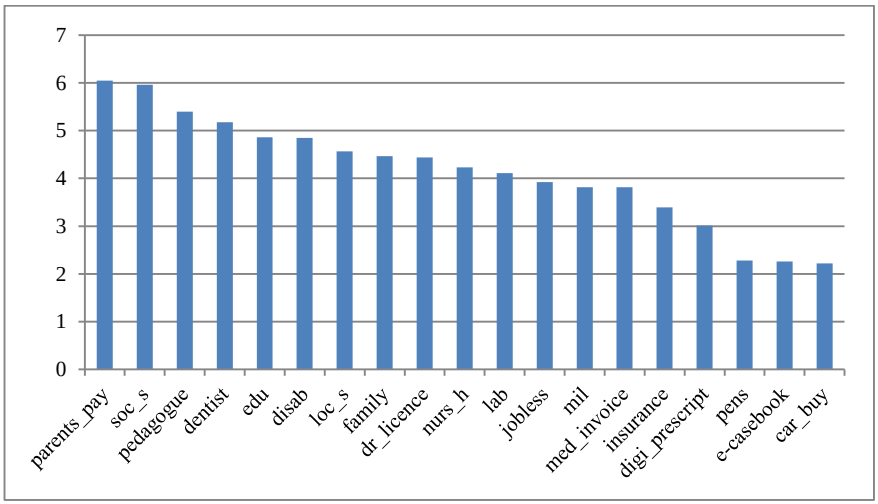


Figure 10. Logarithm of the ratios of activities of CRs and CNRs in all registers.

In defining GSL (see formula (2)), the following options will be used:

1. To take into account simply the SLs, that is to consider all parameters a_i equal to each other.
2. To use weighted SLs where weights are proportional to their ability to differentiate CRs and CNRs measured by ratios of average SLs in CRs and CNRs, see Figure 9.
3. Instead of the ratios of average SLs, the logarithms of the ratios are used as weights, see Figure 10.

In the following we will use and compare these options of calculating the weights. To ensure their comparability, all weights are normed by the mean of simple sum of SLs that equals 4.068, see Figure 11.

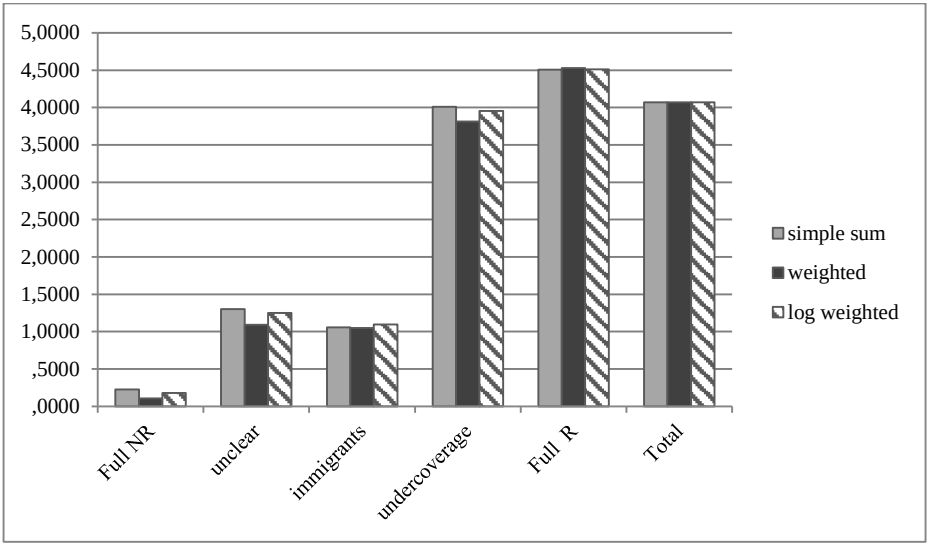


Figure 11. Weighted sum of SLs in different subpopulations of *M* (see Table 1).

It follows that the differences are not big, but the weighted sum is the most sensible.

Calculation of the residence index R1 for the next year

For calculation of the value of the index for the next year, formula (6) was used with parameters c=0.7, d= 0.8 and g = 0.2, and three different sets of weights. Table 2 gives the result of the decision in each case, the number of residents and non-residents, also the percentages.

Table 3. The number of residents and non-residents in the case of different weights for signs of life

	Number of residents	%	Number of non-residents	%	Increase in the number of residents
Simple sum	1 325 258	90.6	137 601	9.4	9 897
Weighted	1 318 385	90.1	144 474	9.9	3 024
Log weighted	1 318 585	90.1	144 274	9.9	3 224

In all the cases, the number of residents has increased by 0.2—0.7 %. In the case of weighted SL, the result is the closest to the supposed real situation.

Conclusion

The residency index is a tool for estimating the residency status of a single person from a population, therefore it can be used for estimating the coverage of a population census and also the population size of a country in an arbitrary year. The residency index uses the so-called signs of life that demonstrate the activity of a person in different registers. The most efficient is the calculation of the residency index in consecutive years, which gives a tool for monitoring the changes of population.

Compared with other types of statistical models, the advantage of this methodology is the possibility of using a large number of different registers, which may have both positive and negative impact on the residency status.

However, the use of indexes might be restricted by the specialities of registers. If the registers are not connected with detailed addresses of living places of persons, then residency indexes cannot be used for describing interior migration and population in small areas.

REFERENCES

Register-based statistics in the Nordic countries - Review of best practices with focus on population and social statistics (United Nations publication). http://unstats.un.org/unsd/censuskb20/KnowledgebaseArticle10220.aspx.

TIIT, E.-M., (2012). 2011. aasta rahva ja eluruumide loenduse alakaetuse hinnang. Eesti Statistika Kvartalikiri, 4/12, Quarterly Bulletin of Statistics Estonia, 4. 12, pp. 110–119.

TIIT, E.-M., MERES, K., VÄHI, M., (2012). Rahvaloenduse üldkogumi hindamine. Eesti Statistika Kvartalikiri, 3, pp. 79–108.

TIIT, E.-M., (2014), 2011. aasta rahva ja eluruumide loendus, Metoodika, 76+19 lk.

Main Results of the UNECE-UNSD Survey on the 2010 Round of Population and Housing Censuses (ECE/CES/GE.41/2009/25).

TIIT, E.-M., (2015). The register-based population and housing census: methodology and developments thereof. Quarterly Bulletin of Statistics Estonia. 3, 15, pp. 42–64.

MAASING, ETHEL, (2015). Eesti alaliste elanike määratlemine registripõhises loenduses**.** http://dspace.utlib.ee/dspace;/handle/10062/47557.

MAASING, ETHEL, (2015). First results in determining permanent residency status in register-based census, banocoss2015/Presentations?preview=#!/preview/149296295/170626623/ Maasing _Abstract.pdf.

TIIT, E.-M., (2015). Residence testing using registers – conceptual and methodological problems, https://wiki.helsinki.fi/display/banocoss2015/Presentations?preview=#!/preview/ 149296295/170626640/Tiit_Abstract.pdf.

LI-CHU, ZHANG, JOHN, DUNNE, Census like population size estimation based on administrative data https://wiki.helsinki.fi/display/banocoss2015/Presentations?preview=/149296295/172987273/CensuslikePopulationSize.pdf.

ZADEH, L. A., (1965). "Fuzzy sets", Information and Control, 8 (3), pp. 338–353.

KLAUA, D., (1965) Über einen Ansatz zur mehrwertigen Mengenlehre. Monatsb, Deutsch, Akad. Wiss, Berlin 7, pp. 859–876.

STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 227–245. DOI 10. 21307

EFFICIENT FAMILY OF RATIO-TYPE ESTIMATORS FOR MEAN ESTIMATION IN SUCCESSIVE SAMPLING ON TWO OCCASIONS USING AUXILIARY INFORMATION

Nazeema T. Beevi[1], **C. Chandran**

ABSTRACT

In this paper, we proposed an efficient family of ratio-type estimators using one auxiliary variable for the estimation of the current population mean under successive sampling scheme. This family of estimators have been studied by Ray and Sahai (1980) under simple random sampling using one auxiliary variable for estimation of the population mean. Using these estimators in successive sampling, the expression for bias and mean squared error of the proposed estimators are obtained up to the first order of approximation. Usual ratio estimator is identified as a particular case of the suggested estimators. Optimum replacement strategy is also discussed. The proposed family of estimators at optimum condition is compared with the simple mean per unit estimator, Cochran (1977) estimator and existing other members of the family. Expressions of optimization are derived and results have been justified through numerical study interpretation.

Key words: auxiliary information, bias, mean square error, optimum replacement.

1. Introduction

Nowadays, it is often seen that sample surveys are not limited to one-time inquiries. A survey carried out on a finite population is subject to change over time if the value of a study character of a finite population is subject to (dynamic) change over time. A survey carried out on a single occasion will provide information about the characteristics of the surveyed population for the given occasion only and can not give any information on the nature or the rate of change of the characteristics over different occasions and the average value of the characteristics over all occasions or the most recent occasion. A part of the sample is retained being replaced for the next occasion (or sampling on successive occasions, which is also called successive sampling or rotation sampling).

The successive method of sampling consists of selecting sample units on different occasions such that some units are common with samples selected on previous

[1]Department of Statistics, University of Calicut, Kerala - 673 635, India.
E-mail: nazeemathaj@gmail.com.

occasions. If sampling on successive occasions is done according to a specific rule, with replacement of sampling units, it is known as successive sampling. Replacement policy was examined by Jessen (1942), who examined the problem of sampling on two occasions, without or with replacement of part of the sample in which what fraction of the sample on the first occasion should be replaced in order that the estimator of $\bar{Y}$ may have maximum precision. Yates (1949) extended Jessen's scheme to the situation where the population mean of a character is estimated on each of $(h > 2)$ occasions from a rotation sample design. These results were generalized by Patterson (1950) and Narain (1953), among others. Rao and Mudhdkar (1967) and Das (1982), used the information collected on the previous occasions for improving the current estimate. Data regarding changing properties of the population of cities or counties and unemployment statistics are collected regularly on a sample basis to estimate the changes from one occasion to the next or to estimate the average over a certain period. An important aspect of continuous surveys is the structure of the sample on each occasion. To meet these requirements, successive sampling provides a strong tool for generating the reliable estimates at different occasions.

Biradar and Singh (2001) proposed an estimator for the population mean on the second of two successive occasions on an auxiliary variate with an unknown population mean. Singh (2005) developed ratio estimators for the population mean on the current occasion using the information for all the units in successive sampling over two occasions. In many situations, information on an auxiliary variate may be readily available on the first as well as the second occasions; for example, tonnage (or seat capacity) of each vehicle or ship is known in survey sampling of transportation and the number of beds in hospital surveys.

Ray and Sahai (1980) have proposed two parameter family of estimators under SRSWOR scheme using auxiliary variate for estimating the population mean $\bar{Y}$ and assuming population mean $\bar{X}$ of auxiliary variate to be known. The objective of this paper is to develop a two parameter family of estimators that estimate the population mean on the current occasion in successive sampling using auxiliary variables on both occasions.

The paper is divided into nine sections. Sample structure and notations have been discussed in section 2 and section 3 respectively. In section 4, the proposed estimators have been formulated. Optimal choice and minimum mean square error of proposed estimator is derived in section 5 and section 6 respectively. In section 7, the optimum replacement policy is discussed and efficiency comparisons are made in section 8. In secection 9, the results have been justified through real data study with interpretation and we give interpretation and conclusion in section 10.

2. Selection of the sample

Consider a finite population $U = (U_1, U_2 \ldots U_N)$ which has been sampled over two occasions. Let x and y be the study variables on the first and second occasions respectively. We further assumed that the information on the auxiliary variable z_1 and z_2, whose population means are known, which is closely related to x and y on the first and second occasions respectively is available on the first as well as on the second occasion. For convenience, it is assumed that the population under consideration is large enough. Allowing SRSWOR (Simple Random Sampling without Replacement) design in each occasions, the successive sampling scheme is proposed as follows:

- We have n units which constitutes the sample on the first occasion. A random sub sample of $n_m = n\lambda$ $(0 < \lambda < 1)$ units is retained (matched) for use on the second occasion, where λ is the fraction of matched sample on the current occasion.

- In the second occasion $n_u = n\mu$ $(= n - n_m)$ $(0 < \mu < 1)$ units are drawn from the remaining $(N-n)$ units of the population, where μ is the fraction of fresh sample on the current occasion.

Therfore the sample size on the second occasion is also n $(= n\lambda + n\mu)$.

3. Description of Notations

We use the following notations in this paper.

$\bar{X}$: The population mean of the study variable on the first occasion.
$\bar{Y}$: The population mean of the study variable on the second occasion.
$\bar{Z}_1$: The population mean of the auxiliary variable on the first occasion.
$\bar{Z}_2$: The population mean of the auxiliary variable on the second occasion.
S_y^2: Population variance of y.
S_x^2: Population variance of x.
$S_{z_1}^2$: Population variance of z_1.
$S_{z_2}^2$: Population variance of z_2.
n_m: The sample size observed on the second occasion and common with the first occasion.
n_u: The sample size of the sample drawn afresh on the second occasion.
n: Total sample size.

$\bar{z}_n$: The sample mean of the auxiliary variable based on n units drawn on the first occasion.

$\bar{z}_{n_u}$: The sample mean of the auxiliary variable based on n_u units drawn on the second occasion.

$\bar{x}_n$: The sample mean of the study variable based on n units drawn on the first occasion.

$\bar{y}_{n_u}$: The sample mean of the study variable based on n_u units drawn afresh on the second occasion.

$\bar{y}_{n_m}$: The sample mean of the study variable based on n_m units common to both occasions and observed on the first occasion.

ρ_{yx}: The correlation coefficient between the variables y on x.

ρ_{xz_1}: The correlation coefficient between the variables x on z_1.

ρ_{yz_2}: The correlation coefficient between the variables y on z_2.

4. Proposed Family of Estimators in Successive sampling

Ray and Sahai (1980) have proposed a two parameter family of estimators under SRSWOR scheme using auxiliary variate for estimating the population mean $\bar{Y}$. The objective of this paper is to develop a two parameter family of estimators that estimate the population mean on the current occasion in successive sampling using auxiliary variables on the both occasions and it is known. To estimate the population mean $\bar{Y}$ on the second occasion, two different estimators are suggested. The first estimator is a family of estimators based on sample of size $n_u(=n\mu)$ drawn afresh on the second occasion given by:

$$t_{n_u}^{(RK\theta)} = \bar{y}_{n_u}\left[\frac{K\bar{Z}_2+\theta\bar{z}_2}{\bar{z}_2+(K+\theta-1)\bar{Z}_2}\right] \tag{1}$$

The second estimator is a family of estimators based on the sample of size $n_m(=n\lambda)$ common with both the occasions defined as

$$t_{n_m}^{(RK\theta)} = \bar{y}_{n_m}\left[\frac{K\bar{x}_n+\theta\bar{x}_{n_m}}{\bar{x}_{n_m}+(K+\theta-1)\bar{x}_n}\right]\left[\frac{K\bar{Z}_2+\theta\bar{z}_2}{\bar{z}_2+(K+\theta-1)\bar{Z}_2}\right], \tag{2}$$

where K is a non-negative constant and $0 \leq \theta \leq 1$. Combining the estimators t_{n_u} and t_{n_m}, we have the final estimator $t_{RK\theta}$ as follows

$$t_{RK\theta} = \psi t_{n_u}+(1-\psi)t_{n_m}, \tag{3}$$

where ψ is an unknown constant to be determined such that $MSE(t_{RK\theta})$ is minimum. We prove theoretically that the estimator is more efficient than the proposed

estimator by Cochran (1977) when no auxiliary variables are used at any occasion. Cochran's classical difference estimator is a widely used estimator to estimate the population mean $\bar{Y}$, in successive sampling. It is given by

$$\bar{y}_2' = \phi_2 \bar{y}_{2u}' + (1-\phi_2)\bar{y}_{2m}',$$

where ϕ_2 is an unknown constant to be determined such that $V(\hat{\bar{Y}})_{opt}$ is minimum and $\bar{y_{2u}}' = \bar{y_{2u}}$ is the sample mean of the unmatched portion of the sample on the second occasion and $\bar{y}_{2m}' = \bar{y}_{2m} + b(\bar{y_1} - \bar{y}_{1m})$ is based on matched portion. The variance for large N of this estimator is

$$Var(\hat{\bar{Y}})_{opt} = [1 + \sqrt{(1-\rho^2)}]\frac{S_y^2}{2n}.$$

Similarly, the variance for large N of the mean per unit estimator is given by

$$Var(\bar{y}) = \frac{S_y^2}{n}.$$

4.1. Properties of $t_{RK\theta}$

Since t_{n_u} and t_{n_m} both are biased estimators of $t_{RK\theta}$, therefore, resulting estimator $t_{RK\theta}$ is also a biased estimator. The bias and MSE up to the first order of approximation are derived as using large sample approximation given below:

$\bar{y}_{n_u} = \bar{Y}(1+e_{\bar{y}_{n_u}}), \quad \bar{y}_{n_m} = \bar{Y}(1+e_{\bar{y}_{nm}}),$
$\bar{x}_{n_m} = \bar{X}(1+e_{\bar{x}_{n_m}}), \quad \bar{x}_n = \bar{X}(1+e_{\bar{x}_n}),$
$\bar{z_1} = \bar{Z}_1(1+e_{\bar{z_1}}), \quad \bar{z}_2 = \bar{Z}_2(1+e_{\bar{z}_2})$

where $e_{\bar{y}_{nu}}, e_{\bar{y}_{nm}}, e_{\bar{x}_{nm}}, e_{\bar{x}_n}, e_{\bar{z_1}}$ and $e_{\bar{z}_2}$ are sampling errors and they are of very small quantities. We assume that

$E(e_{\bar{y}_{nu}}) = E(e_{\bar{y}_{nm}}) = E(e_{\bar{x}_{nm}}) = E(e_{\bar{x}_n}) = E(e_{\bar{z_1}}) = E(e_{\bar{z}_2}) = 0$. Then, for simple random sampling without replacement for both the first and second occasions, we write using the occasion wise operation of expectation as:

$E(e^2_{\bar{y}_{nu}}) = \left(\frac{1}{n_u} - \frac{1}{N}\right) S_y^2, E(e^2_{\bar{y}_{nm}}) = \left(\frac{1}{n_m} - \frac{1}{N}\right) S_y^2,$
$E(e^2_{\bar{x}_{nm}}) = \left(\frac{1}{n_m} - \frac{1}{n}\right) S_x^2, E(e^2_{\bar{x}_n}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_x^2,$
$E(e^2_{\bar{z_1}}) = \left(\frac{1}{m} - \frac{1}{n}\right) S^2_{z_2},$

$$E(e_{\bar{y}_{n_u}} e_{\bar{z}_2}) = \left(\frac{1}{n_u} - \frac{1}{N}\right) S_{yz_2}, \qquad E(e_{\bar{y}_{n_m}} e_{\bar{x}_n}) = \left(\frac{1}{n_m} - \frac{1}{n}\right) S_{yx},$$
$$E(e_{\bar{y}_{n_m}} e_{\bar{x}_{n_m}}) = \left(\frac{1}{n_m} - \frac{1}{n}\right) S_{yx}, \qquad E(e_{\bar{y}_{n_m}} e_{\bar{x}_n}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{yx},$$
$$E(e_{\bar{y}_{n_m}} e_{\bar{z}_1}) = \left(\frac{1}{m} - \frac{1}{n}\right) S_{yz_1}, \qquad E(e_{\bar{x}_{n_m}} e_{\bar{x}_n}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_x^2,$$
$$E(e_{\bar{x}_{n_m}} e_{\bar{z}_1}) = \left(\frac{1}{m} - \frac{1}{n}\right) S_{xz_1}, \qquad E(e_{\bar{x}_n} e_{\bar{z}_1}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{xz_1}$$
$$E(e_{\bar{z}_1} e_{\bar{z}_2}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_z^2.$$

We derive the bias of t_{n_u} and t_{n_m} in lemma 4.3 and lemma 4.5 respectively.

4.2. Lemma

The bias of t_{n_u} denoted by $B(t_{n_u})$ is given by
$B(t_{n_u}) = \bar{Y}\left(\frac{1}{n_u} - \frac{1}{N}\right)\left[Q\left(S_{z_2}^2 (K+\theta)^{-1} - \rho_{yz_2} S_y S_{z2}\right)\right]$,
where $Q = \left[\frac{1-\theta}{K+\theta}\right]$, K is a non-negative constant and $0 \le \theta \le 1$.

Expressing (1) in terms of e's, we have

$$t_{n_u} = \bar{Y}(1 + e_{\bar{y}_{n_u}}) \left[\frac{K\bar{Z}_2 + \theta(1+e_{\bar{z}_2})\bar{Z}_2}{\bar{Z}_2(1+e_{\bar{z}_2}) + (K+\theta-1)\bar{Z}_2}\right],$$

$$t_{n_u} = \bar{Y}(1 + e_{\bar{y}_{n_u}}) \left[1 + \frac{\theta}{a} e_{\bar{z}_2}\right] \left[1 + \frac{e_{\bar{z}_2}}{a}\right]^{-1}, \tag{4}$$

where $a = K + \theta$.
Taking expectation of (4) on both sides, we get

$$E(t_{n_u} - \bar{Y}) = \bar{Y} E\left(e_{\bar{y}_{n_u}} + \frac{\theta}{a} e_{\bar{z}_2} - \frac{e_{\bar{z}_2}}{a} + \frac{e_{\bar{z}_2}^2}{a^2} - \frac{\theta}{a^2} e_{\bar{z}_2}^2 + \frac{\theta}{a} e_{\bar{y}_{n_u}} e_{\bar{z}_2} - \frac{e_{\bar{y}_{n_u}} e_{\bar{z}_2}}{a}\right)$$

$$B(t_{n_u}) = \bar{Y}\left(\frac{1}{n_u} - \frac{1}{N}\right)\left[Q\left(S_{z_2}^2 (K+\theta)^{-1} - \rho_{yz_2} S_y S_{z2}\right)\right],$$

where $Q = \left[\frac{1-\theta}{K+\theta}\right]$, K is a non-negative constant and $0 \le \theta \le 1$.

4.3. Remark

The bias of estimator $B(t_{n_u})$ is the same as the bias of $B(t_{RS})$ proposed by Ray and Sahai (1980).

4.4. Lemma

The bias of t_{n_m} denoted by $B(t_{n_m})$ is given by

$$B(t_{n_m}) = \bar{Y}\left[\left(\frac{1}{n_m}-\frac{1}{n}\right)\left[Q\left(S_x^2(K+\theta)^{-1}-\rho_{yx}S_yS_x\right)\right] + \left(\frac{1}{n}-\frac{1}{N}\right)\left[Q\left(S_{z_1}^2(K+\theta)^{-1}-\rho_{yz_1}S_yS_{z_1}\right)\right]\right],$$

where $Q=\left[\frac{1-\theta}{K+\theta}\right]$, K is a non-negative constant and $0\leq\theta\leq 1$.
Expressing (2) in terms of e's, we have

$$t_{n_m} = \bar{Y}(1+e_{\bar{y}_{n_m}})\left[\frac{K(1+e_{\bar{x}_n})\bar{X}+\theta(1+e_{\bar{x}_{n_m}})\bar{X}}{(1+e_{\bar{x}_{n_m}})\bar{X}+(K+\theta-1)(1+e_{\bar{x}_n})\bar{X}}\right]$$

$$\left[\frac{K\bar{Z}_1+\theta(1+e_{\bar{z}_1})\bar{Z}_1}{(1+e_{\bar{z}_1})\bar{Z}_1+(K+\theta-1)\bar{Z}_1}\right],$$

$$= \bar{Y}(1+e_{\bar{y}_{n_m}})\left[\frac{a+Ke_{\bar{x}_{n_m}}+\theta e_{\bar{x}_{n_m}}}{a+(a-1)e_{\bar{x}_n}+e_{\bar{x}_{n_m}}}\right]\left[\frac{a+\theta e_{\bar{z}_1}}{a+e_{\bar{z}_1}}\right],$$

$$= \bar{Y}(1+e_{\bar{y}_{n_m}})\left[\frac{1+\frac{K}{a}e_{\bar{x}_{n_m}}+\frac{\theta}{a}e_{\bar{x}_{n_m}}}{1+\frac{a-1}{a}e_{\bar{x}_n}+\frac{e_{\bar{x}_{n_m}}}{a}}\right]\left[\frac{1+\frac{\theta}{a}e_{\bar{z}_1}}{1+\frac{e_{\bar{z}_1}}{a}}\right],$$

$$t_{n_m} = \bar{Y}(1+e_{\bar{y}_{n_m}})\left[1+\frac{K}{a}e_{\bar{x}_n}+\frac{\theta}{a}e_{\bar{x}_{n_m}}\right]$$

$$\left[1+\frac{a-1}{a}e_{\bar{x}_n}+\frac{e_{\bar{x}_{n_m}}}{a}\right]^{-1}[1+\frac{\theta}{a}e_{\bar{z}_1}][1+\frac{e_{\bar{z}_1}}{a}]^{-1}. \tag{5}$$

Expanding (5) the right hand side and neglecting higher terms, we get

$$t_{n_m} = \bar{Y}\left[\left(1+\frac{K}{a}e_{\bar{x}_n}+\frac{\theta}{a}e_{\bar{x}_{n_m}}-\frac{e_{\bar{x}_{n_m}}}{a}-\frac{a-1}{a}e_{\bar{x}_n}\right.\right.$$

$$+\frac{e^2_{\bar{x}_{n_m}}}{a^2}+2\frac{a-1}{a^2}e_{\bar{x}_{n_m}}e_{\bar{x}_n}+(\frac{a-1}{a})^2-\frac{K}{a^2}e_{\bar{x}_{n_m}}e_{\bar{x}_n}-K\frac{a-1}{a^2}e^2_{\bar{x}_n}$$

$$-\frac{\theta}{a^2}e^2_{\bar{x}_{n_m}}-\frac{a-1}{a^2}\theta e_{\bar{x}_n}e_{\bar{x}_{n_m}}\Bigg)\left(1+\frac{\theta}{a}e_{\bar{z}_1}-\frac{e_{\bar{z}_1}}{a}+\frac{e^2_{\bar{z}_1}}{a^2}-\frac{\theta}{a^2}e^2_{\bar{z}_1}\right)\Bigg]. \tag{6}$$

Taking expectation of (6) on both sides, we get

$$E(t_{n_m}-\bar{Y})=\bar{Y}\Bigg[\left(\frac{K}{a}e_{\bar{x}_n}+\frac{\theta}{a}e_{\bar{x}_{n_m}}-\frac{e_{\bar{x}_{n_m}}}{a}-\frac{a-1}{a}e_{\bar{x}_n}\right.$$

$$+\frac{e^2_{\bar{x}_{n_m}}}{a^2}+2\frac{a-1}{a^2}e_{\bar{x}_{n_m}}e_{\bar{x}_n}+(\frac{a-1}{a})^2-\frac{K}{a^2}e_{\bar{x}_{n_m}}e_{\bar{x}_n}-K\frac{a-1}{a^2}e^2_{\bar{x}_n}$$

$$\left.-\frac{\theta}{a^2}e^2_{\bar{x}_{n_m}}-\frac{a-1}{a^2}\theta e_{\bar{x}_n}e_{\bar{x}_{n_m}}\Bigg)\left(1+\frac{\theta}{a}e_{\bar{z}_1}-\frac{e_{\bar{z}_1}}{a}+\frac{e^2_{\bar{z}_1}}{a^2}-\frac{\theta}{a^2}e_{\bar{z}_1^2}\right)\right]$$

$$B(t_{n_m})=\bar{Y}\Bigg[\left(\frac{1}{n_m}-\frac{1}{n}\right)\left[Q\left(S^2_x(K+\theta)^{-1}-\rho_{yx}S_yS_x\right)\right]$$
$$+\left(\frac{1}{n}-\frac{1}{N}\right)\left[Q\left(S^2_{z_1}(K+\theta)^{-1}-\rho_{yz_1}S_yS_{z_1}\right)\right]\Bigg],$$

where $Q=\left[\frac{1-\theta}{K+\theta}\right]$. K is a non-negative constant and $0\leq\theta\leq 1$.
Using lemma 4.3 and 4.5, the bias of the estimator $t_{RK\theta}$ can be derived as follows.

4.5. Theorem

The bias of the estimator $t_{RK\theta}$ to the first order approximation is,

$$B(t_{RK\theta}) \quad = \quad \psi B(t_{n_u})+(1-\psi)B(t_{n_m}), \tag{7}$$

where

$$B(t_{n_u})=\bar{Y}\left(\frac{1}{n_u}-\frac{1}{N}\right)\left[Q\left(S^2_{z_2}(K+\theta)^{-1}-\rho_{yz}S_yS_{z_2}\right)\right],$$

and

$$B(t_{n_m}) = \bar{Y}\left[\left(\frac{1}{n_m}-\frac{1}{n}\right)\left[Q\left(S_x^2(K+\theta)^{-1}-\rho_{yx}S_yS_x\right)\right] + \left(\frac{1}{n}-\frac{1}{N}\right)\left[Q\left(S_{z_2}^2(K+\theta)^{-1}-\rho_{yz}S_yS_{z_2}\right)\right]\right],$$

where
$Q = \left[\frac{1-\theta}{K+\theta}\right]$, K is a non-negative constant and $0 \leq \theta \leq 1$.
The bias of the estimator $t_{RK\theta}$ is given by

$$B(t_{RK\theta}) = E(t_{RK\theta}-\bar{Y})$$

$$B(t_{RK\theta}) = \psi E(t_{n_u}-\bar{Y}) + (1-\psi)E(t_{n_m}-\bar{Y}), \tag{8}$$

Using lemmas 4.3 and 4.5 into equation (8), we have the expression for the bias of the estimator $t_{RK\theta}$ as shown in (7).

We derive the MSE of t_{n_u} and t_{n_m} in lemma 4.7 and lemma 4.9 respectively.

4.6. Lemma

The mean square error of t_{n_u} denoted by $M(t_{n_u})$ is given by

$$MSE(t_{n_u}) = \bar{Y}^2\left(\frac{1}{n_u}-\frac{1}{N}\right)\left[S_y^2+Q^2S_{z_2}^2-2Q\rho_{yz}S_yS_{z_2}\right],$$

where $Q\left[=\frac{1-\theta}{K+\theta}\right]$, $0 \leq \theta \leq 1$.
Expressing (2) in terms of e's, we have

$$t_{n_u} = \bar{Y}(1+e_{\bar{y}_{nu}})\left[\frac{K\bar{Z}_2+\theta(1+e_{\bar{z}_2})\bar{Z}_2}{\bar{Z}_2(1+e_{\bar{z}_2})+(K+\theta-1)\bar{Z}_2}\right], \tag{9}$$

where $Q = \left[\frac{1-\theta}{K+\theta}\right]$.

$$t_{n_u} = \bar{Y}(1+e_{\bar{y}_{nu}})\left[K\bar{Z}_2+\theta(1+e_{\bar{z}_2})\bar{Z}_2\right]$$

$$\left[(1+e_{\bar{z}_2})\bar{Z}_2+(K+\theta-1)\bar{Z}_2\right]^{-1},$$

$$= \bar{Y}(1+e_{\bar{y}_{n_u}})[a+\theta e_{\bar{z}_2}][a+e_{\bar{z}_2}]^{-1},$$

$$= \bar{Y}(1+e_{\bar{y}_{n_u}})[a+\theta e_{\bar{z}_2}][a-e_{\bar{z}_2}]. \quad (10)$$

Squaring (10) and expectation, the right hand side and neglecting the terms with power two or greater, we get

$$E(t_{n_u}-\bar{Y})^2 = \bar{Y}^2 E\left[e_{\bar{y}_{n_u}}+\frac{\theta}{a}e_{\bar{z}_2}-\frac{e_{\bar{z}_2}}{a}\right]^2.$$

$$MSE(t_{n_u}) = \bar{Y}^2\left(\frac{1}{n_u}-\frac{1}{N}\right)\left[S_y^2+Q^2S_{z_2}^2-2Q\rho_{yz_2}S_yS_{z_2}\right], \quad (11)$$

where $Q=\left[\frac{1-\theta}{K+\theta}\right]$, $0\le\theta\le 1$.

4.7. Remark

The mean square of estimator $MSE(t_{n_u})$ is the same as the mean square of $MSE(t_{RS})$ proposed by Ray and Sahai (1980).

4.8. Lemma

The mean square error of t_{n_m} denoted by $MSE(t_{n_m})$ is given by

$$MSE(t_{n_m}) = \bar{Y}^2\left[\left(\frac{1}{n_m}-\frac{1}{N}\right)S_y^2+\left(\frac{1}{n_m}-\frac{1}{n}\right)\left[Q^2{S_x}^2-2Q\rho_{yx}S_yS_x)\right]+\left(\frac{1}{n}-\frac{1}{N}\right)\left[Q^2S_{z_1}^2-2Q\rho_{yz_1}S_yS_{z_1}\right]\right],$$

where $Q=\left[\frac{1-\theta}{K+\theta}\right]$, $0\le\theta\le 1$.
Expressing (2) in terms of e's, we have

$$t_{n_m} = \bar{Y}(1+e_{\bar{y}_{nm}})\left[\frac{K(1+e_{\bar{x}_n})\bar{X}+\theta(1+e_{\bar{x}_{n_m}})\bar{X}}{(1+e_{\bar{x}_{n_m}})\bar{X}+(K+\theta-1)(1+e_{\bar{x}_n})\bar{X}}\right]$$

$$\left[\frac{K\bar{Z}+\theta(1+e_{\bar{z}_1})\bar{Z}}{(1+e_{\bar{z}_{1n}})\bar{Z}+(K+\theta-1)\bar{Z}}\right],$$

$$= \bar{Y}(1+e_{\bar{y}_{n_m}})\left[\frac{a+Ke_{\bar{x}_n}+\theta e_{\bar{x}_{n_m}}}{a+(a-1)e_{\bar{x}_n}+e_{\bar{x}_{n_m}}}\right]\left[\frac{a+\theta e_{\bar{z}_1}}{a+e_{\bar{z}_1}}\right],$$

$$= \bar{Y}(1+e_{\bar{y}_{n_m}})\left[\frac{1+\frac{K}{a}e_{\bar{x}_{n_m}}+\frac{\theta}{a}e_{\bar{x}_{n_m}}}{1+\frac{a-1}{a}e_{\bar{x}_n}+\frac{e_{\bar{x}_{n_m}}}{a}}\right]\left[\frac{1+\frac{\theta}{a}e_{\bar{z}_1}}{1+\frac{e_{\bar{z}_1}}{a}}\right],$$

where
$Q = \frac{1-\theta}{K+\theta}$.

$$t_{n_m} = \bar{Y}(1+e_{\bar{y}_{n_m}})\left[1+\frac{K}{a}e_{\bar{x}_n}+\frac{\theta}{a}e_{\bar{x}_{n_m}}\right]$$

$$\left[1-\frac{a-1}{a}e_{\bar{x}_n}-\frac{e_{\bar{x}_{n_m}}}{a}\right][1+\frac{\theta}{a}e_{\bar{z}_1}][1-\frac{e_{\bar{z}_1}}{a}]. \tag{12}$$

Expanding (12) to the right hand side and neglecting the higher terms, we get

$$t_{n_m} = \bar{Y}\left[1+e_{\bar{y}_{n_m}}+Qe_{\bar{x}_n}-Qe_{\bar{x}_{n_m}}-Qe_{\bar{y}_n}\right]. \tag{13}$$

Squaring and taking expectation (13), we get *MSE* of the estimator t_{n_m} up to first order of approximation as,

$$E(t_{n_m}-\bar{Y})^2 = \bar{Y}E\left[1+e_{\bar{y}_{n_m}}+Qe_{\bar{x}_n}-Qe_{\bar{x}_{n_m}}-Qe_{\bar{z}_n}\right]^2, \tag{14}$$

$$= \bar{Y}^2E[e^2_{\bar{y}_{n_m}}+Q^2e_{\bar{x}_n}+Q^2e^2_{\bar{x}_{n_m}}+Q^2e^2_{\bar{z}_1}+2Qe_{\bar{y}_{n_m}}e_{\bar{x}_n}$$

$$-2Qe_{\bar{y}_{n_m}}e_{\bar{x}_{n_m}}-2Qe_{\bar{y}_{n_m}}e_{\bar{z}_n}-2Q^2e_{\bar{x}_n}e_{\bar{y}_{n_m}}-2Q^2e_{\bar{x}_n}e_{\bar{z}_1}+2Q^2e_{\bar{x}_{n_m}}e_{\bar{z}_1}]$$

$$MSE(t_{n_m}) = \bar{Y}^2\left[\left(\frac{1}{n_m}-\frac{1}{N}\right)S_y^2+\left(\frac{1}{n_m}-\frac{1}{n}\right)\left[Q^2{S_x}^2-2Q\rho_{yx}S_yS_x)\right]\right.$$

$$\left.+\left(\frac{1}{n}-\frac{1}{N}\right)\left[Q^2S^2_{z_1}-2Q\rho_{yz}S_yS_{z_1}\right]\right], \tag{15}$$

where $Q=\left[\frac{1-\theta}{K+\theta}\right]$, $0\le\theta\le 1$.
Using lemma 4.7 and lemma 4.9, we derive the MSE of $t_{RK\theta}$.

4.9. Theorem

The estimators of t_{n_u} and t_{n_m} are based on two independent samples of sizes n_u and n_m respectively. The mean square error of the estimator $t_{RK\theta}$ to the first order approximation is,

$$MSE(t_{RK\theta}) = \psi^2 MSE(t_{n_u}) + (1-\psi)^2 MSE(t_{n_m}), \tag{16}$$

where

$$MSE(t_{n_u}) = \bar{Y}^2 \left(\frac{1}{n_u} - \frac{1}{N}\right) \left[S_y^2 + Q^2 S_{z_2}^2 - 2Q\rho_{yz_2} S_y S_{z_2}\right],$$

and

$$MSE(t_{n_m}) = \bar{Y}^2 \left[\left(\frac{1}{n_m} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n_m} - \frac{1}{n}\right) \left[Q^2 {S_x}^2 - 2Q\rho_{yx} S_y S_x)\right] + \right.$$

$$\left. \left(\frac{1}{n} - \frac{1}{N}\right) \left[Q^2 S_{z_1}^2 - 2Q\rho_{yz_1} S_y S_{z_1}\right] \right],$$

where $Q = \left[\frac{1-\theta}{K+\theta}\right]$, $0 \le \theta \le 1$.
The mean square error of the estimator $t_{RK\theta}$ is given by

$$MSE(t_{RK\theta}) = E(t_{RK\theta} - \bar{Y})^2$$

$$MSE(t_{RK\theta}) = E[\psi(t_{n_u} - \bar{Y}) + (1-\psi)(t_{n_m} - \bar{Y})]^2, \tag{17}$$

$$MSE(t_{RK\theta}) = \psi^2 MSE(t_{n_u}) + (1-\psi)^2 MSE(t_{n_m}),$$

using lemma 4.7 and lemma 4.9 into the equation (17), we have the expression for the *MSE* of the estimator $t_{RK\theta}$ as shown in (16).

5. Optimal Choice for Proposed Family of Ratio-Type Estimators $t_{RK\theta}$

The mean square error of the $t_{RK\theta}$ is a function of Q and also the function of K and θ. We minimize for K and θ. We have $\frac{\partial(MSE(t_{RK\theta}))}{\partial K} = 0$ and $\frac{\partial(MSE(t_{RK\theta}))}{\partial \theta} = 0$. This gives $Q = -\frac{U}{V}$, assuming $\frac{\partial Q}{\partial K} \neq 0$ and $\frac{\partial Q}{\partial \theta} \neq 0$. The differential equations are given

below:

$$\frac{\partial(MSE(t_{KR\theta}))}{\partial K} = 0 \Rightarrow \psi^2 \frac{\partial MSE(t_{n_u})}{\partial K} + (1-\psi)^2 \frac{\partial M(t_{n_m})}{\partial K} = 0, \tag{18}$$

$$\frac{\partial(MSE(t_{KR\theta}))}{\partial \theta} = 0 \Rightarrow \psi^2 \frac{\partial MSE(t_{n_u})}{\partial \theta} + (1-\psi)^2 \frac{\partial MSE(t_{n_m})}{\partial \theta} = 0, \tag{19}$$

where ψ is an unknown constant. From (18) and (19), differentiate with respect to K and θ, we get

$$\hat{K} = \frac{1-\theta(1+\rho_{yz_1}\frac{S_y}{S_{z_1}})}{\rho_{yz_1}\frac{S_y}{S_{z_1}}}, \quad \theta \quad fixed.$$

$$\hat{\theta} = \frac{1-K\rho_{yz_1}\frac{S_y}{S_{z_1}}}{1+\rho_{yz_1}\frac{S_y}{S_{z_1}}}, \qquad K \quad fixed.$$

6. Minimum Mean Square Error of $t_{RK\theta}$

To get the optimum value of ψ, we partially differentiate the expression (16) with respect to ψ, and put it equal to zero and we get

$$\psi_{opt} = \frac{MSE\,(t_{n_m})}{MSE\,(t_{n_u}) + MSE\,(t_{n_m})\,.} \tag{20}$$

Substituting the values of $MSE\,(t_{n_u})_{opt}$ and $MSE\,(t_{n_m})_{opt}$ from (11) and (15) in (20), we get

$$\begin{aligned} \psi_{opt} &= \frac{(p_1+\mu p_2)}{(p_1+\mu^2 p_2)} \\ &= \frac{\mu[(p_1+\mu p_2)]}{(p_1+\mu^2 p_2).} \end{aligned}$$

Substitution of ψ_{opt} from (20) into (16) gives optimum value of MSE of $t_{RK\theta}$ as:

$$MSE(t_{RK\theta})_{opt} = \frac{MSE\,(t_{n_m})\,MSE\,(t_{n_u})}{MSE\,(t_{n_u}) + MSE\,(t_{n_m})}. \tag{21}$$

Substituting the values of $MSE(t_{n_u})$ and $MSE(t_{n_m})$ from (11) and (15) in (21), we get

$$MSE(t_{RK\theta})_{opt} = \frac{1}{n}\left[\frac{p_1^2 + \mu p_1 p_2}{p_1 + \mu^2 p_2}\right], \tag{22}$$

where $p_1 = q_1 + q_3$, $p_2 = q_2 - q_3$, $q_1 = {S_y}^2$, $q_2 = Q^2 - 2Q\rho_{yx}S_yS_x$, $q_3 = Q^2 - 2Q\rho_{yz_1}S_yS_{z_1}$, and $\mu = \frac{n_u}{n}$.

7. Replacement Policy of $t_{RK\theta}$

In order to estimate $t_{RK\theta}$ with maximum precision an optimum value of μ should be determined so as to know what fraction of the sample on the first occasion should be replaced. We minimize, $MSE(t_{RK\theta})_{opt}$ in (22) with respect to μ, the optimum value of μ is obtained as,

$$\hat{\mu} = \frac{-p_1 \pm \sqrt{p_1^2 + p_1 p_2}}{p_2}, \tag{23}$$

where $p_1 = q_1 + q_3$, $p_2 = q_2 - q_3$. From (23) it is obvious that for $\rho_{yz_1} \neq \rho_{yx}$ two values of $\hat{\mu}$ are possible, therefore to choose the value of $\hat{\mu}$, it should be remembered that $0 \leq \hat{\mu} \leq 1$. All other values of $\hat{\mu}$ are inadmissible. If both the real values of $\hat{\mu}$ are admissible, the lowest one will be the best choice as it reduces the total cost of the survey. Substituting the value of $\hat{\mu}$ from (23) in (22), we get

$$MSE(t_{\hat{R}\hat{K}\theta})_{opt} = \frac{1}{n}\left[\frac{p_1^2 + \hat{\mu} p_1 p_2}{p_1 + \hat{\mu^2} p_2}\right], \tag{24}$$

μ is the fraction of fresh sample drawn on the current occasion. To estimate the population mean on each occasion 1 is a better choice of μ. However, to estimate the change in the mean from one occasion to the other, μ should be 0.

8. Efficiency Comparisons

In this section, to compare $t_{RK\theta}$ with respect to $\bar{y}$, sample mean of y, when a sample units are selected at second occasion without any matched portion. Since $\bar{y}$ and is unbiased estimators of $\bar{Y}$, its variance for large N is respectively given by

$$Var(\bar{y}) = \frac{S_y^2}{n},$$

The variance of Cochran's (1977) estimator is

$$Var(\hat{\bar{Y}})_{opt} = [1+\sqrt{1-\rho_{yx}^2}]\frac{S_y^2}{2n},$$

the mean square error of ratio estimator is

$$MSE(\bar{y}_R)_{opt} = \frac{1}{n}\bar{Y}^2[S_y^2+S_x^2-2\rho_{yx}S_yS_x].$$

and the variance of Biradar and Singh's estimator is

$$V(\hat{\bar{Y}}_2)_{opt} = \frac{S_{y_2}^2}{n_1}[\frac{(1-\lambda\rho_{x_2,y_2}^2)(1-\mu R_{y_2.y_1x_1x_2}^2)}{1-\mu^2R_{y_2.y_1x_1x_2}^2-\lambda^2\rho_{x_2.y_2}^2}].$$

The relative efficiencies of $t_{RK\theta}$ with respect to $\bar{y}$, $\hat{\bar{Y}}$ and $\bar{y_R}$ are given by

$$R_1 = \frac{Var(\bar{y})}{MSE(t_{RK\theta})_{opt}} \times 100,$$

$$R_2 = \frac{Var(\hat{\bar{Y}})_{opt}}{MSE(t_{RK\theta})_{opt}} \times 100$$

$$R_3 = \frac{MSE(\bar{y}_R)_{opt}}{MSE(t_{RK\theta})_{opt}} \times 100,$$

and

$$R_4 = \frac{V(\hat{\bar{Y}}_2)_{opt}}{MSE(t_{RK\theta})_{opt}} \times 100$$

the estimator $t_{RK\theta}$ (at optimal conditions) is also compared with respect to the estimator $V(\bar{y})$, where

$$MSE(t_{RK\theta})_{opt} = \frac{1}{n}\left[\frac{p_1^2+\hat{\mu}p_1p_2}{p_1+\hat{\mu}^2p_2}\right] \tag{25}$$

and

$$\hat{\mu} = \frac{-p_1 \pm \sqrt{p_1^2+p_1p_2}}{p_2}, \tag{26}$$

where $p_1 = q_1+q_3$, $p_2 = q_2-q_3$.

8.1. Numerical Illustration

The results obtained in previous sections are examined with the help of two natural population sets of data.

Population Source:1 [Free access to the data from the Statistical Abstracts of the United States ($https://www.census.gov/history/www/reference/publications/statistical_abstracts.html$).]

The first population presents $N = 41$ states of the United States. Let y_i (study variable on the second occasion) be the corn production (in million bushels) during 2009 in the i^{th} state of the United States, x_i (study variable on the first occasion) be the corn production (in million bushels) during 2008 in the i^{th} state of the United States and z_i (auxiliary variable) be the corn production (in million bushels) during 2006 in the i^{th} state of the United States. Consider the population, $N = 41$ states of sample size $n = 30$ states were selected using SRSWOR in the year 2008. From each one of the sample drawn (matched samples), $n_m = 25$ states were retained and fresh sample (unmatched samples) $n_u = 5$ states were selected from the remaining $N - n = 41 - 30 = 11$ states using SRSWOR in the year 2009. The results in Table 6.9. show the comparison of the suggested estimator $t_{RK\theta}$ with respect to the estimators $\bar{y}$, $\hat{\bar{Y}}$ and $\bar{y}_R$ respectively for different selection of matched and unmatched samples. For convenience, the different selections of n_m and n_u are considered as different sets in the population, which are given below:

Population Source 2 [Free access data from the data.gov.in.(https://data.gov.in/)]

The population presents, $N = 28$ states in India. Let y_i (study variable on the second occasion) be the infant mortality rate (per 1000 live births) during 2011 in the i^{th} state in India, x_i (study variable on the first occasion) be the infant mortality rate (per 1000 live births) during 2010 in the i^{th} state in India and z_i (auxiliary variable) be the infant mortality rate (per 1000 live births) during 2009 in the i^{th} state in India. Consider the population, $N = 28$ states of sample size $n = 20$ states were selected using SRSWOR in the year 2010. From each one of the sample drawn (matched samples), $n_m = 17$ states were retained and fresh sample (unmatched samples) $n_u = 3$ states were selected from remaining $N - n = 28 - 20 = 8$ states using SRSWOR in the year 2011. The results in Table 6.10. show the comparison of the suggested estimator $t_{RK\theta}$ with respect to the estimators $\bar{y}$, $\hat{\bar{Y}}$ and $\bar{y}_R$ respectively for different selection of matched and unmatched samples. For convenience, the different selections of n_m and n_u are considered as different sets in the population, which are given below:

	$Bias(t_{R\hat{K}\hat{\theta}})$	MSE/Variance	Relative Efficiencies of $t_{RK\theta}$ with respect to $\bar{y}$, $\hat{\bar{Y}}$ $\bar{y}_R$ and$\hat{\bar{Y}}_2$
$n=30, n_m=25, n_u=5,$ $\hat{K}=1.5, \hat{\theta}=0.33$	1.0031	$MSE(t_{R\hat{K}\hat{\theta}})$=0.1197 $V(\bar{y})$= 35.03 $V(\hat{\bar{Y}})_{opt}$= 33.97 $MSE(\bar{y}_R)_{opt}$= 32.12 $V(\hat{\bar{Y}}_2)_{opt}$ = 15.02	 R_1 =292.64 R_2 =283.79 R_3 =268.33 R_4 =125.73
$n=30, n_m=20, n_u=10$ $\hat{K}=2.31, \hat{\theta}=0.17$	3.7410	$MSE(t_{R\hat{K}\hat{\theta}})$=0.2489 $V(\bar{y})$= 39.61 $V(\hat{\bar{Y}})_{opt}$= 37.33 $MSE(\bar{y}_R)_{opt}$= 36.19 $V(\hat{\bar{Y}}_2)_{opt}$ = 17.14	 R_1 =159.14 R_2 =148.77 R_3 =145.39 R_4 =68.89
$n=30, n_m=17, n_u=13$ $\hat{K}=2.84, \hat{\theta}=0.12$	3.9971	$MSE(t_{R\hat{K}\hat{\theta}})$=0.3141 $V(\bar{y})$= 40.73 $V(\hat{\bar{Y}})_{opt}$= 39.88 $MSE(\bar{y}_R)_{opt}$= 38.64 $V(\hat{\bar{Y}}_2)_{opt}$ = 19.25	 R_1 =129.67 R_3 =126.96 R_3 =123.01 R_4 =61.29

Table 8.1. The suggested estimators, $t_{RK\theta}$ is compared to $\bar{y}$, $\hat{\bar{Y}}$, $\bar{y}_R$ and $\hat{\bar{Y}}_2$ for the population by using Real data results

9. Interpretations of Empirical Results of $t_{RK\theta}$

The following conclusions can be made from Table 8.1.

1. The values of R_1, R_2, R_3 and $R_4$are increasing while K is increasing with the decreasing values of θ for fixed values of sample size n.

2. The values of R_1, R_2, R_3 and μ are increasing with increasing values of ρ_{yz}. The behaviour indicates that an agreement with the Sukhatme et.al (1984) results, which explains that more the value of ρ_{yx}, more the fraction of fresh sample is required at the second (current) occasion.

3. For the fixed values of ρ_{yx}, ρ_{yz}, $\hat{K}$ and $\hat{\theta}$ there is an appreciable gain in the performance of the proposed estimator $t_{RK\theta}$ over $\bar{y}$ and $\hat{\bar{Y}}$ with the increasing value of μ.

10. Conclusions

Tables 8.1. clearly indicates that the proposed estimators is more efficient than the simple arithmetic mean estimator, Cochran (1977) estimator and ratio estimator. The following conclusion can be drawn from Tables 8.1. For fixed K, θ, the values of R_1, R_2, R_3 and R_4 are increasing. This phenomenon indicates that a smaller fresh sample at current occasion is required if a highly positively correlated auxiliary characters is available. This means the performance of the precision of the estimates also reduces the cost of the survey.

REFERENCES

BIRADAR, R. S., SINGH, H. P., (2001). Successive sampling using Auxiliary Information on both the Occasions, Calcutta Statistical Association Bulletin, 51, 243–251.

COCHRAN, W.G., (1977). Sampling Techniques. 3^{rd} edition.

DAS, K., (1982). Estimation of population ratio on two occasions. Jour. Ind. Soc. Agri. Stat. 34 (2), 1–9.

JESSEN, R. J., (1942). Statistical Investigation of a sample survey for obtaining farm facts. Iowa Agricultural Experiment Statistical Research Bulletin, 304.

NARAIN, R. D., (1953). On the recurrence formula in sampling on successive occasions. Jour. Ind. Soc. Agri. Stat., 5, 96–99.

PATTERSON, H. D., (1950). Sampling on successive occasions with partial replacement of units. Jour. Roy. Stat. Soc., $B(12)$, 241–255.

RAO, J. N. K., MUDHOLKAR, G. S., (1967). Generalized multivariate estimators for the mean of finite population parameters. Jour. Amer. Stat. Asso., 62, 1008–1012.

RAY, S. K., SAHAI, A., (1980). Efficient families of ratio and product type estimators. Biometrika, 67(1), 211–215.

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

SEN, A. R., (1973). Theory and application of sampling on repeated occasions. Jour. Amer. Stat. Asso., 59, 492–509.

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

YATES, F., (1949). Sampling Methods for Censuses and Surveys. Charles Griffln and Co., London.

STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 247–270, DOI 10. 21307

SEQUENTIAL DATA WEIGHTING PROCEDURES FOR COMBINED RATIO ESTIMATORS IN COMPLEX SAMPLE SURVEYS

Aylin Alkaya[1], H. Öztaş Ayhan[2], Alptekin Esin[3]

ABSTRACT

In sample surveys weighting is applied to data to increase the quality of estimates. Data weighting can be used for several purposes. Sample design weights can be used to adjust the differences in selection probabilities for non-self weighting sample designs. Sample design weights, adjusted for nonresponse and non-coverage through the sequential data weighting process. The unequal selection probability designs represented the complex sampling designs. Among many reasons of weighting, the most important reasons are weighting for unequal probability of selection, compensation for nonresponse, and post-stratification. Many highly efficient estimation methods in survey sampling require strong information about auxiliary variables, *x*. The most common estimation methods using auxiliary information in estimation stage are regression and ratio estimator. This paper proposes a sequential data weighting procedure for the estimators of combined ratio mean in complex sample surveys and general variance estimation for the population ratio mean. To illustrate the utility of the proposed estimator, Turkish Demographic and Health Survey 2003 real life data is used. It is shown that the use of auxiliary information on weights can considerably improve the efficiency of the estimates.

Key words: combined ratio estimator, data weighting, design weight, nonresponse weighting, Post-stratification, weighting, sequential weighting.

1. Introduction

Applying weights to sample survey data is one of the important methods that are used to correct for sampling and nonsampling biases and to improve efficiency of estimations in sample surveys. The use of an insufficient sampling

[1] Department of Business Administration, Nevşehir Hacı Bektaş Veli University, 50300 Nevşehir, Turkey. E-mail: Aylin@nevsehir.edu.tr
[2] Department of Statistics, Middle East Technical University, 06800 Ankara, Turkey. E-mail: OAyhan@metu.edu.tr
[3] Department of Statistics, Gazi University, 06500 Ankara, Turkey. E-mail: Alpesin@gazi.edu.tr

framework, incorrect implementation of the sample selection process, inaccurate data collection and evaluation, nonresponses etc. can lead to biased estimates. The weights are applied to obtain unbiased estimates from the biased sample (Ayhan 1981). The rationale of weighting sample data is to make survey estimates to be representative of the whole population in the cases of selecting units with unequal probabilities; nonresponse; and coverage errors which creates bias and departures between sample and the reference population (Holt and Elliot 1991; Smith 1991). Weighting the data can be conducted sequentially for unequal selection probability, nonresponse, coverage errors, post-stratification as a process. At each step of sequential data weighting the calculated weights are multiplied the previous step weights.

In the first step of sequential data weighting, design weights W_i are assigned to the sampling units. Kish (1992) has stated that, design weights can be either the element's selection probability $W_i = k(1/\pi_i)$ or proportional to that inverse $W_i \propto 1/\pi_i$. It is common to increase the sampling fraction $f = n/N$ to $k\,f$ ($k > 1$) in order to reduce sampling errors in one or more domains, where the domain weights will be $w_h \propto 1/f_h = N_h/n_h$ and sampling fractions will be $f_h = n^*/N_h$. Weighting data by $W_i \propto 1/\pi_i$ is a simple process that should be "always" applied to samples with unequal π_i's (*according to the design based theory*). The general and most useful form of weighting is to assign the weights W_i to the sample cases *i* with $W_i = 1/\pi_i$, *i= 1 . . . N.* The selection probabilities π_i for all sampling units must be known for all probability samples by definition (Kish 1992).

For the sample, *n* units are selected from a finite population size *N* with known but unequal probabilities. Complex sample surveys such as stratification, clustering or multi stage sampling involve unequal selection probabilities. In these surveys to compensate for the differences in the probabilities of selection of samples weighting is introduced, the data is weighted with the inverse of the selection probabilities of units. The purpose is to weight each sampling unit to produce unbiased estimates of population parameters.

The second step of weighting is the adjustment for unit or total nonresponse. Nonresponse leads bias because usually nonrespondents differ from respondents. The lower the response rate, the higher the bias will be. Nonresponse weighting adjustments increase the weights of the sampled units for which data were collected. This means that every responding unit in the survey is assigned a weight, and estimates of population characteristics are obtained by processing weighted observations.

After nonresponse adjustments of the weights, further adjustments for noncoverage can be assigned to the weights as appropriate. Non-coverage refers to the failure of the sampling frame to cover the entire target population. In

practice, to reduce the effect of noncoverage and nonresponse the design weights are generally adjusted by a weighting method of calibration. The method depends on auxiliary variable(s) which uses auxiliary variable information to increase efficiency of the estimators. Calibration is called as a weighting method and in the literature many weighting methods such as raking, post-stratification, generalized regression estimator (GREG) and linear weighting are classified as a calibration weighting method. Efficient weighting for variable values observed in a survey is a topic with a long history. The earliest references to the use of weighting include the iterative proportional fitting technique as named raking by Deming and Stephan (1940). The reference of calibration starts with Deville (1988) and continues with Deville and Särndal (1992), Wu and Sitter (2001), Wu (2003), Estevao and Särndal (2006), Kott (2006). Särndal (2007). Some of the substantial references for GREG are Cassel, Särndal and Wretman (1976), Särndal (1980), Isaki and Fuller (1982), Wright (1983), Deville and Särndal (1992), Deville, Särndal and Sautory (1993), Kalton and Flores-Cervantes (2003), Ardilly and Tillé (2006) and Tikkiwal, Rai and Ghiya (2012) studies.

Post-stratification is a well-known and frequently used weighting method to reduce nonresponse and noncoverage bias. Post-stratification is stratification after selection of the sample in Cochran (1977: 135). Post-stratification studies continued by Guy (1979), Holt and Smith (1979), Bethlehem and Kersten (1985), Bethlehem and Keller (1987), Little (1993), Singh (2003), Lu and Gelman (2003), Cervantes and Brick (2009) and many other studies. The idea behind the post-stratification is to divide population into homogenous strata according to the information gathered from the sample population (Bethlehem and Kersten 1985). Additionally, in the last step of sequential weighting, extreme weights (high or low) can be adjusted using a methodology known as trimming, which is often done to reduce the variance of the weights.

Auxiliary information is used for improving the efficiency of the sample survey design. The most common estimation methods using auxiliary information are regression and ratio estimator. The ratio estimator uses auxiliary variable information to produce efficient estimates. Cochran (1940) was the first to show the contribution of known auxiliary information in improving the efficiency of the estimator of the population mean $\bar{Y}$ in survey sampling (Singh 2003). The quantity that is to be estimated from a sample design is the ratio of two variables both of which vary from unit to unit. In this paper, the population parameter to be estimated is the two variable ratio, *R*. Under stratified random sampling designs, there are two ways to produce ratio estimates, one way is the separate ratio estimator and the second way is the combined ratio estimator. Many large scale complex sample surveys are based on combined ratio mean estimator. "Combined ratio mean" is more practical to compute than the "separate ratio mean".

Sequential data weighting methodology (Deming and Stephan 1940, Stephan 1942) for the combined ratio estimator is handled by Ayhan (1991) and Verma (1991) and was elaborated by Ayhan (2003). The purpose of this paper is to present a combined ratio estimator under sequential weighting procedure rely on

Ayhan (2003)'s combined ratio estimator. In accordance with this purpose, combined ratio estimator is merely to provide an estimator for illustration. Alternative illustrations can also be made for the separate ratio estimators, in another context.

In the proposed estimator, the weights are based on selection probabilities, the observed values of auxiliary variables. Compared to the known combined ratio estimator, this method uses more information about auxiliary variables in regard to determining the weights. It can be expected that Ayhan (2003)'s combined ratio estimator which involves more information in determining weights will give additional gain on the accuracy of the parameter estimation.

Simple variance formulae depend on one variable and for linear estimators are extensively given in the literature. However, in variance estimation of complex estimators which depend on more than one variable or nonlinear estimator (e.g., ratio, regression or calibration estimator) there complex structural variance estimation methods should have to be required. Lu and Gelman (2003) develop a method for estimating the sampling variance of survey estimates with weighting adjustments. This study revealed a general equation for variance estimation of the population ratio estimator under sequential weighting through Lu and Gelman (2003) variance estimation equation.

The paper is organized as follows. In Section 2, ratio estimation in simple random sampling and combined ratio estimation in stratified sampling ratio estimation is introduced. In Section 3, an alternative combined ratio estimator which was proposed depending on Ayhan (2003)'s combined ratio estimator under sequential weighting in complex sample surveys is considered. Section 4 contains a general equation for variance estimation of the population ratio estimator in weighted data depending on Taylor-series method determination. Section 5 covers variance inflation factor in the comparison of the weighting methods. The methodology using the 2003 Turkey Demographic and Health Survey (TDHS 2003) is given in Section 6. The conclusions are summarized in Section 7.

2. Estimation of a two variable ratio

Frequently, the quantity that is to be estimated from a sample design is the ratio of two variables both of which vary from unit to unit. Let U be a finite population consisting of N elements $(u_1, u_2, \ldots, u_N)$ on which the variables y and x are defined. The values of variables (y, x) for U_i be y_i, x_i, $i = 1, \ldots, N$. Denoted by (Y, X) the population totals of (y, x), respectively. The population parameter to be estimated is the two variable ratio,

$$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{Y}{X} \tag{1}$$

and corresponding sample estimate is,

$$r = \frac{\sum_{i=1}^{n} (y_i / \pi_i)}{\sum_{i=1}^{n} (x_i / \pi_i)}. \tag{2}$$

The ratio estimator r determined by Horvitz-Thompson (1952) and be accepted as a Hájek (1971) type estimator, where $\pi_i = P(i \in s)$ defined as sample inclusion probability for unit *i*, $i = 1 \ldots N$.

There are too many reasons to take into account of the ratio estimator, $r = y / x$. One of them is related to a random variable not a sample size n. In addition, in many cases, sampling units are different from the basic units.

The purpose of using auxiliary variables in the estimation stage is to get better estimates. High levels of efficient estimation strategies involve extensive auxiliary information (Särndal et. al. 1992). When y and x are highly correlated, the ratio estimator provides greater reduction in the standard error and increases the accuracy of estimates. The ratio estimator is consistent but a biased estimator, this bias can be neglected. In most of the practical surveys, being a biased estimator seems substantially trivial besides yielding significant reduction of sampling error. When sample size is large enough, the ratio estimator is nearly normally distributed and the formula for its variance is valid. The results may be used if the sample size exceeds 30 (Cochran 1977).

The ratio estimation in SRS, the combined ratio estimation in stratified sampling and the proposed combined ratio estimation in complex sampling designs are presented here.

2.1. Ratio estimation in Simple Random Sampling

Let sampling units based on two correlated measures are y_i and x_i, which are selected from a population by simple random sample of size *n*. Naturally, SRS is a self-weighted sampling design, thus under a SRS design, while obtaining the ratio estimation and its variance we need to assign weights to the data. In SRS without replacement, the design weights are $\pi_i = n / N$ for all sampling units,

$i = 1,..., N$. Hence, from Equation (2) the sample ratio r which is the estimate of R is

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{y}{x}. \tag{3}$$

y and x values are random variables and differ from sample to sample. Here, r is a ratio of two random variables and is obtained from SRS design.

2.2. Ratio estimation in Stratified Random Sampling

When using ratio estimation for R with stratified random sampling, there are two different ways to produce estimates. One is to make a separate ratio estimate of the total of each stratum and add these totals. The second one is the combined ratio estimate that is derived from a single combined ratio. The combined ratio estimation will be taken into account. The combined ratio estimator for R can be defined as the ratio of two totals as

$$r_c = \frac{\hat{Y}_{st}}{\hat{X}_{st}} = \frac{\sum_{h=1}^{H} W_h y_h}{\sum_{h=1}^{H} W_h x_h}, \qquad W_h = \frac{N_h}{n_h} \tag{4}$$

where $\hat{Y}_{st}$ and $\hat{X}_{st}$ are the standard estimates of the population totals Y and X; y_h and x_h are the sample totals of the stratum h for Y and X, respectively (*st* for *stratified*). N_h number of units in the stratum h, n_h sample size corresponding to the stratum h and W_h is hth stratum sample weight.

3. Proposed combined ratio estimator

The combined ratio estimator for R in complex sampling designs suggested by Ayhan (2003) will be continued. The weighting procedures are based on different subclasses (domains) for each type of weighting which is illustrated on Table 1. Design weights and nonresponse weights are obtained at *segregated class* levels, while post-stratification weights are based on either *cross class* or *mixed class* levels (Ayhan 2003). The table is designed to reflect different types of weights for each stage of the weighting operation, which can be considered as a combined conditional approach.

Table 1. Weighting layout for sequential weighting process

Design Weights *for Segregated Classes*			⇒	Nonresponse Weights *for Segregated Classes*			⇒	Post-stratification Weights *for Cross/Mixed Classes*	
${}_AW_1$				${}_AW_1^*$				${}_AW_1^{**}$	
	${}_AW_h$				${}_AW_h^*$				${}_AW_k^{**}$
		${}_AW_H$				${}_AW_H^*$			${}_AW_K^{**}$

Source: Ayhan (2003)

Table 1 illustrates the general sequential weighting process. Here, ${}_AW_h$ design weights, ${}_AW_h^*$ *h*th stratum nonresponse weights and ${}_AW_k^{**}$ *k*th post stratum weights, $k=1,\ldots,K$.

Design weights for non-self-weighting sample designs can be computed for each stratum *h* with the same probability of selection p_h for a combined ratio mean (Ayhan 1991; Verma 1991). Ayhan (2003) extend the combined ratio estimator and design weights and the design weight ${}_AW_h$ for *h*th strata is,

$$ {}_AW_h = P_0 / P_h = \left[\sum_{h=1}^{H} x_h \Big/ \sum_{h=1}^{H} \left(x_h / P_h \right) \right] \Big/ \left[\left(X/x \right) p_h \right] \qquad (5) $$

$$ P_0 = \sum_{h=1}^{H} x_h \Big/ \sum_{h=1}^{H} \left(x_h / P_h \right) $$

$$ P_h = (X/x) p_h \ , \qquad p_h = n_h / N_h $$

Here P_0 is an adjustment factor for the overall weighted and unweighted sample sizes, p_h is the *h*th stratum units selection probability depends on auxiliary variable *x*.

The combined ratio estimator depends on the design weights (5) can be written as

$$ {}_Ar_c = \frac{\sum_{h=1}^{H} {}_AW_h \, y_h}{\sum_{h=1}^{H} {}_AW_h \, x_h} \ . \qquad (6) $$

In sequential data weighting, a weighting procedure for nonresponse is essential for self-weighting and nonself-weighting sample design outcomes.

If there are nonrespondents in the sample, the design weights have to be adjusted for nonresponse. The nonresponse weight, ${}_AW_h^*$ for *h*th strata is

$$ {}_AW_h^* = R_0 / R_h \tag{7}$$

$$R_0 = \sum_{h=1}^{H} \left({}_AW_h \; x_h \right) \Big/ \sum_{h=1}^{H} \left({}_AW_h \; x_h \; / \; R_h \right). \tag{8}$$

where R_h is the response rate in stratum h and R_0 is the overall response rate which is used to adjust the sample sizes to be the same, $\sum_{h=1}^{H} \left({}_AW_h \, {}_AW_h^* \; x_h \right) = x$. The combined ratio mean estimator depends on the design weights from Equation (6) and nonresponse weights from Equation (7) will be,

$${}_Ar_c = \frac{\sum_{h=1}^{H} {}_AW_h \; {}_AW_h^* y_h}{\sum_{h=1}^{H} {}_AW_h \; {}_AW_h^* x_h} \; . \tag{9}$$

Finally, a weighting procedure for post-stratification of a complex sampling scheme requires additional weighting procedures for independent subclasses. Post-stratification weights are given by

$$ {}_AW_k^{**} = X_k / X \tag{10}$$

where $\sum_{k=1}^{K} {}_AW_k^{**} \sum_{h=1}^{H} \left({}_AW_h \;\; {}_AW_h^* \;\; x_k \right) = x$ is the overall sample adjustment procedure (Ayhan 2003). At the last step of sequential data weighting, if design weights are adjusted for nonresponse and post-stratification in complex sampling surveys, the combined ratio estimator is computed as

$${}_Ar_c = \frac{\sum_{k=1}^{K} {}_AW_k^{**} \sum_{h=1}^{H} {}_AW_h \; {}_AW_h^* y_{khR}}{\sum_{k=1}^{K} {}_AW_k^{**} \sum_{h=1}^{H} {}_AW_h \; {}_AW_h^* x_{khR}} \tag{11}$$

where y_{khR} is the *k*th post strata, *h*th stratum sample total from respondents.

4. General variance estimation for the population ratio estimator

Although weighting data or sequential data weighting procedures are commonly used, it can be difficult to estimate sampling variances of associated weighted estimates. Lu and Gelman (2003) proposed a method for estimating the sampling variances of survey estimates with weighting adjustments derived from design-based analytic and Taylor-series variance estimators of population mean estimator in a general way. A natural simplifying assumption is to pretend that the weighting is all inverse-probability, with independent sampling where the probability that unit i is selected with proportional to $\pi_i = 1/W_i$. To compute the variance for inverse-probability weighting, a general variance estimator for $\theta = \bar{Y}$ acknowledged as a ratio form of the weighted mean

$$\hat{\theta} = \frac{\sum_{i=1}^{n} W_i y_i}{\sum_{i=1}^{n} W_i} \tag{12}$$

where the denominator of this expression is 1, but only after the weights have normalized. The variance of $\hat{\theta}$ is given by

$$\hat{V}_{HT}(\hat{\theta}) = \sum_{i=1}^{n} W_i^2 (y_i - \hat{\theta})^2 \tag{13}$$

and $\sum^{n} W_i = 1$ (Lu and Gelman 2003).

Taylor-series method consists of deriving from a complex non-linear statistic, a linear statistic which has the same asymptotic variance

$$\hat{V}(\hat{\theta}) \approx \hat{V}\left(\sum_{i=1}^{n} W_i \; z_i \right) \tag{14}$$

where z_i new variable whose expression depends on $\hat{\theta}$ and called a linearized variable for $\hat{\theta}$. When $\hat{\theta} = y/x$ then $z_i = y_i - \hat{\theta}\, x_i$, $i = 1,...,n$ (Osier and Museux 2006). Mean of this variable is $\bar{Z} = \bar{Y} - R\bar{X} = 0$. Therefore, in a weighted sample the variance estimation of mean $\bar{z}$ is

$$\hat{V}(\bar{z}) = \frac{1}{N^2} \sum_{i=1}^{n} W_i^2 \left[(y_i - \bar{y}\;) - r(x_i - \bar{x}) \right]^2 \tag{15}$$

General variance estimation for the estimators of population ratio $\theta = R$, the linear relation can be expressed by $y = bx$. In weighted data for the variance estimation of population ratio estimator is given by Taylor-series method as using variable $z_i = y_i - \hat{\theta} x_i$, where $\hat{\theta} = r = y/x$, the variable sample mean is

$\bar{z} = \bar{y} - \hat{\theta}\bar{x}$ and population total estimate is $N\ \bar{z}$. Thereby, since $\hat{\theta} = r$ then the variance estimation of r is

$$\hat{V}(r) = \frac{1}{X^2}\sum_{i=1}^{n} W_i^2\left[(y_i - \bar{y}) - r(x_i - \bar{x})\right]^2 \qquad (16)$$

A general equation for variance estimation of the population ratio estimator under stratified random sampling design depending on Taylor-series method can be introduced depending on Equation (16). A new variable defined as $z_{hi} = y_{hi} - \hat{\theta}x_{hi}$, where $\hat{\theta} = r_c$. Since this variable sample mean is $\bar{z}_{st} = \bar{y}_{st} - \hat{\theta}\bar{x}_{st}$ and population total estimate is $N\ \bar{z}_{st}$. Here $\bar{y}_{st} = N^{-1}\sum_{h=1}^{H} N_h \bar{y}_h$ and $\bar{x}_{st} = N^{-1}\sum_{h=1}^{H} N_h \bar{x}_h$ are the standard estimates of the population means $\bar{Y}$ and $\bar{X}$, respectively, made from a stratified sample. Mean of the new variable is $\bar{Z} = \bar{Y} - R\bar{X} = 0$. Thereby, under stratified random sampling design the variance estimation of $\bar{z}_{st}$,

$$\hat{V}(\bar{z}_{st}) = \frac{1}{N^2}\sum_{h=1}^{H} W_h^2 s_{zh}^2, \quad W_h = \frac{N_h}{n_h} \qquad (17)$$

$$\hat{V}(\bar{z}_{st}) = \frac{1}{N^2}\sum_{h=1}^{H} W_h^2 \sum_{i=1}^{n_h}\left[(y_{hi} - \bar{y}_h) - r_c(x_{hi} - \bar{x}_h)\right]^2 \qquad (18)$$

Where n_h is the sample size of stratum *h*, y_{hi} is the *i*th value of variable y in stratum h, x_{hi} is the *i*th value of variable x in stratum h (here $(1 - f_h)$ are neglected where f_h is the sampling fraction for *h*th stratum). Therefore, defining $z_i = y_i - \hat{\theta}x_i$ and $\hat{\theta} = r_c$, the variance estimation can be obtained as

$$\hat{V}(r_{st}) = \frac{1}{X^2}\sum_{h=1}^{H} W_h^2 \sum_{i=1}^{n_h}\left[(y_{hi} - \bar{y}_h) - r_c(x_{hi} - \bar{x}_h)\right]^2 \qquad (19)$$

This general variance estimation formulation for the stratified sampling design can also be extended to be the basis for the other complex sampling designs.

5. Variance inflation factor in the comparison of the weighting methods

The variability of weights increases, thereby the accuracy of estimates decreases. A useful measure of the accuracy of this loss is the variance inflation factor (*VIF*). *VIF* which is adopted to be the variability measure of weights can be used for comparing the weights and weighting methods. The measure *VIF*

represents the multiplying factor that is applied to the variance of a survey estimate due to the variability in the weights where equal weights are optimal (Kalton and Cervantes 2003; Kish 1992). Even though the use of the weights in the analysis of survey data tends to reduce the bias in the estimates, it could also inflate the variances of such estimates. The effect of using weights in the estimation of the population parameters can be defined by the *VIF*,

$$VIF = n \frac{\sum_{i=1}^{n} W_i^2}{\left(\sum_{i=1}^{n} W_i \right)^2} = 1 + CV^2(W_i) \tag{20}$$

W_i is the i th sampling unit weight and $CV^2(W_i)$ indicates the relative loss is defined as the coefficient of variation of weights (Kish 1992).

6. Application of the methodology

In this section we demonstrate the proposed methodology and study the efficiency of the combined ratio estimators by using data from 2003 Turkey Demographic and Health Survey (TDHS-2003). In the selection of the TDHS-2003 sample, weighted multi-stage stratified cluster sampling approach was used. Here, under stratified sampling design, the combined ratio estimator and proposed combined ratio estimator will be used for the estimation of population ratio.

6.1. Survey design

TDHS-2003 is the eighth Turkish national survey carried out by the Institute of Population Studies in Turkey. The major objective of the TDHS-2003 survey was to ensure that the survey would provide estimates with acceptable precision for the domains for most of the important demographic characteristics, such as fertility, infant and child mortality, and contraceptive prevalence, as well as for the health indicators. In TDHS-2003 to represent Turkey nationally and at the urban-rural and regional levels interviews were carried out with 8075 ever-married women in 10836 households. The sample design and sample size of the TDHS–2003 provides to perform analyses for Turkey as a whole, for urban and rural areas and for the five demographic regions of the country (West, South, Central, North and East). The sample of the research also allows for the analysis of 12 geographical regions (NUTS 1), which was established within the second half of the year 2002 within the context of Turkey's move to join the European Union. Among these 12 regions, İstanbul and the Southeastern Anatolian Project regions (due to their special situations) were oversampled. Thereby, settlements are divided into 40 strata, H=40.

From the 2000 Turkish General Population Census the population size for the ever-married women is N = 12630510. In the TDHS-2003 the eligible women

were identified as 8477 of whom 96 percent were interviewed and so interviews were carried out with 8075 ever-married women.

6.2. Two variable ratio estimation

One of the objectives of this paper is to measure a population ratio. Using data from the TDHS-2003, we have decided to examine the ratio of the number of live births to the number of living children of ever-married women. Therefore, y, indicates the number of living children and x, indicates the number of live births. $R = Y / X =$*Number of living children / Number of live births* will be estimated. In the estimation of R the known combined ratio estimator r_c and Ayhan (2003)'s proposed combined ratio estimator ${}_A r_c$ are used and the comparison of the r_c and ${}_A r_c$ estimates are illustrated in the following sections.

The initial information on all places of residences in Turkey was derived from the year 2000 Turkish General Population Census results which provided a computerized list of all settlements (provincial and district, sub-districts and villages), their populations and the households. From 2000 Turkish General Population Census results, the true population ratio for the ever-married women is $R = Y / X =$ 30398682/32713021 = 0.929253 and this means in Turkey nearly 93% of the live birth children are still living. Eligible women design weights, W_h strata design weights, response rates, respondent sample sizes, W_h' *h*th strata adjusted design weights and final design weights by strata information are presented in Table 2.

Table 2. Eligible women design weights and response rates, respondent sample sizes, adjusted design weights and final design weights by strata, Turkey 2003

Strata	Inverse of sampling fraction W_h	Household level $1/r_h^{HH}$	Women level $1/r_h^{WOMEN}$	n_{hR}	Women adjusted design weights in entire sample $W_h'.(N/N_r)$	Women standardized weight in entire sample $(W_h')^s$	Women weight in entire sample $(W_h')^s$ x1000000
1	1160555/960	891/779	672/630	630	1708.8	1.076474	1076474
2	1587651/60	870/682	478/449	449	2602.12	1.659981	1659981
3	24989/100	68/63	52/50	50	324.996	0.272980	272980
4	76858/60	46/46	35/34	34	1209.74	0.962433	962433
5	469931/500	410/391	285/269	269	2455.64	0.802196	802196
6	362247/240	220/218	119/115	115	5688.92	1.150401	1150401
7	685892/400	348/300	195/183	183	1680.23	1.546953	1546953
8	686133/150	144/137	96/94	94	1333.83	3.583791	3583791

Table 2. Eligible women design weights and response rates, respondent sample sizes, adjusted design weights and final design weights by strata, Turkey 2003 (cont.)

Strata	Inverse of sampling fraction W_h	Household level $1/r_h^{HH}$	Women level $1/r_h^{WOMEN}$	n_{hR}	Women adjusted design weights in entire sample $W_h'.(N/N_r)$	Women standardized weight in entire sample $(W_h')^s$	Women weight in entire sample $(W_h')^s$ x1000000
9	667273/240	211/204	139/135	135	1655.51	2.305124	2305124
10	202772/150	129/127	94/89	89	5173.43	1.058475	1058475
11	211704/60	48/47	50/48	48	2235.38	2.739621	2739621
12	352876/400	348/300	225/200	200	598.79	0.840259	840259
13	129118/100	83/75	46/46	46	2734.71	1.042909	1042909
14	109307/60	33/33	27/26	26	3442.6	1.841054	1841054
15	377921/100	90/86	70/62	62	846.61	3.259059	3259059
16	148605/60	56/56	39/38	38	1587.51	1.855263	1855263
17	182284/100	86/85	68/65	65	1305.67	1.408203	1408203
18	65446/60	45/45	21/21	21	867.53	0.796109	796109
19	47999/100	80/77	57/55	55	1349.47	0.377212	377212
20	83237/60	55/55	44/43	43	641.78	1.036076	1036076
21	915073/500	451/386	287/260	260	513.44	1.722755	1722755
22	431779/150	128/124	99/99	99	945.73	2.168697	2168697
23	298404/240	173/172	116/107	107	946.82	1.130884	1130884
24	276431/400	361/349	276/270	270	1527.76	0.533328	533328
25	1052242/900	808/734	593/557	557	1826.15	1.028638	1028638
26	681896/540	470/446	302/286	286	3430.47	1.085906	1085906
27	523267/500	457/438	354/343	343	4348.88	0.822517	822517
28	373756/240	210/205	162/159	159	2191.87	1.186317	1186317
29	336258/500	427/395	275/267	267	2945.04	0.546506	546506
30	318422/240	207/204	156/153	153	1263.75	1.001856	1001856
31	224473/200	180/176	138/136	136	1644.66	0.850111	850111
32	201222/90	82/82	60/59	59	1570.77	1.659488	1659488
33	310851/600	497/474	362/355	355	1628.01	0.404297	404297
34	349165/240	203/199	136/126	126	1883.16	1.169152	1169152
35	212359/500	462/452	392/384	384	1590.35	0.323444	323444
36	218260/240	200/199	158/151	151	2634.27	0.797725	797725
37	371366/500	478/449	383/371	371	1855.92	0.595771	595771
38	257644/240	227/220	208/195	195	1108.02	0.862345	862345
39	756933/1000	922/877	762/742	742	1368.89	0.596458	596458
40	356146 / 480	455 / 449	416 / 403	403	899.22	0.566475	566475

Source: TDHS 2003

The nonresponse adjustments for the sampling weights W_h are conducted at each strata, $h = 1,..., H$.

The adjusted nonresponse weights $W_h(1/r_h^{HH})$ $(1/r_h^{WOMEN})$ are defined by multiplying sampling weights by the inverse of household and women level response ratios. However, to provide equality of the adjusted sampling weights total to the population total, the adjusted sampling weights $W_h(1/r_h^{HH})($ $1/r_h^{WOMEN})$ are multiplied with the value of,

$$N\Big/\left\{\sum_{h=1}^{H}\sum_{i=1}^{nhR} W_h(1/r_h^{HH})(1/r_h^{WOMEN})\right\} = 12630510/10901679 = 1.158584.$$

Thus, the adjusted sampling weights are presented as $W_h'.(N/N_r)$ in Table 2. For example, the calculation for the adjusted value W_h' = 1474.9 from Table 3 is as,
$W_h'.(N/N_h)$ = (1160555/960) (891/ 779) (672 / 630) 1.158584 = 1474.9(1.158584) = 1708.799. Hence, W_h' used for design weights W_h.

Table 3. Unit variances of strata

Strata	W_h'	s_{yh}^2	s_{xh}^2	s_{yxh}	Strata	W_h'	s_{yh}^2	s_{xh}^2	s_{yxh}
1	280.51	3.157	5.763	1.88	21	1405.17	3.147	4.130	1.89
2	443.16	4.064	6.010	2.22	22	1419.55	1.510	1.867	1.32
3	516.83	1.758	2.628	1.38	23	1428.91	0.757	0.973	0.83
4	553.94	1.762	2.119	1.22	24	1450.24	1.181	1.316	1.33
5	730.72	3.243	4.145	1.9	25	1474.9	1.952	2.820	2.37
6	748.78	1.330	1.719	1.24	26	1576.19	1.051	1.384	1.20
7	776.14	9.318	12.35	1.21	27	1601.88	2.861	3.538	3.03
8	816.28	4.600	6.186	3.3	28	1625.4	3.151	4.250	1.76
9	817.22	5.255	7.000	2.2	29	1891.85	1.421	2.106	1.32
10	956.36	6.562	9.490	2.5	30	1929.41	1.028	1.835	1.23
11	1044.15	1.340	1.590	1.18	31	2119.52	1.080	1.467	1.25
12	1090.77	2.747	2.857	1.67	32	2245.95	1.713	2.475	1.44
13	1126.95	2.248	2.935	1.55	33	2273.7	1.874	3.713	1.66
14	1151.26	1.480	1.949	1.28	34	2360.39	1.282	1.653	1.19
15	1164.75	1.837	2.978	1.67	35	2541.93	2.691	3.078	1.7
16	1181.52	7.153	10.17	3.03	36	2960.91	1.595	2.120	1.33
17	1318.64	0.952	0.941	1.97	37	2971.38	1.869	2.272	1.44
18	1355.77	2.294	3.509	1.76	38	3753.62	2.056	3.400	1.63
19	1370.21	2.966	3.833	1.88	39	4465.31	3.188	4.027	1.89
20	1372.67	2.589	3.367	1.71	40	4910.24	1.341	1.737	1.21

6.3. Combined ratio estimator

Combined ratio estimator for *R* is

$$r_c = \frac{\sum_{h=1}^{H} W_h' y_h}{\sum_{h=1}^{H} W_h' x_h} = \frac{30466444}{33214885} = 0.917253.$$

This means that, the ever-married women, 91.7% of live born children are estimated to have lived. A general variance estimation proposed for the population ratio which is given by Equation (19) can be written as,

$$\hat{V}(r_c) = \frac{1}{X^2} \sum_{h=1}^{H} W_h^2 \frac{1}{n_h} \left(s_{yh}^2 - 2 r_c s_{yxh} + r_c^2 s_{xh}^2 \right). \tag{21}$$

The variance estimation of combined ratio estimator (conventional combined ratio estimator) depending on W_h' adjusted weights is

$$\hat{V}(r_c) = \frac{1}{X^2} \sum_{h=1}^{H} W_h'^2 \frac{1}{n_h} \left(s_{yh}^2 - 2 r_c s_{yxh} + r_c^2 s_{xh}^2 \right)$$

$$\hat{V}(r_c) = (3.22)\ 10^{-9}.$$

s_{yh}^2, s_{yxh}, s_{xh}^2 computed unit variance values of the strata are given in Table 3. There is an increase in the variance of the ratio estimate due to the use of design weights W_h', and so that the VIF value is obtained as:

$$VIF\ (W_h') = \left[H \left(\sum_{h=1}^{H} (W_h')^2 \right) \Big/ \left(\sum_{h=1}^{H} W_h' \right)^2 \right] = 1.387289$$

For a VIF≈1.387, i.e., a reduction in the effective sample size of almost 38.7 percent.

6.4. Proposed combined ratio estimator

The estimator was proposed under sequential weighting process and so on the design weights are adjusted for nonresponse and post-stratification in TDHS-2003. First step is to obtain design weights ${}_A W_h$. Second step is to compute nonresponse weights ${}_A W_h^*$. The final step is weighting for post-stratification that is conducted by ${}_A W_k^{**}$. Here, *h=1,..,H, H=40.* The ${}_A W_h$ weight results are

presented in Table 4, calculation of R_h and R_0 results are presented in Table 5. ${}_AW_h^*$ weight results are presented in Table 6. ${}_AW_k^{**}$ weight results are presented in Table 7.

Design weights:

We will start with obtaining ${}_AW_h$ design weights. The adjustment factor P_0 is

$$P_0 = \sum_{h=1}^{H} x_h \Big/ \sum_{h=1}^{H} (x_h / P_h) = 22443.5 / 0.013092223 = 1714262.026$$

for the overall weighted and unweighted sample sizes is to be the same, where p_h, *h*th stratum units selection probability to the sample. The values $X = 32713021$, x = 22443, X / x = 1457.605 and P_h are then computed as below given in Table 4.

Table 4. P_h, x_h / P_h and combined design weights ${}_AW_h$

Strata	Women adjusted design weights $W_h'.(N / N_r)$	x_h	$P_h = (X / x)p_h$	x_h / P_h	${}_AW_h = P_0 / P_h$
1	1708.8	1442.70	2490699.324	0.000579	0.688
2	2602.12	969.84	3792777.695	0.000255	0.452
3	324.996	123.00	473705.116	0.000259	3.618
4	1209.74	508.41	1763283.357	0.000288	0.972
5	2455.64	362.34	3579272.524	0.000101	0.478
6	5688.92	186.12	8292011.470	0.000022	0.206
7	1680.23	197.58	2449056.487	0.000080	0.699
8	1333.83	448.00	1944153.488	0.000234	0.881
9	1655.51	102.12	2413025.303	0.000042	0.710
10	5173.43	135.16	7540647.592	0,000017	0.227
11	2235.38	135.20	3258227.678	0.000041	0.526
12	598.79	125.95	872779.639	0.000144	1.964
13	2734.71	590.20	3986037.189	0000148	0.430
14	3442.6	205.92	5017837.953	0.000041	0.341
15	846.61	650.70	1233995.175	0.005270	1.389
16	1587.51	1425.92	2313910.396	0.000616	0.740
17	1305.67	840.35	1903108.255	0.000441	0900
18	867.53	787.65	1264487.585	0.000622	1.355
19	1349.47	371.28	1966949.916	0.000188	0.871
20	641.78	930.10	935440.667	0.000994	1.832

Table 4. P_h, x_h / P_h and combined design weights ${}_AW_h$ (cont.)

Strata	Women adjusted design weights $W_h'.(N/N_r)$	x_h	$P_h = (X/x)p_h$	x_h / P_h	${}_AW_h = P_0 / P_h$
21	513.44	1025.28	748375.855	0.001370	2.290
22	945.73	953.47	1378469.728	0.000691	1.243
23	946.82	2144.38	1380058.482	0.001553	1,242
24	1527.76	78.88	2226820.459	0.000035	0.769
25	1826.15	354.20	2661745.418	0.000133	0.644
26	3430.47	449.55	5000157.602	0.000089	0.342
27	4348.88	171.84	6338806.459	0.000027	0.270
28	2191.87	85.02	3194808.713	0.000026	0.536
29	2945.04	152.00	4292608.344	0.000035	0.399
30	1263.75	68.04	1842006.830	0.000036	0.930
31	1644.66	107.07	2397210.645	0.000044	0.715
32	1570.77	254.66	2289510.638	0.000112	0.748
33	1628.01	1092.52	2372942,069	0.000460	0.722
34	1883.16	605.79	2744841.608	0.000220	0.624
35	1590.35	602.82	2318049.901	0.000260	0.739
36	2634.27	207.09	3839638.641	0.000053	0.446
37	1855.92	375.48	2705137.342	0.000138	0.633
38	1108.02	418.27	1615019.116	0.000258	1.061
39	1368.89	783.90	1995255.968	0.000392	0.859
40	899.22	1974.70	1310678.047	0.001506	1.307
	Total	22443.50		0.013092	

Nonresponse weights:

In TDHS–2003, there are also non-respondent women in the survey. A weighting procedure for nonresponse is essential so we should adjust the design weights by assigning nonresponse weights to the data. Table 5 presents the calculation of the response rates R_h.

Table 5. Calculation of R_h and the Equation R_0

Strata	R_h	n_{hR}	${}_AW_h\, x_h$	${}_AW_h\, x_h / R_h$	Strata	R_h	n_{hR}	${}_AW_h\, x_h$	${}_AW_h\, x_h / R_h$
1	0.820	630	2465286.0	3007711.90	21	0.775	260	526419.8	678937.77
2	0.737	449	2523640.0	3427234.70	22	0.969	99	901725.2	930813.09
3	0.89	50	39974.5	44872.97	23	0.917	107	2030342.0	2213915.50
4	0.97	34	615043.9	633133.44	24	0.946	270	120509.7	127423.38
5	0.90	269	889776.6	988509.07	25	0.853	557	646822.3	758053.41
6	0.95	115	1058822.0	1105702.20	26	0.899	286	1542168.0	1716072.10
7	0.80	183	331979.8	410348.86	27	0.929	343	747311.5	804735.05
8	0.93	94	597555.8	641451.46	28	0.958	159	186352.8	194499.83

Table 5. Calculation of R_h and the Equation R_0 (cont.)

Strata	R_h	n_{hR}	${}_AW_h\, x_h$	${}_AW_h\, x_h / R_h$	Strata	R_h	n_{hR}	${}_AW_h\, x_h$	${}_AW_h\, x_h / R_h$
9	0.93	135	169060.7	180042.87	29	0.898	267	447646.1	498410.29
10	0.93	89	699240.8	750154.29	30	0.967	153	85985.5	88960.83
11	0.94	48	302223.4	321514.23	31	0.964	136	176093.7	182744.35
12	0.76	200	75417.6	98419.97	32	0.983	59	400012.3	406792.16
13	0.90	46	1614026.0	1786188.60	33	0.935	355	1778633.0	1901711.90
14	0.96	26	708900.2	736165.58	34	0.908	126	1140799.0	1256089.70
15	0.84	62	550889.1	650900.51	35	0.958	384	958694.8	1000319.50
16	0.97	38	2263662.0	2323232.30	36	0.951	151	545531.0	573688.93
17	0.94	65	1097220.0	1161364.90	37	0.910	371	696860.8	765865.46
18	1.00	21	683310.0	683310.00	38	0.909	195	463451.5	510077.56
19	0.92	55	501031.2	539481.08	39	0.926	742	1073073.0	1158541.50
20	0.97	43	596919.6	610801.43	40	0.956	403	1775690.0	1857464.10
	Total						8075	34028101	37725657

From Equation (8)

$$R_0 = \frac{34028101}{37725657} = 0.901988.$$

Further, using R_h and R_0 response rate values, the combined weights for nonresponse ${}_AW_h^*$ from Equation (7) are obtained and given in Table 6.

Table 6. ${}_AW_h^*$ combined weights for nonresponse

Strata	${}_AW_h^*$	${}_AW_h\, {}_AW_h^*$	Strata	${}_AW_h^* = R_0 / R_h$	${}_AW_h\, {}_AW_h^*$
1	1.1004	1880.447	21	1.1633	597.2943
2	1.2249	3187.459	22	0.9310	880.5547
3	1.0125	329.0642	23	0.9835	931.2369
4	0.9285	1123.265	24	0.9537	1457.079
5	1.0020	2460.737	25	1.0570	1930.422
6	0.9419	5358.535	26	1.0037	3443.17
7	1.1149	1873.316	27	0.9712	4224.055
8	0.9682	1291.477	28	0.9414	2063.474
9	0.9605	1590.252	29	1.0042	2957.633
10	0.9676	5006.144	30	0.9331	1179.33
11	0.9595	2144.986	31	0.9360	1539.491
12	1.1770	704.8326	32	0.9172	1440.83

Table 6. ${}_A W_h^*$ combined weights for nonresponse (cont.)

Strata	${}_A W_h^*$	${}_A W_h \, {}_A W_h^*$	Strata	${}_A W_h^* = R_0 / R_h$	${}_A W_h \, {}_A W_h^*$
13	0.9982	2729.789	33	0.9644	1570.06
14	0.9366	3224.615	34	0.9931	1870.249
15	1.0657	902.2662	35	0.9411	1496.759
16	0.9257	1469.597	36	0.9485	2498.724
17	0.9547	1246.549	37	0.9913	1839.783
18	0.9019	782.5019	38	0.9927	1099.969
19	0.9712	1310.616	39	0.9738	1333.067
20	0.9229	592.3403	40	0.9435	848.4382

Weighting for post-stratification:

In TDHS-2003 survey the age group auxiliary variable x_2 is used for post-stratification. The data separated into $k=7$ age groups (post strata, $k = 1,...,7$). Post-stratification weights ${}_A W_k^{**}$ were defined by Equation (10) and the ratio estimator can be obtained by Equation (11). The components $\sum_{h=1}^{H} {}_A W_h \, {}_A W_h^* y_{hkR}$, $\sum_{h=1}^{H} {}_A W_h \, {}_A W_h^* x_{hkR}$ have been computed and presented in Table 7.

Table 7. n_{kR} and N_k distribution, $\sum_{h=1}^{H} {}_A W_h \, {}_A W_h^* y_{hkR}$ and $\sum_{h=1}^{H} {}_A W_h \, {}_A W_h^* x_{hkR}$ weights by age groups

Post strata	x_2 : Age group	N_k	n_{kR}	$\sum_{h=1}^{H} {}_A W_h \, {}_A W_h^* y_{hkR}$	$\sum_{h=1}^{H} {}_A W_h \, {}_A W_h^* x_{hkR}$
1	15–19	453511	240	180.5941	194.5517
2	20–24	1727365	1080	1749.078	1842.8710
3	25–29	2378665	1516	4214.355	4002.4680
4	30–34	2244391	1506	5734.090	5362.3310
5	35–39	2282957	1410	6394.855	5852.9540
6	40–44	1922351	1297	6846.528	6141.9070
7	45–49	1621270	1026	5581.468	4919.2090

In the estimation of R, the population total of the number of live births for the post-stratified sample by age groups must be known. From the 2000 General Census of Population X_k, the *k*th post-stratified population totals are obtained. The population totals and the post-stratification weights ${}_A W_k^{**} = X_k / X$ are presented in Table 8.

Table 8. X_k population totals and ${}_AW_k^{**}$ post-strata weights

x_2 :Age group	X_k	${}_AW_k^{**} = X_k / X$	$W_k^{**}\sum_{h=1}^{H} {}_AW_h\,{}_AW_h^{*} y_{khR}$	$W_k^{**}\sum_{h=1}^{H} {}_AW_h\,{}_AW_h^{*} x_{khR}$
15–19	294628	0.0359	7883.0720	8452.682
20–24	2078364	0.1368	263369.562	275931.954
25–29	4522719	0.1883	848192.363	886883.534
30–34	5700038	0.1777	1033509.385	1099511.889
35–39	7036619	0.1807	1173253.904	1279513.638
40–44	6707033	0.1522	1011205.470	1126042.687
45–49	6394157	0.1284	717786.487	820766.657
Total	32733558	-	5055200.244	5497103.043

$$ {}_Ar_c = \frac{\sum_{k=1}^{K} {}_AW_k^{**}\sum_{h=1}^{H} {}_AW_h\,{}_AW_h^{*} y_{khR}}{\sum_{k=1}^{K} {}_AW_k^{**}\sum_{h=1}^{H} {}_AW_h\,{}_AW_h^{*} x_{khR}} = \frac{5055200.244}{5497103.43} = 0.919612 $$

We can state that, 91.9% of live born children is estimated to have lived. The post-stratification weights and related unit variances are computed and presented on Table 9.

Table 9. Post-stratification weights and unit variances

x_2 : Age group	$W_k^{**}\,{}_AW_h\,{}_AW_h^{*}$	s_{yk}^2	s_{xk}^2	s_{yxk}
15–19	14040.19763	0.50	0.25	1.32
20–24	226920.2164	0.92	0.84	1.70
25–29	448066.179	1.63	2.65	2.11
30–34	415966.1629	1.44	2.09	1.33
35–39	399255.382	3.36	11.33	6.70
40–44	303439.722	1.50	2.25	1.80
45–49	212837.3363	1.39	1.95	2.70

The variance estimation given by Equation (19) can be defined as below for ${}_Ar_c$:

$$ \hat{V}({}_Ar_c) = \frac{1}{X^2}\sum_{k=1}^{K} W_k^2\,\frac{1}{n_h}\left(s_{yk}^2 - 2r_c s_{yxk} + r_c^2 s_{xk}^2\right) \qquad (22) $$

The variance estimation value is

$$\hat{V}({}_{A}r_{c}) = (2.9)\ 10^{-9}$$

where $W_k = W_k^{***}{}_A W_{h\,A} W_h^*$ and s_{yk}^2 s_{xk}^2 unit variances of k th poststrata for y and x, respectively and s_{yxk} is covariance of k th poststrata for y and x. Inflation factor for W_k obtained as $VIF(W_k) = 1.238408$. There is nearly 10% reduction in the VIF value of proposed ratio estimator relative to conventional combined ratio estimator. VIF is reduced from 1.387 with conventional combined ratio estimator to 1.238 with proposed ratio estimator.

The comparison of the conventional combined ratio estimator and the proposed combined ratio estimator results of means, variance estimations and VIF are given on Table 10. In Table 10, we observe the values of mean, variance estimation and VIF of the combined ratio estimator and the proposed combined ratio estimator. From Table 10, it can be concluded that the proposed combined ratio estimator has the minimum variance estimation but it is seen that both have approximate variance estimation values. The variability level of weights according VIF values ${}_{A}r_{c}$ seems as less variable than r_c.

Table 10. The comparisons of combined ratio estimator results

	Mean	$\hat{V}_{HT}(\hat{\theta})$	VIF
Conventional combined ratio estimator	r_c = 0.917	(3.2) 10^{-9}	$VIF\ (W_h')$ = 1.387289
Ayhan (2003)'s combined ratio estimator	${}_{A}r_{c}$ = 0.919	(2.9) 10^{-9}	$VIF\ (W_k^{***}{}_A W_{h\,A} W_h^*)$ = 1.238408

7. Conclusions

Researchers believe that, the weights that provide excellent estimates for auxiliary variables will also provide good estimates for the interest variable. The new weights will continue to give unbiased estimates, but a realistic expectation is to remain near unbiasedness (Deville and Särndal 1992). Using the data weighted according to the auxiliary variable(s) which are known to be related to the interest variable lead to additional gains in the information. The weights in the combined ratio estimator ${}_{A}r_{c}$ are defined on the basis of population and sample sizes and also information on the auxiliary variable. TDHS-2003 results have shown that, the combined ratio estimator which is defined by Ayhan (2003) provided a better

estimate of the parameter, by using auxiliary variable values in the calculation of weights. The proposed estimator has lower variance; it is not enough to prove that it is more efficient. The variance could be underestimated. We can say that, the estimator better reflects the effect of post-stratification.

REFERENCES

ARDILLY, P., TILLE, Y., (2006). Sampling methods: exercises and solutions. Translated from French by Leon Jang. Springer Science+Business Media, USA.

AYHAN, H. Ö., (1981). Sources and bias of nonresponse in the Turkish Fertility Survey 1978, Turkish Journal of Population Studies 2–3, pp. 104–148.

AYHAN, H. Ö., (1991). Post–stratification and weighting in sample surveys. Invited paper, Research Symposium '91, State Institute of Statistics, Ankara.

AYHAN, H. Ö., (2003). Combined weighting procedures for post-survey adjustment in complex sample surveys. Bulletin of the International Statistical Institute, 60 (1), pp. 53–54.

BETHLEHEM, J. G., KERSTEN, H. M. P., (1985). On the treatment of nonresponse in sample surveys. Journal of Official Statistics, 1 (3), pp. 287–300.

BETHLEHEM, J. G., KELLER, W. J., (1987). Linear weighting of the sample survey data. Journal of Official Statistics, 3 (2), pp. 141–153.

CASSEL, C. M., SÄRNDAL, C. E., WRETMAN J. H., (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika, 63, pp. 615–620.

CERVANTES, I. F., BRICK, J. F., (2009). Efficacy of Poststratification in Complex Sample Design. Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 4642–4655.

COCHRAN, W. G., (1977). Sampling Techniques. 3rd ed. Wiley, New York.

DEMING, W. E., STEPHAN, F. F., (1940). On a least squares adjustment of a sample frequency table when the expected marginal totals are known. Annals of Mathematical Statistics, 11 (4), pp. 427–444.

DEVILLE, J. C., SARNDAL, C. E., (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, pp. 376–382.

DEVILLE, J. C., SARNDAL, C. E., SAUTORY, O., (1993). Generalized raking procedure in survey sampling. Journal of the American Statistical Association, 88, pp. 1013–1020.

ESTEVAO, V. M., SÄRNDAL, C. E., (2000). A functional form approach to calibration. Journal of Official Statistics, 16 (4), pp. 379–399.

GUY, P. W., (1979). Small sample theory for poststratification, Doctor of Philosophy, Graduate College of Texas ASM University, Texas.

HAJEK, J., (1971). Comment on a paper of D. Basu, In Foundations of Statistical Inference, Toronto, pp. 236–237.

HOLT, D., ELLIOT, D., (1991). Methods of weighting for unit non-response. The Statistician, 40, pp. 333–342.

HOLT, D., SMITH, T. M. F., (1979). Post stratification. Journal of the Royal Statistical Society, Series A (General), 142 (1), pp. 33–46.

HORVITZ, D. G., THOMPSON, D. J., (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, pp. 663–685.

ISAKI, C. T., FULLER, W. A., (1982). Survey design under the regression superpopulation model. Journal of the American Statistical Association, 77, pp. 89–96.

KALTON, G., CERVANTES, F. I., (2003). Weighting methods. Journal of Official Statistics, 19 (2), pp. 81–97.

KISH, L., (1965). Survey Sampling. John Wiley & Sons, Inc., USA.

KISH, L., (1992). Weighting for unequal Pi. Journal of Official Statistics, 8 (2), pp. 183–200.

KOTT, P. S., (2006). Using calibration weighting to adjust for nonresponse and coverage errors. Survey Methodology, 32, pp. 133–142.

LITTLE, R. J. A., (1993). Post-stratification: A modeler's perspective, Journal of The American Statistical Association, 88, pp. 1001–1012.

LITTLE, R. J. A., VARTIVARIAN, S., (2005). Does weighting for nonresponse increase the variance of survey means? Proceedings of American Statistical Association: Section on Survey Research Methods, Minneapolis, pp. 3897–3904.

LU H., GELMAN, A., (2003). A method for estimating design based sampling variances for surveys with weighting, post-stratification, and raking. Journal of Official Statistics, 19(2), pp. 133–151.

OSIER, G., MUSEUX, J. M., (2006). Variance estimation for EU–SILC complex poverty indicators using linearization techniques. European Conference on Quality in Survey Statistics, Luxembourg, pp. 1–11.

SÄRNDAL, C. E., (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling, Biometrika, 67, pp. 639–650.

SÄRNDAL, C. E., SWENSON, B., WRETMAN, J., (1992). Model assisted survey sampling. Springer–Verlag, New York.

SÄRNDAL, C. E., (2007). The calibration approach in survey theory and practice. Survey Methodology, Statistics Canada, 33 (2), pp. 99–119.

SINGH, S., (2003). Advanced sampling theory with applications. Kluwer Academic Publishers, Dordrecht Boston, London.

SMITH, T. M. F., (1991). Post–stratification. The Statistician, 40, pp. 315–321.

STEPHAN, F. F., (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. Annals of Mathematical Statistics, 13 (2), pp. 166–178.

TDHS, (2004). Turkey Demographic and Health Survey 2003. Hacettepe University, Institute of Population Studies, Ankara, Turkey.

TIKKIWAL, G. C., RAI, P. K., GHIYA, A., (2013). On the Performance of Generalized Regression Estimator for Small Domains, Communication in Statistics: Simulation and Computation, 42, pp. 891–909.

VERMA, V., (1991). Sampling Methods. Manual for Statistical Trainers Number 2, Statistical Institute for Asia and the Pacific, Tokyo, Japan.

VERMA, V., (2007). Recent advances in survey sampling. In: Ayhan, H.Ö. and Batmaz, I. (eds.), Recent Advances in Statistics. Turkish Statistical Institute Press, Ankara, Turkey, pp. 77–101.

WRIGHT, R. L., (1983). Finite population sampling with multivariate auxiliary information. Journal of the American Statistical Association, 78, pp. 879–884.

WU, C., SITTER, R. R., (2001). A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association, 96, pp. 185–193.

WU, C., (2003). Optimal calibration estimators in survey sampling. Biometrika, 90(4), pp. 937–951.

STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 271–290, DOI 10. 21307

A MULTIDIMENSIONAL AND DYNAMISED CLASSIFICATION OF POLISH PROVINCES BASED ON SELECTED FEATURES OF HIGHER EDUCATION IN 2002–2013

Wojciech Łukaszonek[1]

ABSTRACT

For close to two decades after the fall of communism in 1989, Polish higher education enjoyed an unprecedented period of development. Favourable political, economic, social and demographic changes led to a fivefold increase in the number of students and the number of higher educational institutions. The dynamic changes and their effects did not occur uniformly, in either space or time. An attempt is made here to identify and analyse the regional differentiation between Polish provinces in terms of features relating to higher education. To investigate the changes in higher education in the period of economic and social transformation, observations were made of fundamental characteristics of higher education in the years 2002–2013. The applied procedure uses new statistical methods applicable to a space of doubly multivariate data. The covariance matrix used to construct principal components is given the structure of a Kronecker product. The results led to the identification of six groups of provinces, including two consisting of a single province – Mazowieckie and Małopolskie provinces – which contain the largest and the highest-ranked[2] higher educational institutions in Poland: the University of Warsaw and Jagiellonian University.

Key words: higher education, doubly multivariate data, cluster analysis, dendrite method, covariance matrix, Kronecker product.

1. Introduction

The systemic changes that took place in Poland in the late 1980s and early 1990s led to transformation of the country in many different areas, including economic, social and cultural ones (Golinowska, ed. 2005). The process of changes in the 1990s also affected education, and its effects made a strong

[1] Faculty of Management, President Wojciechowski Higher Vocational State School, Kalisz, Poland. E-mail: w.lukaszonek@g.pl.

[2] Rankings by *Perspektywy* and *Rzeczpospolita/Gazeta Prawna* (2002–2015); Shanghai Ranking (2010–2015).

impression on the system of higher education. The economic reforms forced changes in the labour market, which led to increased needs for highly qualified staff. One consequence was a rapid growth in the number of people entering higher education. A degree seemed to be a guarantee of well-paid work, the possibility of further development, economic independence and improved social status (Sikorska 1998, Mach 2003). It had previously been an elite attribute, as reflected in the number of graduates in the population. In the centrally planned Polish economy of the 1970s and 1980s, the higher education system was closely controlled by the authorities, and student numbers were centrally regulated (Wnuk-Lipińska 1996, Antonowicz 2012, Kwiek 2014). In 1990 the percentage of the Polish population holding degrees was approximately 6%, this being a result of the policy applied in previous years. The adoption of democratic principles, giving more freedom to citizens, had a strong impact on social behaviours. There was an increase in Poles' educational aspirations, linked to the economic changes that were reflected in the dynamic expansion of the private sector (Ziółkowski 2000, Kwiek 2014). The increased demand for employment was accompanied by demographic changes, manifested in an increase in the population aged 19–24, the time at which higher education is undertaken. In effect, the number of students increased extremely rapidly. Over the years 1990–2005 the total number increased almost fivefold, from approximately 400,000 to almost 2 million (Fig. 1).

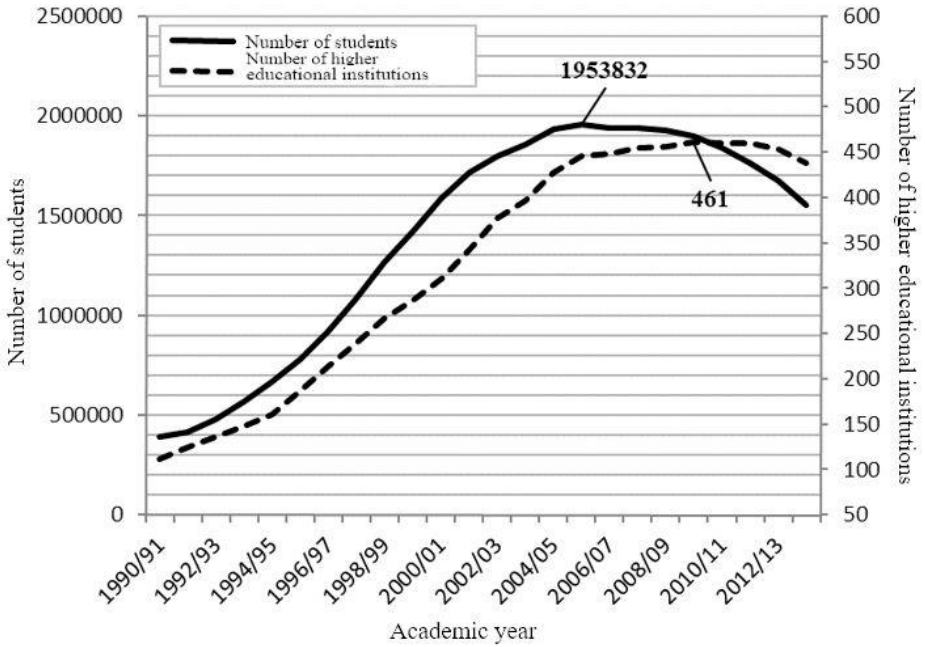


Figure 1. Numbers of students and of higher educational institutions in 1990–2013

Source: based on Higher Education 2014, GUS 2015.

The universities and colleges then existing in Poland were not prepared (particularly in terms of infrastructure) for such a rapid rise in the number of people interested in studying. In 1990 there were 112 higher educational institutions, all of them – apart from the Catholic University of Lublin – being state-run (public) institutions. The free market principles adopted at that time, and the *Act on higher education* (Dz.U. 1990 no. 65 item 385), enabled the foundation of Poland's first private higher educational institutions. The requirements for such an establishment were very liberal and relatively easy to fulfil, which gave an impetus to an unprecedented, and in effect uncontrolled, process of privatisation of higher education (Kwiek 2014). In the 15 years following the fall of communism, a total of 315 non-public (private) higher educational institutions were established in Poland, filling the gap in the market that had arisen due to the increasing public desire to study and the inability of state institutions to meet that need (Misztal 2000, Wasielewski 2013). From 1990 to 2010 the number of higher educational institutions in Poland increased almost fivefold (similarly to the growth in the number of students; Fig. 1).

The educational boom of that period naturally led to an increase in numbers of graduates. The net index of student numbers[3] (among persons aged 19–24) increased from 9.8% in 1990 to 40.8% in 2010, which is in agreement with the growth in the number of students (Fig. 2). Since 2005 the index has remained above 38%, one of the highest values among the countries of the European Union (GUS 2009, 2012, 2015).

[3] On the GUS website, the net index of student numbers is defined as the ratio of the number of persons (in a given age group) in a given level of education at the start of the school year to the total population (at 31 December) in the age group corresponding to that level of education. For example, the net index for the primary school level is calculated by dividing the number of primary school pupils aged 7–12 (the age assigned to that level) at the start of a given school year by the total population aged 7–12 at 31 December of the same year. The result is given as a percentage.

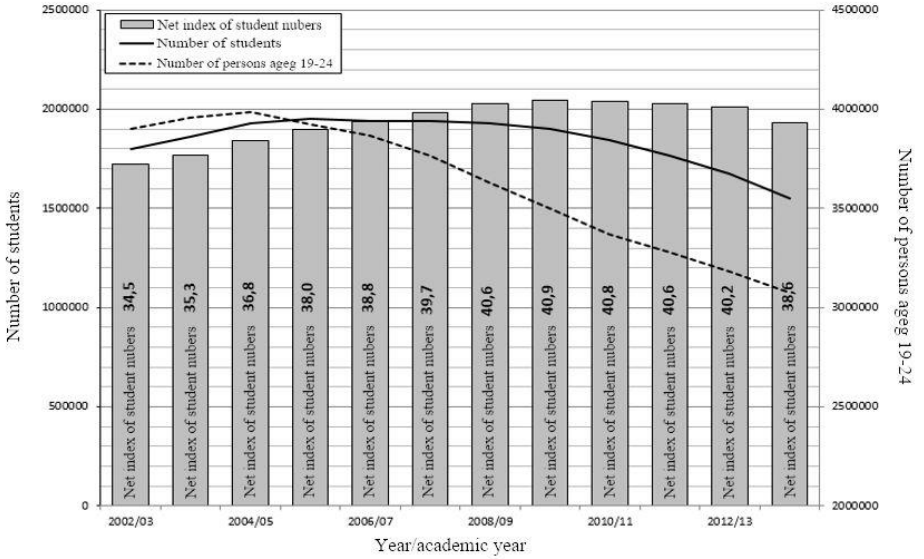


Figure 2. Number of students, number of persons aged 19–24, net index of student numbers, 2002–2013

Source: based on Higher Education 2014, GUS 2015.

The 1990s, which introduced democratic norms into social and political life, and free market values into the economy, brought to light social inequalities (not previously noticed in the post-war period) and the phenomena causing them (including unemployment, educational levels inadequate to the needs of the labour market, physical disability, and other factors). The period saw a growth in the importance of statutory measures aimed at counteracting social exclusion. One such decision was the *Act on vocational colleges* (Dz.U. 1997 no. 96 item 590), whose aims included enabling persons in difficult economic and life situations to undertake higher education. The establishment of national vocational colleges in smaller cities, bringing higher education closer to places distant from large academic centres, enabled the powerful development of local communities. These colleges were intended as a response to the needs of local labour markets, providing teaching oriented towards the professional and practical dimensions of academic subjects. The decision to find such institutions was in line with the international trend towards making higher education more widely available to the general population (Trow 1973).

The speed of changes in the Polish higher education system, whose scale was unprecedented anywhere in Europe, and the consequences of those changes for economic and social development in Poland, provided the motivation for the present work. The aim of the research is to identify and investigate the regional differentiation between Polish provinces in terms of features relating to higher education in the period of economic and social transformation. Observations were

made in doubly multivariate way presenting: (i) fundamental characteristics of higher education and (ii) their changes in the years 2002–2013.

The next section of the article will present the features relating to higher education that were selected as diagnostic variables. The third section will describe the research procedure and statistical methods used. Since the study was based on doubly multivariate data, a method of principal component analysis for data of that type was applied. Delimitation of provinces was performed using the dendrite method (Florek et al. 1951, Kruskal 1956, Prim 1957) applied to the obtained principal components. The results of the classification of Polish provinces obtained by the described algorithm are set out in the fourth section. The final conclusions are preceded by an analysis of the identified clusters.

2. Diagnostic variables

The study was carried out using a set of diagnostic variables that are cited in many reports assessing the state of higher education and in analyses of educational systems[4].

The first variable is the number of higher educational institutions per 10,000 population (X1). During the first decade following the fall of communism, approximately 200 non-public higher educational institutions were established in Poland. These had an impact on the availability and diversity of study courses offered within the provinces in which the institutions were located.

The second variable considered is the number of students per 1000 population (X2). The aforementioned rise in student numbers has led to variation between provinces in terms of features relating to higher education over the past decades. Changes in numbers of students have been closely linked to changes in the number of higher educational institutions (Fig. 1). One consequence of the rising number of students is an increase in the number of graduates per 1000 population, which is taken as variable X3 in the model.

The structure and size of teaching and academic staff affect the prestige enjoyed by higher educational institutions. This is reflected in the academic potential of the provinces in which those institutions are situated. The analysis included two values relating to staffing: the number of academic teachers per 10,000 population (X4) and the number of academic teachers with the title of professor per 10,000 population (X5).

Research activity and specialist education are represented by two variables: the number of post-graduate students per 10,000 population (X6) and the number of doctoral students per 10,000 population (X7).

The analysis of the spatial variation between provinces was based on the set of seven diagnostic variables presented above. The data used are taken from the

[4] For example: *Szkolnictwo Wyższe w Polsce 2013*, Ministry of Science and Higher Education; *Szkoły wyższe i ich finanse 2013*, GUS.

Local Data Bank (http://stat.gov.pl/bdl/), the original source being the annual reports of higher educational institutions[5]. Missing values were acquired from the Statistical Yearbooks of Provinces, published by the Central Statistical Office (GUS).

The available absolute figures were divided by the numbers of inhabitants of the relevant provinces. To ensure correctness of the analysis, zero unitarization was applied (Walesiak 2014).

3. Research procedure

The algorithm for spatial delimitation of provinces consisted of three stages: data normalisation, construction of principal components, and cluster analysis.

Data normalisation was performed using the method of zero unitarization (Walesiak 2014). The fact that all of the observed values are stimulants (having positive impact) meant that a single common normalisation formula could be used. The source values of the observed features were transformed according to (1):

$$z_{jp} = \frac{x_{jp} - min\{x_{jp}\}}{max\{x_{jp}\} - min\{x_{jp}\}} \tag{1}$$

where z_{jp} is the normalised value of the pth variable for the jth object, and x_{jp} is the value of the pth variable for the jth object. The method gives normal values of the observed features in the interval <0; 1>, reducing the effect of disproportions in these values on the principal component analysis carried out in the second stage.

The second stage of the procedure involved principal component analysis of doubly multivariate data for a covariance matrix with Kronecker product structure.

Let us assume that we have an n-element sample consisting of objects characterised by p statistical features measured at T different time points. Data of this type are called doubly multivariate. Let $\mathbf{X}_{jk}$ denote the column vector of measurements of p features on the jth object at the kth time point, j=1,2,…,n, k=1,2,…,T. Let $\mathbf{X}_j=(\mathbf{X}_{j1}, \mathbf{X}_{j2},\ldots, \mathbf{X}_{jT})$ be a $p \times T$ matrix, and $\mathbf{x}_j=\text{vec}(\mathbf{X}_j)$ be a pT-dimensional column vector of measurements of p features for the jth object at successive time points k, j=1,2,…,n.

[5] The reports are denoted in the GUS databases as follows:
F-01/s: Report on revenue, costs and financial results of higher educational institutions;
S-10: Report on higher education;
S-11: Report on material and social assistance to students and doctoral students;
S-12: Report on academic scholarships, post-graduate and doctoral studies and employment in higher educational institutions and scientific and research institutes.

We assume that $\mathbf{x}_j \sim N_{pT}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, $j=1, 2, \ldots, n$, where $\boldsymbol{\Omega}$ is a positive definite covariance matrix. Based on the estimator of the covariance matrix $\boldsymbol{\Omega}$ we construct the principal components (Hotelling 1933). The estimator of the matrix $\boldsymbol{\Omega}$ constructed from an n-element sample is positive definite with probability 1 if and only if $n>pT$ (e.g. Giri 1996). This condition implies a need to have a very large sample, which is not always possible. We, therefore, assume that the matrix $\boldsymbol{\Omega}$ has the structure of a Kronecker product (e.g. Gałecki 1994, Naik and Rao 2001, Roy and Khattree 2005, Krzyśko et al. 2011):

$$\boldsymbol{\Omega} = \mathbf{V} \otimes \boldsymbol{\Sigma}, \tag{2}$$

where $\mathbf{V}$ is the positive definite matrix of covariance between time points, with dimension $T\times T$, and $\boldsymbol{\Sigma}$ is the positive definite matrix of covariance between all statistical features, with dimension $p\times p$. When the matrix $\boldsymbol{\Omega}$ has this structure, its estimator is positive definite (with probability 1) if and only if $n>\max(p, T)$, which significantly weakens the condition on the size of the sample.

Bearing in mind that the matrix $\mathbf{V}$ represents variability over time, we may consider three models:

Model 1. We assume that the observations $\mathbf{x}_j$ are independent and that $\mathbf{x}_j \sim N_{pT}(\boldsymbol{\mu}, \mathbf{V} \otimes \boldsymbol{\Sigma})$, where $\mathbf{V}$ is a $T\times T$ positive definite matrix, $\boldsymbol{\Sigma}$ is a $p\times p$ positive definite matrix, and $n>\max(p, T)$. We do not impose any additional restrictions on $\mathbf{V}$.

Model 2. Weadopt the same assumptions as in Model 1, but also assume that the matrix $\mathbf{V}$ is completely symmetric, that is it has the form:

$$\mathbf{V} = -\frac{1}{1-\rho}\left[(1-\rho)\mathbf{I}_T + \rho \mathbf{1}_T \mathbf{1}_T^{\mathbf{T}}\right], \tag{3}$$

where ρ is the coefficient of correlation, and $\mathbf{1}_T$ is a T-dimensional column vector of ones.

Model 3. We adopt the same assumptions as in Model 1, but also assume that the matrix $\mathbf{V}$ has the structure of a first-order autoregression (Krzyśko et al. 2011), that is it has the form:

$$\mathbf{V} = \frac{1}{1-\rho^2}\left(\rho^{|r-s|}\right)_{r,s=1}^{T}, \tag{4}$$

where ρ is the coefficient of correlation.

In all three models the unknown parameters are estimated by the maximum likelihood method, solving appropriate systems of simultaneous equations iteratively until the selected "stop" criterion is attained (Srivastava et. al 2008, Krzyśko and Skorzybut 2009). We construct principal components based on the

matrix $\widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{V}} \otimes \widehat{\boldsymbol{\Sigma}}$ (Deręgowski and Krzyśko 2009). If n>max(p, T), then the matrix $\widehat{\mathbf{V}} \otimes \widehat{\boldsymbol{\Sigma}}$ is positive definite with probability **1**, and so all eigenvalues are real and positive. If $\alpha_1, \alpha_2, .. , \alpha_T$ are the eigenvalues of $\widehat{\mathbf{V}}$ and $\beta_1, \beta_2, \ldots , \beta_p$ are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$, the eigenvalues of $\widehat{\mathbf{V}} \otimes \widehat{\boldsymbol{\Sigma}}$ are pT numbers of the form $\alpha_r\beta_s$, where r=1, 2, … ,T, s=1, 2, … ,p. Based on the eigenvalues so defined, we construct the principal components of the matrix $\widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{V}} \otimes \widehat{\boldsymbol{\Sigma}}$.

The principal components constructed in this way were used in the cluster analysis that formed the last stage of the study procedure. A hierarchical algorithm was used, based on the Wrocław taxonomy (Florek et al. 1951), involving the construction of the shortest dendrite[6] over a set of n objects, based on a selected measure of dissimilarity (Euclidean distance in this case):

$$\rho(\mathrm{x}_u, \mathrm{x}_v) = \left((\mathrm{x}_u - \mathrm{x}_v)'(\mathrm{x}_u - \mathrm{x}_v)\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{p} (\mathrm{x}_{ui} - \mathrm{x}_{vi})^2\right)^{\frac{1}{2}}, \tag{5}$$

where: u,v=1, 2, …, n, i=1, 2, …, p.

In the shortest dendrite we determine the mean $\bar{\rho}$ and standard deviation s_ρ of the weights of all edges (distances between objects). The critical value, providing a criterion for the removal of edges from the dendrite, was taken to be the sum of $\bar{\rho}$ and s_ρ. The removal of edges whose weight exceeds the critical value leads to a division of the dendrite, and consequently to the separation of clusters.

4. Classification of provinces

Statistical analysis of the higher education data was performed in several stages (steps), with a different number of provinces considered each time. Each of the stages was based on the dendrite method, where the critical value was taken to be the mean length of an edge of the dendrite plus the standard deviation of the lengths. The method produced a division into six groups of provinces in four steps.

In the **first step**, the principal components were constructed for all 16 provinces, taking account of the three models for the structure of the matrix **V**. The goodness criterion was taken to be the index W, being the ratio of the sum of the variances of the first two principal components to the sum of the variances of all principal components, expressed as a percentage (Table 1). Model 2 was found to preserve the largest proportion of the variation of the data (73.55%).

[6] The shortest dendrite is the tree for which the sum of the weights on the edges is the smallest. The weights are taken as the distance between the tree nodes representing the studied objects. The shortest dendrite was constructed using Kruskal's algorithm (1956).

Table 1. The goodness criterion for the models in the first step

	The goodness criterion		
	Model 1	Model 2	Model 3
W index	63.27	73.55	63.4

Source: own calculations.

Figure 3 shows a projection of the provinces in the plane of the first two principal components, together with a dendrite over the points representing the provinces. The dotted line marks an edge longer than the critical value of 2.1591 (dendrite connections and edge lengths are given in Table 2).

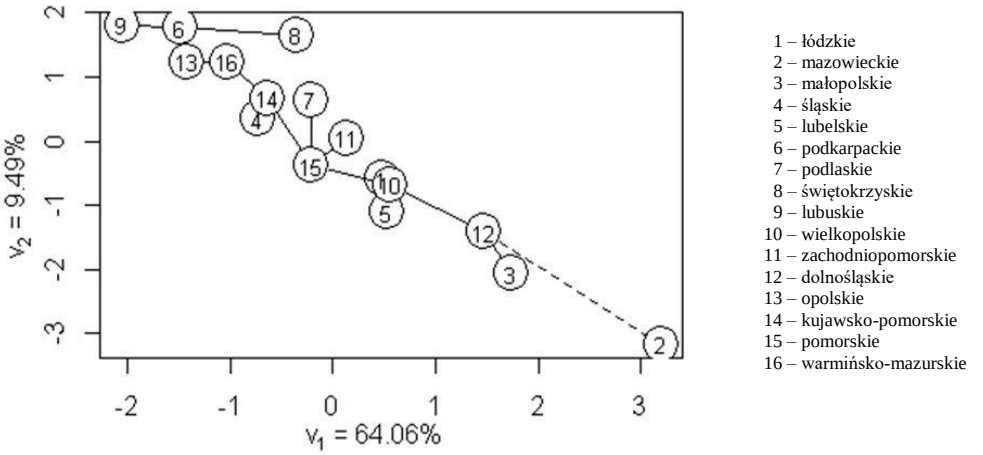


Figure 3. Dendrite over points representing provinces in the plane of the first two principal components, model 2 in the first step

Source: own calculations.

Table 2. Dendrite connections and edge lengths

Pairs of provinces	Edge length
1-10	0.8752
2-12	3.7373
3-12	1.7664
4-14	0.8875
5-10	1.1293
6-8	1.7859
6-9	1.0788
6-13	1.1570

Pairs of provinces	Edge length
7-15	1.4587
10-12	1.4477
10-15	1.2559
11-15	1.3666
13-16	1.2440
14-15	1.5957
14-16	1.3620

Source: own calculations.

The result of the analysis in the **first step** reveals the identification of two clusters, one of which is an isolated (single-element) cluster consisting of Mazowieckie province, further denoted ***Cluster 1***.

The remaining provinces, which make up the second cluster, underwent analysis in the **second step**. The greatest part of the variation, expressed by the index *W*, is preserved by the principal components in the second model (71.16%; Table 3).

Table 3. The goodness criterion for the models in the second step

	The goodness criterion		
	Model 1	Model 2	Model 3
W index	61.65	71.16	61.98

Source: own calculations.

Figure 4 shows a projection of the provinces in the plane of the first two principal components, together with the constructed dendrite. The edges that exceed the critical value (1.5989) are marked by a dotted line.

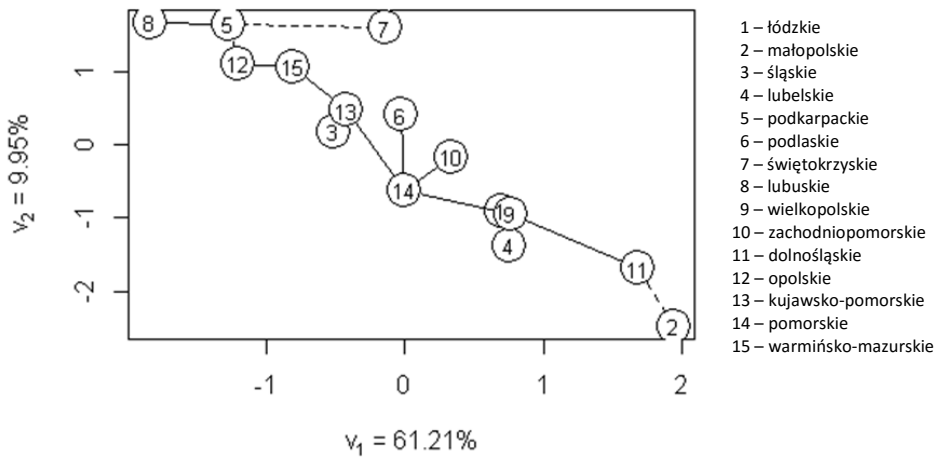


Figure 4. Dendrite over points representing provinces in the plane of the first two principal components, model 2 in the second step

Source: own calculations.

In the second step three clusters were identified, of which two are isolated clusters:

- ***Cluster 2*** consisting of Małopolskie province;
- ***Cluster 6*** consisting of Świętokrzyskie province.

In the **third step**, analysis was applied to the third cluster from the previous stage, consisting of 13 provinces: Śląskie, Podkarpackie, Lubuskie, Opolskie, Kujawsko-Pomorskie, Warmińsko-Mazurskie, Łódzkie, Lubelskie, Podlaskie, Wielkopolskie, Zachodniopomorskie, Dolnośląskie and Pomorskie.

The values of *W* for the three considered cases (Table 4) clearly show that Model 2 is again the most adequate to the data, explaining 67.15% of the variation.

Table 4. The goodness criterion for the models in the third step

	The goodness criterion		
	Model 1	Model 2	Model 3
W index	61.18	67.15	60.37

Source: own calculations.

Figure 5 shows a projection of the 13 provinces in the plane of the first two principal components, together with a dendrite constructed on the points representing them. The edges whose length exceed the critical value (1.462) are marked by dotted lines. This leads to a division into two clusters, consisting of the provinces:

- Łódzkie, Lubelskie, Podlaskie, Wielkopolskie, Zachodniopomorskie, Dolnośląskie, Pomorskie (denoted as ***Cluster 3***);
- Śląskie, Podkarpackie, Lubuskie, Opolskie, Kujawsko-Pomorskie, Warmińsko-Mazurskie (denoted temporarily as *Cluster 4*).

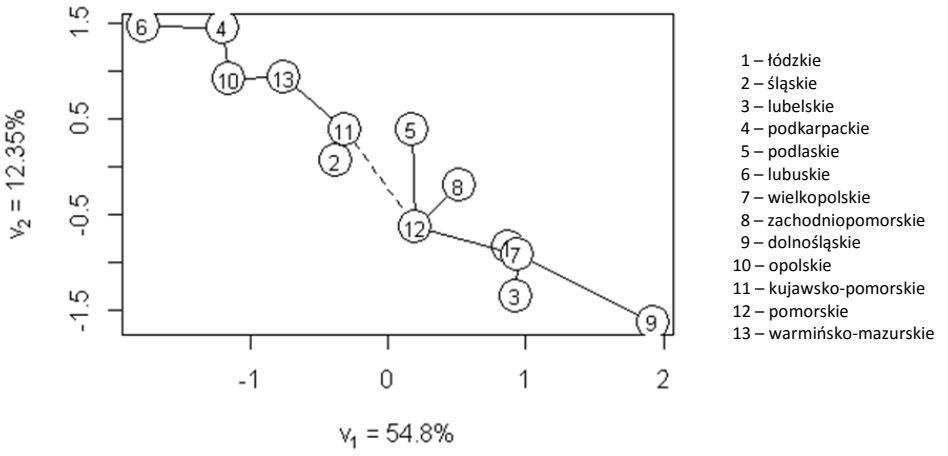


Figure 5. Dendrite over points representing provinces in the plane of the first two principal components, model 2 in the third step

Source: own calculations.

The **fourth step**, the final stage of the analysis, concerned Śląskie, Podkarpackie, Lubuskie, Opolskie, Kujawsko-Pomorskie and Warmińsko-Mazurskie provinces, contained in the temporary *Cluster 4*. The greatest part of the variation (67.75%) is preserved by the first two principal components in Model 2 (Table 5).

Table 5. The goodness criterion for the models in the fourth step

	The goodness criterion		
	Model 1	Model 2	Model 3
W index	55.76	67.75	57.88

Source: own calculations.

The critical edge length value (1.3246) in the dendrite was exceeded by the pairing of Kujawsko-Pomorskie and Warmińsko-Mazurskie provinces. This led to a division of the considered provinces into two clusters:

- ***Cluster 4*** consisting of Śląskie and Kujawsko-Pomorskie;
- ***Cluster 5*** consisting of the remaining provinces: Podkarpackie, Lubuskie, Opolskie and Warmińsko-Mazurskie.

The projection of the six provinces in the plane of the first two principal components, together with the constructed dendrite, is shown in Figure 6.

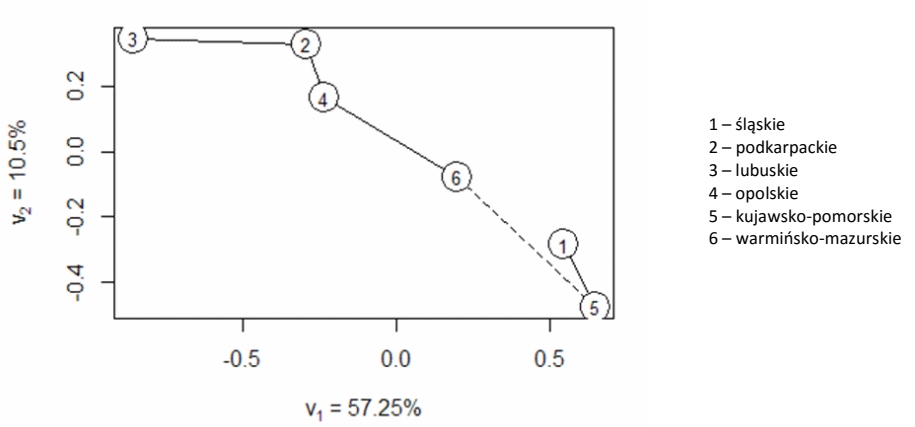


Figure 6. Dendrite over points representing provinces in the plane of the first two principal components, model 2 in the fourth step

Source: own calculations.

The provinces constituting ***Cluster 3*** were also analysed, but no basis was found for any further division of that cluster.

As a result of the four-stage classification process described above, the provinces were divided into a total of six groups (Fig. 7).

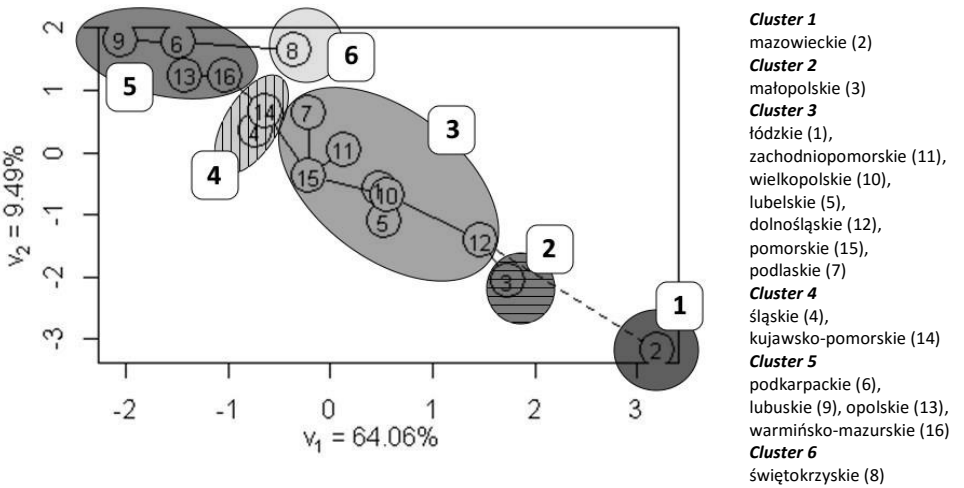


Figure 7. Dendrite for all 16 provinces, with clusters shown

Note: circles contain the numbers of the provinces belonging to the identified groups, and squares contain the numbers used to denote the clusters. In the legend, the numbers assigned to the provinces in the computational procedure are given in brackets.

Source: own calculations.

5. Analysis

In the first two steps of the algorithm, two single-element clusters were identified, consisting of Mazowieckie and Małopolskie provinces (Fig. 8). These regions contain the two largest and most renowned academic centres in Poland: Warsaw and Kraków. In national rankings, the University of Warsaw and Jagiellonian University are the two highest ranked higher educational institutions (*Perspektywy* ranking of higher educational institutions[7], *Polityka* ranking of higher educational institutions). They are also the only Polish institutions to appear on the Shanghai Ranking of the world's 500 leading universities (2015). It should also be noted that Warsaw and Kraków are the largest cities in Poland (in terms of population). As the national capital, Warsaw is also a financial, political and cultural centre. Mention should also be made of other higher educational institutions in these two provinces, which appear in the top ten of the aforementioned ranking: Warsaw University of Technology, the Warsaw School of Economics (SGH), and AGH University of Science and Technology in Kraków. Among non-public institutions offering master's degree courses, the leading ten (in the aforementioned ranking) include six institutions in Warsaw and one in Kraków. The concentration of so many leading institutions in those provinces explains their strong position in the higher education market, and is visible on the dendrite (Fig. 7) in the form of the large distance separating those regions from the remainder.

[7] *Perspektywy* ranking of higher educational institutions 2015.

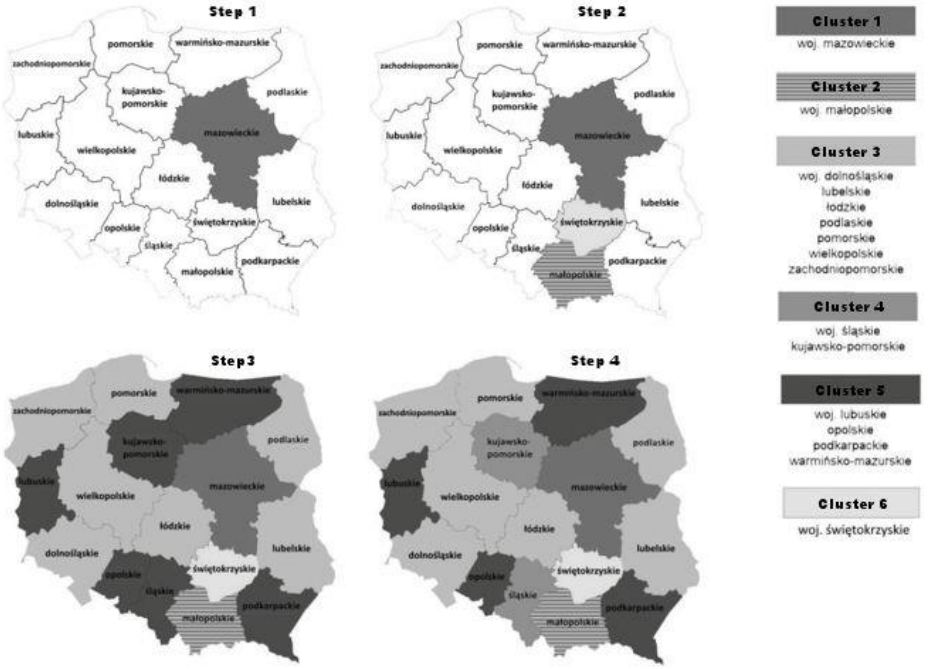


Figure 8. Spatial classification of provinces based on features relating to higher education (2002–2013)

Source: own analysis.

Cluster 3 consists of seven provinces containing higher educational institutions that are well-renowned within Poland and have a long-established tradition[8]. Most of the capitals of provinces in this cluster are among Poland's largest cities: Łódź, Wrocław, Poznań, Gdańsk (together with Gdynia and Sopot), Szczecin. The group also includes Lubelskie and Podlaskie provinces. These two eastern regions owe their membership of this cluster to the presence of higher educational institutions with notable values: the Catholic University of Lublin, Maria Curie-Skłodowska University in Lublin, the University of Białystok,

[8] The top 40 higher educational institutions in the 2015 *Perspektywy* ranking included, in order: Adam Mickiewicz University in Poznań, Wrocław University of Technology, the University of Wrocław, Gdańsk Medical University, Łódź University of Technology, Poznań Medical University, Poznań University of Technology, Wrocław Medical University, the University of Łódź, Gdańsk University of Technology, the University of Gdańsk, Łódź Medical University, Poznań University of Economics, Białystok Medical University, Lublin Medical University, Maria Curie-Skłodowska University in Lublin, the Pomeranian Medical University in Szczecin, Poznań University of Life Sciences, Wrocław University of Environmental and Life Sciences, the Catholic University of Lublin, Lublin University of Technology, and the University of Białystok.

Białystok Medical University, and the theological colleges in Białystok, Łomża and Drohiczyn.

In the course of the delimitation procedure, Cluster 4 was separated from Cluster 5. The dendrite (Fig. 7) shows the closeness of Śląskie and Kujawsko-Pomorskie provinces both to the group of provinces with the smallest potential (Cluster 5) and to the numerous group (Cluster 3), occupying the central part of the diagram.

Cluster 5 contains the Polish provinces with the lowest values of the analysed parameters. The cartogram reveals the peripheral nature of these regions, as well as their relative closeness to regions with higher potential.

The last of the identified clusters, consisting of Świętokrzyskie province, deviates from the axial arrangement of clusters seen on the dendrite. The values of some of the analysed higher education parameters were such as to place this province in the central group (a shift to the right on the horizontal axis of the dendrite) while others indicated that it belonged to the group with the smallest academic potential (a shift upwards on the vertical axis). This dual nature of observed values is well illustrated by the dendrite (Fig. 7).

The above analysis is complemented by a characterisation of the identified clusters in terms of descriptive statistics (Table 6).

Table 6. Descriptive statistics of the diagnostic variables for distinguished (identified) clusters

Variable	Clusters						Overall mean	Coefficient of variation between groups
	1	**2**	**3**	**4**	**5**	**6**		
	Mean value within group							
X1	19.60	9.84	11.60	9.21	6.75	11.20	11.37	38.59%
X2	63.44	60.12	47.33	40.02	34.76	38.59	47.38	25.31%
X3	14.12	12.11	10.84	9.66	8.68	10.80	11.03	17.28%
X4	3.15	3.66	2.69	2.04	1.58	1.43	2.43	36.75%
X5	0.85	0.75	0.59	0.47	0.38	0.39	0.57	34.01%
X6	10.35	4.14	3.37	2.67	2.45	2.86	4.31	70.12%
X7	1.71	1.49	0.87	0.58	0.23	0.08	0.82	80.09%

Meanings of variables: X1 – the number of higher educational institutions per 10,000 population; X2 – the number of students per 1000 population; X3 – the number of graduates per 1000 population; X4 – the number of academic teachers per 1000 population; X5 – the number of academic teachers with the title of professor per 10,000 population; X6 – the number of post-graduate students per 10,000 population; X7 – the number of doctoral students per 10,000 population.

Source: own calculations.

The mean values of the analysed features exhibit variation between the identified clusters. Mazowieckie province has the highest values for six out of the seven features. The single-element Cluster 2, consisting of Małopolskie province, has the highest number of academic teachers per 10,000 population, while in the other categories it lies second only to Mazowieckie province (often coming only slightly behind). The other clusters are separated from the leading two by a significant distance. The values recorded for Świętokrzyskie province clearly reveal its dual nature: the number of higher educational institutions per 10,000 population, the number of graduates per 1000 population and the number of post-graduate students per 10,000 population have values close to those for the high-potential clusters, while the values of number of academic teachers per 1000 population, the number of academic teachers with the title of professor per 10,000 population and the number of doctoral students per 10,000 population would place that province in the weakest group. It should be noted that, in terms of the values of the observed features, Cluster 4 differs from Cluster 3 (with higher potential) to a similar degree as from Cluster 5 (with lower potential), from which it was separated out.

The differentiation of the identified groups is greatest in the case of the variables representing numbers of post-graduate and doctoral students, for which the coefficients of variation are 70% and 80% respectively. The number of graduates per 1000 population, on the other hand, is relatively similar for all clusters, with a coefficient of variation not exceeding 20%.

The large disproportions in the values of variables between the two isolated clusters (Mazowieckie and Małopolskie provinces) and the other groups, the large group of provinces with moderate academic potential containing renowned centres of learning, and the isolated position of Świętokrzyskie province, deviating from the axial arrangement of the other clusters, create a characteristic picture of the spatial variation between Polish provinces based on the selected parameters relating to higher education.

6. Conclusions

The analysis has confirmed the dominance of Mazowieckie and Małopolskie provinces in the Polish higher education market. The higher educational institutions of these regions have been ranked the highest in national rankings for many years, as well as being Poland's only representatives in important international rankings. The applied delimitation model revealed relations between the provinces in terms of the analysed features. A detailed analysis of the results obtained and consideration of additional parameters relating to economic, demographic and social features would enable a better and more comprehensive presentation of the differences between the regions.

A wider-ranging analysis of the Polish regions, covering features relating to human capital and the quality of life, would appear to be a natural development of the research reported here, and will form a part of the author's future work.

REFERENCES

ANTONOWICZ, D., (2012). External influences and local responses. Changes in Polish higher education 1990–2005, in: KWIEK M., MAASSEN P., (eds.), National Education Reforms in a European Context. Comparative Reflections on Poland and Norway. Frankfurt: Peter Lang, pp. 87–111.

DERĘGOWSKI, K., KRZYŚKO, M., (2009). Principal components analysis in case of multivariate repeated measures data. Biometrical Letters, 46 (2), pp. 163–172.

FLOREK, K., ŁUKASZEWICZ, J., PERKAL, J., STEINHAUS, H., ZUBRZYCKI, S., (1951). Taksonomia wrocławska. Przegląd Antropologiczny, 17, pp. 193–211.

GAŁECKI, A. T., (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. Communications in Statistics – Theory and Methods, 23, pp. 3105–3119.

GIRI, N.C., (1996). Multivariate Statistical Analysis. New York: Marcel Dekker, Inc.

GOLINOWSKA, S., ed. (2005). Raport Społeczny Polska 2005. Fundacja im. Friedricha Eberta.

GUS, (2010, 2012, 2015). Rocznik Statystyki Międzynarodowej. Główny Urząd Statystyczny.

HOTELLING, H., (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24 (6), pp. 417–441.

KRUSKAL, J. B., (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of American Mathematical Society, 7, pp. 48–50.

KRZYŚKO, M., SKORZYBUT, M., (2009). Discriminant analysis of multivariate repeated measures data with a Kronecker product structured covariance matrices. Statistical Papers, 50, pp. 817–835.

KRZYŚKO, M., SKORZYBUT, M., WOŁYŃSKI, W., (2011), Classifiers for doubly multivariate data. Discussiones Mathematicae. Probability and Statistics, 31, pp. 5–27.

KWIEK, M., (2009). The two decades of privatization in Polish higher education. Cost-sharing, equity, and access. Sense Publishers.

KWIEK, M., (2014). Structural changes in the Polish higher education system (1990–2010): A synthetic view. European Journal of Higher Education, 4 (3), pp. 266–280.

MACH, B., (2003). Pokolenie historycznej nadziei i codziennego ryzyka: Społeczne losy osiemnastolatków z roku 1989. Wydawnictwo Instytutu Studiów Politycznych PAN.

MISZTAL, B., ed. (2000). Prywatyzacja szkolnictwa wyższego w Polsce: wyzwania w świetle transformacji systemowej. Kraków: Universitas.

NAIK, D. N., RAO, S., (2001). Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix. J. App. Statist. 28, pp. 91–105.

PRIM, R. C., (1957). Shortest connection networks and some generalizations. Bell System Technical Journal, 36, pp. 1389–1401.

ROY, A., KHATTREE R., (2005). On implementation of a test for Kronecker product covariance structure for multivariate repeated measures data. Statistical Methodology, 2, pp. 297–306.

SIKORSKA, J., (1998). Konsumpcja: Warunki, zróżnicowania, strategie. Wydawnictwo Instytutu Filozofii i Socjologii PAN.

SRIVASTAVA, M. S., von ROSEN, T., von ROSEN, D., (2008). Models with a Kronecker product covariance structure: estimation and testing. Mathematical Methods of Statistics, 17 (4), pp. 357–370.

TROW, M., (1973). Problems in the transition from elite to mass higher education. Carnegie Commission on Higher Education, Berkeley.

WALESIAK, M., (2014). Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej. Przegląd Statystyczny, 61 (4), pp. 363–372.

WASIELEWSKI, K., (2013). Caught in the trap of mass education – transformations in the Polish higher education after 1989, in: SZAFRANIEC K. and KONSTANTINOVSKIY D. (eds.). Polish and Russian Youth: Education and Work in Changing Society. IS RAS, Moskwa.

WNUK-LIPIŃSKA, E., (1996). Innowacyjność a konserwatyzm: uczelnie polskie w procesie przemian społecznych. Wydawnictwo Instytutu Studiów Politycznych PAN.

ZIÓŁKOWSKI, M., (2000). Przemiany interesów i wartości społeczeństwa polskiego. Teorie, tendencje, interpretacje. Wydawnictwo Fundacji Humaniora, Poznań.

STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 291–310, DOI 10. 21307

A THREE-PARAMETER WEIGHTED LINDLEY DISTRIBUTION AND ITS APPLICATIONS TO MODEL SURVIVAL TIME

Rama Shanker[1], **Kamlesh Kumar Shukla**[2], **Amarendra Mishra**[3]

ABSTRACT

In this paper a three-parameter weighted Lindley distribution, including Lindley distribution introduced by Lindley (1958), a two-parameter gamma distribution, a two-parameter weighted Lindley distribution introduced by Ghitany et al. (2011) and exponential distribution as special cases, has been suggested for modelling lifetime data from engineering and biomedical sciences. The structural properties of the distribution including moments, coefficient of variation, skewness, kurtosis and index of dispersion have been derived and discussed. The reliability properties, including hazard rate function and mean residual life function, have been discussed. The estimation of its parameters has been discussed using the maximum likelihood method and the applications of the distribution have been explained through some survival time data of a group of patients suffering from head and neck cancer, and the fit has been compared with a one-parameter Lindley distribution and a two-parameter weighted Lindley distribution.

Key words: moments, stochastic ordering, hazard rate function, mean residual life function, maximum likelihood estimation, lifetime data, goodness of fit.

1. Introduction

The probability density function (p.d.f.) of the two-parameter weighted Lindley distribution (WLD), introduced by Ghitany et al. (2011) with parameters α and θ, is given by

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

[1] Department of Statistics, Eritrea Institute of Technology, Asmara, Eritrea. E-mail: shnakerrama2009@gmail.com.
[2] Department of Statistics, Eritrea Institute of Technology, Asmara, Eritrea. E-mail: kkshukla22@gmail.com.
[3] Department of Statistics, Patna University, Patna, India. E-mail: mishraamar@rediffmail.com.

where

$$\Gamma(\alpha)=\int_0^{\infty} e^{-y}\, y^{\alpha-1}\, dy;\ \alpha>0$$

is the complete gamma function. Its structural properties including moments, hazard rate function, mean residual life function, estimation of parameters and applications to modelling survival time data have been discussed by Ghitany et al. (2011). The corresponding cumulative distribution function (c.d.f.) of WLD (1.1) can be obtained as

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

where

$$\Gamma(\alpha,z)=\int_z^{\infty} e^{-y}\, y^{\alpha-1}\, dy;\ \alpha>0,\ z\geq 0 \qquad (1.3)$$

is the upper incomplete gamma function.

It can be easily shown that at $\alpha=1$, WLD (1.1) reduces to Lindley (1958) distribution having p.d.f.

$$f(x;\theta)=\frac{\theta^2}{\theta+1}(1+x)e^{-\theta x}\quad ;x>0,\ \theta>0 \qquad (1.4)$$

It can be easily verified that the p.d.f. (1.4) is a two-component mixture of exponential (θ) and gamma $(2,\theta)$ distributions. Ghitany et al. (2008) have conducted a detailed study about various properties of Lindley distribution including skewness, kurtosis, hazard rate function, mean residual life function, stochastic ordering, stress-strength reliability, among other things; estimation of its parameter and application to model waiting time data in a bank. Shanker and Mishra (2013 a, 2013 b), Shanker and Amanuel (2013), and Shanker *et al.* (2013) have obtained different forms of the two-parameter Lindley distribution and discussed their various properties including skewness, kurtosis, index of dispersion, hazard rate function, mean residual life function, stochastic ordering, mean deviation, stress-strength reliability; estimation of parameters and their applications to model waiting and survival times data. Sankaran (1970) has obtained discrete Poisson-Lindley distribution by mixing Poisson distribution with Lindley (1958) distribution and studied its properties based on moments, estimation of parameters and applications to model count data from biological sciences. Shanker *et al.* (2015) have discussed a comparative study of Lindley and exponential distributions for modelling various lifetime data sets from biomedical science and engineering, and concluded that there are lifetime data where exponential distribution gives better fit than Lindley distribution and in majority of data sets Lindley distribution gives better fit than exponential distribution.

Further, p.d.f. (1.1) can also be expressed as a two-component mixture of gamma (α,θ) and gamma $(\alpha+1,\theta)$ distributions. We have

$$f(x;\theta,\alpha) = p\, f_1(x;\theta,\alpha) + (1-p)\, f_2(x;\theta,\alpha+1), \tag{1.5}$$

where

$$p = \frac{\theta}{\theta+\alpha},\ f_1(x;\theta,\alpha) = \frac{\theta^{\alpha}}{\Gamma(\alpha)}\, e^{-\theta x} x^{\alpha-1}, \text{ and}$$

$$f_2(x;\theta,\alpha+1) = \frac{\theta^{\alpha+1}}{\Gamma(\alpha+1)}\, e^{-\theta x} x^{\alpha+1-1}.$$

Ghitany *et al.* (2011) have discussed the structural properties of WLD including the nature of its p.d.f., hazard rate function, mean residual life function and applications to survival data using maximum likelihood estimation. It has been shown by Ghitany *et al.* (2011) that the shapes of hazard rate function and mean residual life function are decreasing, increasing and bathtub and thus has the potential to model survival time data of different nature. Shanker *et al.* (2016) have discussed some of its important statistical and mathematical properties including central moments, coefficient of variation, skewness, kurtosis, index of dispersion, stochastic ordering and the applications to modelling lifetime data from engineering and biomedical sciences.

In the present paper, a three-parameter weighted Lindley distribution, which includes Lindley (1958) distribution, WLD introduced by Ghitany et al. (2011), two-parameter gamma distribution and exponential distribution as particular cases, has been proposed and discussed. Its moments about origin and central moments, coefficient of variation, skewness, kurtosis and index of dispersion have been derived. The hazard rate function and the mean residual life function of the distribution have been derived and their shapes have been discussed for varying values of the parameters. The estimation of its parameters has been discussed using maximum likelihood method. Finally, the goodness of fit and the applications of the distribution have been explained through some survival data and the fit has been compared with a one-parameter Lindley distribution and the two-parameter WLD.

2. A three-parameter weighted Lindley distribution

A three-parameter weighted Lindley distribution (TPWLD) having parameters θ, α, and β can be defined by the probability density function

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

where α and β are shape parameters and θ is a scale parameter.

It can be easily verified that the Lindley distribution introduced by Lindley (1958) and the two-parameter WLD introduced by Ghitany et al. (2011) are particular cases of (2.1) for $\alpha=\beta=1$ and $\beta=1$ respectively. A two-parameter gamma(α,θ) distribution is a particular case of TPWLD for $\beta\to\infty$. Again, for $\alpha=1$ and $\beta\to\infty$, TPWLD reduces to the one-parameter exponential distribution. Further, the p.d.f. (2.1) can be easily expressed as a two-component mixture of gamma (α,θ) and gamma $(\alpha+1,\theta)$ distributions. We have

$$f(x;\theta,\alpha,\beta)=p\,f_1(x;\theta,\alpha)+(1-p)\,f_2(x;\theta,\alpha+1), \qquad (2.2)$$

where

$$p=\frac{\beta\theta}{\beta\theta+\alpha},\quad f_1(x;\theta,\alpha)=\frac{\theta^{\alpha}}{\Gamma(\alpha)}\,e^{-\theta x}x^{\alpha-1},$$

$$f_2(x;\theta,\alpha+1)=\frac{\theta^{\alpha+1}}{\Gamma(\alpha+1)}\,e^{-\theta x}x^{\alpha+1-1}.$$

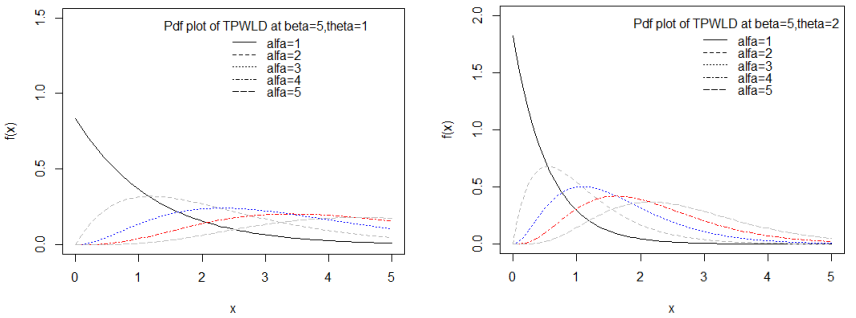
The corresponding cumulative distribution function of TPWLD (2.1) can be obtained as

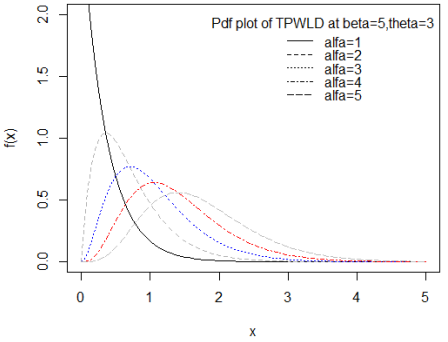
$$F(x;\theta,\alpha)=1-\frac{(\beta\theta+\alpha)\Gamma(\alpha,\theta x)+(\theta x)^{\alpha}e^{-\theta x}}{(\beta\theta+\alpha)\Gamma(\alpha)};$$

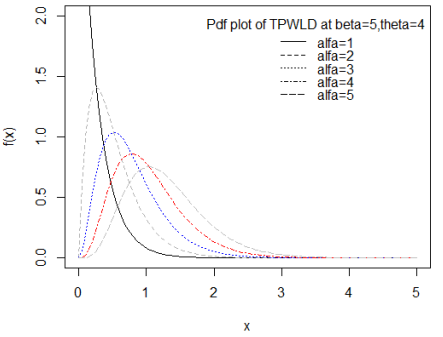
$$x>0,\theta>0,\alpha>0,\beta\theta+\alpha>0 \qquad (2.3)$$

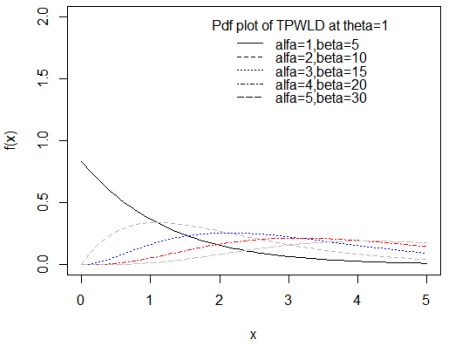
where $\Gamma(\alpha,z)$ is the upper incomplete gamma function defined in (1.3).

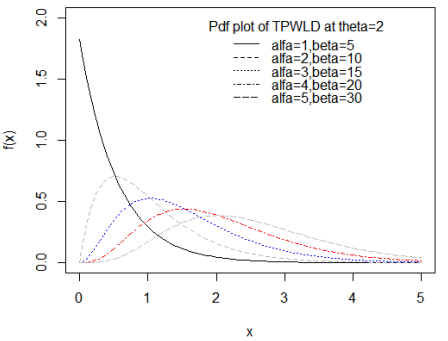
The nature of the p.d.f. of TPWLD for varying values of the parameters has been shown graphically in Figure 1.

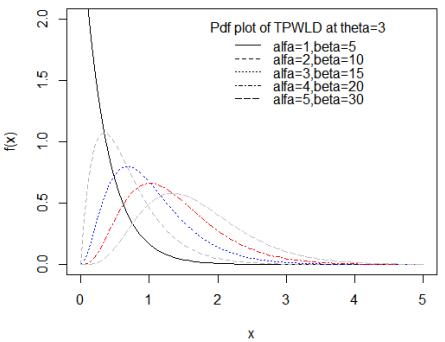


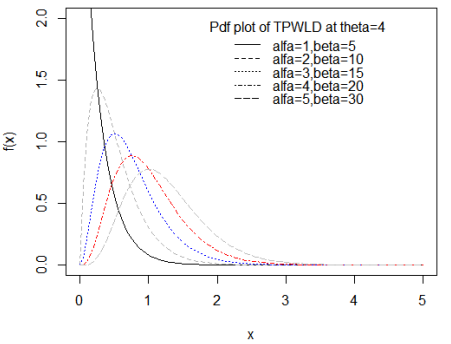

Pdf plot of TPWLD at beta=5,theta=3
alfa=1
alfa=2
alfa=3
alfa=4
alfa=5
f(x)
x

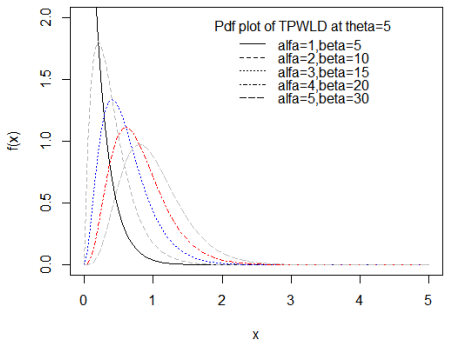

Pdf plot of TPWLD at beta=5,theta=4
alfa=1
alfa=2
alfa=3
alfa=4
alfa=5
f(x)
x

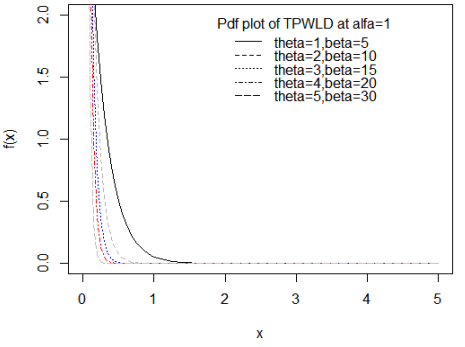

Pdf plot of TPWLD at theta=1
alfa=1,beta=5
alfa=2,beta=10
alfa=3,beta=15
alfa=4,beta=20
alfa=5,beta=30
f(x)
x

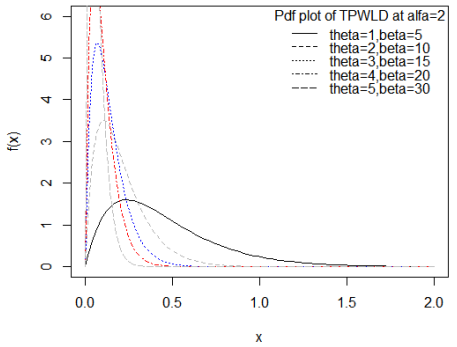

Pdf plot of TPWLD at theta=2
alfa=1,beta=5
alfa=2,beta=10
alfa=3,beta=15
alfa=4,beta=20
alfa=5,beta=30
f(x)
x

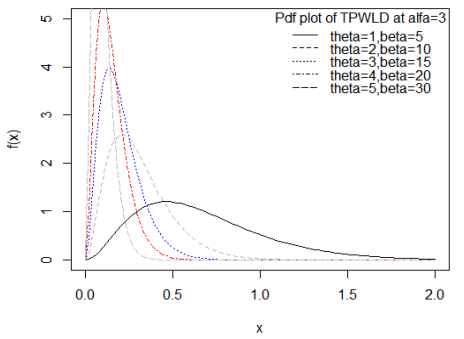

Pdf plot of TPWLD at theta=3
alfa=1,beta=5
alfa=2,beta=10
alfa=3,beta=15
alfa=4,beta=20
alfa=5,beta=30
f(x)
x

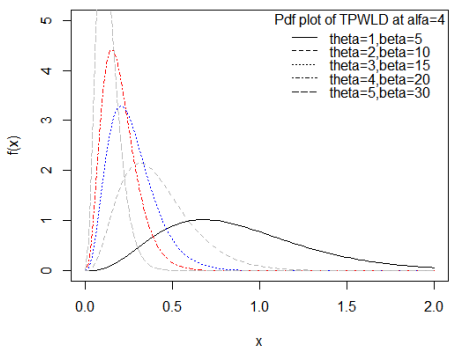

Pdf plot of TPWLD at theta=4
alfa=1,beta=5
alfa=2,beta=10
alfa=3,beta=15
alfa=4,beta=20
alfa=5,beta=30
f(x)
x

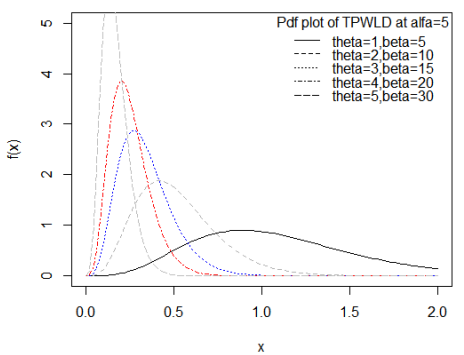

Pdf plot of TPWLD at theta=5
alfa=1,beta=5
alfa=2,beta=10
alfa=3,beta=15
alfa=4,beta=20
alfa=5,beta=30
f(x)
x


Pdf plot of TPWLD at alfa=1
theta=1,beta=5
theta=2,beta=10
theta=3,beta=15
theta=4,beta=20
theta=5,beta=30
f(x)
x


Pdf plot of TPWLD at alfa=2
theta=1,beta=5
theta=2,beta=10
theta=3,beta=15
theta=4,beta=20
theta=5,beta=30
f(x)
x


Pdf plot of TPWLD at alfa=3
theta=1,beta=5
theta=2,beta=10
theta=3,beta=15
theta=4,beta=20
theta=5,beta=30
f(x)
x


Pdf plot of TPWLD at alfa=4
theta=1,beta=5
theta=2,beta=10
theta=3,beta=15
theta=4,beta=20
theta=5,beta=30
f(x)
x


Pdf plot of TPWLD at alfa=5
theta=1,beta=5
theta=2,beta=10
theta=3,beta=15
theta=4,beta=20
theta=5,beta=30
f(x)
x

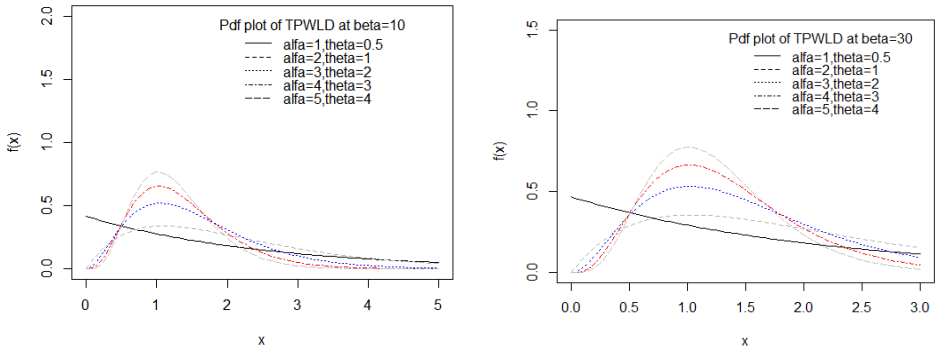


Figure 1. Nature of the p.d.f. of TPWLD for varying values of the parameters

3. Moments and related measures

Using the mixture representation (2.2), the r th moment about origin of TPWLD (2.1) can be obtained as

$$\mu_r' = E\left(X^r\right) = p\int_0^\infty x^r f_1\left(x;\theta,\alpha\right)dx + (1-p)\int_0^\infty x^r f_2\left(x;\theta,\alpha+1\right)dx$$

$$= \frac{\left[\beta\theta+\alpha+r\right]\Gamma\left(\alpha+r\right)}{\theta^r\left(\beta\theta+\alpha\right)\Gamma\left(\alpha\right)};\ r=1,2,3,\ldots \qquad (3.1)$$

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

$$\mu_1' = \frac{\alpha\left(\beta\theta+\alpha+1\right)}{\theta\left(\beta\theta+\alpha\right)}$$

$$\mu_2' = \frac{\alpha\left(\alpha+1\right)\left(\beta\theta+\alpha+2\right)}{\theta^2\left(\beta\theta+\alpha\right)}$$

$$\mu_3' = \frac{\alpha\left(\alpha+1\right)\left(\alpha+2\right)\left(\beta\theta+\alpha+3\right)}{\theta^3\left(\beta\theta+\alpha\right)}$$

$$\mu_4' = \frac{\alpha\left(\alpha+1\right)\left(\alpha+2\right)\left(\alpha+3\right)\left(\beta\theta+\alpha+4\right)}{\theta^4\left(\beta\theta+\alpha\right)}$$

It can be easily verified that these raw moments reduce to the corresponding raw moments of a two-parameter gamma distribution for $\beta\to\infty$.

Again, using the relationship between central moments and moments about origin, the central moments of TPWLD are obtained as

$$\mu_2 = \frac{\alpha\left[\beta^2\theta^2 + 2\beta\theta(\alpha+1) + \alpha(\alpha+1)\right]}{\theta^2(\beta\theta+\alpha)^2}$$

$$\mu_3 = \frac{2\alpha\left[\beta^3\theta^3 + 3\beta^2\theta^2(\alpha+1) + 3\beta\theta\alpha(\alpha+1) + \alpha^2(\alpha+1)\right]}{\theta^3(\beta\theta+\alpha)^3}$$

$$\mu_4 = \frac{3\alpha\left[\begin{array}{l}(\alpha+1)\beta^4\theta^4 + 4(\alpha^2+3\alpha+2)\beta^3\theta^3 + 2\alpha(3\alpha^2+11\alpha+8)\beta^2\theta^2 \\ +4\alpha^2(\alpha^2+4\alpha+3)\beta\theta + \alpha^3(\alpha^2+4\alpha+3)\end{array}\right]}{\theta^4(\beta\theta+\alpha)^4}$$

The expressions for coefficient of variation (C.V.), coefficient of skewness $\left(\sqrt{\beta_1}\right)$, coefficient of kurtosis (β_2), and index of dispersion (γ) of TPWLD (2.1) are thus obtained as

$$C.V. = \frac{\sigma}{\mu_1'} = \frac{\sqrt{\alpha\left[\beta^2\theta^2 + 2\beta\theta(\alpha+1) + \alpha(\alpha+1)\right]}}{\alpha(\beta\theta+\alpha+1)}$$

$$\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{2\alpha\left[\beta^3\theta^3 + 3\beta^2\theta^2(\alpha+1) + 3\beta\theta\alpha(\alpha+1) + \alpha^2(\alpha+1)\right]}{\left[\alpha\left\{\beta^2\theta^2 + 2\beta\theta(\alpha+1) + \alpha(\alpha+1)\right\}\right]^{3/2}}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\left[\begin{array}{l}(\alpha+1)\beta^4\theta^4 + 4(\alpha^2+3\alpha+2)\beta^3\theta^3 + 2\alpha(3\alpha^2+11\alpha+8)\beta^2\theta^2 \\ +4\alpha^2(\alpha^2+4\alpha+3)\beta\theta + \alpha^3(\alpha^2+4\alpha+3)\end{array}\right]}{\alpha\left\{\beta^2\theta^2 + 2\beta\theta(\alpha+1) + \alpha(\alpha+1)\right\}^2}$$

$$\gamma = \frac{\sigma^2}{\mu_1'} = \frac{\left[\beta^2\theta^2 + 2\beta\theta(\alpha+1) + \alpha(\alpha+1)\right]}{\theta(\beta\theta+\alpha)(\beta\theta+\alpha+1)}$$

The nature of the coefficient of variation, coefficient of skewness, coefficient of kurtosis and index of dispersion of TPWLD for varying values of the parameters have been shown in Figure 2.

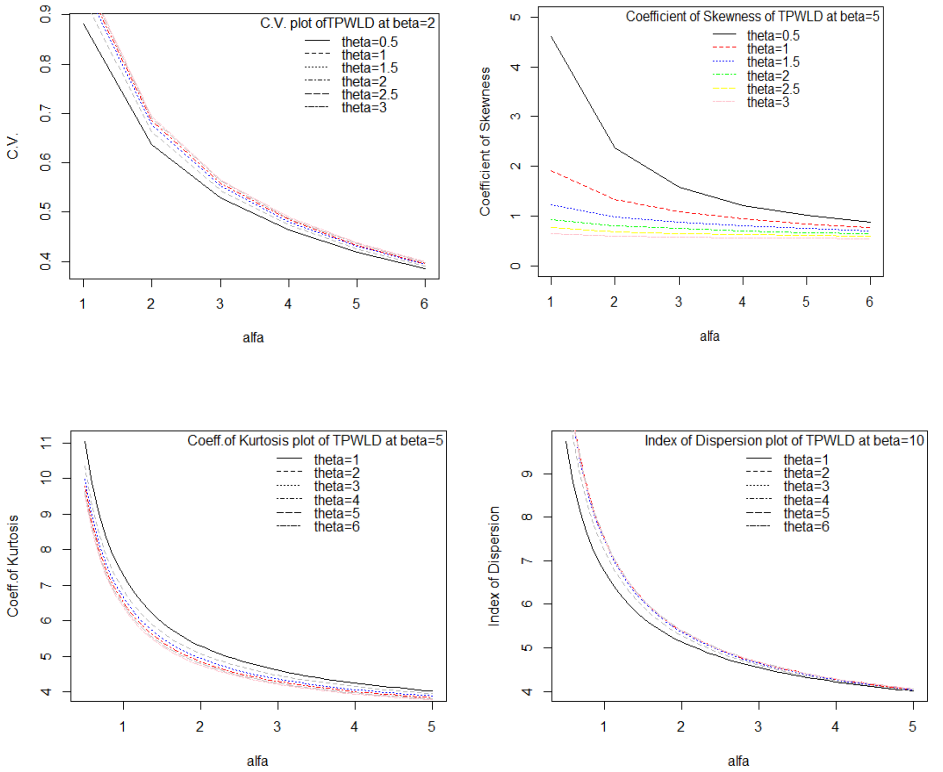


Figure 2. Graphs of coefficient of variation, coefficient of skewness, coefficient of kurtosis and index of dispersion of TPWLD for varying values of the parameters

4. Stochastic ordering

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

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

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

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

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

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

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

TPWLD is ordered with respect to the strongest 'likelihood ratio' ordering as shown in the following theorem:

Theorem: Let X ~ TPWLD $(\theta_1, \alpha_1, \beta_1)$ and Y ~ TPWLD$(\theta_2, \alpha_2, \beta_2)$. Then, the following results hold true

(i) If $\alpha_1 = \alpha_2, \beta_1 = \beta_2$ and $\theta_1 > \theta_2$, then $X \leq_{lr} Y$, $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$.

(ii) If $\alpha_1 < \alpha_2, \beta_1 = \beta_2$ and $\theta_1 = \theta_2$, then $X \leq_{lr} Y$, $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$.

(iii) If $\alpha_1 = \alpha_2, \beta_1 > \beta_2$ and $\theta_1 = \theta_2$, then $X \leq_{lr} Y$, $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$.

(iv) If $\alpha_1 < \alpha_2, \beta_1 > \beta_2$ and $\theta_1 > \theta_2$, then $X \leq_{lr} Y$, $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$.

Proof: We have

$$\frac{f_X(x)}{f_Y(x)} = \frac{\theta_1^{\alpha_1+1}(\beta_2\theta_2+\alpha_2)\Gamma(\alpha_2)}{\theta_2^{\alpha_2+1}(\beta_1\theta_1+\alpha_1)\Gamma(\alpha_1)} x^{\alpha_1-\alpha_2}\left(\frac{\beta_1+x}{\beta_2+x}\right)e^{-(\theta_1-\theta_2)x} \quad ; x>0$$

Now

$$\log\frac{f_X(x)}{f_Y(x)} = \log\left[\frac{\theta_1^{\alpha_1+1}(\beta_2\theta_2+\alpha_2)\Gamma(\alpha_2)}{\theta_2^{\alpha_2+1}(\beta_1\theta_1+\alpha_1)\Gamma(\alpha_1)}\right] + (\alpha_1-\alpha_2)\log x + \log\left(\frac{\beta_1+x}{\beta_2+x}\right) - (\theta_1-\theta_2)x.$$

This gives $$\frac{d}{dx}\log\frac{f_X(x)}{f_Y(x)} = \frac{\alpha_1-\alpha_2}{x} + \frac{\beta_2-\beta_1}{(\beta_1+x)(\beta_2+x)} - (\theta_1-\theta_2).$$

Thus for $\alpha_1 = \alpha_2, \beta_1 = \beta_2$ and $\theta_1 > \theta_2$, $\frac{d}{dx}\log\frac{f_X(x)}{f_Y(x)} < 0$. This means that $X \leq_{lr} Y$ and hence $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$, and thus (i) is verified. Similarly, (ii), (iii) and (iv) can easily be verified.

5. Hazard rate function and mean residual life function

5.1. Hazard rate function

Using the mixture representation (2.2), the survival (reliability) function of TPWLD can be obtained as

$$S(x)=P(X>x)=p\int_{x}^{\infty} f_1(y;\theta,\alpha)\,dy+(1-p)\int_{x}^{\infty} f_2(y;\theta,\alpha+1)\,dy$$

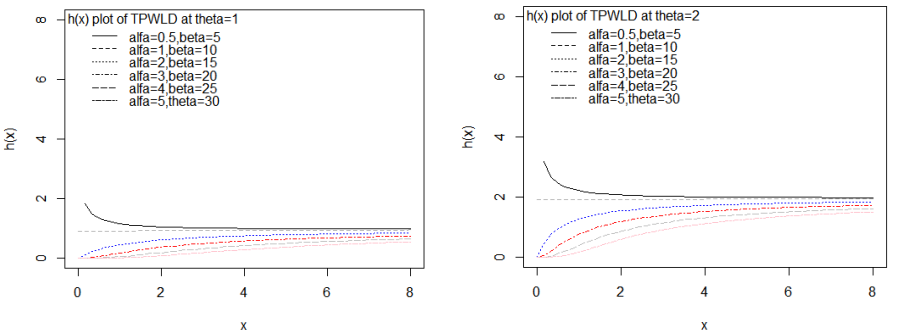
$$=\frac{(\beta\theta+\alpha)\Gamma(\alpha,\theta x)+(\theta x)^{\alpha}e^{-\theta x}}{(\beta\theta+\alpha)\Gamma(\alpha)} \qquad (5.1.1)$$

where $\Gamma(\alpha,z)$ is the upper incomplete gamma function defined in (1.3).

The hazard (or failure) rate function, $h(x)$ of TPWLD is thus obtained as

$$h(x)=\frac{f(x)}{S(x)}=\frac{\theta^{\alpha+1}x^{\alpha-1}(\beta+x)e^{-\theta x}}{(\beta\theta+\alpha)\Gamma(\alpha,\theta x)+(\theta x)^{\alpha}e^{-\theta x}};\ x>0,\theta>0,\alpha>0,\beta\theta+\alpha>0 \qquad (5.1.2)$$

The shapes of the hazard rate function, $h(x)$ of TPWLD for varying values of the parameters are shown in Figure 3.



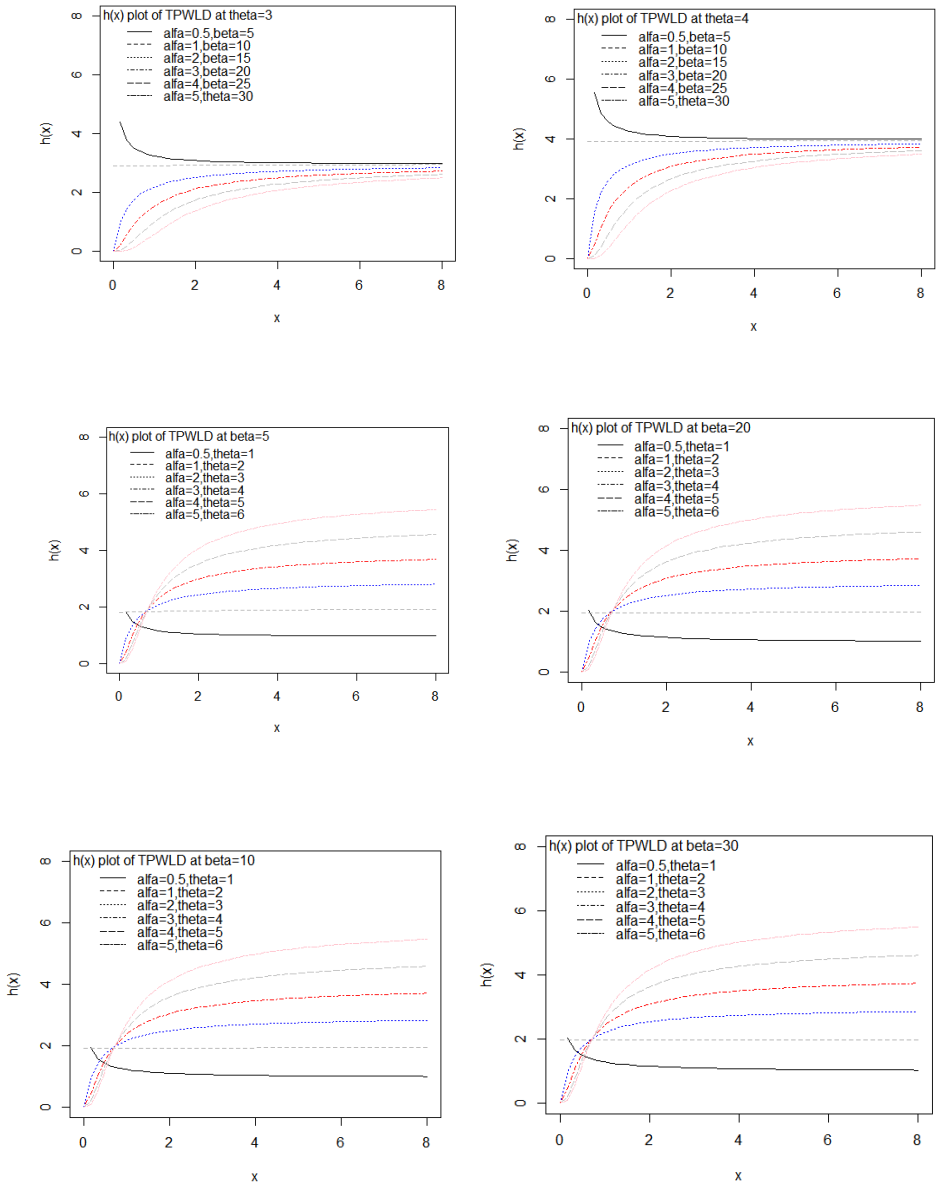


Figure 3. Graphs of the hazard rate function, $h(x)$ of TPWLD for varying values of the parameters

5.2. Mean residual life function

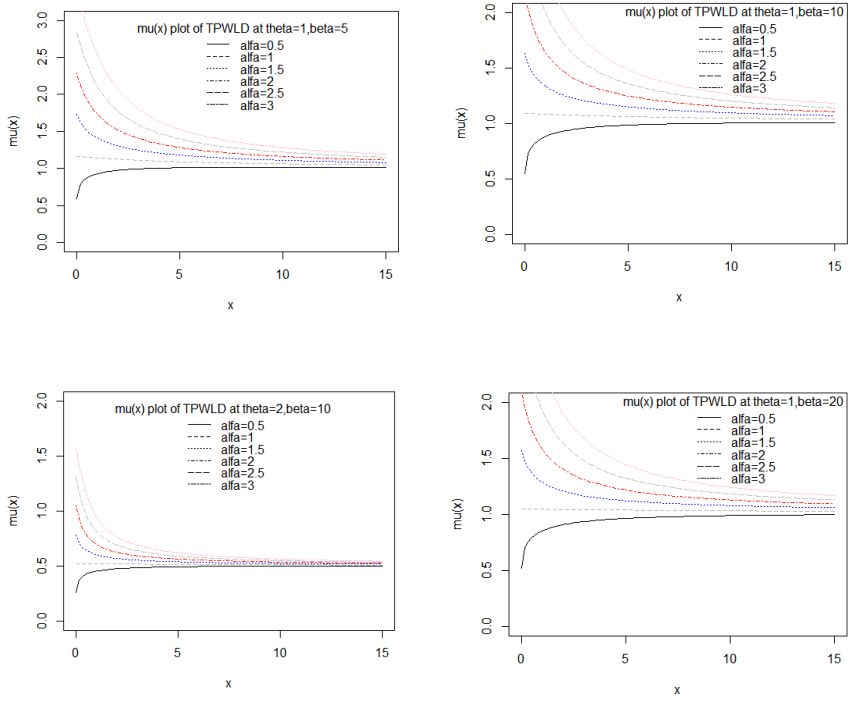
Using the mixture representation (2.2), the mean residual life function $m(x) = E(X - x \mid X > x)$ of TPWLD can be obtained as

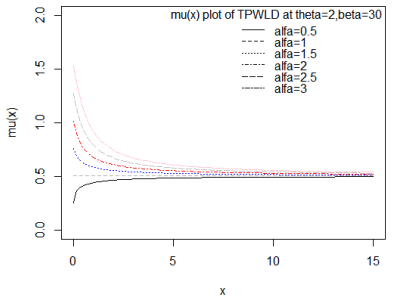
$$m(x) = \frac{1}{S(x)} \int_x^\infty y\, f(y)\, dy - x$$

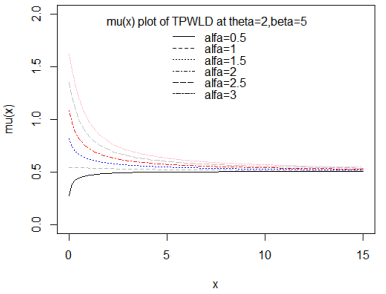
$$= \frac{1}{S(x)} \left[p \int_x^\infty y\, f_1(y;\theta,\alpha)\, dy + (1-p) \int_x^\infty y\, f_2(y;\theta,\alpha+1)\, dy \right] - x$$

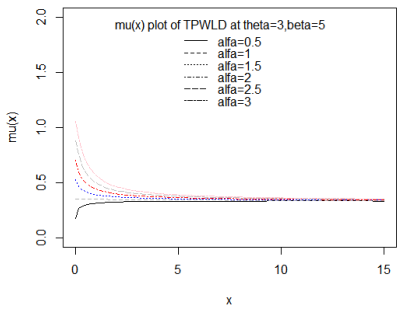
$$= \frac{\left[\alpha(\beta\theta+\alpha+1) - (\beta\theta+\alpha)(\theta x)\right]\Gamma(\alpha,\theta x) + (\beta\theta+\alpha+1)(\theta x)^\alpha e^{-\theta x}}{\theta\left[(\beta\theta+\alpha)\Gamma(\alpha,\theta x) + (\theta x)^\alpha e^{-\theta x}\right]}$$

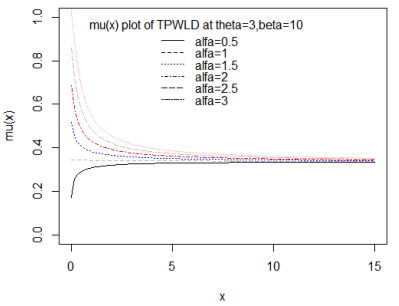
The shapes of the mean residual life function, $m(x)$ of TPWLD for varying values of the parameters are shown in Figure 4.

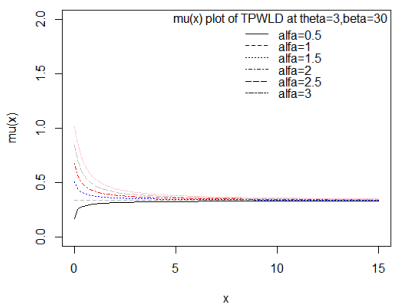


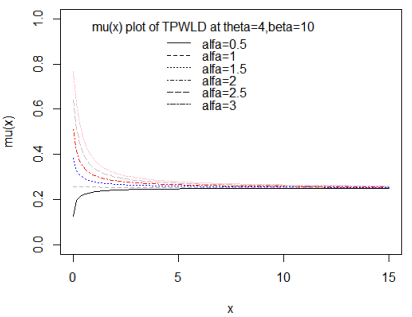

mu(x) plot of TPWLD at theta=2,beta=30
alfa=0.5
alfa=1
alfa=1.5
alfa=2
alfa=2.5
alfa=3
mu(x)
x


mu(x) plot of TPWLD at theta=2,beta=5
alfa=0.5
alfa=1
alfa=1.5
alfa=2
alfa=2.5
alfa=3
mu(x)
x


mu(x) plot of TPWLD at theta=3,beta=5
alfa=0.5
alfa=1
alfa=1.5
alfa=2
alfa=2.5
alfa=3
mu(x)
x


mu(x) plot of TPWLD at theta=3,beta=10
alfa=0.5
alfa=1
alfa=1.5
alfa=2
alfa=2.5
alfa=3
mu(x)
x


mu(x) plot of TPWLD at theta=3,beta=30
alfa=0.5
alfa=1
alfa=1.5
alfa=2
alfa=2.5
alfa=3
mu(x)
x


mu(x) plot of TPWLD at theta=4,beta=10
alfa=0.5
alfa=1
alfa=1.5
alfa=2
alfa=2.5
alfa=3
mu(x)
x

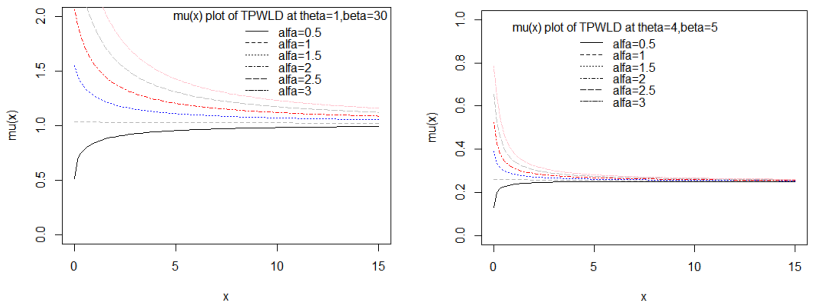


Figure 4. Graphs of the mean residual life function, $m(x)$ of TPWLD for varying values of the parameters

6. Maximum likelihood estimation

Let $(x_1, x_2, x_3, \ldots, x_n)$ be a random sample of size n from TPWLD (2.1). The likelihood function, L of TPWLD is given by

$$L = \left(\frac{\theta^{\alpha+1}}{\beta\theta+\alpha}\right)^n \frac{1}{\left(\Gamma(\alpha)\right)^n} \prod_{i=1}^{n} x_i^{\alpha-1}(\beta + x_i) e^{-n\theta\bar{x}}\ ;\ \bar{x} \text{ being the sample mean.}$$

The natural log likelihood function is thus obtained as

$$\ln L = n\left[(\alpha+1)\ln\theta - \ln(\beta\theta+\alpha) - \ln\left(\Gamma(\alpha)\right)\right] + (\alpha-1)\sum_{i=1}^{n}\ln(x_i) + \sum_{i=1}^{n}\ln(\beta+x_i) - n\theta\bar{x}$$

The maximum likelihood estimates (MLEs) $(\hat{\theta}, \hat{\alpha}, \hat{\beta})$ of parameters (θ, α, β) of TPWLD are the solution of the following nonlinear equations

$$\frac{\partial \ln L}{\partial \theta} = \frac{n(\alpha+1)}{\theta} - \frac{n\beta}{\beta\theta+\alpha} - n\bar{x} = 0$$

$$\frac{\partial \ln L}{\partial \alpha} = n\ln\theta - \frac{n}{\beta\theta+\alpha} - n\psi(\alpha) + \sum_{i=1}^{n}\ln(x_i) = 0$$

$$\frac{\partial \ln L}{\partial \beta} = \frac{-n\theta}{\beta\theta+\alpha} + \sum_{i=1}^{n}\frac{1}{\beta+x_i} = 0$$

where $\bar{x}$ is the sample mean and $\psi(\alpha)=\frac{d}{d\alpha}\ln\Gamma(\alpha)$ is the digamma function. It should be noted that the first equation gives the expression for the mean as $\bar{x}=\frac{\alpha(\beta\theta+\alpha+1)}{\theta(\beta\theta+\alpha)}$.

These three log likelihood equations do not seem to be solved directly. However, Fisher's scoring method can be applied to solve these equations. We have

$$\frac{\partial^2 \ln L}{\partial\theta^2}=\frac{-n(\alpha+1)}{\theta^2}+\frac{n\beta^2}{(\beta\theta+\alpha)^2}$$

$$\frac{\partial^2 \ln L}{\partial\alpha^2}=\frac{n}{(\beta\theta+\alpha)^2}-n\psi'(\alpha)$$

$$\frac{\partial^2 \ln L}{\partial\beta^2}=\frac{n\theta^2}{(\beta\theta+\alpha)^2}-\sum_{i=1}^{n}\frac{1}{(\beta+x_i)^2}$$

$$\frac{\partial^2 \ln L}{\partial\theta\,\partial\alpha}=\frac{n}{\theta}+\frac{n\beta}{(\beta\theta+\alpha)^2}$$

$$\frac{\partial^2 \ln L}{\partial\theta\,\partial\beta}=\frac{-n\alpha}{(\beta\theta+\alpha)^2}$$

$$\frac{\partial^2 \ln L}{\partial\alpha\,\partial\beta}=\frac{n\theta}{(\beta\theta+\alpha)^2}$$

where $\psi'(\alpha)=\frac{d}{d\alpha}\psi(\alpha)$ is the tri-gamma function.

For the maximum likelihood estimates $(\hat{\theta},\hat{\alpha},\hat{\beta})$ of (θ,α,β) of TPWLD (2.1), the following equations can be solved

$$\begin{bmatrix} \frac{\partial^2 \ln L}{\partial\theta^2} & \frac{\partial^2 \ln L}{\partial\theta\,\partial\alpha} & \frac{\partial^2 \ln L}{\partial\theta\,\partial\beta} \\ \frac{\partial^2 \ln L}{\partial\theta\,\partial\alpha} & \frac{\partial^2 \ln L}{\partial\alpha^2} & \frac{\partial^2 \ln L}{\partial\alpha\,\partial\beta} \\ \frac{\partial^2 \ln L}{\partial\theta\,\partial\beta} & \frac{\partial^2 \ln L}{\partial\alpha\,\partial\beta} & \frac{\partial^2 \ln L}{\partial\beta^2} \end{bmatrix}_{\substack{\hat{\theta}=\theta_0\\ \hat{\alpha}=\alpha_0\\ \hat{\beta}=\beta_0}} \begin{bmatrix} \hat{\theta}-\theta_0 \\ \hat{\alpha}-\alpha_0 \\ \hat{\beta}-\beta_0 \end{bmatrix} = \begin{bmatrix} \frac{\partial \ln L}{\partial\theta} \\ \frac{\partial \ln L}{\partial\alpha} \\ \frac{\partial \ln L}{\partial\beta} \end{bmatrix}$$

where $(\theta_0, \alpha_0, \beta_0)$ are the initial values of (θ, α, β) respectively. These equations are solved iteratively using any numerical iterative methods until sufficiently close values of $(\hat{\theta}, \hat{\alpha}, \hat{\beta})$ are obtained.

7. Applications and goodness of fit

In this section, the applications and goodness of fit of TPWLD has been discussed for several lifetime data and the fit is compared with a one-parameter Lindley distribution and a two-parameter WLD. In order to compare TPWLD, WLD and Lindley distribution, $-2\ln L$, AIC (Akaike information criterion), K-S Statistic (Kolmogorov-Smirnov Statistic) and p-value for two data sets have been computed and presented in Table 1. The formulae for AIC and K-S Statistics are as follows: $AIC = -2\log L + 2k$, and $K\text{-}S = \sup_x \left|F_n(x) - F_0(x)\right|$, where k is the number of parameters involved in the respective distributions, n is the sample size and $F_n(x)$ is the empirical distribution function. The best distribution corresponds to the lower values of $-2\ln L$, AIC and K-S statistic and higher p-value.

The goodness of fit of TPWLD, WLD, and Lindley distribution for data set 1 and 2 are based on maximum likelihood estimates (MLE). The data sets 1 and 2 are survival times of a group of patients suffering from head and neck cancer.

Data Set 1: The data set reported by Efron (1988) represent the survival times of a group of patients suffering from Head and Neck cancer disease and treated using radiotherapy (RT)

6.537 10.42 14.48 16.10 22.70 3441.55 4245.28 49.40 53.62 63
64 83 84 91 108 112 129 133 133 139 140 140 146 149 154
157 160 160 165 146 149 154 157 160 160 165 173 176 218 225
241 248 273 277 297 405 417 420 440 523 583 594 1101 1146
1417

Data Set 2: The data set reported by Efron (1988) represent the survival times of a group of patients suffering from Head and Neck cancer disease and treated using a combination of radiotherapy and chemotherapy (RT+CT).

12.20 23.56 23.74 25.87 31.98 37 41.35 47.38 55.46 58.36
63.47 68.46 78.26 74.47 81.43 84 92 94 110 112 119 127 130
133 140 146 155 159 173 179 194 195 209 249 281 319 339
432 469 519 633 725 817 1776

Table 1. MLE's, S.E., -2ln L, AIC, K-S Statistic and p-values of the fitted distributions of data sets 1-2

Data	Model	Parameter Estimate	S.E.	-2ln L	AIC	K-S Statistic	p-value
1	TPWLD	$\hat{\theta} = 0.0046940$ $\hat{\alpha} = 0.05090197$ $\hat{\beta} = -0.0913944$	0.000599 0.009192 0.178881	744.04	**750.04**	**0.172**	**0.064**
	WLD	$\hat{\theta} = 0.0052993$ $\hat{\alpha} = 0.2124576$	0.000797 0.113080	746.79	750.79	0.1826	0.042
	Lindley	$\hat{\theta} = 0.008804$	0.008060	763.74	765.74	0.246	0.002
2	TPWLD	$\hat{\theta} = 0.0047801$ $\hat{\alpha} = 0.0484017$ $\hat{\beta} = -0.077115$	0.0007015 0.0103219 0.1822874	563.45	**569.45**	**0.146**	**0.281**
	WLD	$\hat{\theta} = 0.0054135$ $\hat{\alpha} = 0.2271618$	0.0009513 0.1386034	565.93	569.93	0.161	0.185
	Lindley	$\hat{\theta} = 0.008905$	0.0009409	579.16	581.16	0.219	0.024

It is obvious from the goodness of fit in the above table that TPWLD gives much closer fit than the two-parameter WLD and the one-parameter Lindley distribution, and hence it can be considered as an important tool for modelling survival time data over these distributions.

The fitted plots of TPWLD, WLD and Lindley distribution for data set 1 and 2 are shown in the following Figure 4.

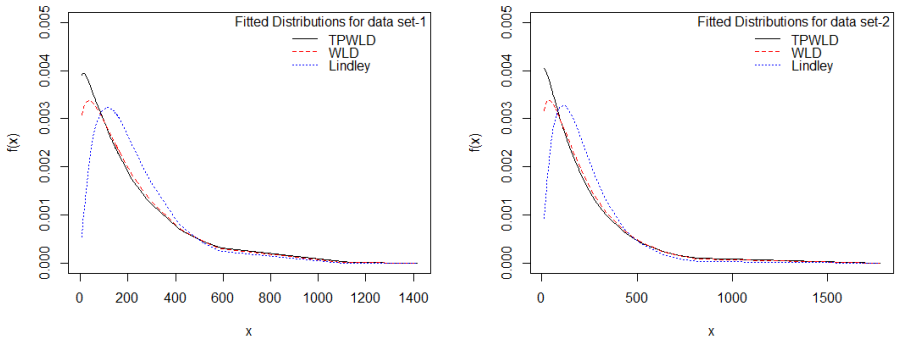


Figure 4. Fitted plots of TPWLD, WLD and Lindley distribution for the data set 1 and 2

8. Concluding remarks

A three-parameter weighted Lindley distribution (TPWLD), which includes two-parameter WLD and Lindley distribution as special cases, has been introduced. Its moments and moments-based expressions, including the coefficient of variation, skewness, kurtosis, and index of dispersion, have been derived and studied. The hazard rate function and the mean residual life function have been obtained and discussed. MLE has been used to estimate the parameters of the distribution. Goodness of fit of TPWLD has been discussed with some survival time data of a group of patients suffering from head and neck cancer and the fit shows a quite satisfactory fit over one-parameter Lindley distribution and two-parameter WLD.

Acknowledgement

Authors are grateful to the editor-in-chief of the journal and anonymous reviewers for their useful comments, which improved the presentation of the paper.

REFERENCES

EFRON, B., (1988). Logistic regression, survival analysis and the Kaplan-Meier curve, Journal of the American Statistical Association, 83, pp. 414–425.

GHITANY, M. E., ATIEH, B., NADARAJAH. S., (2008). Lindley distribution and its Application, Mathematics Computing and Simulation, 78, pp. 493–506.

GHITANY, M. E., ALQALLAF, F., AL-MUTAIRI, D. K., HUSAIN, H. A., (2011). A two-parameter weighted Lindley distribution and its applications to survival data, Mathematics and Computers in simulation, 81, pp. 1190–1201.

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

SANKARAN, M., (1970). The discrete Poisson-Lindley distribution, Biometrics, 26(1), pp. 145–149.

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

SHANKER, R., MISHRA, A., (2013 a). A Two Parameter Lindley Distribution, Statistics in Transition new series, 14 (1), pp. 45–56.

SHANKER, R, MISHRA, A., (2013 b). A Quasi Lindley Distribution, African Journal of Mathematics and Computer Science Research (AJMCSR), 6 (4), pp. 64 – 71.

SHANKER, R., SHARMA, S., SHANKER, R., (2013). A Two Parameter Lindley Distribution for Modeling Waiting and Survival Times Data, Applied Mathematics, 4 (2), pp. 363–368.

SHANKER, R., AMANUEL, A. G., (2013). A New Quasi Lindley Distribution**,** International Journal of Statistics and System, 8 (2), pp. 143–156.

SHANKER, R., HAGOS, F., SUJATHA, S., (2015). On modeling of Lifetimes data using exponential and Lindley distributions, Biometrics & Biostatistics International Journal, 2 (5), pp. 1–9.

SHANKER, R., SHUKLA, K. K., HAGOS, F., (2016). On weighted Lindley distribution and Its applications to model Lifetime data, Jacobs Journal of Biostatistics, 1 (1), pp. 1–9.

STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 311–328, DOI 10. 21307

EXAMINING TESTS FOR COMPARING SURVIVAL CURVES WITH RIGHT CENSORED DATA

Pinar Gunel Karadeniz[1], **Ilker Ercan**[2]

ABSTRACT

Background and objective: In survival analysis, estimating the survival probability of a population is important, but on the other hand, investigators want to compare the survival experiences of different groups. In such cases, the differences can be illustrated by drawing survival curves, but this will only give a rough idea. Since the data obtained from survival studies contains frequently censored observations some specially designed tests are required in order to compare groups statistically in terms of survival. **Methods:** In this study, Logrank, Gehan-Wilcoxon, Tarone-Ware, Peto-Peto, Modified Peto-Peto tests and tests belonging to Fleming-Harrington test family with (p, q) values; (1, 0), (0.5, 0.5), (1, 1), (0, 1) ve (0.5, 2) are examined by means of Type I error rate obtained from a simulation study, which is conducted in the cases where the event takes place with equal probability along the follow-up time. **Results:** As a result of the simulation study, Type I error rate of Logrank test is equal or close to the nominal value. **Conclusions:** When survival data were generated from lognormal and inverse Gaussian distribution, Type I error rate of Gehan-Wilcoxon, Tarone-Ware, Peto-Peto, Modified Peto-Peto and Fleming-Harrington (1,0) tests were close to the nominal value.

Key words: survival analysis, survival curves, comparison of survival curves, right censored observations.

1. Introduction

In survival analysis, investigators frequently want to determine if individuals from one population live longer than individuals from a second population. When all individuals in the population are observed, it is easy to estimate and compare the survival functions of different populations. However, most clinical research is completed in a prespecified time period, and it is not always possible to observe

[1] Uludag University, Faculty of Medicine, Department of Biostatistics Gorukle Campus Bursa/TURKEY 16059. E-mail: gunelpinar@yahoo.com.
[2] Uludag University, Faculty of Medicine, Department of Biostatistics Gorukle Campus Bursa/TURKEY 16059. E-mail: iercan@msn.com.

all individuals in this period until they experience the event. In such cases, censored data are encountered.

Since time-to-event data often include censored observations, some specialized methods are needed to compare the survival experiences of two groups (Dawson and Trapp, 2001). Several methods for testing the equality of two survival curves in the presence of censored data have been proposed.

First, Cox (1953) showed that an F test can be used to test the difference between two groups (whether or not the data include censored observations) when the survival data follow the exponential distribution. Then, the original (unweighted) log-rank test, which extended this test, was proposed by Mantel and Haenszel (1959) with Mantel (1966). Then, the generalized Wilcoxon tests, Gehan-Wilcoxon test (Gehan, 1965), the Peto-Peto test (Peto and Peto, 1972), and the Tarone-Ware test (Tarone and Ware, 1977) were proposed, respectively. Another useful subfamily within the class of weighted log-rank statistics is also proposed by Fleming and Harrington (1981) and Harrington and Fleming (1982).

There are studies in the literature that compare the performances of survival comparison tests. Lee et al. (1975) compared the size and power of the tests using small samples from the exponential and Weibull distributions with and without censoring. They arranged their simulation study with censoring rates and sample sizes of the groups being the same. Latta (1981) extended the simulations to include log-normal distributions, unequal sample sizes and censoring of only one group. Fleming et al. (1987) examined the properties of the tests based on linear rank statistics. Beltangady and Frankowski (1989) focused on the effect of unequal censoring by using various combinations of censoring proportions. Leton and Zuluaga (2001; 2005) compared the performance of various versions of generalized Wilcoxon and log-rank tests under scenarios of early and late hazard differences. Akbar and Pasha (2009) compared the performances of the log-rank and generalized Wilcoxon tests with low and high censoring rates for small and large sample sizes. Jurkiewicz and Wycinka (2011) compared the log-rank, Gehan-Wilcoxon, Tarone-Ware, Peto-Peto and F-H tests when the sample size is small.

Log-rank test is proposed in order to give equal weight to all failures among the follow-up (Lee and Wang, 2003). However, for the log-rank test there is an assumption that the hazard ratio of the groups should be proportional along the follow-up period (Fleming et al., 1987; Lee, 1996; Buyske et al., 2000). Only in this situation is the log-rank test powerful. When the hazard ratio is non-constant, the Gehan-Wilcoxon and Tarone-Ware tests can be more powerful than the log-rank test (Tarone and Ware, 1977; Pepe and Fleming, 1989). The Peto-Peto test is also efficient when proportional hazard assumption is violated (Kleinbaum and Klein, 2005). F-H tests, which are the most flexible tests for choosing weights, are focused on crossing the hazard ratios of groups (Pepe and Fleming, 1989).

The log-rank test, which compares outcomes over the whole time interval, may not adequately detect important differences between groups which occur either early or late in the interval (Klein et al., 2001). In some situations, a

treatment will decrease the hazard for some initial period, but its effect on the hazard becomes negligible later on (Pepe and Fleming, 1989). Therefore, the need to use tests that give more weight to early failures arises. In such cases, the Gehan-Wilcoxon and Tarone-Ware tests, which give more weight to the events that occur earlier, can be used. Likewise, the Peto-Peto and F-H (1,0) tests give more weight to early events as well.

When survival comparison tests are examined in the literature in terms of censoring, the Gehan-Wilcoxon test is powerful if the censoring rate is low (Stevenson, 2009; Martinez and Naranjo, 2010). Nevertheless, if the censoring rate is high, the Gehan-Wilcoxon test has less power. In addition, both the Gehan-Wilcoxon and the Peto-Peto tests have the assumption that censoring distributions of two groups should be same. When this assumption is violated, Efron stated that the Peto-Peto test has better performance than the Gehan-Wilcoxon test. For the log-rank test, it is more efficient when the censoring distribution of groups is different (Wang et al., 2010). This property is an advantage of the log-rank test over the others.

In this study, type I error rates were considered in examining the tests. Weibull, log-normal, exponential and inverse Gaussian distributions with different shape and scale parameters were used in order to generate survival times. The aim of this study is to examine the survival comparison tests in regard to type I error rates with right-censored data in some defined particular cases with events spread equally during the follow-up time.

2. Materials and Methods

2.1. Survival Comparison Tests

In survival analysis, estimating the survival probability of a population is important and investigators also want to compare the survival experiences of different groups. In such cases, the differences between groups can be illustrated by drawing survival curves obtained from the Kaplan-Meier (K-M) method, but this will only give a rough comparison and does not reveal whether the differences are statistically significant or not (Lee and Wang, 2003; Kim and Dailey, 2008).

When there are no censored observations, standard independent sample tests can be used to compare two survival distributions. However, in practice, censored data are frequently encountered. In such cases, in order to analyze the difference between two groups statistically, specially designed tests are used (Lee and Wang, 2003).

In this study, survival comparison tests (log-rank, Gehan-Wilcoxon, Tarone-Ware, Peto-Peto, Modified Peto-Peto, and Fleming-Harrington test family (with (p, q) values: (1, 0), (0.5, 0.5), (1, 1), (0, 1) and (0.5, 2), respectively), which are used to compare survival curves from two groups in cases of right-censored data, were compared in regard to type I error rates in the specific case that events

occurred at equal rates throughout the follow-up time and when the follow-up time fits some specific distributions. The nominal value was considered as 0.05 for type I error rates. When type I error rates were close to the nominal value the false positivity was close to the desired value so that the probability of making a wrong decision when there was not a real difference was at the desired value.

Suppose we have survival data as in Table 1. In order to obtain the general test statistic, which compares survival curves, Table 2 can be generated from Table 1.

Table 1. Sample survival data set

Individual (Patient)	Survival Time (t_j)	Status Variable (1: Event occurred 0: Censored observation)	Group
1	t_1	1	1
2	t_2	1	1
3	t_3	0	2
4	t_4	1	2
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
29	t_{29}	0	1
30	t_{30}	1	2

Table 2. Summary of observations at t_j time period

Group	1	2	Total
Number of events	d_{1j}	d_{2j}	d_j
Number of individuals at risk	r_{1j}	r_{2j}	r_j

Table 2 is generated repeatedly in all time periods in which the event of interest occurs (Bland and Altman, 2004; Kleinbaum and Klein, 2005). That is, by taking Table 1 as reference, at t_1, t_2, t_4, ..., t_{30} time periods in which the event of interest occurred, 2 by 2 tables are obtained. The observed and expected events in each group are considered from these tables. The general test statistic is obtained as the sum of the squared differences of the observed and expected counts scaled by the expected counts (Fisher and Belle, 1993; Klein et al., 2001). The test statistic is as in Equation 1 (Altman, 1991; Stevenson, 2009).

$$Test\ statistic = \frac{\left(\sum_1^k w_j\left(O_{ij} - E_{ij}\right)\right)^2}{var\left(\sum_1^k w_j\left(O_{ij} - E_{ij}\right)\right)} \quad (1)$$

Here,

i, denotes the group; j denotes the time that the event occurred,

O_{ij}, number of observed events in the i^{th} group at the j^{th} time period,

E_{ij}, number of expected events in the i^{th} group at the j^{th} time period.

O_{ij} and E_{ij} are computed as in Equation 2 and Equation 3, respectively (Leton and Zuluaga, 2005):

$$O_{ij} = \sum_{j=1}^{k} d_{ij} \quad (2)$$

$$E_{ij} = \sum_{j=1}^{k} d_j \frac{r_{ij}}{r_j} \quad (3)$$

When Equation 2 and Equation 3 are replaced in Equation 1, the general test statistic equals Equation 4 (Leton and Zuluaga, 2005):

$$Test\ statistic = \frac{\left(\sum_j w_j \left(d_{ij} - d_j \frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k} w_j^2 \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \quad (4)$$

In Equation 4;

d_{ij}, number of individuals who experience the event in group i at time j

d_j, total number of individuals in both groups who experience the event

r_{ij}, number of individuals at risk in group i at time j

r_j, total number of individuals at risk at time j

r_{1j}, number of individuals at risk in group *1*

r_{2j}, number of individuals at risk in group *2*.

The test statistic is compared to a chi-square table with 1 degree of freedom (Altman, 1991; Dawson and Trapp, 2001; Stevenson, 2009). The survival comparison tests are designated according to weight w_j, which is given in Equation 4.

Hypotheses for the survival comparison test are as below (Lee and Wang, 2003; Kleinbaum and Klein, 2005).

H_0: $S_1(t) = S_2(t)$ (survival probability of two groups is equal)

H_1: $S_1(t) \neq S_2(t)$ (survival probability of two groups is different) or

H_1: $S_1(t) < S_2(t)$ (survival probability of the first group is less than the survival probability of second group) or

H_1: $S_1(t) > S_2(t)$ (survival probability of the first group is greater than the survival probability of the second group)

2.1.1. Log-rank Test

The log-rank test, which is also known as the Mantel Log-rank Test, is the most commonly used test for comparing survival curves. It gives equal weight to early and late failures (Stevenson, 2009; Allison, 2010). The test statistic is based on the ranks of the time period in which the event occurred (Lee and Wang, 2003).

It takes $w_j=1$ as the weight in Equation 4. The test statistic turns into Equation 5:

$$Logrank\ test\ statistic = \frac{\left(\sum_j \left(d_{ij} - d_j \frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k} \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \quad (5)$$

The log-rank test assumes that the hazard functions for the two groups are parallel meaning that the hazard ratios of two groups are constant among all time periods (Dawson and Trapp, 2001; Stevenson, 2009).

Survival curves can be used to visualize whether the hazard functions of the two groups are parallel or not (Martinez and Naranjo, 2010).

2.1.2. Gehan Generalized Wilcoxon Test

The Gehan Generalized Wilcoxon Test is a distribution-free two-sample test and it is a generalization of the Wilcoxon test that samples right-censored observations (Gehan, 1965; Lee et al., 1975; Kim and Dailey, 2008).

The Gehan-Wilcoxon test uses the number of individuals at risk at time period t_j as the weight; thus, in Equation 4, $w_j=r_j$.

Since the weight is the number of individuals at risk, the Gehan-Wilcoxon test places more emphasis on the information at the beginning of the survival curve, where the number at risk is larger, allowing early failures to receive more weight than later failures (Tarone and Ware, 1977; Fisher and Belle, 1993; Kleinbaum and Klein, 2005).

The test statistic is as in Equation 6.

$$Gehan - Wilcoxon\ test\ statistic = \frac{\left(\sum_j r_j \left(d_{ij} - d_j \frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k} r_j^2 \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \tag{6}$$

In comparison with the log-rank test, the Gehan-Wilcoxon test does not have the assumption that the hazard functions of two groups are parallel making it a powerful test (Dawson and Trapp, 2001; Stevenson, 2009).

2.1.3. Tarone-Ware Test

The Tarone-Ware test places heavy weight on hazards in the early periods, just as the Gehan-Wilcoxon test does. It uses the square root of the number of individuals at risk as weight $w_j=\sqrt{r_j}$ (Tarone and Ware, 1977; Klein et al, 2001; Kleinbaum and Klein, 2005; Allison, 2010).

The weight used in the Tarone-Ware test is greater than the weight used in the log-rank test but less than the weight used in the Gehan-Wilcoxon test.

The Tarone-Ware test statistic is as in Equation 7.

$$Tarone - Ware\ test\ statistics = \frac{\left(\sum_j \sqrt{r_j} \left(d_{ij} - d_j \frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k} r_j \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \tag{7}$$

2.1.4. Peto-Peto Test

The Peto-Peto test assigns weights that depend on the estimated percentile of the failure time distribution. Failures occurring early, when the estimated survivor function is large, receive larger weights, while those in the right tail of the failure time distribution receive smaller weights (Prentice and Marek, 1979). This test is used when the hazard ratio between groups is not constant (Stevenson, 2009).

The Peto-Peto test uses the estimation of survival function as weight $w_j=\tilde{S}(t)$. The survival function here is a modified version of the K-M estimator (Allison, 2010). The test statistic is given in Equation 8.

$$Peto - Peto\ test\ statistic = \frac{\left(\sum_j \tilde{S}(t_j) \left(d_{ij} - d_j \frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k} \tilde{S}(t_j)^2 \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \tag{8}$$

Here,

$$\tilde{S}(t) = \prod_{t_j<t} \left(1 - \frac{d_j}{r_j + 1}\right) \tag{9}$$

2.1.5. Modified Peto-Peto Test

The Modified Peto-Peto test is an extension of the Peto-Peto test (Allison, 2010). It provides even greater weight to the early events as the Peto-Peto test (Hintze, 2007).

The modified Peto-Peto test uses survival function and the number of individuals at risk as weight $w_j=\tilde{S}(t_j)r_j/(r_j+1)$ (Hintze, 2007).

The test statistic is given in Equation 10.

$$Modified\ Peto-Peto\ test\ statistic$$

$$= \frac{\left(\sum_j \tilde{S}(t_j)r_j/(r_j+1)\left(d_{ij} - d_j\frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k}\left[\tilde{S}(t_j)r_j/(r_j+1)\right]^2 \frac{r_{1j}r_{2j}d_j(r_j-d_j)}{r_j^2(r_j-1)}} \tag{10}$$

2.1.6. Fleming-Harrington Test Family

Fleming-Harrington (F-H) test family comprises weighted log-rank tests. This family was designed in order to test the hypothesis of whether the survival curves of groups are equal or not equal, just as log-rank and other survival comparison tests do (Logan et al., 2008).

F-H tests use $w_j = \hat{S}(t_{j-1})^p \left[1 - \hat{S}(t_{j-1})\right]^q$ equality as weight when $p \geq 0$ and $q \geq 0$ (Oller and Gomez, 2010). Here, $\hat{S}(t)$ is an estimation of the Kaplan-Meier survival function. The test statistic is as below in Equation 11.

$$F-H\ test\ statistic$$

$$= \frac{\left(\sum_j \hat{S}(t_{j-1})^p \left[1 - \hat{S}(t_{j-1})\right]^q \left(d_{ij} - d_j\frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k}\left[\hat{S}(t_{j-1})^p \left[1 - \hat{S}(t_{j-1})\right]^q\right]^2 \frac{r_{1j}r_{2j}d_j(r_j-d_j)}{r_j^2(r_j-1)}} \tag{11}$$

Here, Kaplan-Meier survival function is obtained as follows:

$$\hat{S}(x) = \prod_{t_j\leq t} \left(1 - \frac{d_j}{r_j}\right) \tag{12}$$

In F-H tests, the choice of p and q determines what weight is given to middle and late occurring events (Gomez et al., 2009; Oller and Gomez, 2012). For example, if it is accepted that a treatment has an impact in the earlier periods, then

q is chosen as 0 with increasing values of p to ensure weight is given to earlier events. When p and q are equal, it ensures weight is given to events occurring in the middle of the whole time period. When p equals 0, increasing values of q ensure that more weight is placed on late events (Lee, 1996; Gomez et al., 2009). When p and q are both 0, the test is equivalent to the log-rank test. If p=1 and q=0, the test will be approximately equal to the Peto-Peto test (Harrington and Fleming, 1982). The choice of the weight function in F-H test must be made before evaluating the data and based on clinical expectations for the outcome (Klein et al., 2001; Gomez et al., 2009).

The summary of survival comparison tests and their weights are given in Table 3 (Kleinbaum and Klein, 2005; Jurkiewicz and Wycinka, 2011).

Table 3. Survival comparison tests and their weights

TEST	**WEIGHT (w_j)**	
LOGRANK	*1*	Equal weights throughout the whole time period
GEHAN-WILCOXON	r_j	Places very heavy weight on hazards at the beginning of the study
TARONE-WARE	$\sqrt{r_j}$	Places heavy weight on hazards at the beginning of the study
PETO-PETO	$\tilde{S}(t) = \prod_{t_j<t}\left(1 - \frac{d_j}{r_j + 1}\right)$	Places slightly more weight on hazards at the beginning of the study
MODIFIED PETO-PETO	$\tilde{S}(t) = \prod_{t_j<t}\left(1 - \frac{d_j}{r_j + 1}\right)$	Places slightly more weight on hazards at the beginning of the study
F-H (1,0)	$\hat{S}(t_{j-1})$	Places slightly more weight on hazards at the beginning of the study
F-H (0.5,0.5)	$\hat{S}(t_{j-1})^{0.5}\left[1 - \hat{S}(t_{j-1})\right]^{0.5}$	Places weight on hazards in the middle of the study
F-H (1,1)	$\hat{S}(t_{j-1})\left[1 - \hat{S}(t_{j-1})\right]$	Places weight on hazards in the middle of the study
F-H (0,1)	$1 - \hat{S}(t_{j-1})$	Places weight on hazards at the end of the study
F-H (0.5,2)	$\hat{S}(t_{j-1})^{0.5}\left[1 - \hat{S}(t_{j-1})\right]^{2}$	Places weight on hazards at the end of the study

3. Theory/calculation

3.1. Simulation Study

In this study, in order to examine survival comparison tests, a simulation study with 500 replicates was conducted, and type I error rates were obtained.

Survival times for two groups with sample sizes of n=10, 30, 50, and 100 were generated from the Weibull, log-normal, exponential and inverse Gaussian distributions with different shape and scale parameters. The status variable was generated from the binomial distribution with a probability of p=0.50.

While generating the survival data, other simulation studies in the literature were reviewed and most frequently used distributions with their most frequently used parameters were considered for our simulation study. Additionally, various parameters of the distributions were included. The reason for this choice is that in survival analysis follow-up time data fit generally the aforementioned distributions.

For the exponential distribution, the scale parameter was selected as β= 0.5, 1, 1.5; for the Weibull distribution, the shape parameter was α= 1, 2, 3 and the scale parameter was β= 1.5, 2.5, 3.5; for the log-normal distribution, the shape parameter was σ= 1, 2, 3 and the scale parameter was m= 0; for the inverse Gaussian distribution, the location parameter was μ= 0.5 and the scale parameter was λ= 1, 2, 3.

The data were generated using R software version 3.0.3 and the data were analyzed using NCSS package program with 500 replicates. Winautomation program is used for replicates.

4. Results

Type I error rates according to the simulation study for the sample sizes of n=10, 30, 50 and 100 are given in Table 4.

Table 4. Type I error rates of tests

Tests	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100
	Exponential (0.5)				Exponential (1)				Exponential (1.5)			
Logrank	0.0420	0.0560	0.0560	0.0520	0.0760	0.0420	0.0520	0.0620	0.0600	0.0720	0.0480	0.0500
Gehan-Wilcoxon	0.0360	0.0540	0.0620	0.0320	0.0600	0.0480	0.0460	0.0520	0.0560	0.0520	0.0500	0.0540
Tarone-Ware	0.0440	0.0560	0.0600	0.0460	0.0620	0.0460	0.0460	0.0520	0.0540	0.0680	0.0500	0.0520
Peto-Peto	0.0400	0.0560	0.0560	0.0440	0.0680	0.0440	0.0480	0.0520	0.0500	0.0620	0.0520	0.0520
Mod. Peto-Peto	0.0420	0.0560	0.0580	0.0440	0.0660	0.0460	0.0460	0.0520	0.0500	0.0600	0.0520	0.0520
F-H (1, 0)	0.0400	0.0540	0.0580	0.0440	0.0680	0.0440	0.0500	0.0520	0.0500	0.0640	0.0500	0.0520
F-H (0.5, 0.5)	0.0480	0.0460	0.0520	0.0620	0.0860	0.0360	0.0600	0.0640	0.0440	0.0500	0.0500	0.0480
F-H (1, 1)	0.0500	0.0480	0.0620	0.0560	0.0800	0.0420	0.0640	0.0620	0.0500	0.0480	0.0460	0.0420
F-H (0, 1)	0.0580	0.0760	0.0640	0.0620	0.0860	0.0580	0.0700	0.0680	0.0580	0.0620	0.0420	0.0540
F-H (0.5, 2)	0.0640	0.0860	0.0620	0.0680	0.0840	0.0620	0.0700	0.0640	0.0560	0.0800	0.0480	0.0500

Table 4. Type I error rates of tests (cont.)

Tests	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100
	Weibull (1, 1.5)				Weibull (1, 2.5)				Weibull (1, 3.5)			
Logrank	0.0440	0.0520	0.0440	0.0520	0.0620	0.0600	0.0520	0.0480	0.0620	0.0520	0.0620	0.0480
Gehan-Wilcoxon	0.0320	0.0480	0.0340	0.0520	0.0520	0.0580	0.0440	0.0500	0.0480	0.0440	0.0680	0.0400
Tarone-Ware	0.0340	0.0520	0.0280	0.0460	0.0540	0.0480	0.0420	0.0540	0.0560	0.0420	0.0760	0.0400
Peto-Peto	0.0340	0.0520	0.0280	0.0460	0.0580	0.0500	0.0440	0.0500	0.0500	0.0400	0.0720	0.0440
Mod. Peto-Peto	0.0340	0.0540	0.0260	0.0460	0.0580	0.0480	0.0440	0.0520	0.0520	0.0400	0.0760	0.0440
F-H (1, 0)	0.0320	0.0460	0.0280	0.0460	0.0560	0.0480	0.0440	0.0500	0.0540	0.0380	0.0720	0.0440
F-H (0.5, 0.5)	0.0560	0.0520	0.0560	0.0560	0.0800	0.0700	0.0540	0.0480	0.0640	0.0620	0.0480	0.0500
F-H (1, 1)	0.0660	0.0420	0.0560	0.0620	0.0820	0.0580	0.0560	0.0480	0.0700	0.0560	0.0480	0.0580
F-H (0, 1)	0.0740	0.0580	0.0580	0.0620	0.0860	0.0540	0.0660	0.0680	0.0760	0.0600	0.0400	0.0580
F-H (0.5, 2)	0.0740	0.0700	0.0600	0.0560	0.0920	0.0540	0.0580	0.0640	0.0740	0.0580	0.0460	0.0560
	Weibull (2, 1.5)				Weibull (2, 2.5)				Weibull (2, 3.5)			
Logrank	0.0660	0.0600	0.0600	0.0480	0.0480	0.0560	0.0500	0.0400	0.0540	0.0560	0.0440	0.0500
Gehan-Wilcoxon	0.0520	0.0500	0.0520	0.0420	0.0280	0.0520	0.0460	0.0400	0.0640	0.0700	0.0420	0.0580
Tarone-Ware	0.0640	0.0620	0.0640	0.0480	0.0380	0.0560	0.0380	0.0400	0.0580	0.0640	0.0480	0.0560
Peto-Peto	0.0640	0.0620	0.0620	0.0420	0.0400	0.0560	0.0400	0.0400	0.0580	0.0720	0.0480	0.0560
Mod. Peto-Peto	0.0620	0.0580	0.0620	0.0420	0.0360	0.0560	0.0400	0.0400	0.0560	0.0720	0.0480	0.0540
F-H (1, 0)	0.0620	0.0620	0.0620	0.0420	0.0380	0.0560	0.0400	0.0400	0.0580	0.0700	0.0480	0.0560
F-H (0.5, 0.5)	0.0760	0.0740	0.0540	0.0560	0.0480	0.0560	0.0480	0.0520	0.0600	0.0620	0.0580	0.0620
F-H (1, 1)	0.0880	0.0640	0.0580	0.0600	0.0460	0.0560	0.0520	0.0400	0.0660	0.0480	0.0580	0.0540
F-H (0, 1)	0.0980	0.0700	0.0660	0.0540	0.0540	0.0660	0.0680	0.0460	0.0760	0.0560	0.0620	0.0540
F-H (0.5, 2)	0.0920	0.0760	0.0620	0.0560	0.0560	0.0720	0.0680	0.0520	0.0820	0.0560	0.0580	0.0540
	Weibull (3, 1.5)				Weibull (3, 2.5)				Weibull (3, 3.5)			
Logrank	0.0540	0.0460	0.0540	0.0500	0.0500	0.0560	0.0600	0.0560	0.0480	0.0560	0.0440	0.0380
Gehan-Wilcoxon	0.0600	0.0440	0.0540	0.0520	0.0440	0.0400	0.0500	0.0500	0.0480	0.0520	0.0340	0.0440
Tarone-Ware	0.0580	0.0420	0.0500	0.0540	0.0380	0.0500	0.0360	0.0420	0.0500	0.0480	0.0280	0.0360
Peto-Peto	0.0580	0.0440	0.0540	0.0520	0.0400	0.0480	0.0400	0.0420	0.0480	0.0460	0.0340	0.0360
Mod. Peto-Peto	0.0580	0.0420	0.0500	0.0520	0.0420	0.0480	0.0400	0.0400	0.0480	0.0500	0.0340	0.0360
F-H (1, 0)	0.0580	0.0420	0.0540	0.0500	0.0400	0.0500	0.0400	0.0420	0.0440	0.0460	0.0340	0.0360
F-H (0.5, 0.5)	0.0540	0.0440	0.0580	0.0420	0.0580	0.0720	0.0600	0.0580	0.0580	0.0460	0.0520	0.0380
F-H (1, 1)	0.0580	0.0420	0.0600	0.0400	0.0680	0.0680	0.0500	0.0600	0.0660	0.0460	0.0400	0.0480
F-H (0, 1)	0.0640	0.0560	0.0580	0.0480	0.0820	0.0880	0.0580	0.0500	0.0800	0.0580	0.0580	0.0560
F-H (0.5, 2)	0.0740	0.0580	0.0540	0.0520	0.0820	0.0800	0.0560	0.0460	0.0700	0.0580	0.0680	0.0520

Table 4. Type I error rates of tests (cont.)

Tests	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100
	Lognormal (0, 1)				Lognormal (0, 2)				Lognormal (0, 3)			
Logrank	0.0580	0.0560	0.0540	0.0620	0.0700	0.0640	0.0640	0.0540	0.0520	0.0520	0.0440	0.0560
Gehan-Wilcoxon	0.0420	0.0500	0.0480	0.0520	0.0640	0.0460	0.0560	0.0520	0.0520	0.0480	0.0540	0.0400
Tarone-Ware	0.0500	0.0580	0.0540	0.0540	0.0700	0.0600	0.0620	0.0500	0.0460	0.0540	0.0480	0.0520
Peto-Peto	0.0480	0.0500	0.0520	0.0540	0.0680	0.0600	0.0600	0.0500	0.0480	0.0540	0.0460	0.0500
Mod. Peto-Peto	0.0500	0.0500	0.0500	0.0520	0.0700	0.0600	0.0620	0.0500	0.0480	0.0540	0.0440	0.0520
F-H (1, 0)	0.0440	0.0520	0.0480	0.0560	0.0680	0.0600	0.0620	0.0500	0.0520	0.0540	0.0480	0.0500
F-H (0.5, 0.5)	0.0600	0.0600	0.0540	0.0520	0.0540	0.0560	0.0600	0.0400	0.0520	0.0520	0.0480	0.0500
F-H (1, 1)	0.0580	0.0580	0.0540	0.0520	0.0620	0.0600	0.0540	0.0380	0.0600	0.0540	0.0500	0.0520
F-H (0, 1)	0.0640	0.0660	0.0720	0.0760	0.0720	0.0660	0.0480	0.0320	0.0700	0.0640	0.0500	0.0380
F-H (0.5, 2)	0.0720	0.0680	0.0680	0.0740	0.0700	0.0600	0.0460	0.0420	0.0680	0.0780	0.0620	0.0320
	Inverse Gaussian (0.5, 1)				Inverse Gaussian (0.5, 2)				Inverse Gaussian (0.5, 3)			
Logrank	0.0400	0.0560	0.0600	0.0560	0.0600	0.0460	0.0520	0.0480	0.0540	0.0740	0.0360	0.0380
Gehan-Wilcoxon	0.0420	0.0580	0.0500	0.0600	0.0440	0.0480	0.0360	0.0520	0.0580	0.0620	0.0400	0.0420
Tarone-Ware	0.0360	0.0580	0.0520	0.0520	0.0540	0.0520	0.0520	0.0500	0.0580	0.0660	0.0340	0.0400
Peto-Peto	0.0380	0.0540	0.0520	0.0560	0.0500	0.0500	0.0480	0.0500	0.0540	0.0680	0.0340	0.0400
Mod. Peto-Peto	0.0400	0.0520	0.0500	0.0540	0.0500	0.0540	0.0480	0.0480	0.0540	0.0680	0.0360	0.0400
F-H (1, 0)	0.0360	0.0520	0.0500	0.0560	0.0500	0.0500	0.0480	0.0480	0.0540	0.0680	0.0320	0.0400
F-H (0.5, 0.5)	0.0460	0.0480	0.0600	0.0580	0.0640	0.0500	0.0460	0.0420	0.0620	0.0640	0.0380	0.0300
F-H (1, 1)	0.0500	0.0560	0.0520	0.0600	0.0660	0.0500	0.0460	0.0400	0.0660	0.0560	0.0320	0.0320
F-H (0, 1)	0.0600	0.0420	0.0560	0.0460	0.0700	0.0580	0.0520	0.0500	0.0720	0.0680	0.0420	0.0500
F-H (0.5, 2)	0.0640	0.0480	0.0540	0.0460	0.0720	0.0640	0.0500	0.0520	0.0720	0.0680	0.0460	0.0460

In the case that the event occurs with equal probability along the follow-up time, the type I error rate of the log-rank test is equal or too close to the nominal value (0.05) for all distributions.

5. Discussion

In this study, a simulation was conducted in order to examine the performance of survival comparison tests under various scenarios, and the type I error rates were evaluated.

As a result, in the case that the event occurs with equal probability along the follow-up time, the type I error rate of the log-rank test is equal or too close to the nominal value. This result is in agreement with Lee and Wang (2003), who state that the "log-rank test gives equal weight to all failures." In addition, when the sample size gets larger, the type I error rate approaches the nominal value for all

tests. For the exponential distributions, the best results for all tests were obtained when the scale parameter was 1.5. When the scale parameter was 0.5, the best result was obtained for log-rank test; and the results farthest from the nominal value were obtained for the F-H tests, which give more weight to middle and late events (F-H (0.5,0.5), (1,1), (0,1), (0.5,2)). When the scale parameter was 1 for the exponential distribution, the closest type I error rates to the nominal value were obtained for the tests that give more weight to early events, namely, the Gehan-Wilcoxon, Tarone-Ware, Peto-Peto, Modified Peto-Peto and F-H (1,0). For F-H tests, which give more weight to middle and late events (F-H (0.5,0.5), (1,1), (0,1), (0.5,2)), the type I error rate tended to be greater than the nominal value when the scale parameters of the exponential distribution were 0.5 and 1. When survival data were generated from the Weibull distribution for all parameters of the distribution, the type I error rate of the log-rank test was equal or close to the nominal value. When the shape parameter of the Weibull distribution is 1, the type I error rates obtained were very close to the type I error rates of the exponential distribution. This result supports information found in the literature that a Weibull distribution with a shape parameter of 1 is equivalent to the exponential distribution (Kalbfleisch and Prentice, 2002). When the shape parameter of a Weibull distribution was 2, for almost all tests, the type I error rates were close to the nominal value. When the shape parameter of the Weibull distribution was 3 (which means the distribution is close to a normal distribution), the type I error rate for all tests were found to be close to 0.05. The error rate tended to be smaller than 0.05 only for a Weibull distribution with a shape parameter of 3 and a scale parameter of 3.5. In their study, Lee et al. (1975) demonstrated that if it is known that the survival data fit the exponential or Weibull distributions, the log-rank test has the best result; our simulation results further support this result.

When survival data were generated from the log-normal distribution, type I error rates of the Gehan-Wilcoxon and the Peto-Peto were equal or very close to the nominal value. In his study, Latta (1981) stated that the Gehan-Wilcoxon and the Peto-Peto tests best perform when the survival data fit log-normal distribution; therefore, our result agrees with Latta's result. The type I error rates for the Tarone-Ware, Modified Peto-Peto and F-H (1,0) tests were also close to the nominal value. For the log-normal distribution, the type I error rate of the log-rank test tended to be larger than the nominal value. The Gehan-Wilcoxon, Tarone-Ware and Peto-Peto tests showed suitable results in terms of type I error rate of an inverse Gaussian distribution that is similar to a log-normal distribution in its probability density function and hazard function.

In addition to all these results, it is stated in the literature that while comparing survival curves of two different groups, the hazard ratio should be examined. There have been several graphical methods for assessing the proportional hazards assumption (Martinez and Naranjo, 2010). If hazard ratios are parallel, the log-rank test is more efficient; if the hazard ratio of one group tends to differ more than the other as time progresses, the Tarone-Ware, Peto-Peto

and Gehan-Wilcoxon tests are more efficient (Peto and Peto, 1972; Lee et al., 1975; Harrington and Fleming, 1982). Furthermore, in the case that the hazard ratios of two groups cross, F-H tests are advantageous because the weight of the test may be specified accordingly.

Limitation of this study is that we exceedingly stick to the literature with regards to choosing distributions and their parameters. Although various distributions with various parameters were included in this study, it would be better to evaluate more distributions with more parameters in order to evaluate more different situations that are encountered in practice.

6. Conclusions

As a consequence, when making a choice of methods to compare survival curves, one must pay particular attention to the proportional hazards assumption, the proportion of censoring, the size of the sample under consideration and/or the distribution of the survival data. Besides, as mentioned in the discussion section in detail, when we encountered specific circumstances (specified distribution with specified parameter) that we indicate the type I error rate is close to nominal value, it is suggested to use the stated survival comparison tests.

Once these are taken into account, it is possible to make a more informed decision about the type of test that should be used to compare survival curves.

REFERENCES

AKBAR, A, PASHA, G. R., (2009). Properties of Kaplan-Meier estimator: group comparison of survival curves. European Journal of Scientific Research, 32 (3), pp. 391–397. https://www.researchgate.net/profile/Atif_Akbar2/publication/255648672_Properties_of_Kaplan-Meier_Estimator_Group_Comparison_of_Survival_Curves/links/549a82f00cf2b80371359dd2.pdf.

ALLISON, P. D., (2010). Survival analysis using SAS: a practical guide, 2nd edition, SAS Press, North Carolina.

ALTMAN, D. G., (1991). Practical statistics for medical research, Chapman&Hall, London.

BLAND, J. M., ALTMAN, D. G., (2004). The logrank test. British Medical Journal, 328, pp. 1073. http://www.bmj.com/content/328/7447/1073.long.

BELTANGADY, M. S., FRANKOWSKI, R. F., (1989). Effect of unequal censoring on the size and power of the logrank and Wilcoxon types of tests for survival data. Statistics in Medicine, 8 (8), pp. 937–945. https://www.ncbi.nlm.nih.gov/pubmed/2799123.

BUYSKE, S., FAGERSTROM, R., YING, Z., (2000). A class of weighted log-rank tests for survival data when the event is rare. Journal of the American Statistical Association, 95 (449), pp. 249–258. https://www.jstor.org/stable/2669542?seq=1#page_scan_tab_contents.

COX, D. R., (1953). Some simple approximate tests for poisson variates. Biometrika, 40, pp. 354–360. https://www.jstor.org/stable/2333353?seq=1#page_scan_tab_contents.

DAWSON, B., TRAPP, R. G., (2001). Basic&Clinical Biostatistics. Boston: McGraw Hill.

FISHER, L. D., BELLE, G. V., (1993). Biostatistics, a methodology for the health sciences, John Wiley&Sons Inc, New York.

FLEMING, T. R., HARRINGTON, D. P., (1981). A class of hypothesis tests for one and two samples censored survival data. Communications in Statistics, 10, pp. 763–794. http://www.tandfonline.com/doi/abs/10.1080/03610928108828073?journalCode=lsta20.

FLEMING, T. R., HARRINGTON, D. P., O'Sullivan, M., (1987). Supremum versions of the log-rank and generalized Wilcoxon statistics. Journal of the American Statistical Association, 82 (397), pp. 312–320. https://www.jstor.org/stable/2289169?seq=1#page_scan_tab_contents.

GEHAN, E. A., (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. Biometrika, 52, pp. 203–223. https://academic.oup.com/biomet/article-abstract/52/1-2/203/359447/A-generalized-Wilcoxon-test-for-comparing.

GOMEZ, G., CALLE, M. L., OLLER, R., LANGOHR, K., (2009). Tutorial on methods for interval-censored data and their implementation in R. Statistical Modelling, 9 (4), pp. 259–297. http://journals.sagepub.com/doi/abs/10.1177/1471082X0900900402.

HARRINGTON, DP., FLEMING, T. R., (1982). A class of rank test procedures for censored survival data. Biometrika, 69 (3), pp. 553–566. https://www.jstor.org/stable/2335991.

HEINZE, G., GNANT, M., SCHEMPER, M., (2003). Exact log-rank tests for unequal follow-up. Biometrics, 59, pp. 1151–1157. https://www.ncbi.nlm.nih.gov/pubmed/14969496.

HINTZE, J. L., (2007). NCSS user guide V tabulation, item analysis, proportions, diagnostic tests, and survival / reliability, Published by NCSS, Kaysville, Utah.

JURKIEWICZ, T., WYCINKA, E., (2011). Significance tests of differences between two crossing survival curves for small samples. Acta Universitatis Lodziensis Folia Oeconomica, 255, pp. 114. http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.hdl_11089_690?printView=true.

KALBFLEISCH, J. D., PRENTICE, R. L., (2002). The Statistical Analysis of Failure Time Data. New York: John Wiley&Sons Inc.

KIM, J., KANG, D. R., NAM, C. M., (2006). Logrank-type tests for comparing survival curves with interval-censored data. Computational Statistics & Data Analysis, 50 (11), pp. 3165–3178. http://www.sciencedirect.com/science/article/pii/S0167947305001441

KIM, J. S., DAILEY, R. J., (2008). Biostatistics for oral healthcare, Blackwell Publishing Company, Iowa, pp. 287–291.

KLEIN, J. P., RIZZO, J. D., ZHANG, M. J., KEIDING, N., (2001). Statistical methods for the analysis and presentation of the results of bone marrow transplants. Part I: Unadjusted analysis. Bone Marrow Transplantation, 28, pp. 909–915. https://www.ncbi.nlm.nih.gov/pubmed/11753543.

KLEINBAUM, D. G., KLEIN, M., (2005). Survival Analysis a Self-Learning Text. New York: Springer.

LATTA, R. B., (1981). A monte carlo study of some two-sample rank tests with censored data. Journal of American Statistical Association, 76 (375), pp. 713–719. https://www.jstor.org/stable/2287536?seq=1#page_scan_tab_contents .

LEE, E. T., DESU, M. M., GEHAN, E. A., (1975). A Monte Carlo study of the power of some two-sample tests. Biometrika, 62 (2), pp. 425–432. https://www.jstor.org/stable/2335383?seq=1#page_scan_tab_contents.

LEE, J. W., (1996). Some versatile tests based on the simultaneous use of weighted log-rank statistics. Biometrics, 52 (2), pp. 721–725. http://www.jstor.org/stable/2532911?seq=1#page_scan_tab_contents.

LEE, E. T., WANG, J. W., (2003). Statistical Methods for Survival Data Analysis. New Jersey: John Wiley&Sons Inc.

LETON, E., ZULUAGA, P., (2001). Equivalence between score and weighted tests for survival curves. Communications in Statistics - Theory and Methods, 30 (4), pp. 591–608. http://www.tandfonline.com/doi/abs/10.1081/STA-100002138.

LETON, E., ZULUAGA, P., (2005). Relationships among tests for censored data. Biometrical Journal, 47 (3), pp. 377–387. http://onlinelibrary.wiley.com/doi/10.1002/bimj.200410115/abstract.

LOGAN, B. R., KLEIN, J. P., ZHANG, M. J., (2008). Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. Biometrics, 64 (3), pp. 733–740. https://www.ncbi.nlm.nih.gov/pubmed/18190619.

MANTEL, N., HAENSZEL, W., (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of National Cancer Institute, 22 (4), pp. 719–748. https://www.ncbi.nlm.nih.gov/pubmed/13655060.

MANTEL, N., (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50 (3), pp. 163–170. https://www.ncbi.nlm.nih.gov/pubmed/5910392.

MARTINEZ, RLMC, NARANJO, J. D., (2010). A pretest for choosing between logrank and Wilcoxon tests in the two-sample problem. Metron: International Journal of Statistics, 68 (2), pp. 111–125. https://link.springer.com/article/10.1007/BF03263529.

OLLER, R., GOMEZ, G., (2012). A generalized Fleming and Harrington's class of tests for interval-censored data. The Canadian Journal of Statistics, 40 (3), pp. 501–516. http://onlinelibrary.wiley.com/doi/10.1002/cjs.11139/abstract.

PEPE, M. S., FLEMING, T. R., (1989). Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. Biometrics, 45, pp. 497–507. https://www.jstor.org/stable/2531492?seq=1#page_scan_tab_contents.

PETO, R, PETO, J., (1972). Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society, 135 (2), pp. 185–207. https://www.jstor.org/stable/2344317?seq=1#page_scan_tab_contents.

PRENTICE, R. L., MAREK, P., (1979). A qualitative discrepancy between censored data rank tests. Biometrics, 35 (4), pp. 861–867. https://www.jstor.org/stable/2530120?seq=1#page_scan_tab_contents.

STEVENSON, M., (2009). An Introduction to Survival Analysis, EpiCentre, IVABS. Massey Massey University. http://www.massey.ac.nz/massey/fms/Colleges/College%20of%20Sciences/Epicenter/docs/ASVCS/Stevenson_survival_analysis_195_721.pdf.

TARONE, R. E., WARE, J., (1977). On distribution-free tests for equality of survival distributions. Biometrika, 64, pp. 156–160. https://www.jstor.org/stable/2335790?seq=1#page_scan_tab_contents.

WANG, R., LAGAKOS, S. W., GRAY, R. J., (2010). Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring. Biostatistics, 11 (4), pp. 676–692. https://www.ncbi.nlm.nih.gov/pubmed/20439258.

XIE, J., LIU, C., (2005). Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Statistics in Medicine, 24 (20), pp. 3089–3110.
http://onlinelibrary.wiley.com/doi/10.1002/sim.2174/abstract.

STATISTICS IN TRANSITION-new series, June 2017
Vol. 18, No. 2, pp. 329–360, DOI 10. 21307

MET AND UNMET NEED FOR CONTRACEPTION: SMALL AREA ESTIMATION FOR RAJASTHAN STATE OF INDIA

Piyush Kant Rai[1], **Sarla Pareek**[2], **Hemlata Joshi**[3]

ABSTRACT

Nowadays, Family Planning has become the basic human right, which is closely linked to the empowerment of women and perhaps it is the only treatment that can avert many serious issues, which are an impediment in the advancement of the country like maternal mortality, infant mortality and can exert out the families from poverty and stabilize population growth etc. Increasing use of family planning can be helpful in the reduction of unmet need for family planning by which a substantial proportion of unwanted births ends in childbirths, and which are related to deaths and injuries for both mother and child.

Due to lack of availability of reliable data at the small level (area-wise) specifically in developing countries like India. In this article the small area estimation technique is used for the estimation of met and unmet need for contraception for 187 towns of Rajasthan state of India and for empirical analysis. Data is taken from the District Level Household Survey (DLHS): 2002-04 and the Census 2001 of India.

Key words: contraceptive use, met and unmet need for family planning, smallarea estimation, logit-link function, district level household survey, census of India.

1. Introduction

Adjustment and execution of any voluntary family planning programmes by the government of a country wish to improve the demographic situation ata particular time. If we focus on Millennium Development Goals (MDG's) that the families having lesser number of children can lead to healthy productive lives which can help in the alleviation of poverty (MDG 1). Children are more likely to attend school and attain higher education (MDG 2). Women with few number of pregnancies can lead to take up jobs and be empowered with improved status within their family as well as outside (MDG 3) and reduce the risk of maternal mortality either due to complications of pregnancy or an abortion (MDG 5). Well-spaced births can reduce malnutrition and infant mortality (MDG 4). One contraceptive method, the

[1]Banasthali University. E-mail: raipiyush5@gmail.com
[2]Banasthali University. E-mail: psarla13@gmail.com
[3]Manipal University. E-mail: hemlata.joshi28@gmail.com

condom, prevents both HIV transmission and unwanted pregnancy (MDG 6). Thus the real progress of the country depends on the improved family planning programs and policies and adequate providence of family planning services. As the use of contraception is a part of family planning, in most of the developing countries like India contraception is basically used to reduce fertility and protect couples from other infectious diseases at the time of sexual intercourse. Stover and Ross (2008) argued that contraception plays an important role in reducing maternal morbidity and mortality in the developing world, not only through the reduction of births, but also through the reduction of pregnancies for risk groups, such as teenagers and older women, who already have four or more children (Stover and Ross, 2008).

India launched the National Family Welfare Program in 1951 with the objective of reducing the birth rate to the extent necessary to stabilize the population, consistent with the requirements of the national economy. Since its inception, the program has experienced significant growth in terms of financial investment, service delivery points, type of services, and the range of contraceptive methods offered. Since October 1997, the services and interventions under the Family Welfare Programme and the Child Survival and Safe Motherhood Programme have been integrated with the Reproductive and Child Health Program.

The government of India has been organizing several programs for reducing birth rate. Some of the programs and policies have been successful and the rate of increase has also reduced, but has still to reach the sustainable rate. However, the knowledge of contraception is almost universal in India but still the total met need for family planning is low, i.e. 56.3%, and the total unmet need for contraception is about 12.8% in India (NFHS-3). Unmet need for family planning is a very important indicator for evaluating the potential demand for family planning services and determining the demographic goals in the countries having the fertility level below replacement. Ross and Winfrey (2002) argued that more than 100 million women in developing countries want to avoid pregnancy but are not using any method of family planning (Ross and Winfrey, 2002) and a significant proportion of unintended births ends in pregnancy and child births related injuries and deaths (Sedgh et al. 2007; Stover and Ross 2008). Demographers and health specialists refer to these women as having an “unmet need”for family planning that influences the development of family planning programs. Over the past decade, rising rates of contraception use has reduced the unmet need for family planning in most of the countries. In some less developed countries, where unmet need remains persistently high or is increasing, it is required to have greater efforts to understand and address the causes of unmet need of family planning.

Westoff (1978), using the World Fertility Survey (WFS) data, estimated the unmet need where the exposure was limited to fecund women who wanted no more children and who were not using contraception. Pregnant and amenorrheic women, and women who wanted children within two years, were not included in the definition of unmet need. Further refinement in the estimation of unmet need for spacing and limiting was carried out by Westoff and Pebley (1981). Nortman (1982) advocated inclusion of pregnant, breastfeeding and amenorrheic women in the definition of unmet need. The most widely used measure of unmet need was developed by Westoff and Bankole (1995), based on data from Demographic and Health Surveys (DHSs). However, it has been criticized for its exclusion of married men and unmarried girls and boys, its limited scope in reducing unwanted fertility, its non-addressing of side effects of methods, and its inclusion of traditional methods (Dixon-Muller & Germain (1992); Pritchett (1994); Reddy (2003)).

The concept of unmet need admits the promise of improving the health of population, by reducing fertility and achieving reproductive goals. The Programme of Action of the International Conference on Population and Development, Cairo (1994) recognizes this need and states that 'government goals for family planning should be defined in terms of unmet need for information and services'(United Nations 1994). Also, universal access to reproductive health services, of which unmet need for contraception is a key component, has been acknowledged as one of the main strategies in achieving the millennium development goals (United Nations 2005). Unmet need for contraception has been adopted as one of the monitoring indicators in the 62nd General Assembly of the United Nations in 2007.

Recently, after the Cairo Conference, reproductive health became an essential component of the Indian Family Welfare Programme. As in many other countries, there was a shift in emphasis in the programme in India. Unmet need for contraception became a policy instrument to strengthen the Reproductive and Child Health programme of the country. Meeting the unmet need for contraception has been accorded priority as it has the potential for the reduction of fertility and prevention of induced abortions (Ministry of Health and Family Welfare (MOHFW) 1997). The immediate objective of the National Population Policy is to address the unmet need for contraception, health care services and health infrastructure (MOHFW 2000). The recent policy document, the National Rural Health Mission (2005-2012), also aims at addressing the unmet need for contraception along with other objectives (MOHFW 2005). Thus, the unmet need for contraception is now a well-recognized and useful indicator to steer the programme in India.

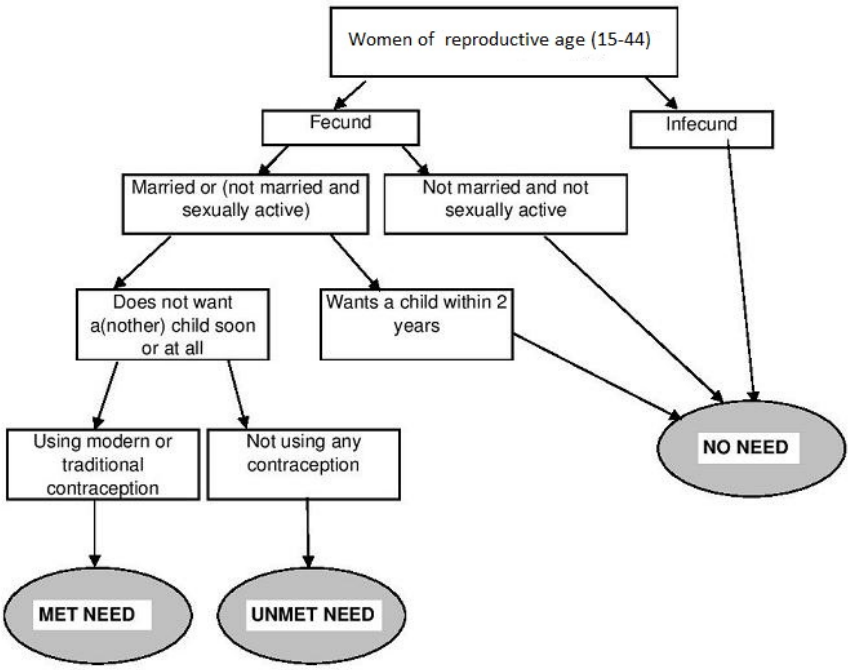
The term "unmet need"indicates the negative kind of concept in the sense that governments are unable to provide family planning programme services to those

who wish to use it but are not getting facilities. The reasons may be due to many socio-economic factors. Therefore here, the concept like KAP (women's knowledge of, attitudes toward, and practice of birth control)- gap came. In most of the cases, where a substantial proportion of women who wanted to stop childbearing or delay in pregnancy but are not practicing contraception, this discrepancy between reproductive preferences and birth control practices is referred to as the"KAP- gap"or the "unmet need"for contraception, (Bongaart,1991). According to NFHS-3, unmet need for spacing includes pregnant women whose pregnancy was mistimed; amenorrhoeic women who are not using family planning and whose last birth was mistimed, or whose last births was unwanted but now say they want more children; and fecund women who are neither pregnant nor amenorrhoeic, who are not using any method of family planning, and say they want to wait 2 or more years for their next birth. Also, unmet need for spacing includes fecund women who are not using any method of family planning and say they are unsure whether they want another child or who want another child but are unsure when to have the birth. Unmet need for limiting refers to pregnant women whose pregnancy was unwanted; amenorrhoeic women who are not using family planning, whose last child was unwanted and who do not want any more children; and fecund women who are neither pregnant nor amenorrhoeic, who are not using any method of family planning, and who want no more children. Pregnant and amenorrhoeic women who became pregnant while using a method (these women are in need of a better method of contraception) are excluded from the unmet need category.

There are key states (Bihar, Uttar Pradesh, Rajasthan, Madhya Pradesh, Jharkhand and Orissa) in India, where met need for family planning is low and unmet need for family planning is high in comparison with the national average, i.e. (56.3%) and 12.8% respectively. Also, the adolescent fertility rates in these states contribute largely to high TFR. Bihar, Uttar Pradesh, Rajasthan, Madhya Pradesh, Jharkhand and Orissa itself constitute 50 per cent of the total population of India. The comparative data suggest that alongside of TFR, the age at marriage, female education, and contraceptive prevalence rate are also lower in these states, whereas Maternal Mortality Rate (MMR) and Infant Mortality Rate (IMR) are quite high compared to the national average. There is an urgent need to focus on spacing and limiting methods of family planning. Thus, it will be very fruitful to us to understand the extent of met and unmet need that can be a powerful tool to manage the family planning programme effectively.

Combining the estimate of unmet need with the contraceptive use provides a picture of the total potential demand for family planning in a country, that is what the demand would be if all married women acted on their stated preferences. For

Figure 1: Conceptual framework of women with unmet need, met need and no need for contraception



family planning services and for policy purposes, this estimate is useful because it helps reveal the size and characteristics of the potential market for contraceptives and to project how much fertility could decline if the additional need for family planning were met.

The present study is conducted in Rajasthan state of India, where met need for family planning is 43.8% and approximately 14.65 per cent women want to delay or avoid pregnancy but are not using an effective method of family planning (NFHS-3). Small-Area Estimation technique is used for estimation of met and unmet need for family planning at town level in Rajasthan state of India. For the analysis, the data has been taken from the District Level Household Survey(DLHS):2002-04 and Census 2001 of 32 districts of Rajasthan, India and the met and unmet need for family planning are estimated for 187 towns of Rajasthan using the traditional small area technique.

Being a developing country, India does not provide complete and reliable statis-

tical, demographic and health data. It is well known to us that, however, census is a complete enumeration and provides demographic information as detailed as the settlement level, but this information is very limited and has some inadequacies. Also it is not possible to get useful information from the registration system even for the nation as a whole while sample surveys provide accurate and detailed demographic information and some basic health information, and this information is limited to nation totals, urban and rural area and at most to geographical regions due to the nature of sample surveys. In recent years there has been an ample growth in the demand for statistical data relating to various subdivisions into the country. Sometimes, the geographical subdivisions of our interest include relatively large units, such as states, and some subdivisions include smaller units, such as towns, rural communities, local government districts, or health service areas. Statistics for geographical subdivisions, commonly referred to as small area statistics, are of great interest in many countries throughout the world (Kalton, et. al, 1993).

The use of small area statistics germinated several centuries ago. Brackstone (1987) mentions the existence of such statistics in 11th century England and 17th century Canada. As well, various powerful statistical methods with theoretical foundations have emerged for the analysis of local data. Indirect Small Area estimation techniques can be classified into two main groups (Rao, 2000, Marker 1999): Traditional Techniques and Model Based Techniques. Synthetic Estimates are considered as one the Traditional Techniques and this technique requires relatively available data from the surveys and censuses.

Deriving small-area statistics for maternal health indicators, such as contraceptive use, unmet need, and satisfied demand for contraception, are particularly important for a country that lacks the infrastructure and resources to mount surveys to collect representative data at the district level. In recent times, policy makers, health care providers, and planners have shown increased interest in small area statistics, particularly where decentralized approaches to health planning and resource allocation have been adopted.

2. Data

The findings in this paper are based on data taken from the District Level Household and Facility Survey(DLHS):2002-04 and Census 2001 of 32 districts of Rajasthan, India, DLHS is designed to provide estimates on maternal and child health, family planning and other reproductive health services.

For the assessment of district level Reproductive and Child Health indicators, Government of India intended to undertake district level household surveys through the non-governmental agencies on an annual basis. The District Level Household

Survey (DLHS) was the result of government's initiative. In Rajasthan, IHMR, India was confided the work of carrying out of the survey. The survey for Phase-1 of the DLHS covering 9 districts of the state was conducted during May 2002 to August 2002. The survey for Phase-2 covering the remaining districts of the state was carried out during Feb 2004 to June 2004. The focus of the survey was on: i) Coverage on antenatal care (ANC) and immunization services, ii) Extent of safe deliveries, iii) Contraceptive prevalence rate and unmet need for family planning, iv) Awareness about RTI/STI and HIV/AIDS, and v) Utilization of government health services and user's satisfaction.

For both the phases together, the data was collected from 33,833 households in Rajasthan. From these households, 32,911 eligible women (usual resident or visitors who stayed in the sample household the night before the interview, currently married aged 15-44 whose marriage was consummated) and 20,980 husbands of eligible women were interviewed.

3. Methodology and Model Diagnostics

In this paper, we are interested in estimating such a model in which the dependent variable is dichotomous in nature and takes values 0 and 1. Here, we want to estimate the met and unmet need for family planning methods as a function of several variables like female literacy rate (FLR), female work participation rate (FWPR), proportion of urban population (PUP), density, any government health facility (GHF), decadal increase during census 1999-2001 (growth rate), number of illiterate persons of age group (7 and above). The most commonly used approach for estimating such models is the linear probability model i.e. Logit model.

The Logit regression analysis is the uni/ multivariate technique which allows us for estimating the probability that an event occurs or not by predicting the binary dependent outcome for a set of independent variables.

The linear probability model is depicted as

$$\begin{aligned} p_i &= E(y=1|X_i) \\ &= \beta_0 + \beta_i X_i \end{aligned}$$

where $y = 1$ is the met need for family planning or the unmet need for family planning and X_i are the independent variables.
Let us consider the following representation of the above model (Logit model)

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_i X_i)}} \tag{3.1}$$

For ease of exposition, we write equation (3.1) as

$$p_i = \frac{1}{1+e^{-z_i}} = \frac{e^z}{1+e^z} \tag{3.2}$$

where $z_i = \beta_0 + \beta_i X_i$

The equation (3.2) is called the logistic distribution function. Here z_i ranges from $-\infty$ to ∞, p_i ranges from 0 to 1, p_i is non linearly related to z_i i.e. X_i, thus satisfying two conditions of the required probability model. Here, an estimation problem has occurred because p_i is not only non linear in X_i but also in $\beta's$. This means that OLS estimator cannot be used to estimate the parameters.

Here, p_i is the probability of using contraception, given in equation 3.2, then $(1-p_i)$, the probability of not using contraception is given as

$$1 - p_i = \frac{1}{1+e^{z_i}} \tag{3.3}$$

Therefore, we can write

$$\frac{p_i}{1-p_i} = \frac{1+e^{z_i}}{1+e^{-z_i}} = e^{z_i} \tag{3.4}$$

$\frac{p_i}{1-p_i}$ is the odds ratio in favour of using contraception. Taking natural log of equation (3.4), we obtain

$$L_i = ln\left(\frac{p_i}{1-p_i}\right) = z_i = \beta_0 + \beta_i X_i \tag{3.5}$$

Let N_i and n_i be the total number of married women of age group 15-44 in the population and number of married women selected in the sample of reproductive age group of any i^{th}district $(i = 1,2,\ldots D)$ respectively, where $D = 32$ districts of Rajasthan state of India. Let y_i be the number of women possessing the given attribute in the district i.e. the number of women having met need for family planning. Also, let y_{si} and y_{nsi} be the women selected in the sample who possess the given characteristic and the women who are not counted in the sample but have the same characteristic. As we have discussed above, that the response variable Y_{si} follows binomial distribution with parameters n_i and p_i. then obviously y_{nsi} will also follow the binomial distribution with parameters $N_i - n_i$ and p_i, where p_i is the probability that a woman possesses the attribute of using contraception in the district i. Then,

$$L_i(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = z_i = \beta_0 + \beta_i X_i + u_i; \quad i = 1,2,\ldots 32 \tag{3.6}$$

Here, β is the k-vector of unknown fixed effects parameters and we assume that u_i are the random effects that accounts for between district variability in the response

not explained by the independent variables in the model and independently identically normally distributed with mean 0 and variance ϕ. Now, from equation (3.6), we can write

$$p_i = e^{z_i}\left(1+e^{z_i}\right)^{-1} = e^{\beta_0+\beta_i X_i+u_i}\left(1+e^{\beta_0+\beta_i X_i+u_i}\right)^{-1} \tag{3.7}$$

Equation (3.6) relates the area-level proportions to area-level covariates. This type of model is often referred to as an area-level model in SAE terminology (Rao 2003). Fay and Herriot's (1979) used this model for the prediction of mean income per head in small geographic areas (less than 500 persons) within counties in the USA. Fay and Herriot method is based on an area-level linear mixed model for small area estimation, which is applicable for a continuous variable. In contrast, equation (3.6) is a special case of Generalized Linear Mixed Model (GLMM) with logit-link function (Breslow and Clayton 1993) and it is suitable for discrete variable, particularly for binary variable. Alternative approaches to estimating the logistic model for the small-area-estimation case include the empirical Bayes and the hierarchical Bayes methods (Rao 2003). Saei and Chambers (2003) described equation (3.6) in the context of SAE.

By definition, means of y_{si} and y_{nsi} are

$$E(y_{si}|u_i) = n_i\left[e^{\beta_0+\beta_i X_i+u_i}\left(1+e^{\beta_0+\beta_i X_i+u_i}\right)^{-1}\right] \tag{3.8}$$

and

$$E(y_{nsi}|u_i) = (N_i-n_i)\left[e^{\beta_0+\beta_i X_i+u_i}\left(1+e^{\beta_0+\beta_i X_i+u_i}\right)^{-1}\right] \tag{3.9}$$

Now, let T_i be the total number of women possessing the characteristic of met need of family planning in the i^{th} district, then T_i can be written as

$$T_i = y_{si}+y_{nsi};\ \ i = 1,2\ldots 32 \tag{3.10}$$

Here, T_i includes all the women with the attribute of using contraception who are selected in the sample (y_{si}) and not selected in the sample but possessing the attribute y (y_{nsi}). In the expression 3.10, the first term y_{si} (i.e. the direct estimate from the survey)is known whereas the second term y_{nsi}, the non sample count is unknown. Thus, the total number of women with met need of family planning in district i can be obtained by replacing y_{nsi} by its estimated value under the model 3.6. So,

$$\hat{T}_i = y_{si} + \hat{y}_{nsi} = (N_i - n_i)\left[e^{\beta_0+\beta_i X_i+u_i}\left(1+e^{\beta_0+\beta_i X_i+u_i}\right)^{-1}\right] \tag{3.11}$$

Sometimes sample data are not available for some districts for which $n_i = 0$ and $y_{si} = 0$. In this context small area estimation technique can be used to derive the estimates for the districts for which data are not available.
Here we have used synthetic type estimator for estimating the T_i, given as,

$$\hat{T}_i = N_i\left[e^{\beta_0+\beta_i X_i+u_i}\left(1+e^{\beta_0+\beta_i X_i+u_i}\right)^{-1}\right] \tag{3.12}$$

The proportion of women with the assignable property (met or unmet need (limiting or spacing) for family planning) in district i (p_i) is obtained by the ratio of the total number of women of the reproductive age group with the particular outcome (met or unmet need (limiting or spacing) for family planning) to the total number of women of the reproductive age group of that t^{th} district. Thus, (p_i) can be written as

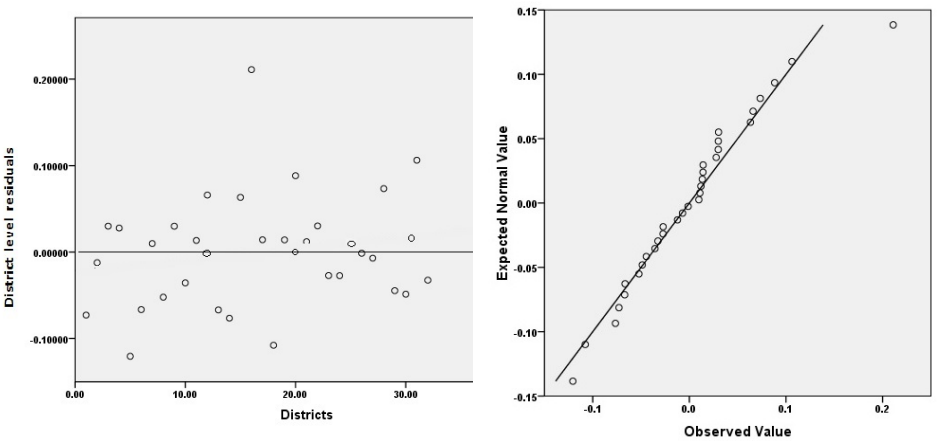
$$\hat{p}_i = \frac{\hat{T}_i}{N_i} \tag{3.13}$$

While using the logit link function, the residuals at district level are assumed to follow normal distribution with mean 0 and variance ϕ. The model diagnostics are used to verify that the model assumptions are satisfied or not. If the model assumptions are satisfied then the residuals are randomly distributed and do not differ significantly from the regression line $y = 0$. Therefore, from equation 3.6, the residuals can be written as $u_i = \hat{z}_i - (\hat{\beta}_0 + \hat{\beta}_i X_i)$, where $i = 1, \ldots 32$. Plots of the distribution of residuals given in Figure 1 show that the residuals are randomly distributed and regression line is not significantly different from the residual line of fit and the qq plot is also satisfying the normality assumption of residuals.

4. Result and Discussion

In the present study, we estimate the proportion of women having met and unmet need (spacing and limiting fertility) of family planning with 95% confidence intervals at town level (187 towns) by using the sample of 32 districts of Rajasthan, the largest state (area-wise) in India, taken from the District Level Household Survey(DLHS): 2002-04 and Census 2001 of India using special case of Generalized Linear Mixed Model with logit-link function (Breslow and Clayton, 1993). We use this model because of the providence of estimates of the binomial nature variables. The detailed discussion of the method is given in Amoako Johnson *et al.* (2010).

Figure 1: Model diagnostic plot and qq plot of residuals for districts



Here, Table 1 describes the estimates of the Parameters and their standard errors from the generalized linear mixed model of met and unmet need for contraception in Rajasthan. The results are also tested at 5% level of significance. It is found that the β-coefficients for the different covariates and random effects with "$*$ "have the significant impact on the met and unmet need for contraception and the percentage of met and unmet need for family planning within the districts can be seen in the Figure 1, which shows that the trend of maximum use of family planning methods is in Hanumangarh (66.5%), Ganganagar (65.4%) and Jaipur (62.4%) districts and the Barmer (24.5%) and Jaisalmer (27%) are using the contraception methods at low level among all the districts of Rajasthan.

The unmet need for spacing lies within the range of 2.5% to 17.2% for the 32 districts and in overall Rajasthan, the unmet need for spacing fertility is approximately 8%. Unmet need for contraception for limiting fertility lies between 5.5% to 21.1% in districts of Rajasthan and the Rajasthan overall has 13.7% unmet need for limiting fertility. In Figure 1, it is also observed that the percentage of unmet need for limiting fertility is higher than the unmet need for spacing fertility among all the districts of Rajasthan. The total unmet need (for limiting and spacing fertility) in Rajasthan is about 21.8%. By using these estimates, the "demand satisfied" $\big[$ a measure obtained by dividing the current use of contraception by the total demand i.e. the sum of met and unmet need for contraception (UNFPA 2010)$\big]$ is also calculated at district level, which is an another indicator of program effectiveness and is being used increasingly. The total demand satisfied by the districts is presented in Figure 2 and it is observed that Barmer and Jaisalmer are the two districts

in which only $40\% - 42\%$ is satisfied and Hanumangarh, Ganganagar and Jaipur have the highest percentage of demand satisfied for any method of contraception, i.e. $89.26\%, 87.08\%$ and 85.01% respectively. Figure 2 also tells in almost all the districts the percentage of demand satisfied for any method vary between 50% to 70%.

To the estimation of met and unmet need for family planning at town level, the small area estimation (SAE) technique is used, which is explained in Section 3 and the obtained results are tabulated in Table 3, which suggest that the maximum use of contraception (any method) in the women in the reproductive age group 15-44 years is in Todra town of Sawai Madhopur (59.98%). It can also be inferred that, out of 100 women of the reproductive age group 15-44 years of Sawai Madhopur district approximately 60 women who are using the contraceptive devices are from Todara town followed by Jaipur (57.94%), i.e. per 100 women of the reproductive age group 15-44 years of Jaipur district 58 women are having met need for contraception and the remaining 42 women are from other towns of the same district, Bikaner town (53.89%) of Bikaner district and Mahwa town (52.18%) of Mahwa district. The minimum use of family planning methods is in Jobner town of Jaipur district (0.27%) i.e. out of 100 women of the reproductive age group 15-44 years of the Jaipur district, not even one woman is able to get the met need for contraception followed by Viratnagar (0.44%), Phulera (0.55%) and Bagru (0.56%) of Jaipur district within state of Rajasthan. In the Jaipur district, only Jaipur town is using 57.94% of the contraceptive methods and the rest of the towns of this district are using contraception methods at a very low level, which is approximately equivalent to zero (Table 3). Similarly, the results can also be seen for the unmet need for contraception from the same Table 3. The percentage of unmet need of family planning methods for limiting fertility lies between 0.22% to 53.82% in the Rajasthan state, which means approximately 54% women of their reproductive age group 15-44 years need family planning devices to limit the fertility but they are suffering from the unmet need for contraception, and the limits for unmet need for spacing fertility within towns are 0.13% and 31.11%, i.e. in overall Rajasthan, 31% of women in their reproductive age group are facing the problem of unmet need for spacing the fertility.

From Figure 3 and Table 3, it can also be concluded that approximately 70% towns of Rajasthan state using family planning (any method) at very low level i.e. approximately 10%. There are only 2.1% of the towns in which 50% of women are able to get the advantage of contraceptive devices, approximately 25 % towns have unmet need for limiting fertility i.e. more than 10 per cent and same is with the un-

met need for spacing fertility, here more than 10 per cent women of the reproductive age group 15-44 years of 17% of the towns out of 187 towns of the Rajasthan are facing the problem of unmet need family planning for spacing fertility.

The comparison of the met and unmet need for family planning for various districts of Rajasthan with the estimated values of the met and unmet need for family planning by the Annual Health Survey(2012-13) has also been done, which can be seen in Figure 4 and Table 2. This shows that the percentage of total unmet need for contraception in some of the districts (Baran, Dausa and Kota) is approximately the same during the period 2002-2013. Hanumangarh district has increased the percent of total unmet need for contraception with the highest points (9.86) followed by Bikaner, Jaisalmer and Sirohi. Churu district has reduced the total percent of unmet need for contraception with 8.46 points followed by Tonk, Jaipur and Bharatpur districts in the time duration (2002-2013). Approximately 70% of the districts reduced the percentage of the unmet need for spacing, out of which Bharatpur district reduced the maximum percentage of unmet need for spacing followed by Jaipur, Ajmer and Bhilwada. Also, in approximately 30% of the districts, there was an increment in the unmet need for spacing (highest increment is seen in Jaisalmer district). The noticeable point is that only 18% of the districts reduced the percentage of unmet need for limiting fertility and in the remaining 82% of districts, the unmet need for limiting increased. The highest reduction in the percentage of the unmet need for limiting fertility is seen in Churu (6.47%) and the highest increment is seen in Sirohi (7.93%) followed by Hanumangarh. This means that for spacing between the births, women are using the contraceptive devices in a very effective manner but for limiting the births, they are not using the contraception methods properly.

Since unmet need for family planning is an indicator to evaluate the effectiveness of the family planning programme, the policy makers and family planning programme planners use it to know the demand for family planning services/supplies. Thus, it is very important to focus on the problems which affected the unmet need for contraception. Also, We cannot refuse the fact that Rajasthan, a prime state of country, is suffering from various serious demographic and social issues like early age at marriage, low level of education, high TFR rate and low rate of contraception use. The nature of the present study supports the planners in implementing policies and programmes based on the results in the large scale, not only within the state but also within the country.

References

[1] AMOAKO, J. F. & MADISE, N. J., (2009). Examining the geographical heterogeneity associated with risk of mistimed and unwanted pregnancy in Ghana. *Journal of Biosocial Science*. 41(2). pp. 249–267.

[2] AMOAKO, J. F., CHANDRA, H., BROWN, J. J. & PADMADAS, S. S., (2010), District-level estimates of institutional births in Ghana: application of small area estimation technique using Census and DHS data. *Journal of Official Statistics*. 26(2). pp. 341–359.

[3] BATTESE, G. E., HARTER, R. M. & FULLER, W. A., (1988). An error-components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*. 83(401). pp. 28–36.

[4] BONGAARTS, J., (1991). The KAP-gap and the unmet need for contraception. *Population and Development Review*. 17(2). pp. 293–313.

[5] BONGAARTS, J., (2006). The causes of stalling fertility transitions, *Studies in Family Planning*. 37(1). pp. 1–16.

[6] BRADLEY, S. E. K., CROFT, T. N., FISHEL, J. D., & WESTOFF, C. F. (2012) Revising unmet need for family planning. DHS Analytical Studies No. 25, Calverton, MA: ICF International.

[7] BRESLOW, N. E. & CLAYTON, D. G., (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistics Association*. 88(421). pp.9–25.

[8] CAMPBELL, O. M. R. & GRAHAM, W. J., (2006). Strategies for reducing maternal mortality: Getting on with what works. *The Lancet*. 368(9543). pp. 1284–1299.

[9] CHOUDHARY, S., SALUJA, N., SHARMA. S., GAUR, D. and PANDEY, S., (2009). A study on the extent and reasons of unmet need for family planning among women of reproductive age group in rural area of Haryana. *Internet J Health*. 12(1).

[10] CLELAND, J., NDUGWA, R. P. & ZULU, E. M., (2011). Family planning in sub-Saharan Africa: Progress or Stagnation? *Bulletin of the World Health Organisation*. 89. pp. 137–143.

[11] DATTA, G. S., GHOSH, M. & WALLER, L. A., (2000). Hierarchical and empirical Bayes method for environmental risk assessment. *Handbook of Statistics*. 18. pp. 223–245.

[12] DEMOMBYNES, G., ELBERS, C., LANJOUW, J. O. & LANJOUW, P., (2007). How good a map? Putting small area estimation to the test. World Bank Policy Research Working Paper 4155. *The World Bank, Washington DC*. Available from: http://papers.ssrn.com/sol3/papers.cfm? abstract_id=967547.

[13] DIXON-MULLER, R. & GERMAIN, A. (1992) Stalking the elusive: unmet need for family planning. *Studies in Family Planning*. 23(25). pp. 330–335.

[14] ELBERS, C., LANJOUW, J. O. & LANJOUW, P., (2002). Micro-level estimation of welfare. World Bank Policy Research Working Paper 2911. *The World Bank, Washington DC*. Available from: http://are.berkeley.edu/courses/ARE251/2004/assignments/mapping.pdf.

[15] ELBERS, C., LANJOUW, J. O. & LANJOUW, P., (2003). Micro-level estimation of poverty and inequality. *Econometrica*. 71(1). pp. 355–364.

[16] FAY, R. E. & HERRIOT, R. A., (1979). Estimation of income from small places: An application of James-Stein procedures to Census data. *Journal of the American Statistics Association*. 74(366). pp. 269–277.

[17] FUJII, T., (2010). Micro-level estimation of child under nutrition indicators in Cambodia. *World Bank Economic Review*. 24(3). pp. 520–553.

[18] GONZALEZ-MANTEIGA, W., LOMBARDIA, I., MOLINA, M. J., MORALES, D. & SANTAMARIA, L., (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics & Data Analysis*. 51(5). pp. 2720–2733.

[19] GULATI, S. C., and DAS, A., (2013). Demand for contraception in EAG states in India: Family planning strategies to meet the unmet need. *Health and Population - Perspectives and Issues*. 36(3 & 4). pp. 115–132.

[20] KUMAR, S., PRIYADARSHNI A., KANT, S., ANAND, K. and YADAV, B. K.,(2005). Attitude of women towards family planning methods and its use - study from a slum of Delhi. *Kathmandu Univ Med J*. 3. pp. 259–62.

[21] MARSTON, C. & CLELAND, J., (2004). The effects of contraception on obstetric outcomes. Working Paper No. 630, Department of Reproductive

Health and Research, *World Health Organisation, Geneva*. Available from: http://whqlibdoc.who.int/publications/2004/9241592257.pdf.

[22] Ministry of Health and Family Welfare (MOHFW), (1997). Reproductive and Child Health Programme: Schemes for Implementation, Department of Family Welfare. MOHFW, New Delhi.

[23] Ministry of Health and Family Welfare (MOHFW), (2000). National Population Policy 2000, Department of Family Welfare. MOHFW, New Delhi.

[24] Ministry of Health and Family Welfare (MOHFW), (2005). National Rural Health Mission (2005-2012) Mission Document, Department of Family Welfare. MOHFW, New Delhi.

[25] NORTMAN, D. L., (1982). Measuring the unmet need for contraception to space and limit births. *International Family Planning Perspectives*. 87(4). pp. 125–134.

[26] PATIL, S. S., RASHID, A. K. and NARAYAN, K. A., (2010). Unmet needs for contraception in married women in a tribal area of India. *Malaysian J Public Health Med.* 10. pp. 44–51.

[27] PFEFFERMANN, D., (2002). Small area estimation: new developments and directions. *International Statistical Review*. 70(1). pp. 125–143.

[28] PRATESI, M. & SALVATI, N., (2008). Small area estimation: The EBLUP estimator with autoregressive random area effects. *Statistical Methods and Application.* 17. pp. 113–141.

[29] PRITCHETT, L. H., (1994). Desired fertility and the impact of population policies. *Population and Development Review*. 20(1). pp. 1–55.

[30] REDDY, P. H., (2003). Unmet need for contraception: how real is it?. *Economic and Political Weekly*. 38(3). pp. 188–191.

[31] RAO, J. N. K., (2003). *Small Area Estimation*. John Wiley & Sons, New Jersey.

[32] ROSS, J. A. & WINFREY, W. L., (2002). Unmet need for contraception in the developing world and the former Soviet Union: an updated estimate. *International Family Planning Perspectives*. 28(3). pp. 138–143.

[33] ROSS, J. A. & WINFREY, W. L., (2001). Contraceptive Use, Intention to Use and Unmet Need during the Extended Postpartum Period. *International Family Planning Perspectives*. 27(1). pp. 20âĂŞ27.

[34] SAEI, A. & CHAMBERS, R., (2003). Small area estimation under linear and generalized linear mixed models with time and area effects. Methodology Working Paper No. M03/15. Southampton Statistical Sciences Research Institute, *University of Southampton, Southampton*.

[35] SEDGH, G., HENSHAW, S., SINGH, S., AHMAN, E. & SHAH, I. H., (2007). Induced abortion: Estimated rates and trends worldwide. *The Lancet.* 370(9595). pp. 1338–1345.

[36] SONFIELD, A. (2006) Working to eliminate the world's unmet need for contraception. *Guttmacher Policy Review.* 9(1). p.10-13.

[37] STOVER, J. & ROSS, J. (2008) How contraceptive use affects maternal mortality. Washington, DC: US Agency for International Development, Health Policy Initiative, Task Order 1. Available from: http://www.healthpolicyinitiative.com/Publications/Documents/668_1_TMIH_FINAL_ 12_19_08.pdf.

[38] United Nations (1994). Programme of Action for the 1994 International Conference on Population and Development. *United Nations, New York.*

[39] United Nations (2005). Investing in Development: A Practical Plan to Achieve the Millennium Development Goals. [online] Available from: http://www.unmillenniumproject.org/reports/ index.htm [Accessed: 24^{th} February 2009].

[40] UNFPA (2010) How Universal is Access to Reproductive Health: A Review of the Evidence. *New York: United Nations Population Fund*. Available from: http://www.unfpa.org/public/home/publications/pid/6526.

[41] WESTOFF, C. F. (1978) The unmet need for birth control in five Asian countries. *Family Planning Perspectives*. 10(3). p.173-181.

[42] WESTOFF, C. F. & BANKOLE, A. (1995) Unmet need: 1990-94, DHS Comparative Studies No. 16. *Macro International Inc.. Calverton, MD*.

[43] WESTOFF, C. F. & PEBLEY, A. R. (1981) Alternative measures of unmet need for family planning in developing countries. *International Family Planning Perspectives*. 7(4). p.124-136.

[44] World Health Organisation. (2009) Achieving Millennium Development Goal 5: Target 5A and 5B on Reducing Maternal Mortality and Achieving Universal Access to Reproductive Health, WHO/RHR/09.06. *Geneva: World Health Organisation*. Available from: http://www. who.int/reproductivehealth/publications/monitoring/rhr_09_06/en/index.html.

Appendix:

Table 1: Parameter estimates and their standard errors from the generalized linear mixed model of met and unmet need for contraception, Rajasthan

	Met need for contraception	Unmet need for contraception	
	Any method	For limiting fertility	For spacing fertility
	β	β	β
Covariates and random effects	SE	SE	SE
Intercept	0.4577*	0.1340*	-0.0107
	(0.1936)	(0.0853)	(0.0951)
Female literacy rate	0.0073*	-0.0015*	0.0011
	(0.0033)	(0.0014)	(0.0016)
Female work participation rate	-0.0028	0.0015	0.0022*
	(0.0032)	(0.0014)	(0.0016)
Percentage of urban population	-0.0024	0.0003*	-0.0002
	(0.0028)	(0.0012)	(0.0014)
Density	0.0000	0.0000	-0.0001*
	(0.0002)	(0.0000)	(0.0001)
Any govt. health facility	-0.0020*	0.0004	0.0008
	(0.0017)	(0.0007)	(0.0008)
Growth rate	-0.0052	0.0000	-0.0007
	(0.0045)	(0.0020)	(0.0022)
No. of illiterate persons	0.0235*	-0.0061	-0.0082*
	(0.0167)	(0.0074)	(0.0082)

Note - $*p < 0.05$; β refers to the parameters and *SE* refers to the standard error.

Table 2: Comparison of observed estimates of met and unmet need for family planning of various districts of Rajasthan with the estimates of Annual Health Survey (2012-13)

Districts	Annual Health Survey (2012-13)			Observed estimates		
	Spacing	Limiting	Total	Spacing	Limiting	Total
Ajmer	5.18	8.84	14.02	9.6	8.3	17.9
Alwar	4.44	7.62	12.06	6.5	7.7	14.2
Banswara	5.66	4.57	10.23	4.6	2.3	6.9
Baran	5.19	8.93	14.12	7.7	6.4	14.1
Barmer	12.56	11.86	24.42	10	9.7	19.7
Bharatpur	3.89	6.63	10.52	10.2	4.8	15
Bhilwara	3.89	6.79	10.68	8	4.8	12.8
Bikaner	10.37	12.81	23.18	8.5	8.7	17.2
Bundi	7.78	13.47	21.25	10.1	8.7	18.8
Chittaurgarh	3.98	6.74	10.72	5.7	2.7	8.4
Churu	3.11	5.33	8.44	5.1	11.8	16.9
Dausa	6.22	10.71	16.93	8.7	7.5	16.2
Dhaulpur	11.27	15.56	26.83	10	15.4	25.4
Dungarpur	6.81	5.66	12.47	5.9	2.9	8.8
Ganganagar	3.11	5.26	8.37	4	2.5	6.5
Hanumangarh	5.19	8.87	14.06	2.8	1.4	4.2
Jaipur	2.83	4.75	7.58	7.8	4.7	12.5
Jaisalmer	15.56	9.56	25.12	12	7.2	19.2
Jalor	10.37	6.66	17.03	11.5	7.2	18.7
Jhalawar	3.89	6.71	10.6	6.1	7.1	13.2
Jhunjhunun	3.46	5.87	9.33	4.3	6.7	11
Jodhpur	10.75	5.45	16.2	9.2	4.3	13.5
Karauli	10.37	17.82	28.19	12.3	12.9	25.2
Kota	3.46	5.83	9.29	5.7	4.5	10.2
Nagaur	7.9	5.33	13.23	7.4	7.7	15.1
Pali	8.67	13.64	22.31	9.2	11.3	20.5
Rajsamand	9.67	7.4	17.07	9.8	6.1	15.9
Sawai Madhopur	7.58	6.44	14.02	8.5	9.9	18.4
Sikar	4.44	7.57	12.01	6	8.2	14.2
Sirohi	7.78	13.43	21.21	10.2	5.5	15.7
Tonk	5.19	8.99	14.18	9	11.8	20.8
Udaipur	5.45	5.9	11.35	8.4	4.6	13

Figure 1: Percentage of women with met and unmet need for contraception within districts of Rajasthan, 2002-2004

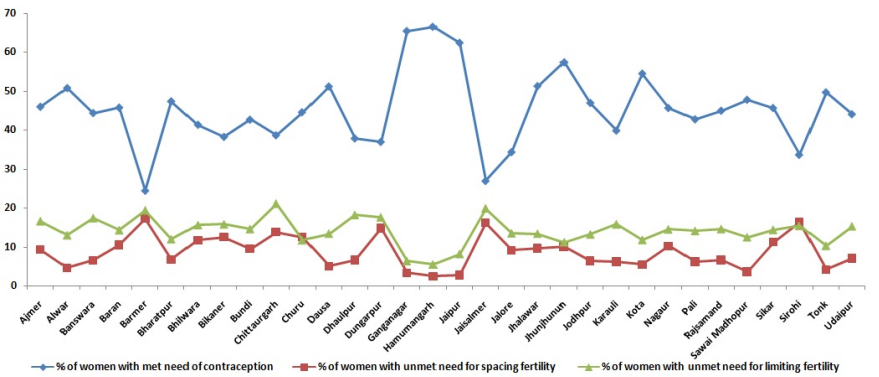


Figure 2: Percentage of women with demand satisfied for contraception within districts of Rajasthan, 2002-2004

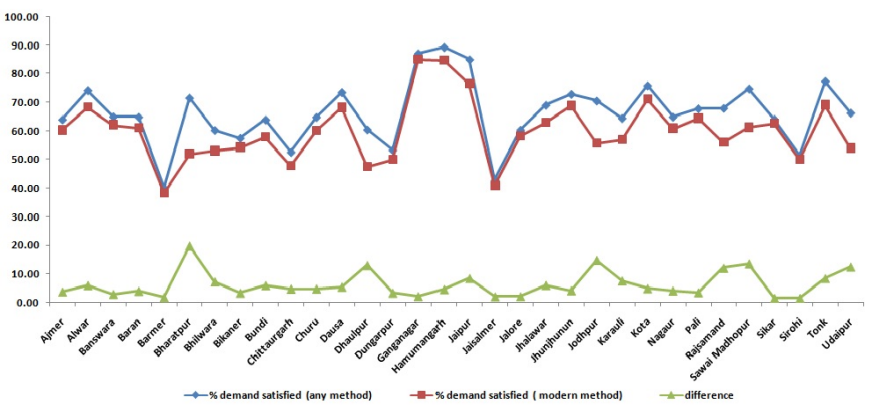


Figure 3: Percentage of towns with met and unmet need for contraception within Rajasthan, 2002-2004

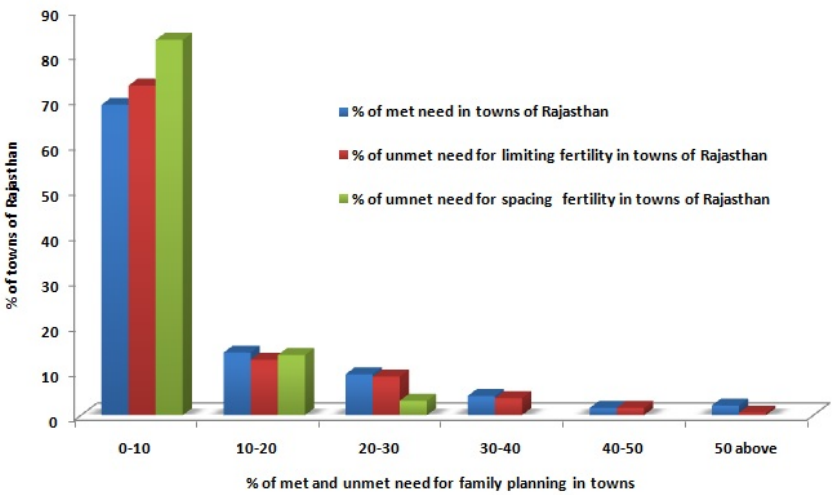


Figure 4: Comparison of observed unmet need for contraception (%) of various districts of Rajasthan with the data of Annual Health Survey (2012-13)

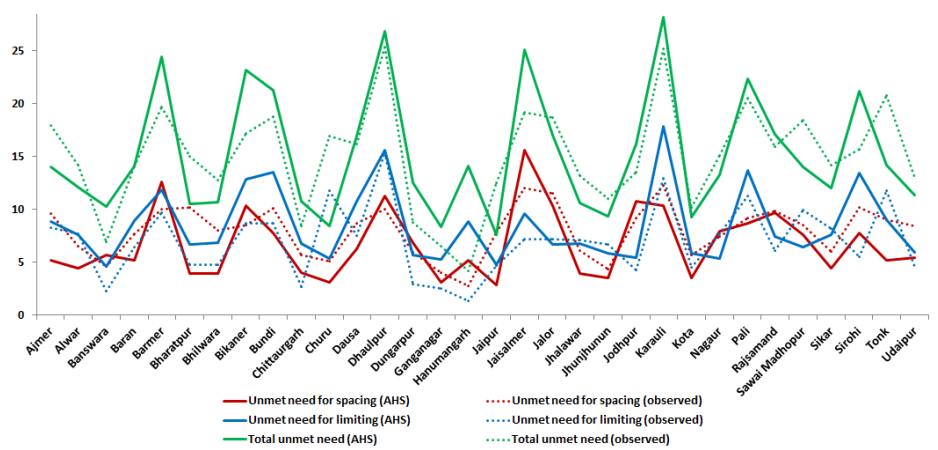


Table 3: Model-based estimates of proportion of women with met and unmet need for contraception in 187 towns within 32 districts of Rajasthan, India

District/Town	**Met need for family planning (any method)**				**Unmet need for family planning**							
	Proportion of women of age group 15-44 in tehsils using contraception	per cent	Lower limit	Upper limit	unmet need for family planning for limiting fertility	Per cent	Lower limit	Upper limit	unmet need for family planning for spacing fertility	Per cent	Lower limit	Upper limit
Ajmer												
Kishangarh (M)	0.29	28.85	0.28	0.29	0.24	24.29	0.24	0.25	0.142	14.25	0.14	0.15
Nasirabad (CB)	0.11	10.92	0.10	0.12	0.09	9.19	0.09	0.10	0.054	5.39	0.05	0.06
Pushkar (M)	0.04	3.58	0.03	0.04	0.03	3.01	0.02	0.04	0.018	1.77	0.01	0.02
Kekri (M)	0.09	8.63	0.08	0.09	0.07	7.26	0.07	0.08	0.043	4.26	0.04	0.05
Sarwar (M)	0.04	4.11	0.03	0.05	0.03	3.46	0.03	0.04	0.020	2.03	0.02	0.03
Vijainagar (M)	0.07	6.93	0.06	0.08	0.06	5.83	0.05	0.06	0.034	3.42	0.03	0.04
Alwar												
Bhiwadi (CT)	0.12	12.19	0.11	0.13	0.10	10.37	0.10	0.11	0.060	6.05	0.05	0.07
Govindgarh (CT)	0.05	4.71	0.04	0.06	0.04	4.01	0.03	0.05	0.023	2.34	0.02	0.03
Khairthal (M)	0.15	14.74	0.14	0.16	0.13	12.55	0.12	0.13	0.073	7.31	0.07	0.08
Kherli (M)	0.07	7.08	0.06	0.08	0.06	6.02	0.05	0.07	0.035	3.51	0.03	0.04
Tijara (M)	0.09	9.27	0.08	0.10	0.08	7.89	0.07	0.09	0.046	4.60	0.04	0.05
Behror (M)	0.10	10.38	0.09	0.11	0.09	8.84	0.08	0.10	0.052	5.15	0.05	0.06
Kishangarh (CT)	0.04	4.35	0.03	0.05	0.04	3.70	0.03	0.05	0.022	2.16	0.02	0.03
Banswara												
Partapur (CT)	0.30	30.39	0.28	0.32	0.27	27.27	0.25	0.29	0.157	15.71	0.14	0.17
Kushalgarh (M)	0.30	29.81	0.28	0.32	0.27	26.76	0.25	0.29	0.154	15.41	0.14	0.17

Table 3 *Continued*

District/Town	Met need for family planning (any method)				Unmet need for family planning							
	Proportion of women of age group 15-44 in tehsils using contraception	per cent	Lower limit	Upper limit	unmet need for family planning for limiting fertility	Per cent	Lower limit	Upper limit	unmet need for family planning for spacing fertility	Per cent	Lower limit	Upper limit
Baran												
Kherliganj (CT)	0.02	2.05	0.01	0.03	0.02	1.81	0.01	0.03	0.011	1.05	0.00	0.02
Mangrol (M)	0.08	7.81	0.07	0.09	0.07	6.91	0.06	0.08	0.040	4.01	0.03	0.05
Baran (M)	0.28	27.70	0.27	0.28	0.25	24.50	0.24	0.25	0.142	14.22	0.14	0.15
Antah (M)	0.09	9.39	0.09	0.10	0.08	8.31	0.08	0.09	0.048	4.82	0.04	0.05
Chhipabarod (CT)	0.06	5.75	0.05	0.07	0.05	5.08	0.04	0.06	0.030	2.95	0.02	0.04
Chhabra (M)	0.08	7.89	0.07	0.09	0.07	6.98	0.06	0.08	0.041	4.05	0.03	0.05
Barmer												
Barmer (M)	0.34	33.86	0.33	0.35	0.31	30.81	0.30	0.32	0.178	17.85	0.17	0.18
Balotra (M)	0.25	25.17	0.24	0.26	0.23	22.91	0.22	0.24	0.133	13.27	0.13	0.14
Bharatpur												
Bhusawar (M)	0.06	5.87	0.05	0.07	0.05	4.93	0.04	0.06	0.029	2.89	0.02	0.03
Bayana (M)	0.10	10.35	0.10	0.11	0.09	8.69	0.08	0.09	0.051	5.10	0.05	0.06
Deeg (M)	0.13	12.65	0.12	0.13	0.11	10.61	0.10	0.11	0.062	6.23	0.06	0.07
Kaman (M)	0.09	9.45	0.09	0.10	0.08	7.93	0.07	0.09	0.047	4.65	0.04	0.05
Kumher (M)	0.06	6.27	0.06	0.07	0.05	5.27	0.05	0.06	0.031	3.09	0.03	0.04
Weir (M)	0.05	5.41	0.05	0.06	0.05	4.54	0.04	0.05	0.027	2.67	0.02	0.03
Nadbai (M)	0.07	6.65	0.06	0.07	0.06	5.58	0.05	0.06	0.033	3.27	0.03	0.04
Nagar (M)	0.07	6.53	0.06	0.07	0.05	5.48	0.05	0.06	0.032	3.22	0.03	0.04

Table 3 *Continued*

District/Town	Met need for family planning (any method)				Unmet need for family planning							
	Proportion of women of age group 15-44 in tehsils using contraception	per cent	Lower limit	Upper limit	unmet need for family planning for limiting fertility	Per cent	Lower limit	Upper limit	unmet need for family planning for spacing fertility	Per cent	Lower limit	Upper limit
Bhilwara												
Gulabpura (M)	0.03	3.40	0.03	0.04	0.03	3.05	0.03	0.04	0.018	1.76	0.01	0.02
Gangapur (M)	0.03	2.52	0.02	0.03	0.02	2.26	0.02	0.03	0.013	1.31	0.01	0.02
Jahazpur (M)	0.03	2.78	0.02	0.03	0.02	2.50	0.02	0.03	0.014	1.44	0.01	0.02
Asind (M)	0.02	2.11	0.02	0.03	0.02	1.89	0.01	0.02	0.011	1.09	0.01	0.01
Bhilwara (MCI)	0.40	40.15	0.40	0.41	0.36	36.03	0.36	0.36	0.208	20.83	0.20	0.21
Beejoliya Kalan (CT)	0.02	1.83	0.01	0.02	0.02	1.64	0.01	0.02	0.009	0.95	0.00	0.01
Mandalgarh (M)	0.03	3.01	0.03	0.04	0.03	2.71	0.02	0.03	0.016	1.56	0.01	0.02
Shahpura (M)	0.04	4.18	0.04	0.05	0.04	3.75	0.03	0.04	0.022	2.17	0.02	0.03
Bikaner												
Bikaner (M CI)	0.54	53.89	0.54	0.54	0.47	47.43	0.47	0.48	0.276	27.61	0.27	0.28
Deshnoke (M)	0.02	1.67	0.01	0.02	0.01	1.47	0.01	0.02	0.009	0.86	0.01	0.01
Nokha (M)	0.05	5.15	0.05	0.06	0.05	4.54	0.04	0.05	0.026	2.64	0.02	0.03
Bundi												
Kaprain (M)	0.07	7.46	0.07	0.08	0.07	6.72	0.06	0.08	0.039	3.88	0.03	0.05
Bundi (M)	0.37	37.08	0.36	0.38	0.33	33.41	0.33	0.34	0.193	19.29	0.19	0.20
Nainwa (M)	0.06	6.38	0.06	0.07	0.06	5.75	0.05	0.07	0.033	3.32	0.03	0.04
Keshoraipatan (M)	0.09	8.88	0.08	0.10	0.08	8.01	0.07	0.09	0.046	4.62	0.04	0.05

Table 3 *Continued*

District/Town	Met need for family planning (any method)				Unmet need for family planning							
	Proportion of women of age group 15-44 in tehsils using contraception	per cent	Lower limit	Upper limit	unmet need for family planning for limiting fertility	Per cent	Lower limit	Upper limit	unmet need for family planning for spacing fertility	Per cent	Lower limit	Upper limit
Chittaurgarh												
Rawatbhata (M)	0.07	7.15	0.07	0.08	0.06	6.38	0.06	0.07	0.037	3.68	0.03	0.04
Pratapgarh (M)	0.07	7.38	0.07	0.08	0.07	6.58	0.06	0.07	0.038	3.80	0.03	0.04
Nimbahera (M)	0.11	11.23	0.11	0.12	0.10	10.02	0.09	0.11	0.058	5.78	0.05	0.06
Kapasan (M)	0.04	3.96	0.03	0.05	0.04	3.54	0.03	0.04	0.020	2.04	0.02	0.02
Begun (M)	0.04	4.06	0.03	0.05	0.04	3.62	0.03	0.04	0.021	2.09	0.02	0.03
Bari Sadri (M)	0.03	3.18	0.03	0.04	0.03	2.84	0.02	0.03	0.016	1.64	0.01	0.02
Chhoti Sadri (M)	0.04	3.52	0.03	0.04	0.03	3.14	0.03	0.04	0.018	1.81	0.01	0.02
Chittaurgarh (M)	0.20	19.93	0.19	0.20	0.18	17.79	0.17	0.18	0.103	10.26	0.10	0.11
Churu												
Rajaldesar (M)	0.03	3.43	0.03	0.04	0.03	3.02	0.03	0.04	0.018	1.76	0.01	0.02
Rajgarh (M)	0.04	3.58	0.03	0.04	0.03	3.15	0.03	0.04	0.018	1.84	0.01	0.02
Ratangarh (M)	0.09	9.49	0.09	0.10	0.08	8.35	0.08	0.09	0.049	4.87	0.05	0.05
Chhapar (M)	0.03	2.71	0.02	0.03	0.02	2.39	0.02	0.03	0.014	1.39	0.01	0.02
Ratannagar (M)	0.02	1.68	0.01	0.02	0.01	1.48	0.01	0.02	0.009	0.86	0.00	0.01
Sardarshahar (M)	0.12	11.98	0.11	0.12	0.11	10.54	0.10	0.11	0.061	6.15	0.06	0.07
Sujangarh (M)	0.13	12.53	0.12	0.13	0.11	11.02	0.11	0.11	0.064	6.43	0.06	0.07
Taranagar (M)	0.04	4.02	0.04	0.05	0.04	3.53	0.03	0.04	0.021	2.06	0.02	0.02
Bidasar (M)	0.04	4.47	0.04	0.05	0.04	3.93	0.03	0.04	0.023	2.29	0.02	0.03
Dungargarh (M)	0.07	6.75	0.06	0.07	0.06	5.94	0.05	0.06	0.035	3.46	0.03	0.04
Dausa												
Dausa (M)	0.17	17.45	0.17	0.18	0.24	24.10	0.23	0.25	0.140	14.00	0.13	0.15
Bandikui (M)	0.07	7.26	0.06	0.08	0.06	6.38	0.06	0.07	0.037	3.70	0.03	0.04
Lalsot (M)	0.13	12.99	0.12	0.14	0.11	11.41	0.11	0.12	0.066	6.63	0.06	0.07
Mandawar (CT)	0.04	4.49	0.04	0.05	0.04	3.94	0.03	0.05	0.023	2.29	0.02	0.03
Mahwa (CT)	0.32	32.18	0.31	0.34	0.08	7.72	0.07	0.09	0.045	4.48	0.04	0.05

Table 3 *Continued*

District/Town	Met need for family planning (any method)				Unmet need for family planning							
	Proportion of women of age group 15-44 in tehsils using contraception	per cent	Lower limit	Upper limit	unmet need for family planning for limiting fertility	Per cent	Lower limit	Upper limit	unmet need for family planning for spacing fertility	Per cent	Lower limit	Upper limit
Dhaulpur												
Rajakhera (M)	0.22	21.80	0.21	0.23	0.19	19.12	0.18	0.20	0.111	11.12	0.10	0.12
Bari (M)	0.39	39.17	0.38	0.40	0.34	34.35	0.33	0.35	0.200	19.99	0.19	0.21
Dungarpur												
Dungarpur (M)	0.31	31.31	0.30	0.32	0.28	27.69	0.27	0.29	0.160	15.97	0.15	0.17
Galiakot (CT)	0.05	5.10	0.04	0.06	0.05	4.51	0.03	0.06	0.026	2.60	0.02	0.03
Sagwara (M)	0.25	24.57	0.24	0.26	0.22	21.73	0.21	0.23	0.125	12.54	0.12	0.13
Ganganagar												
Suratgarh (M)	0.16	16.48	0.16	0.17	0.13	13.46	0.13	0.14	0.080	7.97	0.07	0.08
3STR (CT)	0.03	3.09	0.02	0.04	0.03	2.52	0.02	0.03	0.015	1.49	0.01	0.02
Anupgarh (M)	0.08	8.27	0.08	0.09	0.07	6.76	0.06	0.07	0.040	4.00	0.03	0.05
Gajsinghpur (M)	0.03	2.73	0.02	0.03	0.02	2.23	0.02	0.03	0.013	1.32	0.01	0.02
Karanpur (M)	0.06	5.90	0.05	0.07	0.05	4.82	0.04	0.05	0.029	2.85	0.02	0.03
Kesrisinghpur (M)	0.04	3.80	0.03	0.05	0.03	3.11	0.02	0.04	0.018	1.84	0.01	0.02
Padampur (M)	0.05	4.87	0.04	0.06	0.04	3.98	0.03	0.05	0.024	2.35	0.02	0.03
Raisinghnagar (M)	0.08	7.80	0.07	0.09	0.06	6.37	0.06	0.07	0.038	3.77	0.03	0.04
Sadulshahar (M)	0.06	6.35	0.06	0.07	0.05	5.19	0.05	0.06	0.031	3.07	0.03	0.04
Vijainagar (M)	0.05	5.06	0.04	0.06	0.04	4.14	0.03	0.05	0.024	2.45	0.02	0.03
Hanumangarh												
Bhadra (M)	0.07	7.17	0.07	0.08	0.06	6.24	0.06	0.07	0.036	3.65	0.03	0.04
Hanumangarh (M)	0.26	25.86	0.25	0.26	0.23	22.50	0.22	0.23	0.132	13.15	0.13	0.14
Nohar (M)	0.09	8.63	0.08	0.09	0.08	7.51	0.07	0.08	0.044	4.39	0.04	0.05
Rawatsar (M)	0.06	5.77	0.05	0.06	0.05	5.03	0.04	0.06	0.029	2.94	0.03	0.03
Pilibanga (M)	0.07	6.76	0.06	0.07	0.06	5.88	0.05	0.06	0.034	3.44	0.03	0.04
Sangaria (M)	0.07	6.97	0.06	0.08	0.06	6.07	0.06	0.07	0.035	3.55	0.03	0.04

Table 3 *Continued*

District/Town	**Met need for family planning (any method)**				**Unmet need for family planning**							
	Proportion of women of age group 15-44 in tehsils using contraception	per cent	Lower limit	Upper limit	unmet need for family planning for limiting fertility	Per cent	Lower limit	Upper limit	unmet need for family planning for spacing fertility	Per cent	Lower limit	Upper limit
Jaipur												
Bagru (M)	0.01	0.56	0.00	0.01	0.00	0.45	0.00	0.01	0.003	0.27	0.00	0.02
Chaksu (M)	0.01	0.74	0.01	0.01	0.01	0.60	0.00	0.01	0.004	0.36	0.00	0.01
Chomu (M)	0.01	1.29	0.01	0.02	0.01	1.04	0.01	0.01	0.006	0.62	0.00	0.01
Jaipur (M Corp.)	0.58	57.94	0.58	0.58	0.47	46.70	0.47	0.47	0.278	27.82	0.25	0.28
Kishangarh Renwal (M)	0.01	0.71	0.00	0.01	0.01	0.57	0.00	0.01	0.003	0.34	0.00	0.03
Kotputli (M)	0.01	1.00	0.01	0.01	0.01	0.81	0.01	0.01	0.005	0.48	0.00	0.01
Sambhar (M)	0.01	0.57	0.00	0.01	0.00	0.46	0.00	0.01	0.003	0.27	0.00	0.04
Phulera (M)	0.01	0.55	0.00	0.01	0.00	0.45	0.00	0.01	0.003	0.27	0.00	0.02
Jobner (M)	0.00	0.27	0.00	0.00	0.00	0.22	0.00	0.00	0.001	0.13	0.00	0.02
Shahpura (M)	0.01	0.71	0.00	0.01	0.01	0.58	0.00	0.01	0.003	0.34	0.00	0.01
Viratnagar (M)	0.00	0.44	0.00	0.01	0.00	0.35	0.00	0.01	0.002	0.21	0.00	0.02
Jaisalmer												
Jaisalmer (M)	0.44	44.00	0.43	0.45	0.40	39.89	0.39	0.41	0.231	23.07	0.22	0.24
Pokaran (M)	0.15	15.33	0.14	0.16	0.14	13.90	0.13	0.15	0.080	8.04	0.07	0.09
Jalor												
Jalor (M)	0.23	23.48	0.23	0.24	0.22	21.95	0.21	0.23	0.126	12.56	0.12	0.13
Bhinmal (M)	0.21	20.92	0.20	0.22	0.20	19.55	0.19	0.20	0.112	11.19	0.10	0.12
Sanchore (M)	0.14	13.77	0.13	0.15	0.13	12.87	0.12	0.14	0.074	7.36	0.07	0.08

Table 3 *Continued*

District/Town	Met need for family planning (any method)				Unmet need for family planning							
	Proportion of women of age group 15-44 in tehsils using contraception	per cent	Lower limit	Upper limit	unmet need for family planning for limiting fertility	Per cent	Lower limit	Upper limit	unmet need for family planning for spacing fertility	Per cent	Lower limit	Upper limit
Jhalawar												
Aklera (M)	0.07	6.53	0.06	0.07	0.06	5.80	0.05	0.07	0.034	3.36	0.03	0.04
Bhawani Mandi (M)	0.13	12.67	0.12	0.13	0.11	11.25	0.11	0.12	0.065	6.52	0.06	0.07
Bakani (CT)	0.03	2.87	0.02	0.04	0.03	2.55	0.02	0.03	0.015	1.48	0.01	0.02
Jhalawar (M)	0.17	17.24	0.16	0.18	0.15	15.31	0.15	0.16	0.089	8.87	0.08	0.09
Jhalrapatan (M)	0.11	10.85	0.10	0.12	0.10	9.64	0.09	0.10	0.056	5.58	0.05	0.06
Kolvi Mandi Rajendrapur (CT)	0.03	2.83	0.02	0.04	0.03	2.51	0.02	0.03	0.015	1.46	0.01	0.02
Manoharthana (CT)	0.03	3.34	0.03	0.04	0.03	2.97	0.02	0.04	0.017	1.72	0.01	0.02
Pirawa (M)	0.04	4.11	0.03	0.05	0.04	3.65	0.03	0.04	0.021	2.12	0.02	0.03
Jhunjhunun												
Baggar (M)	0.03	2.76	0.02	0.03	0.02	2.29	0.02	0.03	0.013	1.35	0.01	0.02
Bissau (M)	0.04	4.38	0.04	0.05	0.04	3.63	0.03	0.04	0.021	2.14	0.02	0.03
Chirawa (M)	0.07	7.47	0.07	0.08	0.06	6.19	0.06	0.07	0.036	3.65	0.03	0.04
Jhunjhunun (M)	0.20	20.17	0.20	0.21	0.17	16.72	0.16	0.17	0.099	9.85	0.09	0.10
Mandawa (M)	0.04	4.31	0.04	0.05	0.04	3.58	0.03	0.04	0.021	2.11	0.02	0.03
Surajgarh (M)	0.04	3.76	0.03	0.04	0.03	3.12	0.03	0.04	0.018	1.84	0.01	0.02
Udaipurwati (M)	0.06	5.62	0.05	0.06	0.05	4.66	0.04	0.05	0.027	2.74	0.02	0.03
Mukandgarh (M)	0.04	3.67	0.03	0.04	0.03	3.04	0.02	0.04	0.018	1.79	0.01	0.02
Nawalgarh (M)	0.12	11.55	0.11	0.12	0.10	9.57	0.09	0.10	0.056	5.64	0.05	0.06
Jodhpur												
Bilara (M)	0.33	32.88	0.32	0.34	0.29	28.83	0.28	0.30	0.168	16.82	0.16	0.18
Pipar City (M)	0.28	27.94	0.27	0.29	0.24	24.50	0.23	0.26	0.143	14.29	0.13	0.15

Table 3 *Continued*

District/Town	Met need for family planning (any method)				Unmet need for family planning							
	Proportion of women of age group 15-44 in tehsils using contraception	per cent	Lower limit	Upper limit	unmet need for family planning for limiting fertility	Per cent	Lower limit	Upper limit	unmet need for family planning for spacing fertility	Per cent	Lower limit	Upper limit
Karauli												
Hindaun (M)	0.30	29.65	0.29	0.30	0.26	26.20	0.26	0.27	0.153	15.25	0.15	0.16
Karauli (M)	0.23	23.38	0.23	0.24	0.21	20.66	0.20	0.21	0.120	12.02	0.11	0.13
Todabhim (M)	0.07	7.46	0.07	0.08	0.07	6.60	0.06	0.07	0.038	3.84	0.03	0.04
Kota												
Chechat (CT)	0.05	4.83	0.04	0.06	0.04	3.97	0.03	0.05	0.024	2.35	0.02	0.03
Kaithoon (M)	0.10	9.79	0.09	0.11	0.08	8.04	0.07	0.09	0.048	4.77	0.04	0.05
Kumbhkot (CT)	0.03	2.72	0.02	0.04	0.02	2.23	0.01	0.03	0.013	1.32	0.01	0.02
Modak (CT)	0.04	3.96	0.03	0.05	0.03	3.25	0.02	0.04	0.019	1.93	0.01	0.03
Suket (CT)	0.08	8.01	0.07	0.09	0.07	6.57	0.06	0.07	0.039	3.90	0.03	0.05
Satalkheri (CT)	0.07	6.96	0.06	0.08	0.06	5.71	0.05	0.07	0.034	3.39	0.03	0.04
Ramganj Mandi (M)	0.15	14.68	0.14	0.16	0.12	12.05	0.11	0.13	0.071	7.15	0.07	0.08
Sangod (M)	0.09	8.91	0.08	0.10	0.07	7.32	0.06	0.08	0.043	4.34	0.04	0.05
Udpura (CT)	0.04	4.06	0.03	0.05	0.03	3.33	0.02	0.04	0.020	1.98	0.01	0.03
Nagaur												
Basni Belima (CT)	0.05	4.71	0.04	0.05	0.04	4.11	0.04	0.05	0.024	2.40	0.02	0.03
Didwana (M)	0.09	9.28	0.09	0.10	0.08	8.11	0.08	0.09	0.047	4.73	0.04	0.05
Goredi Chancha (CT)	0.02	2.01	0.01	0.03	0.02	1.75	0.01	0.02	0.010	1.02	0.01	0.01
Kuchaman City (M)	0.11	10.64	0.10	0.11	0.09	9.29	0.09	0.10	0.054	5.43	0.05	0.06
Kuchera (M)	0.04	4.09	0.03	0.05	0.04	3.57	0.03	0.04	0.021	2.09	0.02	0.03
Ladnu (M)	0.12	12.10	0.12	0.13	0.11	10.57	0.10	0.11	0.062	6.17	0.06	0.07
Nawa (M)	0.04	3.75	0.03	0.04	0.03	3.27	0.03	0.04	0.019	1.91	0.01	0.02
Parbatsar (M)	0.03	2.76	0.02	0.03	0.02	2.41	0.02	0.03	0.014	1.41	0.01	0.02
Merta City (M)	0.08	8.25	0.08	0.09	0.07	7.21	0.07	0.08	0.042	4.21	0.04	0.05
MundwÃą (M)	0.03	3.41	0.03	0.04	0.03	2.98	0.02	0.04	0.017	1.74	0.01	0.02

Table 3 *Continued*

District/Town	**Met need for family planning (any method)**				**Unmet need for family planning**							
	Proportion of women of age group 15-44 in tehsils using contraception	per cent	Lower limit	Upper limit	unmet need for family planning for limiting fertility	Per cent	Lower limit	Upper limit	unmet need for family planning for spacing fertility	Per cent	Lower limit	Upper limit
Pali												
Bali (M)	0.03	2.90	0.02	0.03	0.03	2.56	0.02	0.03	0.015	1.49	0.01	0.02
Falna (M)	0.03	3.26	0.03	0.04	0.03	2.89	0.02	0.03	0.017	1.68	0.01	0.02
Jaitaran (M)	0.03	3.00	0.02	0.04	0.03	2.65	0.02	0.03	0.015	1.54	0.01	0.02
Pali (MCI)	0.29	28.67	0.28	0.29	0.25	25.34	0.25	0.26	0.147	14.72	0.14	0.15
Rani (M)	0.02	1.95	0.01	0.02	0.02	1.72	0.01	0.02	0.010	1.00	0.01	0.01
Sadri (M)	0.04	3.97	0.03	0.04	0.04	3.51	0.03	0.04	0.020	2.04	0.02	0.02
Sumerpur (M)	0.05	4.86	0.04	0.05	0.04	4.30	0.04	0.05	0.025	2.50	0.02	0.03
Takhatgarh (M)	0.03	2.51	0.02	0.03	0.02	2.22	0.02	0.03	0.013	1.29	0.01	0.02
Sojat (M)	0.06	6.05	0.06	0.07	0.05	5.35	0.05	0.06	0.031	3.11	0.03	0.03
Sojat Road (CT)	0.02	1.76	0.01	0.02	0.02	1.56	0.01	0.02	0.009	0.91	0.01	0.01
Marwar Junction (CT)	0.02	1.65	0.01	0.02	0.01	1.45	0.01	0.02	0.008	0.85	0.00	0.01
Rajsamand												
Nathdwara (M)	0.24	24.43	0.23	0.25	0.21	21.31	0.20	0.22	0.124	12.41	0.12	0.13
Rajsamand (M)	0.37	36.80	0.36	0.38	0.32	32.09	0.31	0.33	0.187	18.70	0.18	0.19
Sawai Madhopur												
Todra (CT)	0.60	59.98	0.57	0.63	0.54	53.82	0.51	0.57	0.311	31.11	0.28	0.34

Table 3 *Continued*

District/Town	**Met need for family planning (any method)**				**Unmet need for family planning**							
	Proportion of women of age group 15-44 in tehsils using contraception	per cent	Lower limit	Upper limit	unmet need for family planning for limiting fertility	Per cent	Lower limit	Upper limit	unmet need for family planning for spacing fertility	Per cent	Lower limit	Upper limit
Sikar												
Reengus (M)	0.05	5.29	0.05	0.06	0.04	4.50	0.04	0.05	0.026	2.64	0.02	0.03
Neem-Ka-Thana (M)	0.07	6.89	0.06	0.08	0.06	5.86	0.05	0.06	0.034	3.44	0.03	0.04
Ramgarh (M)	0.07	6.94	0.06	0.08	0.06	5.90	0.05	0.06	0.035	3.47	0.03	0.04
Sri Madhopur (M)	0.07	6.71	0.06	0.07	0.06	5.71	0.05	0.06	0.034	3.35	0.03	0.04
Lachhmangarh (M)	0.11	11.26	0.11	0.12	0.10	9.58	0.09	0.10	0.056	5.62	0.05	0.06
Losal (M)	0.06	6.28	0.06	0.07	0.05	5.34	0.05	0.06	0.031	3.14	0.03	0.04
Fatehpur (M)	0.19	18.91	0.18	0.20	0.16	16.09	0.16	0.17	0.094	9.45	0.09	0.10
Sirohi												
Mount Abu (M)	0.12	11.69	0.11	0.13	0.11	10.55	0.10	0.12	0.061	6.11	0.05	0.07
Pindwara (M)	0.12	12.19	0.11	0.13	0.11	11.00	0.10	0.12	0.064	6.37	0.06	0.07
Sheoganj (M)	0.15	14.78	0.14	0.16	0.13	13.34	0.12	0.14	0.077	7.72	0.07	0.08
Sirohi (M)	0.21	20.87	0.20	0.22	0.19	18.84	0.18	0.20	0.109	10.91	0.10	0.12

Table 3 *Continued*

District/Town	Met need for family planning (any method)				Unmet need for family planning							
	Proportion of women of age group 15-44 in tehsils using contraception	per cent	Lower limit	Upper limit	unmet need for family planning for limiting fertility	Per cent	Lower limit	Upper limit	unmet need for family planning for spacing fertility	Per cent	Lower limit	Upper limit
Tonk												
Banasthali (CT)	0.03	2.51	0.02	0.03	0.02	2.27	0.02	0.03	0.013	1.31	0.01	0.02
Deoli (M)	0.05	5.02	0.04	0.06	0.05	4.54	0.04	0.05	0.026	2.62	0.02	0.03
Uniara (M)	0.03	2.86	0.02	0.04	0.03	2.59	0.02	0.03	0.015	1.49	0.01	0.02
Todaraisingh (M)	0.06	5.57	0.05	0.06	0.05	5.04	0.04	0.06	0.029	2.91	0.02	0.03
Niwai (M)	0.08	8.08	0.07	0.09	0.07	7.31	0.07	0.08	0.042	4.21	0.04	0.05
Tonk (M CI)	0.36	35.60	0.35	0.36	0.32	32.20	0.32	0.33	0.186	18.57	0.18	0.19
Udaipur												
Dhariawad (CT)	0.01	1.37	0.01	0.02	0.01	1.18	0.01	0.02	0.007	0.69	0.00	0.01
Bhinder (M)	0.02	2.15	0.02	0.03	0.02	1.86	0.01	0.02	0.011	1.09	0.01	0.01
Kherwara Chhaoni (CT)	0.01	0.85	0.00	0.01	0.01	0.74	0.00	0.01	0.004	0.43	0.00	0.01
Fatehnagar (M)	0.03	2.57	0.02	0.03	0.02	2.22	0.02	0.03	0.013	1.30	0.01	0.02
Kanor (M)	0.02	1.55	0.01	0.02	0.01	1.34	0.01	0.02	0.008	0.78	0.00	0.01
Salumbar (M)	0.02	2.07	0.02	0.03	0.02	1.79	0.01	0.02	0.010	1.05	0.01	0.01
Rikhabdeo (CT)	0.01	1.02	0.01	0.02	0.01	0.88	0.00	0.01	0.005	0.52	0.00	0.01
Udaipur (M CI)	0.49	49.21	0.49	0.50	0.42	42.47	0.42	0.43	0.248	24.84	0.24	0.25
Bhalariya (CT)	0.01	0.82	0.00	0.01	0.01	0.71	0.00	0.01	0.004	0.42	0.00	0.01

STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 361—363, DOI 10. 21307

REPORT

The XXXV International Conference on Multivariate Statistical Analysis, 7–9 November 2016, Łódź, Poland

The 35th edition of the International Conference on Multivariate Statistical Analysis was held in Łódź, Poland, on November 7-9, 2016. The MSA 2016 conference was organized by the Department of Statistical Methods of the University of Lodz, the Institute of Statistics and Demography of the University of Lodz, the Polish Statistical Association and the Committee on Statistics and Econometrics of Polish Academy of Sciences. The Mayor of the City of Łódź, Hanna Zdanowska, took the honorary patronage of the Multivariate Statistical Analysis MSA 2016 conference. Its organization was financially supported by the National Bank of Poland, the Polish Academy of Sciences and StatSoft Polska Sp. z o.o. The Organizing Committee was headed by Professor Czesław Domański. The scientific secretaries included Piotr Szczepocki, M.Sc. and Jacek Białek, Assistant Professor from the Department of Statistical Methods of the University of Lodz.

As all previous Multivariate Statistical Analysis conferences, the 2016 edition was aimed at creating the opportunity for scientists and practitioners of statistics to present and discuss the latest theoretical achievements in the field of the multivariate statistical analysis, its practical aspects and applications. The scientific programme covered various statistical problems, including multivariate estimation methods, statistical tests, non-parametric inference, discrimination analysis, Monte Carlo analysis, Bayesian inference, application of statistical methods in finance and economy, especially methods used in capital market and risk management. The range of topics also included the design of experiments and survey sampling methodology, mainly for the social science purposes.

The conference was attended by 78 participants from main academic centres in Poland (Białystok, Katowice, Kraków, Olsztyn, Opole, Poznań, Rzeszów, Szczecin, Toruń, Warszawa and Wrocław) and from abroad (Italy). In 17 sessions 64 papers were presented, including 2 invited lectures.

The conference was opened by the Head of the Organizing Committee, Professor Czesław Domański. The subsequent speakers at the conference opening included Dominik Rozkrut, the Head of Central Statistical Office of Poland and Assistant Professor Michał Przybyliński, the Deputy Dean of the Faculty of Economics and Sociology of the University of Lodz.

After the opening ceremony, all participants had the opportunity to attend the invited lecture by Professor Józef Pociecha (Cracow University of Economics) *Statistics Of Poland – The First Yearbook Of Polish Lands*. The second invited lecture was presented by Professor Włodzimierz Okrasa (Cardinal Stefan Wyszyński University in Warsaw) and was tilted *The Quality Of Life and The Surroundings.*

Among regular conference sessions, the historical plenary session was held, chaired by Professor Włodzimierz Okrasa (Cardinal Stefan Wyszyński University in Warsaw), and dedicated to eminent Polish scientists. Professor Czesław Domański (University of Lodz) presented statistical threads in the work of Jan Śniadecki. Professor Mirosław Krzyśko (Adam Mickiewicz University in Poznań) recalled his memories of Mikołaj Olekiewicz. Professor Jan Kordos (Warsaw Management University) presented the work of Jan Buga.

Other sessions were chaired respectively by:

SESSION II	Professor Danuta Strahl (Wrocław University of Economics)
SESSION III A	Professor Małgorzata Markowska (Wrocław University of Economics)
SESSION III B	Professor Andrzej Dudek (Wrocław University of Economics)
SESSION IV A	Professor Grzegorz Kończak (University of Economics in Katowice)
SESSION IV B	Professor Grażyna Trzpiot (University of Economics in Katowice)
SESSION V	Professor Bronisław Ceranka (Poznań University of Life Sciences)
SESSION VI A	Professor Grażyna Dehnel (Poznań University of Economics)
SESSION VI B	Professor Tadeusz Bednarski (University of Wrocław)
SESSION VII A	Professor Wojciech Zieliński (Warsaw University of Life Sciences)
SESSION VII B	Professor Jerzy Korzeniewski (University of Łódź)
SESSION VIII A	Professor Alina Jędrzejczak (University of Łódź)
SESSION VIII B	Professor Iwona Markowicz (University of Szczecin)
SESSION IX A	Professor Janusz Wywiał (University of Economics in Katowice)
SESSION IX B	Professor Marek Walesiak (Wrocław University of Economics)
SESSION X	Professor Andrzej Sokołowski (Cracow University of Economics)
SESSION XI	Professor Józef Dziechciarz (Wrocław University of Economics)

The MSA 2016 conference was closed by the Chairman of the Organizing Committee, Professor Czesław Domański, who summarized the Conference and thanked all the guests, conference partners and sponsors.

The next edition of Multivariate Statistical Analysis Conference MSA 2017 is planned on November 6-8, 2017 and will be held in Łódź, Poland. The Chairman of the Organizing Committee, Professor Czesław Domański, informed that this will be the 36th edition of the conference and kindly invited all interested scientists, researchers and students to participate.

Prepared by:

Piotr Szczepocki
Department of Statistical Methods, University of Lódź.

Jacek Białek
Department of Statistical Methods, University of Lódź

STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 365–367, DOI 10. 21307

ABOUT THE AUTHORS

Alkaya Aylin received her PhD from the Gazi University in 2010. She is currently Assistant Professor at Nevşehir Haci Bektaş Veli University (Turkey). Her research activities mainly focus on statistical methods, probability sampling techniques and regression analysis.

Ayhan Öztaş H. received his PhD from the University of Wales in 1978. He is currently Emeritus Professor of Statistics and affiliated faculty at METU (Turkey). His major research interests focus on survey sampling techniques, web survey methodology and components of the total survey error.

Beevi Nazeema T. is a statistical investigator at Directorate of Economics and Statistics and received his/her PhD in the Department of Statistics, University of Calicut. His/her research interests include sampling theory (two-phase sampling, super population model, predictive model), statistical inference and data analysis in particular, small area estimation and item response theory. He/she is a lifelong member of Kerala Statistical Association (KSA) and also a reviewer of Journal of Applied Probability and Statistics (2016).

Bhattacharya Manjima is a Data Analyst at Credit Suisse, a Suisse multinational financial services holding company. She received her MStat in 2016 from Indian Statistical Institute with specialization in Bio-Statistics. At Credit Suisse, she is involved in statistical data analysis in the area of mathematical finance.

Chandran C. is Professor of the Department of Statistics, University of Calicut. His research interests include probability modelling, statistical inference and extreme value theory. He is a lifelong member of an Indian Society of Probability and Statistics and also Kerala Statistical Association (KSA).

Dihidar Kajal is Assistant Professor in Sampling and Official Statistics Unit of Indian Statistical Institute, Kolkata, India. In 2010, she received her PhD in Statistics in the area of survey sampling methodology, in particular in adaptive sampling, model-cum-design based estimation, randomized response techniques, from Indian Statistical Institute under the supervision of Prof. Mausumi Bose. She is interested in theory and estimation in different aspects of sample survey methodologies, small area estimation, protection of privacy in sensitive population characteristic estimation, spatial statistics, multivariate statistical inference, classification methods and data analysis in particular. She is a life member of Calcutta Statistical Association.

Ercan Ilker is Professor at the Department of Biostatistics at Medical School in Uludag University, Bursa, Turkey. He has experience in the analysis of several types of data sets. His research area is principally focused on statistical shape analysis and modern morphometry, and his studies also deal with reliability, allometry and growth models, and biostatistics education. He holds membership in many Journal Editorships and Advisory Boards.

Esin Alptekin received his PhD from the Ankara University in 1973. He is now Emeritus Professor of Statistics at Gazi University (Turkey). His research activities focus on the fields of statistical methods, probability sampling methods and operations research techniques.

Joshi Hemlata is Assistant Professor of Statistics at Manipal University, Rajasthan, India. She received her PhD in Statistics, Mathematical Demography in particular, in 2015 from Banasthali University, Rajasthan, India. She has produced good articles along with one book to her credit. Her research interests include population sciences, fertility and mortality modelling and applied statistics. She is a member of many professional and academic societies and has actively participated in conferences and seminars.

Karadeniz Pinar Gunel is Assistant Professor at the Department of Biostatistics at Medical School in Sanko University, Gaziantep, Turkey. She received her PhD in the field of Biostatistics in 2015 from Institute of Health Science in Uludag University under the supervision of Prof. Ilker Ercan. Her scientific research area is in survival analyses and validity and reliability of health related quality of life questionnaires. She is a member of Turkish Biostatistics Association.

Łukaszonek Wojciech is working as an assistant at President Wojciechowski Higher Vocational State School in Kalisz, Poland. In 2002 he received his MSc in Mathematics from Technical University of Łódź. Currently he is doing a PhD in the field of analysis of higher educational market at Poznań University of Economics and Business.

Pareek Sarla is Professor of Statistics and Head of the Department of Mathematics and Statistics at Banasthali University, Rajasthan, India. Her research interests are in scaling, forecasting and modelling, demography, applied statistics and inventory theory. She has produced 17 PhD and 12 MPhil dissertations to her academic credit. Dr Pareek has published more than 65 articles, 2 books and many book chapters in various journals/publishers of national and international repute. She is a member of several academic societies and a member of editorial boards of many journals. She carries out good academic projects sponsored by government and private organizations. She has more than 30 years of academic and administrative experience, received academic excellence, awards, and actively participated/organized in the several academic workshops/conferences and seminars.

Rai Piyush Kant is Professor of Statistics at the Department of Mathematics and Statistics at Banasthali University, Rajasthan, India. He is interested in sample surveys, small area estimation, Bayesian modelling, regression modelling, mathematical demography and modelling of population data, bio-statistics, survival analysis. He received a gold medal and other awards in his academic career. Dr Rai has published several articles, books, book chapters in various journals/publishers of national and international repute. He works in active collaboration with various societies like ASI, IBS, ISBA, and SSCA, among others. He is a member of editorial boards of many journals. He has handled government-sponsored academic projects and actively participated/organized more than 50 academic conferences/seminars in his career.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, *Key words* (in bold italics) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).