STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

CONTENTS

Volume 20, Number 1, March 2019

EDITORIAL OFFICE **ISSN 1234-7655**

Scientific Secretary
Marek Cierpiał-Wolan, e-mail: m.wolan@stat.gov.pl
Secretary
Patryk Barszcz, e-mail: P.Barszcz@stat.gov.pl, phone number 00 48 22 — 608 33 66
Technical Assistant
Rajmund Litkowiec, e-mail: r.litkowiec@stat.gov.pl

Address for correspondence
Statistics Poland, al. Niepodległości 208, 00-925 Warsaw, Poland, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 1−4

FROM THE EDITOR

This issue of the *Statistics in Transition new series* is slightly modified in strictly editorial terms. Namely, we resign from traditionally maintained distinction between papers grouped, respectively, under the headings 'sample and estimation issues' and 'research papers'. From now on, they will be combined into one section, as original research papers. The other sections remain unchanged. In this way we are getting a possibility to place the paper we consider of a special interest to the readers at the forefront. In order to ensure the appropriate paper is obtained we are arranging - with the help of the distinguished members of our Editorial Board and Associate Editors panel – for an invited paper to be submitted by renewed scholars and leading experts. The paper by Danny Pfeffermann et al., which inaugurates this issue, provides an example of this new approach.

This issue contains a set of nine papers, characterized briefly below.

Danny Pfeffermann's, Dano Ben-Hur's and Olivia Blum's paper ***Planning the next Census for Israel*** is devoted to currently hot topic of designing census while dealing with challenges beyond those considered 'standard' in methodology of such a type of research. They are posed by the fact that despite having fairly accurate population register at the national level (consisting of about 9 million persons), Israel has much less accurate the register for small geographical (statistical) areas, with an average area enumeration error of about 13%. In order to correct the errors at the area level the following three-step procedure is employed: (i) draw a sample from an enhanced register to obtain initial direct sample estimates for the number of persons residing in each area on “census day”; (ii) fit the Fay-Herriot model to the direct estimates in an attempt to improve their accuracy; (iii) compute a final census estimate for each statistical area as a linear combination of the estimate obtained in step (ii) and the register figure. The authors also consider a procedure to deal with not missing at random (NMAR) nonresponse (in step i) - the proposed procedures are illustrated using data from the 2008 Census in Israel.

The paper ***Imputation of missing values using raw moments*** by **Muhammed Umair Sohail, Javid Shabbir** and **Fariha Sohil** presents a method of imputation that has been found to be a cheaper procedure from a cost point of view in a situation when the sample data have missing values. The authors' study shows the improvement of the performance of imputation procedures by utilizing the raw moments of the auxiliary information rather than their ranks, especially, when the ranking of the auxiliary variable is expensive or difficult to achieve. Equations for bias and mean squared error are obtained by large sample-based approximation. Through the numerical and simulation studies it can be easily understood that the proposed method of imputation can outperform their counterparts.

Kumari Priyanka's and **Pidugu Trisandhya's** paper ***Modelling sensitive issues on successive waves*** addresses the problem of estimation of population mean of a sensitive character using non-sensitive auxiliary variable at current wave

in two wave successive sampling. A general class of estimator is proposed and studied under randomized and scrambled response model. Many existing estimators have been modified to work for sensitive population mean estimation. The modified estimators became the members of the proposed general class of estimators. The detail properties of all the estimators have been discussed. Their behaviour under randomized and scrambled response techniques have been elaborated. Numerical illustrations including simulation have been accompanied to judge the performance of different estimators. Finally, suitable recommendations are forwarded.

In the next article, ***Nonrandomized response model for complex survey designs*** by **Raghunath Arnab, Dahud Kehinde Shangodoyin** and **Antonio Arcos** discuss Warner's randomized response (RR) model, which is used to collect sensitive information for a broad range of surveys but possesses several limitations – such as lack of reproducibility and higher costs; and it is not feasible for mail questionnaires. To overcome such difficulties, nonrandomized response (NRR) surveys have been proposed. The proposed NRR surveys are limited to simple random sampling with replacement (SRSWR) design. In this paper, NRR procedures are extended to complex survey designs in a unified set-up, which is applicable to any sampling design and wider classes of estimators. Existing results for NRR can be derived from the proposed method as special cases.

Ramakrishnaiah Y. S., Manish Trivedi and **Konda Satish** in the paper ***On the smoothed parametric estimation of mixing proportion under fixed design regression model*** re-examine the estimator proposed by Boes (1966) – James (1978), herein called BJ estimator, which was constructed for estimating mixing proportion in a mixed model based on independent and identically distributed (i.i.d.) random samples. They also propose a completely new (smoothed) estimator for mixing proportion based on independent and not identically distributed (non-i.i.d.) random samples. The proposed estimator is nonparametric in true sense based on known "kernel function" as described in the introduction. The following results of the smoothed estimator have been checked under the non-i.i.d. set-up: (i) its small sample behaviour as compared with the unsmoothed version (BJ estimator) based on their mean square errors by using Monte-Carlo simulation, and established percentage gain in precision of smoothed estimator over its unsmoothed version measured in terms of their mean square error, (ii) its large sample properties such as almost surely (a.s.) convergence and asymptotic normality of these estimators. These results are completely new in the literature not only under the case of i.i.d., but can be generalised to non-i.i.d. design as well.

The next article, ***The Odd generalized exponential log-logistic distribution group acceptance sampling plan*** by **Devireddy Charana Udaya Sivakumar, Rosaiah Kanaparthi, Gadde Srinivasa Rao,** and **Kruthiventi Kalyani** presents a group acceptance sampling plan (GASP) being developed when the lifetime of the items follows odd generalized exponential log-logistic distribution (OGELLD), and the multiple number of items as a group can be tested simultaneously in a tester. The design parameters such as the minimum group size and the acceptance number are derived under specified the consumer's risk and the test termination time. The operating characteristic (OC) function values are calculated (intended) according to various quality levels, and the minimum ratios of the true average life to the specified average life at the specified producer's risk are derived. The

methodology is illustrated using real data on health and survival times of guinea pigs in days, and results show that the proposed methodology performs well as compared with existing sampling plans.

Soma Dhar, Lipi. B. Mahanta and **Kishore. K. Das** in the paper ***Formulation of the simple Markovian model using fractional calculus approach and its application to analysis of queue behaviour of severe patients*** introduce a fractional order of a simple Markovian model where the arrival rate of the patient is Poisson, i.e. independent of the patient size. Fraction is obtained by replacing the first order time derivative in the difference differential equations, which govern the probability law of the process with the Mittag-Leffler function. The probability distribution of the number N(t) of patients suffering from severe disease at an arbitrary time t is derived. Also, the authors obtain the mean size (number) of the patients suffering from severe disease waiting for service at any given time t, in the form of En 0:5;0:5(t), for different fractional values of server activity status, n = 1;0:95;0:90 and for arrival rates a = b = 0:5. A numerical example is also evaluated and analysed by using the simple Markovian model with the help of simulation techniques.

In the article ***Functional data analysis and its application to local damage detection*** by **Jacek Leśkow** and **Maria Skupien** vibration signals sampled with a high frequency are used as a basic source of information about machine behaviour. Few minutes of signal observations easily translate into several millions of data points to be processed with the purpose of the damage detection. Big dimensionality of data sets creates serious difficulties with detection of frequencies specific for a particular local damage. In view of that, traditional spectral analysis tools like spectrograms should be improved to efficiently identify the frequency bands where the impulsivity is most marked (the so-called informative frequency bands or IFB). The authors propose the functional approach known in modern time series analysis to overcome these difficulties. The data sets are treated as collections of random functions to apply techniques of the functional data analysis (FDA). In effect, massive data sets can be represented by few real-valued functions and corresponding parameters, which are the eigen-functions and eigen-values of the covariance operator describing the signal. Also, a new technique based on the bootstrap resampling is proposed to choose the optimal dimension in representing big data sets under processing. Using real data generated by a gearbox and a wheel bearings, it is demonstrated how these techniques work in practice.

Sukanya Intarapak's and **Thidaporn Supapakorn's** paper, ***An alternative matrix transformation to the F test statistic for clustered data*****,** discusses the problem of regression analysis of clustered data when the error of cluster data violates the independence assumption. Consequently, the OLS based test statistic leads to incorrect inferences. To overcome this shortcoming, the transformation is required to apply to the observations. The authors propose an alternative matrix transformation that adjusts the intra-cluster correlation with Householder matrix and apply it to the F test statistic based on GLS (generalized least squares) procedures for the regression coefficients hypothesis. By Monte Carlo simulations of the balanced and unbalanced data, it is found that the F test statistic based on the GLS procedures, with Adjusted Householder transformation, performs well in terms of the type I error rate and power of the test.

In the last article (in the *other articles* section), ***Survival regression models for single event and competing risks based on pseudo-observations*** by **Ewa Wycinka** and **Tomasz Jurkiewicz,** a survival data problem associated with the presence of censored observations is discussed. If no censoring occurs in the data, standard statistical models could be employed to analyse them. Pseudo-observations can replace censored observations and thereby allow standard statistical models to be used. Authors apply a pseudo-observation approach to single-event and competing-risks analysis. In the empirical part of the study, the use of regression models based on pseudo-observations in credit-risk assessment was investigated. Default, defined as a delay in payment, was considered to be the event of interest, while prepayment of credit was treated as a possible competing risk. Credits that neither default nor are prepaid during the follow-up were censored observations. Typical application characteristics of the credit and creditor were the covariates in the regression model. In a sample of retail credits provided by a Polish financial institution, regression models based on pseudo-observations were built for the single-event and competing-risks approaches. Estimates and discriminatory power of these models were compared to the Cox PH and Fine-Gray models.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 5

SUBMISSION INFORMATION FOR AUTHORS

Statistics in Transition new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and the Statistics 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:
P.Barszcz@stat.gov.pl,
GUS/Statistics Poland,
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/

EDITORIAL POLICY

The broad objective of *Statistics in Transition new series* is to advance the statistical and associated methods used primarily by statistical agencies and other research institutions. To meet that objective, the journal encompasses a wide range of topics in statistical design and analysis, including survey methodology and survey sampling, census methodology, statistical uses of administrative data sources, estimation methods, economic and demographic studies, and novel methods of analysis of socio-economic and population data. With its focus on innovative methods that address practical problems, the journal favours papers that report new methods accompanied by real-life applications. Authoritative review papers on important problems faced by statisticians in agencies and academia also fall within the journal's scope.

***

ABSTRACTING AND INDEXING DATABASES

Statistics in Transition new series is currently covered in:

- BASE – Bielefeld Academic Search Engine
- Central and Eastern European Online Library (CEEOL)
- Central European Journal of Social Sciences and Humanities (CEJSH)
- CNKI Scholar (China National Knowledge Infrastructure)
- Current Index to Statistics (CIS)
- Dimensions
- EconPapers
- Elsevier - Scopus
- ERIH Plus
- Google Scholar
- Index Copernicus
- J-Gate
- JournalGuide
- JournalTOCs
- Keepers Registry
- MIAR
- OpenAIRE
- ProQuest - Summon
- Publons
- RePec
- WorldCat
- Zenodo

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 7–19, DOI 10.21307/stattrans-2019-001

PLANNING THE NEXT CENSUS FOR ISRAEL

Danny Pfeffermann[1], **Dano Ben-Hur**[2] **,Olivia Blum**[3]

ABSTRACT

Like in many countries, Israel has a fairly accurate population register at the national level, consisting of about 9 million persons (not including Israelis living abroad). However, the register is much less accurate for small geographical (statistical) areas, with an average area enumeration error of about 13%. The main reason for the inaccuracy at the area level is that people moving in or out of an area are often late in reporting their change of address, and in some cases, not reporting at all. In order to correct the errors at the area level in our next census, we investigate the use of the following three-step procedure:

A- Draw a sample from an enhanced register to obtain initial direct sample estimates for the number of persons residing in each area on "census day",
B- Fit the Fay-Herriot model to the direct estimates in an attempt to improve their accuracy,
C- Compute a final census estimate for each statistical area as a linear combination of the estimate obtained in Step B and the register figure.

We also consider a procedure to deal with not missing at random (NMAR) nonresponse in Step A. The proposed procedures are illustrated using data from the 2008 Census in Israel.

Key words: direct estimator, Fay-Herriot model, Missing Information Principle, NMAR nonresponse, Root MSE estimation.

1. Introduction

In this article, we propose a new method of running a census, which combines a survey with administrative data. We consider alternative ways of integrating the survey information with the administrative data for forming a single census estimate in small geographical areas, accounting for errors in both data sources and for not missing at random (NMAR) nonresponse. We illustrate our proposed method using data from the 2008 Census in Israel.

[1] National Statistician & Head of Central Bureau of Statistic, Israel, Professor, Hebrew University of Jerusalem, Israel, and University of Southampton, UK.
[2] Senior Methodologist, Central Bureau of Statistic, 66 Kanfey Nesharim street, Jerusalem, Israel, 9546456.
[3] Senior Director, Demography and Census Department, Central Bureau of Statistic, 66 Kanfey Nesharim street, Jerusalem, Israel, 9546456.

1.1. Description of last census in Israel (2008)

Israel has a fairly accurate Central Population Register (CPR); almost perfect at the country level. However, the CPR is much less accurate for small statistical areas, with an average enumeration error of 13% and a 95 percentile of 40%. Israel is divided into about 3,000 statistical areas, and census information such as counts and socio-economic information is required for every area. The main reason for the inaccuracy in the register counts at the area level is that people moving in or out of areas, often report late their change of address, while others who have an address of interest (tax benefits, school area, parking, etc.) do not report their change of address as long as the interest persists. In 2008, the Israel Central Bureau of Census (ICBS) conducted an integrated census, which consisted of the population register, corrected by estimates obtained from two coverage samples for each area. A field (area) sample of people living in the area on census day for estimating the register undercount (the "U sample"), and a sample of people registered in the same area for estimating the register over-count (the "O sample"). The U sample was also used for collecting the socio-economic information.

The final, census estimate has been computed as follows: Denote by N_i the true number of persons residing in area i on census day and by K_i the number of persons registered as living in the area. Let $p_{i,L|R}$ represent the proportion of persons living in area i among those registered as living in the area, and $p_{i,\mathrm{R|L}}$ represent the proportion of persons registered in area i among those living in the area. Then,

$$N_i \times p_{i,R|L} = K_i \times p_{i,L|R} \Rightarrow \hat{N}_i = K_i \times \frac{\hat{p}_{i,L|R}}{\hat{p}_{i,R|L}}. \qquad (1)$$

By the use of Taylor expansion, the conditional (design-based) variance of $\hat{N}_i$ can be approximated as,

$$\mathrm{Var}(\hat{N}_i \mid K_i) \cong K_i^2 \left[\frac{Var(\hat{p}_{i,L|K})}{[E(\hat{p}_{i,K|L})]^2} + \frac{[E(\hat{p}_{i,L|K})]^2}{[E(\hat{p}_{i,K|L})]^4} \times Var(\hat{p}_{i,K|L}) \right]. \qquad (2)$$

Over the last decade, Israel, as many other countries, experienced an accelerated process of using administrative data for the production of official statistics in general, and in particular, it improved its abilities to use administrative data for census purposes. As a result, the 2020 census methodology in Israel will use new data sources and for the first time, a geo-demographic administrative file (GDAF) will serve as the sampling frame for a sample that will be used to correct the administrative data in small statistical areas.

Two key facts enable the shift in the planed methodology: a) entries and departures to and from the country are well recorded; b) all people in the country have administrative records; the citizens are registered in the CPR and the foreigners are reported in functional records like work permits and visas. A conceptual and practical leap towards fully administrative censuses in the future

can be thought of, but unfortunately, not yet in our next census in 2021, with reference census day defined as 31/12/2020.

1.2. New method planned for the next census in Israel

For our 2021 census we plan a different method, which will hopefully get us closer to the use of fully administrative censuses in the future. The census will combine information from a single sample taken from the GDAF, with information available from the register and other administrative files, mainly to correct the counts obtained from the GDAF. The sample will collect geo-demographic information on all members of the household on census day, as well as socio-economic information. It is planned to obtain the information by the Internet, then by phone from people not responding via the internet, and in cases of nonresponse by either of the two modes, by personal interviews.

The direct estimates obtained from the sample will be improved by the use of the Fay-Herriot (F-H) estimator, employing relevant covariate information known at the area level, such as the number of buildings and the total volume of all the buildings in the area, with the volume defined as the building roof area times its height. Other covariates will be used for estimating the area socio-economic means of interest.

For estimating the area counts, we shall combine the F-H estimator with the corresponding GDAF count, to obtain our final, composite, census estimator (see below).

2. Proposed three-stage census estimator

2.1. Direct count estimate (Stage 1)

Denote by N the number of residents in the country on census day and by N_i the number of residents in area i, such that $N = \sum_i N_i$. Let $p_i = N_i / N$ denote the true proportion of residents in the GDAF living in area i, and $\hat{p}_i$ denote the corresponding direct sample estimator, e.g. the sample proportion in the case of simple random sampling. (More efficient sampling designs and direct estimators are presently studied.) Finally, denote by $K \cong N$ the size of the GDAF on census day. The direct estimator for the count of area i is then $\hat{N}_i = K \times \hat{P}_i$.

The conditional design-based variance of $\hat{N}_i$ is,

$$Var_D(\hat{N}_i \mid K) = K^2 Var_D(\hat{P}_i) = \sigma_{Di}^2 .$$

2.2. "Improved" Fay-Herriot estimate (Stage 2)

The (standard) Fay-Herriot (F-H, 1979) model is:

$$\hat{N}_i = \alpha + \mathrm{x}_i'\beta + u_i + e_i , \qquad (3)$$

where $\hat{N}_i$ is the direct sample estimator, x_i represents the area covariates (number of residential buildings in the area and total volume of all the residential buildings in our empirical illustrations; we are presently searching for more powerful covariates), u_i is a random effect and e_i is the sampling error of the direct estimator.

Under the model (3), the improved, empirical best linear unbiased predictor (EBLUP) of the true count is,

$$\hat{N}_{i,IMP} = \hat{\gamma}_i \hat{N}_i + (1-\hat{\gamma}_i)\mathrm{x}_i'\hat{\beta}; \quad \hat{\gamma}_\mathrm{i} = \hat{\sigma}_u^2(\hat{\sigma}_u^2 + \hat{\sigma}_{D_i}^2)^{-1}, \tag{4}$$

where $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_{D_i}^2$ are appropriate sample estimates.

2.3. Final census count estimates

The final count estimate in area i, will be obtained as a weighted average of the improved F-H estimate in (4), and the GDAF count. For this, we assume $K_i \sim Possion(N_i) \Rightarrow Var(K_i) = N_i$. The final composite census estimator is thus,

$$\hat{N}_{i,COM} = \hat{\alpha}_i K_i + (1-\hat{\alpha}_i)\hat{N}_{i,IMP}; \ \hat{\alpha}_\mathrm{i} = \frac{\hat{\sigma}_{i,FH}^2}{\hat{\sigma}_{i,FH}^2 + \hat{Var}(K_i)}. \tag{5}$$

3. Alternative estimation of census counts

3.1. Including the register count among the covariates as fixed numbers

Rather than computing the composite estimator (5), include the GDAF count as an additional covariate in the F-H model (3). Fitting this model “as is”, implies conditioning on the known register count, ignoring its possible error. The final census count estimate is in this case the F-H estimate.

3.2. Accounting for the errors of the register errors

Following Ybarra and Lohr (2008), we add the GDAF count to the set of covariates but account for its possible measurement error by assuming, $K_i \sim N(N_i, Var(K_i))$. Denote, $\tilde{\mathrm{x}}_i = (\mathrm{x}_i', K_i)$. Assuming that all the other covariates are measured without error,

$$C_i = Var(\tilde{\mathrm{x}}_i) = \begin{bmatrix} 0\ldots0, & \ldots, & 0 \\ 0\ldots0, & \ldots, & 0 \\ \cdot\ \cdot\ \cdot\ \cdot & , & \cdot \\ \cdot\ \cdot\ \cdot\ \cdot & , & \cdot \\ \cdot\ \cdot\ \cdot\ \cdot & , & \cdot \\ 0\ldots0, & \ldots, & V(K_i) \end{bmatrix}, \text{ and}$$

$$\hat{N}_{i,YL} = \hat{\delta}_i \hat{N}_i + (1-\hat{\delta}_i)\tilde{\mathrm{x}}_i'\hat{\beta}; \quad \hat{\delta}_i = \frac{\hat{\sigma}_u^2 + \hat{\beta}' C_i \hat{\beta}}{\hat{\sigma}_u^2 + \hat{\beta}' C_i \hat{\beta} + \hat{\sigma}_{Di}^2}. \quad (6)$$

4. Empirical illustrations

To illustrate the method and its various options, we use the Over-count (O) sample drawn from the central population register for the 2008 census. The total sample size is approximately 600,000 persons. We consider the 205 areas of size 1,000-10,000 as estimated in the 2008 census, because these area sizes correspond to the sizes of the statistical areas of interest in our next census. The sample has been drawn by stratified simple random sampling. The covariates used for the models are the number of residential buildings in the area and the total volume of all the residential buildings. The F-H model parameters have been estimated by MLE, using the PROC mixed procedure in SAS, which assumes normality of the random effects and the sampling errors. The 2008 census estimates (based on the O and U the samples) are taken as the true counts (referred to in the figures below as the "Census values").

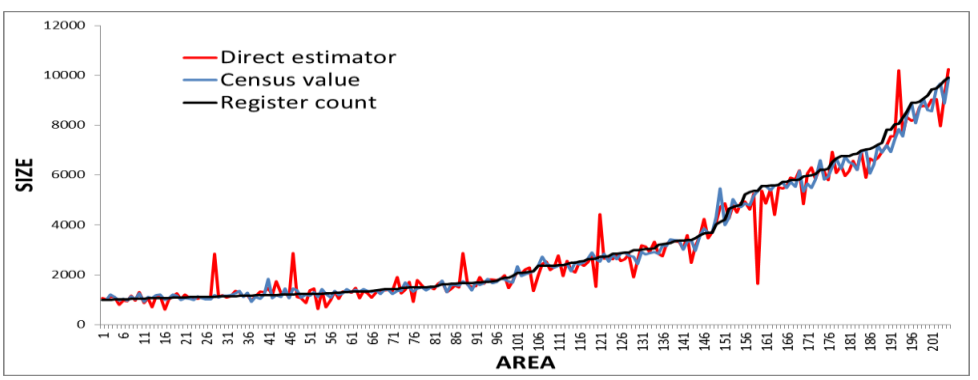


Figure 4-1. Direct estimator, Census value and Register count for the 205 small areas, ordered by their size in the register

As can be seen, the direct estimator is unbiased, but with large variance.

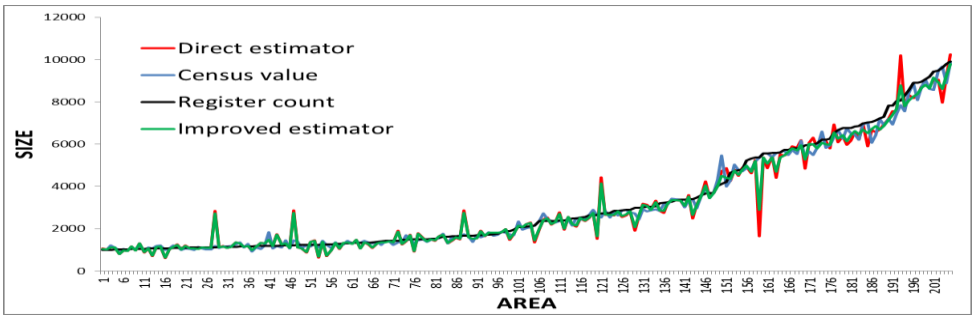


Figure 4-2. Direct estimator, Census value, Register count and Improved (F-H) *estimator*

The improved F-H estimator reduces only mildly the variance of the direct estimator. We are in the process of searching for more powerful covariates. In particular, we expect to get from the electricity company a list of all dwelling apartments and houses in Israel, which should improve the F-H estimator very significantly compared to the use of only the number of buildings.

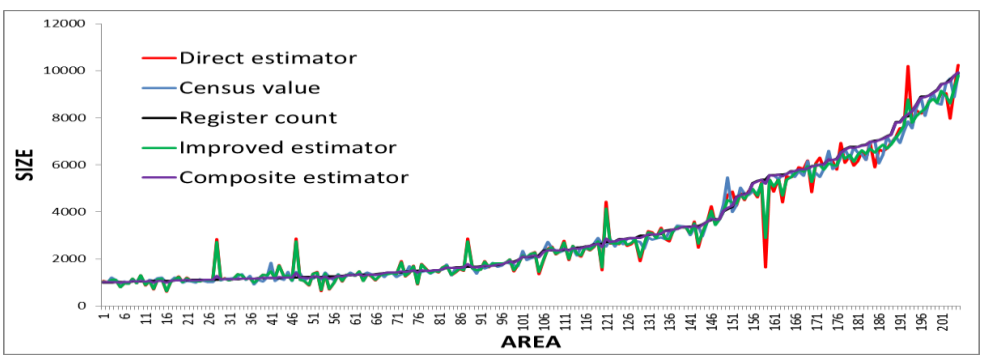


Figure 4-3. Direct estimator, Census value, Register count, Improved estimator and Composite estimator

The Composite estimator is seen to estimate the true counts much more precisely than the other estimators. Table 4.1 exhibits some summary statistics of the performance of the various estimators considered so far.

Table 4-1. Absolute relative distance of estimates from Census values **I**

Estimate	Mean	10th Pctl.	25th Pctl.	50th Pctl.	75th Pctl.	90th Pctl.
Direct	0.1047	0.0101	0.0243	0.0556	0.1084	0.2202
Register count	0.0616	0.0010	0.0151	0.0507	0.0912	0.1344
Improved	0.0946	0.0112	0.0275	0.0573	0.0956	0.1959
Composite	0.0598	0.0056	0.0189	0.0469	0.0834	0.1257

Finally, Figure 4-4 and Table 4-2 exhibit the results obtained when adding the register count as an additional covariate in the F-H model, with (FH_WME) and without (FH_NME) accounting for its measurement error. In the latter case, we estimated σ_u^2 and β by the method of modified least squares, (Ybarra and Lohr, 2008).

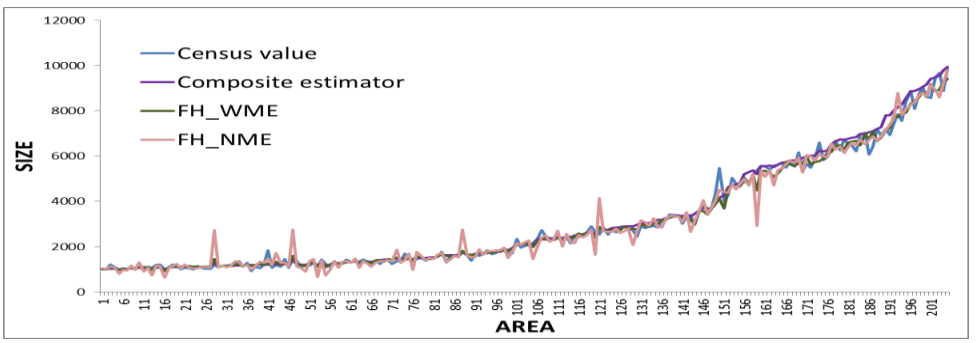


Figure 4-4. Estimates when adding the register count to the covariates of the Fay-Herriot model, with and without accounting for its measurement error

Table 4-2. Absolute relative distance of estimates from census values **II**

Estimate	Mean	10th Pctl.	25th Pctl.	50th Pctl.	75th Pctl.	90th Pctl.
Direct	0.1047	0.0101	0.0243	0.0556	0.1084	0.2202
Register count	0.0616	0.0010	0.0151	0.0507	0.0912	0.1344
Improved	0.0946	0.0112	0.0275	0.0573	0.0956	0.1959
FH_NME	0.0893	0.0100	0.0261	0.0540	0.0931	0.1877
Composite	0.0598	0.0056	0.0189	0.0469	0.0834	0.1257
FH_WME	0.0603	0.0094	0.0227	0.0498	0.0793	0.1230

As clearly noticed, not accounting for the measurement error of the register count yields a census estimator with only minor improvement over the direct sample estimator. Accounting for the error of the register count improves the performance of the F-H estimator very significantly, but quite surprisingly, the composite estimator performs somewhat better, despite of the EBLUP property of the Ybarra and Lohr (2008) estimator. Although only based on a single empirical study, a possible explanation for this result is that in the latter estimator, the same weight is assigned to the register count and the other (fixed) covariates, whereas the composite estimator is more flexible, allowing for different weights for the register count and the other covariates. Further theoretical research and empirical illustrations are required to validate this result.

5. Accounting for Not Missing At Random (NMAR) nonresponse

Sverchkov and Pfeffermann (2018) propose a method that uses the Missing Information Principle of Orchard and Woodbury (1972) for estimating the response probabilities in small areas. The basic idea is as follows: first construct the likelihood that would be obtained if the missing outcome values were known also for the nonrespondents. However, since the missing outcomes are practically unknown, replace the likelihood by its expectation with respect to the distribution of the missing outcomes, given all the observed data. The latter distribution is obtained from the distribution of the observed outcomes, as fitted to the observed values. See Sverchkov and Pfeffermann (2018) for the relationship between the distributions of the observed and the missing outcomes, for given covariates and response probabilities.

Ideally, we would want to show how the method performs in estimating the true number of persons residing in each area on census day, but this information is practically unknown for our test data (the O-sample used so far). Consequently, in what follows we illustrate instead the performance of the method when predicting the true number of divorced persons registered in each area. The O-sample is drawn from the central population register and the true number of divorced persons registered in each area is therefore known.

Define the outcome variable, y_{ij}, to be 1 if person j registered in area i is divorced, and 0 otherwise. Let the response indicator, R_{ij}, take the vlaue 1 if unit j in area i responds, and 0 otherwise. We restrict the analysis to persons aged 20+. The model fitted for the observed outcomes of the responding units and the model assumed for the response probabilities are defined in Equations (7) and (8). The covariates used for this illustration are listed in Table 5.1.

$$\Pr(y_{ij}=1 \mid \mathrm{x}_{ij}, u_i, R_{ij}=1)=\frac{\exp(\beta_0+\mathrm{x}'_{ij}\beta+u_i)}{1+\exp(\beta_0+\mathrm{x}'_{ij}\beta+u_i)};\quad u_i \sim N(0,\sigma_u^2), \tag{7}$$

$$\Pr(R_{ij}=1 \mid y_{ij}, \mathrm{x}_{ij}, \mathrm{u}_i;\gamma)=\frac{\exp(\gamma_0+\mathrm{x}'_{ij}\gamma+\gamma_y y_{ij})}{1+\exp(\gamma_0+\mathrm{x}'_{ij}\gamma+\gamma_y y_{ij})}. \tag{8}$$

Clearly, for $\gamma_y \neq 0$, Equation (8) defines an informative response mechanism.

We first impose $\gamma_y = 0$, thus presuming that being divorced does not affect the probability of response, which corresponds to assuming missing at random (MAR) nonresponse. This is implemented by omitting the marriage status, y_{ij}, from the response model (8). The results are shown in Tables 5-1 and 5-2.

Table 5-1 displays the Odds ratios of the estimated Logistic model of the response probabilities for this case. As expected, the odds ratio for responding increases as the number of telephones belonging to the administrative family increases, and similarly for the administrative family size. The age group with the

smallest response probability is 30-39 (odds ratio=0.87), and people born in Israel have a much higher odds ratio to respond than people born abroad.

Table 5-1. Odds ratios of estimated Logistic model of response probabilities assuming MAR nonresponse

Variable	Odds ratio in case of MAR non-response
# of telephones per family	1.70
Administrative family size	1.15
Age 20-29	0.98
Age 30-39	0.87
40+	1.00
Jew	1.04
Other	1.00
Born in Israel	1.27
Other	1.00

Table 5-2 shows the distribution of the estimated response probabilities under the model of Table 5.1.

Table 5-2. Distribution of estimated response probabilities under the model exhibited in Table 5-1

Marriage status	Mean	5th Pctl	25th Pctl	75th Pctl
Other	0.815	0.489	0.822	0.885
Divorced	0.742	0.359	0.683	0.843
Total	0.812	0.487	0.819	0.885

It is quite clear from Table 5-2 that the supposition $\gamma_y = 0$ is incorrect. The probability of responding among divorced persons is significantly lower than for other persons. Hence, we estimated the response probabilities by including the binary variable "divorced" as an additional explanatory variable.

Table 5-3. Odds ratios of estimated Logistic model of response probabilities allowing for NMAR nonresponse

Variable	Odds ratio in case of MAR non-response	Odds ratio in case of NMAR non-response
# of telephones per family	1.70	1.83
Administrative family size	1.15	1.11
Age 20-29	0.98	0.95
Age 30-39	0.87	0.86
Other age	1.00	1.00
Jew	1.04	1.05
Other	1.00	1.00
Born in Israel	1.27	1.25
Other	1.00	1.00
Divorced	**----**	**0.531**

We notice in Table 5-3 that the odds ratio for responding among divorced persons is about twice as small as for other persons, in correspondence with the results in Table 5-2. Interestingly, the odds ratios of the other covariates are very similar to the odds ratios obtained when assuming MAR nonresponse.

The estimated models in Tables 5-1 and 5-3 allow us to estimate the response probability for each responding person in the sample. By viewing the response as an additional stage of sampling, the estimated response probabilities will be used for predicting the true area means of the target variable (proportion of divorced persons in the present illustration) using standard sampling theory, for example, by employing the approximately design-unbiased estimator,

$$\hat{\bar{Y}}_i^{HB} = \sum_{j,(i,j)\in R} (y_{ij} / \tilde{\pi}_{j|i}) / \sum_{j,(i,j)\in R} (1/\tilde{\pi}_{j|i}); \;\; \tilde{\pi}_{j|i} = \pi_{j|i}\hat{p}_r(y_{ij}, x_{ij}; \hat{\gamma}), \qquad (9)$$

where $\pi_{j|i}$ denotes the sampling probability. Sverchkov and Pfeffermann (2018) derive also the empirical best predictor under the models (7) and (8), but we do not consider this predictor in the present paper.

Figure 5-1 and Tables 5-4 and 5-5 compare the performance of the following three predictors of the true proportion of divorced persons in the various areas: the proportion of divorced persons in the observed sample, ignoring the non-response (hereafter the direct estimator), the estimator obtained when assuming MAR nonresponse, and the estimator obtained when allowing for NMAR nonresponse (Equation 8).

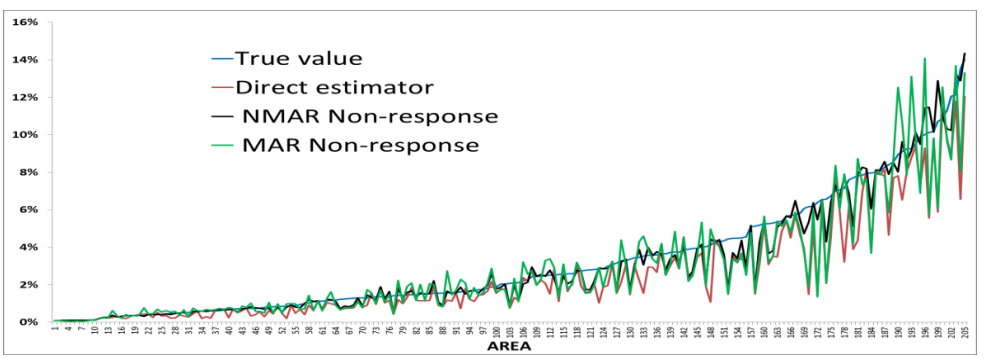


Figure 5-1. Percent of divorced persons in areas: true value, direct estimator and estimators obtained when assuming MAR and NMAR non-response

Table 5-4. Difference between true values and estimates over all the areas

Estimator	Mean	10th Pctl.	25th Pctl.	50th Pctl.	75th Pctl.	90th Pctl.
Direct	0.0075	-0.0005	0.0006	0.0036	0.0099	0.0211
MAR	0.0033	-0.0077	-0.0018	0.0004	0.0057	0.0168
NMAR	0.0019	-0.0027	-0.0004	0.0001	0.0032	0.0094

Table 5-5. Absolute relative distance of estimates from true values

Estimator	Mean	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl
Direct	0.270	0.042	0.121	0.233	0.406	0.551
MAR	0.256	0.032	0.113	0.216	0.379	0.472
NMAR	0.118	0.004	0.022	0.055	0.156	0.362

As indicated by Figure 5.1 and Tables 5-4 and 5-5, the estimates obtained when accounting for NMAR nonresponse have by far the smallest bias and the smallest absolute relative distance from the true values. The direct estimates, which ignore the nonresponse, have large bias and large relative distance from the true values.

6. Root MSE estimation under NMAR nonresponse

Like in any publication of official statistics, we are not only required to publish counts and other socio-demographic estimates, but also evaluate their precision. One set of evaluation measures (but definitely not the only one), is to estimate for each area the root MSE (RMSE) of the final estimate. This is relatively simple if we were to use the direct sample estimates as our final estimates and if all sample units will respond, but this obviously will not happen. Estimation of the RMSE is more complicated when using the Fay-Herriot estimates and accounting for NMAR nonresponse, and even more so, when using the composite estimator described in Section 2.

Sverchkov and Pfeffermann (2018) propose a bootstrap method for estimation of the RMSE of small area estimates under NMAR nonresponse, which accounts for the random processes assumed to generate the population values, and the sampling and response processes. This implies that the target area parameters (the true proportion of persons residing in the area on census day, out of all the persons registered in the CPR in our application), are considered as random, which is different from classical survey sampling applications under which the finite population values and hence the target parameters are viewed as fixed values. Users of sample survey (official statistics) estimates are familiar with measures of error, which only account for the variability originating from the randomness of the sample selection (known as the randomization distribution), and the nonresponse. In other words, users are accustomed to estimates of the design-based (randomization) MSE (denoted hereafter as DMSE), over all possible sample selections, with the population values of the survey variables (and hence the values of the target parameters), held fixed. Estimation and publication of the DMSE (or its square root) is a common routine in national statistical offices all over the world.

In a recent article, Pfeffermann and Ben-Hur (2019) propose a new procedure for estimating the DMSE of model-based small area predictors, which is shown to perform well in an extensive simulation study and outperforming other procedures for DMSE estimation proposed in the literature. We are presently extending this procedure for estimating the DMSE of our proposed composite estimators,

accounting, in particular, for the inevitable NMAR nonresponse and the use of the composite estimator, which combines the survey estimate with the administrative population register count.

7. Concluding Remarks

In this article we consider a new method for running a census, combining sample estimates with administrative data. A major advantage of this method is that it does not require the use of personal interviews, except in the case of non-respondents. Israel still does not have a sufficiently reliable dwelling register, and the use of a field sample requires prior listing of all the dwelling units in a sample of cells in each statistical area, which is rather complicated logistically and very expensive. It also requires verifying that each of the apartments is inhabited.

Under the new method, a single sample of persons is drawn from the GDAF, which is known to be generally accurate at the national level, except for some small "outlying" sub-populations, such as illegal immigrants. We consider alternative ways of combining the survey information with the GDAF to form a single final census estimator, accounting for the sampling errors in the survey, and errors in the addresses in the GDAF. We also propose a simple descriptive procedure of testing the informativeness of the missing sample data, and a way of accounting for NMAR nonresponse. We illustrate all the above topics by the use of real empirical data.

We are currently planning a census rehearsal for next year in two statistical regions of Israel, which will hopefully provide us with another opportunity to test the ideas discussed in the present article, with more up-to-date data.

Leaning more on administrative sources of information opens the way for new opportunities to the census process and outcomes. Referring to an identified population in a known population frame implies a substantial change in the concept of a census. Area boundaries become, in a way, a virtual entity rather than the main physical entity in a census. The theoretical and socio-economic implications of this change, and the influence on policy making, should be further investigated.

REFERENCES

BLUM, O., (2018). From physical area to virtual lists: Toward an administrative census in Israel. UNECE group of experts on population and housing censuses. Geneva.

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 Statistical Association, 74, pp. 269–277.

ORCHARD, T., WOODBURY, M. A., (1972). A missing information principle: theory and application. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, pp. 697–715.

PFEFFERMANN, D., BEN-HUR, D., (2018). Estimation of Randomization Mean Square Error in Small Area Estimation. International Statistical Review, 0, 0, pp. 1–19, DOI:10.1111/insr.12289.

SVERCHKOV, M., PFEFFERMANN, D., (2018), "Small area estimation under informative sampling and not missing at random non-response". Journal of the Royal Statistical Society, Series A, 181, pp. 981–1008.

YBARRA, L. M. R., LOHR, S. L., (2008), Small area estimation when auxiliary is measured with error, Biometrika, 95, pp. 919–931.

STATISTICS IN TRANSITION new series, March 2019
Vol. 20 No. 1, pp. 21–40, DOI 10.21307/stattrans-2019-002

IMPUTATION OF MISSING VALUES BY USING RAW MOMENTS

Muhammed Umair Sohail[1] **Javid Shabbir**[2], **Fariha Sohil**[3]

ABSTRACT

The estimation of population parameters might be quite laborious and inefficient, when the sample data have missing values. In comparison follow-up visits, the method of imputation has been found to be a cheaper procedure from a cost point of view. In the present study, we can enhance the performance of imputation procedures by utilizing the raw moments of the auxiliary information rather than their ranks, especially, when the ranking of the auxiliary variable is expensive or difficult to do so. Equations for bias and mean squared error are obtained by large sample approximation. Through the numerical and simulation studies it can be easily understood that the proposed method of imputation can outperform their counterparts.

Key words: non-response, imputation, raw moments, relative efficiency.
Mathematical classification: 62D05.

1. Introduction

In survey sampling, the common problem which is faced by most of social sciences, economic and scientific studies is the item or unit non-response or missing values. The main reason of the non-response is the sensitive or embarrassing nature of the questions which are relevant to the variable of interest. Usually respondents hesitate to respond to questions related to the sensitive issues, such as age, income, tax returns etc., or due to summer vocations remain a problematic issues in survey sampling. The best available sources need to be utilized for reducing the non-response rate as much as possible. In most of social studies, item or unit non-response mislead the researchers about the effective inference about the problem of interest. Usually the missing values can create a problem, when the follow-up visits are expensive, population is highly dispersed over the frame or difficult to reach. Alternatively, imputation is the most cheapest and easiest procedure to impute the non-responses by appropriate use of the auxiliary information, which is correlated with the variable of interest.

In the last few decades, several methods of imputation have been proposed to handle out such problems in an effective manner. Among them Rubin (1976) was the first who considered a comprehensive examination of non-response and explain

[1]Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan. E-mail: umairsohailch@gmail.com. ORCID ID: https://orcid.org/0000-0002-5440-126X.
[2]Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan.
[3]Department of Education, Government College University, Faisalabad, Pakistan.

the different models under which it would occur, such as missing at random (MAR), observed at random (OAR) and if the prior distributions are specified (PDS). Heitjan and Basu (1996) and Ahmed et al. (2006) provided different imputation procedures by correct use of the auxiliary information after Rubin (1976). The problem of non-response under ranked set sampling, when the ranking of observation units is inexpensive was discussed by Herrera and Al-Omari (2011). They consider the problem of missing values under the hot deck (HD) imputation strategy by the significant use of supplementary information. Grover and Kaur (2014) provide an alternative estimation procedure by combination the features of the proposed estimators by Rao (1991) and Bahl and Tuteja (1991) to provide better results than existing ones. An extensive discussion on item and unit non-response was considered by Little and Rubin (2014) in their text. They explained a different method of imputation in significant manners with suitable real life examples. Recently, Mohamed et al. (2016) provided an efficient model for handling the problem of non-response by using multi auxiliary information. Haq et al. (2017) suggested an estimation procedure for the estimation of population mean by using the ranks of the supplementary information. Sohail et al. (2017) considered the problem of scrambled non-response for the estimation of population mean and suggested a class of estimators by modifying the existing ones.
Motivated by Mohamed et al. (2016) and Sohail et al. (2017), in the present study, we appraise the problem of missing completely at random (MCAR), i.e. the probability of obtaining the response from i^{th} unit does not depend on x_i, y_i or survey design and the respondents units are representative of the selected sample for the estimation of population mean. The objective of the study is to provide an alternative procedure for those situations where the ranking of the auxiliary information is expensive or difficult to create. The proposed model not only provides more better results in terms of efficiency than Grover and Kaur (2014) and Haq et al. (2017) estimators but is also easier to understand than others.

The rest of article is structured as follows: In Section 3, we discuss some existing estimators in the literature for the imputation of missing values. In Section 4, we propose an estimator by utilizing the second raw moment of the auxiliary variable for imputing the missing values. The numerical and empirical studies are considered in Section 6. We conclude our study in Section 7.

2. Notations

Let r^* be the total number of the respondents (individuals or items) who belong to group G in sample (s) and $(n-r^*)$ are those who do not provide the respond, are belong to group G^c. So, $s = G \cup G^c$, and it is also assumed that $\hat{\bar{Y}}_r^* = \frac{1}{r^*}\sum_{j=1}^{r^*}\hat{Y}_j$ is the sample mean of the study variable obtained from respondent units in group G.

Let $\bar{X} = \sum_{j=1}^{N} X_j/N, \bar{R} = \sum_{j=1}^{N} R_j/N$ and $\bar{U} = \sum_{j=1}^{N} U_j/N$ be the population mean of the auxiliary variable, rank variable and second raw moment, respectively, and also let $\bar{x}_{r^*} = \sum_{j=1}^{r^*} x_j/r^*, \bar{r}_{r^*} = \sum_{j=1}^{r^*} r_j/r^*$ and $\bar{u}_{r^*} = \sum_{j=1}^{r^*} u_j/r^*$ be the sample mean of the

auxiliary variable, ranked variable and second raw moment, respectively, from the respondent group.
For evaluating the mathematical expressions for bias and mean square error of the existing and proposed estimators, we defined some useful notations as follows:
Let

$$\begin{aligned} e_0 &= \frac{\hat{\bar{Y}}_{r^*}}{\bar{Y}} - 1, \quad e_1 = \frac{\bar{x}_{r^*}}{\bar{X}} - 1, \quad e_3 = \frac{\bar{r}_{r^*}}{\bar{R}} - 1, \quad e_5 = \frac{\bar{u}_{r^*}}{\bar{U}} - 1, \text{ such that} \\ E(e_i) &= 0 \quad \text{for} \quad i = 0,1,3,5. \end{aligned}$$

and

$$\begin{aligned} E\left(e_0^2\right) &= \theta_{r^*,N} C_y^2, \quad E\left(e_1^2\right) = \theta_{r^*,N} C_x^2,, \quad E\left(e_3^2\right) = \theta_{r^*,N} C_r^2, \quad E\left(e_5^2\right) = \theta_{r^*,N} C_u^2, \\ E\left(e_0 e_1\right) &= \theta_{r^*,N} \rho_{xy} C_x C_y, \quad E\left(e_0 e_3\right) = \theta_{r^*,N} \rho_{ry} C_y C_r, \quad E\left(e_0 e_5\right) = \theta_{r^*,N} \rho_{uy} C_u C_y, \\ E\left(e_1 e_3\right) &= \theta_{r^*,N} \rho_{xr} C_x C_r, \quad E\left(e_1 e_5\right) = \theta_{r^*,N} \rho_{xu} C_u C_x, \quad E\left(e_3 e_5\right) = \theta_{r^*,N} \rho_{ru} C_u C_r, \end{aligned}$$

where

$$\begin{aligned} \tau &= \frac{1}{N}\sum_{j=1}^{N} \tau_j, \quad C_\tau^2 = \frac{\sigma_\tau^2}{\bar{\tau}^2}, \quad \rho_{\tau\psi} = \frac{S_{\tau\psi}}{S_\tau S_\psi}, \quad \theta_{r^*,N} = \left(\frac{1}{r^*} - \frac{1}{N}\right), \\ S_{\tau\psi} &= \frac{1}{N-1}\sum_{j=1}^{N} (\tau_j - \bar{\tau})(\psi_j - \bar{\psi}), \quad \text{where} \quad \tau, \psi = R, U, X, Y. \end{aligned}$$

3. Some Existing Methods of Imputation

In this section, we discuss some existing methods of imputation, which are available in the literature and commonly used for estimation of the population mean. These are defined below.
• Under mean imputation approach

$$\hat{Y}_j = \begin{cases} \hat{Y}_j & \text{if } j \varepsilon G \\ \hat{\bar{Y}}_{r^*} & \text{if } j \varepsilon G^c, \end{cases} \tag{1}$$

The point estimator for population mean ($\bar{Y}$) is given by

$$\hat{\bar{Y}}_M = \frac{1}{n}\left[\sum_{j=1}^{r^*} \hat{Y}_j + \sum_{j=1}^{n-r^*} \hat{Y}_j\right] = \hat{\bar{Y}}_{r^*} \tag{2}$$

The variance of the mean estimator is given by:

$$Var(\hat{\bar{Y}}_M) = \theta_{r^*,N} \bar{Y}^2 C_y^2. \tag{3}$$

• Cochran (1940) suggested the ratio estimator for the estimation of the population mean. We can rewrite it for imputing missing values as:

$$\hat{Y}_j = \begin{cases} \hat{Y}_j & \text{if } j\varepsilon G \\ \frac{1}{1-f_1}\left[\frac{\hat{\bar{Y}}_{r^*}}{\bar{x}_{r^*}}\bar{X} - f_1\hat{\bar{Y}}_{r^*}\right] & \text{if } j\varepsilon G^c, \end{cases} \tag{4}$$

where $f_1 = \frac{r^*}{n}$ and $\bar{X}$ are the population mean of the auxiliary variable. The point estimator is given as:

$$\hat{\bar{Y}}_R = \hat{\bar{Y}}_{r^*}\frac{\bar{X}}{\bar{x}_{r^*}}. \tag{5}$$

The ratio estimator is conditionally more efficient as compared to the mean estimator when the correlation between y and x is positive. The bias and the mean square error are given by

$$Bias(\hat{\bar{Y}}_R) \cong \theta_{r^*,N}\bar{Y}\left(C_x^2 - \rho_{yx}C_yC_x\right) \tag{6}$$

and

$$MSE(\hat{\bar{Y}}_R) \cong \theta_{r^*N}\bar{Y}^2\left(C_y^2 + C_x^2 - 2\rho_{yx}C_yC_x\right). \tag{7}$$

• Bahl and Tuteja (1991) proposed the ratio-exponential type estimator for imputing non-response, is expressed as:

$$\hat{Y}_j = \begin{cases} \hat{Y}_j & \text{if } j\varepsilon G \\ \frac{1}{1-f_1}\left[\hat{\bar{Y}}_{r^*}\exp\left(\frac{\bar{X}-\bar{x}_{r^*}}{\bar{X}+\bar{x}_{r^*}}\right) - f_1\hat{\bar{Y}}_{r^*}\right] & \text{if } j\varepsilon G^c, \end{cases} \tag{8}$$

The point estimator is given by:

$$\hat{\bar{Y}}_{B.T-R} = \hat{\bar{Y}}_{r^*}\exp\left(\frac{\bar{X}-\bar{x}_{r^*}}{\bar{X}+\bar{x}_{r^*}}\right), \tag{9}$$

with bias and mean squared error

$$Bias(\hat{\bar{Y}}_{B.T-R}) \cong \theta_{r^*,N}\bar{Y}\left(\frac{3}{8}C_x^2 - \frac{1}{2}\rho_{yx}C_yC_x\right) \tag{10}$$

and

$$MSE(\hat{\bar{Y}}_{B.T-R}) \cong \frac{1}{4}\theta_{r^*N}\bar{Y}^2\left(4C_y^2 + C_x^2 - 4\rho_{yx}C_yC_x\right). \tag{11}$$

The product-exponential type estimator for imputing the missing values is given by

$$\hat{Y}_j = \begin{cases} \hat{Y}_j & \text{if } j\varepsilon G \\ \frac{1}{1-f_1}\left[\hat{\bar{Y}}_{r^*}\exp\left(\frac{\bar{x}_{r^*}-\bar{X}}{\bar{x}_{r^*}+\bar{X}}\right) - f_1\hat{\bar{Y}}_{r^*}\right] & \text{if } j\varepsilon G^c, \end{cases} \tag{12}$$

The point estimator for the population mean is given as:

$$\hat{\bar{Y}}_{B.T-P} = \hat{\bar{Y}}_{r^*}\exp\left(\frac{\bar{x}_{r^*}-\bar{X}}{\bar{x}_{r^*}+\bar{X}}\right). \tag{13}$$

The bias and mean squared error of $\hat{\bar{Y}}_{B.T-P}$ are

$$Bias(\hat{\bar{Y}}_{B.T-P}) \cong \theta_{r^*,N}\bar{Y}\left(\frac{1}{2}\rho_{yx}C_yC_x - \frac{3}{8}C_x^2\right). \tag{14}$$

and

$$MSE(\hat{\bar{Y}}_{B.T-P}) \cong \frac{1}{4}\theta_{r^*,N}\bar{Y}^2\left(4C_y^2 + C_x^2 + 4\rho_{yx}C_yC_x\right). \tag{15}$$

• The conventional difference estimator is defined as:

$$\hat{Y}_j = \begin{cases} \hat{Y}_j & \text{if } j\varepsilon G \\ \frac{1}{1-f_1}\left[\hat{\bar{Y}}_{r^*} + k(\bar{X}-\bar{x}_{r^*}) - f_1\hat{\bar{Y}}_{r^*}\right] & \text{if } j\varepsilon G^c, \end{cases} \tag{16}$$

where k is an un-known constant. The point estimator for the population mean is defined as:

$$\hat{\bar{Y}}_D = \hat{\bar{Y}}_{r^*} + k(\bar{X}-\bar{x}_{r^*}). \tag{17}$$

The optimum value of k i.e. $k_{opt.} = \rho_{yx}(S_y/S_x)$. The minimum MSE($\hat{\bar{Y}}_D$) is given by

$$MSE(\hat{\bar{Y}}_D)_{min.} \cong \theta_{r^*,N}\bar{Y}^2C_y^2\left(1-\rho_{yx}^2\right). \tag{18}$$

• Rao (1991) difference type estimator can be reformulated for imputing the missing values, as:

$$\hat{Y}_j = \begin{cases} \hat{Y}_j & \text{if } j\varepsilon G \\ \frac{1}{1-f_1}\left[\nu_1\hat{\bar{Y}}_{r^*} + \nu_2(\bar{X}-\bar{x}_{r^*}) - f_1\hat{\bar{Y}}_{r^*}\right] & \text{if } j\varepsilon G^c, \end{cases} \tag{19}$$

where ν_1 and ν_2 are unknown, which are to be determined. The point estimator $\hat{Y}_j$ is given by:

$$\hat{\bar{Y}}_{R.D} = \nu_1\bar{y}_{r^*} + \nu_2(\bar{X}-\bar{x}_{r^*}). \tag{20}$$

The optimum values of v_1 and v_2 are

$$v_{1(opt.)} = \frac{1}{1+\theta_{r^*,N}C_y^2\left(1-\rho_{yx}^2\right)} \quad \text{and} \quad v_{2(opt.)} = \frac{\bar{Y}C_y^2\rho_{yx}}{\bar{X}C_x\left(1+\theta_{r^*,N}C_y^2(1-\rho_{yx}^2)\right)}.$$

The bias and MSE$(\hat{\bar{Y}}_{R.D})_{min.}$ are given by

$$Bias(\hat{\bar{Y}}_{R.D}) \cong \theta_{r^*N}\bar{Y}\left(k_1-1\right) \tag{21}$$

and

$$MSE(\hat{\bar{Y}}_{R.D})_{min.} \cong \frac{\theta_{r^*N}\bar{Y}^2C_y^2\left(1-\rho_{yx}^2\right)}{1+\theta_{r^*N}C_y^2\left(1-\rho_{yx}^2\right)}. \tag{22}$$

• In line with Grover and Kaur (2014), we can reformulate the given procedure for the imputation of missing values, as:

$$\hat{Y}_j = \begin{cases} \hat{Y}_j & \text{if } j\varepsilon G \\ \frac{1}{(n-f_1)}\left[\left(\alpha_1\hat{\bar{Y}}_{r^*}+\alpha_2(\bar{X}-\bar{x}_{r^*})\right)\times\right. & \\ \left.\exp\left(\frac{a(\bar{X}-\bar{x}_{r^*})}{a(\bar{X}+\bar{x}_{r^*})+2b}\right)-f_1\hat{\bar{Y}}_{r^*}\right] & \text{if } j\varepsilon G^c, \end{cases} \tag{23}$$

where α_1 and α_2 are the suitably chosen constants, where a and b are known parameters of the auxiliary variable, see Table 1, which is described below. The point estimator for the population mean is given as:

$$\hat{\bar{Y}}^*_{GK} = \left[\alpha_1\hat{\bar{Y}}_{r^*}+\alpha_2(\bar{X}-\bar{x}_{r^*})\right]\exp\left[\frac{a(\bar{X}-\bar{x}_{r^*})}{a(\bar{X}+\bar{x}_{r^*})+2b}\right]. \tag{24}$$

The optimum values of α_1 and α_2 are defined as:

$$\alpha_{1(opt.)} = \frac{8-\theta_{r^*,N}\theta^2C_x^2}{8[1+\theta_{r^*,N}C_y^2(1-\rho_{yx}^2)]}$$

and

$$\alpha_{2(opt.)} = \frac{\bar{Y}[\theta_{r^*,N}\theta^3C_x^3+8C_y\rho_{yx}-\theta_{r^*,N}\theta^2C_x^2C_y\rho_{yx}-4\theta C_x\{1-\theta_{r^*,N}C_y^2(1-\rho_{yx})\}]}{8\bar{X}C_x[1+\theta_{r^*,N}C_y^2(1-\rho_{yx}^2)]}.$$

where $\theta = \frac{a\bar{X}}{a\bar{X}-b}$. The bias of $\hat{\bar{Y}}^*_{G.K}$ is given as:

$$Bias(\hat{\bar{Y}}^*_{GK}) \cong \theta_{r^*,N}\bar{Y}\left[(\alpha_1 - 1) + \theta_{r^*,N}\theta\alpha_1 C_x\left(\frac{3}{2}C_x - \rho_{yx}C_y\right)\right] + \theta_{r^*,N}\theta\alpha_2\bar{X}C_x^2. \quad (25)$$

Substituting the optimum values of α_1 and α_2, we get the minimum mean squared error of $\hat{\bar{Y}}^*_{GK}$ as follows:

$$MSE(\hat{\bar{Y}}^*_{GK})_{min.} \cong \frac{\theta_{r^*,N}\bar{Y}^2\begin{bmatrix}64C_y^2(1-\rho_{yx}^2) - \theta_{r^*,N}\theta^4C_x^4 \\ -16\,\theta_{r^*,N}\theta^2C_x^2C_y^2(1-\rho_{yx}^2)\end{bmatrix}}{64[1+\theta_{r^*,N}C_y^2(1-\rho_{yx}^2)]}. \quad (26)$$

• Following Haq et al. (2017), the imputation procedure for imputing the missing values is defined as:

$$\hat{Y}_j = \begin{cases} \hat{Y}_j & \text{if } j\varepsilon G \\ \frac{1}{(n-f_1)}\left[\left(\beta_1\hat{\bar{Y}}_{r^*} + \beta_2(\bar{X} - \bar{x}_{r^*}) + \beta_3(\bar{R} - \bar{r}_{r^*})\right)\right. & \\ \left.\exp\left(\frac{a(\bar{X}-\bar{x}_{r^*})}{a(\bar{X}+\bar{x}_{r^*})+2b}\right) - f_1\hat{\bar{Y}}_{r^*}\right] & \text{if } j\varepsilon G^c, \end{cases} \quad (27)$$

where β_1, β_2 and β_3 are the unknown constants, these constant values are determined by minimizing the resultant mean squared error. The point estimator for procedure given in (27) is given as:

$$\hat{\bar{Y}}^*_{Haq.} = \left\{\beta_1\bar{y}_{r^*} + \beta_2(\bar{X} - \bar{x}_{r^*}) + \beta_3(\bar{R} - \bar{r}_{r^*})\right\}\exp\left\{\frac{a(\bar{X}-\bar{x}_{r^*})}{a(\bar{X}+\bar{x}_{r^*})+2b}\right\}. \quad (28)$$

The optimum values of β_1, β_2 and β_3 are given by:

$$\beta_{1(opt.)} = \frac{8 - \theta_{r^*,N}\theta^2C_x^2}{8[1+\theta_{r^*,N}C_y^2(1-\rho_{yx}^2)]},$$

$$\beta_{2(opt.)} = \frac{\bar{Y}\begin{bmatrix}\theta_{r^*,N}\theta^3C_x^3(-1+\rho_{xr_x}^2) + (-8C_y + \theta_{r^*,N}\theta^2C_x^2C_y)(\rho_{yx} \\ -\rho_{xr_x}\rho_{yr_x}) + 4\,\theta C_x(-1+\rho_{xr_x}^2)[-1+\theta_{r^*,N}C_y^2(1-\rho_{y.xr_x}^2)]\end{bmatrix}}{8\bar{X}C_x(-1+\rho_{xr_x}^2)[1+\theta_{r^*,N}C_y^2(1-\rho_{y.xr_x}^2)]}$$

and

$$\beta_{3(opt.)} = \frac{\bar{Y}(8 - \theta_{r^*,N}\theta^2C_x^2)C_y(\rho_{xr_x}\rho_{yx} - \rho_{yr_x})}{8\bar{R}C_r(-1+\rho_{xr_x}^2)[1+\theta_{r^*,N}C_y^2(1-\rho_{yx}^2)]}.$$

where $\rho_{y.xr_x}^2 = \frac{\rho_{yx}^2+\rho_{yr_x}^2-2\rho_{yx}\rho_{yr_x}\rho_{xr_x}}{1-\rho_{xr_x}^2}$ is coefficient of multiple determination of Y on X and R.

The bias and minimum $MSE(\hat{\bar{Y}}_{Haq.})$ are given as:

$$Bias(\hat{\bar{Y}}^{*}_{Haq.}) \cong \frac{1}{8}\Big[-8\bar{Y}+4\theta_{r^*,N}\theta C_x(\bar{X}C_x\beta_1+\bar{U}C_r\beta_3\rho_{rx}) +\bar{Y}\beta_1\Big\{8+\theta_{r^*,N}\theta C_x\big(3\theta C_x-4C_y\rho_{xy}\big)\Big\}\Big]. \tag{29}$$

and

$$MSE(\hat{\bar{Y}}^{*}_{Haq.})_{min.} \cong \frac{\theta_{r^*,N}\bar{Y}^2\left[\begin{array}{c}64C_y^2(1-\rho^2_{y.xr_x})-\theta_{r^*,N}\theta^4C_x^4\\ -16\theta_{r^*,N}\theta^2C_x^2C_y^2(1-\rho^2_{y.xr_x})\end{array}\right]}{64[1+\theta_{r^*,N}C_y^2(1-\rho^2_{y.xr_x})]}. \tag{30}$$

In Section 4, we propose new procedure for imputing the missing values by utilizing some extra auxiliary information like raw moments.

4. Proposed Method of Imputation

Correct use of auxiliary information about the study variable can enhance the performance of the estimation procedure. If the study and auxiliary variables are correlated with each other, then the second raw moment of the auxiliary variable is also correlated with the study variable. The utilization of the second raw moment is more effective than ranking, especially in those situations, when the ranking of the auxiliary information is done at high cost or is difficult. On the basis of this logic, we propose a new class of the estimators for imputing the missing values by utilizing the second raw moment of the auxiliary variable for the estimation of finite population mean. The suggested class of estimators can incorporate the supplementary information in the form of the second raw moment. Let $\rho_{xu}=S_{xu}/(S_xS_u)$ be the correlation coefficient between X and U.

The imputation procedure for the use of the second raw moment of the auxiliary information is described as follows:

$$\hat{Y}_j=\begin{cases}\hat{Y}_j & \text{if } j\varepsilon G\\ \frac{1}{(n-f_1)}\left[\left\{k_1\hat{\bar{Y}}_{r^*}+k_2(\bar{X}-\bar{x}_{r^*})+k_3(\bar{U}-\bar{u}_{r^*})\right\}\right. & \\ \left.\exp\left\{\frac{a(\bar{X}-\bar{x}_{r^*})}{a(\bar{X}+\bar{x}_{r^*})+2b}\right\}-f_1\hat{\bar{Y}}_{r^*}\right] & \text{if } j\varepsilon G^c,\end{cases} \tag{31}$$

The point estimator for the population mean for using the above mentioned imputation procedure in (31), is defined as:

$$\hat{\bar{Y}}^{*}_{Pr}=\left\{k_1\bar{y}_{r^*}+k_2(\bar{X}-\bar{x}_{r^*})+k_3(\bar{U}-\bar{u}_{r^*})\right\}\exp\left\{\frac{a(\bar{X}-\bar{x}_{r^*})}{a(\bar{X}+\bar{x}_{r^*})+2b}\right\}. \tag{32}$$

where k_1, k_2 and k_3 are suitably chosen constants, which can be determined by minimizing the mean square error. We can rewrite the proposed estimator for imputing

the missing values in terms of error as:

$$\hat{\bar{Y}}^{*}_{Pr_1} = \left(k_1\bar{Y}(1+e_0) - k_2\bar{X}e_1 - k_3\bar{U}e_5\right)\left(1 - \frac{\theta}{2}e_1 + \frac{3}{8}\theta^2 e_1^2\right).$$

The bias of the proposed estimator is:

$$\begin{aligned} Bias(\hat{\bar{Y}}^{*}_{Pr}) \;\cong\; & \frac{1}{8}\Big[-8\bar{Y} + 4\theta_{r^*,N}\theta C_x\Big(\bar{X}C_x k_1 + \bar{U}C_u k_3\rho_{ux}\Big) \\ & + \bar{Y}k_1\Big\{8 + \theta_{r^*,N}\theta C_x\Big(3\theta C_x - 4C_y\rho_{xy}\Big)\Big\}\Big]. \end{aligned} \tag{33}$$

The mean squared error of the proposed imputation procedure is given as:

$$\begin{aligned} MSE(\hat{\bar{Y}}^{*}_{Pr}) \;\cong\; & \bar{Y}^2 + \theta_{r^*,N}\bar{X}C_x k_2\Big(-\bar{Y}\theta + \bar{X}k_1\Big) + \theta_{r^*,N}\bar{U}C_u^2 k_3^2 + \theta_{r^*,N}\bar{U}C_x C_u \\ & \Big(-\bar{Y}\theta + 2\bar{X}k_1\Big) + \bar{Y}^2 k_1^2\Big[1 + \theta_{r^*,N}\Big\{C_y^2 + \theta C_x\Big(\theta C_x - 2C_y\rho_{xy}\Big)\Big\}\Big] \\ & + \frac{1}{4}\bar{Y}k_1\Big[-8\bar{Y} + \theta_{r^*,N}C_x\Big\{\theta C_x\Big(-3\bar{Y}\theta + 8\bar{X}k_2\Big) + 8\bar{U}\theta C_u k_3\rho_{xu} \\ & + 4C_y\Big(\bar{Y} - 2\bar{X}k_2\Big)\rho_{xy}\Big\} - 8\bar{U}C_u C_y\theta_{r^*,N}k_3\rho_{uy}\Big]. \end{aligned} \tag{34}$$

The optimum values of the unknown constants $[k_i \text{ for } i = 1,2,3.]$ are determined by minimizing (34), which can be expressed as:

$$k_{1(opt.)} = \frac{8 - \theta_{r^*,N}\theta^2 C_x^2}{8[1 + \theta_{r^*,N}C_y^2(1-\rho_{yx}^2)]},$$

$$k_{2(opt.)} = \frac{\bar{Y}\left[\begin{array}{r}\theta_{r^*,N}\theta^3 C_x^3(-1+\rho_{xu_x}^2) + (-8C_y + \theta_{r^*,N}\theta^2 C_x^2 C_y)(\rho_{yx} - \rho_{xu_x}\rho_{yu_x}) \\ +4\theta C_x(-1+\rho_{xu_x}^2)\{-1 + \theta_{r^*,N}C_y^2(1-\rho_{y.xu_x}^2)\}\end{array}\right]}{8\bar{X}C_x(-1+\rho_{xu_x}^2)[1 + \theta_{r^*,N}C_y^2(1-\rho_{y.xu_x}^2)]}$$

and

$$k_{3(opt.)} = \frac{\bar{Y}(8 - \theta_{r^*,N}\theta^2 C_x^2)C_y(\rho_{xu_x}\rho_{yx} - \rho_{yu_x})}{8\bar{U}C_u(-1+\rho_{xu_x}^2)[1 + \theta_{r^*,N}C_y^2(1-\rho_{yx}^2)]}.$$

where $\rho_{y.xu_x}^2 = \frac{\rho_{yx}^2 + \rho_{yu_x}^2 - 2\rho_{yx}\rho_{yu_x}\rho_{xu_x}}{1-\rho_{xu_x}^2}$ is coefficient of multiple determination of Y on X and U in simple random sampling.

$$MSE(\hat{\bar{Y}}^{*}_{Pr})_{min.} \cong \frac{\theta_{r^*,N}\bar{Y}^2\left[\begin{array}{r}64C_y^2(1-\rho_{y.xu_x}^2) - \theta_{r^*,N}\theta^4 C_x^4 \\ -16\theta_{rN}\theta^2 C_x^2 C_y^2(1-\rho_{y.xu_x}^2)\end{array}\right]}{64[1 + \theta_{r^*,N}C_y^2(1-\rho_{y.xu_x}^2)]}. \tag{35}$$

Table 1: Some special cases of existing and proposed imputation methods

a	b	$\hat{\bar{Y}}^{*}_{GK}$	$\hat{\bar{Y}}^{*}_{Haq.}$	$\hat{\bar{Y}}^{*}_{Pr}$
1	C_x	$\hat{\bar{Y}}^{1}_{GK}$	$\hat{\bar{Y}}^{1}_{Haq.}$	$\hat{\bar{Y}}^{1}_{Pr}$
1	$N\tilde{X}$	$\hat{\bar{Y}}^{2}_{GK}$	$\hat{\bar{Y}}^{2}_{Haq.}$	$\hat{\bar{Y}}^{2}_{Pr}$
$N\tilde{X}$	C_x	$\hat{\bar{Y}}^{3}_{GK}$	$\hat{\bar{Y}}^{3}_{Haq.}$	$\hat{\bar{Y}}^{3}_{Pr}$
C_x	$N\tilde{X}$	$\hat{\bar{Y}}^{4}_{GK}$	$\hat{\bar{Y}}^{4}_{Haq.}$	$\hat{\bar{Y}}^{4}_{Pr}$
1	ρ_{xy}	$\hat{\bar{Y}}^{5}_{GK}$	$\hat{\bar{Y}}^{5}_{Haq.}$	$\hat{\bar{Y}}^{5}_{Pr}$
C_x	ρ_{xy}	$\hat{\bar{Y}}^{6}_{GK}$	$\hat{\bar{Y}}^{6}_{Haq.}$	$\hat{\bar{Y}}^{6}_{Pr}$
ρ_{xy}	C_x	$\hat{\bar{Y}}^{7}_{GK}$	$\hat{\bar{Y}}^{7}_{Haq.}$	$\hat{\bar{Y}}^{7}_{Pr}$
$N\tilde{X}$	ρ_{xy}	$\hat{\bar{Y}}^{8}_{GK}$	$\hat{\bar{Y}}^{8}_{Haq.}$	$\hat{\bar{Y}}^{8}_{Pr}$
ρ_{xy}	$N\tilde{X}$	$\hat{\bar{Y}}^{9}_{GK}$	$\hat{\bar{Y}}^{9}_{Haq.}$	$\hat{\bar{Y}}^{9}_{Pr}$
1	$N\bar{X}$	$\hat{\bar{Y}}^{10}_{GK}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{Pr}$

5. Efficiency Comparison

Here, we define the regulatory conditions under which the proposed estimators can perform better than their existing estimators, which are given by
(i) By (26) and (35), $MSE(\hat{\bar{Y}}_{GK}) - MSE(\hat{\bar{Y}}^{*}_{pr}) > 0$, if

$$\rho_{uy} > \rho_{xu}\rho_{xy} - \sqrt{\rho_{xy}\left(1-\rho^2_{xu}\right)\left(1-\rho_{xy}\right)}. \tag{36}$$

(ii) By (30) and (35), $MSE(\hat{\bar{Y}}_{Haq.}) - MSE(\hat{\bar{Y}}^{*}_{pr}) > 0$, if

$$\rho_{uy} > \frac{\sqrt{\left(1-\rho^2_{xu}\right)}\left(\rho_{wy}-\rho_{xw}\rho_{xy}\right)}{\sqrt{1-\rho^2_{xw}}} + \rho_{xy}\rho_{xu}. \tag{37}$$

Conditions (i) and (ii) are satisfied, then the proposed estimators for imputing the missing responses perform better than their counterparts.

6. Application

For the relative comparison of the proposed class of estimators with existing ones in terms of efficiency, we consider real life as well as simulated data, sets which are discussed in the following subsections.

6.1. Numerical Study

We consider the following four real life data sets for the practical application of the proposed class of estimator and obtained the percentage relative efficiencies of the existing and proposed estimators. The data description is given below as:
Population 1: [Source: Singh (2003)]
y= Estimated number of fish caught by marine recreational fishermen in year 1995 and x = estimated number of fish caught by marine recreational fishermen in year 1994.

$$\begin{aligned} N &= 69, n = 40, \bar{Y} = 14.0225, \bar{X} = 147.0425, \bar{R} = 100.5, \bar{U} = 28955.59, \\ S_y^2 &= 27.22185, S_x^2 = 7370.95, S_w^2 = 3350 S_u^2 = 653591180, S_{xy} = 350.3902, \\ S_{uy} &= 98116.68 S_{xu} = 2123923, S_{ry} = 234.8867, S_{wx} = 4959.526, S_{wu} = 1438183, \\ \rho_{xy} &= 0.7822, \rho_{uy} = 0.7355817, \rho_{wy} = 0.7778165, \rho_{xu} = 0.967662, \rho_{uw} = 0.97193, \\ \rho_{wx} &= 0.998058 \end{aligned}$$

Population 2: [Source: James et al. (2013)]
y= total sales and x = expenditure on TV advertisement

$$\begin{aligned} N &= 200, n = 40, \bar{Y} = 14.0225, \bar{X} = 177.5965, \bar{R} = 100.5, \bar{U} = 73653530, \\ S_y^2 &= 27.22185, S_x^2 = 8057.097, S_u^2 = 4.4 e^{+16}, S_{xy} = 376.3316, S_{uy} = 98116.68, \\ S_{xu} &= 1.4 e^{+12}, S_{ry} = 94080.28, S_{wx} = 106830.7, S_{wu} = 1.4 e^{+12}, \rho_{xy} = 0.9601, \\ \rho_{uy} &= 0.8554, \rho_{wy} = 0.7689, \rho_{xu} = 0.9283, \rho_{uw} = 0.5208, \rho_{wx} = 0.75434 \end{aligned}$$

Population 3: [Source: James et al. (2013)]
y= Income and x = education

$$\begin{aligned} N &= 30, n = 15, \bar{Y} = 16, \bar{X} = 50.1455, \bar{R} = 15.5, \bar{U} = 2946.634, \\ S_y^2 &= 13.2712, S_x^2 = 446.9652, S_w^2 = 77.5 S_u^2 = 4340687, S_{xy} = 74.31184, \\ S_{uy} &= 7344.01, S_{xu} = 43477.52, S_{ry} = 30.7390, S_{wx} = 106830.7, S_{wu} = 18115.9, \\ \rho_{xy} &= 0.9648, \rho_{uy} = 0.9676, \rho_{wy} = 0.9584, \rho_{xu} = 0.9283, \rho_{uw} = 0.9870, \\ \rho_{wx} &= 0.9925 \end{aligned}$$

Population 4: [Source: James et al. (2013)]
y= Income and x = education + seniority

$$\begin{aligned} N &= 30, n = 15, \bar{Y} = 15.5, \bar{X} = 110.2483, \bar{R} = 15.5, \bar{U} = 15249.32, \\ S_y^2 &= 729.7176, S_x^2 = 3201.347, S_w^2 = 77.5, S_u^2 = 179829664, S_{xy} = 872.8027, \\ S_{uy} &= 186487.9, S_{xu} = 741453.5, S_{ry} = 130.5645, S_{wx} = 491.1011, S_{wu} = 1438183, \\ \rho_{xy} &= 0.5710, \rho_{uy} = 0.5148, \rho_{wy} = 0.5490, \rho_{xu} = 0.97720, \rho_{uw} = 0.9494, \\ \rho_{wx} &= 0.98594 \end{aligned}$$

For the relative efficiencies of the proposed and existing imputation procedures, we consider the following expression:

$$PRE(.) = \frac{Var(\hat{\bar{Y}}_M)}{MSE(\hat{\bar{Y}}_k)} \quad \text{for} \quad k = G.K, Haq., Pr. \tag{38}$$

To check the relative performance of the given procedures, we consider the response rate between 25% to 80% in all of the four populations. By the use of varying response rate, we are able to illustrate the relative performance of the imputation procedure in an significant way. Based on the results given in Table 2 and 3, we conclude that the estimator $\hat{\bar{Y}}_{GK}$, $\hat{\bar{Y}}_{Haq.}$ and $\hat{\bar{Y}}_{pr}$ remain better as compared to $\hat{\bar{Y}}_M$. At varying response rate, the inter-class efficiency of the available estimators is varying slightly over their entire range. After observing Table 2 and 3 in detail, we can say that there exists an inverse relationship between the response rate and PRE's. At low response rate, all the given estimators can perform better as compared to the mean estimator than a high response rate. For intra-class efficiency, we can observe that the proposed estimators can outperform the existing estimators. At the response rate (25%), PRE of the $\hat{\bar{Y}}_{GK}$, and $\hat{\bar{Y}}_{Haq.}$ is $1411.1340, 1502.4550$ and $261.4669, 262.9224$ for the first and second population, but at the same point PRE of $\hat{\bar{Y}}_{pr}$ is 1608.0930 and 266.3743 respectively. In population 3 and 4, PRE of the existing one is $1507.4520, 1508.4190$ and $154.8800, 156.4693$ respectively. The PRE value of the suggested estimator is 1741.5110 and 164.7871 respectively.
Overall, we can say that, the utilization of the second raw moment of the auxiliary variable has significant effect on the estimation of population parameters rather than utilizing the ranks of the supplementary information, even when the ranking of the auxiliary information is inexpensive.

6.2. Empirical Study

An empirical study of any strategy or procedure is helpful to draw the actual picture of the performance for the respective phenomena by assuming some known value of the population parameters. For empirical illustration of the existing and proposed methods of imputing non-response, we consider the following steps to generate the artificial data sets, which are defined as follows:
• We can generate first two artificial data sets by using the bivariate normal population with mean $\mathbf{A} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$ and variance $\mathbf{V} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$, and last two data sets are generated by using the gamma distribution with $\mathbf{Q} = \begin{bmatrix} a \\ b \end{bmatrix}$ under following parametric values:
• **Artificial Data Set 1:**

$$\mathbf{A} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} 6 & 3 \\ 3 & 8 \end{bmatrix}$$

Table 2: PRE(.) of the existing and proposed estimators by real life data sets

				Population 1			Population 2		
r^*	Estimators			$\hat{\bar{Y}}^*_{GK}$	$\hat{\bar{Y}}^*_{Haq.}$	$\hat{\bar{Y}}^*_{pr}$	$\hat{\bar{Y}}^*_{GK}$	$\hat{\bar{Y}}^*_{Haq.}$	$\hat{\bar{Y}}^*_{pr}$
10	$\hat{\bar{Y}}^1_{GK}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	1411.1340	1502.5440	1608.0930	261.4669	262.9224	266.3743
	$\hat{\bar{Y}}^2_{GK}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	1295.5650	1375.8530	1468.0350	258.8331	260.2699	263.6769
	$\hat{\bar{Y}}^3_{GK}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	1411.2400	1502.6620	1608.2260	261.4941	262.9498	266.4023
	$\hat{\bar{Y}}^4_{GK}$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	1295.6230	1375.9150	1468.1010	258.8329	260.2696	263.6767
	$\hat{\bar{Y}}^5_{GK}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	1411.1680	1502.5820	1608.1360	261.4577	262.9132	266.3649
	$\hat{\bar{Y}}^6_{GK}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	1411.1900	1502.6060	1608.1630	261.4323	262.8875	266.3388
	$\hat{\bar{Y}}^7_{GK}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	1411.1290	1502.5390	1608.0870	261.4594	262.9149	266.3666
	$\hat{\bar{Y}}^8_{GK}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	1411.2400	1502.6620	1608.2260	261.4941	262.9498	266.4023
	$\hat{\bar{Y}}^9_{GK}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	1295.5600	1375.8480	1468.0300	258.8330	260.2697	263.6768
	$\hat{\bar{Y}}^{10}_{GK}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	1295.5170	1375.8020	1467.9810	258.8331	260.2699	263.6769
20	$\hat{\bar{Y}}^1_{GK}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	1331.6930	1416.2790	1513.6120	259.2250	260.6695	264.0950
	$\hat{\bar{Y}}^2_{GK}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	1286.4050	1366.6910	1458.8710	258.1407	259.5774	262.9845
	$\hat{\bar{Y}}^3_{GK}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	1331.7320	1416.3220	1513.6610	259.2361	260.6807	264.1064
	$\hat{\bar{Y}}^4_{GK}$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	1286.4290	1366.7170	1458.8990	258.1406	259.5773	262.9844
	$\hat{\bar{Y}}^5_{GK}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	1331.7060	1416.2930	1513.6280	259.2213	260.6657	264.0912
	$\hat{\bar{Y}}^6_{GK}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	1331.7140	1416.3020	1513.6380	259.2108	260.6552	264.0805
	$\hat{\bar{Y}}^7_{GK}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	1331.6910	1416.2770	1513.6100	259.2219	260.6664	264.0919
	$\hat{\bar{Y}}^8_{GK}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	1331.7320	1416.3220	1513.6610	259.2361	260.6807	264.1064
	$\hat{\bar{Y}}^9_{GK}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	1286.4030	1366.6890	1458.8690	258.1406	259.5774	262.9844
	$\hat{\bar{Y}}^{10}_{GK}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	1286.3850	1366.6700	1458.8490	258.1407	259.5774	262.9845
30	$\hat{\bar{Y}}^1_{GK}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	1306.9350	1389.4510	1484.3000	258.4836	259.9244	263.3413
	$\hat{\bar{Y}}^2_{GK}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	1283.3520	1363.6380	1455.8170	257.9099	259.3466	262.7537
	$\hat{\bar{Y}}^3_{GK}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	1306.9560	1389.4730	1484.3250	258.4895	259.9303	263.3473
	$\hat{\bar{Y}}^4_{GK}$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	1283.3650	1363.6510	1455.8310	257.9098	259.3466	262.7536
	$\hat{\bar{Y}}^5_{GK}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	1306.9420	1389.4580	1484.3080	258.4816	259.9224	263.3392
	$\hat{\bar{Y}}^6_{GK}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	1306.9460	1389.4620	1484.3130	258.4761	259.9169	263.3336
	$\hat{\bar{Y}}^7_{GK}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	1306.9340	1389.4500	1484.2990	258.4820	259.9228	263.3396
	$\hat{\bar{Y}}^8_{GK}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	1306.9560	1389.4730	1484.3250	258.4895	259.9303	263.3473
	$\hat{\bar{Y}}^9_{GK}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	1283.3510	1363.6370	1455.8160	257.9099	259.3466	262.7536
	$\hat{\bar{Y}}^{10}_{GK}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	1283.3420	1363.6270	1455.8050	257.9099	259.3466	262.7537

Table 3: PRE(.) of the existing and proposed estimators by real life data sets

				Population 3			Population 4		
r^*	Estimators			$\hat{\bar{Y}}^*_{GK}$	$\hat{\bar{Y}}^*_{Haq.}$	$\hat{\bar{Y}}^*_{pr}$	$\hat{\bar{Y}}^*_{GK}$	$\hat{\bar{Y}}^*_{Haq.}$	$\hat{\bar{Y}}^*_{pr}$
4	$\hat{\bar{Y}}^1_{GK}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	1507.4520	1508.4190	1741.5110	154.8800	156.4693	164.7871
	$\hat{\bar{Y}}^2_{GK}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	1449.6250	1450.5270	1667.1470	152.4071	153.9680	162.1362
	$\hat{\bar{Y}}^3_{GK}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	1509.2460	1510.2150	1743.8630	152.4055	153.9664	162.1345
	$\hat{\bar{Y}}^4_{GK}$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	1449.6130	1450.5150	1667.1330	154.9062	156.4958	164.8154
	$\hat{\bar{Y}}^5_{GK}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	1505.2450	1506.2100	1738.6210	154.8770	156.4663	164.7840
	$\hat{\bar{Y}}^6_{GK}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	1500.3310	1501.2900	1732.1930	154.8499	156.4388	164.7547
	$\hat{\bar{Y}}^7_{GK}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	1507.3880	1508.3550	1741.4270	154.8606	156.4496	164.7662
	$\hat{\bar{Y}}^8_{GK}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	1509.2450	1510.2140	1743.8610	154.9062	156.4958	164.8154
	$\hat{\bar{Y}}^9_{GK}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	1449.6240	1450.5260	1667.1460	152.4057	153.9666	162.1346
	$\hat{\bar{Y}}^{10}_{GK}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	1449.6240	1450.5270	1667.1470	152.4071	153.9680	162.1362
8	$\hat{\bar{Y}}^1_{GK}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	1472.8720	1473.8010	1697.1490	151.1097	152.6824	160.9131
	$\hat{\bar{Y}}^2_{GK}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	1448.9680	1449.8710	1666.4890	150.0889	151.6498	159.8179
	$\hat{\bar{Y}}^3_{GK}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	1473.5970	1474.5270	1698.0940	150.0883	151.6492	159.8172
	$\hat{\bar{Y}}^4_{GK}$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	1448.9630	1449.8660	1666.4830	151.1204	152.6933	160.9246
	$\hat{\bar{Y}}^5_{GK}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	1471.9800	1472.9080	1695.9870	151.1085	152.6812	160.9118
	$\hat{\bar{Y}}^6_{GK}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	1469.9870	1470.9130	1693.3920	151.0974	152.6700	160.8998
	$\hat{\bar{Y}}^7_{GK}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	1472.8460	1473.7750	1697.1160	151.1017	152.6744	160.9045
	$\hat{\bar{Y}}^8_{GK}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	1473.5960	1474.5260	1698.0930	151.1204	152.6933	160.9246
	$\hat{\bar{Y}}^9_{GK}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	1448.9680	1449.8700	1666.4890	150.0883	151.6492	159.8173
	$\hat{\bar{Y}}^{10}_{GK}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	1448.9680	1449.8710	1666.4890	150.0889	151.6498	159.8179
12	$\hat{\bar{Y}}^1_{GK}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	1461.6890	1462.6060	1682.8530	149.8684	151.4358	159.6378
	$\hat{\bar{Y}}^2_{GK}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	1448.7490	1449.6520	1666.2700	149.3162	150.8771	159.0451
	$\hat{\bar{Y}}^3_{GK}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	1462.0780	1462.9950	1683.3590	149.3159	150.8768	159.0448
	$\hat{\bar{Y}}^3_{GK}4$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	1448.7460	1449.6490	1666.2670	149.8742	151.4416	159.6440
	$\hat{\bar{Y}}^5_{GK}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	1461.2090	1462.1260	1682.2290	149.8678	151.4351	159.6371
	$\hat{\bar{Y}}^6_{GK}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	1460.1370	1461.0520	1680.8350	149.8618	151.4290	159.6306
	$\hat{\bar{Y}}^7_{GK}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	1461.6750	1462.5920	1682.8350	149.8641	151.4314	159.6332
	$\hat{\bar{Y}}^8_{GK}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	1462.0780	1462.9950	1683.3590	149.8742	151.4416	159.6440
	$\hat{\bar{Y}}^9_{GK}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	1448.7490	1449.6520	1666.2700	149.3159	150.8768	159.0448
	$\hat{\bar{Y}}^{10}_{GK}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	1448.7490	1449.6520	1666.2700	149.3162	150.8771	159.0451

• **Artificial Data Set 2:**

$$\mathbf{A} = \begin{bmatrix} 6 \\ 8 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} 8 & 4 \\ 4 & 10 \end{bmatrix}$$

• **Artificial Data Set 3:**

$$\mathbf{Q} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

• **Artificial Data Set 4:**

$$\mathbf{Q} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$

The main purpose of generating the two different data sets from the same distribution is to find the pattern of PRE with respect to their parametric values. In Data sets 3 and 4, the study variable is generated as $y = (r_{xy} \times x) + e$, where $e \sim N(0,1)$ and r_{yx} is the sample correlation coefficient between y and x.

• Here, we can select the sample of size n form N units, randomly, and select randomly r units out of n sample units and impute the dropped units by using the above mentioned imputation procedures, then compute the relevant statistics.

• Repeat the process 30000 (say H) times and obtain the value of $\hat{\bar{Y}}_k^*$. The mean squared error of the given estimator is obtained by using the following expression, as:

$$MSE(\hat{\bar{Y}}_k^*) = \frac{1}{H} \sum_{i=1}^{H} \left((\hat{\bar{Y}}_k^*)_i - \bar{Y} \right)^2 \tag{39}$$

At the specified values of parameters and $n = 50$, the behaviour of normal distribution, gamma distribution and self-generated study variable is shown in Appendix (Figure: 1). By utilizing the artificial data sets, mean squared errors of the given procedures are reported below. On the behalf of numerical findings, which are reported in Tables 4 and 5, we see that the relative performance of the existing and proposed imputation method is similar to the reported results in Table 2 and 3. By the use of simulated data sets (which are generated by bivariate normal and gamma distribution under certain regulatory conditions) the performances of the existing and proposed estimators are better than the mean estimator. As given by the reported results in Table 2 and 3, PRE of respective imputation procedure decreases as the response rate increases, but as a whole these are better than traditional estimators. After comprehensive examination of Tables 4 and 5, we can easily understand that our proposed class of estimators performs significantly better than existing and mean imputation procedures even in high response rate. As the parametric values of the population constants increase in normal population, the performance of all the estimator increase. But in the case of positively dispersed population, there is an inverse relationship between PRE's and parametric values.

Table 4: PRE(.) of existing and proposed estimators by using artificial data set (.).

				Artificial Data Sets 1			Artificial Data Sets 2		
r^*	Estimator			$\hat{\bar{Y}}^*_{G.K}$	$\hat{\bar{Y}}^*_{Haq.}$	$\hat{\bar{Y}}^*_{pr}$	$\hat{\bar{Y}}^*_{G.K}$	$\hat{\bar{Y}}^*_{Haq.}$	$\hat{\bar{Y}}^*_{pr}$
10	$\hat{\bar{Y}}^1_{G.K}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	118.6620	105.9573	119.6891	124.0394	127.1311	132.0750
	$\hat{\bar{Y}}^2_{G.K}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	120.2631	106.3291	122.0707	135.0018	134.6481	138.5286
	$\hat{\bar{Y}}^3_{G.K}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	101.4103	117.0941	123.6927	115.5257	119.5066	126.8564
	$\hat{\bar{Y}}^4_{G.K}$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	120.2941	106.3447	122.1189	136.0608	135.7178	141.6042
	$\hat{\bar{Y}}^5_{G.K}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	109.6636	116.9660	121.5692	125.0806	128.1636	132.6883
	$\hat{\bar{Y}}^6_{G.K}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	113.8250	114.6901	115.1804	126.1819	127.2321	131.1512
	$\hat{\bar{Y}}^7_{G.K}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	119.6756	100.1994	120.9550	127.8607	130.1780	132.4522
	$\hat{\bar{Y}}^8_{G.K}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	102.8872	117.2912	120.0964	115.6963	119.6452	126.6902
	$\hat{\bar{Y}}^9_{G.K}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	115.7024	105.5199	120.5186	136.5683	136.1019	139.8346
	$\hat{\bar{Y}}^{10}_{G.K}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	120.7377	106.3721	121.5206	137.3745	136.8175	138.5315
20	$\hat{\bar{Y}}^1_{G.K}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	117.8110	105.2151	119.8508	120.7829	123.9798	129.3023
	$\hat{\bar{Y}}^2_{G.K}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	121.2069	105.0492	120.9690	134.9703	134.5377	137.2016
	$\hat{\bar{Y}}^3_{G.K}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	105.2075	119.1198	127.4633	118.2557	121.9277	128.7035
	$\hat{\bar{Y}}^4_{G.K}$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	120.3429	105.1115	122.1641	134.4532	134.1529	139.9697
	$\hat{\bar{Y}}^5_{G.K}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	110.8191	116.4726	118.5771	125.1330	128.1416	132.7055
	$\hat{\bar{Y}}^6_{G.K}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	114.3393	114.3480	115.7193	128.5056	130.6469	132.4355
	$\hat{\bar{Y}}^7_{G.K}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	120.1568	100.1455	124.3512	128.6391	130.8064	132.7405
	$\hat{\bar{Y}}^8_{G.K}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	103.7974	105.2274	106.1870	116.6377	120.5083	127.8116
	$\hat{\bar{Y}}^9_{G.K}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	119.6394	105.0234	120.5388	134.5111	134.2217	135.1581
	$\hat{\bar{Y}}^{10}_{G.K}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	116.2761	105.1145	119.2251	134.6683	134.4694	137.5448
30	$\hat{\bar{Y}}^1_{G.K}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	119.7622	105.1340	120.6659	122.5217	125.6854	130.4396
	$\hat{\bar{Y}}^2_{G.K}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	119.0610	104.6887	119.7440	134.7711	134.4483	135.1757
	$\hat{\bar{Y}}^3_{G.K}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	105.2325	118.6783	127.5033	117.8911	121.6284	128.4568
	$\hat{\bar{Y}}^4_{G.K}$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	119.8297	104.5238	123.7315	133.8278	133.4614	136.3154
	$\hat{\bar{Y}}^5_{G.K}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	112.3861	116.9674	118.0800	124.5436	127.5873	132.0415
	$\hat{\bar{Y}}^6_{G.K}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	114.4807	114.5614	115.8958	131.6862	133.5503	135.1994
	$\hat{\bar{Y}}^7_{G.K}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	118.2454	100.0115	119.6349	129.0047	131.1589	133.2060
	$\hat{\bar{Y}}^8_{G.K}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	105.5192	106.8852	107.7843	117.4968	121.3278	128.5287
	$\hat{\bar{Y}}^9_{G.K}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	120.2420	104.7962	120.4320	133.3648	134.0260	135.7968
	$\hat{\bar{Y}}^{10}_{G.K}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	115.5720	104.6800	119.5149	132.7433	132.2578	136.8836

Table 5: PRE(.) of existing and proposed estimators by using artificial data set (.).

				Artificial Data Sets 3			Artificial Data Sets 4		
r^*	Estimator			$\hat{\bar{Y}}^*_{G.K}$	$\hat{\bar{Y}}^*_{Haq.}$	$\hat{\bar{Y}}^*_{pr}$	$\hat{\bar{Y}}^*_{G.K}$	$\hat{\bar{Y}}^*_{Haq.}$	$\hat{\bar{Y}}^*_{pr}$
10	$\hat{\bar{Y}}^1_{G.K}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	192.8977	192.5636	193.1662	136.4316	136.2749	136.7016
	$\hat{\bar{Y}}^2_{G.K}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	194.0581	194.0559	194.9967	140.5252	140.5304	140.6584
	$\hat{\bar{Y}}^3_{G.K}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	188.0287	187.8559	188.1561	136.3089	136.3474	136.8219
	$\hat{\bar{Y}}^4_{G.K}$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	191.5353	191.5554	191.7439	140.5487	140.5449	141.2544
	$\hat{\bar{Y}}^5_{G.K}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	189.2338	189.3205	189.9834	138.3192	138.3297	139.3096
	$\hat{\bar{Y}}^6_{G.K}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	191.3141	191.5745	191.9685	139.2955	139.2150	139.4283
	$\hat{\bar{Y}}^7_{G.K}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	192.1016	192.8594	192.8723	139.1179	139.1532	139.4509
	$\hat{\bar{Y}}^8_{G.K}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	187.5624	187.3498	187.7200	138.2495	138.2914	140.1726
	$\hat{\bar{Y}}^9_{G.K}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	194.5011	194.4885	194.4346	140.8154	140.8160	141.8146
	$\hat{\bar{Y}}^{10}_{G.K}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	194.1044	194.0683	194.5535	138.5544	138.5559	138.7474
20	$\hat{\bar{Y}}^1_{G.K}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	153.2724	152.6555	153.8245	118.6956	118.5047	119.0081
	$\hat{\bar{Y}}^2_{G.K}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	154.8089	154.7715	155.7848	140.5252	140.5304	140.7584
	$\hat{\bar{Y}}^3_{G.K}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	151.1041	150.8019	151.3284	122.4189	122.4683	122.7324
	$\hat{\bar{Y}}^4_{G.K}$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	153.5339	153.4891	153.7509	125.0272	125.0842	125.8300
	$\hat{\bar{Y}}^5_{G.K}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	153.1577	153.3651	153.8868	123.5171	123.5145	123.5189
	$\hat{\bar{Y}}^6_{G.K}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	148.0252	149.0603	149.9335	123.0634	122.0864	123.1859
	$\hat{\bar{Y}}^7_{G.K}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	157.3814	157.4595	157.6159	124.3142	124.3697	124.6227
	$\hat{\bar{Y}}^8_{G.K}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	152.6729	152.4070	152.8817	122.5352	122.5805	123.9481
	$\hat{\bar{Y}}^9_{G.K}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	155.2700	155.2155	156.2721	124.9516	124.9523	125.0465
	$\hat{\bar{Y}}^{10}_{G.K}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	154.7365	154.6740	154.7385	124.1267	124.1129	124.1383
30	$\hat{\bar{Y}}^1_{G.K}$	$\hat{\bar{Y}}^1_{Haq.}$	$\hat{\bar{Y}}^1_{pr}$	137.6494	136.8790	138.3680	112.2768	112.0773	112.6044
	$\hat{\bar{Y}}^2_{G.K}$	$\hat{\bar{Y}}^2_{Haq.}$	$\hat{\bar{Y}}^2_{pr}$	141.3239	141.2753	142.2912	118.8472	118.8500	118.9965
	$\hat{\bar{Y}}^3_{G.K}$	$\hat{\bar{Y}}^3_{Haq.}$	$\hat{\bar{Y}}^3_{pr}$	139.0538	138.7795	139.2670	116.4861	116.5465	116.7900
	$\hat{\bar{Y}}^4_{G.K}$	$\hat{\bar{Y}}^4_{Haq.}$	$\hat{\bar{Y}}^4_{pr}$	141.5240	141.4602	141.5615	119.0015	119.0045	119.0104
	$\hat{\bar{Y}}^5_{G.K}$	$\hat{\bar{Y}}^5_{Haq.}$	$\hat{\bar{Y}}^5_{pr}$	139.7605	139.9683	140.4578	118.0333	118.0357	118.1293
	$\hat{\bar{Y}}^6_{G.K}$	$\hat{\bar{Y}}^6_{Haq.}$	$\hat{\bar{Y}}^6_{pr}$	130.9078	132.1338	132.6111	116.6496	116.5642	116.8055
	$\hat{\bar{Y}}^7_{G.K}$	$\hat{\bar{Y}}^7_{Haq.}$	$\hat{\bar{Y}}^7_{pr}$	139.3605	139.5683	139.8578	118.9265	118.9745	119.8550
	$\hat{\bar{Y}}^8_{G.K}$	$\hat{\bar{Y}}^8_{Haq.}$	$\hat{\bar{Y}}^8_{pr}$	138.2340	137.9276	138.4662	116.2475	116.3058	116.4495
	$\hat{\bar{Y}}^9_{G.K}$	$\hat{\bar{Y}}^9_{Haq.}$	$\hat{\bar{Y}}^9_{pr}$	140.9842	140.9393	141.9538	118.7914	118.7869	118.7950
	$\hat{\bar{Y}}^{10}_{G.K}$	$\hat{\bar{Y}}^{10}_{Haq.}$	$\hat{\bar{Y}}^{10}_{pr}$	142.1794	142.1593	142.2399	118.2523	118.2476	118.6504

7. Conclusions

In this study, we assume that the non-response which occurred in the study is MCAR. Our main objective is to introduce the idea of utilizing the second raw moment of the auxiliary variable for the imputation of missing values, especially for those situations when the ranking of the auxiliary information is difficult or expensive. The proposed imputation method provides better results in terms of efficiency than the existing procedures. From Tables 2, 3, 4 and 5, it can be easily understand that the proposed imputation procedure performs better than Grover and Kaur (2014) and Haq et al. (2017) estimators. Thus, we recommend the proposed estimator for the imputation of missing values and for a precise estimation of the population mean.

The current work can easily be extended to other domains of survey sampling such as the estimation population quartiles (Q_1 and Q_3) and population variance under the stratified and other sampling schemes. Another possible extension of the current work is to estimate the population parameter of the sensitive variable with the non-sensitive auxiliary variable, when the non-response occurs after the utilization of the randomized response model, as in Mohamed et al. (2016) and Sohail et al. (2017). This work is deferred to the later article, which is currently in progress for handling the non-response.

REFERENCES

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

BAHL, S., TUTEJA, R., (1991). Ratio and product type exponential estimators. Journal of Information and Optimization Sciences, 12 (1), pp. 159–164.

COCHRAN, W., (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. The Journal of Agricultural Science, 30 (2), pp. 262–275.

GROVER, L. K., KAUR, P., (2014). A generalized class of ratio type exponential estimators of population mean under linear transformation of auxiliary variable. Communications in Statistics-Simulation and Computation, 43 (7), pp. 1552–1574.

HAQ, A., KHAN, M., HUSSAIN, Z., (2017). A new estimator of finite population mean based on the dual use of the auxiliary information. Communications in Statistics-Theory and Methods, 46 (9), pp. 4425–4436.

HEITJAN, D. F., BASU, S., (1996). Distinguishing "missing at random" and "missing completely at random". The American Statistician, 50 (3), pp. 207–213.

HERRERA, C. N. B., AL-OMARI, A. I., (2011). Ranked set estimation with imputation of the missing observations: the median estimator. Revista Investigacion Operacional, 32 (1), pp. 30–37.

JAMES, G., WITTEN, D., HASTIE, T., AND TIBSHIRANI, R., (2013). An introduction to statistical learning, Vol. 6, Springer.

LITTLE, R. J., RUBIN, D. B., (2014). Statistical analysis with missing data. John WileySons.

MOHAMED, C., SEDORY, S. A., SINGH, S., (2016). Imputation using higher order moments of an auxiliary variable. Communications in Statistics-Simulation and Computation, (just-accepted), pp. 00–00.

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

RUBIN, D. B., (1976). Inference and missing data. Biometrika, 63 (3), pp. 581–592.

SINGH, S., (2003). Advanced Sampling Theory With Applications: How Michael Selected Amy, Vol. 2, Springer ScienceBusiness Media.

SOHAIL, M. U., SHABBIR, J., AHMED, S., (2017). Modified class of ratio and regression type estimators for imputing scrambling response, Pakistan, Journal of Statistics, 33 (4), pp. 277–300.

APPENDIX

In Figure 1, we can show the shape of different distributions according to their respective parametric values. In Figure (a), the behaviour of normal distribution is shown according to their respective population parameters. The shape of gamma distribution is expressed in Figure (b) and standard normal distribution is shown in Figure (c). The trend of study variable is shown under the normal and gamma distribution in Figure (d) and (e) respectively. In both Figures, the study variable has an increasing trend.

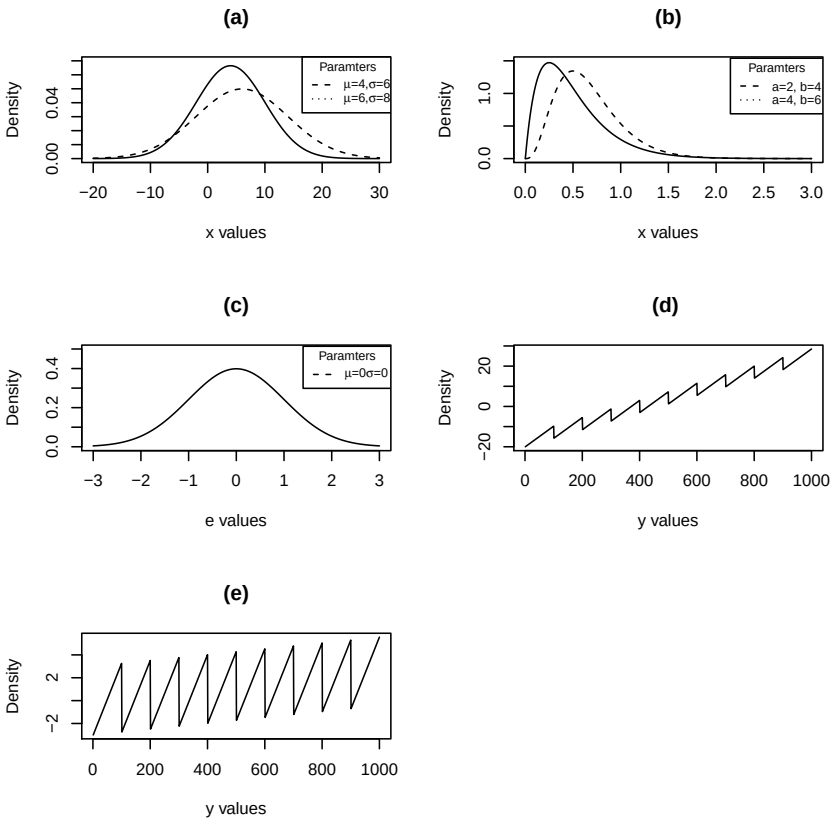


Figure 1: Shape of different distribtions according to their parametric values

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 41–65, DOI 10.21307/stattrans-2019-003

MODELLING SENSITIVE ISSUES ON SUCCESSIVE WAVES

Kumari Priyanka[1], **Pidugu Trisandhya**[2]

ABSTRACT

This paper addresses the problem of estimation of population mean of sensitive character using non-sensitive auxiliary variable at current wave in two wave successive sampling. A general class of estimator is proposed and studied under randomized and scrambled response model. Many existing estimators have been modified to work for sensitive population mean estimation. The modified estimators became the members of proposed general class of estimators. The detail properties of all the estimators have been discussed. Their behaviour under randomized and scrambled response techniques have been elaborated. Numerical illustrations including simulation have been accompanied to judge the performance of different estimators. Finally suitable recommendations are forwarded.

Key words: Sensitive variable, Successive waves, Scrambled Response model, Class of estimators, Population mean, Bias, Mean squared error, Optimum matching fraction.

1. Introduction

The occurrence of unpleasant phenomenon are plenty and abundance in the human society. We as a part of it, are sometimes obliged to take serious notice of and spread their occurrences among the conscientious public. This phenomenon necessitates assemblage of truthful and reliably adequate and accurate date. But the usual survey practices were not enough to elicit human responses through queries about sensitive and stigmatized issues.

Some of the features like gambling habits, alcoholism, illegal drug use, tax evasion, rash driving of motorized vehicles, conjugal malpractices and domestic violence etc., people like to hide from the human communities.

Hence, to deal with sensitive issues, an alternative technique has been introduces by Warner (1965), which is to obtain responses through a randomized response (RR) survey where every sampled unit is asked to give a response through an RR device as per instruction from the investigator. One can refer to Greenberg *et al.* (1971), Barlev *et al.* (2004), Diana and Perri (2011) and Arcos *et al.* (2015), etc.

[1]Department of Mathematics , Shivaji College, University of Delhi, New Delhi-110 027, India. E-mail: priyanka.ism@gmail.com.
[2]Department of Mathematics , Shivaji College, University of Delhi, New Delhi-110 027, India. E-mail: trisandhya.09@gmail.com.

for a comprehensive review of such RR procedure. However, there is another approach to deal with sensitive issue called scrambled response technique introduced by Pollock and Bek (1976). Many researchers such as Eichhorn and Hayre (1983), Saha (2007) and Diana and Perri (2010), etc. considered the scrambled response models to deal with sensitive issues.

There are many situations were one need to study the variable over time as they may opt of change by time. Jessen (1942) inaugurated the journey of research program related to variables which changes by time. Later Patterson (1950), Sen (1973), Feng and Zou (1997), Singh and Priyanka (2008), Priyanka and Mittal (2014, 2015a, 2015b), and Priyanka *et al.* (2015), etc. added sub-sensitive literature in this area.

However, if the variable which opt to change by time is also sensitive in nature, then their arises a need to apply randomized/scrambled response techniques on successive waves. Arnab and Singh (2013), Yu *et al.* (2015), Naeem and Shabbir (2016) and Singh *et al.* (2017) have put their efforts to deal with sensitive issues on successive waves.

In the present work a general class of estimators have been proposed for estimating sensitive population mean at current wave in two wave successive sampling using a non-sensitive auxiliary variable. The proposed estimators have been studied under both the randomized and scrambled response technique. Many existing estimators in successive sampling literature such as estimators by Jessen (1942), Singh and Priyanka (2008), Singh and Karna (2009) and Singh and Prasad (2010) when modified to work for sensitive population mean estimation, becomes the members of proposed general class of estimators. The modified estimators have also been checked for their applicability under considered randomized and scrambled response models. The proposed general class of estimators have been compared with the members of its class in terms of percent relative efficiency. Simulation study has also been carried out to show the practicability of proposed methods. Finally, suitable concluding remarks have been forwarded.

2. Survey Strategies and Analysis

2.1. Background

Let P be finite population of N units which has been considered for two successive waves. The sensitive study variable be named as x at the first wave and y at second wave. Whereas z is assumed to be non-sensitive auxiliary variable which is available at the both the successive waves. A simple random sample without replacement of size n is drawn at the first wave and at the second wave two independent samples are drawn by considering the partial overlapping case, one is matched sample of size $m = n\lambda$ drawn as sub sample from the sample of size n from first wave and

another is unmatched simple random sample of size $u = (n-m) = n\mu$ drawn afresh at the second (current) wave so that the sample size at both the wave is n. On first(second) wave the sensitive variables $x(y)$ are switched to $x'(y')$ with the aid of scrambling variables W_1, W_2 and W_3. The scrambling variable are considered such that they may follow any distribution. The following notations to be considered further are presented below:

$\bar{X}$, $\bar{Y}$, $\bar{Z}$, $\bar{X}_i^{'}$, $\bar{Y}_i^{'}$, $\bar{W}_1$, $\bar{W}_2$, $\bar{W}_3$: Population means of the variables x, y, z, $x_i^{'}$, $y_i^{'}$, W_1, W_2 and W_3 respectively where $i = 1$ and 2 corresponds to randomized and scrambled response models respectively.

$\bar{x}_{ui}^{'}$, $\bar{y}_{mi}^{'}$, $\bar{x}_{mi}^{'}$, $\bar{y}_{ni}^{'}$: Sample mean of the variate based on sample sizes shown in suffices under i^{th} model.

$\bar{z}_u$, $\bar{z}_m$, $\bar{z}_n$: Sample mean of the non-sensitive auxiliary variate based on sample sizes shown in suffice.

ρ_{yx}, ρ_{xz}, ρ_{yz}, $(\rho_{x'y'})_i$, $(\rho_{y'z})_i$, $(\rho_{x'z})_i$: Correlation coefficient between the variables shown in suffices and $'i'$ denote the scrambled and randomized response model.

C_x, C_y, C_z : Coefficient of variation of variables shown in suffices.

S_x^2, S_y^2, S_z^2, $S_{W_1}^2$, $S_{W_2}^2$, $S_{W_3}^2$: Population mean squared error of variables x, y, z, W_1, W_2 and W_3 respectively.

2.2. Randomized Response Techniques on successive waves

A unified approach for randomized response technique has been proposed by Arcos *et al.*(2015). Their technique say M_{AR} is modified to be applied on successive wave for estimation of population mean of sensitive variable. Each respondent on first(second) wave is asked to rotate a spinner bearing the following statements
- Report the real value of variable $x_i[y_i]$
- Report the scrambled response $(x_iW_1 + W_2)$ $[(y_iW_1 + W_2)]$
- Report a value of variable W_3

with corresponding probabilities p_1, p_2 and $(1-p_1-p_2)$ respectively on first [second] waves. Using above randomization devise, response given by j^{th} respondent on first and second wave respectively are described as

$$X_{1j}^{'} = \begin{cases} \mathrm{x}_j & \text{with probability } p_1 \\ \mathrm{x}_j W_1 + W_2 & \text{with probability } \mathrm{p}_2 \\ W_3 & \text{otherwise} \end{cases}, \quad Y_{1j}^{'} = \begin{cases} \mathrm{y}_j & \text{with probability } \mathrm{p}_1 \\ y_j W_1 + W_2 & \text{with probability } \mathrm{p}_2 \\ W_3 & \text{otherwise} \end{cases}$$

Therefore applying M_{AR} on two successive waves, the sensitive variable $x(y)$ are perturbed to $x^{'}(y^{'})$ and are given by

$$X_1^{'} = Xp_1 + (XW_1 + W_2)p_2 + W_3(1 - p_1 - p_2)$$

and

$$Y_1^{'} = Yp_1 + (YW_1 + W_2)p_2 + W_3(1 - p_1 - p_2)$$

such that

$$(\bar{Y})_1 = \frac{\bar{Y}_1^{'} - p_2\bar{W}_2 - (1 - p_1 - p_2)\bar{W}_3}{p_1 + p_2\bar{W}_1} \tag{1}$$

$\left(\rho_{y'x'}\right)_1 = \frac{p_1^2\rho_{yx}S_yS_x + 2p_1p_2\rho_{yx}S_yS_x\bar{W}_1 + p_2^2(\rho_{yx}S_yS_xS_{W_1}^2 + \rho_{yx}S_yS_x\bar{W}_1^2 + \bar{X}\bar{Y}S_{W_1}^2) + (1-p_1-p_2)^2S_{W_3}^2}{\sqrt{I_1}\sqrt{I_2}}$,
$\left(\rho_{y'z}\right)_1 = \frac{(p_1 + p_2\bar{W}_1)\rho_{yz}S_y}{\sqrt{I_2}}$, $\left(\rho_{x'z}\right)_1 = \frac{(p_1 + p_2\bar{W}_1)\rho_{xz}S_x}{\sqrt{I_1}}$

where,
$I_1 = p_1^2S_x^2 + p_2^2(S_x^2S_{W_1}^2 + S_x^2\bar{W}_1^2 + S_{W_1}^2\bar{X}^2 + S_{W_2}^2) + (1 - p_1 - p_2)^2S_{W_3}^2 + 2p_1p_2\bar{W}_1S_x^2$
and
$I_2 = p_1^2S_y^2 + p_2^2(S_y^2S_{W_1}^2 + S_y^2\bar{W}_1^2 + S_{W_1}^2\bar{Y}^2 + S_{W_2}^2) + (1 - p_1 - p_2)^2S_{W_3}^2 + 2p_1p_1\bar{W}_1S_y^2$

Many other randomised response models such as Greenberg *et al.* (1971) (M_G), Barlev *et al.* (2004) (M_B), Diana and Perri (2010) (M_{DP1}) and scrambled response models by Pollock and Bek (1976) (M_{PB}), Eichhorn and Hayre (1983) (M_{EH}), Saha (2007) (M_{SH}) and Diana and Perri (2010) (M_{DP2}) can be viewed as particular cases of above described techniques and are presented in Table 1 .

Table 1. Particular cases

Name of the Model	p_1	p_2	W_1	W_2	W_3
M_G	p	$1-p$	0	W_2	0
M_{PB}	0	1	1	W_2	0
M_{EH}	0	1	W_1	0	0
M_B	p	$1-p$	W_1	0	0
M_{SH}	0	1	W_1	W_1W_2	0
M_{DP1}	p	$1-p$	W_1	W_1W_2	0
M_{DP2}	0	1	$(1-\chi^*)W_1$	$\chi^*W_1W_2$	0

Note:$\chi^* \varepsilon [0,1]$ and $0 \leq p \leq 1$

2.3. Scrambled Response Techniques on successive waves

Considering a convex combination of the multiplicative and additive scrambled response model, Diana and Perri (2010) proposed scrambled response model. Their underlying idea was to combine the two models giving them a different weight according to the problem at hand. Therefore, their model say M_{DP} is modified to be applied on two successive waves, the sensitive variable $x(y)$ is perturbed to $x^{'}(y^{'})$ in the light of this model as:

$$X_2^{'} = \varphi_x^*(X + W_2) + (1 - \varphi_x^*)W_1X \ ; \ where \ \varphi_x^* \ \varepsilon \ [0,1]$$

and

$$Y_2^{'} = \varphi_y^*(Y+W_2)+(1-\varphi_y^*)W_1 Y \ ; \ where \ \varphi_y^* \ \varepsilon \ [0,1]$$

such that

$$(\bar{Y})_2 = \frac{\bar{Y}_1^{'} - \varphi_y^* \bar{W}_1}{\varphi_y^* + (1-\varphi_y^*)\bar{W}_2}, \tag{2}$$

$\left(\rho_{y'x'}\right)_2 = \frac{\varphi_x^*\varphi_y^*[I_3]+[I_4]\left(\varphi_x^*+\varphi_y^*\right)+I_5}{\sqrt{I_6 I_7}}$, $\left(\rho_{y'z}\right)_2 = \frac{\rho_{yz}S_y\left[\varphi_y^*(1-\bar{W}_1)+\bar{W}_1\right]}{\sqrt{I_6}}$,

$(\rho_{x'z})_2 = \frac{\rho_{xz}S_x[\varphi_x^*(1-\bar{W}_1)+\bar{W}_1]}{\sqrt{I_7}}$, $\left(C_{x'}^2\right)_2 = \frac{I_7}{\bar{X}'^2}$ and $\left(C_{y'}^2\right)_2 = \frac{I_6}{\bar{Y}'^2}$

where,

$I_3 = \rho_{yx}S_yS_x + S_{W_2}^2 - 2\rho_{yx}\bar{W}_1S_yS_x + S_{W_1}^2\left[\rho_{yx}S_yS_x + \bar{X}\bar{Y}\right] + \rho_{yx}\bar{W}_1^2S_yS_x,$

$I_4 = \bar{W}_1\rho_{yx}S_yS_x - S_{W_1}^2\left[\rho_{yx}S_yS_x + \bar{X}\bar{Y}\right] - \bar{W}_1^2\rho_{yx}S_yS_x,$

$I_5 = S_{W_1}^2\left[\rho_{yx}S_yS_x + \bar{X}\bar{Y}\right] + \bar{W}_1^2\rho_{yx}S_yS_x,$

$I_6 = \left(\varphi_y^*\right)^2\left[S_y^2 + S_{W_2}^2\right] + \left(1-\varphi_y^*\right)\left[S_{W_1}^2(1+\bar{Y}^2) + S_y^2(1+\bar{W}_1^2)\right] + 2\varphi_y^*\left(1-\varphi_y^*\right)\bar{W}_1S_y^2,$

$I_7 = (\varphi_x^*)^2\left[S_x^2 + S_{W_2}^2\right] + (1-\varphi_x^*)\left[S_{W_1}^2(1+\bar{X}^2) + S_x^2(1+\bar{W}_1^2)\right] + 2\varphi_x^*(1-\varphi_x^*)\bar{W}_1S_x^2.$

Remark 1. *If $\varphi_x^*(\varphi_y^*) = 0$, then the model M_{DP} reduces to multiplicative model and if $\varphi_x^*(\varphi_y^*) = 1$, it reduces to additive scramble response model.*

Remark 2. *The scrambling variables W_1, W_2 and W_3 are such that $E(W_1) = \bar{W}_1$, $E(W_2) = \bar{W}_2$, $E(W_3) = \bar{W}_3$, $V(W_1) = S_{W_1}^2$, $V(W_2) = S_{W_2}^2$, $V(W_3) = S_{W_3}^2$, $S_{y_i'}^2 = \begin{pmatrix} C_{y_i'}^2 \\ \bar{Y}_i'^2 \end{pmatrix}$, $S_{x_i'}^2 = \begin{pmatrix} C_{x_i'}^2 \\ \bar{X}_i'^2 \end{pmatrix}$.*

Remark 3. *$(\bar{Y})_i$, $i = 1$ and 2 denote population mean of sensitive variable y at current wave under i^{th} model in two wave successive sampling.*

Remark 4. *Suitable estimator of population mean of coded response variable $\bar{Y}_i^{'}$ need to be investigated and replaced in equation 1 and 2 respectively in order to obtain appropriate estimator of sensitive population mean at current wave under five different models in two wave successive sampling.*

2.4. General Class of estimators on Successive waves

For estimating the population mean of coded response variable at current wave in two wave successive sampling under randomized as well as scrambled response models described in section 2.2 and 2.3 respectively, two classes of estimators have been proposed based on sample of size u and m respectively. The final estimators

is the general class of estimator formulated by considering convex linear combination of two classes of estimators based on sample size u and m respectively under two consider models.

2.4.1 Class of Estimators based on unmatched sample on the second wave

The literature on successive sampling reveals that in general difference, regression, ratio, product, exponential ratio or product type estimator can be modified for the estimation of population mean of coded response variable. Some of them can be seen as:

$L_{u1i} = \bar{y}'_{ui}$, if no additional non-sensitive auxiliary information is used at any wave.
$L_{u2i} = \bar{y}'_{ui} + k(\bar{z}_u - \bar{Z})$,
$L_{3i} = \bar{y}'_{ui} + \beta_{qiz}(\bar{z}_u - \bar{Z})$,
$L_{u4i} = \bar{y}'_{ui} \frac{\bar{Z}}{\bar{z}_u}$,
$L_{u5i} = \bar{y}'_{ui} \frac{\bar{z}_u}{\bar{Z}}$,
$L_{u6i} = \bar{y}'_{ui} (\frac{\bar{z}_u}{\bar{Z}})^{\theta_1}$,
$L_{u7i} = \bar{y}'_{ui}[2 - (\frac{\bar{z}_u}{\bar{Z}})^{\theta_2}]$,
$L_{u8i} = \bar{y}'_{ui}\ exp\ (\frac{\bar{Z} - \bar{z}_u}{\bar{Z} + \bar{z}_u})$,
$L_{u9i} = \bar{y}'_{ui}\ exp\ (\frac{\bar{z}_u - \bar{Z}}{\bar{z}_u + \bar{Z}})$,
$L_{u10i} = \bar{y}'_{ui} + \beta_{y'iz}(\bar{Z} - \bar{z}_u)$,
$L_{u11i} = \bar{y}'_{ui} + b_{y'iz}(u)(\bar{Z} - \bar{z}_u)$,
etc.,

where, k, θ_1 and θ_2 are constants chosen suitably, so that the mean squared errors of L_{u2i}, L_{u6i} and L_{u7i} may be optimized respectively.

Therefore, following Srivastava (1980) and Tracy *et al.*(1996) a class of estimator have been proposed which may contain the above discussed estimators as its members, under the considered randomized and scrambled response models based on unmatched sample as:

$$L_{ui} = U_i\ (\bar{y}'_{ui},\ \bar{z}_u) \tag{3}$$

where $i = 1$ and 2 denote the randomized and scrambled response model respectively given in section 2.2 and 2.3 and $U_i\ (\bar{y}'_{ui},\ \bar{z}_u)$ is a function of $\bar{y}'_{ui}$ and $\bar{z}_u$ such that

(i) The point $(\bar{y}'_{ui},\ \bar{z}_u)$ assumes the value in a closed convex subset R_2 of two dimensional real space containing the point $(\bar{Y}'_i, \bar{Z})$.

(ii) The function $U_i\ (\bar{y}'_{ui},\ \bar{z}_u)$ is continuous and bounded in R_2.

(iii) $U_i(\bar{Y}'_i,\ \bar{Z}) = \bar{Y}'_i$ and $U_{1i}(\bar{Y}'_i,\ \bar{Z}) = \frac{\partial U_i(\bar{y}'_{ui}, \bar{z}_u)}{\partial \bar{y}'_{ui}} = 1$, i.e., First order partial derivative of U_i with respect to $\bar{y}'_{ui}$ at $U_i\ (\bar{Y}'_i, \bar{Z}) = \bar{Y}'_i \Rightarrow U_{1i}\ (K_i) = \frac{\partial U_i(.)}{\partial \bar{y}'_{ui}}|_{K_i} = 1$, where $K_i = (\bar{Y}'_i, \bar{Z})$.

(iv) The first and second order partial derivatives of $U_i\ (\bar{y}'_{ui}, \bar{z}_u)$ exist and are continuous and bounded in R_2.

2.4.2 Estimators Based on the matched sample at current wave

For the matched sample of size m retained from previous wave, it is clear that there are two kind of auxiliary information available, one is non-sensitive additional auxiliary information (z) and other is information from previous wave based on sample of size n. Hence, motivated by Senapati and Sahoo(2006) let $f_{1i} = g_i\ (\bar{x}_{mi}^{'},\ \bar{z}_m,\ \bar{z}_n)$ and $f_{2i} = h_i\ (\bar{x}_{ni}^{'},\ \bar{z}_n)$ are two different classes of estimators of $\bar{X}_i^{'}$ through samples of sizes m and n respectively such that $g_i\ (\bar{X}_i^{'},\ \bar{Z},\ \bar{Z}) = h_i(\bar{X}_i^{'},\ \bar{Z}) = \bar{X}_i^{'}$. Let $(\bar{y}_{mi}^{'},\ f_{1i},\ f_{2i})$ assumes values in a closed convex subspace R_3 of 3-dimensional real space containing the point $(\bar{Y}_i^{'},\ \bar{X}_i^{'},\ \bar{X}_i^{'})$. Also suppose $T_i\ (\bar{y}_{mi}^{'},\ f_{1i},\ f_{2i})$ is a known function of $\bar{y}_{mi}^{'},\ f_{1i},\ f_{2i}$ such that $T_i\ (\bar{Y}_i^{'},\ \bar{X}_i^{'},\ \bar{X}_i^{'}) = \bar{Y}_i^{'}$ and the three functions $g_i,\ h_i$, and T_i satisfies the regularity conditions stated by Srivastava (1980). Hence, a general class of estimators based on sample size m at current wave for estimating sensitive population mean under two models may be defined as

$$L_{mi} = T_i\ (\bar{y}_{mi}^{'},\ f_{1i},\ f_{2i}) \tag{4}$$

where $i = 1$ and 2 denote the randomized and scrambled response models respectively quoted in section 2.2 and 2.3. Many well known estimators when modified for estimation of sensitive population mean can become a member of proposed class of estimators. Some of them are listed in Table 2.

Table 2. Estimators based on sample size m

Member	Estimator	Functional Form
L_{m1i}	$[\bar{y}_{1mi}^{'} + k(\bar{x}_{ni}^{'} - \bar{x}_{mi}^{'})]$	when no additional non-sensitive auxiliary information is used then $f_{1i} = \bar{x}_{mi}^{'} \& f_{2i} = \bar{x}_{ni}^{'}$
L_{m2i}	$[\bar{y}_{1mi}^{*'} + \beta_{y_i x_i}^{'}(\bar{x}_{1ni}^{*'} - \bar{x}_{1mi}^{*'})]$, where, $\bar{y}_{1mi}^{*'} = [\bar{y}_{mi}^{'} + \beta_{y_i z}^{'}(\bar{Z} - \bar{z}_m)],\ \bar{x}_{1ni}^{*'} = [\bar{x}_{ni}^{'} + \beta_{x_i z}^{'}(\bar{Z} - \bar{z}_n)]$ & $\bar{x}_{1mi}^{*'} = [\bar{x}_{mi}^{'} + \beta_{x_i z}^{'}(\bar{Z} - \bar{z}_m)$	$\bar{y}_{1mi}^{*'} + \beta_{y_i x_i}^{'}(f_{2i} - f_{1i})]$
L_{m3i}	$\left(\frac{\bar{y}_{2mi}^{*'}}{\bar{x}_{2mi}^{*'}}\right)\bar{x}_{2ni}^{*'}$, where, $\bar{y}_{2mi}^{*'} = \bar{y}_{mi}^{'} + b_{y_i z}^{'}(m)(\bar{Z} - \bar{z}_m),\ \bar{x}_{2ni}^{*'} = \bar{x}_{ni}^{'} + b_{x_i z}^{'}(n)(\bar{Z} - \bar{z}_n)$ & $\bar{x}_{2mi}^{*'} = \bar{x}_{mi}^{'} + b_{x_i z}^{'}(m)(Z - \bar{z}_m)$	$\frac{\bar{y}_{2mi}^{*'}}{f_{1i}} f_{2i}$
L_{m4i}	$\bar{y}_{2mi}^{*'} + b_{y_i x_i}^{'}(m)(\bar{x}_{3ni}^{*'} - \bar{x}_{3mi}^{*'})$, where, $\bar{x}_{3ni}^{*'} = \frac{\bar{x}_{ni}^{'}}{\bar{z}_n}\bar{Z},\ \bar{x}_{3mi}^{*'} = \frac{\bar{x}_{mi}^{'}}{\bar{z}_m}\bar{Z}$	$\bar{y}_{2mi}^{*'} + b_{y_i x_i}^{'}(m)(f_{2i} - f_{1i})$
L_{m5i}	$\bar{y}_{4mi}^{*'} + b_{y_i x_i}^{'}(m)(\bar{x}_{3ni}^{*'} - \bar{x}_{3mi}^{*'})$, where, $\bar{y}_{4mi}^{*'} = \frac{\bar{y}_{mi}^{'}}{\bar{z}_m}Z$	$\bar{y}_{4mi}^{*'} + b_{y_i x_i}^{'}(m)(f_{2i} - f_{1i})$
L_{m6i}	$\frac{\bar{y}_{2mi}^{*'}}{\bar{x}_{2mi}^{*'}}\bar{x}_{3ni}^{*'}$	$\frac{\bar{y}_{2mi}^{*'}}{f_{1i}} f_{2i}$

2.4.3 Combined General Class of Estimators

Considering the convex linear combinations of the two classes of estimators L_{ui} and L_{mi} based on sample of size u and m respectively, the final estimator for population mean of coded response variable is given as

$$L_i = \Psi_i^* L_{ui} + (1-\Psi_i^*) L_{mi} \; ; \; i = 1 \; and \; 2 \tag{5}$$

where the class of estimators L_{ui} and L_{mi} are defined in equations 3 and 4 respectively and $\Psi_i^* \; \varepsilon \; [0,1]$ is a scalar quantity to be chosen suitably.
Many existing estimators for population mean at current wave by eminent researches in successive sampling can be the members of the proposed class when modified to work for estimation of sensitive population mean of coded response variable at current wave. Some of them are modified and given as:

$L_{1i} = \Psi_{1i}^* L_{u1i} + (1-\Psi_{1i}^*) L_{m1i}$, (Modified Jessen (1942) estimator)
$L_{2i} = \Psi_{1i}^* L_{u10i} + (1-\Psi_{1i}^*) L_{m2i}$, (Modified Singh and Priyanka (2008))
$L_{3i} = \Psi_{2i}^* L_{u11i} + (1-\Psi_{2i}^*) L_{m3i}$, (Modified Singh and Karna (2009) estimator)

$$\left.\begin{aligned} L_{4i} &= \Psi_{4i}^* L_{u4i} + (1-\Psi_{4i}^*) L_{m4i}, \\ L_{5i} &= \Psi_{5i}^* L_{u4i} + (1-\Psi_{5i}^*) L_{m5i}, \\ L_{6i} &= \Psi_{6i}^* L_{u4i} + (1-\Psi_{6i}^*) L_{m6i}. \end{aligned}\right\} \text{(Modified Singh and Prasad (2010) estimator)}$$

etc.,

3. Features of proposed General Class of Estimators

3.1. Bias and Mean Squared Error

The bias and mean squared error of class of estimators L_{ui} and L_{mi} are derived up to first order approximations under large sample assumptions and using the following transformations.
$\bar{y}_{ui}^{'} = \bar{Y}_i^{'} (1+e_{1i})$, $\bar{y}_{mi}^{'} = \bar{Y}_i^{'} (1+e_{2i})$, $\bar{x}_{mi}^{'} = \bar{X}_i^{'} (1+e_{3i})$, $\bar{x}_{ni}^{'} = \bar{X}_i^{'} (1+e_{4i})$, $\bar{z}_m = \bar{Z} (1+e_5)$, $\bar{z}_u = \bar{Z} (1+e_6)$, $\bar{z}_n = \bar{Z} (1+e_7)$, $\bar{x}_{ui}^{'} = \bar{X}_i^{'} (1+e_{8i})$, $s_{x_i^{'}}^2(m) = S_{x_i^{'}}^2 (1+e_{9i})$, $s_{y_i^{'}z}(u) = S_{y_i^{'}z} (1+e_{10i})$, $s_{y_i^{'}z}(m) = S_{y_i^{'}z} (1+e_{10i}^*)$, $s_z^2(u) = S_z^2 (1+e_{11})$, $s_z^2(m) = S_z^2 (1+e_{11}^*)$, $s_z^2(n) = S_z^2 (1+e_{11}^{**})$, $s_{x_i^{'}z}(n) = S_{x_i^{'}z} (1+e_{12i})$, $s_{x_i^{'}z}(m) = S_{x_i^{'}z} (1+e_{12i}^*)$, such that, $E(e_{si}) = 0$; $|e_{si}| < 1$; $E(e_k) = 0$; $|e_k| < 1$ where, $i = 1$ and 2; $s = 1, 2, 3, 4, 8, 9, 10$ and 12 and $k = 5, 6, 7$ *and* 11.

3.1.1 The Bias and Mean Squared Error of L_{ui}

The expressions of bias and mean squared error of the class of estimators L_{ui} are derived as

$$L_{ui} = U_i(\bar{y}_{ui}^{'}, \bar{z}_u)$$

Expanding $U_i\ (\bar{y}'_{ui},\ \bar{z}_u)$ about the point $K_i\ =\ (\bar{Y}'_i, \bar{Z})$ in a first order Taylor series, we have

$$L_{ui} = [\ U_i(K_i) + \ (\bar{y}'_{ui} - \ \bar{Y}'_i)\ d_{i1} + (\bar{z}_u - \ \bar{Z})\ d_{i2} + \frac{1}{2}\ \left\{(\bar{y}'_{ui} - \ \bar{Y}'_i)^2\ d_{i11} + (\bar{z}_u - \ \bar{Z})^2\ d_{i22} + 2\ (\bar{y}'_{ui} - \ \bar{Y}'_i)(\bar{z}_u - \ \bar{Z})\ d_{i12}\right\} + \ldots] \quad (6)$$

where,
$d_{i1} = \ \frac{\partial U_i}{\partial \bar{y}'_{ui}}|_{K_i},\ d_{i2} = \ \frac{\partial U_i}{\partial \bar{z}_u}|_{K_i},\ d_{i11} = \ \frac{\partial^2 U_i}{\partial \bar{y}'^2_{ui}}|_{K_i},\ d_{i22} = \ \frac{\partial^2 U_i}{\partial \bar{z}_u^2}|_{K_i},$

$d_{i12} = \ \frac{\partial^2 U_i}{\partial \bar{y}'_{ui}\partial \bar{z}_u}|_{K_i}$; $K_i = \ (\bar{Y}'_i,\ \bar{Z})$ and $i = \ 1$ and 2
Applying large sample approximations in equation 6, and retaining terms up to first order approximations we have,

$$\left(L_{ui}\ -\ \bar{Y}'_i\right) = \ \left[\bar{Y}'_i e_{1i} + \bar{Z} e_6 d_{i2} + \frac{1}{2}\left\{\bar{Y}'^2_i e^2_{1i} d_{i11} + \bar{Z}^2 e^2_6 d_{i22} + 2\bar{Y}'_i \bar{Z} e_{1i} e_6 d_{i12}\right\}\right] \quad (7)$$

Taking expectations on both sides in the above equation 7 and assuming the population size is sufficiently large, we get bias of L_{ui} up to first order approximation as

$$B(L_{ui}) = \frac{1}{u}\left[\frac{1}{2}(d_{i11}\bar{Y}'^2_i C^2_{y'_i} + \bar{Z}^2 C^2_z d_{i22}) + (\rho_{y'_i z} C_{y'_i} C_z \bar{Y}'_i \bar{Z} d_{i12})\right] \quad (8)$$

Now, squaring both sides of above equation 7 and retaining terms up to first order of approximations, we have

$$\left(L_{ui}\ -\ \bar{Y}'_i\right)^2\ =\ \left[\bar{Y}'^2_i e^2_{1i} + \bar{Z}^2 e^2_6 d^2_{i2} + 2\bar{Y}'_i \bar{Z} e_{1i} e_6 d_{i2}\right]$$

Taking expectations on both sides in the above equation and assuming the population is very large i.e.,$N \to \infty$, the mean squared error of L_{ui} is obtained as

$$M(L_{ui}) = \frac{1}{u}\left[\bar{Y}'^2_i C_{y'_i 2} + \bar{Z}^2 C^2_z d^2_{i2} + 2\rho_{y'_i z} C_{y'_i} C_z \bar{Y}'_i \bar{Z} d_{i2}\right]$$

which is optimized for $d_{i2} = -\rho_{y'_i z}$, Further, substituting optimized value of d_{i2} in the above equation we obtain the required optimum mean squared error of L_{ui} as

$$M(L_{ui})_{opt.} = \frac{1}{u}\left[\bar{Y}'^2_i C_{y'_i 2} + \bar{Z}^2 C^2_z \rho^2_{y'_i z} - 2\rho_{y'_i z} C_{y'_i} C_z \bar{Y}'_i \bar{Z} \rho_{y'_i z}\right]$$

Remark 5. *Since x' and y' are the same variables over two waves and z is the stable auxiliary variable so as pointed out by Murthy (1967), Cochran(1977), Reddy(1978), Feng and Zou (1996) and Singh and Ruiz-Espejo (2003) the coefficient of variation is stable in nature, so we assume that the coefficients of variation x', y' and z are almost equal (i.e., $C_{y'} \cong C_{x'} \cong C_z$).*

From the above remark 5 we state the following theorem.

Theorem 3.1. *To the first degree of approximations, the bias and mean squared error of L_{ui} under assumption given in remark 5 is*

$$B(L_{ui}) = \frac{1}{2u}\left[d_{i11} + d_{i22} + 2d_{i12}\rho_{y_i'z}\right] S^2_{y_i'} \tag{9}$$

and

$$M(L_{ui})_{opt.} = \frac{1}{u}\left(1 - \rho^2_{y_i'z}\right) S^2_{y_i'} \tag{10}$$

which is similar to the variance of linear regression estimator for population mean.

3.1.2 The Bias and Mean Squared Error of L_{mi}

For deriving the bias and mean squared error of class of estimators L_{mi}, $f_{1i} = g_i(\bar{x}'_{mi}, \bar{z}_m, \bar{z}_n)$ and $f_{2i} = h_i(\bar{x}'_{ni}, \bar{z}_n)$ have been expanded around the points $(\bar{X}'_i, \bar{Z}, \bar{Z})$ and $(\bar{X}'_i, \bar{Z})$ respectively by first order Taylor's series and neglecting the remainder terms we get,
$f_{1i} = \bar{X}'_i + G_1(\bar{x}'_{mi} - \bar{X}'_i) + G_2[(\bar{z}_m - \bar{Z}) - (\bar{z}_n - \bar{Z})]$ and
$f_{2i} = \bar{X}'_i + H_1(\bar{x}'_{ni} - \bar{X}'_i) + H_2(\bar{z}_n - \bar{Z})$.
Following Senapati and Sahoo(2006), we assume $H_1 = 1$ because $h_i(\bar{X}'_i, \bar{Z}) = \bar{X}'_i$ and $G_1 = 1$, $G_2 = -G_3$ because $g_i(\bar{x}'_{mi}, \bar{z}_m, \bar{z}_n)$ and $g_i(\bar{x}'_{mi}, \bar{z}_n, \bar{z}_m)$ assume the same value i.e., $\bar{X}'_i$ at $(\bar{X}'_i, \bar{Z}, \bar{Z})$. Hence we have

$$f_{1i} = \bar{X}'_i + (\bar{x}'_{mi} - \bar{X}'_i) + G_2[(\bar{z}_m - \bar{Z}) - (\bar{z}_n - \bar{Z})] \tag{11}$$

and

$$f_{2i} = \bar{X}'_i + (\bar{x}'_{ni} - \bar{X}'_i) + H_2(\bar{z}_n - \bar{Z}) \tag{12}$$

Similarly, observing $F_1 = 1$, $F_2 = -F_3$ and expanding $T_i(\bar{y}'_{mi}, f_{1i}, f_{2i})$ around the point $(\bar{Y}'_i, \bar{X}'_i, \bar{X}'_i)$ by first order Taylor's series, we have,
$L_{mi} = \bar{Y}'_i + F_1(\bar{y}'_{mi} - \bar{Y}'_i) + F_2[(f_{1i} - \bar{X}'_i) - (f_{2i} - \bar{X}'_i)]$, i.e.,

$$\begin{aligned} L_{mi} = {} & \bar{Y}'_i + (\bar{y}'_{mi} - \bar{Y}'_i) + F_2[(f_{1i} - \bar{X}'_i) - (f_{2i} - \bar{X}'_i)] + \\ & \frac{1}{2}[(\bar{y}'_{mi} - \bar{Y}'_i)^2 F_{11} + (f_{1i} - \bar{X}'_i)^2 F_{22} + (f_{2i} - \bar{X}'_i)^2 F_{33} + \\ & 2(\bar{y}'_{mi} - \bar{Y}'_i)(f_{1i} - \bar{X}'_i) F_{12} + 2(\bar{y}'_{mi} - \bar{Y}'_i)(f_{2i} - \bar{X}'_i) F_{13} + \\ & 2(f_{1i} - \bar{X}'_i)(f_{2i} - \bar{X}'_i) F_{23}]. \end{aligned} \tag{13}$$

where,
$F_1 = \frac{\partial \bar{f}_i}{\partial \bar{y}'_{mi}}|_{s^*_{1i}} = 1$, $F_2 = \frac{\partial \bar{f}_i}{\partial \bar{x}'_{mi}}|_{s^*_{1i}}$, $F_3 = \frac{\partial \bar{f}_i}{\partial \bar{x}'_{ni}}|_{s^*_{1i}}$, $F_{11} = 0$, $F_{22} = \frac{\partial^2 \bar{f}_i}{\partial \bar{x}'^2_{mi}}|_{s^*_{1i}}$,
$F_{33} = \frac{\partial^2 \bar{f}_i}{\partial \bar{x}'^2_{ni}}|_{s^*_{1i}}$, $F_{12} = \frac{\partial^2 \bar{f}_i}{\partial \bar{y}'_{mi} \partial \bar{x}'_{mi}}|_{s^*_{1i}}$, $F_{13} = \frac{\partial^2 \bar{f}_i}{\partial \bar{y}'_{mi} \partial \bar{x}'_{ni}}|_{s^*_{1i}}$, $F_{23} = \frac{\partial^2 \bar{f}_i}{\partial \bar{x}'_{mi} \partial \bar{x}'_{ni}}|_{s^*_{1i}}$,

$G_1 = \frac{\partial \bar{g}_i}{\partial \bar{y}'_{mi}}|s^*_{2i} = 1,\ G_2 = \frac{\partial \bar{g}_i}{\partial \bar{x}'_{mi}}|s^*_{2i},\ G_3 = \frac{\partial \bar{g}_i}{\partial \bar{x}'_{ni}}|s^*_{2i},\ G_{11} = 0,$
$G_{22} = \frac{\partial^2 \bar{g}_i}{\partial \bar{x}'^2_{mi}}|s^*_{2i},\ G_{33} = \frac{\partial^2 \bar{g}_i}{\partial \bar{x}'^2_{ni}}|s^*_{2i},\ G_{12} = \frac{\partial^2 \bar{g}_i}{\partial \bar{y}'_{mi}\partial \bar{x}'_{mi}}|s^*_{2i},\ G_{13} = \frac{\partial^2 \bar{g}_i}{\partial \bar{y}'_{mi}\partial \bar{x}'_{ni}}|s^*_{2i},$
$G_{23} = \frac{\partial^2 \bar{g}_i}{\partial \bar{x}'_{mi}\partial \bar{x}'_{ni}}|s^*_{2i},\ H_1 = \frac{\partial \bar{h}_i}{\partial \bar{y}'_{mi}}|s^*_{3i} = 1,\ H_2 = \frac{\partial \bar{h}_i}{\partial \bar{x}'_{mi}}|s^*_{3i},\ H_3 = \frac{\partial \bar{h}_i}{\partial \bar{x}'_{ni}}|s^*_{3i},$
$H_{11} = 0,\ H_{22} = \frac{\partial^2 \bar{h}_i}{\partial \bar{x}'^2_{mi}}|s^*_{3i},\ H_{33} = \frac{\partial^2 \bar{h}_i}{\partial \bar{x}'^2_{ni}}|s^*_{3i},\ H_{12} = \frac{\partial^2 \bar{h}_i}{\partial \bar{y}'_{mi}\partial \bar{x}'_{mi}}|s^*_{3i},$
$H_{13} = \frac{\partial^2 \bar{h}_i}{\partial \bar{y}'_{mi}\partial \bar{x}'_{ni}}|s^*_{3i},\ H_{23} = \frac{\partial^2 \bar{h}_i}{\partial \bar{x}'_{mi}\partial \bar{x}'_{ni}}|s^*_{3i}$

where $S^*_{1i} = (\bar{Y}'_i,\ \bar{X}'_i,\ \bar{X}'_i),\ S^*_{2i} = (\bar{X}'_i,\ \bar{Z},\ \bar{Z})$ and $S^*_{3i} = (\bar{X}'_i,\ \bar{Z})$; $i = 1$ and 2 correspond to randomized and scrambled response models considered.

After applying large sample approximations in equation 13 taking relevant expectations, simplifying and retaining terms up to first order of approximation we get the bias and mean squared error of L_{mi} for large N as:

$$
\begin{aligned}
B(L_{mi}) = \frac{1}{2}[\frac{1}{m}((\bar{X}'^2_i C_{x'_i} G_{11} + \bar{Z}^2 C_{z^2} G_{22} + \bar{X}'_i \bar{Z} \rho_{x'_i z} C_{x'_i} C_z G_{12}) F_2 + F_{11} \bar{Y}'^2_i C_{y'^2_i} + \\
(\bar{X}'^2_i C^2_{x'_i} + \bar{Z}^2 C^2_z G^2_2 + 2\bar{X}'_i \bar{Z} \rho_{x'_i z} C_{x'_i} C_z G_2) F_{22} + 2(\bar{X}'_i \bar{Y}'_i \rho_{y'_i x'_i} C_{x'_i} C_{y'_i} + \\
\bar{Y}'_i \bar{Z} \rho_{y'_i z} C_{y'_i} C_z G_2) F_{12}) + \frac{1}{n}([(\bar{Z}^2 C^2_z G_{33} + \bar{X}'_i \bar{Z} \rho_{x'_i z} C_{x'_i} C_z G_{13} + \bar{Z}^2 C^2_z G_{23}) F_2 - \\
(\bar{X}'_i C_{x'^2_i} H^2_1 + \bar{Z}^2 C^2_z H^2_2 + \bar{X}'_i \bar{Z} \rho_{x'_i z} C_{x'_i} C_z H_{12})] F_2 - (\bar{Z}^2 G^2_2 C^2_z + 2\bar{X}'_i \bar{Z} G_2 \rho_{x'_i z} C_{x'_i} C_z) F_{22} + \\
(\bar{X}'_i C_{x'^2_i} + \bar{Z}^2 C^2_z H^2_2 + 2\rho_{x'_i z} C_{x'_i} C_z \bar{X}'_i \bar{Z} H_2) F_{33} - 2\rho_{y'_i z} C_{y'_i} C_z \bar{Y}'_i \bar{Z} G_2 F_{12} + \\
2(\rho_{y'_i x'_i} C_{y'_i} C_{x'_i} \bar{X}'_i \bar{Y}'_i + \bar{Y}'_i \bar{Z} H_2 \rho_{y'_i z} C_{y'_i} C_z) F_{13} + 2(\bar{X}'^2_i C^2_{x'_i} + \rho_{x'_i z} \bar{X}'_i \bar{Z} C_{x'_i} C_z H_2) F_{23})]
\end{aligned}
\tag{14}
$$

$$
\begin{aligned}
M(L_{mi}) = \bar{Y}'_i \frac{1}{m} C_{y'^2_i} + \bar{X}'_i F^2_2 (\frac{1}{m} C_{x'^2_i} - \frac{1}{n} C_{x'^2_i}) + F^2_2 G^2_2 \bar{Z}^2 (\frac{1}{m} C_{z^2} - \frac{1}{n} C_{z^2}) + \\
F^2_2 H^2_2 \bar{Z}^2 \frac{1}{n} C^2_z - 2\bar{Y}'_i \bar{X}'_i F_2 (\frac{1}{m} \rho_{x'_i y'_i} C_{y'_i} C_{x'_i} - \frac{1}{n} \rho_{x'_i y'_i} C_{y'_i} C_{x'_i}) + \\
2\bar{Y}'_i \bar{Z} F_2 G_2 (\frac{1}{m} \rho_{y'_i z} C_{y'_i} C_z - \frac{1}{n} \rho_{y'_i z} C_{y'_i} C_z) + 2\bar{Y}'_i F_2 H_2 \bar{Z} \frac{1}{n} \rho_{y'_i z} C_{y'_i} C_z - \\
2\bar{X}'_i \bar{Z} F^2_2 G_2 (\frac{1}{m} \rho_{x'_i z} C_{x'_i} C_z - \frac{1}{n} \rho_{x'_i z} C_{x'_i} C_z)
\end{aligned}
\tag{15}
$$

which is further optimized for

$(F_2)_{opt.} = \frac{\rho_{y'_i x'_i} - \rho_{x'_i z} \rho_{y'_i z}}{\rho^2_{x'_i z} - 1}$ $(say\ F^*_2)$, $(G_2)_{opt.} = \frac{\rho_{y'_i x'_i} \rho_{x'_i z} - \rho_{y'_i z}}{\rho_{x'_i z} \rho_{y'_i z} - \rho_{y'_i x'_i}}$ $(say\ G^*_2)$

and $(H_2)_{opt.} = \frac{\rho_{y'_i z} (\rho^2_{x'_i z} - 1)}{\rho_{y'_i x'_i} - \rho_{x'_i z} \rho_{y'_i z}}$ $(say\ H^*_2)$.

Further, substituting minimum value of F_2, G_2 and H_2 in the above equation we obtain the optimum mean squared error of L_{mi} as

$$M(L_{mi})_{opt.} = \bar{Y}_i' \frac{1}{m} C_{y_i}^2 + \bar{X}_i' F_2^{*2} (\frac{1}{m} C_{x_i}^2 - \frac{1}{n} C_{x_i}^2) + F_2^2 G_2^{*2} \bar{Z}^2 (\frac{1}{m} C_z^2 - \frac{1}{n} C_z^2) +$$
$$F_2^{2*} H_2^{*2} \bar{Z}^2 \frac{1}{n} C_z^2 - 2\bar{Y}_i' \bar{X}_i' F_{*2} (\frac{1}{m} \rho_{x_i' y_i'} C_{y_i'} C_{x_i'} - \frac{1}{n} \rho_{x_i' y_i'} C_{y_i'} C_{x_i'}) +$$
$$2\bar{Y}_i' \bar{Z} F_2^* G_2^* (\frac{1}{m} \rho_{y_i' z} C_{y_i'} C_z - \frac{1}{n} \rho_{y_i' z} C_{y_i'} C_z) + 2\bar{Y}_i' F_2^* H_2^* \bar{Z} \frac{1}{n} \rho_{y_i' z} C_{y_i'} C_z -$$
$$2\bar{X}_i' \bar{Z} F_2^{*2} G_2^* (\frac{1}{m} \rho_{x_i' z} C_{x_i'} C_z - \frac{1}{n} \rho_{x_i' z} C_{x_i'} C_z)$$

From the remark 5 we state the following theorem.

Theorem 3.2. *To the first degree of approximations, the bias and mean squared error of L_{mi} under assumption given in remark 5, is given by*

$$B(L_{mi}) = \left[\frac{1}{m} (a^*) + \frac{1}{n} (b^*) \right] \frac{S_{y_i}^2}{2} \tag{16}$$

Where,

$$a^* = \left(G_{11} + G_{22} + G_{12} \rho_{x_i' z} \right) F_2 + F_{11} + \left(1 + G_2^2 + 2\rho_{x_i' z} G_2 \right) F_{22} +$$
$$2 \left(\rho_{y_i' x_i'} + \rho_{y_i' z} G_2 \right) F_{12},$$
$$b^* = \left[\left(G_{33} + \rho_{x_i' z} G_{13} + G_{23} \right) - \left(H_1^2 + H_2^2 + \rho_{x_i' z} H_{12} \right) \right] F_2 - \left(G_2^2 + 2\rho_{x_i' z} G_2 \right) F_{22}$$
$$+ \left(1 + H_2^2 + 2\rho_{x_i' z} H_2 \right) F_{33} + 2\rho_{y_i' z} F_{12} + 2 \left(\rho_{y_i' x_i'} + \rho_{y_i' z} H_2 \right) F_{13} +$$
$$2 \left(1 + H_2 \rho_{x_i' z} \right) F_{23}.$$

and

$$M(L_{mi})_{opt.} = [(\frac{1}{m} - \frac{1}{n})(F_2^{*2} + F_2^{*2} G_2^{*2} + 2\rho_{y_i' x_i'} F_2^* + 2\rho_{y_i' z} G_2^* F_2^* + 2\rho_{x_i' z} G_2^* F_2^{*2}) +$$
$$\frac{1}{n} (F_2^{*2} + H_2^{*2} - 2\rho_{y_i' z} F_2^* H_2^*) + \frac{1}{m}] S_{y_i}^2 \tag{17}$$

Theorem 3.3. *Bias of the general class of estimators L_i to the first order of approximations are obtained as*

$$B(L_i) = \Psi_i^* B(L_{ui}) + (1 - \Psi_i^*) B(L_{mi}) \tag{18}$$

Substituting the values of $B(L_{ui})$ and $B(L_{mi})$ from the equations 9 and 16 in the above equation, we have the expression for the bias of the general class of estimators L_i in equation 18.

Theorem 3.4. *Mean squared error of the general class of estimators L_i to first order of approximations are obtained as*

$$M(L_i) = \Psi_i^{*2} M(L_{ui})_{opt.} + (1-\Psi_i^*)^2 M(L_{mi})_{opt.} \tag{19}$$

The optimized values of $M(L_{ui})$ and $M(L_{mi})$ are computed in equation 10 and equation 17 respectively and as the two classes of estimators L_{ui} and L_{mi} are based on two non-overlaping samples of sizes u and m respectively.
So, $cov(L_{ui},\ L_{mi}) = 0$. Hence, using these values in above equation 19 we get the mean squared error of L_i.

3.2. Optimum Mean Squared Error of the Proposed class of Estimator

The mean squared error of class of estimators L_i is a function of unknown constant Ψ_i^* therefore, it is minimized with respect to Ψ_i^* and hence the optimum value of Ψ_i^* is obtained as

$$\Psi_{iopt.}^* = \frac{M[L_{mi}]_{opt.}}{M[L_{ui}]_{opt.} + M[L_{mi}]_{opt.}} \tag{20}$$

Substituting the value of $\Psi_{iopt.}^*$ from equation 20 in equation 19, we get the optimum mean squared error of the class of estimator L_i as

$$M[L_i]_{opt.} = \frac{M[L_{ui}]_{opt.} \times M[L_{mi}]_{opt.}}{M[L_{ui}]_{opt.} + M[L_{mi}]_{opt.}} \tag{21}$$

Further, substituting the values $M[L_{ui}]_{opt.}$ and $M[L_{mi}]_{opt.}$ from equations 10 and equation 17 in equation 21, the simplified values of $M[L_i]_{opt.}$ is derived as

$$M[L_i]_{opt.} = \frac{B_{1i}^* \mu_i + B_{2i}^*}{\mu_i^2 A_{3i}^* - \mu_i B_{3i}^* + A_{1i}^*} \left(\frac{S_{y_i'}^2}{n} \right) \tag{22}$$

where,
$A_{1i}^* = 1 - \rho_{y_i' z}^2,\ A_{2i}^* = d^* + 1,\ A_{3i}^* = d^* - H_2^{*2} F_2^{*2} + 2\rho_{y_i' z} H_2^* F_2^*,$
$d^* = F_2^{*2} + G_2^{*2} F_2^{*2} + 2\rho_{y_i' x_i'} F_2^* + 2\rho_{y_i' z} F_2^* G_2^* + 2\rho_{x_i' z} F_2^{*2} G_2^*,$
$B_{1i}^* = A_{1i}^* A_{3i}^*,\ B_{2i}^* = A_{1i}^* A_{2i}^* - A_{1i}^* A_{3i}^*,\ B_{3i}^* = A_{1i}^* - A_{2i}^* + A_{3i}^*.$

3.3. Optimum Rotation Rate

Rotation rate is an important aspect in successive sampling as it is directly related to total cost of survey. More the sample rotated/ matched from previous wave, lesser number of units will be required to be drawn at current wave. Hence, mean squared error of the estimator L_i $(i = 1 \text{ and } 2)$ derived in equation 22 which is a function of μ_i, have been optimized with respect to μ_i $(i = 1 \text{ and } 2)$. The optimum value of

μ_i say $\hat{\mu}^*_{fi}$ have been obtained which satisfies the condition given as:

$$0 < min\left\{\frac{-C^*_{2i}+\sqrt{C^{*2}_{2i}+C^*_{1i}C^*_{3i}}}{C^*_{1i}},\ \frac{-C^*_{2i}-\sqrt{C^{*2}_{2i}+C^*_{1i}C^*_{3i}}}{C^*_{1i}}\right\} < 1 \qquad (23)$$

where, $C^*_{1i} = B^*_{1i}A^*_{3i}$, $C^*_{2i} = A^*_{3i}B^*_{2i}$ and $C^*_{3i} = A^*_{1i}B^*_{1i} + B^*_{3i}B^*_{2i}$.
Substituting the applicable value of $\hat{\mu}^*_{fi}$ in equation 23, we have the optimum value of the mean squared error of the general class of estimators L_i as,

$$M(L_i)_{opt.^*} = \frac{B^*_{1i}\mu^*_{fi} + B^*_{2i}}{\mu^{2*}_{fi}A^*_{3i} - \mu^*_{fi}B^*_{3i} + A^*_{1i}}\left(\frac{S^2_{y_i}}{n}\right)\ ;\ i = 1\ and\ 2. \qquad (24)$$

4. Performance of Proposed Composite class of estimator

The proposed general class of estimator have been compared with the member of its class listed in section 2.4.3. Therefore their optimum fraction of sample to be drawn afresh at current wave and the optimum mean squared error have been computed and are presented below in Table 3 and Table 4 respectively.

Table 3. Optimum rotation rate for proposed estimators

Estimator	Optimum Rotation Rate
L_{1i}	$\hat{\mu}_{Ji}$ satisfies $0 < min\left\{\frac{1+\sqrt{1-\rho^2_{y_ix_i}}}{\rho^2_{y_ix_i}},\ \frac{1-\sqrt{1-\rho^2_{y_ix_i}}}{\rho^2_{y_ix_i}}\right\} < 1$
L_{2i}	$\hat{\mu}_{spi}$ satisfies $0 < min\left\{\frac{-A^*_{1i}+\sqrt{A^*_{1i}(A^*_{1i}+D^*_{1i})}}{D^*_{1i}},\ \frac{-A^*_{1i}-\sqrt{A^*_{1i}(A^*_{1i}+D^*_{1i})}}{D^*_{1i}}\right\} < 1$
L_{3i}	$\hat{\mu}_{ski}$ satisfies $0 < min\left\{\frac{I_{2i}+\sqrt{I^2_{2i}-I_{1i}I_{3i}}}{I_{1i}},\ \frac{I_{2i}-\sqrt{I^2_{2i}-I_{1i}I_{3i}}}{I_{1i}}\right\} < 1$
L_{4i}	$\hat{\mu}_{sp1i}$ satisfies $0 < min\left\{\frac{I_{12i}+\sqrt{I^2_{12i}-I_{11i}I_{13i}}}{I_{11i}},\ \frac{I_{12i}-\sqrt{I^2_{12i}-I_{11i}I_{13i}}}{I_{11i}}\right\} < 1$
L_{5i}	$\hat{\mu}_{sp2i}$ satisfies $0 < min\left\{\frac{I_{22i}+\sqrt{I^2_{22i}-I_{21i}I_{23i}}}{I_{21i}},\ \frac{I_{22i}-\sqrt{I^2_{22i}-I_{21i}I_{23i}}}{I_{21i}}\right\} < 1$
L_{6i}	$\hat{\mu}_{sp3i}$ satisfies $0 < min\left\{\frac{I_{32i}+\sqrt{I^2_{32i}-I_{31i}I_{33i}}}{I_{31i}},\ \frac{I_{32i}-\sqrt{I^2_{32i}-I_{31i}I_{33i}}}{I_{31i}}\right\} < 1$

Table 4. Mean Squared Error

Estimator	Optimum Mean Squared Error
L_{1i}	$\left(\frac{1-\hat{\mu}_{ji}\rho^2_{y_i x_i}}{1-\hat{\mu}^2_{ji}\rho^2_{y_i x_i}}\right)\frac{S^2_{y_i}}{n}$
L_{2i}	$\left(\frac{[A^*_{1i}(A^*_{1i}+\hat{\mu}_{spi}D^*_{1i})]}{A^*_{1i}+\hat{\mu}^2_{spi}D^*_{1i}}\right)\left(\frac{S^2_{y_i}}{n}\right)$ where $D^*_{1i}=2\rho^2_{y_i z}\rho_{y_i x_i}-\rho^2_{y_i x_i}(1+\rho^2_{y_i z})$
L_{3i}	$\left(\frac{\hat{\mu}_{ski}g_{1i}-g_{2i}}{\hat{\mu}^2_{ski}K_{3i}-\hat{\mu}_{ski}g_{3i}-K_{1i}}\right)\left(\frac{S^2_{y_i}}{n}\right)$ where $k_{1i}=1-\rho^2_{y_i z}$, $k_{2i}=2-\rho^2_{y_i z}-2\rho_{y_i x_i}-\rho^2_{x_i z}+2\rho_{y_i z}\rho_{y_i z}$, $k_{3i}=2\rho_{y_i x_i}+\rho^2_{x_i z}-2\rho_{x_i z}\rho_{y_i z}-1$, $g_{1i}=k_{1i}k_{3i}, g_{2i}=k_{1i}k_{2i}+k_{1i}k_{3i}$, $g_{3i}=k_{2i}-k_{1i}+k_{3i}$, $I_{1i}=k_{3i}g_{1i}$, $I_{2i}=k_{3i}g_{2i}$ and $I_{3i}=k_{1i}g_{1i}+g_{2i}g_{3i}$
L_{4i}	$\left(\frac{\hat{\mu}_{sp1i}g_{11i}-g_{12i}}{\hat{\mu}^2_{sp1i}K_{13i}-\hat{\mu}_{sp1i}B_{13i}-K_{11i}}\right)\left(\frac{S^2_{y_i}}{n}\right)$ where $k_{11i}=2-2\rho_{y_i z}$, $k_{12i}=1-\rho^2_{y_i z}-2\rho^2_{y_i x_i}\rho_{x_i z}+2\rho_{y_i z}\rho_{y_i x_i}\rho_{x_i z}$, $k_{13i}=2\rho^2_{y_i x_i}\rho_{x_i z}-2\rho_{y_i z}\rho_{y_i x_i}\rho_{x_i z}$, $g_{11i}=k_{11i}k_{13i}, g_{12i}=k_{11i}k_{12i}+k_{11i}k_{13i}$, $g_{13i}=k_{12i}-k_{11i}+k_{13i}$, $I_{11i}=k_{13i}g_{11i}$, $I_{12i}=k_{13i}g_{12i}$ and $I_{13i}=k_{11i}g_{11i}+g_{12i}g_{13i}$
L_{5i}	$\left(\frac{\hat{\mu}_{sp2i}g_{21i}-g_{22i}}{\hat{\mu}^2_{sp2i}K_{23i}-\hat{\mu}_{sp2i}g_{23i}-K_{11i}}\right)\left(\frac{S^2_{y_i}}{n}\right)$ where $k_{22i}=2-2\rho^2_{y_i x_i}\rho_{x_i z}-2\rho_{y_i z}-2\rho_{y_i x_i}+2\rho_{y_i x_i}\rho_{x_i z}+2\rho_{y_i x_i}\rho_{y_i z}$, $k_{23i}=2\rho^2_{y_i x_i}\rho_{x_i z}+2\rho_{y_i x_i}-2\rho_{y_i x_i}\rho_{x_i z}-2\rho_{y_i x_i}\rho_{y_i z}$, $g_{21i}=k_{11i}k_{23i}$, $g_{22i}=k_{11i}k_{22i}+k_{11i}k_{23i}$, $g_{23i}=k_{22i}-k_{11i}+k_{23i}, I_{21i}=k_{23i}g_{21i}$, $I_{22i}=k_{23i}g_{22i}$ and $I_{23i}=k_{11i}g_{21i}+g_{22i}g_{23i}$
L_{6i}	$\left(\frac{\hat{\mu}_{sp3i}B_{31i}-g_{32i}}{\hat{\mu}^2_{sp3i}K_{33i}-\hat{\mu}_{sp3i}B_{33i}-K_{11i}}\right)\left(\frac{S^2_{y_i}}{n}\right)$ where $k_{32i}=2-\rho^2_{x_i z}-\rho^2_{y_i z}-2\rho_{y_i x_i}+2\rho_{y_i z}\rho_{x_i z}$, $k_{33i}=2\rho^2_{x_i z}-2\rho_{x_i z}+2\rho_{y_i x_i}-\rho_{y_i z}\rho_{x_i z}$, $g_{31i}=k_{11i}k_{33i}, g_{32i}=k_{11i}k_{32i}+k_{11i}k_{33i}$, $g_{33i}=k_{32i}-k_{11i}+k_{33i}, I_{31i}=k_{33i}g_{31i}, I_{32i}=k_{33i}g_{32i}$ and $I_{33i}=k_{1i}g_{31i}+g_{32i}g_{33i}$

5. Estimators for sensitive population mean at current wave

Replacing the population mean of coded response $\bar{Y}_i'$ $(i=1,\ 2)$ in equation 1 and equation 2 by its estimators L_i and L_{ij} $(i=1,\ 2\ ;\ j=1,\ 2,\ 3,\ 4,\ 5,\ 6)$, the corresponding estimators for sensitive population mean at current wave $\hat{\bar{Y}}_i$ and $\hat{\bar{Y}}_{ij}$ respectively is obtained and are given in Table 5.

Since, the estimators $\hat{\bar{Y}}_i$ and $\hat{\bar{Y}}_{ij}$ are biased, the mean squared errors of sensitive population mean estimators $\bar{Y}_{ji}$; $j=1,\ 2,\ 3,\ 4,\ 5,\ 6$ has also been computed under two considered models and are presented in Table 5.

Table 5. Sensitive population mean estimators $\hat{\bar{Y}}_i,\ \hat{\bar{Y}}_{ij}$ and Mean squared error of the estimators $\hat{\bar{Y}}_i,\ \hat{\bar{Y}}_{ij}$ under the models M_{AR} and M_{DP}

i	Model	Sensitive population mean estimator	Mean squared error of sensitive population mean
1	M_{AR}	$\hat{\bar{Y}}_1 = \frac{L_1 - p_2\bar{W}_2 - (1-p_1-p_2)\bar{W}_3}{p_1+p_2W_1}$	$M[\hat{\bar{Y}}_1] = \frac{M[L_1]_{opt.}*}{[p_1+p_2\bar{W}_1]^2}$
		$\hat{\bar{Y}}_{1j} = \frac{L_{1j} - p_2\bar{W}_2 - (1-p_1-p_2)\bar{W}_3}{p_1+p_2W_1}$	$M[\hat{\bar{Y}}_{1j}] = \frac{M[L_{1j}]_{opt.}*}{[p_1+p_2\bar{W}_1]^2}$
2	M_{DP}	$\hat{\bar{Y}}_2 = \frac{L_2 - \varphi_y^*\bar{W}_1}{\varphi_y^*+(1-\varphi_y^*)\bar{W}_2}$	$M[\hat{\bar{Y}}_2] = \frac{M[L_2]_{opt.}*}{[\varphi_y^*+(1-\varphi_y^*)\bar{W}_2]^2}$
		$\hat{\bar{Y}}_{2j} = \frac{L_{2j} - \varphi_y^*W_1}{\varphi_y^*+(1-\varphi_y^*)\bar{W}_2}$	$M[\hat{\bar{Y}}_{2j}] = \frac{M[L_{2j}]_{opt.}*}{[\varphi_y^*+(1-\varphi_y^*)\bar{W}_2]^2}$

6. Comparison

The percent relative efficiency of proposed general class of estimator for sensitive population mean $\hat{\bar{Y}}_i$ with respect to the estimator $\hat{\bar{Y}}_{ij}$ have been computed as

$$E_{ji} = \frac{M(Y_{ij})}{M(Y_i)} \times 100\ ;\ \forall\ i\ =\ 1\ and\ 2\ and\ j =\ 1,\ 2,\ 3,\ 4,\ 5\ and\ 6. \tag{25}$$

Remark 6. *In the present paper we have considered additive, multiplicative and upshot of additive and multiplicative type scrambled response models on two wave successive sampling. The three scrambling variable $W_1,\ W_2$ and W_3 used to perturb the true response through randomized or scrambled response models may follow any distribution. Hence, following Pollock and Bek (1976), Eichhorn and Hayre(1983) and Arcos et al.(2015), we consider scrambling variable W_1 to follow normal distribution with mean 1 and variance 1. However, the scrambling variable W_2 has been assumed to follow normal distribution with mean 0 and variance 1 and W_3 has been assumed to follow normal distribution with mean 1 and variance 2.*

7. Numerical Presentation

Population Source:[Priyanka and Mittal (2016)]
The population comprise of $N = 315$ units. Let x and y denote the average monthly expenditure on drug usage by undergraduate students in 2015 and 2016 respectively. However z denote the average monthly pocket money of undergraduate students from all sources. The parameters of considered population are computed as:
$N\ =\ 315,\ S_x^2\ =\ 1.2463\times10^6,\ S_y^2\ =\ 2.1926\times10^6,\ S_z^2\ =\ 1.4670\times10^7,$
$\bar{X} = 370.5238,\ \bar{Y} = 504.8095,\ \bar{Z} = 4.0233\times10^3,\ \rho_{yx} = 0.8937,$
$\rho_{xz} = 0.6491,\ \rho_{yz} = 0.7012.$
The artificial data for $W_1,\ W_2$ and W_3 have also been generated as per assumption in remark 6. It is observed that $\bar{W}_1 = 1.0871,\ S_{W_1}^2 = 0.5832,\ \bar{W}_2 = -0.0248,\ S_{W_2}^2 = 1.1695.$ and $\bar{W}_3 = 0.9731,\ S_{W_3}^2 = 4.4527$
The optimum values of $\hat{\mu}_i's$ for L_i and L_{ji} and percent relative efficiencies E_{ji} have been computed for the above data under two considered models and are presented in Figure 1 to Figure 13 and Table 6.

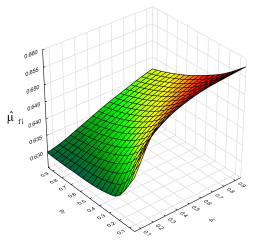


Figure 1: Optimum value of fraction of sample drawn afresh for estimator $\hat{\bar{Y}}_1$

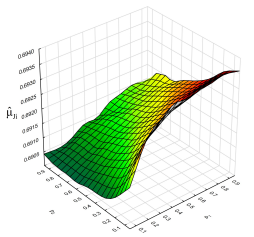


Figure 2: Optimum value of fraction of sample drawn afresh for estimator $\hat{\bar{Y}}_{11}$

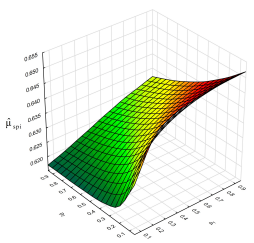


Figure 3: Optimum value of fraction of sample drawn afresh for estimator $\hat{\bar{Y}}_{12}$

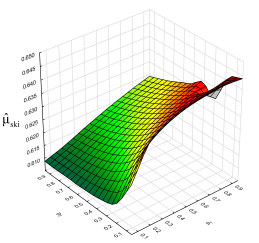


Figure 4: Optimum value of fraction of sample drawn afresh for estimator $\hat{\bar{Y}}_{13}$

Table 6. Optimum fraction of sample drawn afresh and percent relative efficiencies under scrambled response model. where $\alpha = \varphi_y^* = \varphi_x^*$

i	M_{DP}	$\hat{\mu}_{f2}$	$\hat{\mu}_{J2}$	$\hat{\mu}_{sp2}$	$\hat{\mu}_{sk2}$	$\hat{\mu}_{sp12}$	$\hat{\mu}_{sp22}$	$\hat{\mu}_{sp32}$	E_{12}	E_{22}	E_{32}	E_{42}	E_{52}	E_{62}
	α													
2	0.1	0.6562	0.6935	0.6501	0.6430	0.3782	0.6798	0.6880	142.97	100.93	102.05	105.51	122.06	138.91
	0.3	0.6489	0.6926	0.6414	0.6339	0.3996	0.6694	0.6793	151.88	101.18	102.37	104.13	119.76	137.04
	0.5	0.6413	0.6916	0.6320	0.6242	0.4149	0.6584	0.6706	162.02	101.47	102.74	103.25	117.78	135.40
	0.7	0.6345	0.6908	0.6234	0.6153	0.4240	0.6485	0.6630	171.89	101.78	103.11	102.83	116.30	134.17
	0.9	0.6302	0.6904	0.6179	0.6096	0.4278	0.6423	0.6584	178.46	101.99	103.37	102.70	115.51	133.51

8. Simulation Study

The simulation study have been carried out by considering $10,000$ different samples using Monte Carlo simulation for the data mentioned in section 7. The simulated percent relative efficiency E_{sji} of $\hat{\bar{Y}}_i$ with respect to $\hat{\bar{Y}}_{ij}$; $j = 1, 2, \ldots, 6$ and $i = 1$ and 2 respectively have been computed for many combinations of constants and the results are presented in Table 7.

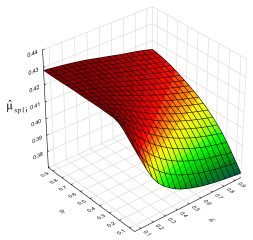


Figure 5: Optimum value of fraction of sample drawn afresh for estimator $\hat{\bar{Y}}_{14}$

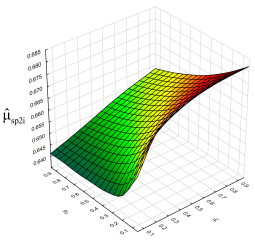


Figure 6: Optimum value of fraction of sample drawn afresh for estimator $\hat{\bar{Y}}_{15}$

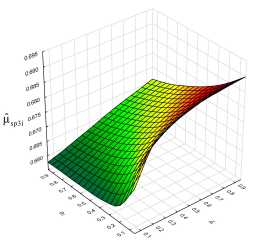


Figure 7: Optimum value of fraction of sample drawn afresh for estimator $\hat{\bar{Y}}_{16}$

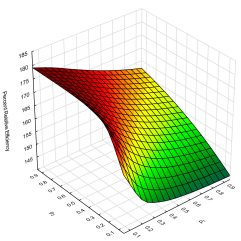


Figure 8: Percent Relative Efficiency E_{11}

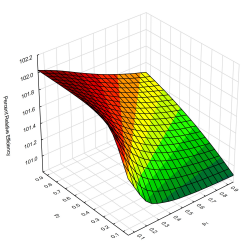


Figure 9: Percent Relative Efficiency E_{21}

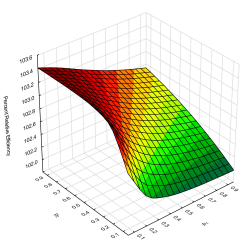


Figure 10: Percent Relative Efficiency E_{31}

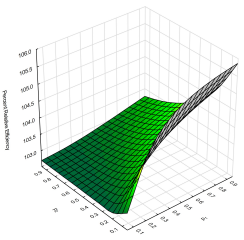


Figure 11: Percent Relative Efficiency E_{41}

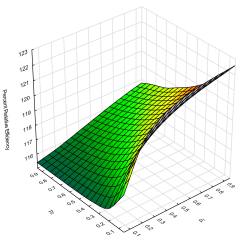


Figure 12: Percent Relative Efficiency E_{51}

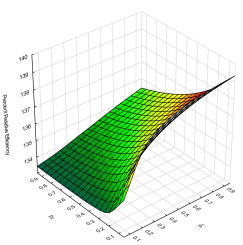


Figure 13: Percent Relative Efficiency E_{61}

Table 7. Simulation results for E_{sji} ; $j = 1, 2, \ldots, 6$;$i = 1$ and 2, $\alpha = \varphi_y^* = \varphi_x^*$

i	Model		E_{s1i}	E_{s2i}	E_{s3i}	E_{s4i}	E_{s5i}	E_{s6i}
1	M_{AR}							
	(p_2)	(p_1)						
	0.1	**0.1**	169.69	100.65	224.69	120.45	131.53	137.86
		0.5	186.63	100.91	196.39	119.78	128.99	135.80
		0.9	188.52	100.93	199.80	120.15	129.17	136.02
	0.3	**0.1**	155.46	100.47	274.59	121.17	134.37	140.23
		0.5	176.95	100.76	210.07	120.13	130.34	136.88
		0.9	183.39	100.85	205.21	120.23	129.75	136.46
	0.5	**0.1**	151.21	100.42	298.82	121.37	135.35	141.05
		0.5	169.71	100.65	227.63	120.60	131.67	137.99
		0.9	177.92	100.78	206.66	119.97	130.08	136.66
	0.7	**0.1**	149.21	100.39	313.36	121.47	135.84	141.47
		0.5	164.75	100.59	230.77	120.36	132.12	138.31
		0.9	173.36	100.71	214.69	120.16	130.79	137.24
	0.9	**0.1**	148.02	100.38	303.94	121.13	135.77	141.37
		0.5	161.38	100.54	258.97	121.24	133.46	139.50
		0.9	169.70	100.65	225.38	120.48	131.56	137.89
2	M_{DP}	φ_y^*						
		0.1	148.01	100.38	295.23	120.92	135.58	141.19
		0.5	169.68	100.66	221.400	120.27	131.36	137.70
		0.9	188.53	100.93	2010.6	120.25	129.27	136.11

9. Scrambling implementation Versus pseudonymous/incognito Questionnaires

As it is familiar that randomized and scrambled response estimators are less efficient than estimators obtained using direct questioning method. Here we discussed that the data is collected by scrambled response which is to be compared with data collected with pseudonymous/incognito questionnaire. For ascertaining the privacy protection additional cost has to be incurred.

In order to evaluate the data scrambling benefits, the estimator under randomized and scrambled response model have been compared with direct questioning method. If no scrambling mechanism have been used at any wave then the similar estimator under direct method is proposed as

$$L_D = \chi L_{uD} + (1-\chi) L_{mD} \; ; \; \chi \; \varepsilon \; [0, \; 1] \tag{26}$$

where

$$L_{uD} = V^* (\bar{y}_u, \; \bar{z}_u), \tag{27}$$

$$L_{mD} = T^* (\bar{y}_m, \; f_1^*, \; f_2^*) \tag{28}$$

where, $V^* (\bar{y}_u, \; \bar{z}_u)$ is a function of $(\bar{y}_u, \; \bar{z}_u)$ such that

$V^* (\bar{Y}, \; \bar{Z}) = \bar{Y} \Rightarrow V_1^*(K^*) = \frac{\partial V^*(.)}{\partial \bar{y}_u}|_{K^*} = 1$ with $K^* = (\bar{Y}, \; \bar{Z})$ and $V^*(\bar{y}_u, \; \bar{z}_u)$ satisfies the following conditions:

1. The function $V^* (\bar{y}_u, \; \bar{z}_u)$ is continuous and bounded in R.
2. The first, second and third partial derivatives of $V (\bar{y}_u, \; \bar{z}_u)$ exist and are continu-

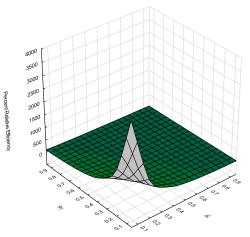


Figure 14: Percent Relative Efficiency E_{1D}

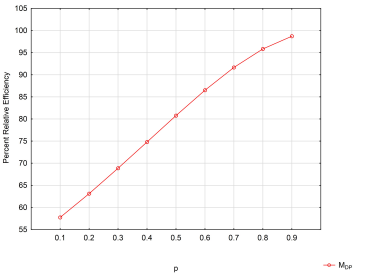


Figure 15: Percent Relative Efficiency E_{2D}

ous and bounded in R.
$T^*\ (\bar{y}_m,\ f_1^*,\ f_2^*)$ is a function of $(\bar{y}_m,\ f_1^*,\ f_2^*)$ such that $f_1^* = g^*\ (\bar{x}_m,\ \bar{z}_m,\ \bar{z}_n),\ f_2^* = h^*\ (\bar{x}_n,\ \bar{z}_n)$ and $T^*\ (\bar{y}_m,\ f_1^*,\ f_2^*)$ is a function of $(\bar{y}_m,\ f_1^*,\ f_2^*)$ such that $T^*\ (\bar{Y},\ \bar{X},\ \bar{X}) = \bar{Y},\ g^*\ (\bar{X},\ \bar{Z},\ \bar{Z}) = h^*(\bar{X},\ \bar{Z}) = \bar{X}$ and three functions $T^*,\ g^*,\ h^*$ satisfy the regularity conditions as considered for L_{ui} given in equation 10.
The minimum mean squared error of the class of estimator L_D to the first order approximations is given as

$$M[L_D]_{opt.^*} = \frac{B_{d1}^*\hat{\mu}_d + B_{d2}^*}{\hat{\mu}_d^2 A_{d3}^* - \hat{\mu}_d B_{d3}^* + A_{d1}^*}\left(\frac{S_y^2}{n}\right) \tag{29}$$

where,
$A_{d1}^* = 1-\rho_{xz}^2,\ A_{d2}^* = d_d + 1,\ A_{d3}^* = d_d - (H_{d2}^*)^2(F_{d2}^*)^2 + 2H_{d2}^*F_{d2}^*\rho_{yz},$
$d_d = (F_{d2}^*)^2 + (G_{d2}^*)^2(F_{d2}^*)^2 + 2F_{d2}^*\rho_{yx} + 2F_{d2}^*G_{d2}^*\rho_{yz} + 2(F_{d2}^*)^2G_{d2}^*\rho_{xz},\ B_{d1}^* = A_{d1}^*A_{d3}^*,$
$B_{d2}^* = A_{d1}^*A_{d2}^* - A_{d1}^*A_{d3}^*,\ B_{d3}^* = A_{d1}^* - A_{d2}^* + A_{d3}^*$ and

$\hat{\mu}_d$ satisfies

$$0 \ < \ min\left\{\frac{-C_{d2}^* + \sqrt{C_{d2}^{*2} + C_{d1}^*C_{d3}^*}}{C_{d1}^*},\ \frac{-C_{d2}^* - \sqrt{C_{d2}^{*2} + C_{d1}^*C_{d3}^*}}{C_{d1}^*}\right\} < 1 \tag{30}$$

where $C_{d1}^* = B_{d1}^*A_{d3}^*,\ C_{d2}^* = A_{d3}^*B_{d2}^*$ and $C_{d3}^* = A_{d1}^*B_{d1}^* + B_{d3}^*B_{d2}^*$.

$$E_{iD} \ = \frac{M(L_D)_{opt.^*}}{M(\hat{\bar{Y}}_i)_{opt.^*}} \times 100 \tag{31}$$

The percent relative efficiencies have been computed for the data represented in section 7 for different choices of $\{p_1,\ p_2,\ \varphi_y^*\}\varepsilon\{0.1,\ 0.2,\ 0.3,\ 0.4,\ \ldots,\ 0.9\}$ and are presented in graphical form in Figure 14 to Figure 15 for the two considered models respectively.

10. Demonstration of Results

1. From Figure.1 - Figure.13, following can be concluded.
(i) It can be seen that optimum value of fraction of sample to be drawn afresh exists for all considered estimators under both the randomized and scrambled response model.
(ii) The proposed general class of estimators performs appreciably good in terms of percent relative efficiency under the considered models when compared with other modified estimators L_{ij} ; $j = 1, 2, \ldots, 6$ which are also the members of proposed class of estimators L_i ; $i = 1$ and 2.
(iii) For M_{AR}, it can be seen that for fixed value of p_1 if p_2 increases E_{1j} decreases. However for fixed value of p_2 if p_1 increases E_{1j} also increases.
(iv) Both the models, M_{AR} and M_{DP} are performing almost similar in terms of percent relative efficiency.
(v) The Scrambled response model M_{DP} performs appreciably good in terms of optimum fraction of sample to be drawn afresh than the model M_{AR}.
(vi) The Randomized response model M_{AR} is more general as it provide wider scope to the respondents and moderate optimum fraction of sample to be drawn afresh as well as percent relative efficiency.
(vii) Out of scrambled and randomized response models, M_{DP} is showing stable behaviour as per assumptions of successive sampling.

2. From the simulation result in Table 7 it can be focused that the proposed general class of estimator is efficient than others considered under both randomized and scrambled response model.

3. From Figure 14 and Figure 15 it is indicated that when the proposed estimator is compared with direct method, for some combinations percent relative loss has been observed which is in accordance with the theory as scrambling or randomization procedures in general yields loss in efficiency.

11. Epilogue

The propounded general class of estimator for estimating sensitive population mean at current wave under considered scrambled and randomized response models accomplishes good percent relative efficiency when proposed general class of estimator L_i is compared with modified estimators L_{ij} ; $j = 1, 2, \ldots, 6$ and $i = 1$ and 2. Out of the two considered techniques, the model under scrambled response technique proves more stable in context of successive sampling with proposed estimator on two successive waves. However, depending on the sensitive nature of the character under study the two available techniques can be explored with the proposed general class of estimator. Therefore, depending on the given situation the scrambled or randomized response models may be selected to be applied with proposed general class of estimators on successive waves.

Acknowledgements

Authors are thankful to honourable reviewers for deeply reading the paper which lead to improvement over the earlier version of the paper. Authors are also thankful to SERB, New Delhi, India for providing the financial assistance to carry out the present work.

REFERENCES

ARNAB, R, SINGH, S., (2013). Estimation of mean of sensitive characteristics for successive sampling, Comm. Statist.-Theo. Meth., 42, pp. 2499–2524.

ARCOS, A., RUEDA M., SARJINDER SINGH, (2015). A generalized approach to randomized response for quantitative variables, Qual Quant, Vol. 49, Issue 3, pp 1239–1256.

BAR-LEV, S. K., BOBOVITCH, E., BOUKAI, B., (2004). A note on randomized response models for quantitative data. Metrica, 60, pp. 255–260.

COCHRAN, W. G., (1977). Sampling Techniques. Third edition. New York:Johan Wiley.

DIANA, G., PERRI, P. F., (2010). New Scrambled response models for estimating the mean of a sensitive quantitative character, J. App. Statis., 37 (11), pp. 1875–1890.

DIANA, G., PERRI, P. F., (2011). A class of estimators for quantitative sensitive data, Stat. Pap., 52, pp. 633–650.

EICHHORN, B. H., HAYRE, L. S., (1983). Scrambled randomized response method for obtaining sensitive quantitative data. J. Statist. Plann. Infer., 7, pp. 307–316.

FENG, S., ZOU, G., (1997). Sample rotation method with auxiliary variable, Comm. Stat.- Theo. Meth., 26, pp. 1497–1509.

GREENBERG, B. G., KUBLER, R. R.,ABERNATHY, J. R., (1971). Horvitz, D.G., Application of RR technique in obtaining quantitative data, J. Amer. Statist. Asso., 66, pp. 243–250.

JESSEN, R. J., (1942). Statistical investigation of a sample survey for obtaining farm facts, Iowa Agri. Exp. Stat. Road Bull., 304, pp. 1–104.

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

NAEEM, N., SHABBIR, J., (2016). Use of scrambled responses on two occasions successive sampling under non-response, Hacettepe University Bulletin of Natural Sciences and Engineering Series B: Mathematics and Statistics, 46. Available at http://www.hjms.hacettepe.edu.tr/uploads/32c18b80-275f-4b5a-a28d-6a5890eecac3.pdf.

PATTERSON, H. D., (1950). Sampling on successive occasions with partial replacement of units, J. Royal Statis. Soci. 12, pp. 241–255.

POLLOCK, K. H., BEK, Y., (1976). A comparison of three randomized response models for quantitative data. J. Amm. Sta. Asso., 71, pp. 884–886.

PRIYANKA K., MITTAL, R., (2014). Effective rotation patterns for median estimation in successive sampling. Statis. Trans., 15, pp. 197–220.

PRIYANKA, K., MITTAL R., MIN-KIM, J., (2015). Multiariate rotation design for population mean in sampling on successive occasions. Comm. Statis. Appli. Meth., 22, pp. 445–462.

PRIYANKA K., MITTAL, R., (2015a). Estimation of population median in two-occasion Rotation Sampling. J. Stat. App. Prob. Lett. 2, pp. 205–219.

PRIYANKA, K., MITTAL, R., (2015b). A class of estimators for population median in two occasion rotation sampling. HJMS, 44, pp. 189–202.

PRIYANKA, K., MITTAL, R., (2016). Search of Good Rotation Patterns on Successive Occasions and its Applications. UGC sponsored Project report, No.: [42–42 (2013)/SR].

REDDY, V. N., (1978). A Study on the use of prior knowledge on certain population parameters in estim ation. Sankhya C., 40, pp. 29–37.

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

SRIVASTAVA, S. K., (1980). A class of estimators using auxiliary information in sample surveys, Cand. Jour.Stat., (8), pp. 253–254.

SINGH, H. P., RUIZ-ESPEJO, M. R., (2003). On linear regression and ratio-product estimation of finite population mean. The Statistician., 52 (1), pp. 59–67.

SENAPATI, S.C., SAHOO, L. N., (2006). An alternative class of estimators in double sampling. Bull. Malays. Math. Sci. Soc., (2), 29 (1), (2006), 89–94.

SAHA, A., (2007). A simple randomized response technique in complex surveys. Metron, LXV, pp. 59–66.

SINGH, G. N., PRIYANKA, K., (2008). Search of good rotation patterns to improve the precision of estimates at current occasion, Comm. Stat. Theo. Meth., 37, pp. 337–348.

SINGH, G. N., KARNA, J. P., (2009). Estimation of population mean on current occasion in two occasion successive sampling. Metron, LXVII (1), pp. 87–103.

SINGH, G. N., PRASAD, S., (2010). Some estimates of population mean in two-occasion rotation patterns. A.M.S.E., 47 (2), 1–18.

SINGH, G. N., SUMAN, S., KHETAN, M., PAUL, C., (2017). Some estimation procedures of sensitive character using scrambled response techniques in successive sampling, Comm.Statist-Theory and Methods,
DOI: 10.1080/03610926.2017.1327073.

TRACY, D. S.,SINGH, H. P., SINGH, R., (1996). An alternative to the ratio-cum-product estimator in sample surveys, Journal of Statistical Planning and Inference (53), pp. 375–397.

WARNER, S. L., (1965). Randomized response: a survey technique for eliminating evasive answer bias, J. Amer. Statist. Asso., 60, pp. 63–69.

YU, B., JIN, Z., TIAN, J., GAO, G., (2015). Estimation of sensitive proportion by randomized response data in successive sampling, Compu. Mathemat. Meth. Med, DOI: 10.1155/2015/172918.

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 67–86, DOI 10.21307/stattrans-2019-004

NONRANDOMIZED RESPONSE MODEL FOR COMPLEX SURVEY DESIGNS

Raghunath Arnab[1], **Dahud Kehinde Shangodoyin**[2], **Antonio Arcos**[3]

ABSTRACT

Warner's randomized response (RR) model is used to collect sensitive information for a broad range of surveys, but it possesses several limitations such as lack of reproducibility, higher costs and it is not feasible for mail questionnaires. To overcome such difficulties, nonrandomized response (NRR) surveys have been proposed. The proposed NRR surveys are limited to simple random sampling with replacement (SRSWR) design. In this paper, NRR procedures are extended to complex survey designs in a unified setup, which is applicable to any sampling design and wider classes of estimators. Existing results for NRR can be derived from the proposed method as special cases.

Key words: complex survey designs, parallel model, randomized response, probability proportional to size, varying probability sampling.

Mathematics Subject Classification: 62D05

1. Introduction

In epidemiological, medical and sociological surveys among others, information is often collected on highly sensitive issues such as induced abortion, HIV/AIDS, drug addiction, domestic violence and cheating in examination, etc. In such situations, direct response (DR) surveys where sensitive questions are asked directly to the respondents, the respondents often provide wrong answers, or refuse to answer because of social stigma and/or fear. Under such circumstances the randomized response (RR) techniques may be used to collect more reliable data, protect respondents' confidentiality and avoid unacceptable rate of nonresponse. The RR technique was introduced by Warner (1965). Warner's technique was later modified by Horvitz et al. (1967), Greenberg et al. (1969), Raghavrao (1978), Franklin (1989), Arnab (1990, 1996), Kuk (1990) and Rueda et al. (2015) to increase co-operations from respondents and improve efficiencies of the proposed estimators. The applications of the RR technique to real life situations were reported by many researchers: Greenberg et al. (1969)

[1] University of Botswana, Botswana and University of KwaZulu-Natal, South Africa. E-mail: arnabr@mopipi.ub.bw. ORCID ID: https://orcid.org/0000-0001-5755-5857.
[2] University of Botswana, Botswana. E-mail: shangodoyink@mopipi.ub.bw. ORCID ID: https://orcid.org/0000-0002-0449-9510.
[3] University of Granada, Spain. Email: arcos@ugr.es.

with regard to illegitimacy of offspring; Abernathy et al. (1970) concerning incidence of induced abortions; Goodstadt and Gruson (1975) concerning drug uses, Folsom et al. (1973) concerning drinking and driving; and Arnab and Mothupi (2015) concerning sexual habits of University students.

In all randomised response models proposed in the literature, respondents have to perform randomized response experiments using devices such as spinners, the drawing of cards and the drawing of random numbers. So, in a survey involving RR methods, the investigators have to describe the methods and supply randomized devices to the respondents, which make the survey more expensive and time consuming rather than the direct response surveys. Tan et al. (2009) pointed out a few further limitations of RR which include (i) lack of reproducibility in the sense that the same respondent may provide different response depending on the outcome of the RR trial, (ii) uneven implementation of RR devices, which make it difficult to convince the respondents that their privacy is protected, (iii) some of the questions are alternative to sensitive questions (e.g. Warner (1965) model) and (iv) unfeasible for mail questionnaire. To overcome some of the aforementioned difficulties, nonrandomized response (NRR) model was proposed by Tian et al. (2007), Yu et al. (2008), Tan et al. (2009), Tian (2014) among others. In the proposed NRR models, independent non-sensitive questions were used to obtain indirect answers on sensitive issues. Obviously, NRR models reduce costs and are feasible for mail questionnaire. Tan et al. (2009) and Tian (2014) reported that the NRR model is more efficient than the RR model for estimating population characteristics. NRR techniques in real life surveys were used by Tang et al. (2014) to investigate homosexual experience among college students; Tian (2014) to investigate sexual behaviour and on plagiarism; and Wu and Tang (2016) to investigate pre-marital sex experience.

All the NRR models available in the literature are limited to simple random sampling with replacement (SRSWR) sampling design only. However, in practice most surveys are complex and multi-character surveys. A sampling design other than simple random sampling is called a complex sampling design. Complex sampling often involves clustering, stratification and unequal probability sampling among others, while in multi-character surveys information of more than one character is collected at a time. Some of the characters are of a confidential nature and others are not. For example, Household Income and Expenditure Survey 2002/03 (HIES 2002/03) conducted by CSO (2004), Botswana, involved a selection of first stage units by inclusion probability proportional to size (IPPS) sampling design, and the second stage units by a systematic sampling procedure. The same survey design was used by Statistics South Africa (2005) for HIES 2005/06 survey, Botswana Aids Impact Surveys (BAIS (2008)) conducted by CSO (2009) to collect data relating to sensitive issues such as sexual behaviour along with non-sensitive items such as socio-economic conditions.

In this paper, we have extended Tian (2014) NRR model called “The parallel model” for estimating population characteristics when the data is collected using complex survey designs. The estimator of the population proportion, its variance and unbiased estimators of variances of the estimators are derived in a unified setup, which is applicable to any sampling design and estimators. The estimators of the population proportions, their variances and unbiased estimators of the variances for the existing NRR models can be obtained from the proposed

method as special cases. It was found that under the SRSWR sampling, both the estimator and variance of the estimator of the population proportion π_y for the Greenberg et al. (1969) and Tian (2014) are the same. However, for the simple random sampling without replacement (SRSWOR) estimators of π_y are the same while the variance of Greenberg et al. (1969) estimator is higher than the Tian (2014) estimator. Two pioneering RR techniques are described below.

1.1. Warner's model

In Warner's (1965) pioneering method, a sample of size n was selected from a population by SRSWR method. Each of the respondents selected in the sample was asked to draw a card at random from a pack of well scaffolded cards consisting of two types of cards with known proportions and identical in appearance. Card type 1, with proportion $P_1(\neq 1/2)$ contains the question "Do you belong to the sensitive group A?" while card type 2 with proportion $1-P_1$ contains the question "Do you belong to group $\overline{A}$?" where A is a sensitive group such as HIV positive and $\overline{A}$ is the complement of group A (HIV negative). The respondent will supply a truthful answer "Yes" or "No" for the question mentioned in the selected card. The experiment is performed in the absence of the interviewer and hence the privacy of the respondent is maintained because the interviewer will not know which of the two questions the respondent has answered (See Arnab, 2017).

1.2. Greenberg et al. model

Greenberg et al. (1969) modified Warner's method by incorporating a sensitive question (character y) along with a non-sensitive question (character x). In this method, a sample of n units is selected by SRSWR method and each of the respondents selected in the sample has to pick a card at random (unobserved by the interviewer) from a pack containing two types of identical cards with known proportions as in Warner's model. The type 1 cards bear the sensitive question "Do you belong to the sensitive group A?" with proportion $P_2(\neq 0)$ while card type 2 (with proportion $1-P_2$) bears a question of unrelated or non-sensitive characteristic B such as "Are you an African?". Here also, the respondent will supply a truthful answer "Yes" or "No" for the question mentioned in the selected card (See Arnab, 2017).

2. Tian's NRR model

Tian (2014) proposed the following NRR model called "The parallel model", where the respondents need not require RR devices to provide responses. In this parallel model, respondents fill the questionnaire themselves unobserved by the interviewer. The questionnaire is a mixture of sensitive and non-sensitive questions. The parallel method is described below.

2.1. Parallel method

Let A denote the group of people who possess a sensitive characteristic y (such as HIV positive) and $\bar{A}$ denotes the people who do not possess the sensitive characteristic y (HIV negative). Further, let x and w be two non-sensitive dichotomous variates, such that y, x and w are mutually independent. For example, $x = 1(0)$ if the respondent's birthday 1 to 15 (16-31) days of a month while $w = 1(0)$ if the respondent is born between July and December (January to June) of a year. Clearly x and w are independent of the HIV infection status y such that $\pi_x = \Pr ob(x = 1) \cong 0.5$ and $1 - p = \Pr ob(w = 1) \cong 0.5$. Here a respondent has to answer truthfully "Yes" or "No" the unrelated non-sensitive question $Q1$ if his/her birthday falls in the first half of the year, i.e. $(w = 0)$ or a sensitive question $Q2$ if his/her birthday falls within the second half of the year, i.e. $(w = 1)$. The respondent should provide the answer "Yes" or "No" without disclosing which question he/she has answered. Hence, the confidentiality of the respondent is maintained.

For example, the questions $Q1$ and $Q2$ are as follows:

$Q1$: Are you a vegetarian?

$Q2$: Are you HIV positive?

2.2. Sampling design and methods of estimation

Tian (2014) used SRSWR method of sampling for the selection of a sample. Let n be the number of respondents selected and n' be the number of respondents who answered "Yes". Here, the probability of obtaining "Yes" answer from a respondent is

$$\theta_t = \Pr ob\{w = 0 \cap x = 1\} + \Pr ob\{w = 1 \cap y = 1\}$$

$$= (1 - p)\pi_x + p\pi_y \tag{2.1}$$

Noting that n' follows binomial distribution, Tian (2014) obtained an unbiased estimator of π_y as

$$\hat{\pi}_{ty} = \frac{\hat{\lambda}_t - \pi_x(1 - p)}{p} \tag{2.2}$$

where $\hat{\lambda}_t = n'/n$ = proportion of "Yes" answers.

The variance $\hat{\pi}_{ty}$ is given by

$$Var\left(\hat{\pi}_{ty}\right) = \frac{\pi_y(1 - \pi_y)}{n} + \frac{(1 - p)g(\pi_x \mid \pi_y, p)}{np^2} \tag{2.3}$$

where $g(\pi_x \mid \pi_y, p) = (p-1)\pi_x^2 + (1-2\pi_y p)\pi_x + \pi_y p$.

For $\pi_x = 1/2$, $g(\pi_x \mid \pi_y, p) = \frac{(p-1)}{4} + \frac{1}{2}$.

3. Parallel models for Complex survey designs

In this section we propose a methodology of estimating population proportion of a sensitive characteristic of a complex multi-character survey design where the data of the sensitive characteristic is collected by using the parallel method.

Consider a finite population $U = \{1,..,i,..,N\}$ of N units from which a sample s of size n units is selected with probability $p(s)$ using a sampling design $\mathcal{P}$. Let $\pi_i = \sum_{s \supset i} p(s)$ and $\pi_{ij} = \sum_{s \supset i,j} p(s)$ be the inclusion probabilities for the ith, and ith and jth $(i \neq j)$ units of the population. From each of the units in the sample s, the information on the sensitive characteristic y is obtained by using a parallel method. Let $B(\bar{B})$ be the group of respondents whose birthday falls between first half of a month i.e. 01 and 15 days (after 15th day of a month) of a month; $W(\bar{W})$ be the group of respondents born in the second half of the year, i.e. between July and December (January and June) and $A(\bar{A})$ be the group of respondents who do (do not) possess the sensitive characteristic y. Define

$$x_i = \begin{cases} 1 & \text{if the ith unit } \in B \\ 0 & \text{if the ith unit } \in \bar{B} \end{cases}, \quad w_i = \begin{cases} 1 & \text{if the unit i} \in W \\ 0 & \text{if the unit i} \in \bar{W} \end{cases}, \quad y_i = \begin{cases} 1 & \text{if the ith unit } \in A \\ 0 & \text{if the ith unit } \in \bar{A} \end{cases}$$

and $z_i = \begin{cases} 1 & \text{if the ith unit answers "}Yes\text{"} \\ 0 & \text{if the ith unit answers "}No\text{"} \end{cases}$

Under the parallel model, if a respondent belongs to the group $\bar{W}$, he/she answers the question $Q1$. In this case if the respondent's birthday falls between 01 and 15th day of a month, the respondent provides "Yes" answers with probability one. Otherwise if the respondent is born after 15th day of a month, the respondent supplies "No" answers with probability 1. Hence,

$$z_i = x_i \qquad \text{if } i \in \bar{W} \tag{3.1}$$

Similarly, if a respondent belongs to the group W, then the respondent answers the question $Q2$. In this case the respondent answers "Yes" with probability one if he/she belongs to the sensitive group A (HIV positive). On the other hand, if the respondent belongs to the complementary group $\bar{A}$ (HIV negative), then he/she supplies response answer "No" with probability one. Hence, in this cas

$$z_i = y_i \qquad i \in W \tag{3.2}$$

Equations (3.1) and (3.2) yield

$$z_i = w_i y_i + (1-w_i)x_i \tag{3.3}$$

and

$$\begin{aligned} Z &= \sum_{i\in U} z_i \\ &= \sum_{i\in \bar{W}} x_i + \sum_{i\in W} y_i \\ &= N_{\bar{W}B} + N_{WA} \end{aligned}$$

where $N_{\bar{W}B}\,(N_{WA})$ is the number of individuals of the population belonging to the groups $\bar{W}\cap B\ (W\cap A)$.

Assuming that the membership of an individual to the group $A(\bar{A})$, $W(\bar{W})$, and $B(\bar{B})$ is mutually independent, we make the following assumptions:

$$\pi_{WB} = p\pi_x;\ \pi_{\bar{W}B} = (1-p)\pi_x;\ \pi_{WA} = p\pi_y;\ N_{\bar{W}A} = (1-p)\pi_y \tag{3.4}$$

where

$$\pi_{WB} = \frac{N_{WB}}{N},\ \pi_{\bar{W}B} = \frac{N_{\bar{W}B}}{N},\ \pi_{WA} = \frac{N_{WA}}{N}, \pi_{\bar{W}A} = \frac{N_{\bar{W}A}}{N}\ , \pi_x = \frac{N_B}{N},\ \pi_y = \frac{N_A}{N}$$ and

$p = \frac{N_W}{N}$; N_F and N_{FG} denote the number of individuals belonging to the group F and $F\cap G$; $F, G = A, \bar{A}, B, \bar{B}, W, \bar{W}$.

Under the assumption (3.4), we have

$$\bar{Z} = Z/N = p\pi_y + (1-p)\pi_x \tag{3.5}$$

Here, we propose a linear homogeneous unbiased estimator of $\bar{Z}$ as

$$\hat{\bar{Z}} = \frac{1}{N}\sum_{i\in s} b_{si} z_i \tag{3.6}$$

Where $\sum_{i\in s}$ denotes the sum over distinct units in s, b_{si}'s are known constants satisfying the unbiasedness condition

$$\sum_{s\supset i} b_{si}\, p(s) = 1\,. \tag{3.7}$$

The variance of $\hat{\bar{Z}}$ is

$$V\left(\hat{\bar{Z}}\right) = V\left(\sum_{i\in s} b_{si} z_i\right)/N^2$$

$$= \left[E\left(\sum_{i \in s} b_{si} z_i \right)^2 - Z^2 \right] / N^2$$

$$= \frac{1}{N^2} E\left[\sum_s \left(\sum_{i \in s} b_{si}^2 z_i^2 + \sum_{i \neq} \sum_{j \in s} b_{si} b_{sj} z_i z_j \right) p(s) \right] - \bar{Z}^2$$

(where $p(s)$ is the probability of the selection of the sample s)

$$= \frac{1}{N^2} \left[\sum_{i \in U} \alpha_i z_i^2 + \sum_{i \neq} \sum_{j \in U} \alpha_{ij} z_i z_j \right] - \bar{Z}^2 \tag{3.8}$$

where

$$\alpha_i = \sum_{s \supset i} b_{si}^2 p(s) \text{ and } \alpha_{ij} = \sum_{s \supset i} b_{si} b_{sj} p(s).$$

The expression (3.8) yields

$$V\left(\hat{\bar{Z}}\right) = \sum_{i \in U} \alpha_i^* z_i^2 + \sum_{i \neq} \sum_{j \in U} \alpha_{ij}^* z_i z_j \tag{3.9}$$

where $\alpha_i^* = \frac{1}{N^2}(\alpha_i - 1)$ and $\alpha_{ij}^* = \frac{1}{N^2}(\alpha_{ij} - 1)$

From expression (3.9), we set an unbiased estimator of $V\left(\hat{\bar{Z}}\right)$ as

$$\hat{V}\left(\hat{\bar{Z}}\right) = \sum_{i \in s} c_{si} z_i^2 + \sum_{i \neq} \sum_{j \in s} c_{sj} z_i z_j \tag{3.10}$$

where c_{si} and c_{sij} are suitably chosen constants satisfying unbiasedness conditions

$$\sum_{s \supset i} c_{si} p(s) = \alpha_i^* \text{ and } \sum_{s \supset i} c_{sij} p(s) = \alpha_{ij}^* \tag{3.11}$$

We may choose c_{si} and c_{sij} in various ways. One of the obvious choices is $c_{si} = \alpha_i^* / \pi_i$ and $c_{sij} = \alpha_{ij}^* / \pi_{ij}$.

Substituting $z_i = w_i y_i + \bar{w}_i x_i$, $\bar{w}_i = 1 - w_i$ in equation (3.9) and noting that $w_i, \bar{w}_i, y_i$ and x_i are indicator variables, we have the following simplifications:

$$V\left(\hat{\bar{Z}}\right) = \sum_{i \in U} \alpha_i^* \left(w_i y_i + \bar{w}_i x_i \right) + \sum_{i \neq} \sum_{j \in U} \alpha_{ij}^* \left(w_i y_i + \bar{w}_i x_i \right) \left(w_j y_j + \bar{w}_j x_j \right)$$

$$= \left\{ \sum_{i \in W} \alpha_i^* y_i + \sum_{i \neq} \sum_{j \in W} \alpha_{ij}^* y_i y_j \right\} + \left\{ \sum_{i \in \bar{W}} \alpha_i^* x_i + \sum_{i \neq} \sum_{j \in \bar{W}} \alpha_{ij}^* x_i x_j \right\}$$

$$+\left\{\sum_{i\in W} y_i \sum_{j(\neq i)\in\bar{W}} \alpha_{ij}^* x_j + \sum_{j\in W} x_j \sum_{i(\neq j)\in\bar{W}} \alpha_{ij}^* y_j\right\}$$

$$=\left\{\sum_{i\in W\cap A} \alpha_i^* + \sum_{i\neq} \sum_{j\in W\cap A} \alpha_{ij}^*\right\}+\left\{\sum_{i\in \bar{W}\cap B} \alpha_i^* + \sum_{i\neq} \sum_{j\in \bar{W}\cap B} \alpha_{ij}^*\right\}$$

$$+\left\{\sum_{i\in W\cap A} \sum_{j(\neq i)\in \bar{W}\cap B} \alpha_{ij}^* + \sum_{j\in W\cap B} \sum_{i(\neq j)\in \bar{W}\cap A} \alpha_{ij}^*\right\}.$$

The above results lead to the following theorem.

Theorem 3.1.

Under assumptions (3.4),

(i) $\hat{\pi}_y = \dfrac{\hat{\bar{Z}} - (1-p)\pi_x}{p}$ is an unbiased estimator of π_y when the population proportion π_x is assumed to be known.

(ii) The variance of $\hat{\pi}_y$ is

$$V(\hat{\pi}_y) = \frac{1}{p^2}\left[\left\{\sum_{i\in W\cap A} \alpha_i^* + \sum_{i\neq} \sum_{j\in W\cap A} \alpha_{ij}^*\right\}+\left\{\sum_{i\in \bar{W}\cap B} \alpha_i^* + \sum_{i\neq} \sum_{j\in \bar{W}\cap B} \alpha_{ij}^*\right\}+\left\{\sum_{i\in W\cap A} \sum_{j(\neq i)\in \bar{W}\cap B} \alpha_{ij}^* + \sum_{j\in W\cap B} \sum_{i(\neq j)\in \bar{W}\cap A} \alpha_{ij}^*\right\}\right]$$

(iii) An unbiased estimator of $V(\hat{\pi}_y)$ is

$$\hat{V}(\hat{\pi}_y) = \frac{1}{p^2}\left[\sum_{i\in s} c_{si} z_i + \sum_{i\neq} \sum_{j\in s} c_{sij} z_i z_j\right].$$

We now present expressions of $\hat{\pi}_y$, $V(\hat{\pi}_y)$ and $\hat{V}(\hat{\pi}_y)$ for various sampling strategies as special cases of Theorem 3.1.

3.1. Arbitrary sampling design with Horvitz-Thompson estimator

For $b_{si} = 1/\pi_i$, we have $\alpha_i^* = \dfrac{1}{N^2}\left(\dfrac{1}{\pi_i} - 1\right)$, $\alpha_{ij}^* = \dfrac{1}{N^2}\left(\dfrac{\pi_{ij}}{\pi_i \pi_j} - 1\right)$ and the expression of the Horvitz- Thompson estimator for π_y as

$$\hat{\pi}_{hte} = \frac{\sum_{i\in s} \dfrac{z_i}{N\pi_i} - (1-p)\pi_x}{p} \tag{3.12}$$

The expression of the variance of and its unbiased estimators are obtained from the Theorem 3.1 as follows:

$$V\left(\hat{\pi}_{hte}\right)=\frac{1}{N^2p^2}\left[\begin{array}{l}\left\{\sum_{i\in W\cap A}\left(\frac{1}{\pi_i}-1\right)+\sum_{i\neq}\sum_{j\in W\cap A}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)\right\}+\left\{\sum_{i\in \bar{W}\cap B}\left(\frac{1}{\pi_i}-1\right)+\sum_{i\neq}\sum_{j\in \bar{W}\cap B}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)\right\}\\+\left\{\sum_{i\in W\cap A}\sum_{j(\neq i)\in \bar{W}\cap B}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)+\sum_{j\in W\cap B}\sum_{i(\neq j)\in \bar{W}\cap A}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)\right\}\end{array}\right] \tag{3.13}$$

and

$$\hat{V}\left(\hat{\pi}_{hte}\right)=\frac{1}{N^2p^2}\left[\sum_{i\in s}\frac{1}{\pi_i}\left(\frac{1}{\pi_i}-1\right)z_i+\sum_{i\neq j}\sum_{\in s}\frac{1}{\pi_{ij}}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)z_iz_j\right] \tag{3.14}$$

3.2. Simple random sampling without replacement (SRSWOR)

For SRSWOR, $\pi_i=n/N$, $\pi_{ij}=n(n-1)/\{N(N-1)\}$, $\alpha=\left(\frac{1}{\pi_i}-1\right)=\frac{N-n}{n}$ and $\beta=\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)=-\frac{N-n}{n(N-1)}$. In this case, the expressions $\hat{\pi}_{hte}$, $V\left(\hat{\pi}_{hte}\right)$ and $\hat{V}\left(\hat{\pi}_{hte}\right)$ come out as follows:

$$\hat{\pi}_{wor}=\frac{\bar{z}_s-(1-p)\pi_x}{p} \tag{3.15}$$

where $\bar{z}_s=\sum_{i\in s}z_i/n=\lambda_s$ = proportion of "Yes" answers in the sample s.

$$V(\hat{\pi}_{wor})=\frac{1}{N^2p^2}\left[\left\{\alpha\sum_{i\in W\cap A}+\beta\sum_{i\neq}\sum_{j\in W\cap A}\right\}+\left\{\alpha\sum_{i\in \bar{W}\cap B}+\beta\sum_{i\neq}\sum_{j\in \bar{W}\cap B}\right\}+\beta\left\{\sum_{i\in W\cap A}\sum_{j(\neq i)\in \bar{W}\cap B}+\sum_{j\in W\cap B}\sum_{i(\neq j)\in \bar{W}\cap A}\right\}\right]$$

$$=\frac{1}{N^2p^2}\left[\left\{N_{WA}\alpha+N_{WA}\left(N_{WA}-1\right)\beta\right\}+\left\{\alpha N_{\bar{W}B}+\beta N_{\bar{W}B}\left(N_{\bar{W}B}-1\right)\right\}+\beta\left\{N_{WA}N_{\bar{W}B}+N_{WB}N_{\bar{W}A}\right\}\right]$$

$$=\frac{N-n}{Nnp^2}\left[\left\{p\pi_y-p\pi_y\left(Np\pi_y-1\right)\frac{1}{(N-1)}\right\}+\left\{(1-p)\pi_x-(1-p)\pi_x\left(N(1-p)\pi_x-1\right)\frac{1}{(N-1)}\right\}\right]$$

$$-2\frac{N-n}{n(N-1)}\frac{(1-p)}{p}\pi_x\pi_y$$

$$=\frac{N(1-f)}{n(N-1)p^2}\left[\left\{p\pi_y(1-p\pi_y)\right\}+\left\{(1-p)\pi_x\left(1-(1-p)\pi_x\right)\right\}-2p(1-p)\pi_x\pi_y\right]$$

$$\left(\text{where } f=n/N\right)$$

$$= \frac{N(1-f)}{(N-1)}\left[\frac{\pi_y(1-\pi_y)}{n} + \frac{(1-p)}{np^2}\left\{(p-1)\pi_x^2 + \left(1-2p\pi_y\right)\pi_x + p\pi_y\right\}\right] \tag{3.16}$$

From the expression (3.15), we set an unbiased estimator of $V(\hat{\pi}_{wor})$ as

$$\hat{V}(\hat{\pi}_{wor}) = \frac{1}{p^2}\frac{(1-f)}{n}\frac{1}{n-1}\sum_{i\in s}(z_i - \bar{z}_s)^2$$

$$= \frac{1}{p^2}\frac{(1-f)}{n-1}\lambda_s(1-\lambda_s) \tag{3.17}$$

3.3. Probability proportional to size with replacement (PPSWR)

Let a sample of size n be selected from the population by PPSWR method using normed size measure $p_i\left(>0, \sum p_i = 1\right)$ attached to the ith unit. Let $z(r)$ be the response obtained from the respondent selected at the r th $(r = 1,..,n)$ draw with probability $p(r)$ so that $z(r) = z_j$ and $p(r) = p_j$ if rth draw produces the jth unit. The Hansen-Hurwitz estimator of the population proportion π_y is given by

$$\hat{\pi}_{hh} = \frac{\frac{1}{N}\left(\frac{1}{n}\sum_{r=1}^{n}\frac{z(r)}{p(r)}\right) - (1-p)\pi_x}{p} \tag{3.18}$$

Noting that $E\left\{\frac{z(r)}{p(r)}\right\} = \sum_{i=1}^{N} z_i = \sum_{i=1}^{N}\left\{w_i y_i + \bar{w}_i x_i\right\} = N\left\{p\pi_y + (1-p)\pi_x\right\}$, we find that $\hat{\pi}_{hh}$ is an unbiased estimator of π_y.

The variance of $\hat{\pi}_{hh}$ is

$$V(\hat{\pi}_{hh}) = \frac{1}{N^2p^2}V\left(\frac{1}{n}\sum_{r=1}^{n}\frac{z(r)}{p(r)}\right)$$

$$= \frac{1}{N^2p^2n}\left(\sum_{i=1}^{N}\frac{z_i^2}{p_i} - Z^2\right)$$

$$= \frac{1}{N^2p^2n}\left[\sum_{i=1}^{N}\frac{\left\{w_i y_i + \bar{w}_i x_i\right\}}{p_i} - N^2\left\{p\pi_y + (1-p)\pi_x\right\}^2\right] \tag{3.19}$$

Further noting that $\frac{z(r)}{p(r)}$ are independently distributed random variables, we find an unbiased estimator of $\hat{\pi}_{hh}$ as

$$\hat{V}(\hat{\pi}_{hh}) = \frac{1}{N^2p^2n(n-1)}\sum_{r=1}^{n}\left\{\frac{z(r)}{p(r)} - \frac{1}{n}\sum_{r=1}^{n}\frac{z(r)}{p(r)}\right\}^2 \tag{3.20}$$

3.4. Simple random sampling with replacement (SRSWR)

The PPSWR sampling scheme reduces to SRSWR sampling scheme if $p_i = 1/N$ for $i = 1,...,N$. Substituting $p_i = 1/N$ in the expressions (3.18), we find an unbiased estimator of π_y for SRSWR sampling method as

$$\hat{\pi}_{swr} = \frac{\lambda_s - (1-p)\pi_x}{p} \tag{3.21}$$

The expression of the variance of $\hat{\pi}_{swr}$ and its unbiased estimator come out as follows:

$$\begin{aligned} V(\hat{\pi}_{swr}) &= \frac{1}{p^2 n}\left[\frac{1}{N}\sum_{i=1}^{N}\{w_i y_i + \bar{w}_i x_i\} - \{p\pi_y + (1-p)\pi_x\}^2\right] \\ &= \frac{1}{p^2 n}\left[\{p\pi_y + (1-p)\pi_x\} - \{p\pi_y + (1-p)\pi_x\}^2\right] \\ &= \frac{\pi_y(1-\pi_y)}{n} + \frac{(1-p)}{np^2}\left[(p-1)\pi_x^2 + (1-2p\pi_y)\pi_x + p\pi_y\right] \end{aligned} \tag{3.22}$$

and

$$\hat{V}(\hat{\pi}_{swr}) = \frac{\lambda_s(1-\lambda_s)}{p^2(n-1)} \tag{3.23}$$

Expressions (3.21), (3.22) and (3.23) are the same as those obtained by Tian (2014).

3.5. Stratified multi-stage sampling design

Consider a population comprising of H strata. The $h(=1,..,H)$th stratum consists of M_h first-stage units (fsus) and the ith fsu of the hth stratum consists of $M_{hi}\,(i = 1,..,M_h)$ second-stage units (ssus). The total number of ssus in the population is $\sum_{h=1}^{H}\sum_{i=1}^{Mh} M_{hi} = M$. From the h th stratum, a sample s_h of size n_h fsus is selected by using a suitable sampling scheme with $\pi_{i|h}$ and $\pi_{ij|h}$ as inclusion probabilities for the i th, and i th and $j\,(j \neq i)$th fsus. If the ith fsu is selected in the sample s_h, a sub-sample s_{hi} of size n_{hi} ssus is selected from it by using a suitable sampling scheme with inclusion probabilities $\pi_{k|hi}$ and $\pi_{kl|hi}$ for the k th, and k and $l(l \neq k)$th ssus. We denote the j th ssu of the i th fsu of the h th stratum as hij th unit. We define the following notations similar to the Section 3.

$$x_{hij} = \begin{cases} 1 & \text{if } hij\text{th unit} \in B \\ 0 & \text{if } hij\text{th unit} \in \bar{B} \end{cases}, \quad w_{hij} = \begin{cases} 1 & \text{if } hij\text{th unit} \in W \\ 0 & \text{if } hij\text{th unit} \in \bar{W} \end{cases}, \quad y_{hij} = \begin{cases} 1 & \text{if } hij\text{th unit} \in A \\ 0 & \text{if } hij\text{th unit} \in \bar{A} \end{cases},$$

$$z_{hij} = \begin{cases} 1 & \text{if } hij\text{th unit answers "}Yes\text{"} \\ 0 & \text{if } hij\text{th unit answers "}No\text{"} \end{cases}.$$

Now, writing $z_{hij} = w_{hij}y_{hij} + (1- w_{hij})x_{hij}$ and using the assumption similar to (3.4), we find that

$$Z = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{J=1}^{N_{hi}} z_{ijk} = M[p\pi_y + (1-p)\pi_x] \quad (3.24)$$

Further, noting that $\hat{Z}_{hte} = \sum_{h=1}^{H}\sum_{i\epsilon s_h}\frac{\hat{Z}_{i|h}}{\pi_{i|h}}$ with $\hat{Z}_{i|h} = \sum_{j\epsilon s_{hi}}\frac{z_{hij}}{\pi_{j|hi}}$ is an unbiased estimator of Z, we get the following theorem.

Theorem 3.2.

(i) $\hat{\pi}_y = \frac{1}{p}\left[\frac{\hat{Z}_{hte}}{M} - (1-p)\pi_x\right]$ is an unbiased estimator of π_y.

(ii) The variance of $\hat{\pi}_y$ is

$$V(\hat{\pi}_y) = \frac{1}{p^2M^2}\sum_{h=1}^{H}\left[\sum_{i\neq}^{M_h}\sum_{j=1}^{M_h}(\pi_{i|h}\pi_{j|h} - \pi_{ij|h})\left(\frac{Z_{i|h}}{\pi_{i|h}} - \frac{Z_{j|h}}{\pi_{j|h}}\right)^2 + \sum_{i=1}^{M_h}\frac{\sigma_{i|h}^2}{\pi_{i|h}}\right]$$

where

$Z_{i|h} = \sum_{j=1}^{M_{hi}} z_{hij}$ and $\sigma_{i|h}^2 = V(Z_{i|h}) = \sum_{k\neq}^{M_{hi}}\sum_{l=1}^{M_{hi}}(\pi_{k|hi}\pi_{l|hi} - \pi_{kl|hi})\left(\frac{Z_{hik}}{\pi_{k|hi}} - \frac{Z_{hil}}{\pi_{l|hi}}\right)^2$

(iii) An unbiased estimator of $V(\hat{\pi}_y)$ is

$$\hat{V}(\hat{\pi}_y) = \frac{1}{p^2M^2}\sum_{h=1}^{H}\left[\sum_{i\neq}\sum_{j\in s_h}\left(\frac{\pi_{i|h}\pi_{j|h} - \pi_{ij|h}}{\pi_{ij|h}}\right)\left(\frac{\hat{Z}_{i|h}}{\pi_{i|h}} - \frac{\hat{Z}_{j|h}}{\pi_{j|h}}\right)^2 + \sum_{i\in s_h}^{M_h}\frac{\hat{\sigma}_{i|h}^2}{\pi_{i|h}}\right]$$

where

$$\hat{\sigma}_{i|h}^2 = \sum_{k\neq}\sum_{l\in s_{hi}}\frac{(\pi_{k|hi}\pi_{l|hi} - \pi_{kl|hi})}{\pi_{kl|hi}}\left(\frac{z_{hik}}{\pi_{k|hi}} - \frac{z_{hil}}{\pi_{l|hi}}\right)^2$$

is an unbiased estimator of $\sigma_{i|h}^2$.

Proof:

(i) $E(\hat{\pi}_y) = \frac{1}{p}\left[\frac{E(\hat{Z}_{hte})}{M} - (1-p)\pi_x\right]$

$= \frac{1}{p}\left[\frac{Z}{M} - (1-p)\pi_x\right]$

$= \pi_y$

(ii) $V(\hat{\pi}_y) = \frac{1}{M^2p^2}\sum_{h=1}^{H} V(\hat{Z}_h)$

$$= \frac{1}{M^2p^2}\sum_{h=1}^{H}\left[V\{E(\hat{Z}_h|s_h)\} + E\{V(\hat{Z}_h|s_h)\}\right]$$

$$= \frac{1}{M^2p^2}\sum_{h=1}^{H}\left[V\left\{\sum_{i\in s_h}\frac{Z_{i|h}}{\pi_{i|h}}\right\} + E\left\{\sum_{i\in s_h}\frac{\sigma_{i|h}^2}{\pi_{i|h}^2}\right\}\right]$$

$$= \frac{1}{M^2p^2}\sum_{h=1}^{H}\left[\sum_{i\neq}^{M_h}\sum_{j=1}^{M_h}(\pi_{i|h}\pi_{j|h} - \pi_{ij|h})\left(\frac{Z_{i|h}}{\pi_{i|h}} - \frac{Z_{j|h}}{\pi_{j|h}}\right)^2 + \sum_{i=1}^{M_h}\frac{\sigma_{i|h}^2}{\pi_{i|h}}\right]$$

(iii) $E[\hat{V}(\hat{\pi}_y)]$

$$= \frac{1}{M^2p^2}\sum_{h=1}^{H} E\left[\sum_{i\neq}\sum_{j\in s_h}\left(\frac{\pi_{i|h}\pi_{j|h} - \pi_{ij|h}}{\pi_{ij|h}}\right)E\left\{\left(\frac{\hat{Z}_{i|h}}{\pi_{i|h}} - \frac{\hat{Z}_{j|h}}{\pi_{j|h}}\right)^2|s_h\right\} + E\left(\sum_{i\in s_h}\frac{\hat{\sigma}_{i|h}^2}{\pi_{i|h}}|s_h\right)\right]$$

$$= \frac{1}{M^2p^2}\sum_{h=1}^{H} E\left[\sum_{i\neq}\sum_{j\in s_h}\left(\frac{\pi_{i|h}\pi_{j|h} - \pi_{ij|h}}{\pi_{ij|h}}\right)\left\{\left(\frac{Z_{i|h}}{\pi_{i|h}} - \frac{Z_{j|h}}{\pi_{j|h}}\right)^2 + \frac{\sigma_{i|h}^2}{\pi_{i|h}^2} + \frac{\sigma_{j|h}^2}{\pi_{j|h}^2}\right\} + \sum_{i\in s_h}\frac{\sigma_{i|h}^2}{\pi_{i|h}}\right]$$

$$= \frac{1}{M^2p^2}\sum_{h=1}^{H}\left[\sum_{i\neq}^{M_h}\sum_{j=1}^{M_h}(\pi_{i|h}\pi_{j|h} - \pi_{ij|h})\left\{\left(\frac{Z_{i|h}}{\pi_{i|h}} - \frac{Z_{j|h}}{\pi_{j|h}}\right)^2 + \frac{\sigma_{i|h}^2}{\pi_{i|h}^2} + \frac{\sigma_{j|h}^2}{\pi_{j|h}^2}\right\} + \sum_{i=1}^{M_h}\sigma_{i|h}^2\right]$$

Now, noting that $\sum_{i=1}^{M_h}\pi_{i|h} = n_h$ and $\sum_{j(\neq i)=1}^{M_h}\pi_{ij|h} = (n_h - 1)\pi_{i|h}$, we find $E[\hat{V}(\hat{\pi}_y)] = V(\hat{\pi}_y)$.

4. Comparison with Greenberg RR model

Consider the Greenberg et al. (1969) model described in Section 1.2 with $P_2 = p$. Let $y_i = 1(0)$ if the ith unit does (does not) belong to the sensitive group A, $x_i = 1(0)$ if the ith unit possesses (does not possess) the non-sensitive characteristic B and $z_i = 1(0)$ if the ith respondent answers "Yes" ("No"). Denoting

$E_R(V_R)$ as expectation (variance) with respect to the RR model and noting x_i and y_i are indicator variables, one finds that

$$E_R(z_i) = py_i + (1-p)x_i = E_R(z_i^2) \tag{4.1}$$

$$V_R(z_i) = py_i + (1-p)x_i - \{py_i + (1-p)x_i\}^2$$
$$= p(1-p)(x_i + y_i - 2x_iy_i) \tag{4.2}$$

Let a sample s of size n be selected from the population using SRSWR method, $\lambda_s = \frac{1}{n}\sum_{i\in s} z_i$ be the proportion of "Yes" answers in the population and $\sum_{i\in s}$ denote the sum over the units in s with repetition. In this case we have the following theorem:

Theorem 4.1.

Under SRSWR sampling

(i) $\hat{\pi}_G = \frac{1}{p}[\lambda_s - (1-p)\pi_x]$ is an unbiased estimator of π_y when π_x is known.

(ii) The variance of $\hat{\pi}_G$ is

$$V(\hat{\pi}_G) = \frac{\pi_y(1-\pi_y)}{n} + \frac{1-p}{p^2n}\left[(p-1)\pi_x^2 + (1-2p\pi_y)\pi_x + p\pi_y\right]$$

(iii) An unbiased estimator of $V(\hat{\pi}_G)$ is

$$\hat{V}(\hat{\pi}_G) = \frac{1}{p^2n}\left[\frac{1}{n-1}\sum_{i\in s}(z_i - \lambda_s)^2\right] = \frac{\lambda_s(1-\lambda_s)}{(n-1)p^2}$$

Proof:

(i) $E(\hat{\pi}_G) = \frac{1}{p}[E(\bar{z}) - (1-p)\pi_x]$

$$= \frac{1}{p}\left[E_p\left\{\frac{1}{n}\sum_{i\in s} E_R(z_i)\right\} - (1-p)\pi_x\right]$$

$$= \frac{1}{p}\left[\frac{1}{N}\sum_{i\in U}\{py_i + (1-p)x_i\} - (1-p)\pi_x\right]$$

$$= \pi_y$$

(ii) $V(\hat{\pi}_G) = V_p[E_R(\hat{\pi}_G)] + E_p[V_R(\hat{\pi}_G)]$

$$= V_p\left[\frac{1}{np}\sum_{i\in s} E_R(z_i) - \frac{(1-p)}{p}\pi_x\right] + E_p\left[\frac{1}{(np)^2}\sum_{i\in s} V_R(z_i)\right]$$

$$= V_p\left[\frac{1}{np}\sum_{i\in s}\{py_i + (1-p)x_i\}\right] + E_p\left[\frac{1-p}{n^2p}\sum_{i\in s}(x_i + y_i - 2x_iy_i)\right]$$

$$= \frac{1}{np^2}\left[\frac{1}{N}\sum_{i\in U}\{py_i + (1-p)x_i\}^2 - \{p\pi_y + (1-p)\pi_x\}^2\right] + \frac{1-p}{npN}\sum_{i\in U}(x_i + y_i - 2x_iy_i)$$

$$= \frac{1}{np^2}\left[p^2\pi_y(1-\pi_y) + (1-p)^2\pi_x(1-\pi_x) + 2p(1-p)(\pi_{xy} - \pi_x\pi_y)\right] + \frac{1-p}{np}(\pi_x + \pi_y - 2\pi_{xy})$$

Noting that $\pi_{xy} = \pi_x\pi_y$, as x and y are independent, we obtain

$$V(\hat{\pi}_G) = \frac{\pi_y(1-\pi_y)}{n} + \frac{1-p}{p^2n}\left[(p-1)\pi_x^2 + (1-2p\pi_y)\pi_x + p\pi_y\right]$$

(iii) Further, z_i's, $i = 1,2,..,n$ are independent and identically distributed random variables, one finds that $E[\hat{V}(\hat{\pi}_G)] = E[V(\hat{\pi}_G)]$.

Here, we note that for the SRSWR sampling, the expressions $\hat{\pi}_G$ and $\hat{V}(\hat{\pi}_G)$ of the Greenberg et al. (1969) model are respectively the same as the expressions $\hat{\pi}_{swr}$ (Eq. 3.21) and $V(\hat{\pi}_{swr})$ (Eq. 3.22) in the Parallel model proposed by Tian (2014).

Consider the situation where a sample s of size n is selected by the SRSWOR method and from each of the selected respondents randomized responses were obtained by using Greenberg et al. (1969) RR technique. Let $\lambda_s = \bar{z}_s = \sum_{i\in s} z_i/n$ denote the proportion of "Yes" answers in the sample. In this case we have the following results:

Theorem 4.2.

Under SRSWOR sampling,

(i) $\hat{\pi}_G^* = \frac{1}{p}[\lambda_s - (1-p)\pi_x]$ is an unbiased estimator of π_y.

(ii) The variance of $\hat{\pi}_G^*$ is

$$V(\hat{\pi}_G^*) = \frac{N-n}{(N-1)n}\left[\pi_y(1-\pi_y) + \frac{1-p}{p^2}\pi_x(1-x)\right] + \frac{1-p}{np}(\pi_x + \pi_y - 2\pi_x\pi_y)$$

(iii) An unbiased estimator of $V(\hat{\pi}_G^*)$ is

$$\hat{V}(\hat{\pi}_G^*) = \frac{N-n}{p^2Nn}\frac{1}{n-1}\sum_{i\in s}(z_i-\bar{z})^2 + \frac{1-p}{p}(\hat{\pi}_G + \pi_x - 2\pi_x\hat{\pi}_G)$$

$$= \frac{N-n}{p^2N}\frac{\lambda_s(1-\lambda_s)}{(n-1)} + \frac{1-p}{p}(\hat{\pi}_G + \pi_x - 2\pi_x\hat{\pi}_G)$$

Proof:

(i) $E(\hat{\pi}_G^*) = \frac{1}{p}[E(\lambda_s) - (1-p)\,\pi_x\,]$

$$= \frac{1}{p}\left[E_p\left\{\frac{1}{n}\sum_{i\in s}E_R(z_i)\right\} - (1-p)\pi_x\right]$$

$$= \frac{1}{p}\left[\frac{1}{N}\sum_{i\in U}\{py_i + (1-p)x_i\} - (1-p)\pi_x\right]$$

$$= \pi_y$$

(ii) $V(\hat{\pi}_G^*) = V_p[E_R(\hat{\pi}_G^*)] + E_p[V_R(\hat{\pi}_G^*)]$

$$= V_p\left[\frac{1}{np}\sum_{i\in s}E_R(z_i) - \frac{(1-p)}{p}\pi_x\right] + E_p\left[\frac{1}{(np)^2}\sum_{i\in s}V_R(z_i)\right]$$

$$= \frac{N-n}{np^2}\left[\frac{1}{N}\sum_{i\in U}\{py_i + (1-p)x_i\}^2 - \left\{p\pi_y + (1-p)\pi_x\right\}^2\right]$$

$$+\frac{1-p}{npN}\sum_{i\in U}(x_i + y_i - 2x_iy_i)$$

$$= \frac{N-n}{np^2}\left[p^2\pi_y(1-\pi_y) + (1-p)^2\pi_x(1-\pi_x) + 2p(1-p)(\pi_{xy} - \pi_x\pi_y)\right]$$

$$+\frac{1-p}{np}(\pi_x + \pi_y - 2\pi_{xy})$$

Now, noting that, $\pi_{xy} = \pi_x\pi_y$ we find that

$$V(\hat{\pi}_G^*) = \frac{N-n}{(N-1)n}\left[\pi_y(1-\pi_y) + \frac{1-p}{p^2}\pi_x(1-\pi_x)\right] + \frac{1-p}{np}(\pi_x + \pi_y - 2\pi_x\pi_y)$$

(iii) $E\left[\hat{V}(\hat{\pi}_G^*)\right] = \frac{N-n}{p^2Nn}\frac{1}{n-1}E_p\left[\sum_{i\in S}E_R(z_i^2) - \frac{\sum_{i\in S}E_R(z_i^2) + \sum_{i\neq}\sum_{j\in S}E_R(z_i)E_R(z_j)}{n}\right]$

$$+\frac{1-p}{p}(\pi_x + \pi_y - 2\pi_x\pi_y)$$

$$= \frac{N-n}{p^2Nn}\left[\sum_{i\in U}\{E_R(z_i)\}^2 + \sum_{i\in U} V_R(z_i) - \frac{1}{N}\sum_{i\neq}\sum_{j\in U} E_R(z_i)\,E_R(z_j)\right]$$

$$+ \frac{1-p}{p}(\pi_x + \pi_y - 2\pi_x\pi_y)$$

$$= V(\hat{\pi}_G^*)$$

From the expressions of $V(\hat{\pi}_G^*)$ and (3.16), we find that

$$V(\hat{\pi}_G^*) - V(\hat{\pi}_{wor}) = \frac{n-1}{N-1}\frac{1-p}{np}\left[\pi_x(1-\pi_y) + \pi_y(1-\pi_x)\right]$$

$$\geq 0 \qquad (4.3)$$

From the Eq. (4.3), we conclude for the SRSWOR sampling, Tian's (2014) estimator $\hat{\pi}_{wor}$ based on NRR method is more efficient than the Greenberg et al.'s (1969) estimator $\hat{\pi}_G^*$ based on RR technique for estimating the population proportion π_y. However, for large N, both are equally efficient. The percentage relative efficiency of $\hat{\pi}_{wor}$ with respect to $\hat{\pi}_G^*$ under SRSWOR sampling assuming $\frac{N-1}{N} \cong 1$ is given by

$$\frac{V(\hat{\pi}_G^*)}{V(\hat{\pi}_{wor})} \times 100$$

$$= \frac{(1-f)\left[\pi_y(1-\pi_y) + \frac{1-p}{p^2}\pi_x(1-\pi_x)\right] + \frac{1-p}{p}(\pi_x + \pi_y - 2\pi_{xy})}{(1-f)\left[\pi_y(1-\pi_y) + \frac{1-p}{p^2}\{(p-1)\pi_x^2 + (1-2p\pi_y)\pi_x + p\pi_y\}\right]} \times 100 \qquad (4.4)$$

The percentage relative efficiency (E) for different values of π_x, π_y, p and f is given in the Table 4.1. For the given values of π_x,π_y, the efficiency increases with p until $p = 0.50$, then it decreases. Efficiency increases with the increase in the sampling fraction f. The maximum efficiency 148.6 is attained when

f = 0.40, π_x= 0.10, π_y = 0.75 and p= 0.40.

Table 4.1. Efficiency of $\hat{\pi}_G^*$ with respect to $\hat{\pi}_{wor}$

		f = 0.1						f = 0.2				
π_y	π_x	p						p				
		0.1	0.25	0.4	0.5	0.75		0.1	0.25	0.4	0.5	0.75
0.10	0.10	102.0	104.2	105.3	105.6	104.2		104.5	109.4	112	112.5	109.4
	0.25	101.7	103.7	105.2	105.8	105.3		103.8	108.4	111.7	113.0	111.9
	0.40	101.8	104	105.6	106.2	106.1		104.1	109.0	112.5	114.0	113.6
	0.50	102.0	104.3	105.9	106.6	106.5		104.5	109.8	113.4	114.9	114.6
	0.75	103.2	106.0	107.5	108.0	107.5		107.3	113.5	116.8	117.9	116.9

Table 4.1. Efficiency of $\hat{\pi}_G^*$ with respect to $\hat{\pi}_{wor}$ (cont.)

π_y	π_x	f = 0.1						f = 0.2				
		p						p				
		0.1	0.25	0.4	0.5	0.75		0.1	0.25	0.4	0.5	0.75
0.25	0.10	102.9	105.3	106.0	105.8	103.7		106.6	111.9	113.4	113.0	108.4
	0.25	102.0	104.2	105.3	105.6	104.2		104.5	109.4	112,0	112.5	109.4
	0.40	101.9	104.1	105.3	105.7	104.6		104.3	109.1	112.0	112.8	110.3
	0.50	102.0	104.2	105.6	105.9	104.8		104.5	109.5	112.5	113.3	110.9
	0.75	103.0	105.6	106.7	106.9	105.6		106.7	112.5	115.2	115.6	112.5
0.40	0.10	103.7	106.1	106.5	106.2	104.0		108.4	113.6	114.7	114.0	109.0
	0.25	102.3	104.6	105.6	105.7	104.1		105.2	110.3	112.6	112.8	109.1
	0.40	102.0	104.2	105.3	105.6	104.2		104.5	109.4	112.0	112.5	109.4
	0.50	102.0	104.2	105.4	105.6	104.3		104.5	109.4	112.1	112.6	109.6
	0.75	102.7	105.1	106.2	106.3	104.6		106.1	111.5	113.9	114.1	110.3
0.50	0.10	104.2	106.5	106.9	106.6	104.3		109.3	114.6	115.6	114.9	109.8
	0.25	102.5	104.8	105.9	105.9	104.2		105.6	110.9	113.2	113.3	109.5
	0.40	102.1	104.3	105.4	105.6	104.2		104.7	109.6	112.2	112.6	109.4
	0.50	102.0	104.2	105.3	105.6	104.2		104.5	109.4	112.0	112.5	109.4
	0.75	102.5	104.8	105.9	105.9	104.2		105.6	110.9	113.2	113.3	109.5
0.75	0.10	105.1	107.5	108.1	108.0	106.0		111.4	116.9	118.2	117.9	113.5
	0.25	103.0	105.6	106.7	106.9	105.6		106.7	112.5	115.2	115.6	112.5
	0.40	102.2	104.6	105.9	106.3	105.1		105.0	110.3	113.3	114.1	111.5
	0.50	102.0	104.2	105.6	105.9	104.8		104.5	109.5	112.5	113.3	110.9
	0.75	102.0	104.2	105.3	105.6	104.2		104.5	109.4	112.0	112.5	109.4
		f = 0.3						f = 0.4				
0.10	0.10	107.7	116.1	120.6	121.4	116.1		112.0	125.0	132.0	133.3	125.0
	0.25	106.4	114.4	120.1	122.3	120.3		110.0	122.4	131.2	134.6	131.6
	0.40	106.9	115.4	121.4	124.0	123.4		110.8	123.9	133.3	137.3	136.4
	0.50	107.8	116.7	122.9	125.5	125.1		112.1	126.0	135.7	139.7	139.1
	0.75	112.5	123.2	128.8	130.7	129.1		119.5	136.1	144.8	147.7	145.2
0.25	0.10	111.4	120.3	123.0	122.3	114.4		117.7	131.6	135.7	134.6	122.4
	0.25	107.7	116.1	120.6	121.4	116.1		112.0	125.0	132.0	133.3	125.0
	0.40	107.3	115.6	120.6	122.0	117.7		111.4	124.3	132.1	134.2	127.5
	0.50	107.7	116.3	121.4	122.9	118.7		112.0	125.4	133.3	135.6	129.1
	0.75	111.5	121.4	126.0	126.8	121.4		117.9	133.3	140.4	141.7	133.3
0.40	0.10	114.3	123.4	125.2	124.0	115.4		122.3	136.4	139.2	137.3	123.9
	0.25	108.9	117.7	121.6	122.0	115.6		113.9	127.5	133.7	134.2	124.3
	0.40	107.7	116.1	120.6	121.4	116.1		112.0	125.0	132.0	133.3	125.0
	0.50	107.7	116.1	120.7	121.6	116.4		112.0	125.1	132.2	133.7	125.6
	0.75	110.4	119.8	123.8	124.1	117.7		116.2	130.7	137.0	137.5	127.5
0.50	0.10	116.0	125.1	126.7	125.5	116.7		124.9	139.1	141.6	139.7	126.0
	0.25	109.7	118.7	122.6	122.9	116.3		115.0	129.1	135.2	135.6	125.4
	0.40	108.0	116.4	120.9	121.6	116.1		112.4	125.6	132.5	133.7	125.1
	0.50	107.7	116.1	120.6	121.4	116.1		112.0	125.0	132.0	133.3	125.0
	0.75	109.7	118.7	122.6	122.9	116.3		115.0	129.1	135.2	135.6	125.4
0.75	0.10	119.6	129.1	131.3	130.7	123.2		130.5	145.2	148.6	147.7	136.1
	0.25	111.5	121.4	126.0	126.8	121.4		117.9	133.3	140.4	141.7	133.3
	0.40	108.6	117.7	122.8	124.1	119.8		113.4	127.5	135.4	137.5	130.7
	0.50	107.7	116.3	121.4	122.9	118.7		112.0	125.4	133.3	135.6	129.1
	0.75	107.7	116.1	120.6	121.4	116.1		112.0	125.0	132.0	133.3	125.0

5. Conclusion

The Randomized Response technique was introduced by Warner (1965) to collect data on sensitive characteristics. In this technique, the respondents have to perform randomized response experiments using devices which make the survey more expensive and time-consuming than the direct response surveys. Apart from these limitations, the procedure may yield different response depending on the outcome of the RR trial and it is unfeasible for mail questionnaire. To overcome some of the aforementioned difficulties, nonrandomized response (NRR) model was proposed by Tian et al. (2007), Yu et al. (2008), Tan et al. (2009), Tian (2014), among others. All the proposed procedures are limited to SRSWR sampling design and are unusable in real life complex multi-character surveys. In this paper, NRR models have been extended to complex surveys in a unified setup, which is applicable to any sampling design and estimators. The estimators of the population proportions, their variances and unbiased estimators of the variances for the existing NRR models can be obtained from the proposed method as special cases. It has been found for the SRSWR sampling, expressions of the estimators of the population proportion π_y, its variance for the Greenberg et al. (1969) and Tian (2014) are the same. However, for the SRSWOR sampling, the variance of Tian (2014) estimator is smaller than that of the Greenberg et al. (1969) estimator. But for large population they are equal.

Acknowledgements

The authors are grateful to the anonymous reviewers, whose thoughtful suggestions led to the substantial improvement of the earlier version of the manuscript.

REFERENCES

ABERNATHY, J. R., GREENBERG, B. G., HORVITZ D. G., (1970). Estimates of induced abortion in urban North Carolina, Demography, 7, pp. 19–29.

ARNAB, R., (1990). On commutativity of design and model expectations in randomized response surveys. Communications in Statistics, Theory & Methods, pp. 3751–2757.

ARNAB, R., (1996). Randomized response trials: a unified approach for qualitative data, Commun. Statist. Theory & Methods 25 (6), p. 1173.

ARNAB, R., (2017). Survey Sampling Theory and Applications. Academic Press, Oxford.

ARNAB, R., MOTHUPI, T., (2015). Randomized response techniques: A case study of the risky behaviors' of students of a certain University, Model Assisted Statistics and Applications, 10, pp. 421–430.

CENTRAL STATISTICAL OFFICE, (2004). Household Income and Expenditure Survey 2002/03, Republic of Botswana.

CENTRAL STATISTICS OFFICE, (2009). Botswana Aids Impact Survey III (2008), Statistical Report.

FOLSOM, S. A., (1973). The two alternative questions randomized response model for human surveys. J. Amer. Statist. Assoc., 68, pp. 525-530.

FRANKLIN, L. A., (1989). A comparison of estimators for randomized response sampling with continuous distribution from dichotomous populations. Commun. Statist. Theory and methods 18, pp. 489–505.

GOODSTADT, M. S., GRUSON, V., (1975). The randomized response technique; a test on drug use. J. Amer. Statist. Assoc., 70, pp. 814–818

GREENBERG, B. G., ABUL-ELA, A. L. A., SIMMONS, W. R., HORVITZ, D. G., (1969). The unrelated question randomized response model: Theoretical framework. J. Amer. Statist. Assoc. 64, pp. 520–539

HORVITZ, D. G., SHAH, B. V., SIMMONS, W. R., (1967). The unrelated question randomized response model. Proceedings of Social Statistical section, Amer. Statist. Assoc. pp. 65–72.

KUK, A. Y., (1990). Asking sensitive question indirectly. Biometrika 77, 436-438.

RAGHAVRAO, D., (1978). On estimation problem in Warner's randomized response techniques. Biometrics 34, pp. 87–90.

RUEDA, M., COBO, B., ARCOS, A., (2015). Package 'RRTCS': Randomized Response Techniques for Complex Surveys, http://cran.r-project.org/web/packages/RRTCS/.

STATISTICS SOUTH AFRICA, (2005). Income and Expenditure of households 2005/2006, Republic of South Africa.

TAN, G.L., YU, J. W., TANG, M. L., (2009). Sample survey with sensitive questions: a non-randomized response approach, The American Statistician, 63, pp. 9–16.

TANG, M., WU, Q., TIAN, G., GUO, J., (2014). Two-sample Non Randomized Response Techniques for Sensitive Questions. Commun. Statist. Theory & Methods, 43, pp. 408–425.

TIAN, G. L., YU, J. W., TANG, M. L., GENG, Z., (2007). A new non-randomized model for analysing sensitive question with binary outcomes. Statistics in Medicine, 26, pp. 4238–4252.

TIAN, G. L., (2014). A new non-randomized response model: the parallel model. Statistica Neerlandica, 68, pp. 293–323.

WARNER, S. L., (1965). Randomize response: a survey technique for eliminating evasive answer bias. J. Amer. Statist. Assoc. 60, pp. 63–69.

WU, Q., TANG, M., (2016). Non-randomized response model for sensitive survey with noncompliance. Statistical Methods in Medical Research, 25, pp. 2827–2839.

YU, J. W., TIAN, G. L., TANG, M. L., (2008). Two new models for survey sampling with sensitive characteristics: Design and Analysis. Metrika, 67, pp. 251–263.

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 87–102, DOI 10.21307/stattrans-2019-005

ON THE SMOOTHED PARAMETRIC ESTIMATION OF MIXING PROPORTION UNDER FIXED DESIGN REGRESSION MODEL

Ramakrishnaiah Y. S.[1], **Manish Trivedi**[2], **Konda Satish**[3]

ABSTRACT

The present paper revisits an estimator proposed by Boes (1966) – James (1978), herein called BJ estimator, which was constructed for estimating mixing proportion in a mixed model based on independent and identically distributed (i.i.d.) random samples, and also proposes a completely new (smoothed) estimator for mixing proportion based on independent and not identically distributed (non-i.i.d.) random samples. The proposed estimator is nonparametric in true sense based on known "kernel function" as described in the introduction. We investigated the following results of the smoothed estimator under the non-i.i.d. set-up such as (a) its small sample behaviour is compared with the unsmoothed version (BJ estimator) based on their mean square errors by using Monte-Carlo simulation, and established the percentage gain in precision of smoothed estimator over its unsmoothed version measured in terms of their mean square error, (b) its large sample properties such as almost surely (a.s.) convergence and asymptotic normality of these estimators are established in the present work. These results are completely new in the literature not only under the case of i.i.d., but also generalises to non-i.i.d. set-up.

Key words: mixture of distributions, mixing proportion, smoothed parametric estimation, fixed design regression model, mean square error, optimal band width, strong consistency, asymptotic normality.

1. Introduction

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and not identically distributed (non-i.i.d.) random variables with continuous distribution functions (d.f.s) $\{F_i(x), 1 \le i \le n\}$. Let H(x) be a continuous cumulative distribution function (cdf) of mixture of component cdfs $H_1(x), \ldots, H_m(x)$, ($m \ge 2$) such that $H(x) = \sum_{j=1}^{m} p_j H_j(x)$, where $\{p_j;\ 1 \le j \le m\}$ is a set of mixing proportions satisfying (i) $0 < p_j < 1$, (ii) $\sum_{j=1}^{m} p_j = 1$. Let $\bar{H}_n(x) = n^{-1} \sum_{i=1}^{n} F_i(x) \rightarrow H(x)$ as $n \rightarrow \infty$ and $\bar{H}_j(x) = n_j^{-1} \sum_{i=1}^{n_j} F_{ji}(x) \rightarrow H_j(x)$,

[1] Faculty of Statistics, Osmania University in Hyderabad. E-mail: ysrkou@gmail.com.
[2] Faculty of Statistics, School of Sciences, Indira Gandhi National Open University in New Delhi. E-mail: Manish_trivedi@ignou.ac.in. ORCID ID: https://orcid.org/0000-0002-5790-6546.
[3] Faculty of Statistics, Aurora College in Hyderabad. E-mail: statishlaks@gmail.com. ORCID ID: https://orcid.org/0000-0003-4119-8773.

$n_j \to \infty$, j=1,2,…,m; H(x), H_j(x) are known d.f.s. The problem of estimation of mixing proportions p_j in a mixture

$$\mathbf{H(x)} = \boldsymbol{p_1 H_1}(x) + \boldsymbol{p_2 H_2}(x) + \ldots + \boldsymbol{p_m H_m}(x) \tag{1.1}$$

of m known distributions H_j(x) is investigated based on independent random samples of sizes n, n_j generated from the fixed design regression models

$$X_i = \beta t_i + \epsilon_i \text{ , } 1 \le i \le n, \ \epsilon_i \sim \text{i.i.d.F(x)} \tag{1.2}$$

$$X_{ji} = \beta_j t_{ji} + \epsilon_{ji}, \ 1 \le i \le n_j, \ j = 1, 2, \ldots, m, \ \epsilon_{ij} \sim \text{i.i.d. } F_j(x), \tag{1.3}$$

β's and t's are known reals satisfying the model conditions

$$\beta_j > 0, \ \sum_{i=1}^{n} t_i = 0 \text{ and } \frac{1}{n}\sum t_i^2 = o(n^{-1}). \tag{1.4}$$

Note that $t_i = \mp \frac{i}{n^{\delta}}$, i =∓1, ∓2,…, ∓n, $\delta \ge \frac{3}{2}$ fulfill (1.4) and $\bar{H}_n(x)=n^{-1}\sum_{i=1}^{n} F_i(x)$ = F(x) +O($\frac{1}{n}\sum t_i^2$) + … = H(x) + o(n^{-1}) and $\bar{H}_{n_j}(x)=n_j^{-1}\sum_{i=1}^{n_j} F_{ji}(x) \to H_j(x)$, j=1,2 as $n_j \to \infty$.

Mixture distributions have been used in a wide variety of numerous applications in such diversified fields as physics, chemistry, biology, social sciences and others. Many typical problems in which such mixtures occur have been well described in a series of research papers. Karl Pearson (1894) dealt with the application of normal mixtures to the theory of evolution, which considered the first paper in the mixtures of distributions. Acheson and McElwee (1951), who identified failures in an electronic tube in gaseous defects, mechanical defects and normal deterioration of the cathode. One can find the proportion of the population which will fail in each cause to redesign the system or to improve the methods of manufacturing process. Apart from this, it would be desirable to know the distribution of defectives for each cause. Mendenhall and Hader (1958) studied censored life testing as a mixed failure populations. They suggested an example that the engineer may identify the product as defective/failure and nondefective by two or more different types of causes. Hosmer (1973) studied characteristics such as sex, age, and length of halibut (fish). Odell and Basu (1976) applied them in the field of remote sensing to estimate the crop acreages from remote sensors on orbiting satellites.

We shall show some of the typical problems which were described in Choi and Bulgren (1968):

1. In fishery biology, it is often desired to measure certain characteristics in a natural population of fish. For this purpose samples of fish are taken and the desired trait is measured for each fish in the sample. However, many characteristics vary markedly with the age of the fish. Then, the trait has a distinct distribution for each age group so that the population has a mixture of distributions.
2. A geneticist analyses the inheritance of qualitative characters. In general, such characters vary continuously over some intervals of real numbers so that a given genotype may be able to produce phenotypic values over an interval of real numbers. Then, the phenotypic value which the geneticist observes has a mixture of distributions, each of which is given by a genotype.

3. In photographing the absorption spectrum of an ionized atom, we obtain a photograph of a constantly varying intensity distribution on the photographic plate, and not a series of discrete "lines". This phenomenon is caused by several effects (such as the Doppler effect) and it is accepted in spectroscopy that an intensity distribution whose graph can be approximated closely by that of a normal density function belongs to every theoretical "line". Then the graph of the whole spectrum section can be considered as a mixture of normal density functions.

Other references to mixed failure populations are given in papers by Davis (1952), Epstein (1953), Herd (1953), Steen and Wilde (1952), Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), Lindsay (1995) and McLachlan and Peel (2000), Fu (1968-Pattern Recognition), Varli et al. (1975-Pattern Recognition), Clark (1976-Geology), Macdonald and Pitcher (1979-Fisheries), Bruni et al. (1985-Genetics), Merz (1980-Physics) and Christensen et al. (1980-Nuclear Physics).

The applications of finite mixture distributions describing mixture populations for non- i.i.d. sequence of variables are given below:

Area	**Characteristic X_i**	**Distribution function $F_i(x)$**
Survival Analysis	life time of components produced by i^{th} machine operated by i^{th} foreman.	Life time distribution of various products
Nutritional Studies	weight for age/height for age/ weight for height of i^{th} infant of i^{th} origin or group.	Distribution of weight for age/height of i^{th} infant
Fisheries	Fish length or weight of age of i^{th} fish.	Distribution of weight/length of i^{th} fish

The mixing model with two component populations becomes

$$\mathbf{H(x) = p}\boldsymbol{H_1}\mathbf{(x) + (1-p)}\boldsymbol{H_2}\mathbf{(x).}$$

Here, X_i is the characteristic with distribution function $F_i(x)$ assuming

$$\bar{F}_n(x) = n^{-1}\sum_{i=1}^{n} F_i(x) = \bar{H}_n(x) \rightarrow H(x)$$

$$\bar{F}_{jn}(x) = n^{-1}\sum_{i=1}^{n} F_{ji}(x) = \bar{H}_{jn}(x) \rightarrow H_j(x) \text{ as } n\rightarrow\infty$$

and n, n_1 and n_2 are independent random sample sizes from mixed and component populations selected in such a way that $n = n_1 + n_2$ with $n_1 = [\rho n]$, $n_2 = [(1-\rho)n]$, $0< \rho <1$.

More details on such examples can be found in Choi and Bulgreen (1968), Harris (1958), Blischke (1965), Fu (1968-Pattern Recognition), Varli et al. (1975-Pattern Recognition), Clark (1976-Geology), Macdonald and Pitcher (1979-Fisheries), Odell and Basu (1976-Remote sensing), Bruni et al. (1985-Genetics), Merz (1980-Physics) and Christensen et al. (1980-Nuclear Physics), etc.

i.i.d case: The mixing model for i.i.d. case is

$$F(x) = pG_1(x)+(1-p)G_2(x) \tag{1.5}$$

where F(x), $G_j(x)$; j=1,2 are cdfs of mixed and component populations respectively. The following estimator is studied in the literature.

Boes (1966)-James (1978) (BJ) estimator: Let $F_n(x)= n^{-1} \sum_{i=1}^{n} I(X_i \leq x)$ be the empirical distribution function of a random sample X_i, $1 \leq i \leq n$ from a mixture of two known component distribution functions G_j, j=1,2. Boes (1966) proposed an estimator of p given by

$$\boldsymbol{p_{n,1}}(x) = \frac{\widetilde{F}_n(x) - G_2(x)}{G_1(x) - G_2(x)} \tag{1.6}$$

and shown as a minimax unbiased estimator, and derived the Cramer-Rao lower bound. James (1978) considered the problem of estimating the mixing proportion in a mixture of two known normal distributions. He studied the simple estimators based on (a) the number of observations lower than a fixed point r, (b) the numbers lower than s and greater than t, and (c) the sample mean. Van Houwelingen (1974) used Boes (1966) estimator to estimate the mixing proportion by using frequency densities and obtained the Cramer-Rao lower bound. Jayalakshmi (2002) used BJ estimator using kernel-based empirical distribution and established that smoothing improves efficiency when the component distributions are known.

In the present work, we extend the idea of estimation of mixing proportion p in two directions:

- The estimators based on kernel based empirical d.f. called smoothed estimations are proposed under regression models (1.2) - (1.3).
- The proposed parametric estimators are based on independent, but not identically distributed (non-i.i.d.) samples generated by the fixed design regression models described by (1.2)-(1.3).

The main object of the present paper is to confine attention to m=2 case in the model (1.1) and to construct parametric estimators when component distributions are known, based on the usual empirical and kernel-based distribution functions defined by

$$\widetilde{H}_n(x) = n^{-1} \sum_{i=1}^{n} I(X_i \leq x),\ \widehat{H}_n(x)= n^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{a_n}\right), \tag{1.7}$$

$\{a_n\}$ being the smoothing sequence satisfying $0< a_n \to 0$, $na_n \to \infty$ defined by

$$p_{n,1}(x) = \frac{\widetilde{H}_n(x) - H_2(x)}{H_1(x) - H_2(x)},\quad p_{n,2}(x) = \frac{\widehat{H}_n(x) - H_2(x)}{H_1(x) - H_2(x)}. \tag{1.8}$$

We study the small and large sample behaviour of the proposed parametric estimators and establish the superiority of smoothed estimator $p_{n,2}(x)$ over unsmoothed one $p_{n,1}(x)$ in the sense of minimum mean square error. The results of the present investigations for the non-i.i.d. sequences are completely new in the literature.

In section 2, we obtain the exact expressions for MSEs of the proposed estimators in order to establish the superiority of $p_{n,2}(x)$ over $p_{n,1}(x)$ for some fixed x. Furthermore, large sample behaviour, such as asymptotic normality and rates

of a.s. convergence of the proposed estimators is also established. In section 3, the crucial choice of smoothing parameter 'a_n' in kernel based estimator $p_{n,2}$(x) is discussed and its value is determined by employing minimum mean square criterion. Section 4 deals with establishing superiority of $p_{n,2}$(x) over $p_{n,1}$(x). Section 5 explains the small sample comparisons by Monte Carlo method based on the samples generated by regression models.

2. Asymptotics of $p_{n,1}$(x) and $p_{n,2}$(x)

We now present the properties of both estimators $p_{n,1}$(x) and $p_{n,2}$(x) under the fixed design regression model (1.3). The properties such as mean square errors (MSEs), rates of a.s. convergence, and asymptotic normality of $p_{n,1}$(x) and $p_{n,2}$(x) are established. We first consider the representations of the proposed estimators: from the mixing model,

$$\mathrm{H(x)} = \mathrm{p}H_1(\mathrm{x}) + (1\text{-}\mathrm{p})H_2(\mathrm{x}) \tag{2.1}$$

H(x) – H_2(x) = d_{12}(x)p for each x, where d_{12}(x)= H_1(x) - H_2(x)

A.S. Representations to $p_{n,1}$(x), $p_{n,2}$(x): from (1.8) and (2.1),

$$\mathrm{d_{12}(x)}[p_{n,1}(\mathrm{x})\text{-}\mathrm{p}] = \widetilde{\mathrm{H}}_{\mathrm{n}}(\mathrm{x}) - \mathrm{H(x)} = n^{-1}\sum_{i=1}^{n}(\mathrm{I}(X_i \leq \mathrm{x}) - F_i(\mathrm{x}) + F_i(\mathrm{x}) - \mathrm{H(x)})$$

$$=: n^{-1}\sum_{i=1}^{n} Z_{1i}(\mathrm{x}) + \tau_{n1}(\mathrm{x}) \tag{2.2}$$

where by (1.4) and by the assumption on regression model

$$\tau_{n1}(\mathrm{x}) = n^{-1}\sum_{i=1}^{n} F_i(\mathrm{x}) - \mathrm{H(x)} = \mathrm{o}(\mathrm{n}^{-1}\textstyle\sum_1^n t_i^2) = \mathrm{o}(\mathrm{n}^{-1}) \tag{2.3}$$

Similarly, from (1.8) and (2.1)

$$\mathrm{d_{12}(x)}[\mathrm{p}_{n,2}(\mathrm{x})\text{-}\mathrm{p}] = \widehat{\mathrm{H}}_{\mathrm{n}}(\mathrm{x}) - \mathrm{H(x)} = \widehat{\mathrm{H}}_{\mathrm{n}}(\mathrm{x}) - \mathrm{E}\,\widehat{\mathrm{H}}_{\mathrm{n}}(\mathrm{x}) + \mathrm{E}\,\widehat{\mathrm{H}}_{\mathrm{n}}(\mathrm{x}) - \mathrm{H(x)}$$

$$=: \mathrm{n}^{-1}\sum_{i=1}^{n}[K(\tfrac{\mathrm{x}-\mathrm{X_i}}{\mathrm{a_n}}) - \mathrm{E}\,K(\tfrac{\mathrm{x}-\mathrm{X_i}}{\mathrm{a_n}})] + \tau_{n2}(\mathrm{x})$$

$$= n^{-1}\sum_{i=1}^{n} Z_{2i}(\mathrm{x}) + \tau_{n2}(\mathrm{x}) \tag{2.4}$$

with τ_{n2}(x) = o(n^{-1}) as in (2.3).

2.1. Mean Square Error of $p_{n,1}$(x) and $p_{n,2}$(x)

We first consider the small sample property, i.e. MSEs of both estimators $p_{n,j}$(x), j=1,2 for each x in the following, which will help in establishing the

superiority of smoothed estimator over unsmoothed version based on a random sample of exact size n under the regression model.

Theorem 2.1. Let $\{X_i$: 1≤i≤n$\}$ be a sequence of non-i.i.d. random variables with corresponding sequence of uniformly continuous distribution functions $\{F_i$(x): 1≤i≤n$\}$. If $\{F_i\}$, the kernel density function k and $\{a_n\}$ in (1.7) satisfy

AI: i) F_i(x) is uniformly continuous distribution function with finite q[th] derivatives $F_i^{(q)}$(x) < ∞, 1 ≤ i ≤ n and

$\bar{H}_n^{(q)}$(x) = $\frac{1}{n}$ $\sum F_i^{(q)}$(x), q = 2,4,6

ii) $\bar{H}_n$(x) = $n^{-1}\sum_{i=1}^{n} F_i$(x) → H(x) as n→∞

AII: i) The kernel function satisfies μ_{2j}(K) = $\int_{-\infty}^{\infty} t^{2j}$ dK(t) ≠ 0 and

μ_j(K) = $\int_{-\infty}^{\infty} t^j$ dK(t) = 0 for j=1,3,…

ii) ψ_j(K) = 2 $\int_{-\infty}^{\infty} t^j$ K(t) dK(t) < ∞ , j = 0, 1, 2, 3, 4

AIII: $\{a_n\}$ is a sequence of bandwidths such that

i) $0 < a_n \downarrow 0$; $na_n \to \infty$ as $n \to \infty$

ii) $na_n^4 \to 0$ as $n \to \infty$

then

$$\textbf{MSE } [p_{n,2}(x)] = \textbf{MSE}(p_{n,1}(x)) - d_{12}^{-2}(x)\left[\frac{a_n}{n}\bar{H}_n^{(1)}(x)\psi_1(K) - \frac{a_n^4}{4}\bar{H}_n^{(2)^2}(x)\mu_2^2(K)\right] + O\left(\frac{a_n^2}{n}\right) + o(a_n^4)$$

Proof: From (2.4),

$$d_{12}^2(x)\,nMSE[p_{n,2}(x)] = Var(n^{-1/2}\sum_{i=1}^{n} Z_{2i}(x)) + n\tau_{n2}^2 \qquad (2.5)$$

where σ_{n2}^2 = $Var(n^{-1/2}\sum_{i=1}^{n} Z_{2i}(x))$

= $n^{-1}\sum_{i=1}^{n} E\, Z_{2i}^2(x)$

= $n^{-1}\sum_{i=1}^{n} \sigma_{2i}^2$

σ_{2i}^2 = $E\,K^2\left(\frac{x-X_i}{a_n}\right)$ - $E^2\,K\left(\frac{x-X_i}{a_n}\right)$

= I_{1i} - I_{2i}^2 (say)

where I_{1i} = $E\,K^2\left(\frac{x-X_i}{a_n}\right) = \int K^2\left(\frac{x-u}{a_n}\right) dF_i(u) = \int F_i(x-a_nt)\, dK^2(t)$

= $F_i(x)\int dK^2(t) - F_i^{(1)}(x)a_n\int t\,dK^2(t) + \frac{a_n^2}{2!}F_i^{(2)}(x)\int t^2 dK^2(t) - \frac{a_n^3}{3!}F_i^{(3)}(x)\int t^3 dK^2(t)$

$+ \frac{a_n^4}{4!}F_i^{(4)}(x)\int t^4 dK^2(t) + o(a_n^4)$

$$= F_i(x)\,\psi_0(K) - a_n F_i^{(1)}(x)\,\psi_1(K) + \frac{a_n^2}{2!} F_i^{(2)}(x)\,\psi_2(K) - \frac{a_n^3}{3!} F_i^{(3)}(x)\,\psi_3(K)$$

$$+ \frac{a_n^4}{4!} F_i^{(4)}(x)\,\psi_4(K) + o(a_n^4) \qquad (2.6)$$

where $\psi_0(K) = 2\int_{-\infty}^{\infty} K(t)dK(t) = 2\int_0^1 y\,dy = 1$, while

$$I_{2i} = E\,K\left(\frac{x - X_i}{a_n}\right) = \int F_i(x - a_n t)dK(t)$$

$$= F_i(x) + \frac{a_n^2}{2!} F_i^{(2)}(x) \int_{-\infty}^{\infty} t^2 dK(t) + \frac{a_n^4}{4!} F_i^{(4)}(x) \int_{-\infty}^{\infty} t^4 dK(t) + o(a_n^4)$$

$$=: F_i(x) + \frac{a_n^2}{2} F_i^{(2)}(x)\mu_2(K) + \frac{a_n^4}{4!} F_i^{(4)}(x)\,\mu_4(K) + o(a_n^4) \qquad (2.7)$$

From (2.5) and (2.7),

$$\sigma_{n2}^2 = n^{-1} \sum_{i=1}^n \{[F_i(x) - a_n F_i^{(1)}(x)\psi_1(K) + \frac{a_n^2}{2} F_i^{(2)}(x)\psi_2(K) - \frac{a_n^3}{3!} F_i^{(3)}(x)\psi_3(K)$$

$$+ \frac{a_n^4}{4!} F_i^{(4)}(x)\psi_4(K) + o(a_n^4)]$$

$$- (F_i(x) + \frac{a_n^2}{2} F_i^{(2)}(x)\mu_2(K) + \frac{a_n^4}{4!} F_i^{(4)}(x)\,\mu_4(K) + o(a_n^4))^2\}$$

$$= n^{-1} \sum_{i=1}^n F_i(x)(1 - F_i(x)) - a_n \bar{H}_n^{(1)}(x)\psi_1(K) + \frac{a_n^2}{2} \bar{H}_n^{(2)}(x)\psi_2(K)$$

$$- \frac{a_n^3}{6} \bar{H}_n^{(3)}(x)\,\psi_3(K) - a_n^2\, n^{-1} \sum_{i=1}^n F_i(x) F_i^{(2)}(x)\mu_2(K)$$

$$- \frac{a_n^4}{12} n^{-1} \sum_{i=1}^n F_i(x) F_i^{(4)}(x)\mu_4(K) + o(\frac{a_n^4}{n})$$

$$= n^{-1} \sum_{i=1}^n F_i(x)(1 - F_i(x)) - a_n \bar{H}_n^{(1)}(x)\psi_1(K)$$

$$+ a_n^2\, (\frac{1}{2} \bar{H}_n^{(2)}(x)\psi_2(K) - n^{-1} \sum_{i=1}^n F_i(x) F_i^{(2)}(x)\mu_2(K)) + O(a_n^3) \qquad (2.8)$$

and from (2.7),

$$\tau_{n2}^2 = (n^{-1} \sum_{i=1}^n E\,K\left(\frac{x - X_i}{a_n}\right) - H(x))^2$$

$$= [\bar{H}_n(x) - H(x) + \frac{a_n^2}{2} \bar{H}_n^{(2)}(x)\mu_2(K) + \frac{a_n^4}{4!} \bar{H}_n^{(4)}(x)\mu_4(K) + o(a_n^4)]^2$$

$$=: [\xi_{n,0}(x) + a_n^2\, \xi_{n,2}(x) + a_n^4 \xi_{n,4}(x) + o(a_n^4)]^2$$

$$= \xi_{n,0}(x)[\xi_{n,0}(x) + 2a_n^2 \xi_{n,2}(x) + 2a_n^4 \xi_{n,4}(x) + o(a_n^4)] + a_n^4 \xi_{n,2}^2(x) + o(a_n^4) \qquad (2.9)$$

$$= a_n^4 \xi_{n,2}^2(x) + O(\frac{a_n^2}{n})$$

in view of $\xi_{n,0}(x) = o(n^{-1})$, where $\xi_{n,2}(x) = \frac{1}{2} \bar{H}_n^{(2)}(x)\mu_2(K)$.

Thus, from (2.5), (2.8) and (2.9),

$$MSE(p_{n,2}(x)) = d_{12}^{-2}(x) n^{-1} [n^{-1} \sum_{i=1}^n F_i(x)(1 - F_i(x)) - a_n \bar{H}_n^{(1)}(x)\psi_1(K)$$

$$+ a_n^2 (\frac{1}{2} \bar{H}_n^{(2)}(x)\psi_2(K) - n^{-1} \sum_{i=1}^n F_i(x) F_i^{(2)}(x)\mu_2(K))]$$

$$+ d_{12}^{-2}(x)\xi_{n,0}(x)[\xi_{n,0}(x) + 2a_n^2\xi_{n,2}(x) + 2a_n^4\xi_{n,4}(x)]$$

$$+ a_n^4\ \xi_{n,2}^2(x) + o(\tfrac{a_n^4}{n})$$

$$= d_{12}^{-2}(x)[n^{-1}\textstyle\sum_{i=1}^{n} F_i(x)(1- F_i(x))/n - \frac{a_n}{n}\bar{H}_n^{(1)}(x)\psi_1(K) + a_n^4\ \xi_{n,2}^2(x)]$$

$$+ O(\tfrac{a_n^2}{n})$$

as $\xi_{n,0}(x) = \bar{H}_n(x) - H(x) = o(n^{-1})$ as $n \to \infty$

$$n^{-1}\textstyle\sum_{i=1}^{n} F_i(x)(1-F_i(x)) = n^{-1}\sum_{i=1}^{n} F_i(x) - n^{-1}\sum_{i=1}^{n} F_i^2(x)$$

$$= \bar{H}_n(x) - \bar{H}_n^2(x) - n^{-1}(\textstyle\sum_{i=1}^{n} F_i^2(x) - n\bar{H}_n^2(x))$$

$$= \bar{H}_n(x)(1-\bar{H}_n(x)) - (n^{-1}\textstyle\sum_{i=1}^{n}(F_i(x) - \bar{H}_n(x))^2$$

$$= \bar{H}_n(x)(1 - \bar{H}_n(x)) - V_{nF}(x) \qquad (2.10)$$

where $V_{nF}(x) = n^{-1}\sum(F_i(x) - \bar{H}_n(x))^2 > 0$ and by considering the terms containing $\frac{a_n^2}{n}$ as of higher order,

$$MSE(p_{n,2}(x)) = d_{12}^{-2}(x)[\tfrac{\bar{H}_n(x)[1-\bar{H}_n(x)]}{n} - \tfrac{V_{nF}(x)}{n}] - d_{12}^{-2}(x)[\tfrac{a_n}{n}\bar{H}_n^{(1)}(x)\psi_1(K)$$

$$+ \tfrac{a_n^4}{4}\bar{H}_n^{(2)^2}(x)\mu_2^2(K)] + O(\xi_{n,0}(x)\ a_n^2) + o(a_n^4)$$

Corollary 2.1: Under the conditions of Theorem 2.1 on $\{F_i(x)\}$,

$$MSE(p_{n,1}(x)) = d_{12}^{-2}(x)[\tfrac{\bar{H}_n(x)[1-\bar{H}_n(x)]}{n} - \tfrac{V_{nF}(x)}{n}] + O(n^{-2})$$

where $V_{nF}(x) = n^{-1}\sum(F_i(x) - \bar{H}_n(x))^2 > 0$

Proof: This proof follows exactly the similar line of argument as for the proof of Theorem 2.1.

2.2. Asymptotic Normality of $p_{n,1}(x)$ and $p_{n,2}(x)$

We now consider the limiting distribution of BJ estimators $p_{n,j}(x)$, j=1,2 of p using Lyapunov CLT to the sequence $\{Z_{2i}(x)\}$ of independent random variables in the following Theorem.

Theorem 2.2: Under the conditions AI – AIII on $\{F_i\}$, the kernel function k, and the sequence $\{a_n\}$ for each fixed x,

$$\sqrt{n}\ (p_{n,2}(x) - p) \xrightarrow{L} N(0, \tfrac{\tau^2}{d_{12}^2(x)}) \text{ as } n \to \infty$$

where $\tau^2 = \lim_{n\to\infty}[\bar{F}_n(x)(1 - \bar{F}_n(x)) - V_{nF}(x)]$, $V_{nF}(x) = n^{-1}\sum(F_i(x) - \bar{F}_n(x))^2 > 0$

Proof: Note from (2.4)

$$d_{12}(x)[p_{n,2}(x)-p] = n^{-1}\textstyle\sum_{i=1}^{n} Z_{2i}(x) + \tau_{n2}(x)$$

with $\tau_{n2}(x) = n^{-1}\sum_1^n$ [E K($\frac{x-X_i}{a_n}$) – H(x)]

$= \bar{H}_n(x) - H(x) + \frac{a_n^2}{2}\bar{H}_n^{(2)}(x)\,\mu_2(K) + o(a_n^2) \to 0$ as $n\to\infty$

$Z_{2i}(x) = K(\frac{x-X_i}{a_n}) - E\,K(\frac{x-X_i}{a_n})$, $|Z_{2i}(x)| \le 2\,\|K\| = M < \infty$

$\sigma_{2i}^2 = Var\ Z_{2i}(x)$

$\frac{s_{n2}^2}{n} = \sum\frac{\sigma_{2i}^2}{n} = n^{-1}\sum[F_i(x)(1-F_i(x)) - a_n F_i^{(1)}(x)\psi_1(K) + O(a_n^2)]$

$= \bar{H}_n(x)(1-\bar{H}_n(x)) - V_{nF}(x) - a_n\bar{H}_n^{(1)}(x)\psi_1(K) + O(a_n^2/n)$

$s_{n2}^2 = O(n)$

In order to apply Lyapunov CLT to the sequence $\{Z_{2i}(x)\}$, consider the Lyapunov condition

$$\frac{1}{s_{n2}^3}\sum_1^n E\,|Z_i|^3 = \frac{n}{s_{n2}^3}[n^{-1}\sum_1^n E\,|Z_{2i}(x)|^3]$$

$$= O(\frac{n}{n^{3/2}}) \to 0 \text{ as } n\to\infty.$$

Now, Lyapunov condition is satisfied, and Lyapunov CLT to the sequence $\{Z_{2i}(x)\}$ holds. As $\tau_{n2}(x)\to 0$ as $n \to \infty$

$$n^{-1/2}\sum_1^n Z_{2i}(x) \xrightarrow{L} N(0,1)$$

$$\frac{s_{n2}}{n^{1/2}} \to [H(x)(1-H(x)) - V(x)]^{1/2} = \tau$$

where $V(x) = \lim_{n\to\infty} V_F(x) = \lim_{n\to\infty} n^{-1}\sum(F_i(x) - \bar{F}_n(x))^2$

$$\text{i.e. } d_{12}(x)\sqrt{n}\,(p_{n,2}(x) - p) \xrightarrow{L} N(0, \tau^2)$$

$$\sqrt{n}\,(p_{n,2}(x) - p) \xrightarrow{L} N(0, (\tau/d_{12}(x))^2)$$

Thus, the Theorem is proved.

Corollary 2.2: Under the conditions of Theorem 2.1 on $\{F_i(x)\}$,

$$\sqrt{n}\,(p_{n,1}(x) - p) \xrightarrow{L} N(0, \tfrac{\tau^2}{d_{12}^2(x)}) \text{ as } n \to \infty$$

Proof: This proof follows exactly the similar line of argument as for the proof of Theorem 2.2. ■

2.3. Rates of strong convergence of $p_{n,1}(x)$ and $p_{n,2}(x)$

We now establish a.s. convergence of the BJ type estimators $p_{n,1}(x)$ and $p_{n,2}(x)$ defined in (1.8) under non-i.i.d. set-up in the following result:

Theorem 2.3: Under the conditions of Theorem 2.1,

I. $p_{n,2}(x) - p = O(\frac{\log n}{n})^{\frac{1}{2}}$ **a.s.**

II. $\hat{\lambda}_n(x) = \hat{H}_n(x) - H(x) = O(\frac{\log n}{n})^{\frac{1}{2}}$ **a.s. as n** $\to \infty$

Proof: Note that from (2.4) and (2.8)

$$E(Z_{2i}^2(x)) = \sigma_{2i}^2 = F_i(x)(1-F_i(x)) - a_n F_i^{(1)}(x)\psi_1(K) + O(a_n^2)$$

$$\leq \frac{1}{4} + |\psi_1(K)| = C < \infty$$

$$\sigma_{n2}^2 = \frac{1}{n}\sum E[Z_{2i}^2(x)] = \frac{\sum F_i(x)(1-F_i(x))}{n} - \frac{a_n\psi_1(K)\sum F_i^{(1)}(x)}{n} + O(a_n^2)$$

$$= \bar{H}_n(x)(1-\bar{H}_n(x)) - V_{nF}(x) - a_n\bar{H}_n^{(1)}(x)\psi_1(K) + O(a_n^2) < \infty$$

By applying Bernstein (1946) inequality to $\{Z_{2i}(x)\}$ with M=2,

$$P(n^{-1}\sum Z_{n2}(x) > t) \leq \exp\left(-\frac{\frac{nt^2}{2}}{C+\frac{2}{3}t}\right)$$

setting $t = \left(\frac{4C\log n}{n}\right)^{\frac{1}{2}}$

$$P(n^{-1}\sum Z_{n2}(x) > t) \leq \exp\left[-\frac{\frac{n4C\log n}{2n}}{C+\frac{2}{3}\left(\frac{4C\log n}{n}\right)^{\frac{1}{2}}}\right]$$

$$= \exp\left[-\frac{\frac{n4C\log n}{2n}}{C\left(1+\frac{2}{3C}\left(\frac{4\log n}{n}\right)^{\frac{1}{2}}\right)}\right]$$

$$= \exp\left[\frac{-2\log n}{1+\frac{2}{3C}\left(\frac{4\log n}{n}\right)^{\frac{1}{2}}}\right]$$

$\leq n^{-2}$ for sufficiently large.

$\Rightarrow \sum_{n\geq 1}^{\infty} P(\bar{Z}_{n2} > t) \leq \sum_{n\geq 1}^{\infty} n^{-2} < \infty$

By Borel–Cantelli lemma, we conclude that $\bar{Z}_{n2} = O(\frac{logn}{n})^{1/2}$ as $n\to\infty$. Therefore, as $\tau_{n2}(x)\to 0$ as $n \to \infty$,

$$(p_{n,2}(x) - p)\, d_{12}(x) = \bar{Z}_{n2} + \tau_{n2} \overline{\overline{a.s.}}\, O(\frac{logn}{n})^{1/2}$$

i.e. $p_{n,2}(x) - p = O(\frac{logn}{n})^{1/2}$ a.s. for each x as $n\to\infty$.

(II) is an immediate consequence of part(I). Hence the result follows.

Corollary 2.3: Under the conditions of Theorem 2.1 on $\{F_i(x)\}$,

I. $p_{n,1}(x) - p = O(\frac{\log n}{n})^{\frac{1}{2}}$ **a.s.**

II. $\tilde{\lambda}_n(x) = \tilde{H}_n(x) - H(x) = O(\frac{\log n}{n})^{\frac{1}{2}}$ **a.s. as n** $\to \infty$

Proof: The proof follows exactly the similar line of argument as for the proof of Theorem 2.3.

3. Optimal bandwidth $a_{n,opt}$

We select the optimal $a_{n,opt}$ as that a_n for which MSE ($p_{n,2}$(x)) is the minimum. Solving the equation $\frac{\partial \text{MSE}\,(p_{n,2}(\text{x}))}{\partial x}$ = 0 for a_n;

$$\text{i.e. M} = \text{MSE}(p_{n,2}(\text{x})) = \frac{(\bar{H}_n(x)[1-\bar{H}_n(x)]-v_{nF}(x))}{n} - \frac{a_n}{n}\,\xi_{n,1}(\text{x}) + a_n^4\,\xi_{n,2}^2(\text{x})$$

$$\frac{\partial \text{M}}{\partial a_n} = 0 = -\frac{1}{n}\,\xi_{n,1}(\text{x}) + 4a_n^3\,\xi_{n,2}^2(\text{x})$$

so that

$$a_{n,opt} = \left[\frac{\xi_{n,1}(x)}{4\xi_{n,2}^2(\text{x})}\right]^{1/3} \cdot n^{-1/3} \tag{3.1}$$

where $\xi_{n,1}(\text{x}) = \bar{H}_n^{(1)}(\text{x})\psi_1(\text{K})$, $\xi_{n,2}^2(\text{x}) = \frac{1}{4}\,\bar{H}_n^{(2)^2}(\text{x})\mu_2^2(\text{K})$, $\psi_1(\text{K}) = 2\int t\text{K}(t)d\text{K}(t)$

4. Comparisons between the estimators

We first compare the performance of the proposed smoothed estimator $p_{n,2}$(x) with Boes-James type estimator $p_{n,1}$(x), when H_1(x), H_2(x) are known based on the minimum mean square error (MSE) criterion under non-i.i.d. set-up. Note that from Theorem 2.1 and Corollary 2.1,

$$\text{MSE}\,(p_{n,2}(\text{x})) < \text{MSE}\,(p_{n,1}(\text{x}))$$

$$\text{If } \frac{a_n}{n}\,\bar{H}_n^{(1)}(\text{x})\psi_1(\text{K}) > \frac{a_n^4}{4}\,\bar{H}_n^{(2)^2}(\text{x})\mu_2^2(\text{K})$$

$$\text{If } a_n\bar{H}_n^{(1)}(\text{x})\,\psi_1(\text{K}) > \text{n}a_n^4\,[\frac{1}{4}\,\bar{H}_n^{(2)^2}(\text{x})\mu_2^2(\text{K})] \tag{4.1}$$

for finite values of n. Since both terms on the left side of the above inequality are always positive and in view of $\text{n}a_n^4 \to 0$ for moderate n, (4.1) holds. The gain in precision of $p_{n,2}$(x) over $p_{n,1}$(x) is defined as

$$\frac{\text{MSE}\,(\text{p}_{\text{n},1}(x))-\text{MSE}\,(p_{n,2}(\text{x}))}{\text{MSE}(p_{n,1}(\text{x}))} \times 100.$$

5. Monte Carlo Simulation

A simulation study is carried out in the estimation of p by $p_{n,j}$(x); j=1,2 when two component distributions are known and are estimated by using empirical distribution function and kernel distribution function for Normal and Exponential populations. The procedure is given in appendix A.

Table 5.1. Simulation results of $p_{n,j}(x_0)$ and $p_{n,j}(x_0)$ for different sets N of sample size n with p=0.5 and X_0 = -2, -1, -0.5, 0.5, 1, 2 and X_0 = 0.2, 0.3, 0.33, 0.4, 0.5, 0.6 for Exponential population

p=0.5	N	H_1(x)=N(βt_{1i},0.5^2), H_2(x)= N(βt_{2i},3^2), H(x)=N(βt_i,$(2.151)^2$)					p=0.5	H_1(x)=Exp(2), H_2(x)=Exp(3), H(x)=Weibull(1.25,k=0.5)				
		$p_{n,1}(x_0)$	$p_{n,2}(x_0)$	$\widehat{MSE}$		Efficiency		$p_{n,1}(x_0)$	$p_{n,2}(x_0)$	$\widehat{MSE}$		Efficiency
				$p_{n,1}(x_0)$	$p_{n,2}(x_0)$					$p_{n,1}(x_0)$	$p_{n,2}(x_0)$	
X_0=-2 n =12	10	0.538	0.119	0.026	0.006	78.03	X_0=0.2 n=12	0.69	0.76	0.11	0.02	80.11
	25	0.525	0.140	0.035	0.011	70.04		0.67	0.76	0.12	0.03	70.94
	50	0.525	0.144	0.039	0.013	66.12		0.58	0.74	0.12	0.04	61.31
	75	0.516	0.148	0.044	0.014	67.36		0.54	0.76	0.11	0.04	59.82
	100	0.531	0.154	0.039	0.014	64.12		0.50	0.75	0.10	0.05	48.81
X_0=-1 n=12	10	0.369	0.228	0.051	0.017	67.21	X_0=0.3 n=12	0.1889	0.865	0.055	0.011	79.59
	25	0.431	0.248	0.059	0.015	74.57		0.211	0.874	0.062	0.010	83.34
	50	0.383	0.249	0.054	0.024	55.66		0.211	0.868	0.062	0.012	80.19
	75	0.377	0.252	0.047	0.025	47.14		0.211	0.869	0.062	0.012	81.22
	100	0.376	0.249	0.048	0.024	49.31		0.211	0.867	0.062	0.012	80.19
X_0= -0.5 n=12	10	0.460	0.300	0.106	0.038	64.14	X_0=0.33 n=12	0.57	0.86	0.1810	0.0614	66.09
	25	0.362	0.265	0.081	0.032	60.68		0.32	0.87	0.0680	0.0127	81.37
	50	0.405	0.257	0.078	0.039	49.74		0.34	0.89	0.0734	0.0104	85.83
	75	0.402	0.255	0.074	0.039	45.84		0.33	0.89	0.0717	0.0082	88.62
	100	0.396	0.248	0.077	0.038	49.66		0.35	0.88	0.0757	0.0115	84.84
X_0=0.5 n=12	10	0.463	0.388	0.041	0.033	19.42	X_0=0.4 n=12	0.50	0.85	0.08	0.02	73.61
	25	0.446	0.360	0.042	0.028	32.43		0.40	0.89	0.07	0.01	83.50
	50	0.438	0.342	0.051	0.031	38.39		0.34	0.90	0.05	0.01	85.67
	75	0.475	0.335	0.066	0.032	51.79		0.30	0.90	0.04	0.01	75.67
	100	0.461	0.325	0.068	0.033	50.95		0.28	0.90	0.03	0.01	65.91
X_0=1 n=12	10	0.58	0.16	0.081	0.021	73.55	X_0=0.5 n=12	0.47	0.35	0.08	0.05	42.27
	25	0.52	0.16	0.067	0.017	74.28		0.49	0.47	0.09	0.07	19.86
	50	0.51	0.19	0.061	0.021	65.92		0.39	0.40	0.08	0.06	26.23
	75	0.47	0.17	0.061	0.022	64.30		0.44	0.43	0.08	0.07	21.28
	100	0.47	0.18	0.056	0.024	57.39		0.40	0.41	0.08	0.06	24.23
X_0=2 n=12	10	0.439	0.127	0.045	0.012	73.97	X_0=0.6 n=12	0.071	0.879	0.034	0.011	67.70
	25	0.512	0.135	0.036	0.012	67.50		0.083	0.836	0.039	0.025	36.48
	50	0.525	0.169	0.039	0.022	43.46		0.071	0.822	0.034	0.028	17.11
	75	0.494	0.163	0.053	0.024	53.91		0.075	0.827	0.036	0.027	24.21
	100	0.518	0.153	0.047	0.021	55.75		0.071	0.822	0.034	0.028	17.16

6. Comments

In the paper it is shown that when the component normal populations with parameters are $N(\beta t_{1i}, 0.25)$, $N(\beta t_{2i}, 9)$ and the mean value of estimate $p_{n,1}(x_0)$ and $p_{n,2}(x_0)$ is close to its actual value p**.** The simulation results show that $\widehat{MSE}$ for smoothed parametric estimator is less than that of unsmoothed estimator for different values of x, uniformly for all samples. Thus, the smoothed estimator is a better estimator in terms of minimum MSE when compared to BJ estimator. The average gain in efficiency due to smoothing lies between 19% to 88% for different sets N of size n.

REFERENCES

ACHESON, M. A., McELWEE, E. M., (1951). Concerning the reliability of electron tubes, Proceedings of the IRE, Vol. 40, pp. 1204–1206.

BERNSTEIN, S. N., (1946). The Theory of Probabilities, Gastehizdat Publishing House, Moscow.

BOES, D. C., (1966). On the estimation of mixing distributions, Ann. Math. Statist. 37, pp. 177–188.

BLISCHKE, W. R., (1965). Mixtures of discrete distributions, Collected papers presented at the Int. Simp. Ed. By G. P. Patil, Oxford, Pergamon.

BRUNI, C. et al., (1985). On the inverse problem in cytofluorometry-recovering DNA distribution from FMF date, Cell Bio. Physics 5, pp. 5–19.

CHOI, K., BULGREN, W. G., (1968). An estimation procedures for mixtures distributions, J.R.S.S. Series B (Methodological), Vol. 30, No. 3, pp. 444–460.

CLARK, M. V., (1976). Some methods for statistical analysis of multi-model distributions and their application to grain-size data, J. Math. Geol. 8, pp. 267–282.

CHRISTENSEN, P. R. et al., (1980). Gamma-ray multiplicity moments from 86 kr reactions on 144, 154 at 490, Mev. Nuclear Phys. A 349, pp. 217–257.

DAVIS, D. J., (1952). An analysis of some failure date, Jour. Amer.Stat.Assoc., 47, pp. 113–150.

EPSTEIN, B., (1953). Statistical problems in life testing, proceeding of The Seventh Annual Convention, American Society of Quality Control.

EVERITT, B., HAND, D., (1981). Finite mixture distributions. Chapman and Hall, London.

FU, K. S., (1968). Sequential methods in pattern recognition and machine learning, Academic Press, New York.

HARRIS, E. K., (1958). On the probability of survival of bacteria in sea water, Biometrics, 14, pp. 195–206.

HERD, G. R., (1953). Heterogeneous distributions. Unpublished technical note.

HOSMER W. DAVID, (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics, 29, pp. 761–770.

JAMES, I. R., (1978). Estimation of the Proportion in a mixture of two normal distributions from simple, rapid measurements, Biometrics, pp. 265–275.

JAYALAKSHMI, C., SUDHAKAR RAO, M., (2002). Estimation of mixing proportion using smooth distribution function in two population mixture model, Aligarh Jl. of Statistics, Vol. 22, pp. 91–99.

JEWEL, N. P., (1982). Mixture of Exponential distributions, Annals of Statistics, Vol. 10, No. 2, pp. 479–484.

KARL PEARSON, (1894). Contributions to the mathematical theory of evolution, Phil. Trans. Roy Soc. 185 A, pp. 71–110.

LINDSAY, B. G., (1995). Mixture models: Theory geometry and applications. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, Institute of Mathematical Statistics, Hayward.

MENDENHALL, W., HADER, R. J., (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data, Biometrika 45(3-4), pp. 504–520.

MACDONALD, P. D. M., PITCHER, T. J., (1979). Age groups from size frequency data, J. Fish. Res. Bd. Cano. 36, pp. 987–1008.

McLACHLAN, G. J., BASFORD, K. E., (1988). Mixture Models: Inference and Applications to Clustering. New York: Marcel Dekker.

McLACHLAN, G. J., PEEL, D., (2000). Finite mixture models, Wiley, New York.

MERZ, P. H., (1980). Determination of absorption energy distribution by regularization and a characterization of certain absorption isotherms, J. Comput. Phys. 34, pp. 64–85.

STEEN, J. R., (1952). Life testing of electronic tubes, Unpublished paper presented at a summer statistics conference, University of North Carolina.

TITTERINGTON, D. M. et al., (1985). Statistical analysis of finite mixture distributions. John Wiley & Sons.

WILDE, R. D., (1952). Nature of rate of failure curves, Unpublished paper presented at a summer statistics conference, University of North Carolina.

ODELL, P. L., BASU, J. P., (1976). Concerning several methods for estimations crop acreages using remote sensing data, Commun. Statist. A 5, pp. 1091–1114.

VAN HOUWELINGEN, J. C., DEVRIES, (1974). Minimax estimation of mixing proportion of two known distributions, J. American Stat. Asso., 397, pp. 300–304.

VARLI, Y. et al., (1975). A statistical model for position emission tomography, JASA, 80, pp. 8–20.

APPENDIX A

Random samples of sizes n_1=6 and n_2=6 are generated from the two component mixtures of the normal populations with parameters $(\mu_1,\mu_2)=(\beta t_{1i},\beta t_{2i})$ and $(\sigma_1^2,\sigma_2^2)=(0.5^2,3^2)$ and with parameters (θ_1,θ_2) = (2,3) for Exponential populations. The mixed sample of size n = n_1+n_2 = 12 is generated from the normal population with parameters ($\mu = p\mu_1+q\mu_2$) = ($\beta t_i=p\beta t_{1i}+q\beta t_{2i}$) and σ^2 = ($p\sigma_1^2+q\sigma_2^2$), and in the case of Exponential population, the mixed sample is drawn from Weibull population with shape parameter k less than 1. Since Weibull distribution with shape parameter k<1 arises as a mixture of Exponential distributions (Jewel 1982), the samples of sizes n, n_1, n_2 are independent. Taking p=q= 0.5, β=0.1 and $t_i = \mp \frac{i}{n^{\delta}}$; j = 1, 2, δ =1.5 are selected in such a way that $\sum t_i$ = 0 and $\sum t_i^2 \rightarrow 0$. The present simulation study is to estimate parametric estimators such as $p_{n,1}(x)$ and $p_{n,2}(x)$ for $x=x_0$ defined as follows.

$$\mathbf{p}_{n,1}(\boldsymbol{x_0}) = \frac{\widetilde{\mathbf{H}}_{\mathbf{n}}(\boldsymbol{x_0})-\mathbf{H_2}(\boldsymbol{x_0})}{\mathbf{H_1}(\boldsymbol{x_0})-\mathbf{H_2}(\boldsymbol{x_0})} \quad \text{and} \quad \mathbf{p}_{n,2}(\boldsymbol{x_0}) = \frac{\widehat{\mathbf{H}}_{\mathbf{n}}(\boldsymbol{x_0})-\mathbf{H_2}(\boldsymbol{x_0})}{\mathbf{H_1}(\boldsymbol{x_0})-\mathbf{H_2}(\boldsymbol{x_0})}$$

where $\widetilde{\mathrm{H}}_{\mathrm{n}}(x_0)$, $\widehat{\mathrm{H}}_{\mathrm{n}}(x_0)$ are estimated by the usual empirical and kernel-based distribution functions and $H_j(x)$, j=1,2 such as

$$\widetilde{\mathrm{H}}_{\mathrm{n}}(x) = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} I(X_i \le x) \text{ , if } I(X_i \le x), \text{ assign 1 otherwise 0.}$$

$$\hat{H}_{\mathrm{n}}(x_0) = \mathrm{n}^{-1} \sum_{i=1}^{n} K\left(\frac{x_0 - X_i}{a_n}\right); \; H_j(x) = \frac{1}{n_j} \sum_{i=1}^{n_j} F_{ji}(x_0).$$

Here, we used the Epanechnikov kernel function as $k(u) = \frac{3}{4}(1 - u^2)$; $|u| \le 1$ for Normal distribution, and for Exponential distribution we used Gaussian kernel function as $k(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$. The distribution function of the Epanechnikov kernel function is

$$K\left(\frac{x_0 - X_i}{a_n}\right) = \frac{3}{4}\left[\frac{x_0 - X_i}{a_n} - \frac{1}{3}\left(\frac{x_0 - X_i}{a_n}\right)^3 + \frac{2}{3}\right]$$

Thus, the estimators becomes

$$p_{n,1}(x_0) = \frac{\frac{1}{n}\sum_{i=1}^{n} I(X_i \le x_0) - \frac{1}{n_2}\sum_{i=1}^{n_2} F_{2i}(x_0)}{\frac{1}{n_1}\sum_{i=1}^{n_1} F_{1i}(x_0) - \frac{1}{n_2}\sum_{i=1}^{n_2} F_{2i}(x_0)}$$

$$p_{n,2}(x_0) = \frac{n^{-1}\sum_{i=1}^{n} K\left(\frac{x_0 - X_i}{a_n}\right) - \frac{1}{n_2}\sum_{i=1}^{n_2} F_{2i}(x_0)}{\frac{1}{n_1}\sum_{i=1}^{n_1} F_{1i}(x_0) - \frac{1}{n_2}\sum_{i=1}^{n_2} F_{2i}(x_0)} \qquad (5.1)$$

where $F_{ji}(x)$ are the cumulative distribution functions of Normal distribution and Exponential distribution.

All computations are done by using MS Excel, and the procedure is as follows.

1. Generate n uniform random numbers between (0,1).
2. Generate the cumulative distribution function of Normal and Exponential distribution by taking different means βt_{ji} and variances σ_j^2; j=1,2 at x = x_0
3. Generate the mixed normal and Weibull observations by taking different means $\beta t_i = p\beta t_{1i} + q\beta t_{2i}$ and variance $\sigma^2 = p\sigma_1^2 + q\sigma_2^2$ and $\theta = p\theta_1 + q\theta_2$ respectively with k<1.
4. Calculate (5.1) by taking X_0 = -2, -1, -0.5, 0.5, 1, 2 related to Normal distribution and X_0= 0.2, 0.3, 0.33, 0.4, 0.5, 0.6 related to Exponential distribution.

Generate N=100 mixed and component independent sample sets with sample sizes n=12 and n_1= 6 and n_2=6, so that $n = n_1+n_2$ and calculate $p_{n,1}(x_0)$ and $p_{n,2}(x_0)$, their mean values $\bar{p}_{n,j}(x_0) = \frac{1}{N}\sum p_{n,j}(x_0)$ and their mean square errors $\widehat{MSE}\ (p_{n,j}(x_0)) = \frac{1}{N}\sum_{i=1}^{N}(p_{n,j}(x_0) - \bar{p}_{n,j}(x_0))^2$; j=1,2. Sets are ignored when p ≥1. The results are presented in table 5.1.

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 103–116, DOI 10.21307/stattrans-2019-006

ODD GENERALIZED EXPONENTIAL LOG-LOGISTIC DISTRIBUTION GROUP ACCEPTANCE SAMPLING PLAN

Devireddy Charana Udaya Sivakumar[1], **Konda Rosaiah**[2], **Gadde Srinivasa Rao**[3], **Kruthiventi Kalyani**[1]

ABSTRACT

In this manuscript, a group acceptance sampling plan (GASP) is developed when the lifetime of the items follows odd generalized exponential log-logistic distribution (OGELLD), the multiple number of items as a group can be tested simultaneously in a tester. The design parameters such as the minimum group size and the acceptance number are derived when the consumer's risk and the test termination time are specified. The operating characteristic (OC) function values are calculated (intended) according to various quality levels and the minimum ratios of the true average life to the specified average life at the specified producer's risk are derived. The methodology is illustrated through real data.

Key words: odd generalized exponential log-logistic distribution, group acceptance sampling plan, truncated life test.

1. Introduction

In the present highly competitive global market, different items/products are categorized on several factors of end products. One such factor is quality/durability of a product, which can be examined through most of statistical quality control techniques, which are the two important statistical tools for ensuring the quality of the product, which are (i) Process control and (ii) Product control. In acceptance sampling plans for a truncated life test, the utmost issue is to determine the sample size from a lot under cogitation. In most of the statistical quality control experiments, it is not possible to perform 100% inspection due to various reasons. It is implicitly assumed in the usual sampling plan; the decision of accepting or rejecting a lot is on the basis of a sample of items. To save cost and time in the life test, it is very often to put a number of items in a tester. In this life test, a tester is called a group and the number of items in each tester is called the group size. The acceptance sampling via the group life test is called the group acceptance sampling plan (GASP), which is also often enacted under a truncated life test. For such a type of test, the determination of the sample size is equivalent

[1] UGC BSR Fellows, Department of Statistics, Acharya Nagarjuna University, Guntur – 522 510, India.
[2] Department of Statistics, Acharya Nagarjuna University, Guntur – 522 510, India.
[3] Department of Statistics, The University of Dodoma, P.O.Box: 259, Tanzania.
E-mail:gaddesrao@gmail.com. ORCID ID: https://orcid.org/0000-0002-3683-5486.

to fixation of the number of groups. This type of testers is customarily used in the case of the so-called sudden death testing, which is discussed by Pascual and Meeker (1998) and Vlcek *et al*. (2004). Jun *et al.* (2006) introduced this group concept into acceptance sampling plan and developed variable sampling plans for sudden death testing for the Weibull distribution. GASPs under a truncated life test have been studied by many researchers for different lifetime distributions.

A group sampling plan based on truncated life test based on the gamma distributed items was contemplated by Aslam *et al.* (2009). Aslam and Jun (2009a, 2009b) developed a group acceptance sampling plan for truncated life tests based on the inverse Rayleigh, log-logistic and Weibull distributions. Balamurali and Jun (2009) proposed a repetitive group sampling procedure for variables inspection. Rao (2009, 2010) presented a group acceptance sampling plans for lifetimes following a generalized exponential distribution and Marshall-Olkin extended Lomax distribution. Group acceptance sampling plans for Pareto distribution of the second kind was discussed by Aslam *et al.* (2010). Radhakrishnan and Alagirisamy (2011) developed a group acceptance sampling plan using weighted binomial distribution. Ramaswamy and Anburajan (2012) developed a group acceptance sampling plan using weighted binomial on truncated life tests for inverse Rayleigh and log-logistic distributions. A group acceptance sampling plan using weighted binomial for a truncated life test when the lifetime of an item follows exponential and Weibull distributions was contemplated by Anburajan and Ramaswamy (2015). Rao and Ramesh (2015) considered a group acceptance sampling plans for exponentiated half logistic distribution. Rao *et al.* (2016) studied group acceptance sampling plans for lifetimes following an exponentiated Fréchet distribution. Rao and Rao (2016) developed a two-stage group acceptance sampling plan based on life tests for half logistic distribution. Group acceptance sampling plans for odds exponential log-logistic distribution and Type-II generalized log-logistic distribution were contemplated by Rosaiah *et al.* (2016a, 2016b).

To reduce the experimental time and cost, GASPs have been used (see Jun *et al.,* (2006)). In this case, the total number of items *n* to be tested is divided into equal-sized groups according to the number of available experimental testers. Suppose '*r'* items are in each group and there are a total of '*g'* groups, then $n = rg$. The items in each group are tested independently under identical environmental conditions. Moreover, all the testers run simultaneously. The experiment is stopped at a pre-specified time *t*. If '*c'* is the acceptance number for this experiment, then a lot is accepted if the recorded number of failures in each group is less than '*c'* during the experimental time *t*. The single sampling plans handle this problem by assuming a parametric model for the lifetime distribution and then deriving the minimum sample size '*n'* needed to ensure certain mean or median life of the items under investigation. It is further assumed that the experimental time and the number of items in each group are prefixed in advance. Since *n=rg*, determining '*n'* is equivalent to determining '*g'*.

The main purpose of this manuscript is to develop the GASPs for the odd generalized exponential log-logistic distribution (OGELLD). Gupta (1962) suggested that for a skewed distribution, the median represents a better quality parameter than the mean. On the other hand, for symmetric distribution, mean is preferable to use as a quality parameter. Since OGELLD is a skewed distribution,

we prefer to use the percentile point as the quality parameter and it will be denoted by t_q.

The rest of this manuscript is organized as follows. In Section 2, we describe concisely the OGELLD distribution. The design of group acceptance sampling for lifetime percentiles under a truncated life test is discussed in Section 3. In Section 4, description of the proposed methodology with real data example is presented. A comparison of distributions is discussed in Section 5. Finally, conclusions are made in Section 6.

2. The Odd Generalized Exponential Log-Logistic Distribution (OGELLD)

In this section, we provide a brief summary of the odd generalized exponential log-logistic distribution (OGELLD). The OGELLD was introduced and studied quite extensively by Rosaiah *et al.* (2016c). The probability density function (pdf) and cumulative distribution function (cdf) of OGELLD respectively are given as follows:

$$f(\mathrm{t};\sigma,\lambda,\theta,\gamma)=\frac{\gamma\theta}{\lambda\sigma}\left(t/\sigma\right)^{\theta-1}\left[1-e^{\frac{-1}{\lambda}(t/\sigma)^{\theta}}\right]^{\gamma-1}e^{\frac{-1}{\lambda}(t/\sigma)^{\theta}},\ t>0,\sigma,\lambda,\theta,\gamma>0 \tag{1}$$

and $$F(\mathrm{t};\sigma,\lambda,\theta,\gamma)=\left[1-e^{\frac{-1}{\lambda}(t/\sigma)^{\theta}}\right]^{\gamma},\ t>0,\ \sigma,\lambda,\theta,\gamma>0. \tag{2}$$

where σ, λ are the scale parameters and θ,γ are the shape parameters. The 100*q*-th percentile of the OGELLD is given as:

$$t_q=\sigma\eta_q,\quad \text{where}\quad \eta_q=\left[-\lambda\ln(1-q^{1/\gamma})\right]^{1/\theta}. \tag{3}$$

Hence, for the fixed values of $\lambda=\lambda_0$ and $\theta=\theta_0$, the quantile t_q given in Equation (3) is the function of scale parameter $\sigma=\sigma_0$, that is $t_q\geq t_q^0\Leftrightarrow\sigma\geq\sigma_0$,

where $$\sigma_0=\frac{t_q^0}{\left[-\lambda_0\ln(1-q^{1/\gamma_0})\right]^{1/\theta_0}}. \tag{4}$$

Note that σ_0 also depends on λ_0 and θ_0, to build up acceptance sampling plans for the OGELLD ascertain $t_q\geq t_q^0$, equivalently that σ exceeds σ_0.

3. The Group Acceptance Sampling Plan (GASP)

In this section, we provide group acceptance sampling plans (GASPs) when a lifetime of the product is an OGELLD with known scale and shape parameters λ, θ. We propose the GASP under the truncated life test, which is based on the total number of failures from all groups. The procedure of the proposed plan is as follows [Aslam *et al*. (2011a)]:

- Step 1: Randomly draw a sample of size *n* from a production lot, allocate *r* items to each of *g* groups (or testers) so that *n* = *rg* and put them on test until the pre-determined t_0 units of time.
- Step 2: Accept the lot when the number of failures from all *g* groups is smaller than or equal to *c*. Truncate the test and reject the lot as soon as the number of failures from all *g* groups is larger than *c* before t_0.

The probability of accepting a lot for the group sampling plan based on the number of failures from all groups under a truncated life test at the test time schedule t_0 is

$$\mathrm{P}_a(p) = \sum_{i=0}^{c} \binom{rg}{i} p^i \left(1-p\right)^{rg-i} \tag{5}$$

where *'g'* is the number of groups, *'c'* is the acceptance number, *'r'* is the group size, and *'p'* is the probability of getting a failure within the life test schedule, t_0. Since the product lifetime follows OGELLD, we have $p = F\left(t_0\right)$. Usually, it would be convenient to determine the experiment termination time, t_0, as $t_0 = \delta_q^0 t_q^0$ for a constant δ_q^0 and the targeted 100q-th lifetime percentile, t_q^0. Let t_q be the true 100q-th lifetime percentile. Then, *p* can be rewritten as

$$p = \left[1 - e^{-\frac{1}{\lambda}\left(\frac{t_0}{\sigma}\right)^{\theta}}\right]^{\gamma} = \left[1 - e^{(-1/\lambda)\left(\eta_q \delta_q^0 / \left(t_q / t_q^0\right)\right)^{\theta}}\right]^{\gamma} \tag{6}$$

In order to find the design parameters of the proposed GASP, we prefer the approach based on two points on the OC curve by considering the producer's and consumer's risks. In our approach, the quality level is measured through the ratio of its percentile lifetime to the lifetime, t_q / t_q^0. These percentile ratios are very helpful to the producer to enhance the quality of products. From the producer's perspective, the probability of lot acceptance should be at least $1-\alpha$ at the acceptable reliability level (ARL), p_1. Thus, the producer demands that a lot should be accepted at various levels, say t_q / t_q^0 =2, 4, 6, 8, 10 in Equation (6). On the other hand, from the consumer's viewpoint, the lot rejection probability should

be at most β at the lot tolerance reliability level (LTRL), p_2. In this way, the consumer considers that a lot should be rejected when t_q / t_q^0 =1, in Equation (6).

$$\sum_{i=0}^{c} \binom{rg}{i} p_1^i (1-p_1)^{rg-i} \geq 1-\alpha \tag{7}$$

$$\sum_{i=0}^{c} \binom{rg}{i} p_2^i (1-p_2)^{rg-i} \leq \beta \tag{8}$$

where p_1 and p_2 are given by

$$p_1 = \left[1-e^{(-1/\lambda)\left(\eta_q \delta_q^0 / \left(t_q/t_q^0\right)\right)^\theta}\right]^\gamma \text{ and } p_2 = \left[1-e^{(-1/\lambda)\left(\eta_q \delta_q^0\right)^\theta}\right]^\gamma \tag{9}$$

The plan parametric quantities for distinct values of parameters λ, θ and γ are constructed. Given the producer's risk $\alpha = 0.05$ and termination time schedule $t_0 = \delta_q t_q^0$ with $\delta_q^0 = 0.5$ or 1, the three parameters of the proposed group acceptance sampling plan under the truncated life test at the pre-specified time, t_0, with $\lambda=\theta=\gamma=2$ are obtained according to the consumer's risk $\beta = 0.25, 0.10, 0.05$ and 0.01 for 50th and 25th percentiles, which are shown in Tables 1 to 4.

4. Description of the proposed methodology with real data example

We demonstrate the application of the proposed group acceptance sampling plan for the OGELLD using real lifetime data set from Dey *et al*. (2018), which represent the survival times (in days) of 72 guinea pigs infected with virulent tubercle bacilli. Guinea pigs are known to have high susceptibility to human tuberculosis, which is one of the reasons to choose guinea pigs for such a study. Here, we consider only the study where all animals in a single cage are under the same regimen. The data were observed and reported by Bjerkedal (1960). For ready reference the data set is given below:

0.1, 0.33, 0.44, 0.56, 0.59, 0.72, 0.74, 0.77, 0.92, 0.93, 0.96, 1, 1, 1.02, 1.05, 1.07, 1.07, 1.08, 1.08, 1.08, 1.09, 1.12, 1.13, 1.15, 1.16, 1.2, 1.21, 1.22, 1.22, 1.24, 1.3, 1.34, 1.36, 1.39, 1.44, 1.46, 1.53, 1.59, 1.6, 1.63, 1.63, 1.68, 1.71, 1.72, 1.76, 1.83, 1.95, 1.96, 1.97, 2.02, 2.13, 2.15,

2.16, 2.22, 2.3, 2.31, 2.4, 2.45, 2.51, 2.53, 2.54, 2.54, 2.78, 2.93, 3.27, 3.42, 3.47, 3.61, 4.02, 4.32, 4.58, 5.55.

Before illustrating the methodology, our model is tested for goodness of fit. The maximum likelihood estimates of the three parameters of OGELLD for the survival times of guinea pigs data are $\hat{\lambda} = 1.1513, \hat{\theta} = 1.1606$ and $\hat{\gamma} = 2.6538$. Using the Kolmogorov-Smirnov test, we found that the maximum distance between the data and the fitted OGELLD is 0.089 with p-value 0.617. Thus, the three parameter OGELLD provides a reasonable fit for the survival times of guinea pigs data. The goodness of fit for our model is emphasized by plotting the density plot and Q-Q plot displayed in Figure 1. The plan parameters are also computed at fitted parametric values and are displayed in Tables 5 for 50th percentiles. Suppose that it is desired to develop a group acceptance sampling plan to satisfy that the 50th percentile lifetime is greater than survival times of guinea pigs 0.20 days through the experiment to be completed survival times of guinea pigs by 0.40 days to protect the producer's risk at 5%. For $\hat{\lambda} = 1.1513, \hat{\theta} = 1.1606$ and $\hat{\gamma} = 2.6538$, the consumer's risk is $\beta = 0.25$, r=5, $\delta_q^0 = 0.5$ and t_q / t_q^0=2, the minimum number of groups and the acceptance number are g=6 and c=2 from Table 5. Thus, the design can be implemented as follows: select a total of 30 guinea pigs and allocate five guinea pigs to each of the 6 groups. We can accept the lot when no more than two failures occur before survival times of guinea pigs 0.40 days from each of the 6 groups. According to this plan, the survival times of guinea pigs could have been accepted because there are only two failures before the termination time 0.40 days.

5. Comparison of distributions

In Table 7, we compare the plan parameters of the proposed group acceptance sampling plan with the generalized log-logistic distribution (GLLD) studied by Aslam *et al*. (2011b) and odds exponential log-logistic distribution (OELLD) studied by Rosaiah *et al.* (2016a), when $\beta = 0.10$ and $r = 5, \delta_q^0 = 0.5$. The acceptance number for the OGELLD is smaller as compared to GLLD and OELLD for 50th percentile.

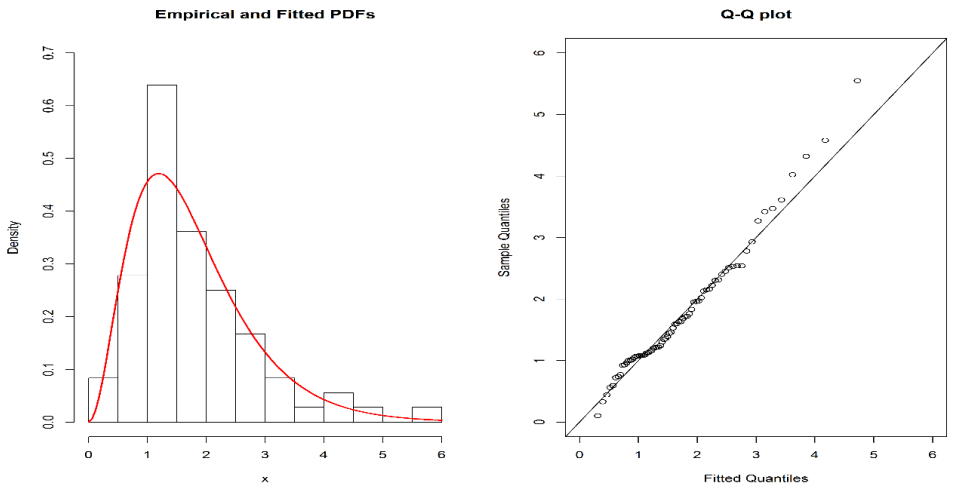


Figure 1. The density plot and Q-Q plot of the fitted OGELLD for the survival times of guinea pigs data.

6. Conclusions

In this article, a group acceptance sampling plan is developed when the lifetime of the product follows OGELLD. The plan parametric quantities like the number of groups, '*g',* and the acceptance number '*c'* are determined by considering the consumer's risk and producer's risk simultaneously. Our proposed plan noticed that if the percentile ratio increases, the number of groups '*g'* reduces and as '*r*' increases the number of groups reduces for all the parametric combinations considered in this article. The proposed plan is illustrated with a real lifetime data set in health sciences, survival times of guinea pigs in days, and results show that our methodology performs well as compared with existing sampling plans.

REFERENCES

ANBURAJAN, P., RAMASWAMY, A. S., (2015). A group acceptance sampling plan for weighted Binomial on truncated life tests using exponential and Weibull distributions, Journal of Progressive Research in Mathematics, 2 (1), pp. 80–88.

ASLAM M., JUN C. H., (2009a). A group acceptance sampling plans for truncated life tests based on the inverse Rayleigh and log-logistic distributions, Pakistan Journal of Statistics, 25 (2), pp. 107–119.

ASLAM, M., JUN, C. H., (2009b). A group acceptance sampling plan for truncated life test having Weibull distribution, Journal of Applied Statistics, 36 (9), pp 1021–1027.

ASLAM, M., JUN C. H., AHMAD. M., (2009). A group sampling plan based on truncated life tests for gamma distributed items, Pakistan Journal of Statistics, 25 (3), pp. 333–340.

ASLAM, M., MUGHAL, A. R., AHMAD, M., YAB, Z., (2010). Group acceptance sampling plans for Pareto distribution of the second kind, Journal of Testing and Evaluation, 38 (2), pp. 143–150.

ASLAM, M., KUNDU, D., JUN, C. H., AHMAD, M., (2011a). Time truncated group acceptance sampling plans for generalized exponential distribution, Journal of Testing and Evaluation, 39 (4), pp. 671–677.

ASLAM, M., SHOAIB, M., JUN, C. H., NADIA, S., (2011b). Time truncated group acceptance sampling plans for lifetime percentiles under generalized log-logistic distributions, International Journal of Current Research and Review, 3 (11), pp. 23–35.

BALAMURALI, S., JUN, C. H., (2006). Repetitive group sampling procedure for variables inspection, Journal of Applied Statistics, 33 (10), pp. 327–338.

BJERKEDAL,T., (1060) Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli. American Journal of Hygiene, 72 (1) (1960), pp. 130–148.

DEY, S., NASSAR, M., KUMAR, D., (2018). Alpha power transformed inverse Lindley distribution: A distribution with an upside-down bathtub-shaped hazard function, Journal of Computational and Applied Mathematics, 348, pp. 130–145.

GUPTA, S. S., (1962). Life test sampling plans for normal and lognormal distribution, Technometrics, 4 (2), pp. 151–175.

JUN, C.-H., BALAMURALI, S., LEE, S.-H., (2006). Variables sampling plans for Weibull distributed lifetimes under sudden death testing, IEEE Transactions on Reliability, 55 (1), pp. 53–58.

PASCUAL, F. G., MEEKER, W. Q., (1998). The modified sudden death test: Planning life tests with a limited number of test positions, Journal of Testing and Evaluation, 26 (5), pp. 434–443.

RADHAKRISHNAN, R., ALAGIRISAMY, K., (2011). Construction of group acceptance sampling plan using weighted binomial distribution, International Journal of Recent Scientific Research, 2 (7), pp. 229–231.

RAMASWAMY, A. S., ANBURAJAN, P., (2012). Group acceptance sampling plan using weighted binomial on truncated life tests for inverse Rayleigh and log-logistic distributions, IOSR Journal of Mathematics, 2 (3), pp. 33–38.

RAO G. S., (2009). A group acceptance sampling plans for lifetimes following a generalized exponential distribution, Economic Quality Control, 24 (1), pp. 75–85.

RAO, G. S., (2010). A group acceptance sampling plans based on truncated life tests for Marshall-Olkin extended Lomax distribution, Electronic Journal of Applied Statistical Analysis, 3 (1), pp. 18–27.

RAO, G. S., RAMESH, CH. N., (2015). An exponentiated half logistic distribution to develop a group acceptance sampling plans with truncated time, Journal of Statistics and Management Systems, 18 (6), pp. 519–531.

RAO, G. S., ROSAIAH, K., SRIDHAR BABU, M., (2016). Group Acceptance sampling plans for lifetimes following an exponentiated Fréchet distribution, International Journal of Applied Research and Studies, 5 (3), pp. 1–13.

RAO, B. S., RAO, G. S., (2016). A two-stage group acceptance sampling plan based on life tests for half logistic distribution, Model Assisted Statistics and Applications, 11 (3), pp. 203–211.

ROSAIAH, K., RAO, G. S., KALYANI, K., SIVAKUMAR, D. C. U., (2016a). Group acceptance sampling plans for lifetimes following an odds exponential log-logistic distribution, Sri Lankan Journal of Applied Statistics, 17 (3), pp. 201–216.

ROSAIAH, K., RAO, G. S., PRASAD, S. V. S. V. S. V., (2016b). A group acceptance sampling plans based on truncated life tests for Type-II generalized log-logistic distribution, Prob Stat Forum, 9, pp. 88–94.

ROSAIAH, K., RAO, G. S., SIVAKUMAR, D. C. U., KALYANI, K., (2016c). The odd generalized exponential log logistic distribution, International Journal of Mathematics and Statistics Invention, 4 (5), pp. 21–29.

VLCEK, B. L., HENDRICKS, R. C., ZARETSKY, E.V., (2004). Monte Carlo Simulation of Sudden Death Bearing Testing, Tribology Transactions, 47 (2), pp. 188–199.

APENNDIX

Table 1. GASP for OGELLD with $\lambda = 2, \theta = 2$ and $\gamma = 2$ for 50th percentile

β	t_q / t_q^0	r=5						r=10					
		δ_q =0.50			δ_q =1.0			δ_q =0.50			δ_q =1.0		
		c	g	$P_a(p)$	c	g	$P_a(p)$	c	g	$P_a(p)$	c	g	$P_a(p)$
0.25	2	1	8	0.9797	3	4	0.9531	1	4	0.9797	-	-	-
	4	0	4	0.9928	0	1	0.9730	0	2	0.9928	1	2	0.9947
	6	0	4	0.9986	0	1	0.9944	0	2	0.9986	0	1	0.9888
	8	0	4	0.9995	0	1	0.9982	0	2	0.9995	0	1	0.9964
0.10	2	1	11	0.9635	3	4	0.9531	1	6	0.9572	-	-	-
	4	0	7	0.9874	0	1	0.9730	0	4	0.9857	1	2	0.9947
	6	0	7	0.9975	0	1	0.9944	0	4	0.9971	0	1	0.9888
	8	0	7	0.9992	0	1	0.9982	0	4	0.9991	0	1	0.9964
0.05	2	2	18	0.9866	3	4	0.9531	2	9	0.9866	-	-	-
	4	0	9	0.9839	0	1	0.9730	0	5	0.9821	1	2	0.9947
	6	0	9	0.9968	0	1	0.9944	0	5	0.9964	0	1	0.9888
	8	0	9	0.9990	0	1	0.9982	0	5	0.9989	0	1	0.9964
0.01	2	2	24	0.9715	3	4	0.9531	2	12	0.9715	-	-	-
	4	0	13	0.9768	1	3	0.9970	0	7	0.9750	1	2	0.9947
	6	0	13	0.9953	0	2	0.9888	0	7	0.9950	0	1	0.9888
	8	0	13	0.9985	0	2	0.9964	0	7	0.9984	0	1	0.9964

Table 2. GASP for OGELLD with $\lambda = 2, \theta = 1.5$ and $\gamma = 1.5$ for 50th percentile

β	t_q / t_q^0	*r=5*						*r=10*					
		δ_q =0.50			δ_q =1.0			δ_q =0.50			δ_q =1.0		
		c	*g*	$P_a(p)$	*c*	*g*	$P_a(p)$	*c*	*g*	$P_a(p)$	*c*	*g*	$P_a(p)$
0.25	2	3	7	0.9502	-	-	-	4	5	0.9513	-	-	-
	4	1	4	0.9864	2	3	0.9798	1	2	0.9864	4	5	0.9513
	6	0	2	0.9643	1	2	0.9885	1	2	0.9976	1	2	0.9564
	8	0	2	0.9810	1	2	0.9966	1	2	0.9993	1	2	0.9864
0.10	2	4	10	0.9513	-	-	-	4	5	0.9513	-	-	-
	4	1	5	0.9792	2	3	0.9798	1	3	0.9707	4	5	0.9513
	6	1	5	0.9963	1	2	0.9885	1	3	0.9946	1	2	0.9564
	8	1	5	0.9989	1	2	0.9966	1	3	0.9985	1	2	0.9864
0.05	2	5	13	0.9548	-	-	-	6	8	0.9591	-	-	-
	4	1	6	0.9707	2	3	0.9798	1	3	0.9707	4	5	0.9513
	6	1	6	0.9946	1	2	0.9885	1	3	0.9946	1	2	0.9564
	8	1	6	0.9985	1	2	0.9966	1	3	0.9985	1	2	0.9864
0.01	2	7	19	0.9634	-	-	-	7	10	0.9528	-	-	-
	4	1	8	0.9504	2	3	0.9798	1	4	0.9504	4	5	0.9513
	6	1	8	0.9906	1	3	0.9746	1	4	0.9906	1	2	0.9564
	8	1	8	0.9973	1	3	0.9923	1	4	0.9973	1	2	0.9864

Table 3. GASP for OGELLD with $\lambda = 2, \theta = 2$ and $\gamma = 2$ for 25th percentile

β	t_q / t_q^0	*r=5*						*r=10*					
		δ_q =0.50			δ_q =1.0			δ_q =0.50			δ_q =1.0		
		c	*g*	$P_a(p)$	*c*	*g*	$P_a(p)$	*c*	*g*	$P_a(p)$	*c*	*g*	$P_a(p)$
0.25	2	1	22	0.9830	1	2	0.9748	1	11	0.9830	2	3	0.9604
	4	0	11	0.9936	0	1	0.9910	0	6	0.9931	0	1	0.9822
	6	0	11	0.9987	0	1	0.9980	0	6	0.9986	0	1	0.9964
	8	0	11	0.9996	0	1	0.9994	0	6	0.9996	0	1	0.9988
0.10	2	1	31	0.9678	2	4	0.9866	1	16	0.9659	2	3	0.9604
	4	0	18	0.9896	0	2	0.9822	0	9	0.9896	0	1	0.9822
	6	0	18	0.9979	0	2	0.9964	0	9	0.9979	0	1	0.9964
	8	0	18	0.9993	0	2	0.9988	0	9	0.9993	0	1	0.9988
0.05	2	1	38	0.9535	2	5	0.9754	1	19	0.9535	2	3	0.9604
	4	0	24	0.9862	0	3	0.9734	1	12	0.9862	0	2	0.9647
	6	0	24	0.9972	0	3	0.9946	0	12	0.9972	0	2	0.9928
	8	0	24	0.9991	0	3	0.9983	0	12	0.9991	0	2	0.9977
0.01	2	2	66	0.9777	3	8	0.9818	2	33	0.9777	3	4	0.9818
	4	0	36	0.9793	0	4	0.9647	0	18	0.9793	0	2	0.9647
	6	0	36	0.9959	0	4	0.9928	0	18	0.9959	0	2	0.9928
	8	0	36	0.9987	0	4	0.9977	0	18	0.9987	0	2	0.9977

Table 4. GASP for OGELLD with $\lambda = 2, \theta = 1.5$ and $\gamma = 1.5$ for 25th percentile

β	t_q / t_q^0	*r=5*						*r=10*					
		δ_q =0.50			δ_q =1.0			δ_q =0.50			δ_q =1.0		
		c	*g*	$P_a(p)$	*c*	*g*	$P_a(p)$	*c*	*g*	$P_a(p)$	*c*	*g*	$P_a(p)$
0.25	2	3	16	0.9663	3	04	0.9603	3	8	0.9663	-	-	-
	4	1	8	0.9923	1	02	0.9905	1	4	0.9923	1	2	0.9636
	6	0	5	0.9673	1	02	0.9983	1	4	0.9987	1	2	0.9932
	8	0	5	0.9827	1	02	0.9995	1	4	0.9996	1	2	0.9980
0.10	2	4	24	0.9635	4	06	0.9547	4	12	0.9635	-	-	-
	4	1	12	0.9832	1	03	0.9788	1	6	0.9832	1	2	0.9636
	6	0	7	0.9545	1	03	0.9962	1	6	0.9970	1	2	0.9936
	8	0	7	0.9757	1	03	0.9989	1	6	0.9992	1	2	0.9936
0.05	2	5	32	0.9640	5	08	0.9534	5	16	0.9640	-	-	-
	4	1	14	0.9775	1	04	0.9636	1	7	0.9775	1	2	0.9636
	6	1	14	0.9960	1	04	0.9932	1	7	0.9960	1	2	0.9932
	8	1	14	0.9989	1	04	0.9980	1	7	0.9989	1	2	0.9980
0.01	2	7	48	0.9687	7	12	0.9564	7	24	0.9687	-	-	-
	4	1	20	0.9568	2	07	0.9841	1	10	0.9568	4	5	0.9773
	6	1	20	0.9920	1	05	0.9895	1	10	0.9920	1	2	0.9850
	8	1	20	0.9977	1	05	0.9969	1	10	0.9977	1	2	0.9950

Table 5. GASP for OGELLD with $\hat{\lambda}=1.1513, \hat{\theta}=1.1606$ and $\hat{\gamma}=2.6538$ for 50th percentile

β	t_q/t_q^0	r=5						r=10					
		δ_q=0.50			δ_q=1.0			δ_q=0.50			δ_q=1.0		
		c	g	$P_a(p)$	c	g	$P_a(p)$	c	g	$P_a(p)$	c	g	$P_a(p)$
0.25	2	2	6	0.9553	-	-	-	2	3	0.9553	-	-	-
	4	0	2	0.9620	1	2	0.9724	0	1	0.9620	2	3	0.9553
	6	0	2	0.9882	0	1	0.9569	0	1	0.9882	1	2	0.9868
	8	0	2	0.9950	0	1	0.9808	0	1	0.9950	0	1	0.9620
0.10	2	3	9	0.9687	-	-	-	3	5	0.9563	-	-	-
	4	1	6	0.9939	1	2	0.9724	1	3	0.9939	2	3	0.9553
	6	0	3	0.9824	0	1	0.9569	0	2	0.9766	1	2	0.9868
	8	0	3	0.9925	0	1	0.9808	0	2	0.9900	0	1	0.9620
0.05	2	4	13	0.9706	-	-	-	4	7	0.9613	-	-	-
	4	1	7	0.9918	1	2	0.9724	1	4	0.9894	2	3	0.9553
	6	0	4	0.9766	0	1	0.9569	0	2	0.9766	1	2	0.9868
	8	0	4	0.9900	0	1	0.9808	0	2	0.9900	0	1	0.9620
0.01	2	5	18	0.9669	-	-	-	5	9	0.9669	-	-	-
	4	1	9	0.9867	2	3	0.9933	1	5	0.9838	2	3	0.9553
	6	0	11	0.9823	0	2	0.9713	0	6	0.9808	0	1	0.9713
	8	0	11	0.9929	0	2	0.9884	0	6	0.9922	0	1	0.9884

Table 6. Comparison between GLLD, OELLD and OGELLD

t_q/t_q^0	GLLD			OELLD			OGELLD		
	c	g	$P_a(p_1)$	c	g	$P_a(p_1)$	c	g	$P_a(p_1)$
2	1	69	0.9959	5	12	0.9587	4	10	0.9513
4	0	41	0.9990	1	5	0.9705	1	5	0.9792
6	0	41	0.9999	1	5	0.9936	1	5	0.9963
8	0	41	0.9999	0	3	0.9602	1	5	0.9989

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 117–129, DOI 10.21307/stattrans-2019-007

FORMULATION OF THE SIMPLE MARKOVIAN MODEL USING FRACTIONAL CALCULUS APPROACH AND ITS APPLICATION TO ANALYSIS OF QUEUE BEHAVIOUR OF SEVERE PATIENTS

Soma Dhar[1], **Lipi B. Mahanta**[2], **Kishore Kumar Das**[3]

ABSTRACT

In this paper, we introduce a fractional order of a simple Markovian model where the arrival rate of the patient is Poisson, $i.e.$ independent of the patient size. Fraction is obtained by replacing the first order time derivative in the difference differential equations which govern the probability law of the process with the Mittag-Leffler function. We derive the probability distribution of the number $N(t)$ of patients suffering from severe disease at an arbitrary time t. We also obtain the mean size (number) of the patients suffering from severe disease waiting for service at any given time t, in the form of $E^{\nu}_{0.5,0.5}(t)$, for different fractional values of server activity status, $\nu = 1, 0.95, 0.90$ and for arrival rates $\alpha = \beta = 0.5$. A numerical example is also evaluated and analysed by using the simple Markovian model with the help of simulation techniques.

Key words: fractional order, arrival rate, patients, fractional calculus.

1 Introduction

From the historical point of view, fractional calculus may be described as an extension of the concept of a derivative operator from integer order n to arbitrary order α, where α is a real or complex value, or even more complicated, a complex valued function,

$$\alpha \quad = \quad \alpha(x,t) \tag{1}$$

Despite the fact that this concept has been discussed since the days of Leibniz (1695) and since then has occupied the great mathematicians of their times, no other research area has resisted as much a direct application for centuries. Abel's treatment of the tautochrone problem from 1823 stood for a long time as a singular example of an application for fractional calculus. Abel (1823) define the equation as

$$\frac{d^n}{dx^n} \longrightarrow \frac{d}{dx} \tag{2}$$

[1]Gauhati University. E-mail: somadhar7@gmail.com.
[2]Institute of Advanced Study in Science and Technology. E-mail: lbmahanta@iasst.gov.in.
ORCID ID: https://orcid.org/0000-0002-7733-5461.
[3]Department of Statistics, Gauhati University, Guwahati, India. E-mail: daskkishore@gmail.com.

Differentiation and integration are usually regarded as discrete operations, in the sense that we differentiate or integrate a function once, twice, or any whole number of times. But in the case of integer order functions, the question is how to differentiate or integrate the same. Fractional calculus is useful to evaluate the integer order function.

Fractional Calculus is a significant topic in mathematical analysis as a result of its increasing range of applications, that grows out of the traditional definition of the integer order calculus of derivatives and integrals. It provides several tools for solving differential and integral equations of fractional order. In the recent years, fractional calculus has played a very important role in various fields, based on the wide applications in engineering and sciences such as physics, mechanics, chemistry, biology, applied mathematics, probability and statistics etc.

The application-oriented approaches of fractional calculus are given in many textbooks. For examples, Oldham(1974), Samko (1993), Miller (1993), Kiryakova (1994), Rubin (1996), Gorenflo (1997), Podlubny (1998), Hilfer (2000), Hilfer (2008), Mainardi (2010), Herrmann (2014). These books are explicitly devoted to the practical consequences of using fractional calculus.

There have been few studies related to point processes governed by difference-differential equations containing fractional derivative operators. These processes are direct generalizations of the classical $M/M/1$ queue and the linear birth-death processes. It is well known that a fractional derivative operator induces a non-Markovian behaviour into a system as derived by Veillette (2010). Srivastava (2001) studied a systematic (and historical) investigations carried out by various authors in the field of fractional calculus and its applications. Srinivasan (2008) has considered a brief elementary and introductory approach to the theory of fractional calculus and its applications especially in developing solutions of certain families of ordinary and partial fractional differential equations.

Moreover, parameter estimation and path generation algorithms of these new fractional stochastic models were derived. It is to be noted that the proposed fractional point models (with Markovian and non-Markovian properties) are parsimonious, which makes them desirable for modelling real-world non-Markovian queuing systems. Dhar (2014) studied the comparison between single and multiple Markovian queuing model in an outpatient department. Also, Mahanta (2016) proposed a single server queueing model for severe diseases especially in outpatient department. Further, consider the infinite server queues with time-varying arrival and departure pattern when the parameters are varying with time derived by Dhar(2017).

It is further observed that more recently fractional point processes driven by fractional difference-differential equations such as the fractional Poisson, the fractional birth, the fractional death, and the fractional birth-death processes have already been gaining attention as studied by Beghin (2009), Cahoy (2010), Laskin (2003), Orsingher (2011).

Recently, Uchaikin (2008), Orsingher (2010), Cahoy (2013) have developed the generalizations of the classical birth and death processes by using the techniques of fractional calculus. A major advantage of these models over their classical coun-

terparts is that they can capture both Markovian and non-Markovian structures of a growing or decreasing system.
Uchaikin (2008) partially investigated the fractional linear birth process by using the Riemann-Liouville derivative operator but it was generalised by Orsingher (2010) using the Caputo derivative. Cahoy (2013) derived the inter-birth time distribution using simulation method to simulate the ${}_fY_p$.
A situation of fractional calculus may be applicable in queuing system when the server is found not working, either from the start or in between. A classic example may be the absence of doctor(s) or his/her leaving the hospital in between for other works, despite patients waiting for treatment.
To date, no practical implementation for any real-life problem has been attempted using the theories as mentioned above. In this paper, an attempt is made to develop a model using the concept of fractional calculus on a queuing system for emergency service of severe patients.
In certain departments, like outpatient department, of many public hospitals, unavailability of doctors during working hours has become a trend these days. These doctors come to their department only at a time convenient for them. The outpatient department of a hospital is visited by patients of all types of disease. Some of these diseases require immediate medical attention as severe complications may arise if treatment is delayed. This delay is commonly due to server inactivity, which may be total or in fractions. By 'fractions' we imply that some portions of the server is active while some is not. Examples may be like, i) doctor is present whereas registration desk personnel is not, or ii) all personnel are present but there is some technical lapse, or iii) registration desk is in order but doctor is not present, and so on. Patients coming from far-off places, postponing their own schedule and engagements, are thus deprived of timely medical services. A system, therefore, must be put in place to make the irregularity of doctors fall in line, so that there is a check on the server system functioning as doctors of these hospitals.

2 Basic Preliminaries

The basic definitions and properties of the fractional calculus theory used in this study are given below.
Definition 1: Let $y = f(t)$ be a continuous (but not necessarily differentiable) function and let partition $h > 0$ in the interval $[0,1]$. Then, the fractional derivative is defined by Podlubny (1998)

$$D^n(f) = \frac{d^n f}{dt^n} = \lim_{h\to 0} \frac{\sum_{j=0}^{n}(-1)^j \binom{n}{j} f(t-jh)}{h^n}$$

If n is fixed then $D^n f \to 0$ as $h \to 0$.
Definition 2: Grunwald Letnikov differential integral of arbitrary order q is defined

by Podlubny(1998)

$$D_a{}^q f(t) = \lim_{N\to\infty} h_N{}^{-q} \left[\sum_{j=0}^{N} (-1)^j \binom{q}{j} f(t - jh_N) \right]$$

where

$$\begin{aligned} \binom{q}{0} &= 1 \\ \binom{q}{j} &= \frac{q(q-1)\dots(q-j+1)}{j!},\ j \in N \end{aligned}$$

lemma 1:

$$\frac{d^n}{dt^n} D_a{}^q f(t) = D_a{}^{n+q} f(t)$$

Definition 3: The Riemann-Liouville fractional integral operator of order $a > 0$ is defined Mathai (2008) as

$$I_a{}^q f(t) = \frac{1}{\Gamma(q)} \int_a^t (t-\tau)^{q-1} f(\tau) d\tau,\ \ t > a$$

Definition 4: The Riemann-Liouville fractional derivative operator of order a is defined [Haubold (2011)] as

$$D_a{}^q f(t) = \frac{d^n}{dt^n} \left[\frac{1}{\Gamma(n-q)} \int_a^t (t-\tau)^{n-q-1} f(\tau) d\tau \right]$$

2.1 Mittag-Leffler function

The Mittag-Leffler function, which plays a very important role in the fractional differential equations was in fact introduced by Mittag-Leffler in 1903. It is a generalization of the exponential series, *i.e.* if $\alpha = 1$ then we have the exponential series. The Mittag-Leffler function $E_a(t)$ is defined by the power series (3)

$$E_a(t) = \sum_{n=0}^{\infty} \frac{t^n}{\Gamma(an+1)},\ a > 0 \tag{3}$$

which gives the generalized Mittag-Leffler function (1.4) as defined

$$E_{\alpha,\beta}(t) = \sum_{n=0}^{\infty} \frac{t^n}{\Gamma(\alpha n+\beta)},\ \alpha,\beta > 0 \tag{4}$$

This generalization was studied by Saxena (2005) and Haubold (2011).

3 Generation of single server queuing model applying fractional concept

Consider a single-server queue with inter-arrival time and service time which are exponentially distributed with rates λ and μ, respectively. Let $N(t)$ be the number of patients in the system at time t. We define

$$p_n(t) \quad = \quad Pr[N(t)=n|N(0)=i], \quad i \geq 0 \tag{5}$$

$M/M/1$ is a special case of the general birth-and-death model with $\lambda_n = \lambda$ and $\mu_n = \mu$.
The generator matrix is given by (state space: $0,1,2,\dots$)

$$M = \begin{bmatrix} -\lambda & \lambda & \dots & 0 \\ \mu & -(\lambda+\mu) & \lambda & \dots & 0 \\ \vdots & & \ddots & \vdots & \end{bmatrix}$$

Then, the governing differential-difference equations of the system under consideration are given by

$$\begin{cases} \frac{\delta p_0(t)}{\delta t} = -\lambda p_0(t) + \mu p_1(t) \\ \frac{\delta p_n(t)}{\delta t} = -(\lambda+\mu)p_n(t) + \mu p_{n+1}(t) + \lambda p_{n-1}(t), n \geq 1 \end{cases} \tag{6}$$

The generator matrix is given by (state space: $0,1,2,\dots$) the matrix below, using the generalized Mittag-Leffler function.

$$E_{\alpha,\beta}(Mt^\alpha) \quad = \quad \sum_{n=0}^{\infty} \frac{(Mt^\alpha)^n}{\Gamma(\alpha n+\beta)} \begin{bmatrix} -\lambda^n & {C_n}^1\lambda^{n-1} & \dots & {C_n}^{n-1}\lambda^{n-k-1} \\ \mu^n & -(\lambda+\mu)^n & {C_n}^1\lambda^{n-1} & \dots & 0 \\ \vdots & & \ddots & \vdots & \end{bmatrix}$$

$$= \quad \sum_{n=0}^{\infty} \frac{(t^\alpha)^n}{\Gamma(\alpha n+\beta)} \begin{bmatrix} -\lambda^n & {C_n}^1\lambda^{n-1} & \dots & {C_n}^{n-1}\lambda^{n-k-1} \\ \mu^n & -(\lambda+\mu)^n & {C_n}^1\lambda^{n-1} & \dots & 0 \\ \vdots & & \ddots & \vdots & \end{bmatrix}$$

$$= \quad \begin{bmatrix} -\sum_{n=0}^{\infty} \frac{(t^\alpha)^n\lambda^n}{\Gamma(\alpha n+\beta)} & \sum_{n=0}^{\infty} \frac{(t^\alpha)^n{C_n}^1\lambda^{n-1}}{\Gamma(\alpha n+\beta)} & \dots & \sum_{n=0}^{\infty} \frac{(t^\alpha)^n{C_n}^{n-1}\lambda^{n-k-1}}{\Gamma(\alpha n+\beta)} \\ \sum_{n=0}^{\infty} \frac{(t^\alpha)^n(\lambda+\mu)^n}{\Gamma(\alpha n+\beta)}\mu^n & -\sum_{n=0}^{\infty} \frac{(t^\alpha)^n{C_n}^1\lambda^{n-1}}{\Gamma(\alpha n+\beta)} & \dots & \\ \vdots & & \ddots & \vdots & \end{bmatrix}$$

$$= \quad \begin{bmatrix} -E_{\alpha,\beta}(t^\alpha\lambda) & \frac{1}{1!}\frac{d}{d\lambda}E_{\alpha,\beta}(t^\alpha\lambda) & \dots & \frac{1}{(k-1)!}(\frac{d}{d\lambda})^{k-1}E_{\alpha,\beta}(t^\alpha\lambda) \\ E_{\alpha,\beta}(t^\alpha\mu) & E_{\alpha,\beta}(t^\alpha(\lambda+\mu)) & \dots & \\ \vdots & & \ddots & \vdots & \end{bmatrix}$$

Here, we assume to satisfy the difference-differential equations for the state probabilities with arrival rate $\lambda > 0$, service rate $\mu > 0$ and $i \geq 0$ initial patients, and we get,

$$\begin{cases} \frac{\delta^{\nu} p_0^{\nu}(t)}{\delta t^{\nu}} = -E_{\alpha,\beta}(t^{\alpha}\lambda)p_0^{\nu}(t) + E_{\alpha,\beta}(t^{\alpha}\mu)p_1^{\nu}(t) \\ \frac{\delta^{\nu} p_n^{\nu}(t)}{\delta t^{\nu}} = -E_{\alpha,\beta}(t^{\alpha}(\lambda+\mu))p_n^{\nu}(t) + E_{\alpha,\beta}(t^{\alpha}\mu)p_{n+1}^{\nu}(t) \end{cases} \tag{7}$$

$$+E_{\alpha,\beta}(t^{\alpha}\lambda)p_{n-1}^{\nu}(t),\ n \geq 1, 0 \leq \nu \leq 1$$

According to Bailey (1954, 1990), $p_n^{\nu}(t)$ is the probability that there are n patients in the queue at time t and the probability generating function is $G^{\nu}(z,t)$, *i.e.*

$$G^{\nu}(z,t) = \sum_{n=0}^{\infty} z^{n\nu} p_n^{\nu}(t),\ |z| \leq 1 \tag{8}$$

and

$$\bar{G}^{\nu}(z,t) = \sum_{n=0}^{\infty} z^{n\nu} \bar{p_n}^{\nu}(t)$$

Multiplying equation (7) by $\sum_{n=0}^{\infty} z^{n\nu}$, $n = 0, 1, 2, \ldots,$ we get

$$\begin{aligned} \sum_{n=0}^{\infty} z^{n\nu} \frac{\delta^{\nu} p_n^{\nu}(t)}{\delta t^{\nu}} &= -E_{\alpha,\beta}(t^{\alpha}(\lambda+\mu)) \sum_{n=0}^{\infty} z^{n\nu} p_n^{\nu}(t) + E_{\alpha,\beta}(t^{\alpha}\mu) \sum_{n=0}^{\infty} z^{n\nu} p_{n+1}^{\nu}(t) \\ &\quad + E_{\alpha,\beta}(t^{\alpha}\lambda) \sum_{n=0}^{\infty} z^{n\nu} p_{n-1}^{\nu}(t) \\ \Longrightarrow \sum_{n=0}^{\infty} z^{n\nu} \frac{\delta^{\nu} p_n^{\nu}(t)}{\delta t^{\nu}} &= -E_{\alpha,\beta}(t^{\alpha}\lambda) \sum_{n=0}^{\infty} z^{n\nu} p_n^{\nu}(t) - E_{\alpha,\beta}(t^{\alpha}\mu) \sum_{n=0}^{\infty} z^{n\nu} p_n^{\nu}(t) \\ &\quad + E_{\alpha,\beta}(t^{\alpha}\mu) \sum_{n=0}^{\infty} z^{n\nu} p_{n+1}^{\nu}(t) + E_{\alpha,\beta}(t^{\alpha}\lambda) \sum_{n=0}^{\infty} z^{n\nu} p_{n-1}^{\nu}(t) \end{aligned}$$

And adding with equation (8), we get

$$\begin{aligned} \frac{d}{dt} G^{\nu}(z,t) &= \left(\frac{\mu}{z^{n\nu}} + E_{\alpha,\beta}(t^{\alpha}\lambda)z^{n\nu} - E_{\alpha,\beta}(t^{\alpha}(\lambda+\mu)) \right) (G^{\nu}(z,t) - p_0^{\nu}(t)) \\ &\quad + \frac{E_{\alpha,\beta}(t^{\alpha}\lambda)(z^{n\nu}-1)}{z^{n\nu}} p_0^{\nu}(t) \end{aligned} \tag{9}$$

Applying the Laplace transformation

$$\bar{G}^{\nu}(z,s) = \int_0^{\infty} e^{-st} p_n{}^{\nu}(z,t)\ dt$$

in equation (9), we get

$$\left(s^{\nu}\bar{G}^{\nu}(z,s)-s^{\nu-1}G^{\nu}(z,0)\right) = \left(\frac{E_{\alpha,\beta}(t^{\alpha}\mu)}{z^{n\nu}}+E_{\alpha,\beta}(t^{\alpha}\lambda)z^{n\nu}-E_{\alpha,\beta}(t^{\alpha}(\lambda+\mu))\right)\left(\bar{G}^{\nu}(z,s)-\bar{p_0}^{\nu}(s)\right)$$

where $\bar{p_0}^{\nu}(s)=\int_0^{\infty}e^{-st}p_0{}^{\nu}(t)\,dt$
After simplification, we get
$s^{\nu}\bar{G}^{\nu}(z,s)-s^{\nu-1}G^{\nu}(z,0)$

$$= \left(\frac{E_{\alpha,\beta}(t^{\alpha}\mu)}{z^{n\nu}}+E_{\alpha,\beta}(t^{\alpha}\lambda)z^{n\nu}-E_{\alpha,\beta}(t^{\alpha}\lambda)\right)\bar{G}^{\nu}(z,s)-\left(\frac{E_{\alpha,\beta}(t^{\alpha}\mu)}{z^{n\nu}}+E_{\alpha,\beta}(t^{\alpha}\lambda)z^{n\nu}-E_{\alpha,\beta}(t^{\alpha}\lambda)\right)\bar{p_0}^{\nu}(s)$$

$$\left\{s^{\nu}-\left(\tfrac{E_{\alpha,\beta}(t^{\alpha}\mu)}{z^{n\nu}}+E_{\alpha,\beta}(t^{\alpha}\lambda)z^{n\nu}-E_{\alpha,\beta}(t^{\alpha}\lambda)\right)\right\}\bar{G}^{\nu}(z,s)$$

$$= s^{\nu-1}z^{n\nu i+1}-\left(\frac{E_{\alpha,\beta}(t^{\alpha}\mu)}{z^{n\nu}}+E_{\alpha,\beta}(t^{\alpha}\lambda)z^{n\nu}-E_{\alpha,\beta}(t^{\alpha}\lambda)\right)p_0{}^{\nu}$$

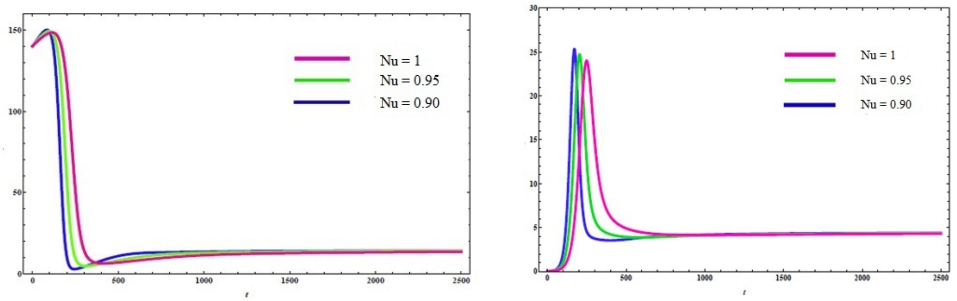
$$\Longrightarrow \bar{G}^{\nu}(z,s) = \frac{s^{\nu-1}z^{n\nu i+1}-\left(\frac{E_{\alpha,\beta}(t^{\alpha}\mu)}{z^{n\nu}}+E_{\alpha,\beta}(t^{\alpha}\lambda)z^{n\nu}-E_{\alpha,\beta}(t^{\alpha}\lambda)\right)p_0{}^{\nu}}{s^{\nu}-\left(\frac{E_{\alpha,\beta}(t^{\alpha}\mu)}{z^{n\nu}}+E_{\alpha,\beta}(t^{\alpha}\lambda)z^{n\nu}-E_{\alpha,\beta}(t^{\alpha}\lambda)\right)}$$

Now, $\bar{G}^{\nu}(z,s)$ converges in the region $|z|\leq 1$, the zero of the numerator and denominator of $\bar{G}^{\nu}(z,s)$ must coincide.

4 Numerical Example

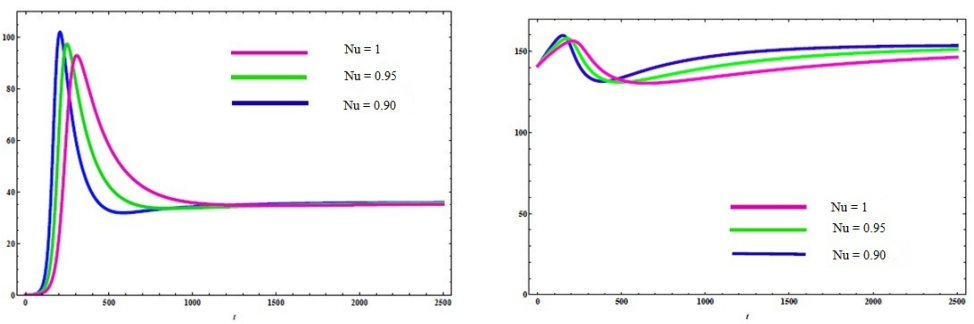
The results obtained below are implemented for estimating the number of arrivals of patients with severe diseases from different departments of public hospital under the assumption mentioned above.
For the numerical solutions of a system of fractional differential equations we use the real data sets, such as, a) the patients waiting time; b) the service time; c) number of patients with severe disease. The data has been collected directly from a public hospital by using the direct observational method. These data (collected for 500 patients) contains all the relevant information regarding each patient.
The simulated solution of the mean size of the arrival patients, the expected service rate, queue size and total patients in the system over the time are displayed in Figures (1)-(4) for $\nu=1,0.95,0.90$ and $\alpha=\beta=0.5$. ν represents server activity status. When $\nu=1$, it means the server is completely (100%) active. $\nu=0.95$ and 0.90 implies 95% and 90% of the server is active respectively. Further, we denote the

rate of arrival of patients who belong to non-severe and severe category as α and β.



(a) (b)

Figure 1: (a) The mean size of the arrival of patients during the time for different values of $\nu = 1, 0.95, 0.90$ and $\alpha = \beta = 0.5$ ($i.e.\ E^{\nu}_{0.5,0.5}(t)$). (b) The expected service time for different values of ν.



(a) (b)

Figure 2: (a) The mean queue size of the arrival of patients during the time for different values of $\nu = 1, 0.95, 0.90$ and $\alpha = \beta = 0.5$ ($i.e.\ E^{\nu}_{0.5,0.5}(t)$). (b) The total patients in the system over the time for different values of ν.

Figure (1) represents the mean size of the arrival of patients during time for different values of ν. Here, X-axis denotes the time in minutes and the number of arrivals of patients suffering from severe disease is represented by the Y-axis. Further, curves are derived by taking different values of $\nu = 1, 0.95, 0.90$ and $\alpha = \beta = 0.5$ ($i.e.\ E^{\nu}_{0.5,0.5}(t)$) and it was observed that curves are upward increasing at a particular point of time and then start to decreases. After decreasing to a certain point of time, curves run parallel to time axis. Also, the graph reveals that the mean size of the arrival of the patient in the hospital from 0 to 120 minutes is high at $\nu = 0.90$ as compared to the value of $\nu = 0.95$ and $\nu = 0.90$.
Figure (2) depicts the relation between the expected service time and the number

of patients who are in queue for getting service under the different values of ν. It can be observed that the service time corresponding to $\nu = 1$ is largest followed by that of $\nu = 0.95$ and $\nu = 0.90$, which also shows that service time increases to the other value of ν and slowly decreases at a point of time.
Figure(3) shows the mean queue size of the patients suffering from severe disease during the time for different values of ν. Here, it is observed that the queue size is stable for all values of ν at a fixed point of time.
Figure (4) defines the total number of patients in the system over the time and shows that each curve peaks at a certain point of time and then starts downward slope. And if $\nu = 1$, the total number of patients remaining in the system is lesser than the value of $\nu = 0.95$ and $\nu = 0.90$.
Table(1) and Table(2) below shows the values of the properties of the queue when implemented to the real-life data.
As revealed from the tables below the pattern of mean arrival of patients, expected

Table 1: The mean arrival of patients and expected service times of $E^{\nu}_{0.5,0.5}(t)$ for different values of ν at different time periods when $\alpha = \beta = 0.5$

Time	Mean arrival of patients			expected service time		
	ν			ν		
	1	0.95	0.90	1	0.95	0.90
0	28.517 ≈ 29	28.266 ≈ 28	28.015≈ 28	0.00	0.00	0.00
50	25.416≈ 25	25.290≈ 25	25.164≈ 25	2.24	4.42	3.49
150	23.021≈ 23	22.937≈ 23	22.85≈ 23	1.64	2.10	1.97
200	21.119≈ 21	21.057≈ 21	20.994≈ 21	1.36	1.58	1.53
250	19.566≈ 20	19.516≈ 20	19.46≈ 19	1.18	1.32	1.29
300	18.268≈ 18	18.226≈ 18	18.184≈ 18	1.06	1.16	1.14
350	17.163≈ 17	17.128≈ 17	17.092≈ 17	0.97	1.04	1.03
400	16.209≈ 16	16.178≈ 16	16.146≈ 16	0.90	0.96	0.95
450	15.375≈ 15	15.347≈ 15	15.319≈ 15	0.84	0.89	0.88
500	14.638≈ 15	14.613≈ 15	14.588≈ 15	0.80	0.85	0.86

service time, mean queue size and the total number of patients in the queue at time t conforms to the findings of the simulated data. Here, it is observed that the mean queue size of the developed queue model is not equal to zero as per the simulated results because in real-life patients arrive at the system substantially before the service starts. Henceforth, all values decrease with time and reach a cusp at $t = 450$.

Table 2: The mean queue size and total number of patients $E^{\nu}_{0.5,0.5}(t)$ for different values of ν at different time periods when $\alpha = \beta = 0.5$

Time	Mean queue size			Total number of patients		
	ν			ν		
	1	0.95	0.90	1	0.95	0.90
0	40.296≈ 40	40.045≈ 40	39.793≈ 40	39.642≈ 40	39.591≈ 40	39.551≈ 40
50	34.532≈ 35	34.406≈ 34	34.280≈ 34	34.205≈ 34	34.180≈ 34	34.159≈ 34
100	30.459≈ 30	30.375≈ 30	30.292≈ 30	30.241≈ 30	30.224≈ 30	30.211≈ 30
150	27.402≈ 27	27.339≈ 27	27.277≈ 27	27.239≈ 27	27.226≈ 27	27.216≈ 27
200	25.005≈ 25	24.955≈ 25	24.904≈ 25	24.874≈ 25	24.864≈ 25	24.856≈ 25
250	23.063≈ 23	23.021≈ 234	22.979≈ 23	22.954≈ 23	22.945≈ 23	22.939≈ 23
300	21.451≈ 21	21.415≈ 21	21.379≈ 21	21.357≈ 21	21.350≈ 21	21.344≈ 21
350	20.086≈ 20	20.055≈ 20	20.023≈ 20	20.004≈ 20	19.998≈ 20	19.993≈ 20
400	18.913≈ 19	18.885≈ 19	18.857≈ 19	18.841≈ 19	18.835≈ 19	18.831≈ 19
450	16.195≈ 16	16.174≈ 16	15.153≈ 15	16.140≈ 16	16.136≈ 16	16.133≈ 16
500	15.579≈ 16	15.560≈ 16	15.441≈ 15	15.729≈ 16	15.625≈ 16	15.522≈ 16

5 Conclusion

The subject of fractional calculus is as old as differential calculus, but remains unexplored outside its theoretical bounds. Here, we attempt to apply that concept to queueing theory and develop its properties on the simple Markovian model. The resultant characteristics of the proposed model are implemented both with simulated and real-life values. It is revealed that the concept put forward conforms to both and fits very well into the theory of queues, particularly when the server is not found to function as it should.

Acknowledgement

We are sincerely thankful to UGC-BSR Scheme, Government of India for granting us the financial assistance to carry out this research work and also thankful to the anonymous reviewers for their valuable comments and suggestions, which helped to improve this paper.

REFERENCES

ABEL, N. H., (1823). Solution de quelques problemesa laide dintegrales definies. Mag. Naturvidenskaberne, 2, pp. 63–68.

BAILEY, N. T., (1954). Queueing for medical care. Applied Statistics, pp. 137–145.

BAILEY, N. T., (1990). The elements of stochastic processes with applications to the natural sciences, volume 25. John Wiley & Sons.

BEGHIN, L., ORSINGHER, E., (2009). Fractional poisson processes and related planar random motions. Electronic Journal of Probability, 14 (61), pp. 1790–1826.

CAHOY, D. O., POLITO, F., PHOHA, V., (2013). Transient behavior of fractional queues and related processes. Methodology and Computing in Applied Probability, pp. 1–21.

CAHOY, D. O., UCHAIKIN, V. V., WOYCZYNSKI, W. A., (2010). Parameter estimation for fractional poisson processes. Journal of Statistical Planning and Inference, 140 (11), pp. 3106–3120.

DHAR, S, DAS, K. K., MAHANTA, L. B., (2014). Comparative study of waiting and service costs of single and multiple server system: A case study on an outpatient department. International Journal of Scientific Footprints, 3 (2), pp. 18–30.

DHAR, S., DAS, K. K., MAHANTA, L. B., (2017). An infinite server queueing model with varying arrival and departures rates for health care system.International Journal of Pure and Applied Mathematics, 113 (5), pp. 583–593.

GORENO, R., MAINARDI, F., (1997). Fractional calculus. Springer.

HAUBOLD, H. J., MATHAI, A. M., SAXENA, R. K., (2011). Mittag-Leffler functions and their applications. Journal of Applied Mathematics, 2011.

HERRMANN, R., (2014). Fractional calculus: an introduction for physicists. World Scientific.

HILFER, R., (2000). Applications of fractional calculus in physics. World Scientific.

HILFER, R., et al., (2008). Threefold introduction to fractional derivatives. Anomalous transport: Foundations and applications, pp. 17–73.

KIRYAKOVA, V., (1994). Generalized fractional calculus and applications longman (pitman res. notes in math. ser. 301).

LASKIN, N., (2003). Fractional poisson process. Communications in Nonlinear Science and Numerical Simulation, 8 (3), pp. 201–213.

MAHANTA, L. B., DAS, K. K., DHAR, S., (2016). A queuing model for dealing with patients with severe disease. Electronic Journal of Applied Statistical Analysis, 9 (2), pp. 362–370.

MAINARDI, F., (2010). Fractional calculus and waves in linear viscoelasticity: an introduction to mathematical models. World Scientific.

MATHAI, A. M., HAUBOLD, H. J., (2008). Special functions for applied scientists, Vol. 4. Springer.

MILLER, K. S., ROSS, B., (1993). An introduction to the fractional calculus and fractional differential equations.

OLDHAM, K., SPANIER, J., (1974). The fractional calculus. 1974.

ORSINGHER, E., POLITO, F., et al., (2011). On a fractional linear birth-death process. Bernoulli, 17 (1), pp. 114–137.

ORSINGHER, E., POLITO, F., SAKHNO, L., (2010). Fractional non-linear, linear and sublinear death processes. Journal of Statistical Physics, 141 (1), pp. 68–93.

PODLUBNY, I., (1998). Fractional differential equations: an introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications, Vol. 198. Academic press.

RUBIN, B., (1996). Fractional integrals and potentials, pitman monogr. Surv. Pure Appl. Math, 82.

SAMKO, S. G., KILBAS, A. A., MARICHEV, O. I., et al., (1993). Fractional integrals and derivatives. Theory and Applications, Gordon and Breach, Yverdon, 1993.

SAXENA, R., SAIGO, M., (2005). Certain properties of fractional calculus operators associated with generalized Mittag-Leffer function. Fractional calculus and applied analysis, 8 (2), pp. 141–154.

SRINIVASAN, A. V., (2008). Managing a modern hospital. SAGE Publications, India.

SRIVASTAVA, H. M., SAXENA, R. K., (2001). Operators of fractional integration and their applications. Applied Mathematics and Computation, 118 (1), pp. 1–52.

UCHAIKIN, V. V., CAHOY, D. O., SIBATOV, R. T., (2008). Fractional processes: from poisson to branching one. International Journal of Bifurcation and Chaos, 18 (09), pp. 2717–2725.

VEILLETTE, M., TAQQU, M. S., (2010). Numerical computation of first-passage times of increasing Levy processes. Methodology and Computing in Applied Probability, 12 (4), pp. 695–729.

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 131–151, DOI 10.21307/stattrans-2019-008

AN APPLICATION OF FUNCTIONAL DATA ANALYSIS TO LOCAL DAMAGE DETECTION

Jacek Leśkow[1], **Maria Skupień**[2]

ABSTRACT

Vibration signals sampled with a high frequency constitute a basic source of information about machine behaviour. Few minutes of signal observations easily translate into several millions of data points to be processed with the purpose of the damage detection. Big dimensionality of data sets creates serious difficulties with detection of frequencies specific for a particular local damage. In view of that, traditional spectral analysis tools like spectrograms should be improved to efficiently identify the frequency bands where the impulsivity is most marked (the so-called *informative frequency bands* or IFB). We propose the functional approach known in modern time series analysis to overcome these difficulties. We will process data sets as collections of random functions to apply techniques of the functional data analysis. As a result, we will be able to represent massive data sets through few real-valued functions and corresponding parameters, which are the eigenfunctions and eigenvalues of the covariance operator describing the signal. We will also propose a new technique based on the bootstrap resampling to choose the optimal dimension in representing big data sets that we process. Using real data generated by a gearbox and a wheel bearings we will show how these techniques work in practice.

Key words: damage detection, functional data, functional principal components, informative frequency band.

1. Introduction

In recent years, extensive research has been focused on big data problems related to statistical signal processing. The big data problem arises when a structural health monitoring system is supported by on-line sensors producing a signal observed with e.g. 20 kHz frequency. After several hours of observations we have millions of data points that can be used for processing. So far, many practical applications have been based on selecting some segments of data and classical analyses have then been conducted on selected segments. However, modern statistical inference can be based on the whole multi-million points sample when the functional data analysis approach is used (see, for example Horváth and Kokoszka, 2012). This is especially suitable when we deal with time-varying systems and when techniques related to time-frequency analysis are used. For example, in Yang and Nagarajaiah (2014), results are presented on independent component analysis with

[1]Cracow University of Technology, Poland. E-mail: jleskow@pk.edu.pl.
[2]Pedagogical University of Cracow, Poland. E-mail: marysia.skupien@gmail.com. ORCID ID: https://orcid.org/0000-0003-1480-0810.

wavelet transform. Interesting exploratory studies on frequency response function (FRF) can be found in Staszewski and Wallace (2014).

In the case of big data and time varying systems, it is convenient to consider data as curves. For example, for a rotating element the natural data curve would be generated in the interval of the length of a cycle. For other signals, a natural interval may be a second or a minute. From this perspective, the spectrogram (see, e.g. Gryllias, et al., 2017 or Khadersab and Shivakumar, 2018) is a way of converting a big signal into segments via Fourier analysis on sliding blocks creating a time-frequency map. Such a map is exactly a collection of random curves. From this point of view, a multimillion data points segment of a signal generated by a sensor attached to some structure is seen as a collection of curves.

In recent years there is a significant research dedicated to wheel bearing diagnostics. Randall and Antoni (2011) presents a review of contemporary techniques. Other publications like (Liu, et al., 2018) or (Jia, et al., 2016) present contemporary artificial intelligence technique and their applications in the wheel bearing diagnosis. However, according to our knowledge, so far no one has implemented modern statistical inference tools, based on functional data approach, to the diagnosis of wheel bearings. In this context we would like to mention research in Spiridonakos and Fassois (2014) dedicated to functional time series and their applications to non-stationary random vibrations, where a functional approach is proposed in a different context.

The main line of our article is to show how to use modern statistical tools like functional data analysis or bootstrap to efficiently process big data sets and identify significant frequencies. We propose a new perspective in looking at a very popular tool in signal analysis such as the spectrogram. In Subsection 3.1. we explain the difference between classical spectrogram and new functional one proposed by us. Usually, the spectrogram is generated by a signal coming from an excavating machine or by a signal generated by a wheel bearing (see, for example Cioch, et al. 2013). Then, the proper functioning of the tested system is diagnosed by identifying the frequency band where the signal impulsivity is most marked, called in the sequel informative frequency bands (IFB) (see Randall and Antoni, 2011 or Obuchowski, et al., 2014).

We propose to view the spectrogram as a collection of random curves. Using the functional data analysis approach, we are able to process such data and quickly solve the problem of identifying the IFB. The main advantage of our method is a possibility of using large data sets to generate few dimension of the diagnostic analysis. Millions of data points are represented as random curves, then in the appropriately defined infinitely dimensional Hilbert space the covariance operator is considered and its empirical counterpart is studied. Finally, only few eigenvalues and eigenfunctions of the empirical covariance operator are sufficient to efficiently represent the signal at hand. We propose a novel technique of using bootstrap resampling in deciding on dimensionality reduction. Another main advantage of our method is that no matter what frequency of signal sampling might be, our method provides a uniform result. On the diagram in Figure 1 we show the logic of our

approach.

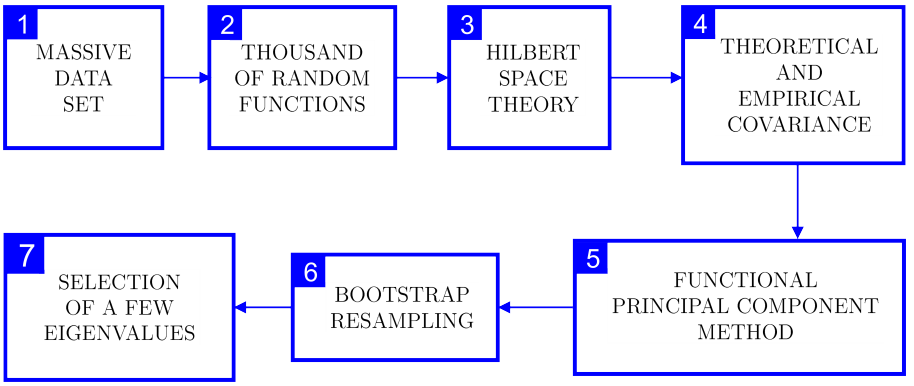


Figure 1: Flowchart of the main idea of our paper.

Our article is organized as follows. In Section 2, we present main elements of the functional data analysis approach as applied to signals. Basic tools such as functional principal components, Hilbert space valued random transformations, variance and covariance operators are presented there. In Section 3, we apply this approach to the problem of identifying the informative frequency bands for a spectrogram. In that context we show that without discarding any data points we are able to reduce the dimensionality of data to just a few eigenvalues and few eigenfunctions and retain more than 80% of its energy. Finally, in Section 4 we provide a short discussion of our results.

2. Functional data approach in statistical signal processing

Our starting point here is the new perspective on statistical signal processing from the functional data analysis point of view. To start, assume that we observe a signal $\{X(s) : s \in [0,T]\}$ and the collect time T is really huge, e.g. in the order of several millions of individual data points. We will view such a signal as a collection of random curves $\{X_n(s), s \in [n, n+w]\}$ each defined on the interval $[n, n+w]$ with the width w. These curves may be considered independent or correlated, depending on the model and a context of study. For example, the spectrogram technique Gryllias, et al., (2017), Khadersab and Shivakumar, (2018) transforms a long signal $\{X(s) : s \in [0,T]\}$ into a collection $\{x_1(f), \ldots, x_N(f)\}$ of spectral densities defined on a common frequency interval $[0,\Lambda]$, where $f \in [0,\Lambda]$. In the Subsection 2.3, in the last algorithm, we explain what our functional observations are and how they were obtained from a discrete vibration signal. We assume those observations to be independent. We are aware that the technic of the overlapping window may introduce some dependence into our data structure. At this point we neglect this dependence and we proceed as if the data were independent. However, in the

literature there is a number of cases in which methods of functional data analysis have been adapted to series or signals by cutting them as if they were curves observed independently (see Ramsay and Silverman (2002) and (2005)).

To simplify our notation and with no loss of generality, we will assume that $\Lambda = 1$ so all data curves are defined on the unit interval $[0,1]$. The observed curves will be assumed to be square integrable. A natural choice of the realization space will therefore be the Hilbert space $H = L^2[0,1]$. This is a consequence of the expansion methods (see Ramsay and Silverman (2002) and (2005)), where the functional form of curves is obtained by a linear span of base functions. The relevant coefficients are then estimated from discrete observation of curves at different time points by least squares methods. This point of view allows us to introduce a Hilbert space of squared integrable functions, where the theory of functional principal component analysis (and many other functional methods) can be applied. From this perspective our initial signal $\{X(s) : s \in [0,T]\}$ can be viewed as a collection of random curves $\{X_n\}$, each in the space H. For such random curves, we will now introduce concepts of mean, variance and covariance.

Note that each random curve X is as a random element acting from some probability space $(\Omega, \mathscr{F}, P)$ onto $L^2[0,1]$. If X is integrable, then there is a unique function $\mu \in L^2$ such that $\mathbb{E}\langle y, X\rangle = \langle y, \mu\rangle$ for each $y \in L^2$. It follows that $\mu(t) = \mathbb{E}[X(t)]$ for all $t \in [0,1]$. Here $\langle \cdot, \cdot \rangle$ is the scalar product in the Hilbert space H defined as $\langle x, y\rangle = \int_0^1 x(s)y(s)ds$ with the norm defined as: $\|f\| = \sqrt{\int_0^1 f^2(t)\, dt}$ for all $f \in H$. For more mathematical details regarding statistics on Hilbert space the reader is referred to Horváth and Kokoszka (2012).

We recall here the notion of spectral decomposition for matrixes and functional operator.

Theorem: 1. *Suppose $\boldsymbol{A}$ is a symmetric, positive definited $k \times k$ matrix. Then, there is an orthonormal matrix $\boldsymbol{U} = [\boldsymbol{u_1}, \ldots, \boldsymbol{u_k}]$ whose columns are the eigenvectors of $\boldsymbol{A}$, i.e.*

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} \qquad \text{and} \qquad \boldsymbol{A}\boldsymbol{u_j} = \lambda_j \boldsymbol{u_j}$$

Moreover, $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U} = Diag[\lambda_1, \ldots, \lambda_k]$ The orthonormality of $\boldsymbol{U}$ is equivalent to the assertion that the vectors $\boldsymbol{u_1}, \ldots, \boldsymbol{u_k}$ form an orthonormal basis in the Euclidean space $\mathbb{R}^k$. Theorem 1 implies that

$$\underset{(k\times k)}{\boldsymbol{A}} = \sum_{i=1}^{k} \lambda_i \underset{(k\times 1)(1\times k)}{\boldsymbol{u_i}\ \boldsymbol{u_i}^T} = \underset{(k\times k)}{\boldsymbol{U}}\ \underset{(k\times k)}{\boldsymbol{\Lambda}}\ \underset{(k\times k)}{\boldsymbol{U}^T},$$

a representation known as a spectral decomposition of $\boldsymbol{A}$.

The above ideas can be easily extended to a separable Hilbert space. Suppose $\boldsymbol{\Psi}$ is a symmetric positive–definite Hilbert–Schmidt operator in L^2. Covariance operator (1) and its sample counterpart (2) are in this class, provided $\mathbb{E}\|X\|^4 < \infty$. The

operator $\boldsymbol{\Psi}$ then admits the functional counterpart of spectral decomposition (1)

$$\langle \boldsymbol{\Psi}(x), x \rangle = \left\langle \sum_{i=j}^{\infty} \lambda_j \langle x, v_j \rangle v_j, x \right\rangle = \sum_{i=j}^{\infty} \lambda_j \langle x, v_j \rangle^2$$

where scalars λ_j are eigenvalues and v_j corresponding eigenfunctions, satisfying equation $\boldsymbol{\Psi}(v_j) = \lambda_j v_j$.

2.1. Theoretical and empirical covariance operators

Since we are adopting the Hilbert space approach, the usual covariance will be an operator, that is a transformation from the Hilbert space to the Hilbert space. This is analogous to the traditional concept of covariance, where a real-valued signal X generates a covariance function transforming real values to real values. Let us have a closer look at the formal definition of the covariance operator.

For X integrable and $\mathbb{E}X = 0$, the covariance operator of X is defined by

$$C(x) = \mathbb{E}[\langle X, x \rangle X], \quad x \in L^2, \tag{1}$$

where

$$\begin{aligned} C(x)(t) =& \mathbb{E}[\langle X, x \rangle X(t)] = \mathbb{E} \int_0^1 X(s)x(s)ds X(t) = \\ =& \int_0^1 \underbrace{\mathbb{E}[X(s)X(t)]}_{=c(s,t)} x(s)ds = \int_0^1 c(s,t)x(s)ds. \end{aligned}$$

In the sequel, the covariance operator C will be our central point of a study as it fully describes the energy generated by the random element X, which in turn represents a signal under study. While studying the covariance operator, we will focus on characterizing its eigenvalues. They will be important in reducing the dimensionality of C to just a few of non-negative numbers. For more theoretical properties of the covariance operators see Horváth and Kokoszka (2012).

The main task of statistical signal processing in the functional data analysis context will be to introduce an empirical covariance operator $\hat{C}$ that is fully defined by random curves $x_1, \ldots, x_N$ and for sufficiently large sample size N approximates the theoretical covariance operator C that describes the signal of interest. Therefore, assume that a sample of random functions $x_1, \ldots, x_N$ corresponds to the signal X. Recall that the spectrogram can be viewed as a collection of random curves with arguments in the frequency interval. In general, however, such random functions can represent segments of signals from different time intervals or replica of signals collected via some transformations.

For $x_1,\ldots,x_N$ we define the sample covariance operator as:

$$\hat{C}(x) = \frac{1}{N}\sum_{i=1}^{N}\langle x_i, x\rangle x_i,\ x \in H. \tag{2}$$

It is important to note that in the formula (2) the symbol x corresponds to any function x from the Hilbert space H while x_i is the *observed* random function x_i generated by the signal of interest. In such a way the estimator $\hat{C}$ given in (2) approximates the theoretical covariance operator C defined in (1). For more mathematical theory related to this approximation the reader is refereed to (Bosq, 2000).

For covariance operators which are symmetric, positive defined and are defined on a Hilbert space and are Hilbert-Schmidt operators we have a very interesting property. Suppose $\mathbf{\Psi}$ is a symmetric, positive definite Hilbert–Schmidt operator with eigenfunctions v_j and eigenvalues λ_j, satisfying $\lambda_1 > \lambda_2 > \cdots$. Then,

$$\sup_{\|x\|=1}\{\langle \mathbf{\Psi}(x), x\rangle : \langle x, v_j\rangle = 0,\ 1 \le j \le i-1, i < p\} = \lambda_i$$

and the supremum is reached if $x = v_i$. The maximizing function x is unique up to a sign. The upper bound p for index i is defined in Subsection 2.2.

Of course, the empirical covariance operator $\hat{C}$ given in (2) satisfies the above property. This in turn gives us the following facts fundamentally important in the subsequent statistical considerations:

- the empirical covariance operator $\hat{C}$ defined in (2) is fully defined by its eigenfunctions and eigenvalues,
- the covariance and the total variance of the sample (thus the signal) will be described by the estimated eigenvalues.

In what follows, we will show how to apply these facts. To start, recall that the random functions $x_1,\ldots,x_N$ correspond to the signal of interest. Now, fix the integer number $p \ll N$. Next, choose the basis $u_1, u_2, \ldots$ in H such that:

$$\hat{S}^2 = \sum_{i=1}^{N}\left\| x_i - \sum_{k=1}^{p}\langle x_i, u_k\rangle u_k \right\|^2 \leftarrow \min$$

Then, each curve x_i can be approximated by $\sum_{k=1}^{p}\langle x_i, u_k\rangle u_k = \sum_{k=1}^{p} c_k u_k$. Hence, an infinite dimensional curve x_i is represented by a p variate vector $(\langle x_i, u_1\rangle, \ldots, \langle x_i, u_p\rangle)$. Now we will use the fundamental fact that the basis elements $u_1,\ldots,u_p$ can be chosen to correspond to the eigenfunctions of the sample covariance $\hat{C}$. More precisely, functions $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_p$ minimizing $\hat{S}^2$ are equal (up to a sign) to normalized eigenfunctions of the sample covariance operator $\hat{C}$.

Note that scores $\langle x, \hat{u}_k\rangle = \int_0^1 x(t)\hat{u}_k(t)dt$ measure the importance of the kth function $\hat{u}_k$ in the representation

$$x \approx \sum_{k=1}^{p} \langle x, \hat{u}_k \rangle \hat{u}_k. \tag{3}$$

In the sequel, we will call $\hat{u}_k$ the $k-th$ *functional principal component*. One of its important properties is orthonormality.

The eigenvalues are extremely important in describing the total energy of the signal. For a random element X with values in the Hilbert space H we have:

$$\mathbb{E}\|X\|^2 = \sum_{j=1}^{\infty} \mathbb{E}\langle X, v_j \rangle^2 = \sum_{j=1}^{\infty} \langle C(v_j), v_j \rangle = \sum_{j=1}^{\infty} \lambda_j.$$

The quantity $\mathbb{E}\|X\|^2$ can be called *theoretical total variance*. Its empirical equivalent, *sample total variance*, based on a sample of random functions $x_1, \ldots, x_N$ is defined as:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\|x_i\|^2 &= \frac{1}{N}\sum_{i=1}^{N}\langle x_i, x_i\rangle = \frac{1}{N}\sum_{i=1}^{N}\Big\langle \sum_{j=1}^{N}\langle x_i, \hat{u}_j\rangle \hat{u}_j, \sum_{j=1}^{N}\langle x_i, \hat{u}_j\rangle \hat{u}_j \Big\rangle = \\ &= \sum_{j=1}^{N}\frac{1}{N}\sum_{i=1}^{N}\langle x_i, \hat{u}_j\rangle^2 = \sum_{j=1}^{N}\langle \hat{C}(\hat{u}_j), \hat{u}_j\rangle = \sum_{j=1}^{N}\sum_{k=1}^{\infty}\hat{\lambda}_k \underbrace{\langle \hat{u}_j, \hat{u}_k\rangle^2}_{=\delta_{jk}} = \sum_{j=1}^{N}\hat{\lambda}_j,\end{aligned}$$

where δ_{jk} is the Kronecker delta - a function of two variables, defined as follows: $\delta_{jk} = \delta(j,k) = \begin{cases} 1, & \text{if} \quad k = j \\ 0, & \text{if} \quad k \neq j \end{cases}$ and $\hat{\lambda}_j$ is interpreted as variance in the direction $\hat{u}_j$. In other words, the empirical functional principal component $\hat{u}_j$ generated by the empirical covariance operator $\hat{C}$ explains the fraction of the total sample variance equal to $\hat{\lambda}_j / \sum_{k=1}^{N} \hat{\lambda}_k$. The above approach will be referred to as the Functional Principal Component Analysis or FPCA for short.

2.2. Reduction of dimensionality

While working with big data sets generated by signals, the crucial point is to select the number p of eigenvalues that give a reasonable approximation of the sample total variance. One of the methods of selecting p, for which function x_i has the best approximation given by the formula $\sum_{j=1}^{p} \langle x_i, \hat{u}_j \rangle \hat{u}_j$ is the CPV method. This method is based on calculating the cumulative percentage of the total variance

(CPV) explained by the first p empirical functional principal components

$$CPV(p) = \frac{\sum_{i=1}^{p} \hat{\lambda}_i}{\sum_{i=1}^{N} \hat{\lambda}_i}. \tag{4}$$

We choose p for which $CPV(p)$ exceeds a desired level. Ideally, one would like to recover 100% of the total variance, however in practical situations we usually settle with 80% or higher. Such approach has a dramatic effect on our ability to process big data generated by the signals observed with high frequency over long periods of time. First, we split the signal into a sequence of random functions and then we follow the approach above to identify the first p eigenvalues. In the following section we will show that for vibration data coming from the gearbox of the excavating machine choosing p as small as 8 retrieves a large percentage of the total variance.
Below, we present our original method, based on bootstrap technique, which allows us to precisely evaluate the percentage of variance explained with a confidence interval.

CPV bootstrap algorithm.

Step 1. We start from the initial sample of random functions $x_1,\ldots,x_N$. We sample with replacement the first bootstrap sample $x_1^{*1},\ldots,x_N^{*1}$ from the initial set $x_1,\ldots,x_N$. It is important that the bootstrap sample is of the same size as the original one. For such bootstrap sample we calculate the first bootstrap value $CPV(p)^{*1}$ of $CPV(p)$ (see formula (4)).

Step 2. We repeat Step 1 B times. Usually, we take B = 1000. As a result, we get B bootstrap replications $\{CPV(p)^{*1},\ldots,CPV(p)^{*B}\}$.

Step 3. We produce a 95% confidence interval for $CPV(p)$ using 2.5% and 97.5% empirical quantiles from the replications $\{CPV(p)^{*1},\ldots,CPV(p)^{*B}\}$.

We illustrate the logic of our bootstrap procedure on the diagram in Figure 2.
The below procedure is admissible from the statistical point of view as it is reconstructing the true unknown distribution of the $CPV(p)$, which in turn is based on the unknown distribution of eigenvalues. For a more detailed discussion related to eigenvalues distribution in the functional approach see e.g. (Mas, 2002). We would like to emphasize that our method allows us to analyse big data sets generated by signals observed over a long period of time using just 8 eigenvalues and 8 associated eigenfunctions. In general, one can start with even 2 eigenvalues, calculate the $CPV(3)/CPV(2)$ and bootstrap it to get its confidence intervals and then see whether adding third eigenvalue significantly improves CPV(3) as compared to

CPV(2). The reader can find an explanation of this improvement in Subsection 3.2 on the example of analysed data set. Additional argument is provided by studying the convergence of the ratio $CPV(p+1)/CPV(p)$ and identifying the proper p, where it starts to stabilize (Figure 3).

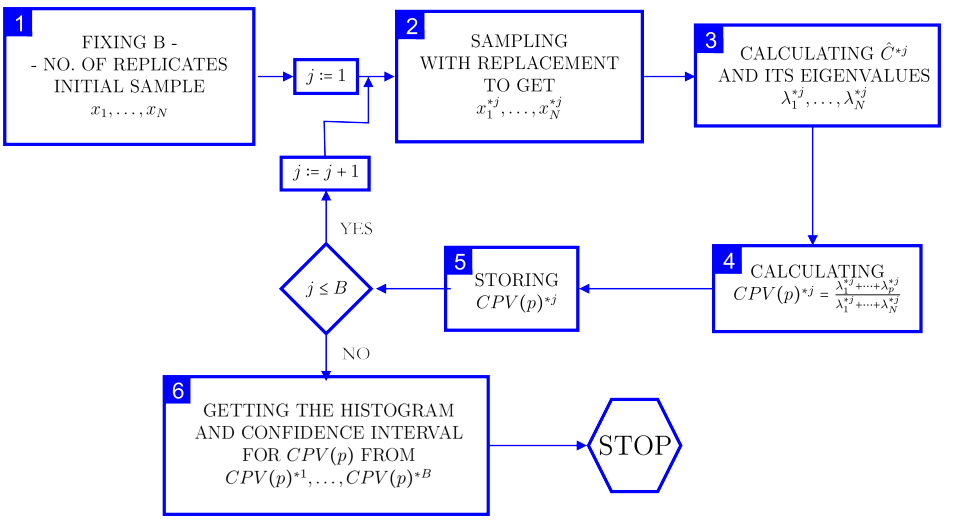


Figure 2: Flowchart of finding the empirical distribution of $CPV(p)$ via bootstrap algorithm.

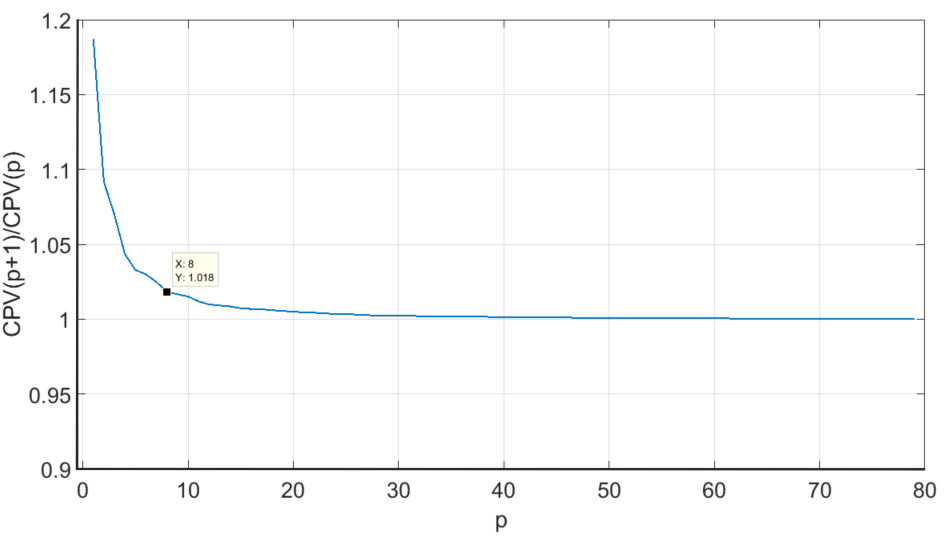


Figure 3: Visualization of the ratio $CPV(p+1)/CPV(p)$ based on experimental data set.

In the following subsection, we will show how to apply the dimensionality reduc-

tion obtained from FPCA to the problem of identification of informative frequency bands (IFB).

2.3. FPCA and informative frequency bands for spectrogram

Let us start with a definition. We define the Informative Frequency Band (IFB) generated by a series of squared absolute of Short-Time Fourier Transforms $|STFT(t,f)|^2$, $t=0,\ldots,T$ and $f\in[0,\Lambda]$ as such a subset $A\subset[0,\Lambda]$ that

$$\frac{IE}{E}(A)\stackrel{def}{=}\frac{\sum\limits_{f\in A}\sum\limits_{t=0}^{T}|STFT(t,f)|^2}{\sum\limits_{f\in[0,\Lambda]}\sum\limits_{t=0}^{T}|STFT(t,f)|^2}\geq L, \tag{5}$$

where E is the total energy of the signal and IE is the energy within the frequency set A and $STFT(t,f)$ defined in (6). The threshold value L, $0\leq L\leq 1$ is usually selected to be bigger than 80%.

To simplify the search for the frequency set A defined in (5), we start from a one-element subset that contains the most energy and we augment it successively adding element by element in the order of the energy contribution. This is represented by the algorithm below.

Identification of IFB by STFT.

Step 1. Identify first f_1 such that $\sum_{t=0}^{T}|STFT(t,f_1)|^2$ is the biggest.
Step 2. Obtain the ranking of frequencies f_i via

$$\sum_{t=0}^{T}|STFT(t,f_1)|^2\geq\ldots\sum_{t=0}^{T}|STFT(t,f_i)|^2\geq\ldots\sum_{t=0}^{T}|STFT(t,f_I)|^2\ .$$

where I is the cardinality of a set of frequencies (discretizated interval of frequencies).
Step 3. From the above ranking we identify the subset $A_{spectr}=\{f_1,f_2,\ldots,f_K\}$ induced by spectrogram such that

$$\frac{IE}{E}(A_{spectr})\geq L,$$

where $\frac{IE}{E}(\cdot)$ was defined in (5).

The above procedure does not involve FPCA, it simply ranks the frequencies in the decreasing order of their influence on the total variability generated by STFT. We will now show how FPCA and the reduction of dimensionality obtained via $CPV(p)$ helps identify the impulsive frequencies.

IFB identification algorithm via FPCA.

Step 1. Represent the raw data produced by the spectrogram as random functions $x_1,\ldots,x_N$. In the Subsection 3.1, we explain how to obtain curves $\{x_i,\ldots,x_N\}$. Using the functional approach, find the empirical covariance estimator $\hat{C}$ (see (2)), the corresponding eigenvalues $\{\hat{\lambda}_i, i=1,\ldots,N\}$ and the functional principal components $\{\hat{u}_i(f), i=1,\ldots,N;\ f\in[0,\Lambda]\}$ (see (3)). Using the $CPV(p)$ technique identify your choice of p. We still use formula (5) with one modification - in place of $|STFT(t,f)|^2$ we insert FPC representation of the curve, i.e. $|x_t(f)|^2 = |\sum_{j=1}^{p}\langle\hat{u}_j,x_t\rangle\hat{u}_j(f)|^2$

Step 2. Identify first f_1 such that $\sum_{t=0}^{T}|x_t(f_1)|^2$ is the biggest.

Step 3. Obtain the ranking of frequencies f_i via

$$\sum_{t=0}^{T}|x_t(f_1)|^2 \geq \sum_{t=0}^{T}|x_t(f_2)|^2 \geq \ldots \geq \sum_{t=0}^{T}|x_t(f_I)|^2$$

Step 4. From the above ranking we identify the FPCA induced subset $A_{FPCA} = \{f_1,f_2,\ldots,f_R\}$ such that

$$\frac{IE}{E}(A_{FPCA}) \geq L,$$

where $\frac{IE}{E}(\cdot)$ was defined in (5).

In the next Section dedicated to applications we will show how close the sets A_{spectr} and A_{FPCA} are.

3. Application to gearbox and wheel bearing data

In this Section we will show the application of the functional data approach presented in the previous section to the spectrogram in the context of identifying the informative frequency band (IFB). Recall that the spectrogram represents the signal as a collection of short-time Fourier transforms (STFT). Using time-frequency plots produced by sequences of STFTs one tries to identify the frequency bands where the excitation (energy) of the signal of interest is the most significant. From our perspective, however, the spectrogram is a collection of random curves. To make our point more precise let X be a signal of interest. Recall that the STFT is defined as:

$$\text{STFT}(t,f) = \int_{-\infty}^{\infty} w(t-\tau)X(\tau)e^{-2\pi i f\tau}d\tau, \tag{6}$$

where $w(\cdot)$ is the window function, $t\in[0,T]$ and the frequency $f\in[0,\Lambda]$. The discrete version of STFT is defined as follows:

$$\mathrm{STFT}(t,f) = \sum_{k=0}^{M-1} X_k w(t-k) e^{-2\pi i f k/M}. \tag{7}$$

3.1. Spectrogram and FPCA on data

A spectrogram is a visual representation of the spectrum of frequencies in a signal as it varies with time or some other variable. A common format is a graph with two geometric dimensions: the vertical axis represents frequency, the horizontal axis is time; a third dimension indicating the amplitude of a particular frequency at a particular time is represented by the intensity or colour of each point in the image. Here we will analyse the spectrogram generated by the open-pit excavating machine from a Polish brown coal mine. The signal is the acceleration signal generated in the gearbox of this machine. The raw signal that combines four signals from four sensors has a length of 20480 data points. The impulses are convoluted with the raw signal and then the spectrogram of the result is analysed. The resulting signal $\{X(s) : s \in [0,T]\}$ is represented in Figure 4.

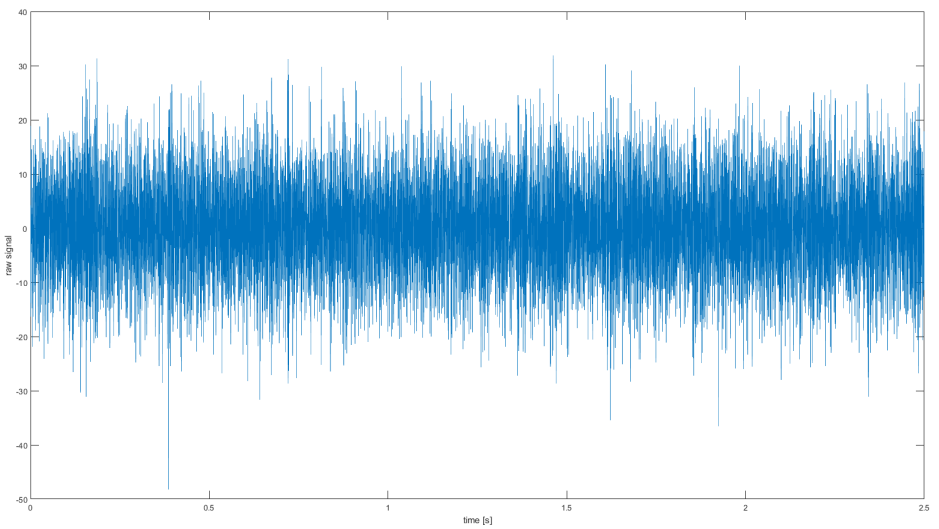


Figure 4: Acceleration signal with the impulses.

The spectrogram corresponding to the above signal is represented in Figure 5 on the left panel. The spectrogram of the signal obtained after the dimensionality reduction done by FPCA method is shown on the right panel. Here, we clarify how to create particular spectrograms.

Traditional data spectrogram: First, decompose the signal $\{X(s) : s \in [0,T]\}$ into the set of overlap narrowband sub-signals $\{X_t(s) : s \in [t,t+w]\}_{t=0}^{N-1}$. Next, use Fourier transform (FFT) to calculate the magnitude of the frequency spectrum for each sub-signal (FFT is a digital process). Vertical (or horizontal) line in the image cor-

responds to each sub-signals; a measurement of magnitude vs. frequency for a specific moment in time. Finally, these spectrums or time plots are then "laid side by side" to form the image or a three-dimensional surface, or slightly overlapped (windowing) in various ways. To sum up, a spectrogram is a frequency-time domain map, representing an energy of signal (power spectrum). Its values are stored in a huge matrix $\{|\text{STFT}(t,f)|^2\}_{f\in[0,\Lambda];\, t\in[0,T]}$ with entries defined in (7).

Spectrogram via FPCA method: We process a vibration signal similarly to traditional method with the difference that each sub-signal selected by windowing is converted into function, hence we view a spectrogram as a collection of random curves. According to our notation, for each $t = 0,\ldots,N-1$ we convert $\{|STFT(t,f)|^2\}_{f=0}^{\Lambda}$ into random curves $\{x_t(f) : f \in [0,\Lambda]\}_{t=0,\ldots,N-1}$, where Λ is fixed maximum frequency, and $x_t \in L^2[0,\Lambda]$. Conversion of vectors (here, rows of the matrix $\{|STFT(t,f)|^2\}_{f=0}^{\Lambda}$) to curves is carried out with the use of basis expansion in $L^2[0,1]$ space. Here, we used the program `R` and its package `fda.usc` with the function `fdata2fd` to produce functional object. The type of functional basis is Bspline by default, but of course it is possible to change basis to any other, for example Fourier basis. Next, we reduce dimensionality of those curves using FPCA. For curves, represented by combination of several eigenfunctions and scores we calculate STFT defined in (6) achieving again a matrix $\{|\text{STFT}(t,f)|^2\}_{f\in[0,\Lambda];\, t\in[0,T]}$ which has graph representation in Figure 5 on right panel. From our perspective, the spectrogram is a collection of random curves indexed by the parameter t (time). In other words, the spectrogram is a set which define the energy of functions. Consequently, IFB is a subset of those functions whose energy within this band is close to the total energy.

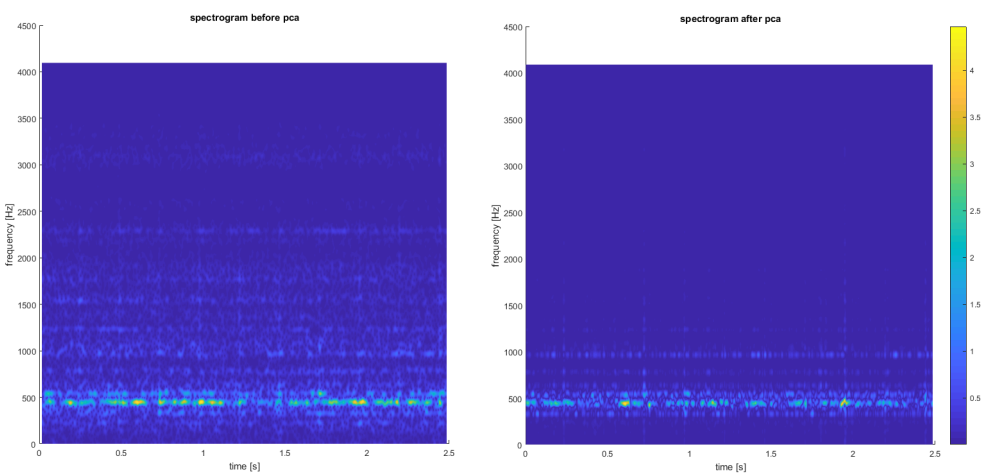


Figure 5: Traditional data spectrogram (left) and spectrogram via FPCA method (right).

Our data set is a relatively big 513×1265 matrix with entries $\{|STFT(t,f)|^2\}$. The second dimension (frequencies f) range is $513 = 2^9 + 1$ and it refers to the sample

frequency. The first dimension (time t) is 1265 and it refers to the number of windows N covering a signal of length 20480 ($16 \cdot N = 20480 - 256$). Here, we have used Hamming window (a function describing the way of sampling within a signal, given by formula $w(t-k) = 0.53836 - 0.46164\cos(\frac{2\pi(t-k)}{M-1})$; $k = 1,\ldots,M$) of length $256 = 2^8$ and an overlap $M = 240$. Next, each row of $\{|STFT(t,f)|^2\}$ is converted to function $x_t(f)$, $t = 0,\ldots,1-N$, which forms our functional data set. The last step is to approximate those curves with principal components expansion.

We will now show how the FPCA method works in practice. We choose $p = 8$ first functional principal components generated by the empirical covariance operator $\hat{C}$ induced by the random curves $x_1,\ldots,x_N$ corresponding to the spectrogram. How to obtain the empirical operator was explained in previous Section 2, i.e. in formula (2). It is worth to notice, that x in (2) is any element of the Hilbert space H, so the easiest way to choose this element is to take any basis element. Moreover, $\hat{C}$ is fully described by its eigenfunction and eigenvalues, which are known from the data. We illustrate our eight empirical principal components in Figure 6, which is a fragment of a 8×8 matrix of small pictures. On the diagonal, first four from the eight functional principal components $\hat{u}_i$, $i = 1,\ldots,8$ are represented. Outside the diagonal, we show the scatterplots of scores $\{(PC_i^k, PC_j^k)\}_{k=1}^N$, $i \neq j$, $i,j = 1,\ldots,8$. Recall that scores $PC_i^k = \langle \hat{u}_i, x_k \rangle$, where $\hat{u}_i$ are the eigenfunctions of the empirical covariance operator $\hat{C}$ and x_k, $k = 1,\ldots,N$ is a function from the sample $\{x_1,\ldots,x_N\}$. In our case $N = 1265$. The more the scatterplots are irregular, the less correlation between them. Ideally, one would like to have a zero correlation between them since we want our functional principal components to be orthogonal. Our graphs confirm the idea of the weak correlation between the scores, hence our empirical FPC are indeed orthogonal.

In Table 1, we show the percentage of variance explained by each of the first eight functional principal components $FPC(i), i = 1,\ldots,8$.

Table 1: Percentage of variability explained by each empirical functional component.

FPC1	FPC2	FPC3	FPC4	FPC5	FPC6	FPC7	FPC8
52.62%	9.91%	5.48%	4.07%	3.36%	2.38%	2.10%	1.96%
Total explained:	81.88%						

First four empirical functional components are quite informative, explaining as much as 72.08% of variability so we show the corresponding 4×4 matrix of plots of the scores, which are highly uncorrelated confirming orthogonality of empirical FPC (see values of correlation in the top of the boxes outside the diagonal).

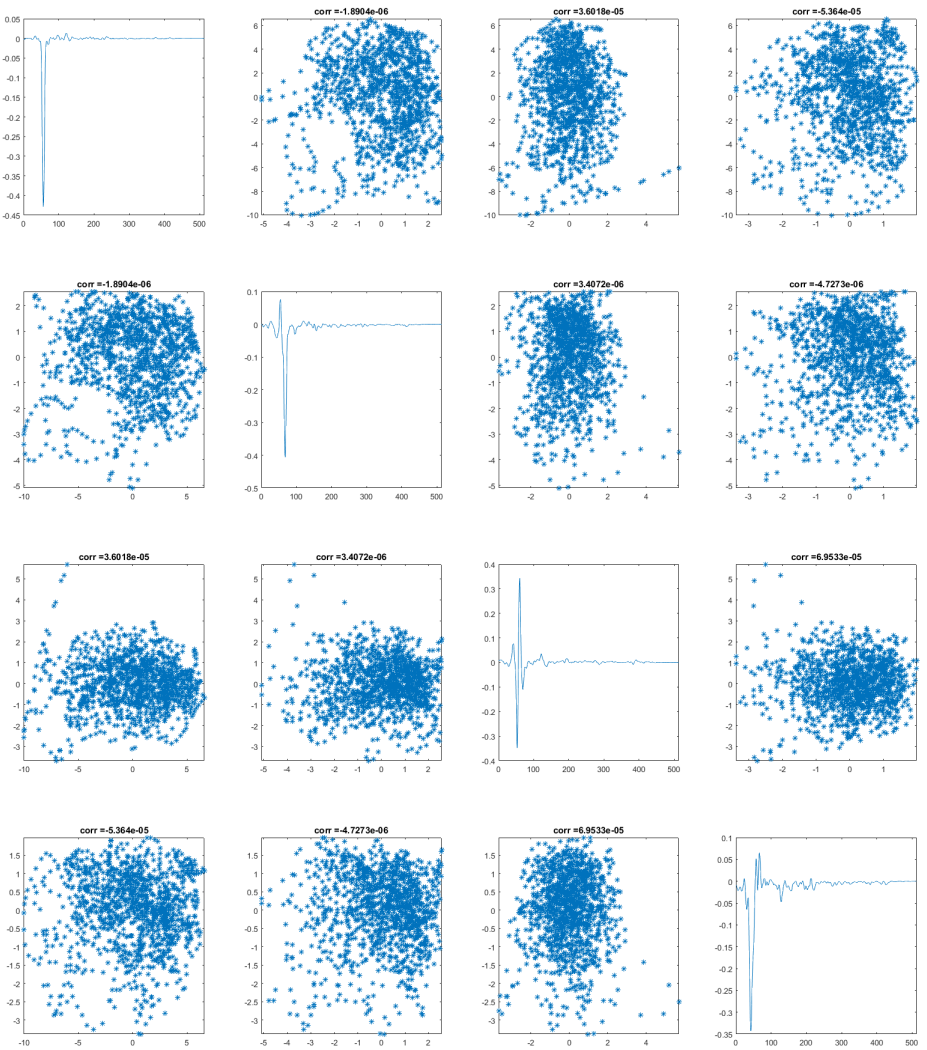
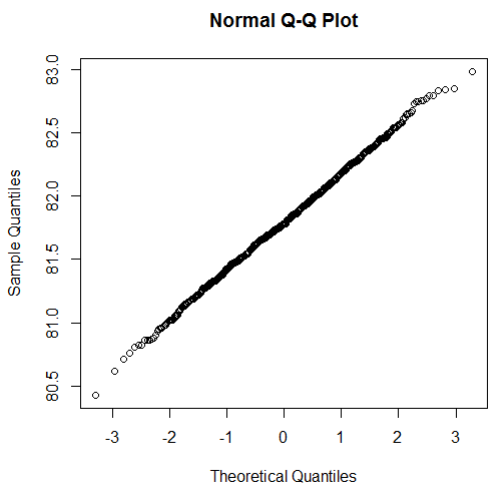


Figure 6: First four FPC with scatterplots of scores.

3.2. Cumulative percentage of variance (CPV) study

The previous subsection was devoted to illustrating how our FPCA method works in practice. For the first eight empirical functional eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_8$ we have obtained a quite reassuring result: they represent as much as 81.88% of variance. From the statistical perspective, however, we would like to get more information on the variability of the $CPV(p)$ coefficient. In other words, we would like to be able to measure the variability of our estimate with the point value of 81.88%. To answer this question, we will apply the CPV bootstrap algorithm introduced in the previous Section 2.
The statistical features of the bootstap distribution of CPV(8) are shown in Figure 7.

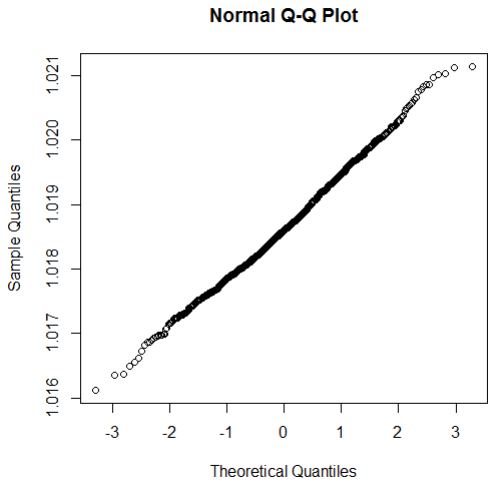


Descriptive statistics of CPV(8)	Values
Minimum	80.43000
2.5% quantile	81.02000
Median	81.78000
Mean	81.79348
97.5% quantile	82.54025
Maximum	82.98000
Variance	0.14648
Skewness	−0.00285
Kurtosis	3.07514

Figure 7: Descriptive statistics for $CPV(8)$ based on bootstrap samples.

The most important message from the above calculations is that the 95% confidence interval for CPV(8) is from 81.02% (the 2.5% quantile) to 82.54% (the 97.5% quantile). This means that CPV as a random variable is quite concentrated around its point value 81.88% and that our results are quite reliable and have only a small spread.

To analyse the speed of convergence $CPV(p+1)/CPV(p) \stackrel{p\to\infty}{\longrightarrow} 1$ we have performed the bootstrap distribution study of this ratio and have obtained the results presented in Figure 8.



Descriptive statistics of CPV(9)/CPV(8)	Values
Minimum	1.01612
2.5% quantile	1.01719
Median	1.01859
Mean	1.01864
97.5% quantile	1.02023
Maximum	1.02113
Variance	$6.56e-07$
Skewness	0.19956
Kurtosis	2.95591

Figure 8: Normality of $CPV(9)/CPV(8)$ based on bootstrap samples.

Again, we see the usefulness of the bootstrap method. From the bootstrap method we get the 95% confidence interval for the proportion $CPV(9)/CPV(8)$ is $[1.017; 1.020]$ which means that increasing p from 8 to 9 we will get only 2% more of the variance explained. This is sufficient argument to stop at $p = 8$. Adding more FPC's dose not improve significantly total variance explained.

3.3. Application to informative frequency bands

In our experiment, we set the threshold L defined in (5) as 80%. Therefore, for our vibration data coming from the gearbox we will be looking for two sets: A_{spectr} and A_{FPCA} such that $\frac{IE}{E}(A_{spectr}) \geq 80\%$ and $\frac{IE}{E}(A_{FPCA}) \geq 80\%$. We will see how close those two sets are on real applications using vibration signals.

Searching for A_{spectr} that satisfies (5) may be quite time consuming as we have to consider all possible subsets of frequencies from the set $[0, \Lambda]$. For example, for the vibration gearbox data one would have to deal with 2^{513} combinations! As described in the previous Section, we start the search by identifying the frequency f_1 that maximizes $\sum_{t=0}^{T} |STFT(t,f)|^2$, then select the second in the order of the energy contribution and so on.

The analysis based on $CPV(p)$ presented in the previous subsection has shown that $p = 8$ first functional principal components reproduce as much as 81.88% of the total variability in the data. Therefore, the FPCA method induced by the spectrogram of gearbox signal creates the eight-dimensional vector of scores, taken from the functional expansion. This means reducing the dimensionality of our problem

from 1265 windows with 513 frequencies in each to 8×513 considering spectrogram data as a functions of frequencies. This means, that finding A_{FPCA} is much faster than A_{spectr}, especially for large data sets.

Below, we provide a listing of all frequencies pertaining to A_{spectr} and A_{FPCA}.

$$\begin{aligned} A_{spectr} =\{&447.13, 455.11, 439.14, 463.1, 431.16, 471.08, 423.17, 479.06, 415.19, \\ &542.94, 550.92, 534.96, 487.05, 526.97, 558.91, 518.99, 566.89, 407.2, \\ &495.03, 511, 574.88, 503.02, 335.35, 343.33, 327.36, 399.22, 351.31, \\ &319.38, 359.3, 391.24\}. \end{aligned}$$

A_{spectr} has 30 elements.

$$\begin{aligned} A_{FPCA} =\{&447.13, 455.11, 439.14, 463.1, 431.16, 471.08, 423.17, 479.06, 415.19, \\ &542.94, 534.96, 550.92, 487.05, 526.97, 558.91, 407.2, 518.99, 566.89\}. \end{aligned}$$

A_{FPCA} has 18 elements.

Note that the first 10 frequencies coincide, up to the second decimal point (in Hz). If we consider frequencies from the intersection $B = A_{spectr} \cap A_{FPCA}$, then the percentage of signal energy describing IFB is already at the level of 74,03%, not very far from the threshold of 80%.

We have applied the above analysis to the data set generated by a wheel bearing and described in Cioch et al. (2013). The sets A_{spectr} and A_{FPCA} are shown below for these data

$$\begin{aligned} A_{spectr} = \{&1272.51, 1291.23, 1253.8, 1309.94, 3106.43, 3125.15, 3087.72, 1235.09, 1328.65 \\ &3143.86, 3069.01, 1347.37, 3162.57, 3050.29, 1216.37, 1366.08, 3031.58, 3181.29 \\ &1197.66, 1384.8, \} \end{aligned}$$

$$\begin{aligned} A_{FPCA} = \{&1272.51, 1291.23, 1253.8, 1309.94, 3106.43, 1235.09, 3125.15, 1328.65, 3087.72 \\ &3143.86, 3069.01, 1347.373162.57, 1216.37, 3050.29, 1366.08, 3181.29, 3031.58 \\ &1384.8, 1197.66\}. \end{aligned}$$

A careful examination of the above listings for both data sets shows that both sets contain the same frequencies. They are shown in the order of their importance in energy explained in Section 2. Therefore, the only change we have using the

FPCA is the change of the order of the energy importance of the frequency. Observe, however, that such a change is not dramatic. FPCA preserves the order of the first five frequencies and then makes only small changes, never bigger than two places in the order of energy.

4. Conclusions

Our article is devoted to introducing the functional data approach to analyse big data generated by signals available for structural health monitoring. We show that applying the FPCA - the functional principal component approach - we can reduce the dimensionality of the data from several millions to several thousands. Using such an approach we show the importance of the eigenfunctions and eigenvalues calculated for functions generated by observing the signal. It turns out that the popular coefficient - the cumulative percentage of the variance explained (CPV) exceeds 80 per cent for the initial few functional components. We show that this approach applied to the signal generated by the excavating machine can be helpful in identifying informative frequency bands. Moreover, applying the bootstrap approach we can show that the CPV has a relatively small dispersion, which proves the numerical stability of our results.

Acknowledgement

The Authors would like to express their gratitude to Professor Radoslaw Zimroz from Wrocław for providing access to data regarding gearbox of the excavating machine.

REFERENCES

BOSQ, D., (2000). Linear Processes in Function Spaces, Springer Verlag.

CHU, F., FENG, Z., LIANG, M., (2013). Recent advances in time-frequency analysis methods for machinery fault diagnosis: a review with application examples, Mechanical Systems and Signal Processing, Vol. 38, No. 1, pp. 165–205.

CIOCH, W., KNAPIK, O., LEŚKOW, J., (2013). Finding a frequency signature for a cyclostationary signal with applications to wheel bearing diagnostics, Mechanical Systems and Signal Processing, Vol. 38, pp. 55–64.

GRYLLIAS, K., ANDRE, H., LECLERE, Q., ANTONI, J., (2017). Condition monitoring of rotating machinery under varying operating conditions based

on Cyclo- Non-Stationary Indicators and a multi-order probabilistic approach for instantaneous angular speed tracking, IFAC papers online, Vol. 50-1, pp. 4708–4712.

GUO, X, CHEN, L., SHEN, CH., (2016). Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis, Measurement, Vol. 93, pp. 490–502.

HORVÁTH, L., KOKOSZKA, P., (2012). Inference for Functional Data with Applications, Springer-Verlag, New York etc.

JIA, F., LEI, Y., LIN, J., ZHOU, X., LU, N., (2016). Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data, Mechanical Systems and Signal Processing, Vol. 72-73, pp. 303–315.

KHADERSAB, A., SHIVAKUMAR, DR. S., (2018). Vibration Analysis Techniques for Rotating Machinery and its effect on Bearing Faults, Procedia Manufacturing, Vol. 20, pp. 247–252.

KRUCZEK, P., WODECKI, J., WYŁOMAŃSKA, A., (2017). Novel method of informative frequency band selection for vibration signal using nonnegative matrix factorization of short-time fourier transform, IEEE 11th International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives (SDEMPED).

LIU, R., YANG, B., ZIO, E., CHEN, X., (2018). Artificial intelligence for fault diagnosis of rotating machinery: A review, Mechanical Systems and Signal Processing, Vol. 108, pp. 33–47.

MAS, A., (2002). Weak convergence for the covariance operators of a Hilbertian linear process, Stochastic Processes and their Applications, 99, pp. 117–135.

OBUCHOWSKI, J. WYŁOMAŃSKA, A., ZIMROZ, R., (2014). Selection of informative frequency band in local damage detection in rotating machinery, Mechanical Systems and Signal Processing, 48, pp. 138–152.

RANDALL, B., ANTONI, J., (2011). Rolling element bearing diagnostics - A tutorial, Mechanical Systems and Signal Processing, Vol. 25, pp. 485–520.

RAMSAY, J. O., SILVERMAN, B. W., (2002). Applied functional data analysis, Springer-Verlag.

RAMSAY, J. O., SILVERMAN, B. W., (2005). Functional data analysis, Springer-Verlag.

SPIRIDONAKOS, M. D., FASSOIS, S. D., (2014). Non-stationary random vibration modelling and analysis via functional time-dependent ARMA (FS-TARMA) models - A critical survey, Mechanical Systems and Signal Processing, Vol. 47, pp. 175–224.

STASZEWSKI, W. J., WALLACE, D. M., (2014). Wavelet-based Frequency Response Function for time-variant systems - An exploratory study, Mechanical Systems and Signal Processing, Vol. 47, pp. 35–49.

YANG, Y., NAGARAJAIAH, S., (2014). Blind identification of damage in time-varying systems using independent component analysis with wavelet transform, Mechanical Systems and Signal Processing, Vol. 47, pp. 3–20.

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 153–169, DOI 10.21307/stattrans-2019-009

AN ALTERNATIVE MATRIX TRANSFORMATION TO THE F TEST STATISTIC FOR CLUSTERED DATA

Sukanya Intarapak[1], **Thidaporn Supapakorn** [2]

ABSTRACT

For the regression analysis of clustered data, the error of cluster data violates the independence assumption. Consequently, the test statistic based on the ordinary least square method leads to incorrect inferences. To overcome this issue, the transformation is required to apply to the observations. In this paper we propose an alternative matrix transformation that adjusts the intra-cluster correlation with Householder matrix and apply it to the F test statistic based on generalized least squares procedures for the regression coefficients hypothesis. By Monte Carlo simulations of the balanced and unbalanced data, it is found that the F test statistic based on generalized least squares procedures with Adjusted Householder transformation performs well in terms of the type I error rate and power of the test.

Key words: adjusted Householder, clustered data, F test statistic, generalized least squares, intra-cluster correlation.

1. Introduction

Clustered data arise in many situations such as health research (multiple patients within a hospital) (see Miall and Oldham (1955) and Ng et al. (2004)), education study (multiple students within a school) (see McCulloch and Shayle (2001)) and biological science (multiple children within a family) (see Agarwal et al. (2005)). Clustered data are characterized as data that can be classified into a number of distinct groups or clusters (see Galbraith et al. (2010)). Any two responses from different clusters are independent, but pairs of responses within clusters are correlated, and the correlation is the same for all pairs of individuals from the same cluster, which is called the intra-cluster correlation (see Eldridge et al. (2009)). In general, the regression technique assumes that the errors in observations are independent, identically and normally distributed. This assumption will not be always held for clustered data. Battese et al. (1988) proposed a regression method for analysing clustered data, which is called the nested error regression model.

The nested error regression model is expressed as

$$y_{ij} = \mathbf{x}_{ij}\beta + u_i + e_{ij}, \quad i = 1,\ldots,c; j = 1,\ldots,n_i, \tag{1}$$

[1]Department of Mathematics, Faculty of Science, Srinakharinwirot University, Bangkok, Thailand. E-mail: sukanyain@g.swu.ac.th. ORCID ID: https://orcid.org/0000-0002-6770-9503.
[2]Corresponding Author, Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand. E-mail: fscitdps@ku.ac.th. RCID ID: https://orcid.org/0000-0003-0019-9884.

where y_{ij} is the observed response for the jth sample unit in the ith cluster, $\mathbf{x}_{ij} = (x_{ij0}, x_{ij1}, \ldots, x_{ij,k-1})$ is the $n \times k$ matrix of explanatory variables and x_{ij0} is the $n \times 1$ column vector where entries are all 1, $\beta = (\beta_0, \beta_1, \ldots, \beta_{k-1})'$ is the k vector of regression coefficients and n_i is the number of sample units observed in the ith cluster ($\sum_{i=1}^{c} n_i = n$). The random effect u_i and random error e_{ij} are assumed to be independent of each other and distributed as $N(0, \sigma_u^2)$ and $N(0, \sigma_e^2)$, respectively.

The model (1) can be written as

$$y = \mathbf{X}\beta + \varepsilon, \tag{2}$$

where $y = (y_1, \ldots, y_c)'$ with $y_i = (y_{i1}, \ldots, y_{in_i})$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_c)'$ with $\mathbf{X}_i = (X_{i10}, \ldots, X_{in_i,k-1})$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_c)'$ with $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})$. Further, $\varepsilon_{ij} = u_i + e_{ij}, \varepsilon \sim N(0, \sigma^2\mathbf{V}), \sigma^2 = \sigma_u^2 + \sigma_e^2$, $\mathbf{V}$ has block-diagonal variance-covariance matrix with $\mathbf{V}_i = (1-\rho)\mathbf{I}_{n_i} + \rho\mathbf{J}_{n_i}$ for the ith cluster where $\rho = \sigma_u^2/\sigma^2$ is the intra-cluster correlation, $\mathbf{I}_{n_i}$ is the $n_i \times n_i$ identity matrix and $\mathbf{J}_{n_i}$ is the $n_i \times n_i$ matrix consisting of all 1s.

For testing the hypothesis about regression coefficients in the nested error regression model, formerly, Wu et al. (1988), Rao et al. (1993) and Lahiri and Li (2009) showed that the F test statistic based on ordinary least squares procedures leads to highly inflated type I error rate. Wu et al. (1988) proposed a modification of the F test statistic with known intra-cluster correlations, which is much better than the F test statistic based on ordinary least squares procedures by type I error rate. Rao et al. (1993) presented the F test statistic based on generalized least squares procedures with the Fuller–Battese transformation (see Galbraith et al. (2010)) in order to make observations independent and then applied the F test statistic to the observations under valid assumption. The F test statistic with the Fuller–Battese transformation performs similar to the modification of the F test statistic in controlling the type I error rate. Furthermore, the power of F test statistic with the Fuller–Battese transformation increases as the intra-cluster correlation increases, whereas the power of the modification of F test statistic decreases. The power of the F test statistic based on ordinary least squares procedures is not comparable because of type I error rate inflation.

Recently, Lahiri and Li (2009) suggested the transformation for the F test statistic based on generalized least squares procedures that is part of the Helmert matrix (see Lancaster (1965)), unlike the Fuller–Battese transformation. Like previous work, the F test statistic with part of Helmert matrix performs as well as in controlling the type I error rate, but the power of the test is not considered.

In this paper, we propose an alternative transformation for the generalized least squares procedures by applying Householder matrix (see Householder (1958)). In Section 2 we review several F test statistics for testing linear hypothesis regarding the regression coefficients under the nested error regression model. Monte Carlo study concerning the type I error rate and the power of the F test statistic is conducted in Section 3 as well as the real application. The results of the simulation are presented in Section 4.

2. The F test statistics

2.1. Prior F test statistics

Under model (1), suppose that the hypothesis of interest is $H_0 : \mathbf{C}\beta = \mathbf{q}$, where $\mathbf{C}$ is a known $m \times k$ matrix of rank $m(< k)$, and $\mathbf{q}$ is a known $m \times 1$ constant vector. The F-statistic based on ordinary least squares procedures is

$$F_{OLS} = \frac{(\mathbf{C}\widehat{\beta} - \mathbf{q})'\{\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\}^{-1}(\mathbf{C}\widehat{\beta} - \mathbf{q})/m}{(\ \mathbf{y} - \mathbf{X}\widehat{\beta})'(\mathbf{y} - \mathbf{X}\widehat{\beta})/(n-k)}$$

where $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

F_{OLS} leads to highly inflated type I error rate when the intra-cluster correlation increases. When the intra-cluster correlation is known, Wu et al. (1988) proposed a modification of the F test statistic by multiplying the numerator and the denominator of F_{OLS} by a chi-squared distribution with $n-k$ and m degrees of freedom, respectively.

The modification of the F test statistic is

$$F_{WU} = F_{OLS} \times \frac{\{n - \mathrm{tr}(\mathbf{PV})\}/(n-k)}{\mathrm{tr}(\mathbf{P}_C\mathbf{V})/m},$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \mathbf{P}_C = \mathbf{X}_C(\mathbf{X}'_C\mathbf{X}_C)^{-1}\mathbf{X}'_C, \mathbf{X}_C = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$ and tr is the trace operator.

Afterwards Rao et al. (1993) presented the F test statistic based on generalized least squares procedures with the Fuller–Battese transformation under model (1).

Define $\mathbf{T}_i = \mathbf{I}_{n_i} - n_i^{-1}(1 - [\{1-\rho\}/\{1+(n_i-1)\rho\}]^{1/2})\mathbf{J}_{n_i}, \mathbf{y}_i^* = \mathbf{T}_i\mathbf{y}_i, \mathbf{X}_i^* = \mathbf{T}_i\mathbf{X}_i$, and $\varepsilon_i^* = \mathbf{T}_i\varepsilon_i$. Then the transformed model can be written as $\mathbf{y}^* = \mathbf{X}^*\beta + \varepsilon^*$, where $\varepsilon \sim \mathbf{N}(0, \sigma_e^2\mathbf{I}_n)$ and $\sigma_e^2 = \sigma^2(1-\rho)$. Thus, the F test statistic based on generalized least squares procedures with the Fuller–Battese transformation is

$$F_{RAO} = \frac{(\mathbf{C}\beta^* - \mathbf{q})'(\mathbf{X}_C^{*'}\mathbf{X}_C^*)^{-1}(\mathbf{C}\beta^* - \mathbf{q})/m}{(\mathbf{y}^* - \mathbf{X}^*\beta^*)'(\mathbf{y}^* - \mathbf{X}^*\beta^*)/(n-k)},$$

where $\beta^* = (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{y}^*$ and $\mathbf{X}_C^* = \mathbf{X}^*(\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{C}'$.

Recently, unlike the Fuller–Battese transformation, Lahiri and Li (2009) proposed the transformation for the F test statistic based on generalized least squares procedures, which is part of the Helmert matrix. Generally, the Helmert matrix is orthogonal (see Farhadian and Asadian (2017)), but Lahiri and Li (2009) used the Helmert matrix by ignoring the first row. Thus, the part of the Helmert matrix is not orthogonal.

Let $\mathbf{G}_i$ be an $(n_i - 1) \times n_i$ matrix which is part of the Helmert matrix by ignoring the first row, i.e. $1'_{n_i}/\sqrt{n_i}$. Multiplying both sides of the model (2) by $\mathbf{G}_i$ then the transformed model is written as $\mathbf{y}^\star = \mathbf{X}^\star\beta + \varepsilon^\star$, where $\varepsilon^\star \sim \mathbf{N}(0, \sigma_e^2\mathbf{I}_{n-c})$ and $\sigma_e^2 = \sigma^2(1-\rho)$. The F test statistic based on generalized least squares procedures with

part of the Helmert transformation is

$$F_{LAH} = \frac{(\mathbf{C}\beta^\star - \mathbf{q})'(\mathbf{X}_C^{\star'}\mathbf{X}_C^\star)^-(\mathbf{C}\beta^\star - \mathbf{q})/m}{(\mathbf{y}^\star - \mathbf{X}^\star\beta^\star)'(\mathbf{y}^\star - \mathbf{X}^\star\beta^\star)/(n-c-k)},$$

where $\beta^\star = (\mathbf{X}^{\star'}\mathbf{X}^\star)^{-1}\mathbf{X}^{\star'}\mathbf{y}^\star$ and $\mathbf{X}_C^\star = \mathbf{X}^\star(\mathbf{X}^{\star'}\mathbf{X}^\star)^{-1}\mathbf{C}'$.

2.2. F test statistic with an alternative transformation

For the F test statistic based on generalized least squares procedures, the transformation matrix is the necessary part to make the observations independent. Unlike the previous transformations, we propose an alternative transformation that adjusts the Householder matrix.

The Householder matrix (see Appendix) is taken into account because of its orthogonal, symmetry and idempotent properties, which are necessary for the transformation matrix. The Householder matrix for the ith cluster, denoted by $\mathbf{H}_i$, can be written in a simple form as

$$\mathbf{H}_i = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & \frac{-1}{\sqrt{n_i-1}} & \frac{-1}{\sqrt{n_i-1}} & \cdots & \frac{-1}{\sqrt{n_i-1}} \\ 0 & \frac{-1}{\sqrt{n_i-1}} & \frac{(n_i-1)(n_i-3)+\sqrt{n_i-1}}{(n_i-1)(n_i-2)} & \cdots & \frac{-1}{(n_i-1)+\sqrt{n_i-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \frac{-1}{\sqrt{n_i-1}} & \frac{-1}{(n_i-1)+\sqrt{n_i-1}} & \cdots & \frac{(n_i-1)(n_i-3)+\sqrt{n_i-1}}{(n_i-1)(n_i-2)} \end{bmatrix}.$$

Even if $\mathbf{H}_i$ is orthogonal, the error term of the transformed model is still fallacious, that is the error term is not independent. Therefore, $\mathbf{H}_i$ is required to be adjusted. Let $\mathbf{D}_i$ be an $n_i \times n_i$ matrix, which is defined as

$$\mathbf{D}_i = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\rho\sqrt{n_i-1} & \sqrt{1+(n_i-2)\rho-(n_i-1)\rho^2} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{1-\rho} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{1-\rho} \end{bmatrix}$$

such that $(\mathbf{H}_i\mathbf{D}_i)(\mathbf{H}_i\mathbf{D}_i)' = \mathbf{V}_i$. Now, we have the alternative matrix transformation, $\mathbf{P}_i$, which is the inverse of matrix $(\mathbf{H}_i\mathbf{D}_i)$, also, $\mathbf{y}_i^{AH} = \mathbf{P}_i\mathbf{y}_i, \mathbf{X}_i^{AH} = \mathbf{P}_i\mathbf{X}_i$ and $\varepsilon_i^{AH} = \mathbf{P}_i\varepsilon_i$.

Then, the transformed model can be written as $\mathbf{y}^{AH} = \mathbf{X}^{AH}\beta + \varepsilon^{AH}$, where

$$\mathbf{y}^{AH} = (\mathbf{y}_1^{AH}, \ldots, \mathbf{y}_c^{AH})' \quad \text{with} \quad \mathbf{y}_i^{AH} = (y_{i1}^{AH}, \ldots, y_{in_i}^{AH}),$$

$$\mathbf{X}^{AH} = (\mathbf{X}_1^{AH}, \ldots, \mathbf{X}_c^{AH})' \quad \text{with} \quad \mathbf{X}_i^{AH} = (X_{i10}^{AH}, \ldots, X_{in_i,k-1}^{AH})$$

$$\text{and} \quad \varepsilon^{AH} = (\varepsilon_1^{AH}, \ldots, \varepsilon_c^{AH})' \quad \text{with} \quad \varepsilon_i^{AH} = (\varepsilon_{i1}^{AH}, \ldots, \varepsilon_{in_i}^{AH}).$$

Currently, the assumption of the error is valid, i.e. $\mathrm{var}(\varepsilon_i^{AH}) = \mathrm{var}(\mathbf{P}_i\varepsilon_i) = \mathrm{var}\{(\mathbf{H}_i\mathbf{D}_i)^{-1}\varepsilon_i\} =$

$\sigma^2(\mathbf{H}_i\mathbf{D}_i)^{-1}\mathbf{V}_i\,(\mathbf{H}_i\mathbf{D}_i)^{-1'} = \sigma^2\mathbf{I}_{n_i}$, and $\mathrm{cov}(\varepsilon_i^{AH}, \varepsilon_j^{AH}) = 0$ for $i \neq j$, that is $\mathrm{cov}(y_i^{AH}, y_j^{AH}) = 0$ for $i \neq j$.

Ultimately, the F test statistic based on generalized least squares procedures with Adjusted Householder transformation is

$$F_{AH} = \frac{(\mathbf{C}\beta^{AH} - \mathbf{q})'(\mathbf{X}_C^{AH'}\mathbf{X}_C^{AH})^{-1}(\mathbf{C}\beta^{AH} - \mathbf{q})/m}{(\mathbf{y}^{AH} - \mathbf{X}^{AH}\beta^{AH})'(\mathbf{y}^{AH} - \mathbf{X}^{AH}\beta^{AH})/(n-k)},$$

where $\hat{\beta}^{AH} = (\mathbf{X}^{AH'}\mathbf{X}^{AH})^{-1}\mathbf{X}^{AH'}\mathbf{y}^{AH}$ and $\mathbf{X}_C^{AH} = \mathbf{X}^{AH}(\mathbf{X}^{AH'}\mathbf{X}^{AH})^{-1}\mathbf{C}'$.

3. Simulation study

3.1. A Monte Carlo simulation

In this section the data sets are randomly generated to illustrate how various methods of statistical inference perform for analysing the clustered data. Following Wu et al. (1988), Rao et al. (1993) and Lahiri and Li (2009), the nested error regression model with two covariates (i.e. x_1 and x_2) is considered:

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + u_i + e_{ij}, \quad i = 1,\ldots,c; j = 1,\ldots,n_i. \tag{3}$$

The data sets of (x_{ij1}, x_{ij2}) are generated from the bivariate normal distribution with additional random effects components to allow for the intra-cluster correlations ρ_{x_1} and ρ_{x_2} on both x_1 and x_2, respectively:

$$x_{ij1} = \mu_{x_1} + u_{x_1 i} + e_{x_1 ij}, \quad x_{ij2} = \mu_{x_2} + u_{x_2 i} + e_{x_2 ij},$$

where $u_{x_1 i} \sim N(0, \sigma^2_{ux_1})$, $e_{x_1 i} \sim N(0, \sigma^2_{ex_1})$, $u_{x_2 i} \sim N(0, \sigma^2_{ux_2})$, $e_{x_2 i} \sim N(0, \sigma^2_{ex_2})$, $\rho_{x_1} = \sigma^2_{ux_1}/\sigma^2_{x_1}$, $\rho_{x_2} = \sigma^2_{ux_2}/\sigma^2_{x_2}$, $\sigma^2_{x_1} = \sigma^2_{ux_1} + \sigma^2_{ex_1}, \sigma^2_{x_2} = \sigma^2_{ux_2} + \sigma^2_{ex_2}, \sigma^2 = \sigma^2_u + \sigma^2_e, \sigma^2_u = \sigma^2_{ux_1} + \sigma^2_{ux_2}, \sigma^2_e = \sigma^2_{ex_1} + \sigma^2_{ex_2}$. u_{x_1}, u_{x_2} and e_{ij} are independent. Moreover, $u_{x_1 i}$ and $u_{x_2 i}$ are correlated with covariance $\sigma_{ux_1x_2}$, and $e_{x_1 i}$ and $e_{x_2 i}$ are correlated with covariance $\sigma_{ex_1x_2}$.

Let $\rho_{x_1x_2} = \sigma_{ux_1x_2}/\sigma_{x_1}\sigma_{x_2}$ and $\mathrm{corr}(x_1, x_2) = \sigma_{x_1x_2}/\sigma_{x_1}\sigma_{x_2}$, where $\sigma_{x_1x_2} = \sigma_{ux_1x_2} + \sigma_{ex_1x_2}$ and $\mathrm{corr}(x_1, x_2)$ denote the correlation between x_{ij1} and x_{ij2}. For the nested error regression model with two covariates, the parameters are set accordingly to the previous researchers (see Wu et al. (1988), Rao et al. (1993) and Lahiri and Li (2009)). Then, without loss of generality, $\sigma^2_{x_1} = \sigma^2_{x_2} = 20$, $\rho_{x_1} = 0.1, \rho_{x_2} = 0.5$, $\rho_{x_1x_2} = 0, \mathrm{corr}(x_1, x_2) = -0.33$, $\mu_{x_1} = 100, \mu_{x_2} = 200, \beta_0 = 10, \beta_1 = \beta_2 = 0$ and $\sigma^2 = 10$.

Given (x_{ij1}, x_{ij2}), y_{ij} is generated by model (3) with five different values for intra-cluster correlation (ρ = 0, 0.05, 0.1, 0.3 and 0.5) and five different numbers of clusters (c = 3, 4, 5, 10 and 15) for the balanced data. When the data is unbalanced, there are three clusters (c = 3) and the sets of the sample size are varied. The simulated data $(y_{ij}, x_{ij1}, x_{ij2})$ are repeated 10,000 replications for all conditions and the F test statistics are computed for each replication to obtain the type I error rate and the power of the test.

3.2. Type I error rate

Type I error rate is obtained by the proportion of times that the p-value for the F test statistic is smaller than the nominal level. The measurements are the binary variables corresponding to the rejection regions of the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$. We then test the null hypothesis of no effect of the regression coefficients at 5% and 10% nominal levels and the confidence interval of the type I error rate ($\hat{\alpha}$) is calculated from $\hat{\alpha} + Z_{\alpha/2}\sqrt{\hat{\alpha}(1-\hat{\alpha})/10000}$ (see Lahiri and Li (2009)). If the nominal level is 5% and 10%, the type I error rate should not exceed 5.43% and 10.49%, respectively.

In Table 1, for the balanced data with the sample sizes (n) of 20, 30, 100 and 150, the results show that F_{AH}, F_{LAH}, F_{RAO} and F_{WU} perform as well as in controlling the type I error rate. Under the F test statistic based on generalized least squares procedures, F_{AH} performs as well as F_{RAO} in terms of the type I error rate. The F test statistic with the Fuller–Battese transformation and the F test statistic with the Adjusted Householder transformation are slightly different but they come up with the same hypothesis testing conclusion, consequently, the type I error rate of F_{RAO} is disregarded. Furthermore, the type I error rates of F_{WU} exceed the limit for large intra-cluster correlation (ρ = 0.5) and F_{OLS} leads to highly inflated the type I error rate for almost situations.

When the sample sizes of each cluster are unbalanced, the results show that F_{AH}, F_{LAH} and F_{WU} perform well in controlling the type I error rate for a small sample size (n=15) and small intra-cluster correlations ($\rho \leq$0.1) as shown in Table 2. Under the unbalanced data and large sample sizes (n=30 and 90), F_{AH}, F_{LAH} and F_{WU} maintain the nominal level for small intra-cluster correlations ($\rho \leq$0.1) when the sample sizes of each cluster are slightly different, such as (n_1, n_2, n_3)=(9,10,11), (29,30,31). While the sample sizes of each cluster are widely varied, such as (n_1, n_2, n_3)=(5,10,15), (3,3,24), (10,20,60), (3,3,84), all F test statistics lead to highly inflated type I error rate for all intra-cluster correlations, except for ρ = 0.

3.3. Power of the test

Power of the test is obtained by the proportion of times that rejects the null hypothesis when the alternative hypothesis is true at the nominal level. Table 3 reports the power of the F test statistic of the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against the specified alternative at nominal 5% and 10% levels for the balanced data. For large intra-cluster correlations, the power of F_{AH} gains over the others. For very small intra-cluster correlations ($\leq$0.05), the power of F_{AH} performs as well as F_{WU}, on the contrary, the power of F_{LAH} is the lowest as shown in Figures 1 - 4. For example, when $c_i \times n_i = 10 \times 10$ and ρ =0.05, the power of F_{LAH} is approximately 53% compared to the power of F_{AH}, which is 74%. Note that a slight decrease of the power of F_{AH} occurs when ρ increases from 0 to 0.1.

For a small sample size, the powers of F_{AH} and F_{WU} are similar when $\rho \leq 0.1$ as shown clearly in Figure 1 and 2, whereas the power of F_{AH} is higher than that of F_{WU} when $\rho \geq 0.3$. For a large sample size, Figure 3 and 4 confirm that the power of F_{AH}

is higher than that of F_{WU} and the powers of F_{AH} and F_{LAH} increase as ρ increases, while the power of F_{WU} decreases as ρ increases, theoretically corresponding to the power of test established by Rao and Wang (1995). The illustration is shown in Figure 4(b). That is, when the nominal level is 10% and the alternative hypothesis is $H_1 : \beta_1 = \beta_2 = 0.2$, the power of F_{AH} increases from 87.07% to 90.42%, but the power of F_{WU} decreases from 87.07% to 54.14% as ρ increases for $c_i \times n_i = 10 \times 10$.

Similar to the balanced data, when the sample sizes of each cluster are unbalanced, the powers of F_{AH} and F_{WU} are higher than the power of F_{LAH} for almost all situations as shown in Table 4.

Table 1. Type I error rates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ at nominal 5% and 10% levels for the balanced data

		Nominal level 5%				Nominal level 10%			
$c \times n_i$	ρ	F_{OLS}	F_{WU}	F_{LAH}	F_{AH}	F_{OLS}	F_{WU}	F_{LAH}	F_{AH}
4×5	0	5.05	5.05	4.28	5.05	10.15	10.15	8.73	10.15
	0.05	5.57*	4.83	3.69	4.88	11.23*	10.11	8.23	9.95
	0.1	6.29*	4.73	3.55	4.69	12.14*	9.75	7.70	9.72
	0.3	9.84*	4.76	4.19	4.83	16.54*	9.44	8.37	9.77
	0.5	14.74*	4.82	3.88	5.03	23.22*	9.52	8.43	9.89
3×10	0	5.20	5.20	4.42	5.20	9.84	9.84	9.24	9.84
	0.05	6.44*	4.85	4.29	4.97	12.24*	10.10	9.17	9.86
	0.1	8.15*	4.74	4.29	4.99	14.59*	9.99	9.17	9.91
	0.3	15.30*	4.37	4.29	4.96	23.04*	9.21	9.17	10.04
	0.5	22.74*	4.04	4.29	5.11	31.83*	8.72	9.17	9.97
5×20	0	4.88	4.88	4.68	4.88	9.97	9.97	9.46	9.97
	0.05	8.99*	4.75	4.82	4.73	15.71*	9.80	9.86	9.94
	0.1	13.38*	5.13	5.05	4.86	21.22*	10.00	9.86	10.06
	0.3	28.99*	5.02	4.75	4.68	38.41*	9.47	9.45	9.54
	0.5	43.21*	5.72*	5.04	5.38	51.83*	10.89*	10.20	10.27
10×10	0	5.18	5.18	4.76	5.18	10.42	10.42	10.15	10.42
	0.05	7.28*	5.43	4.53	5.12	12.74*	9.90	9.32	9.97
	0.1	9.14*	4.76	4.92	4.82	16.06*	9.67	9.93	9.86
	0.3	19.11*	5.11	4.54	4.85	27.73*	10.18	9.40	9.62
	0.5	28.47*	5.49*	4.99	5.21	37.31*	10.11	9.86	9.86
15×10	0	5.01	5.01	4.87	5.01	9.93	9.93	9.78	9.93
	0.05	7.09*	4.91	5.17	4.94	13.17*	9.93	10.07	9.79
	0.1	10.17*	5.26	5.20	5.40	16.69*	10.29	9.93	10.49
	0.3	19.15*	5.04	4.71	4.85	27.92*	9.99	9.35	9.59
	0.5	28.05*	5.50*	5.20	5.28	37.03*	10.10	10.18	10.01

*indicates that the type I error rate exceeded the limit

Table 2. Type I error rates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ at nominal 5% and 10% levels for the unbalanced data

		Nominal level 5%				Nominal level 10%			
$c \times n_i$	ρ	F_{OLS}	F_{WU}	F_{LAH}	F_{AH}	F_{OLS}	F_{WU}	F_{LAH}	F_{AH}
4,5,6	0.0	5.41	5.41	3.73	5.41	10.33	10.33	7.76	10.33
	0.05	5.50*	5.09	3.60	5.19	10.84*	10.26	7.65	10.32
	0.1	5.74*	5.00	3.27	4.91	11.10*	10.03	7.52	10.30
	0.3	8.85*	6.05*	4.47	6.90*	15.75*	11.73*	9.09	13.14*
	0.5	13.05*	7.15*	5.11	8.95*	21.18*	12.98*	10.21	15.71*
3,3,9	0.0	4.76	4.76	3.43	4.76	10.04	10.04	7.54	10.04
	0.05	5.46*	5.23	3.71	5.28	10.53*	10.19	7.93	10.13
	0.1	6.62*	6.13*	4.23	6.19*	11.87*	11.27*	8.65	11.66*
	0.3	8.85*	7.35*	5.76*	8.76*	15.75*	13.26*	11.13*	15.27*
	0.5	12.90*	9.53*	7.48*	12.09*	20.90*	16.33*	13.82*	19.90*
9,10,11	0.0	4.64	4.64	4.36	4.64	9.77	9.77	8.87	9.77
	0.05	5.90*	4.79	4.05	4.97	11.58*	9.70	9.13	9.63
	0.1	8.53*	5.79*	4.38	5.88*	14.82*	11.19*	9.49	10.93*
	0.3	15.52*	5.81*	5.32	6.89*	24.06*	11.53*	10.29	12.97*
	0.5	23.11*	6.05*	5.80*	7.79*	32.43*	11.88*	11.18*	14.09*
5,10,15	0.0	4.83	4.83	4.25	4.83	9.60	9.60	8.47	9.60
	0.05	6.32*	5.68*	5.22	6.01*	11.87*	11.11*	10.16	11.36*
	0.1	7.81*	6.54*	5.88*	7.03*	14.29*	12.42*	10.98*	13.17*
	0.3	16.03*	10.98*	10.16*	12.50*	23.78*	17.84*	16.76*	20.06*
	0.5	23.94*	13.89*	14.91*	17.97*	33.20*	22.08*	22.66*	26.56*
3,3,24	0.0	4.69	4.69	4.02	4.69	9.60	9.60	8.59	9.60
	0.05	6.19*	5.82*	5.08	5.85*	11.79*	11.55*	10.07	11.39*
	0.1	7.76*	7.19*	6.37*	7.38*	14.60*	13.77*	12.36*	13.99*
	0.3	15.66*	12.99*	11.96*	14.40*	24.21*	20.94*	19.41*	22.28*
	0.5	23.63*	17.91*	17.59*	20.93*	32.39*	26.32*	26.12*	29.77*
29,30,31	0.0	4.86	4.86	4.87	4.86	10.19	10.19	10.07	10.19
	0.05	10.22*	5.12	5.06	5.34	17.59*	10.37	9.97	10.49
	0.1	15.97*	6.02*	5.04	5.85*	24.07*	11.13*	10.29	11.27*
	0.3	32.75*	6.06*	5.68*	6.53*	41.29*	10.71*	10.80*	12.14*
	0.5	45.53*	4.89	6.15*	7.11*	53.74*	9.93	11.60*	12.68*
10,20,60	0.0	5.19	5.19	4.88	5.19	10.12	10.12	9.60	10.12
	0.05	10.50*	7.50*	7.95*	8.03*	17.17*	13.40*	13.90*	13.91*
	0.1	15.18*	9.15*	10.97*	11.04*	22.81*	15.27*	18.26*	18.30*
	0.3	32.84*	14.70*	23.37*	23.80*	41.13*	22.05*	31.27*	31.56*
	0.5	46.12*	17.70*	34.42*	35.03*	54.39*	24.33*	42.84*	43.77*
3,3,84	0.0	5.16	5.16	5.04	5.16	10.23	10.23	10.01	10.26
	0.05	10.80*	10.65*	10.00*	10.71*	17.95*	17.78*	16.86*	17.67*
	0.1	15.88*	15.61*	14.47*	15.57*	23.78*	23.32*	22.29*	23.51*
	0.3	33.25*	32.08*	30.92*	31.96*	41.91*	40.65*	39.75*	40.85*
	0.5	45.95*	43.53*	43.56*	44.61*	54.28*	52.10*	51.78*	52.87*

*indicates that the type I error rate exceeded the limit

Table 3. Power estimates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ versus specified alternatives at nominal 5% and 10% levels for the balanced data

				Nominal level 5%			Nominal level 10%		
$c \times n_i$	β_1	β_2	ρ	F_{WU}	F_{LAH}	F_{AH}	F_{WU}	F_{LAH}	F_{AH}
4×5	0.1	0.1	0	8.18	5.41	8.18	14.63	10.50	14.63
			0.05	7.90	5.00	7.91	14.88	10.52	14.78
			0.1	7.61	4.89	7.39	13.93	10.02	13.88
			0.3	7.40	5.95	7.78	13.67	11.34	14.50
			0.5	7.25	6.41	8.63	13.43	12.20	15.44
	0.2	0.2	0	18.30	9.28	18.30	28.02	16.51	28.02
			0.05	18.19	9.39	17.91	28.21	17.27	28.04
			0.1	16.93	9.83	16.80	26.63	17.73	27.16
			0.3	15.38	12.02	17.29	25.32	20.24	27.62
			0.5	15.20	15.18	21.13	24.13	25.37	32.14
3×10	0.1	0.1	0	9.38	6.80	9.38	17.02	12.93	17.02
			0.05	9.55	7.37	9.48	16.50	14.02	16.52
			0.1	9.34	7.09	9.11	15.79	13.46	16.25
			0.3	8.17	8.07	9.53	14.71	14.52	17.03
			0.5	-	10.02	11.31	-	17.26	18.75
	0.2	0.2	0	26.02	15.71	26.02	37.86	25.40	37.86
			0.05	24.55	16.92	24.19	36.20	27.36	36.01
			0.1	23.21	16.94	23.93	34.16	27.42	35.42
			0.3	19.32	21.36	25.76	29.40	32.44	37.35
			0.5	-	29.28	32.19	-	41.92	45.22
5×20	0.1	0.1	0	25.94	16.45	25.94	37.39	25.86	37.39
			0.05	20.84	16.40	21.48	31.47	26.33	32.06
			0.1	18.13	17.29	20.97	28.02	27.34	31.13
			0.3	14.13	22.11	23.83	21.23	32.29	34.65
			0.5	-	28.62	30.15	-	40.66	41.95
	0.2	0.2	0	76.63	53.93	76.63	84.63	66.49	84.63
			0.05	65.53	55.20	68.96	75.89	67.79	79.17
			0.1	58.24	58.35	67.68	69.56	70.21	77.99
			0.3	39.62	69.27	72.96	50.81	79.21	82.22
			0.5	-	83.74	85.32	-	90.69	91.72
10×10	0.1	0.1	0	26.73	15.21	26.73	38.15	24.72	38.15
			0.05	23.88	16.43	23.91	34.81	25.35	35.13
			0.1	22.65	17.44	23.71	33.89	27.58	35.35
			0.3	17.86	21.33	25.23	26.34	31.74	36.14
			0.5	-	26.76	29.42	22.16	38.90	41.54
	0.2	0.2	0	79.46	51.02	79.46	87.07	63.60	87.07
			0.05	73.96	52.92	74.30	82.52	65.02	83.35
			0.1	69.77	56.65	73.14	80.17	68.67	82.64
			0.3	53.22	66.54	74.73	64.90	77.44	83.65
			0.5	-	80.28	84.00	54.14	87.68	90.42

-indicates that power of the test cannot be compared

Table 3. Power estimates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ versus specified alternatives at nominal 5% and 10% levels for the balanced data (cont.)

				Nominal level 5%			Nominal level 10%		
$c \times n_i$	β_1	β_2	ρ	F_{WU}	F_{LAH}	F_{AH}	F_{WU}	F_{LAH}	F_{AH}
15×10	0.1	0.1	0	39.85	21.93	39.85	52.53	33.09	52.53
			0.05	35.06	22.85	35.22	47.41	33.09	47.74
			0.1	31.40	23.75	33.93	43.96	35.84	46.54
			0.3	23.01	29.53	34.22	32.92	41.73	47.22
			0.5	-	39.55	43.25	28.74	52.81	56.39
	0.2	0.2	0	93.48	69.79	93.48	96.56	80.00	96.56
			0.05	90.04	72.15	90.59	94.54	82.09	95.01
			0.1	86.16	74.57	89.01	91.89	83.93	93.76
			0.3	70.50	85.17	90.52	79.78	91.41	94.76
			0.5	-	94.34	96.12	70.21	97.13	98.12

-indicates that power of the test cannot be compared

Table 4. Power estimates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ versus specified alternatives at nominal 5% and 10% levels for the unbalanced data

				Nominal level 5%			Nominal level 10%		
n_1, n_2, n_3	β_1	β_2	ρ	F_{WU}	F_{LAH}	F_{AH}	F_{WU}	F_{LAH}	F_{AH}
4,5,6	0.1	0.1	0	7.11	4.49	7.11	13.59	7.38	13.59
			0.05	6.56	4.11	6.51	12.56	7.01	12.64
			0.1	7.21	4.52	7.10	13.90	7.96	14.02
			0.3	-	5.59	-	-	9.96	-
			0.5	-	7.21	-	-	13.35	-
	0.2	0.2	0	13.27	9.22	13.27	22.12	14.38	22.12
			0.05	12.42	8.77	12.50	21.59	14.16	21.73
			0.1	13.70	9.73	13.78	22.98	15.23	23.28
			0.3	-	11.07	-	-	18.28	-
			0.5	-	13.73	-	-	22.43	-
3,3,9[a]	0.1	0.1	0	6.63	4.33	6.63	13.53	7.82	13.53
			0.05	6.83	4.59	6.82	14.08	8.53	13.90
			0.1	-	5.55	-	-	9.93	-
	0.2	0.2	0	13.13	9.26	13.13	22.40	15.60	22.40
			0.05	15.81	12.76	16.00	35.21	27.59	35.43
			0.1	-	11.04	-	-	17.28	-
9,10,11[b]	0.1	0.1	0	9.47	6.93	9.47	26.05	17.06	26.05
			0.05	9.06	6.58	8.91	23.75	16.79	23.87
			0.1	-	7.95	-	-	18.73	-
			0.3	-	9.57	-	-	23.76	-
	0.2	0.2	0	17.08	13.21	17.08	37.99	26.94	37.99
			0.05	17.03	13.73	16.97	36.61	28.11	36.86
			0.1	-	14.19	-	-	29.18	-
			0.3	-	16.11	-	-	34.60	-

-indicates that power of the test cannot be compared

For $H_1 : \beta_1 = \beta_2 = 0.1$ and $H_1 : \beta_1 = \beta_2 = 0.2$,

[a] when ρ = 0.3 and 0.5, all F-statistics cannot control the type I error rate,

[b] when ρ = 0.5, all F-statistics cannot control the type I error rate

Table 4. Power estimates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ versus specified alternatives at nominal 5% and 10% levels for the unbalanced data (cont.)

				Nominal level 5%			Nominal level 10%		
n_1, n_2, n_3	β_1	β_2	ρ	F_{WU}	F_{LAH}	F_{AH}	F_{WU}	F_{LAH}	F_{AH}
5,10,15[c]	0.1	0.1	0	9.38	7.73	9.38	25.64	19.10	25.64
			0.05	-	9.20	-	-	21.60	-
	0.2	0.2	0	16.58	13.93	16.58	37.61	30.03	37.61
			0.05	-	16.14	-	-	32.97	-
3,3,24[c]	0.1	0.1	0	9.35	7.92	9.35	25.36	20.95	25.36
			0.05	-	8.95	-	-	22.23	-
	0.2	0.2	0	16.30	14.33	16.30	37.15	31.68	37.15
			0.05	-	15.76	-	-	33.09	-
29,30,31	0.1	0.1	0	22.06	15.51	22.06	66.77	50.09	66.77
			0.05	18.79	16.52	19.84	56.97	52.79	61.31
			0.1	-	16.58	-	-	54.96	-
			0.3	-	-	-	-	-	-
			0.5	10.33	-	-	28.31	-	-
	0.2	0.2	0	32.76	24.81	32.76	77.10	63.12	77.10
			0.05	28.29	25.82	30.38	68.79	65.08	72.92
			0.1	-	26.16	-	-	67.57	-
			0.3	-	-	-	-	-	-
			0.5	17.41	-	-	38.58	-	-
10,20,60[d]	0.1	0.1	0	22.26	18.53	22.26	68.03	58.74	68.03
	0.2	0.2	0	33.35	28.68	33.35	77.79	70.11	77.79
3,3,84[d]	0.1	0.1	0	22.00	21.31	22.00	66.90	64.93	66.91
	0.2	0.2	0	32.80	31.34	32.79	77.35	75.29	77.35

-indicates that power of the test cannot be compared

For $H_1 : \beta_1 = \beta_2 = 0.1$ and $H_1 : \beta_1 = \beta_2 = 0.2$,

[c] when ρ = 0.1, 0.3 and 0.5, all F-statistics cannot control the type I error rate,

[d] when ρ = 0.05, 0.1, 0.3 and 0.5, all F-statistics cannot control the type I error rate

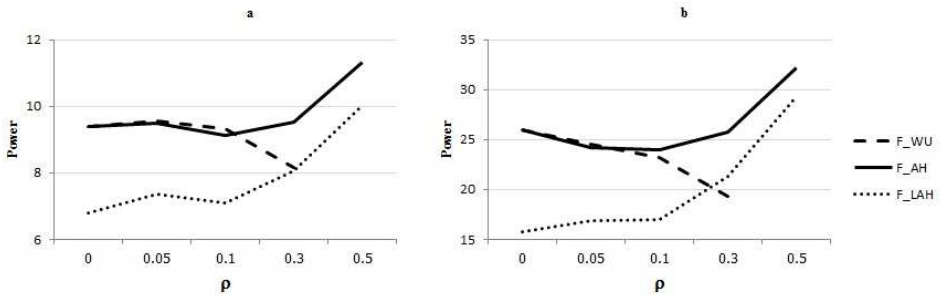


Figure 1: Power estimates (%) of F test statistic versus intra-cluster correlation at nominal 5% level and $c \times n_i = 3 \times 10$ corresponding to the alternatives hypothesis (a) $H_1 : \beta_1 = \beta_2 = 0.1$ and (b) $H_1 : \beta_1 = \beta_2 = 0.2$

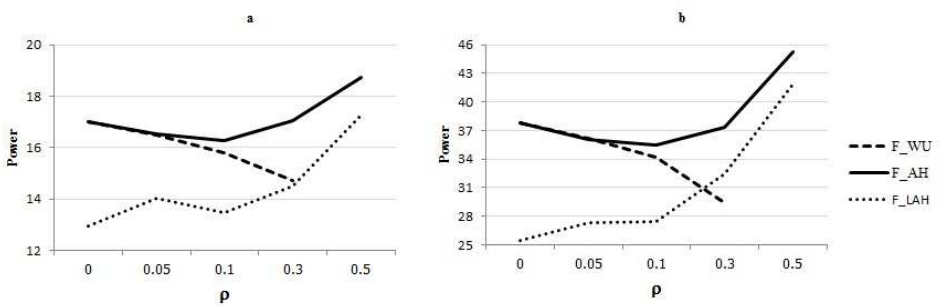


Figure 2: Power estimates (%) of F test statistic versus intra-cluster correlation at nominal 10% level and $c \times n_i = 3 \times 10$ corresponding to the alternatives hypothesis (a) $H_1 : \beta_1 = \beta_2 = 0.1$ and (b) $H_1 : \beta_1 = \beta_2 = 0.2$

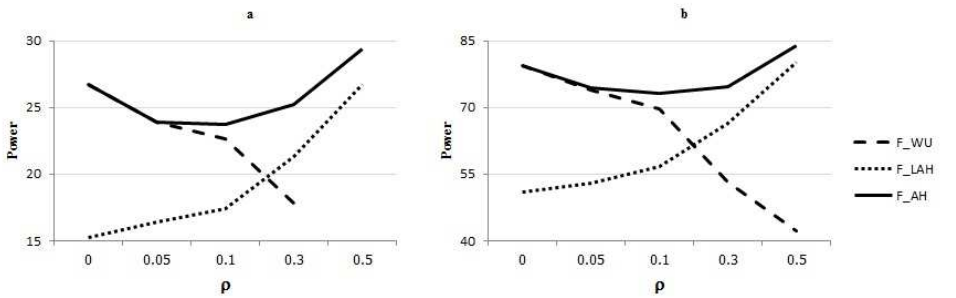


Figure 3: Power estimates (%) of F test statistic versus intra-cluster correlation at nominal 5% level and $c \times n_i = 10 \times 10$ corresponding to the alternatives hypothesis (a) $H_1 : \beta_1 = \beta_2 = 0.1$ and (b) $H_1 : \beta_1 = \beta_2 = 0.2$

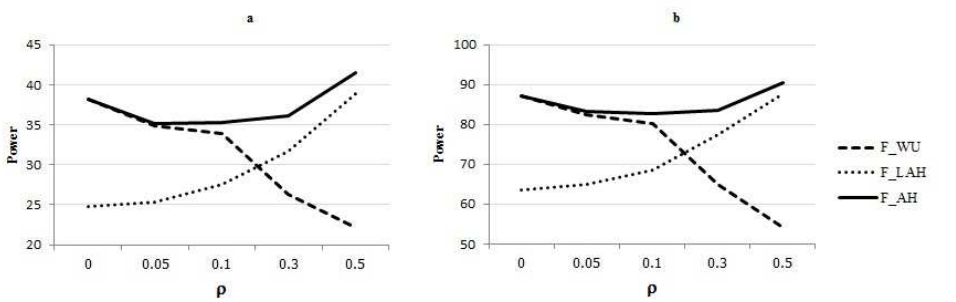


Figure 4: Power estimates (%) of F test statistic versus intra-cluster correlation at nominal 10% level and $c \times n_i = 10 \times 10$ corresponding to the alternatives hypothesis (a) $H_1 : \beta_1 = \beta_2 = 0.1$ and (b) $H_1 : \beta_1 = \beta_2 = 0.2$

3.4. An application

In this section, we consider the data from Smith (1980). This data set covers the values of pattern intensity on soles of 14 families chosen from Polish family data (see Table 5). The families consist of siblings, together with their mothers and fathers. Here, $\mathbf{y}$ is the 49×1 vector of values of pattern intensity on soles of feet of siblings and $\mathbf{X}$ is the 49×3 matrix of ones in the first column and values of pattern intensity on soles of feet of mother and father. In the real data, the intra-cluster correlation is usually unknown and it must be estimated for the F test statistic. The Srivastava estimator of the intra-cluster correlation, 0.4922, (see Srivastava and Katapa (1986)) is applied in this section. In order to test the regression coefficients for the nested error regression model, the p-value of F_{AH}, F_{WU}, F_{LAH} and F_{OLS} are less than 0.05, then we reject the null hypothesis (H_0) at the significance level of 5%. This indicates that at least one regression coefficient is significant to the model. Note that the intra-cluster correlation estimator, the important characteristic of the clustered data, is not used to compute F_{OLS} and F_{LAH}. In addition, the errors in observations of F_{OLS} do not correspond to the assumption of regression analysis even the power of the test is quite high. Therefore, the F_{AH} and F_{WU} using Srivastava estimator of the intra-cluster correlation are suggested and the power of F_{AH} is higher than F_{WU} for applying to this application.

Table 5. Values of pattern intensity on soles of feet in 14 families

Family no.	Mother	Father	Siblings
1	2	3	2,2
2	2	3	2,3
3	2	3	2,2,2
4	2	4	2,2,2,2,2
5	6	7	6,6
6	4	3	4,3,3
7	4	3	2,2,3,6,3,5,4
8	3	7	2,4,7,4,4,7,8
9	5	5	5,6
10	5	4	4,5,4
11	5	6	5,3,4,4
12	2	4	2,4
13	6	3	4,3,3,3
14	2	3	2,2,2

Table 6. The F test statistics, p-values and powers for data set in Table 5

Method	F	p-value	power
F_{OLS}	21.2343	0.0000003	0.9999997
F_{WU}	4.7388	0.0135	0.9865
F_{LAH}	7.4865	0.0021	0.9979
F_{AH}	6.4756	0.0033	0.9967

4. Conclusion

For clustered data analysis with compound symmetry correlation structure of known intra-cluster correlation, the proposed transformation by Adjusted Householder matrix can be used to adjust the correlation of the error term and then allowed to be applied to the F test statistic based on generalized least squares procedures. The simulation study shows that the F test statistic with Adjusted Householder transformation performs as well as the other methods for the balanced and unbalanced data, except for the F test statistic based on standard ordinary least squares procedures, in controlling the type I error rate regarding regression coefficients hypothesis testing for small and large sample sizes. The power of the F test statistic with Adjusted Householder transformation is always higher than that with part of the Helmert transformation for the balanced and unbalanced data. Also, the power of the F test statistic with Adjusted Householder (F_{AH}) and part of the Helmert (F_{LAH}) transformations are the increasing functions of the intra-cluster correlation whereas the power of the modification of the F test statistic (F_{WU}) is the decreasing function.

REFERENCES

AQARWAL, G. G., AWASTHI, S., WALTER, S. D., (2005). Intra-class Correlation Estimates for Assessment of Vitamin A Intake in Children. Journal of Health, Population, and Nutrition, 23 (1), pp. 66–73.

BATTESE, G. E., HARTER, R. M., FULLER, W. A., (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. Journal of the American Statistical Association, 83 (401), pp. 28–36.

ELDRIDGE, S. M., UKOUMUNNE, O. C., CARLIN, J. B., (2009). The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. International Statistical Review, 77 (3), pp. 378–394.

FARHADIAN, R., ASADIAN, N., (2017). On the Helmert Matrix and Application in Stochastic Processes. International Journal of Mathematics and Computer Science, 12 (2), pp. 107–115.

GALBRAITH, S., DANIEL, J. A., VISSEL, B., (2010). A Study of Clustered Data and Approaches to Its Analysis. The Journal of Neuroscience, 30 (32), pp. 10601–10608.

HOUSEHOLDER, A. S., (1958). Unitary Triangularization of a Nonsymmetric Matrix. Journal of the ACM, 5, pp. 339–342.

LAHIRI, P., LI, Y., (2009). A New Alternative to the Standard F Test for Clustered Data. Journal of Statistical Planning and Inference, 139, pp. 3430–3441.

LANCASTER, H. O., (1965). The Helmert Matrices. The American Mathematical Monthly, 72, pp. 4–12.

MCCULLOCH, C. E., SHAYLE, R. S., (2001). Generalized, linear, and mixed models, New York: John Wiley and Sons.

MIALL, W. E., OLDHAM, P. D., (1955). A Study of Arterial Blood Pressure and Its Inheritance in a Sample of the General Population. Clinical Science, 14 (3), pp. 459–488.

NG, S. K., MCLACHLAN, G. J., YAU, K. K. W., LEE, A. H., (2004). Modeling the Distribution of Ischaemic Stroke-specific Survival Time using an EM-based Mixture Approach with Random Effects Adjustment. Statistics in Medicine, 23, pp. 2729–2744.

RAO, J. N. K., SUTRADHAR, B. C., YUE, K., (1993). Generalized Least Squares F Test in Regression Analysis with Two-Stage Cluster Samples. Journal of the American Statistical Association, 88 (424), pp. 1388–1391.

RAO, J. N. K., WANG, S. G., (1995). On the Power of F Tests under Regression Models with Nested Error Structure. Journal of Multivariate Analysis, 53, pp. 237–246.

SMITH, C. A. B., (1980). Estimating Genetic Correlations. Ann. Human Genetics, 43, pp. 265–284.

SRIVASTAVA, M. S., KATAPA, R. S., (1986). Comparison of Estimators of Interclass and Intraclass Correlations from Familial Data. The Canadian Journal of Statistics, 14 (1), pp. 29–42.

WU, C. F. J., HOLT, D., HOLMES, D. J., (1988). The Effect of Two-Stage Sampling on the F Statistic. Journal of the American Statistical Association, 83 (401), pp. 150–159.

APPENDIX

Householder matrix

For any $n \times n$ symmetric matrix, in this paper we consider the variance-covariance matrix $\mathbf{V} = [v_{ij}]; i = 1,2,\ldots,n$ and $j = 1,2,\ldots,n$, which is the compound symmetry correlation structure. The corresponding Householder matrix, $\mathbf{H}$, is a symmetry and orthogonal matrix in the form

$$\mathbf{H} = \mathbf{I}_n - 2\mathbf{w}\mathbf{w}'.$$

Let $\mathbf{w} = (w_1,\ldots,w_n)'$ be a column vector which is a unit vector of Euclidean norm where

$$w_1 = 0, w_2 = \frac{v_{21} - \alpha}{2\gamma} \quad \text{and} \quad w_l = \frac{v_{l1}}{2\gamma}; l = 3,\ldots,n.$$

α and γ are determined by

$$\alpha = -\text{sgn}(v_{21})\sqrt{\Sigma_{i=2}^{n} v_{i1}^2} \quad \text{where} \quad \text{sgn}(v_{21}) = \begin{cases} -1 & \text{for} \quad v_{21} < 0 \\ +1 & \text{for} \quad v_{21} > 0 \end{cases}$$

and $\gamma = \sqrt{\frac{1}{2}(\alpha^2 - v_{21}\alpha)}$.

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 171–188, DOI 10.21307/stattrans-2019-010

SURVIVAL REGRESSION MODELS FOR SINGLE EVENTS AND COMPETING RISKS BASED ON PSEUDO-OBSERVATIONS

Ewa Wycinka[1], **Tomasz Jurkiewicz**[2]

ABSTRACT

Survival data is a special type of data that measures the time to an event of interest. The most important feature of survival data is the presence of censored observations. An observation is said to be right-censored if the time of the observation is, for some reason, shorter than the time to the event. If no censoring occurs in the data, standard statistical models can be used to analyse the data. Pseudo-observations can replace censored observations and thereby allow standard statistical models to be used.
In this paper, a pseudo-observation approach was applied to single-event and competing-risks analysis, with special attention paid to the properties of the pseudo-observations. In the empirical part of the study, the use of regression models based on pseudo-observations in credit-risk assessment was investigated. Default, defined as a delay in payment, was considered to be the event of interest, while prepayment of credit was treated as a possible competing risk. Credits that neither default nor are prepaid during the follow-up were censored observations. Typical application characteristics of the credit and creditor were the covariates in the regression model. In a sample of retail credits provided by a Polish financial institution, regression models based on pseudo-observations were built for the single-event and competing-risks approaches. Estimates and discriminatory power of these models were compared to the Cox PH and Fine-Gray models.

Key words: generalised estimating equations, cumulative incidence function, probability of default, credit risk, survival analysis.

1. Introduction

In the past few decades, survival analysis methods have become more widely used, not only in biostatistics, where their roots are, but also in many other branches of science, including economics, and the social sciences. Survival analysis is a term that covers a vast collection of different methods that focus on timing and duration prior to an event's occurrence (Mills, 2011). Among these methods are parametric and non-parametric estimation of survival time

[1] University of Gdańsk, Faculty of Management. E-mail: ewa.wycinka@ug.edu.pl. ORCID ID: https://orcid.org/0000-0002-5237-3488.
[2] University of Gdańsk, Faculty of Management. E-mail: tomasz.jurkiewicz@ug.edu.pl. ORCID ID: https://orcid.org/0000-0001-7066-5196.

distributions, and parametric and semiparametric regression models. The common goal of these methods is to handle censored observations that are inevitable in time-to-event analysis. A quite new and innovative approach to the problem of censoring is the idea of pseudo-observations that can replace both complete and censored actual observations. Pseudo-observations can be applied to many different objectives; this paper focuses on the usefulness of pseudo-observations in the development of regression models for survival functions in the case that a single event is analysed and in the competing risk analysis. The first objective was to review the properties of pseudo-observations in these two situations. The second goal was to compare the results and performance of the regression models for pseudo-observations with some of the more classical survival models that are currently most popular – in this case, the Cox Proportional Hazards model for single events (Cox, 1972) and the Fine-Gray model for competing risks (Fine and Gray, 1999).

2. Pseudo-observations for single events and competing risks

The methodology of pseudo-observations was first proposed by Andersen et al. (2003). The main idea of this approach is to replace censored observations by the function of event times $f(T)$, for which an expected value is $E(f(T))$. The condition is that an unbiased estimator $\hat{\theta}$ of $\theta = E(f(T))$ exists. Let n be the sample size $(i = 1, \dots, n)$. A pseudo-observation for $f(T)$ for individual i at a predefined series of time points $t = 1, \dots, H$ is defined as

$$\hat{\theta}_i(t) = n\hat{\theta}(t) - (n-1)\hat{\theta}^{(-i)}(t) \tag{1}$$

and is evaluated by the leave-one-out method. $\hat{\theta}(t)$ is the estimator in the sample of size n at time t, and $\hat{\theta}^{(-i)}(t)$ is the estimator at time t in the sample of size $n-1$, consisting of all units except the i-th individual. The pseudo-observation is then a contribution of the i-th unit to the $E(f(T))$ estimate in the sample of size n. Although the aim of using pseudo-observations is to replace the censored observations, pseudo-observations are calculated for all units in the sample (both completed and censored observations). Therefore, an $n \times H$ matrix of pseudo-observations is obtained. Subsequently, pseudo-observations are used as dependent variables in a generalised regression model with some link function g:

$$g\big(E(f(t)|X)\big) = \beta_0 + \sum \beta_j X_j = \beta^T X. \tag{2}$$

For each unit H pseudo-observations are calculated. Multiple measurement is a source of correlation in the data set; a possible solution to this deficiency would be to use generalised estimating equations (GEE), which are the generalisation of regression models for the case of correlated data (Andersen et al., 2003).

2.1. Single event

Assume that there is only one type of event and T is the time to that event, while T_C is the time to censoring. Due to the right censoring, we can observe $\min(T, T_C)$. The survival function is the probability that the unit does not experience the event until time t

$$S(t) = P(T > t). \tag{3}$$

In the survival analysis to the assumed sole type of event (single event), the survival function $S(t)$ can be estimated with the use of the Kaplan-Meier (KM) estimator

$$\hat{S}(t) = \prod_{t_j \le t}(1 - \frac{D_j}{N_j}), \tag{4}$$

where D_j is the number of events at time t_j, N_j is the number at risk just prior to time t_j, and t_j for $j = 1, \ldots, r$ $(r \le n)$ are distinct event times. The KM estimator is a maximum likelihood estimator (Klein and Moeschberger, 2003).

The i-th pseudo-observation based on the survival function is

$$\hat{\theta}_i(t) = n\hat{S}(t) - (n-1)\hat{S}^{(-i)}(t), \tag{5}$$

where $\hat{S}(t)$ is the estimated survival function at time t in a sample of size n and $\hat{S}^{-i}(t)$ is the estimated survival function derived from the $n-1$ sample (without the i-th observation) (Andersen and Perme 2010). At $t = 0$, the pseudo-observations for survival functions for all units are equal to one.

As t increases, the values of pseudo-observations for units in the cohort increase at each event time observed in the cohort (see Figure 1). Between any two successive event times, the values of pseudo-observations do not change. As a result, the curve of pseudo-observations over time for a particular unit is a step function with a varying length of steps depending on the successive event times. If the event for a unit is observed, the pseudo-observation drops below zero at the event time. At the subsequent time points, the unit that has just been excluded from the cohort has negative and increasing pseudo-values. If the unit is censored, then, beginning at the next event time after censoring, the values of pseudo-observations for that unit start decreasing. They remain, however, positive until the end of the follow-up (see Figure 1).

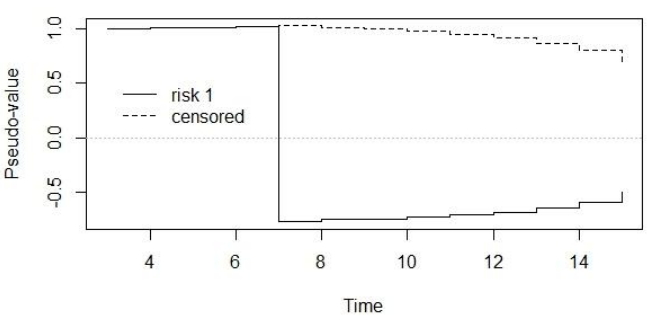


Figure 1. The pseudo-observations for the survival function over time in a censored data set for the individual with event time t=7 (risk 1) and the individual with censored time t_c=7 (censored)

As long as the units are in the cohort, they have similar pseudo-values. The values of pseudo-observations increase at each event time observed in the cohort. Therefore, the value of the pseudo-observation for the unit at its event

time is greater if the event occurred later in time. The later the event occurs, the greater the drop is (see Figure 2). The same pattern is observed if the observation is censored (see Figure 3).

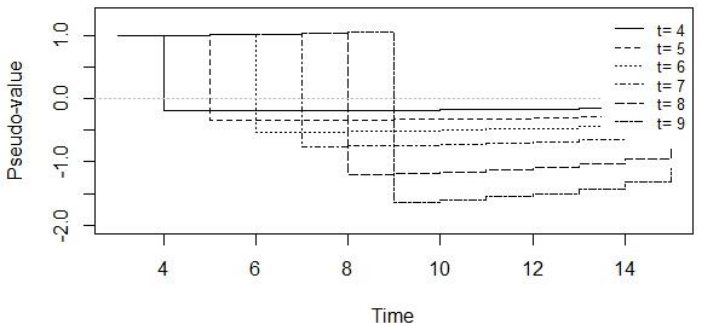


Figure 2. Pseudo-observations for the survival function over time for the units with event times t=4,..9

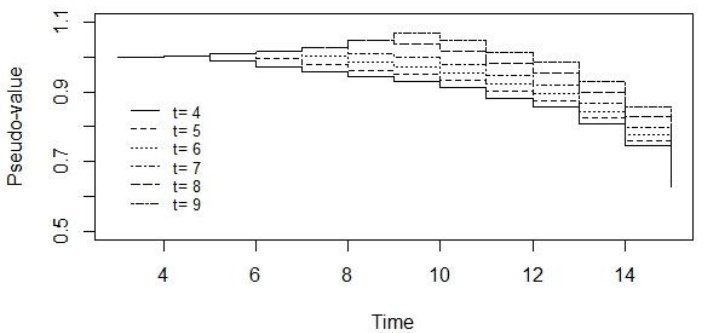


Figure 3. Pseudo-observations for the survival function over time for the units with censoring times C=4,..9

In the absence of censoring, the pseudo-value at time t reduces to the indicator that $T > t$. Therefore, the pseudo-observations are equal as long as the unit is observed in the cohort; after the event, the value of the pseudo-observation falls to zero and is constant until the end of the follow-up (see Figure 4). In this case, pseudo-observations are also independent.

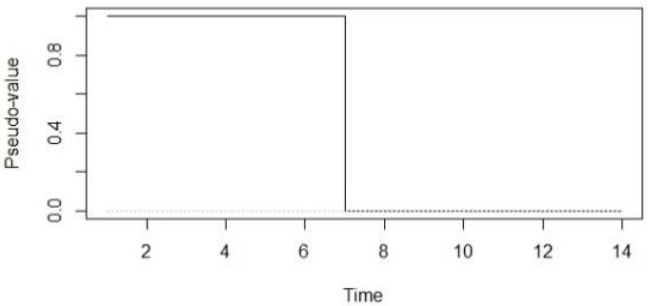


Figure 4. Pseudo-observations for the survival function over time for an individual with a survival time t=7 in a data set with no censoring

2.2. Competing risks

Let (T, C) be a bivariate random variable, such that T is a continuous variable representing the time of the first event, and $C = k$ $(k = 1, \dots, p)$ is a discrete variable denoting the type of event. If the time of the observation for some units is shorter than the time of the first event, we encounter right censoring. In such a situation, $C = 0$ and T_c is the time at which the observation was censored; what we only know is that $T > T_c$. Due to the right censoring, the variable (T, C) is only partially observable, and we observe a pair $(\min\{T, T_c\}, C)$. As a result, the joint distribution of (T, C) is difficult to identify and can be estimated only by making some unverifiable assumptions (Pintilie, 2006, p. 41).

The subdistribution of event k (cumulative incidence function, CIF) is the probability, until time t, that event k will occur

$$F_k(t) = P(T \leq t, C = k). \tag{6}$$

The subdistribution is not a proper distribution because

$$\lim_{t \to \infty} F_k(t) = P(C = k) \leq 1. \tag{7}$$

The equality $P(C = k) = 1$ holds if there is only one type of event (no competing risks). The sum of the subdistributions for all types of events is a marginal distribution of the variable T

$$F(t) = P(T \leq t) = \sum_{k=1}^{p} F_k(t). \tag{8}$$

The maximum likelihood estimator of the subdistribution is

$$\hat{F}_k(t) = \sum_{t_j \leq t} \hat{h}_{kj} \hat{S}(t_{j-1}), \tag{9}$$

where $\hat{h}_{kj}$ is the cause-specific hazard at time t_j for event k. This can be defined as $\hat{h}_{kj} = \frac{D_{kj}}{N_j}$, where D_{kj} is the number of events of type k at time t_j, N_j is the number at risk just prior to time t_j, and t_j, for $j = 1, \dots, r$ $(r \leq n)$ are distinct event times. $\hat{S}(t_{j-1})$ is the survival function for all types of events just before time t_j.

It is worth noting that the estimate depends not only on the number of individuals who have experienced the k-th type of event, but also on the number of individuals who have not experienced any type of event (Binder et al., 2014). Usually, only one type of event is of interest and other types of event are treated as competing risks to it. In such a situation, it is reasonable to consider only two types of event: the event of interest (risk 1) and every other event combined (risk 2). This approach will be considered later in this paper.

The pseudo-observation for the unit i, at time t, for the event type k, based on the CIF, has the form

$$\hat{\theta}_{ik}(t) = n\hat{F}_k(t) - (n-1)\hat{F}_k^{(-i)}(t). \tag{10}$$

Here, $\hat{F}_k(t)$ is the estimated CIF for the k-th event at time t using all observations, and $\hat{F}_k^{(-i)}(t)$ is the estimated CIF derived from all but the i-th observation. When units are in a cohort, have the same pseudo-observation values for the CIF at subsequent times. At $t = 0$, a pseudo-observation for the CIF equals zero. Then, as time increases, pseudo-observations decrease, taking negative values. Figure 4 shows pseudo-observations over time for a unit that leaves the cohort at time 7. If the unit leaves the cohort due to an event of type 1, the pseudo-observation jumps above one at the time of the event, and then, at subsequent times, gradually decreases towards one. When the unit leaves the cohort due to an event of type 2, the pseudo-values remain negative and decreasing at all subsequent times (Andersen and Perme, 2010). If an individual is censored at time t, the pseudo-observations start increasing as of the next event time recorded in the data set (see Figure 5).

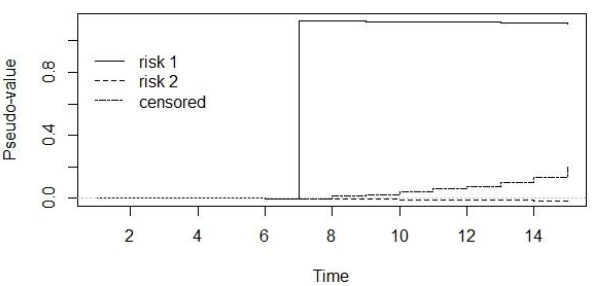


Figure 5. A comparison of the development of pseudo-values over time for three units that leave a cohort at time 7 due to either risk 1, risk 2, or censoring

As we compare the pseudo-observations for units with the same cause of leaving a cohort but at different times, we can see greater changes in the pseudo-values for later departures (see Figure 6). Jumps in the values of pseudo-observations are higher if the event of type 1 happens later, due to the reduction in time of the risk set.

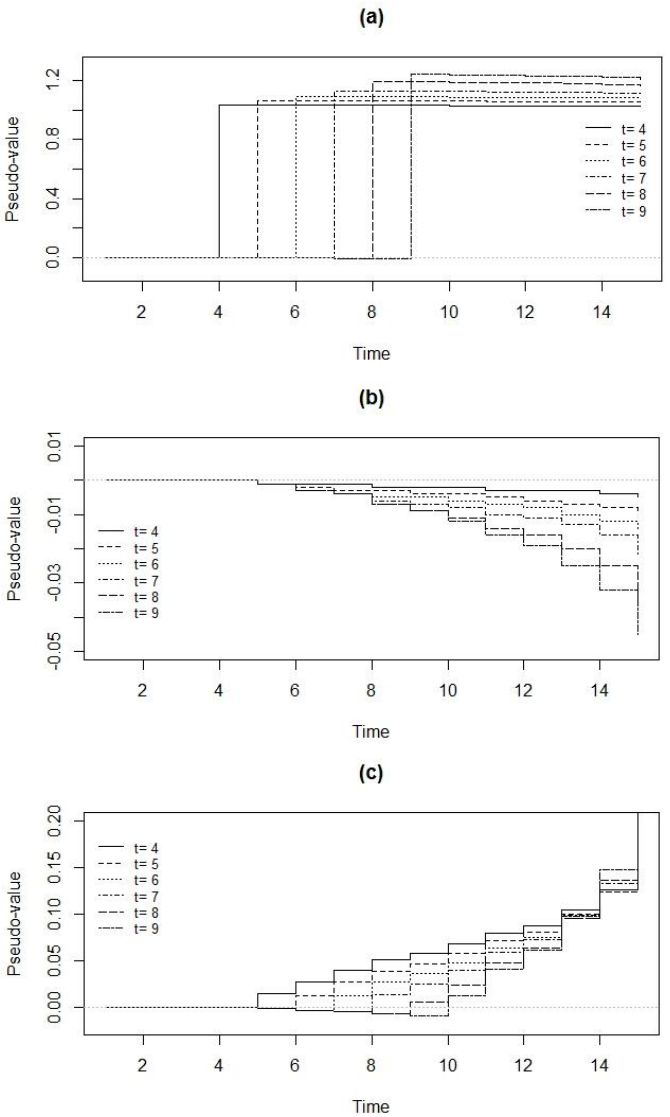


Figure 6. Pseudo-observations for units that experienced (a) risk 1 or (b) risk 2 or were (c) censored at different times (t=4,...9). Different scales are used in each figure

In a special case with no competing risks, the estimated CIF for the type-1 event reduces to the estimation of the distribution function ($\hat{F}_{CIF}$), and the survival function can be estimated as $\hat{S}_{CIF}(t) = 1 - \hat{F}_{CIF}(t)$. If survival functions are estimated directly with a Kaplan-Meier estimator or as $\hat{S}_{CIF}(t)$, the estimations are equal. However, as we show in the empirical part of the study, in the case of pseudo-observations based on these estimators, this equality no longer holds.

If no censoring occurs in the data set, the pseudo-observations for risk 1 reduce to the indicator $F_1(t) = 1[T_1 \leq t]$. They equal zero as long as the unit is in the cohort and rise towards one as the event of type 1 (risk 1) happens. The pseudo-observations for risk 2 equal zero at all time points, even after the occurrence of the event of type 2 (risk 2) (see Figure 7).

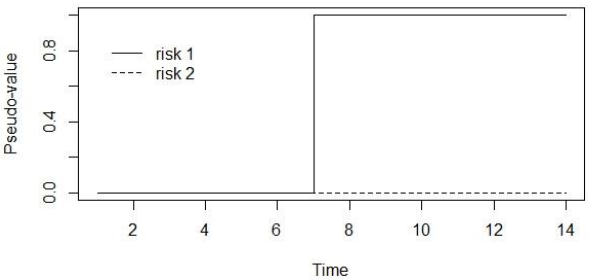


Figure 7. Pseudo-observations for the CIF for risk 1 and risk 2 over time in a data set with no censoring

3. Regression models based on pseudo-observations

For each unit there are H pseudo-observations – one for each predefined point in time. As a result, the data transformed into pseudo-observations is no longer independent, and generalised linear models (GLM) cannot be applied. Generalised estimating equations (GEEs) are the generalisation of GLM models for correlated data, as introduced by Liang and Zeger (1986). This is a method for analysing data collected in clusters where observations within a cluster may be correlated, but observations from different clusters are independent. The variance is a function of the expectation, and a monotone transformation of the expectation is linearly related to the explanatory variable (Højsgaard et al., 2005). The pseudo-observations are dependent variables in GLMs for a given link function $g(.)$. The regression model is

$$g\left(\hat{\theta}_k(t)\middle|X\right) = \beta_0 + \sum_{j=1}^{m+H} \beta_j X_j^* \tag{11}$$

Here, the vector X^* includes indicators of time points $X = (X_{m+1}, \ldots, X_{m+H})$ for $t = 1, \ldots, H$ (as dummy variables), as well as the covariates $X = (X_1, \ldots, X_m)$. When a complementary log-log link function is used, such as $g(x) = \log(-\log(x))$ for a single event, then the regression model has the form

$$\log\left(-\log\left(S(t|X)\right)\right) = \beta_0 + \sum_{j=1}^{m+H} \beta_j X_j^* \tag{12}$$

and can be depicted as

$$S(t|X) = \exp(-\exp(\beta_0 + \sum_{j=1}^{m+H} \beta_j X_j^*)). \tag{13}$$

Estimated coefficients for time points can be put into the model as time-dependent coefficients $\beta_0(t)$:

$$S(t|X) = \exp(-\exp(\beta_0 + \beta_0(t) + \sum_{j=1}^{m} \beta_j X_j)). \tag{14}$$

Finally, the survival function can be expressed as

$$S(\mathrm{t}|\mathrm{X}) = S_0(\mathrm{t})^{\exp\sum_{j=1}^{m}\beta_j X_j}, \tag{15}$$

which is a formula for the Cox PH model. Coefficients can be interpreted as a logarithm of a proportional hazards ratio.

In the case of competing risks, the link function $g(x) = \log(-\log(1-x))$ is used, and the regression model has the form

$$\log\big(-\log\big(1-F_k(t|X)\big)\big) = \beta_0 + \sum_{j=1}^{m+H}\beta_j X_j^*, \tag{16}$$

which can also be expressed in the form

$$F_k(t|X) = 1 - \exp(-\exp(\beta_0 + \beta_0(t) + \sum_{j=1}^{m}\beta_j X_j)). \tag{17}$$

This form is analogous to the proportional hazard model on the subdistribution hazard function in the Fine-Gray model. Coefficients β_j can be interpreted as logarithms of the subdistribution hazard ratios, if all covariates are time independent (Haller et al., 2013, p. 44).

Estimations of the parameters are based on the estimating equations

$$\sum_i \left(\frac{\partial}{\partial\beta} g^{-1}(\beta^T X_i^*)\right)^T V_i^{-1}\left(\hat{\theta}_i - g^{-1}(\beta^T X_i^*)\right) = 0. \tag{18}$$

Here, V_i is a working covariance matrix. The efficiency of the estimators depends on the choice of V_i matrix, which should resemble the true covariance. The GEE method fits marginal mean models and, as a result, only the correct specification of marginal means is required for the parameter estimation to be consistent and asymptotically normal (Højsgaard et al., 2005). The covariance structure does not need to be specified correctly; however, it is necessary to make an assumption about the type of this structure (considered the *working covariance matrix* or *working correlation matrix*). Four different types of working correlation matrix are usually considered.

The simplest – the independent working correlation structure – assumes that $\rho_{t_1,t_2} = corr\left(\hat{\theta}(t_1), \hat{\theta}(t_2)\right) = 0$ for each pair $\left(\hat{\theta}(t_1), \hat{\theta}(t_2)\right)$ and $t_1 \neq t_2$. The compound symmetry (exchangeable) structure treats $\rho_{t_1,t_2} = corr\left(\hat{\theta}(t_1), \hat{\theta}(t_2)\right)$ for all pairs as equal but unknown. The autoregressive structure of order 1 (AR1) has the form $corr\left(\hat{\theta}(t_1), \hat{\theta}(t_2)\right) = \rho^{t_1 - t_2}$, which reflects that observations further apart in time are less correlated. Finally, the unstructured working correlation matrix consists of a set of $corr\left(\hat{\theta}(t_1), \hat{\theta}(t_2)\right)$ that differs for each pair.

Agresti (2007) pointed out that if correlations are small, all working correlation structures yield similar estimates of parameters in GEE models and similar standard errors. In the Monte Carlo study, Klein and Andersen (2005) showed that there are no significant differences in estimations of GEE models for pseudo-observations with different working covariance matrices and recommended the use of the independent working covariance matrix.

The choice of the number of time points has little influence on the model fit. In the Monte Carlo simulations, Klein and Andersen (2005) showed that it is enough to choose five to ten time points, equally spaced on the event scale, to

evaluate pseudo-observations for the fitting model for the entire curve. Parameter estimates are quite insensitive to the number of time points. However, Andersen and Perme (2010) suggested that, nevertheless, all time points should be used if possible.

One of the problems with the implementation of GEE models is that GEE is a non–likelihood-based method. Therefore, information criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) cannot be directly applied, which creates problems with the choice of best model. The GEE models for pseudo-observations with the log-log link function are analogues to the Cox PH and Fine-Gray models; therefore, in the empirical study, a variable selection, and consequently a choice of models, was performed for the last models. The Akaike selection criterion (Akaike 1974) was used to choose the best subset of covariates separately in the Cox PH and Fine-Gray models (Kuk and Varadhan, 2013). Subsequently, these sets of covariates were used in the equivalent GEE models.

4. Empirical study

We considered a cohort of 5,000 retail credits granted during 12 consecutive months by a Polish financial institution. All credits were granted for a fixed term of 24 months. The cohort was followed for 15 months from the moment the first credit was granted. Each credit could terminate in one of two ways: being completely paid back earlier than scheduled (early repayment) or by defaulting. A defaulted credit was considered one that had a delay in instalment payments of at least 90 days. We observed both types of termination in the cohort, as well as censoring. Censored observations were credits for which neither default nor early repayment were observed during the follow-up. That is, for those credits, all instalments were paid on time or with a delay shorter than 90 days. Figure 8 shows the distribution of events and censoring over the months of the credits' life, observed at the end of the follow-up.

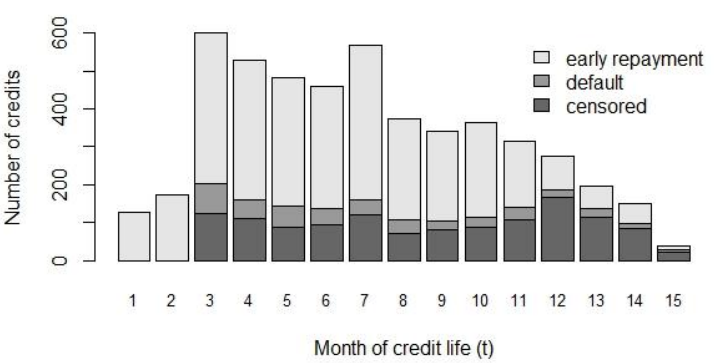


Figure 8. Distribution of the causes of termination during the follow-up of credits

Due to its definition, default cannot be observed for the first three months after credit granting. Early repayments and censoring were not lagged. Through the specificity of the analysed problem in the data set, we observed single events for particular time points with no censoring (t=1,2) and competing events with censoring (t=3,…,15). The objective was to evaluate one model of the probability of default for all time points in the presence of competing risk and heavy censoring.

To evaluate the probability of default, we laid down pseudo-observations and built GEE models for those pseudo-observations. Variables describing a creditor and a credit at the time of credit granting were used as covariates in these models. These variables included information such as age of the applicant, property, educational level, purpose of the credit, amount and instalment payments. To comply with the requirements of the financial institution sharing the data, the names of the variables were anonymised and are denoted in this paper by the letter X and one or more numbers.

All the variables were categorised and included in models as dummy variables. Two approaches were applied that resulted in different methods of assessing the pseudo-observations. The first assumed that the only type of analysed event was default; all other reasons for leaving the cohort of credits were considered to be censoring. In this approach, pseudo-observations were evaluated for the survival function (formula 5).

The second approach considered two causes of events: default and early repayment. Regular payments were handled as censored observations. Pseudo-observations were calculated (formula 10) for all event times due to the small number of analysed time points. To choose the set of covariates for GEE models, variable selection for the Cox PH model for single events, and the Fine-Gray model for competing events, was conducted at the first step with AIC using a stepwise algorithm (Venables and Ripley, 2002). Four different working covariance matrices were then applied.

Parameter estimations for all of the types of matrices were very close. Differences were only observed for estimates of parameters of dummy variables for time points, but these differences had no influence on the models' fit. The independence matrix was a slightly better fit for the model, and the results for this matrix are presented in the latter part of the paper. Table 1 shows the results of estimations of the GEE model for the CIF and estimates of the Fine-Gray model with the same set of covariates. Estimates of the parameters in both models are very similar. The GEE model, apart from covariates, also includes dummy variables for time points for which pseudo-values were calculated. Time point 1 (t1) is not included in the model because it is a reference group.

The CIF changes not only at the time of the considered event, but also at the time of the competing event. This is why time point 2 is included in the model, despite no default having occurred – this is one of the differences between competing- and single-event approaches. Values of the estimates for the subsequent time points increase, which is associated with higher wages for the units leaving a cohort later (compare Figure 6). Standard errors of estimators of covariates in the GEE model are slightly higher than for the Fine-Gray model; this observation is consistent with the findings of Andersen and Perme (2010).

Table 1. Estimates of the GEE model for the CIF and for the Fine-Gray model, both for the risk of default.

	GEE model for CIF								Fine-Gray model			
Time point	β	SE(β)	p-value	Cov.	β	SE(β)	p-value	Cov	β	SE(β)	p-value
Int.	-20.11	0.24	0.000	.	.	.	.	.	.	.	.
t2	-4.67	0.40	0.000	X1_2	-0.34	0.11	0.0027	X1_2	-0.30	0.10	0.0036
t3	17.27	0.40	0.000	X2_2	-0.19	0.13	0.1434	X2_2	-0.17	0.11	0.1300
t4	17.79	0.39	0.000	X2_3	-0.19	0.17	0.2651	X2_3	-0.31	0.15	0.0330
t5	18.18	0.39	0.000	X3_2	-0.30	0.12	0.0098	X3_2	-0.40	0.10	0.0001
t6	18.40	0.39	0.000	X4_2	0.28	0.12	0.0168	X4_2	0.21	0.10	0.0330
t7	18.57	0.39	0.000	X5_2	0.41	0.13	0.0013	X5_2	0.44	0.11	0.0000
t8	18.71	0.39	0.000	X6_2	-0.46	0.15	0.0016	X6_2	-0.61	0.13	0.0000
t9	18.78	0.39	0.000	X6_3	-1.51	0.16	0.0000	X6_3	-1.59	0.15	0.0000
t10	18.87	0.39	0.000	X7_1	-0.49	0.12	0.0000	X7_1	-0.52	0.10	0.0000
t11	18.98	0.39	0.000	X7_2	1.33	0.63	0.0355	X7_2	1.21	0.48	0.0120
t12	19.05	0.39	0.000	X8_1	-0.15	0.28	0.5859	X8_1	-0.31	0.22	0.1700
t13	19.19	0.39	0.000	X8_2	-0.55	0.26	0.0308	X8_2	-0.55	0.22	0.0110
t14	19.33	0.39	0.000	X9_1	0.22	0.15	0.1385	X9_1	0.27	0.13	0.0400
t15	19.58	0.39	0.000	X9_2	0.13	0.13	0.3032	X9_2	0.23	0.11	0.0350

Cov – covariate, Int. – intercept, t – dummy variable for a time point.

The purpose of a credit-risk assessment is not to find the size of the effect of a particular predictor on the risk of default, but to create a model which has the highest discriminatory ability and which allows prediction of the probability of default over the credit's life. To compare the performance of the above models, the following discrimination measures were used: area under the ROC curve (AUC), Kolmogorov-Smirnov test (KS), and Hand measure (H) (Hand, 2009). Additionally, significance tests of the differences between AUCs of both models were calculated (DeLong et al. 1988). Table 2 presents each model's performance at each of the event times. Both models have good and comparable discriminatory power through the whole credit-life. However, the best discrimination was achieved for the first months of credit-life; the slight advantage for the Fine-Gray model according to AUC is significant only for the last six months (see last column of Table 2).

Table 2. Measures of the performance of models for the CIF.

Month	H 95% CI		K-S 95% CI		AUC 95% CI		
	GEE	F-G	GEE	F-G	GEE	F-G	p-value
3	0.215 0.142-0.318	0.221 0.155-0.322	0.422 0.320-0.526	0.406 0.340-0.530	0.751 0.688-0.804	0.754 0.699-0.806	0.457
4	0.236 0.183-0.315	0.236 0.184-0.320	0.458 0.381-0.532	0.456 0.390-0.540	0.775 0.729-0.812	0.778 0.739-0.815	0.357
5	0.219 0.176-0.289	0.216 0.178-0.29	0.419 0.358-0.494	0.424 0.360-0.500	0.764 0.726-0.797	0.765 0.731-0.800	0.750
6	0.206 0.169-0.272	0.205 0.172-0.271	0.399 0.345-0.470	0.407 0.350-0.470	0.754 0.718-0.788	0.756 0.728-0.792	0.316
7	0.205 0.173-0.264	0.207 0.175-0.267	0.400 0.352-0.463	0.407 0.360-0.470	0.756 0.724-0.785	0.759 0.733-0.790	0.128
8	0.200 0.167-0.254	0.201 0.173-0.258	0.389 0.345-0.446	0.394 0.360-0.460	0.753 0.722-0.779	0.757 0.733-0.784	0.073
9	0.189 0.161-0.241	0.191 0.167-0.246	0.375 0.340-0.433	0.384 0.350-0.440	0.747 0.718-0.773	0.75 0.728-0.778	0.069
10	0.182 0.157-0.234	0.183 0.162-0.238	0.366 0.329-0.423	0.372 0.340-0.440	0.741 0.714-0.767	0.745 0.723-0.773	0.046
11	0.176 0.151-0.225	0.177 0.158-0.228	0.355 0.320-0.410	0.359 0.330-0.420	0.735 0.708-0.760	0.739 0.718-0.765	0.019
12	0.171 0.147-0.22	0.173 0.154-0.224	0.350 0.315-0.404	0.357 0.33-0.41	0.731 0.705-0.756	0.736 0.715-0.762	0.014
13	0.174 0.152-0.222	0.176 0.157-0.227	0.355 0.320-0.408	0.362 0.330-0.420	0.733 0.708-0.757	0.737 0.717-0.764	0.023
14	0.174 0.150-0.220	0.175 0.155-0.226	0.35 0.319-0.403	0.359 0.330-0.410	0.73 0.706-0.754	0.734 0.715-0.761	0.013
15	0.174 0.150-0.219	0.174 0.155-0.224	0.351 0.321-0.404	0.36 0.330-0.410	0.73 0.707-0.754	0.733 0.715-0.759	0.037

GEE- generalized estimating equations, F-G – Fine-Gray model, 95% CI – 95% confidence intervals as percentiles form 1000 bootstrapped samples

The application of the competing-risks methodology to credit-risk assessment is quite a recent idea (c.f. Watkins et al., 2014); it is more common to use single-event models (see Dirick et al., 2017). In the single-event approach, only time to default is considered, whereas credits that do not default until data-gathering are censored observations. However, in a credit-risk context, as a loan reaches maturity, default can no longer occur. Moreover, a very large proportion of the population will not go into default; hence, the basic principle in the survival analysis of one event type, that $S(t) \to 0$, does not hold. Therefore, in our study,

we should expect worse performance of single-event models than competing-events models for default. To verify this hypothesis, pseudo-observations for the survival functions were calculated with formula 5. Variable selection for the single-event model was performed using the AIC selection criterion for the Cox PH model. Estimates of the parameters of the Cox PH model and the GEE model for the survival function were calculated (see Table 3).

Table 3. Estimates of the GEE model for the survival function and estimates of the Cox PH model, both for the risk of default.

	GEE model for the survival function							Cox model			
Time point	β	SE(β)	p-value	Cov	β	SE(β)	p-value	Cov	β	SE(β)	p-value
Int	-2.42	0.29	0.000	X1_2	-0.48	0.15	0.001	X1_2	-0.34	0.10	0.001
t4	0.57	0.11	0.000	X2_2	-0.20	0.16	0.188	X2_2	-0.21	0.11	0.058
t5	1.03	0.14	0.000	X2_3	-0.11	0.25	0.653	X2_3	-0.35	0.15	0.020
t6	1.32	0.15	0.000	X3_2	-0.24	0.16	0.138	X3_2	-0.38	0.10	0.000
t7	1.55	0.15	0.000	X4_4	-0.21	0.29	0.460	X4_4	-0.37	0.15	0.010
t8	1.78	0.16	0.000	X5_2	0.28	0.16	0.074	X5_2	0.27	0.11	0.015
t9	1.89	0.16	0.000	X6_2	-0.70	0.23	0.003	X6_2	-0.59	0.13	0.000
t10	2.08	0.17	0.000	X6_3	-1.85	0.28	0.000	X6_3	-1.65	0.15	0.000
t11	2.32	0.17	0.000	X7_1	-0.44	0.16	0.006	X7_1	-0.48	0.11	0.000
t12	2.48	0.18	0.000	X7_2	1.68	0.79	0.033	X7_2	1.83	0.52	0.000
t13	2.82	0.19	0.000	X8_1	-0.04	0.41	0.931	X8_1	-0.35	0.23	0.123
t14	3.12	0.21	0.000	X8_2	-0.64	0.33	0.053	X8_2	-0.53	0.22	0.015
t15	3.68	0.29	0.000	X9_1	0.16	0.21	0.454	X9_1	0.26	0.13	0.055
.	.	.	.	X9_2	0.19	0.17	0.273	X9_2	0.33	0.11	0.004

Cov – covariate, Int. – intercept, t – dummy variable for a time point.

Dummy variables in a single-event approach were evaluated only for time points from 4 to 15. Time point 3 was omitted as the reference group, while time points 1 and 2 were not event times. The Akaike selection criterion applied to the Fine-Gray and Cox PH models gave almost the same set of covariates for both. The only difference was that variable X4_4 was applied to the single-event models instead of X4_2, which was used in the competing-events models. As a result, estimations of the parameters for all covariates in the models can be directly compared. As in the case of competing events, the GEE model for both the survival function and for the Cox PH model gave close estimations of parameters. The fit of the models also does not differ (see Table 4).

For both approaches, an interesting regularity was observed. For most of the covariates, p-values are greater for the GEE models for pseudo-observations than for Cox PH and Fine-Gray models; for some covariates, this resulted in a lack of significance, i.e. X2_3, X9_1, and X9_2 (compare Tables 1 and 3).

Table 4. Measures of the performance of models for the survival function.

Month	H 95% CI		K-S 95% CI		AUC 95% CI		
	GEE	F-G	GEE	F-G	GEE	F-G	p-value
3	0.199 0.130-0.302	0.203 0.148-0.317	0.433 0.314-0.526	0.418 0.329-0.528	0.751 0.680-0.800	0.749 0.694-0.805	0.832
4	0.229 0.163-0.296	0.226 0.175-0.309	0.454 0.365-0.525	0.449 0.384-0.534	0.773 0.720-0.803	0.773 0.734-0.811	0.923
5	0.207 0.157-0.267	0.207 0.171-0.277	0.42 0.346-0.484	0.414 0.359-0.492	0.762 0.718-0.788	0.761 0.728-0.797	0.892
6	0.199 0.148-0.252	0.195 0.163-0.264	0.406 0.332-0.459	0.401 0.351-0.468	0.751 0.708-0.777	0.752 0.722-0.787	0.778
7	0.195 0.152-0.244	0.197 0.166-0.259	0.403 0.339-0.456	0.403 0.360-0.469	0.753 0.714-0.777	0.755 0.729-0.785	0.531
8	0.193 0.152-0.241	0.193 0.166-0.251	0.391 0.333-0.445	0.390 0.352-0.458	0.751 0.714-0.773	0.753 0.729-0.781	0.489
9	0.184 0.147-0.23	0.184 0.161-0.24	0.380 0.327-0.433	0.379 0.347-0.447	0.746 0.711-0.769	0.748 0.725-0.775	0.452
10	0.179 0.142-0.222	0.180 0.159-0.235	0.370 0.318-0.421	0.369 0.337-0.434	0.74 0.706-0.763	0.744 0.722-0.771	0.226
11	0.173 0.138-0.218	0.175 0.155-0.225	0.356 0.306-0.405	0.357 0.325-0.420	0.733 0.698-0.755	0.738 0.717-0.764	0.101
12	0.169 0.133-0.211	0.171 0.151-0.223	0.35 0.301-0.400	0.353 0.320-0.412	0.729 0.695-0.751	0.734 0.714-0.761	0.056
13	0.172 0.139-0.215	0.174 0.154-0.225	0.354 0.305-0.403	0.358 0.325-0.415	0.731 0.700-0.753	0.736 0.715-0.762	0.070
14	0.17 0.140-0.214	0.172 0.152-0.221	0.35 0.303-0.400	0.354 0.323-0.411	0.727 0.697-0.751	0.732 0.714-0.759	0.064
15	0.171 0.141-0.215	0.169 0.149-0.218	0.353 0.307-0.401	0.354 0.321-0.408	0.728 0.700-0.751	0.731 0.713-0.757	0.210

GEE- generalized estimating equations, F-G – Fine-Gray model, 95% CI – 95% confidence intervals as percentiles form 1000 bootstrapped samples.

For the single-event approach, we also applied the method based on a reduction of the CIF to the case of one type of event. This led to the use of the $\hat{S}_{CIF}(t) = 1 - \hat{F}_{CIF}(t))$ estimator of the survival function, instead of the $\hat{S}_{KM}(t)$ estimator, in the calculation of the pseudo-observations. We observed that, for the pseudo-observations, the relation $\hat{S}_{KM}(t) = 1 - \hat{F}_{CIF}(t)$ does not hold. The differences between estimates were very low, but irregular. Figure 8 shows box-plots for the differences $\hat{S}_{KM}(t) - (1 - \hat{F}_{CIF}(t))$ for all the units at all time points. However, one should note that all the differences are very close to zero.

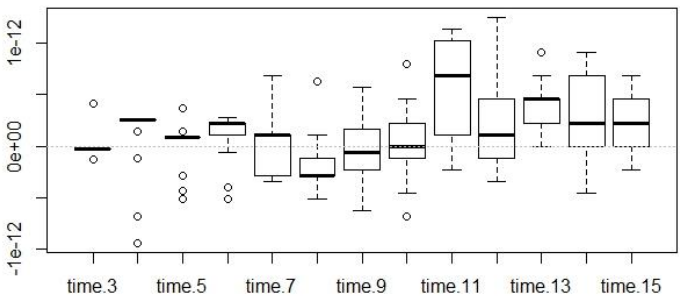


Figure 9. Distribution of differences between the KM estimator and the complement to CIF estimator for the survival function, both estimated by pseudo-observations

The GEE models for survival functions built for pseudo-observations based on both estimators gave exactly the same results. Thus, in spite of the above differences, the methods are fully interchangeable.

5. Conclusions

Pseudo-observations are a method that can be considered competitive with other survival analysis techniques. As shown in section 2, the values of pseudo-observations depend both on the type and time of event. Regression models for pseudo-observations correctly evaluate the whole survival curve, and the use of the log(-log) link function causes the GEE models for both single and competing approaches to simply mimic the results of the Cox PH and Fine-Gray models, respectively.

This observation is consistent with the results of earlier studies by other authors and argues against the use of a more cumbersome pseudo-values approach instead of more classic methods. However, because the independence matrix happened to be the best choice for the GEE model in all of the studies, it is suggested that pseudo-observations could be used as dependent variables in other methods for complete, independent data, such as classification trees.

In application to credit-risk assessment, competing-risks models had more discriminatory power than single-event models, which supports the use of competing-risks models in preference to models for single events. Further studies should focus on the variable-selection method that could be applied to the GEE models.

Acknowledgments

The authors gratefully acknowledge helpful feedback from an anonymous reviewer.

REFERENCES

AGRESTI, A., (2007). Logistic Regression, in An Introduction to Categorical Data Analysis, Second Edition, John Wiley & Sons, Inc., Hoboken, NJ, USA.

AKAIKE, H., (1974). A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19 (6), pp. 716–723.

ANDERSEN, P. K., KLEIN, J. P., ROSTHØJ, S., (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models, Biometrika, 90 (1), pp. 15–27.

ANDERSEN, P. K., PERME, M., (2010). Pseudo-observations in survival analysis, Statistical Methods in Medical Research 19 (1), pp. 71–99.

BINDER, N., GERDS, T. A., ANDERSEN, P. K., (2014). Pseudo-observations for competing risks with covariate dependent censoring, Lifetime data analysis, 20(2), pp. 303–315.

COX, D., (1972). Regression Models and Life-Tables, Journal of the Royal Statistical Society, Series B (Methodological), 34 (2), pp. 187–220

DELONG, E., DELONG, D., CLARKE-PEARSON, D., (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44, pp. 837–845.

DIRICK, L., CLAESKENS, G., BAESENS, B., (2017). Time to default in credit scoring using survival analysis: a benchmark study, J Oper Res Soc 6, pp. 652–655.

FINE, J., GRAY, R., (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk, Journal of the American Statistical Association, 94 (446), pp. 496–509.

HALLER, B., SCHMIDT, G., ULM, K., (2013). Applying competing risks regression models: an overview, Lifetime Data Anal 19, pp. 33–58.

HAND, D. J., (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve, Mach Learn 77, pp. 103–123.

HØJSGAARD, S., HALEKOH, U., YAN, J., (2005). The R Package geepack for Generalized Estimating Equations. Journal of Statistical Software, 15:2, pp. 1–11.

KLEIN, J., MOESCHBERGER, M., (2003). Survival Analysis: Techniques for Censored and Truncated Data, Statistics for Biology and Health, 2nd ed., Springer, New York.

KLEIN, J. P., ANDERSEN, P. K., (2005). Regression Modelling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function, Biometrics, 61 (1), pp. 223–229.

KUK, D., VARADHAN R., (2013). Model selection in competing risks regression, Statistics in Medicine 32, pp. 3077–3088.

LIANG, K., ZEGER, S., (1986). Longitudinal Data Analysis Using Generalized Linear Models. Biometrika, 73 (1), pp. 13–22.

MILLS, M., (2011). Introducing Survival and Event History Analysis, Sage, Los Angeles.

PINTILIE, M., (2006). Competing Risks: A Practical Perspective, Wiley.

VENABLES, W. N., RIPLEY, B. D., (2002). Modern Applied Statistics with S. Fourth edition, Springer.

WATKINS, J. G. T., VASNEV, A. L., GERLACH, R., (2014). Multiple Event Incidence and Duration Analysis for Credit Data Incorporating Non-Stochastic Loan Maturity, J. Appl. Econ., 29, pp. 627–648.

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 189

On behalf of the organizers of the first, envisioned as cyclic, international conference on

Methodology of Statistical Research

– the Statistics Poland (GUS) and the Polish Statistical Association – we are pleased to forward an invitation to this event. which will be held on 3–5 July 2019 in Warsaw, Poland

In the he preliminary program of the event a series of thematic sessions is being projected, embracing the following issues:

- Mathematical Statistics
- Survey Sampling and Small Area Estimation
- Population Statistics
- Social Statistics
- Economic Statistics
- Regional Statistics
- Data Analysis and Classification
- Statistical Data
- National Statistics in the International Context
- History of Polish Statistics, Statistical Research in Historical Perspective
- Statistics Communication and Statistical Education

Information on the conference venue and registration procedure is on the conference portal: http://met2019.stat.gov.pl/en/

Abstract can be sent by May 10 to:

https://rejestracja.stat.gov.pl/Konferencja2019/en/Account/Register

STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 191–195

ABOUT THE AUTHORS

Arcos Antonio PhD is a full Professor at the Department of Statistics and Operational Research, University of Granada. He specializes in the design and analysis of complex surveys. His research has focused on using auxiliary information in surveys in order to improve the accuracy of survey estimates.

Arnab Raghunath is a Professor of Statistics, University of Botswana and an Honorary Professor of the University of KwaZulu-Natal, South Africa. His area of interest is Survey Sampling Theory. He is the author of the book "Survey Sampling Theory and Applications" and over 100 articles in international journals.

Ben-Hur Dano is a senior statistician at the Israeli Central Bureau of Statistics. He has been specializing for 17 years in various statistical methods for estimating regional sizes and other parameters for the Israeli population censuses. He holds a BA degree in statistics and economics and an MA degree in statistics from the Hebrew University of Jerusalem.

Blum Olivia is a senior director of Demography & Census Department in the Israeli Central Bureau of Statistics (ICBS). Her main work and interest in official statistics include traditional, integrated and administrative censuses, population estimates of stocks and flows, health statistics and infrastructure registers. She has been also engaged in developing official statistics, among which is quality frameworks, and in promoting cooperation and partnerships with international organizations and other national statistics offices. Olivia studied Economics and Sociology & Anthropology in Tel-Aviv University and in SUNY at Stony-Brook.

Das Kishore Kumar is a Professor of the Department of Statistics, Gauhati University, Guwahati, Assam, India. Professor Das started his professional career from Karimganj College as a lecturer, then joined the Institute of Advanced Study in Science and Technology, Guwahati. Thereafter he joined the Department of Mathematical Sciences, Tezpur University, Tezpur before joining Gauhati University. He has served as the Head, Department of Statistics, Gauhati University and also worked as an Honorary Director of the Population Research Centre (PRC), Gauhati University. During his directorship, 9 research projects were completed under his supervision. Moreover, he has completed six individual major research projects mostly on socio-economic development, health, communicating diseases, etc. Further, he visited the Sampling and Official Statistics Unit of Indian Statistical Institute, Kolkata, as an Honorary Visiting Scientist. He is actively involved in teaching, research and consultancy. He has over 75 research papers published in reputed journals and co-authored eight books. He was conferred the INSA Visiting Fellow, and IBS International Travel Grants.

Dhar Soma is a guest lecturer in the Department of Statistics at B. Borooah College, Guwahati, Assam, India. She completed her PhD in Statistics in the year

2018 from Gauhati University, Guwahati, Assam, India. Her research area of interest is in the field of actuarial statistics, queuing theory, distribution theory, operations research and statistical modelling.

Intarapak Sukanya graduated from Naresuan University (in Mathematics, 2002) and Chiang Mai University (in Applied Statistics, 2005); PhD in Mathematics from Mahidol University (2016). Now, she is a lecturer in statistics at the Department of Mathematics, Faculty of Science, Srinakharinwirot University. She is a researcher in the field of statistical inference, multivariate statistics and data mining. She is a reviewer of reputed national journals of sciences (Srinakharinwirot Science Journal and Journal of Science and Technology).

Jurkiewicz Tomasz is an Associate Professor at the Statistics Department, Faculty of Management, University of Gdansk. His research interests include small area statistics and their applications in economy, simulation methods and their practical in such areas as applications in insurance, risk modelling using the R language. He has published over 50 papers in Polish and English mainly on applied statistics in small area estimation.

Kalyani Kruthiventi holds an MSc in Statistics, 2013, and is pursuing a PhD in Statistical quality control and reliability under the guidance of Prof K. Rosaiah, Department of Statistics, Acharya Nagarjuna University, Guntur, India. She has published 7 articles in national and international journals.

Kanaparthi Rosaiah has completed a MSc in Statistics in 1982 (Gold Medallist) and holds a PhD in Statistics in 1990 from Acharya Nagarjuna University, Guntur, India. Joined as a lecturer at the Department of Statistics, Acharya Nagarjuna University in 1985 and promoted as a Professor in 2002. Published 71 research articles in reputed peer-reviewed journals like Journal of Applied Statistics, Communications in Statistics Simulation and Computation, Journal of Testing and Evaluation, International Journal of Quality & Reliability Management, Economic Quality Control and Journal of Statistical Computation and Simulation. Reviewer for various National and International Journals. Also published two text books. Research interests areas are statistical inference, statistical quality control and reliability.

Leśkow Jacek is currently the Director of NASK – the Polish National Research Institute of Scientific and Academic Computer Networks located in Warsaw, Poland. His main area of competence is statistical signal processing, artificial intelligence and Big Data. Moreover, he has done extensive research and teaching in risk management, operations management, financial risk management and cross cultural management. His professional experience includes working full-time for the University of California, USA; Polish-American Business School in Nowy Sącz, Poland and Cracow University of Technology, Cracow, Poland. He was a principal investigator of NATO grants for securing the telecommunication signals and also held a position of principal investigator of the Polish Research Agency grant on statistical signal processing. In his career, Jacek Leśkow was a Visiting Professor in: USA, Mexico, France, Brazil, Ukraine, Kyrgistan, Sweden.

Mahanta Lipi B. is an Associate Professor working in the Central Computational Sciences Division of Institute of Advanced Study in Science and Technology, Guwahati, Assam, India. She obtained her PhD degree in Statistics in 2004 from Gauhati University, Guwahati, Assam. After collecting experience as a teacher of Computer Science for more than 20 years she started research work in different areas of health care like medical image processing and pattern recognition, epidemiology and development of operations research techniques with applications of health care. She has published over 50 papers in national and international journals in these various fields.

Pfeffermann Danny is a Professor of statistics at the University of Southampton, UK, and Professor Emeritus at the Hebrew University of Jerusalem, Israel. As of 2013, Danny is the Government Statistician and the Director General of the Central Bureau of Statistics in Israel. He is a past President of the Israel Statistical Society and a past President of the International Association of Survey Statisticians (IASS). For the last 20 years, he serves also as a consultant for the US Bureau of Labor Statistics. Danny is a fellow of the American Statistical Association and an elected member of the International Statistical Institute (ISI). He received a BA degree in Mathematics and Statistics and MA and PhD degrees in Statistics from the Hebrew University of Jerusalem. He is a recipient of multiple prestigious awards, including Waksberg award in 2011, the West Medal by the Royal Statistical Society in 2017, the Julius Shiskin Memorial Award for Economic Statistics in 2018, and the SAE 2018 Award for his distinguished contribution to the SAE methodology and to the advancement of Official Statistics in the Central Bureau of Statistics in Israel.

Priyanka Kumari is an Assistant Professor in Department of Mathematics, Shivaji College (University of Delhi), New Delhi, India. Her research interest includes sampling theory, statistical inference, sensitive estimation theory and statistical modelling. She has published 32 research papers in peer-reviewed journals and is the recipient of two major research projects funded by Government of India. She has also published one book and is a reviewer of many journals. She is the life member of many academic societies and associations.

Ramakrishnaiah Y. S. is a retired Professor of the Department of Statistics in Osmania University, Hyderabad, Telangana, India. He is a Visiting Professor for Research in Statistics, actively involved in collaborative research projects at the University of Albert, Edmonton, Canada. He is a well-known researcher in the field of distribution theory, non-parametric inference, statistical modelling and resampling theory. He is an active member in various Statistical Associations in India and the USA. His publications include four text books/monographs and over fifty research papers in reputed journals.

Rao Gadde Srinivasa received his MSc in Statistics (1988), MPhil. in Statistics (1994) and PhD in Statistics (2002) from the Acharya Nagarjuna University, Guntur, India. He is presently working as a Professor of Statistics at the Department of Statistics, The University of Dodoma, Tanzania. He boasts over 100 publications in different peer-reviewed journals in national and international well-reputed journals including, for example, Journal of Applied Statistics, International Journal of Advanced Manufacturer Technology, Communications

in Statistics Theory and Methods, Communications in Statistics Simulation and Computation, Journal of Testing and Evaluation, Arabian Journal for Science and Engineering, International Journal of Quality & Reliability Management, Economic Quality Control and Journal of Statistical Computation and Simulation. He is a reviewer of various reputed international journals. His research interests include statistical inference, statistical process control, applied Statistics, acceptance sampling plans and reliability estimation.

Satish Konda is working as an Associate Professor in the Department of Statistics of Aurora's Degree and PG College recognised by Osmania University, Telangana, Hyderabad, India. He is a researcher in the field of non-parametric inference, statistical modelling and data analysis in particular. He is a member of Indian Science Congress and has published research papers in reputed journals of statistics.

Shabbir Javid is a Professor at the Department of Statistics in Quaid-i-Azam University Islamabad, Pakistan. His research interests are mathematical statistics, rank set sampling, data imputation, randomized response technique and systematic sampling in particular. Professor Javid has published over 170 articles in national/international journals and conferences. Professor is an active member of scientific professional bodies. He is also the Chief Editor of Journal of Statistical Theory and Practices (JSTP).

Shangodoyin Dahud Kehinde is a Professor of Statistics at the University of Botswana, Botswana. His areas of specialty are time series, econometrics and statistical computing. Professor Shangodoyin is currently the President of African Statistical Association.

Sivakumar Devireddy Charana Udaya has completed a MSc in Statistics (QR & OR) in 2007 with first class with distinction. He is pursuing a PhD in Statistics under the guidance of Prof K. Rosaiah, Department of Statistics, Acharya Nagarjuna University, Guntur, India. He has published 11 research articles.

Skupień Maria has graduated from Cracow University of Technology in mathematics. In 2016 she defended her master thesis entitled "Functional Data Analysis and Its Application" under the supervision of Prof. Jacek Leśkow. She is a PhD student and a teaching assistant at the Mathematical Institute, Pedagogical University of Cracow. Her research interests are functional data analysis in general and their applications. She is also an author of popular science articles: "Nontrivial beauty of Jordan curves", PK 2014 and "Discrete Jordan Curve Theorem", UJ 2015, showing her interest not only in statistics but also in topology, discrete mathematics, graph theory and algebra.

Sohail Muhammad Umair is a PhD student at the Department of Statistics in Quaid-i-Azam University, Islamabad, Pakistan. His research interests are mathematical statistics, missing values, data imputation and data analysis in particular. He published many articles in the field of survey sampling in last few years.

Sohil Fariha is a PhD student at the Department of Education in Government College University, Faisalabad, Pakistan. His research interests are missing values and machine learning.

Supapakorn Thidaporn is an Assistant Professor at the Department of Statistics, Faculty of Science, Kasetsart University in Bangkok, Thailand. She has completed her doctorate in Mathematics and Statistics from Missouri University of Science and Technology (formerly, University of Missouri), Rolla, USA. Her research interests are generalized linear and mixed models, statistical theory and inference and data analysis. She is the member of Thai Statistical Association. She also is the member of the Editorial Board of Journal of Thai Statistical Association, American Journal of Theoretical and Applied Statistics, Journal of Computer Simulation in Application and Journal of Insight Statistics.

Trisandhya Pidugu is pursuing her PhD from University of Delhi, India and is also working as project fellow under SERB, New Delhi, India sponsored Major Research Project under the supervision of Dr Kumari Priyanka. Her research interest includes sampling theory and statistical inference.

Trivedi Manish is an academic statistician. He obtained his PhD in Statistics from Central University of Sagar (MP) in the area of Sampling Design in 2002 and a Master Degree in Statistics from Devi Ahilya Vishwavidyalaya, Indore(MP) in 1997. An ardent advocate of working in a proactive and challenging environment in academic and research field in pure and applied statistics. He has been working as an Associate Professor (18.11.2013 to present) and has worked as Reader (18.11.2010 to 17.11.2013) School of Sciences, IGNOU, New Delhi, INDIA. He also worked as Assistant Professor in Senior Scale (April 01, 2008 to November 16, 2010) and as an Assistant Professor (July 11, 2003 to March 31, 2008) in Birla Institute of Technology Mesra, Ranchi, INDIA. He has been Associated with Design and Development of Graduate, Post Graduate and Research Degree/Diploma programmes in Statistics and Distance Teaching/Learning Materials. Beside this he has been associated with the writing of study learning materials for almost all the courses of PGDAST programme and written 20 units as sole as well as joint authorship. He is a well-known researcher in the field of sampling theory, statistical modelling and applied area of statistics. He is a member of Executive Committee and Associations of many international journals and a reviewer of reputed international and national journals of Statistics and Biostatistics. He has over 35 research papers in reputable journals of statistics, mathematics and computer applications.

Wycinka Ewa is an Associate Professor at the Statistics Department, Faculty of Management, University of Gdansk. Her research interests include survival analysis and competing risks and their applications in economy and social sciences. She published over 40 papers in Polish and English mainly on applied statistics in insurance and credit scoring. She is a reviewer in a few of journals in the area of statistics and operations research.