Volume 24, Number 5, December 2023 ISSN 1234-7655

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

Savitsky T. D., Williams M. R., Gershunskaya J., Beresovsky V., Methods for combining probability and nonprobability samples under unknown overlaps

Abdullahi U. K., Ugwuowo F. I., Lawson N., Power ratio cum median-based ratio estimator of finite population mean with known population median

Das P., Singh G. N., Bandyopadhyay A., Ratio estimation of two population means in two-phase stratified random sampling under a scrambled response situation

Oullada O., Ben Ali M., Adri A., Rifai S., Model for measuring the impact of good pharmacovigilance practices of COVID-19 patients on hcp reactivity: Morocco case study

Khoshkhoo Amiri Z., MirMostafaee S. M. T. K., Analysis for the xgamma distribution based on record values and inter-record times with application to prediction of rainfall and COVID-19 records

Szymkowiak M., Roychowdhury A., Misra S. K., Giri R. L., Bhattacharjee S., A study of a survival data using kernel estimates of hazard rate and aging intensity functions

Baral M. M., Chittipaka V., Pal S. K., Mukherjee S., Shyam H. S., Investigating the factors of blockchain technology influencing food retail supply chain management: a study using TOE framework

Klochko R., Piskunova O., Marketing segmentation of banks' corporate clients based on data mining technique

Kończak G., Stąpor K., Changepoint detection with the use of the RESPERM method-a Monte Carlo study

Korczyński A., Bayesian predictive probability design – theory and practical example in a prospective study

Wójciak W., Another solution for some optimum allocation problem

ISSN 1234-7655

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

CONTENTS

Submission information for authors III

From the Editor VII

Invited paper

Savitsky T. D., Williams M. R., Gershunskaya J., Beresovsky V., Methods for combining probability and nonprobability samples under unknown overlaps 1

Research articles

Abdullahi U. K., Ugwuowo F. I., Lawson N., Power ratio cum median-based ratio estimator of finite population mean with known population median..... 35

Das P., Singh G. N., Bandyopadhyay A., Ratio estimation of two population means in two-phase stratified random sampling under a scrambled response situation..... 45

Oullada O., Ben Ali M., Adri A., Rifai S., Model for measuring the impact of good pharmacovigilance practices of COVID-19 patients on hcp reactivity: Morocco case study..... 63

Khoshkhoo Amiri Z., MirMostafaee S. M. T. K., Analysis for the xgamma distribution based on record values and inter-record times with application to prediction of rainfall and COVID-19 records 89

Szymkowiak M., Roychowdhury A., Misra S. K., Giri R. L., Bhattacharjee S., A study of a survival data using kernel estimates of hazard rate and aging intensity functions..... 109

Baral M. M., Chittipaka V., Pal S. K., Mukherjee S., Shyam H. S., Investigating the factors of blockchain technology influencing food retail supply chain management: a study using TOE framework 129

Klochko R., Piskunova O., Marketing segmentation of banks' corporate clients based on data mining technique..... 147

Other articles

XXXX Multivariate Statistical Analysis 2022, Lodz, Poland. Conference Papers

Kończak G., Stąpor K., Changepoint detection with the use of the RESPERM method-a Monte Carlo study 167

XXXI Scientific Conference of the Classification and Data Analysis Section (SKAD 2022)

Korczyński A., Bayesian predictive probability design – theory and practical example in a prospective study..... 185

Research Communicates and Letters

Wójciak W., Another solution for some optimum allocation problem .....

About the Authors 221

Acknowledgments to reviewers (2023)..... 227

Index of Authors (2023)..... 235

STATISTICS IN TRANSITION new series, *December 2023*
Vol. 24, No. 5, pp. III

Submission information for Authors

Statistics in Transition new series (SiTns) is an international journal published jointly by the Polish Statistical Association (PTS) and 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:
sit@stat.gov.pl,
GUS/Statistics Poland,
Al. Niepodległości 208, R. 296, 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: https://sit.stat.gov.pl/ForAuthors.

STATISTICS IN TRANSITION new series, *December 2023*
Vol. 24, No. 5, pp. V–VI

Policy Statement

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

CEEOL – Central and Eastern European Online Library

CEJSH (The Central European Journal of Social Sciences and Humanities)

CNKI Scholar (China National Knowledge Infrastructure)

CNPIEC – cnpLINKer

CORE

Current Index to Statistics

Dimensions

DOAJ (Directory of Open Access Journals)

EconPapers

EconStore

Electronic Journals Library

Elsevier – Scopus

ERIH PLUS (European Reference Index for the Humanities and Social Sciences)

Genamics JournalSeek

Google Scholar

Index Copernicus

J-Gate

JournalGuide

JournalTOCs

Keepers Registry

MIAR

Microsoft Academic

OpenAIRE

ProQuest – Summon

Publons

QOAM (Quality Open Access Market)

ReadCube

RePec

SCImago Journal & Country Rank

TDNet

Technische Informationsbibliothek (TIB) – German National Library of Science and Technology

Ulrichsweb & Ulrich's Periodicals Directory

WanFang Data

WorldCat (OCLC)

Zenodo

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. VII–XI

From the Editor

Presented to our readers – as the last in this year – December issue of Statistics in Transition new series, contains a set of twelve articles written by thirty-four authors from nine countries (in the order of appearance): USA, Nigeria, Thailand, India, Nepal, Morocco, Iran, Poland, and Ukraine.

The geographical diversity of this issue is accompanied by its thematic richness. From the volume's opening article – an invited paper – demonstrating practical solutions to the increasingly common problem of choosing between probability and nonprobability sampling, through selected issues of parameter estimation and distribution analysis in the subsequent papers, to specific approaches applied in different studies, including those based on conference presentations.

Taking advantage of the opportunity that this is the last issue of this year's issue, I would like to personally, and on behalf of the entire editorial team and our advisory bodies – the Editorial Board and Associate Editors – express deep gratitude to all our stakeholders – authors and readers, but above all, peer-reviewers. The knowledge and commitment of the reviewers helped our authors improve their work while contributing significantly to our efforts to ensure that articles published in SiTns meet the highest quality standards; hence the list of their names is included in the Acknowledgments at the end of this issue.

Invited paper

The opening text in this volume, as the invited paper, by **Terrance D. Savitsky, Matthew R. Williams, Julie Gershunskaya,** and **Vladislav Beresovsky,** entitled ***Methods for combining probability and nonprobability samples under unknown overlaps*** presents innovative perspective to this complex issue, including a novel approach that derives the propensity score for the observed sample as a function of inclusion probabilities for the reference and convenience samples as the main result. The presented approach allows specification of a likelihood directly for the observed sample as opposed to the approximate or pseudo likelihood. The authors construct a Bayesian hierarchical formulation that simultaneously estimates sample propensity scores and the convenience sample inclusion probabilities. A Monte Carlo simulation study to compare the likelihood based results with the pseudo likelihood based approaches considered in the literature was used.

Research articles

Umar K. Abdullahi, Fidelis I. Ugwuowo, and **Nuanpan Lawson** in the article ***Power ratio cum median-based ratio estimator of finite population mean with known population median*** propose the power ratio cum median-based ratio estimator of the finite population mean, which is a function of two ratio estimators in the form of an average. The estimator assumes the population to be homogeneous and skewed, while the properties (i.e. the bias and the Mean Squared Error – MSE) were derived alongside its asymptotically optimum MSE. The efficiency of the developed estimator jointly with its efficiency conditions by comparing it to selected estimators described in the literature were demonstrated. Empirically, a real-life data set from the literature and a simulation study from two skewed distributions (Gamma and Weibull) were used to examine the efficiency gain. Results from both the real-life dataset and the simulation study show efficiency gain for the proposed estimator that incorporates median of study variable (while for the other estimators the result show efficiency loss).

The paper by **Pitambar Das, Garib Nath Singh,** and **Arnab Bandyopadhyay,** ***Ratio estimation of two population means in two-phase stratified random sampling under a scrambled response situation*****,** describes the development of an effective two-phase stratified random sampling estimation procedure in a scrambled response situation. Two different exponential, regression-type estimators were formed separately for different structures of two-phase stratified sampling schemes. The authors have studied the properties of the suggested strategy. The performance of the proposed strategy has been demonstrated through numerical evidence based on a data set of a natural population and a population generated through simulation studies. Taking into consideration the encouraging findings, suitable recommendations for survey statisticians are prepared for the application of the proposed strategy in real-life conditions.

The next paper, by **Oumaima Oullada, Mohamed Ben Ali, Ahmed Adri,** and **Said Rifai,** entitled ***Model for measuring the impact of good pharmacovigilance practices of COVID-19 patients on hcp reactivity: Morocco case study*** presents a conceptual model used to evaluate how the improvement of good pharmacovigilance practices of COVID-19 patients influences the reactivity of the healthcare personnel (HCP) in the Draa Tafilalet region in Morocco. The empirical study is based on a survey submitted to a sample of a total of 180 HCP and on the application of latent variable structural modelling through the partial least squares (PLS) method focusing on the reliability and validity of the proposed model. The study shows that the improvement of good pharmacovigilance practices impacts positively the reactivity of HCP in terms of adverse drug reactions (ADRs) reporting. The reliability of the measurement was > 0.7, which allowed testing the internal and external validity of the conceptual model; several hypotheses were validated against two invalid derivative hypotheses.

Zahra Khoshkhoo Amiri's and **S.M.T.K. MirMostafaee's** article, ***Analysis for the xgamma distribution based on record values and inter-record times with application to prediction of rainfall and COVID-19 records,*** discusses the problem of classical and Bayesian estimation of the unknown parameter of the xgamma distribution based on record values and inter-record times. The problem of Bayesian prediction of future record values based on record values and interrecord times was also discussed. A new lifetime distribution, called "xgamma distribution", which can be used as an alternative to other lifetime distributions, like the exponential one, was introduced. A simulation study has been performed to compare the performance of the proposed estimators and the approximate Bayes predictors, complemented by two real data sets related to rainfall and COVID-19 records.

Magdalena Szymkowiak, Anasuya Roychowdhury, Satya Kr. Misra, Rajib Lochan Giri, and **Subarna Bhattacharjee** present the paper ***A study of a survival data using kernel estimates of hazard rate and aging intensity functions.*** The authors primarily focus on Aging Intensity (AI) and Hazard Rate (HR) functions estimated using four different kernels. They apply them to a case study of patients with primary malignant tumors of sternum with the right-censored data. It turned out that kernel estimates of HR and AI functions for patients with high grade tumor (HGT) are higher than for patients with low grade tumor (LGT), as expected. The authors believe that their study opens up a new direction for applying AI and HR functions in analyzing health and engineering related problems.

In the paper ***Investigating the factors of blockchain technology influencing food retail supply chain management: a study using TOE framework,*** **Manish Mohan Baral, Venkataiah Chittipaka, Surya Kant Pal, Subhodeep Mukherjee,** and **Hari Shankar Shyam** discuss the factors affecting blockchain adoption in the food retail supply chain and create awareness among retail managers for its adoption in their operations. A structured literature review was conducted to identify the TOE factors used in the research. TOE factors were used in many previous studies on technology adoption, like RFID, IoT, cloud computing, intelligent agent technology, and many more. With these factors, a questionnaire was developed for the survey. The questionnaires were sent to retail stores across India through online mode. The results were analysed using EFA and SEM techniques. The findings shown that TOE factors contribute to blockchain adoption by keeping the intention to adopt the technology as a mediating variable.

Rostyslav Klochko's and **Olena Piskunova's** article ***Marketing segmentation of banks' corporate clients based on data mining technique*** aims to segment a bank's corporate client base and develop a pricing strategy for each of the groups that have been singled out in the process. The study sample consisted of 4,500 corporate clients

of a Ukrainian bank who were active users of euro accounts. The k-means data mining algorithm was used to develop marketing segments. The optimal number of clusters was determined by weighing the results of calculating 26 indices from the NbClust package and the bank's business requirements. The study found that clusters 1st and 2nd were a concentration of unprofitable customers for whom an introduction of a service fee was urgently needed. Marketing segments 3 and 4 were customers who did not record net losses but with whom it was deemed necessary to work to improve their profitability. The remaining two segments were 'healthy' users of euro accounts. With regard to these customers, it was recommended no additional service fees should be imposed.

Other articles

XXXX Multivariate Statistical Analysis 2022, Lodz, Poland. Conference Papers

Grzegorz Kończak and **Katarzyna Stąpor** in the paper ***Changepoint detection with the use of the RESPERM method – a Monte Carlo study*** use RESPERM (residuals permutation-based method) as a single changepoint detection method based on regression residuals permutation, which can be applied to many physiological situations where the regression slope can change suddenly at a given point. The article presents the results of a Monte Carlo study on the properties of the RESPERM method for single changepoint detection in a linear regression model. The proposed method was compared with a well-known segmented method for detection breakpoint in linear models. In the simulation study six variants of noise were considered from normal, uniform and two variants of beta distributions together with two cases of equal and unequal variances. Three levels of variance in the distribution of random errors were taken into account: minor, major and dominant errors. The simulations were performed for different locations of changepoint in time series. The Monte Carlo study showed that when the input data are very noisy, the RESPERM method outperforms the segmented approach in terms of variance, and in the case of bias, the results of the two methods are comparable.

XXXI Scientific Conference of the Classification and Data Analysis Section (SKAD 2022)

Adam Korczyński's paper entitled ***Bayesian predictive probability design – theory and practical example in a prospective study*** provides theoretical background and the practical perspective, pointing out the statistical properties but also technical aspects in conducting a trial with predictive design. Also, sensitivity of the design to the choice of prior distribution was considered. The Bayesian predictive design allows to draw conclusions on the prognosis given the actual results. Their theoretical properties are appealing as a tool for detecting the treatment sig. The practical application has shown

the usefulness of the approach from the perspective of the timing of the decision. This accords with argument for adaptive design allowing for reducing the overall sample size, cost of the study, drug development time length. The final decision would still require larger sample size, although the Bayesian design seems to have a supportive role.

Research Communicates and Letters

In the Research Communicates and Letters section an article by **Wojciech Wójciak** analyses ***Another solution for some optimum allocation problem***. The study derives optimality conditions for the optimum sample allocation problem in stratified sampling, formulated as the determination of the fixed strata sample sizes that minimize the total cost of the survey, under the assumed level of variance of the stratified π estimator of the population total (or mean) and one-sided upper bounds imposed on sample sizes in strata. In this context, the author presumes that the variance function is of some generic form that, in particular, covers the case of the simple random sampling without replacement design in strata. The optimality conditions mentioned above are derived from the Karush-Kuhn-Tucker conditions, and the study formulates the LRNA in such a way that it also provides the solution to the classical optimum allocation problem of minimization of the estimator's variance under a fixed total cost (under one-sided lower bounds imposed on sample sizes in strata). In such a case, the LRNA can be considered as a counterparty to the popular recursive Neyman allocation used to solve the classical problem of sample allocation with added one-sided upper bounds.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 1–34, https://doi.org/10.59170/stattrans-2023-061
Received – 24.08.2023; accepted – 11.11.2023

Methods for combining probability and nonprobability samples under unknown overlaps

Terrance D. Savitsky[1], **Matthew R. Williams**[2], **Julie Gershunskaya**[3], **Vladislav Beresovsky**[4]

Abstract

Nonprobability (convenience) samples are increasingly sought to reduce the estimation variance for one or more population variables of interest that are estimated using a randomized survey (reference) sample by increasing the effective sample size. Estimation of a population quantity derived from a convenience sample will typically result in bias since the distribution of variables of interest in the convenience sample is different from the population distribution. A recent set of approaches estimates inclusion probabilities for convenience sample units by specifying reference sample-weighted pseudo likelihoods. This paper introduces a novel approach that derives the propensity score for the observed sample as a function of inclusion probabilities for the reference and convenience samples as our main result. Our approach allows specification of a likelihood directly for the observed sample as opposed to the approximate or pseudo likelihood. We construct a Bayesian hierarchical formulation that simultaneously estimates sample propensity scores and the convenience sample inclusion probabilities. We use a Monte Carlo simulation study to compare our likelihood based results with the pseudo likelihood based approaches considered in the literature.

Key words: Survey sampling, Nonprobability sampling, Data combining, Inclusion probabilities, Exact sample likelihood, Bayesian hierarchical modeling.

1. Introduction

1.1 Motivation

With the proliferation of powerful computers and internet technologies, private data aggregators and research organizations gained the ability to relatively easily collect and store information from samples of respondents. Usually such opportunistic or "convenience" samples are not selected using a probability based sampling design. The non-random participation of units in such a convenience sample limits its ability to be used to construct an estimator (e.g., average income) of a target population quantity because the convenience sample, in general, is not expected to be representative of that population.

[1]Office of Survey Methods Research, U.S. Bureau of Labor Statistics, USA.
E-mail: Savitsky.Terrance@bls.gov. ORCID: https://orcid.org/0000-0003-1843-3106.
[2]RTI International, USA. E-mail: mrwilliams@rti.org. ORCID: https://orcid.org/0000-0001-8894-1240.
[3]OEUS Statistical Methods Division, U.S. Bureau of Labor Statistics, USA.
E-mail: Gershunskaya.Julie@bls.gov. ORCID: https://orcid.org/0000-0002-0096-186X.
[4]Office of Survey Methods Research, U.S. Bureau of Labor Statistics, USA.
E-mail: Beresovsky.Vladislav@bls.gov. ORCID: https://orcid.org/0009-0002-8375-5195.

By contrast, probability based samples or random surveys of units represent the gold standard for cost-effectively sampling a population in a manner that allows provable guarantees about the population representativeness of target estimators (e.g., total employment, vaccination rate) composed from the observed sample where units are randomly invited to participate. We term such a random-inclusions sample as a “reference" sample.

Yet, probability based samples are often relatively small, especially at finer domain levels; hence, probability based sample estimators often have large variances. In other cases, reference samples may not include particular variables of interest, while such variables may be collected with the convenience sample.

Statistical agencies and other survey administrators are increasingly seeking ways to leverage convenience samples to construct estimators of target population quantities with measurable statistical properties. This paper focuses on a class of approaches that suppose the nonrandom convenience sample was drawn from an unknown or latent random sampling design process such that we may treat the convenience sample as a “pseudo" random sample. The sampling design for the random reference sample is set by the governing statistical agency and is encoded in known sample inclusion probabilities assigned to the population of units. These inclusion probabilities are used to form inverse probability sampling weights that are published with other variables collected in the reference sample. So, the task for combining the convenience sample with the reference sample to strengthen estimation (and lower the variance of estimators) is in *estimation* of the unknown convenience sample inclusion probabilities to form “pseudo" weights. We assume the existence of covariates, measured on both the reference and convenience samples, that encode the sampling design. Then, estimated convenience sample pseudo weights may be used with any response variable to form a weighted estimator of the target population quantity.

1.2 Literature Review

Early attempts to address estimation of the convenience sample inclusion probabilities using combined convenience and reference probability samples include Elliott (2009), Valliant and Dever (2011), DiSogra et al. (2011). See recent reviews in Valliant (2020) and Beaumont (2020), and Wu (2022).

Our goal in this paper is to estimate the convenience sample inclusion probabilities based on observed indicator z_i that is defined on the combined convenience and probability samples set as $z_i = 1$ for a unit in the convenience sample, and 0 for a unit in the reference sample.

Elliott (2009) and Elliott and Valliant (2017) consider Bernoulli variable z_i and uses relationship between $\pi_{zi} = P\{z_i = 1\}$, on the one hand, and the convenience and reference sample inclusion probabilities, π_{ci} and π_{ri} (respectively), on the other hand. One is then able to specify a logistic regression for estimation of π_{ci}. While their result implies a practical approach, their derivation requires an assumption that the convenience and reference samples must be disjoint. That is, no unit may be included in *both* the convenience and reference samples. They also use a two-step model estimation process that is suboptimal and often produces unbounded estimates for π_{ci}. A more efficient, one-step likelihood based estimation procedure, was proposed by Beresovsky (2019).

More recently, Chen et al. (2020) approached the problem by considering the convenience sample inclusion indicator R_i, where $R_i = 1$ for unit i in the convenience sample, and 0 for unit i in the finite population less those units which are members of the convenience sample. R_i is a Bernoulli variate; however, convenience sample inclusion probabilities $\pi_{ci} = P\{R_i = 1\}$ cannot be estimated directly from the Bernoulli likelihood of R_i because the finite population is not generally available and indicator R_i is not observed for the whole population; in particular, one does not know which units from the finite population are selected into the convenience sample. To overcome this difficulty, they partition the log-likelihood of R_i into two terms: the sum over convenience sample units and the sum over the finite population. The latter term is approximated by a ''pseudo" likelihood, using inverse probability based weights, defined by observed reference sample inclusion probabilities.

There are two shortcomings in Chen et al. (2020)'s approach. First, the pseudo likelihood approximation is suboptimal because it is a noisy approximation on the observed sample that will produce a higher estimation variance. Second, convenience sample membership indicators R_i are generally not observable. The partitioning proposed by Chen et al. (2020) implies the existence of a different, *observable*, indicator that is defined as follows. Stack together the convenience sample and finite population, so that the sample units appear in the stacked set twice: as part of the population and as the added set; let indicator $Z_i = 1$ for unit i in the convenience sample, and 0 for any unit i in the finite population (regardless of whether it is also a part of the convenience sample). Note, however, that Chen et al. (2020)'s likelihood does not treat observed Z_i as a Bernoulli variate, thus potentially leading to suboptimal results.

Wang et al. (2021) propose an improvement of Chen et al. (2020) by formulating the Bernoulli likelihood for Z_i and providing a formula specifying a relationship between probabilities $P\{Z_i = 1\}$ and convenience sample inclusion probabilities π_{ci}. Would the finite population be observed, this approach would lead to efficient estimation of π_{ci} based on the likelihood of observed Z_i. However, since the finite population is not observed, they still have to rely on the pseudo likelihood approach in their estimation. Wang et al. (2021) apply a two-step estimation procedure, which can be improved by solving the pseudo-likelihood based estimating equations using the one-step approach of Beresovsky (2019).

1.3 Contribution of this Paper

We use first principles to derive a relationship between probability of being in the convenience sample set π_{zi}, on the one hand, and the convenience and reference sample inclusion probabilities, π_{ci} and π_{ri} (respectively), on the other hand. The result of Elliott (2009) can be viewed as a special case of our formula. Importantly, our approach dispenses with the requirement of disjointness between the two sample arms. We show that our method for estimating π_{ci} is valid under *any* degree of overlapping units among the two sampling arms. Unlike Chen et al. (2020) and Wang et al. (2021), our result is defined directly on the observed pooled sample with no approximation required. So, the resulting estimator of π_{ci} from our method is more efficient than the approximate, pseudo likelihoods.

Differently from the two-step estimation process of Elliott (2009) or Wang et al. (2021),

we construct a Bayesian hierarchical modeling formulation discussed in the sequel that estimates both (π_{zi}, π_{ci}) in a single step. Our method accounts for all sources of uncertainty to produce more accurate uncertainty quantification.

Our approach is fully Bayesian for estimation of the unknown inclusion probabilities for the convenience sample units. Notions of informativeness do not apply because the likelihood is formulated directly on the observed set. The model-estimated inclusion probabilities are subsequently used to compute sampling weights and those weights and the response variable are together used to construct a survey-based population estimator (such as the population mean of y).

We introduce notation and list assumptions in Section 2. In Section 3, we detail the setup and provide the proof of the main formula underlying the proposed approach. Namely, we derive the relationship between the propensity score (defined as the probability of belonging to the convenience sample for a unit from the pooled sample), on the one hand, and the inclusion probabilities for the reference and convenience samples, on the other hand. We construct a Bayesian hierarchical modeling formulation in Section 4 that simultaneously estimates all unknown quantities, including unknown reference sample inclusion probabilities for convenience units, in a single step that accounts for all sources of uncertainty. A Monte Carlo simulation study to compare our approach with competitor methods is presented in Section 5. In Section 6, we apply the proposed method to the Current Employment Statistics data, where we estimate pseudo weights for the non-probability based sample for local government in California and compute domain estimates based on these weights. We conclude with a discussion in Section 7.

2. Preliminaries

We begin by introducing notation used in the exposition of our method developed in the following section. We follow by listing common assumptions used to develop the method.

Our set-up consists of a sample acquired under a random sampling design that we label as a “reference" sample to contrast with availability of a nonrandom “convenience" sample. We term the observed pooled sample as a “two-arm" sample with one arm denoting the reference (probability) sample and the other arm the convenience (nonprobability) sample.

Let S_c represent a non-probability (convenience) sample set drawn from sampling frame or population U_c, where $|U_c| = N_c$ and $|S_c| = n_c$ represent the number of units in sets U_c and S_c, respectively; let S_r denote a probability (reference) sample drawn from population U_r, with $|U_r| = N_r$ and $|S_r| = n_r$, the number of units, respectively, in U_r and S_r.

Let $\tilde{U} = U_r + U_c$ denote an imaginary combined set. Operator "+" here is meant to signify that sets U_r and U_c are ”stacked together" in such a way that overlapping units, that belong to both sets U_r and U_c, would be included into $\tilde{U}$ twice. Similarly, let $S = S_r + S_c$ denote a pooled (stacked) sample. Under such a setup, $|\tilde{U}| = N_r + N_c = N$ and $|S| = n_r + n_c = n$.

In an abuse of notation, we index a unit contained in any population or observed sample realization by i, which may indicate a unit in any of the sample or population sets where the context is clear.

Let $\pi_c(\mathbf{x}_i) = P(i \in S_c | i \in U_c, \mathbf{x}_i)$ denote the probability of inclusion into observed sample set S_c from U_c conditional on associated design variables, $\mathbf{x}_i$. We will use the term "conditional inclusion probability" for an inclusion probability whose specification or estimation is conditioned on a set of design variables, $X = \{\mathbf{x}_i\}$. These variables are used to construct the sampling design that governs the observed samples. The design variables typically don't include one or more response variables, y_i, of inferential interest because they are not observed for the full underlying population (such their estimation motivates the administration of the survey).

Let $\pi_r(\mathbf{x}_i) = P(i \in S_r | i \in U_r, \mathbf{x}_i)$ denote the conditional inclusion probability in S_r from U_r.

Let indicator variable z_i on set S take a value of 1 when $i \in S_c$, and 0 when $i \in S_r$; and let $\pi_z(\mathbf{x}_i)$ denote probabilities of $z_i = 1$, given $\mathbf{x}_i$: $\pi_z(\mathbf{x}_i) = P\{z_i = 1 | \mathbf{x}_i\} = P\{i \in S_c | i \in S, \mathbf{x}_i\}$. We label $\pi_z(\mathbf{x}_i)$ as the "propensity score" that measures the propensity or probability for a unit in the *observed* joint sample, S, to be included in S_c.

We will use π_{ci} as a shorthand notation for $\pi_c(\mathbf{x}_i)$ in the sequel when the context is clear and the same for π_{ri}.

(C1) (Latent Random Mechanism)
The observed convenience sample, S_c, is governed by an underlying, latent random mechanism with unknown sample inclusion probabilities, π_{ci}.

(C2) (Design Variables)
$p \times 1$ variables, $X \in \mathscr{X}$, fully determine the unit conditional inclusion probabilities into S_r and S_c for the random selection mechanisms. A consequence of the above set-up is that both U_c and U_r contain variables $\{X_r, X_c\} \in \mathscr{X}$ on the same measure space.

(C3) (Overlapping Populations)
Populations, (U_c, U_r), may overlap where units are jointly contained in each set such that overlapping units will each *appear exactly twice* in $\tilde{U}$. As a result, observed samples (S_c, S_r) may also contain overlapping units such that overlapping units each appear twice in S.

(C4) (Independence of Samples)
Conditional on X, $S_r \perp S_c \mid X$. Inclusions of units into each sample arm are independent, no matter the degree of overlap between U_r and U_c.

(C5) (Positive Inclusion Probabilities)
For all $i \in 1, \ldots, n$ and for all $\mathbf{x} \in \mathscr{X}$, conditional inclusion probabilities in each sampling arm are strictly positive / non-zero, such that $P(i \in S_r \mid \mathbf{x}_i, i \in \tilde{U}) > 0$, $P(i \in S_c \mid \mathbf{x}_i, i \in \tilde{U}) > 0$, which leads to $P(i \in S \mid \mathbf{x}_i, \tilde{U}) > 0$. These conditions result in $P(i \in S \mid i \in \tilde{U}) = \int P(i \in S \mid \mathbf{x}, i \in \tilde{U}) F(d\mathbf{x}) > 0$.

Assumption (C1) states that the non-random convenience sample may be understood as governed by a latent random process that we seek to uncover. The focus of this paper is the estimation of unknown inclusion probabilities into the convenience sample.

Convenience sample inclusion probabilities, π_{ci}, are generally not observed for units in the convenience sample; e.g., $\forall i \in S_c$. Reference sample inclusion probabilities are not generally observed for those units sampled solely into the convenience sample (and not included in the reference sample); e.g., $\forall j \in S_c \setminus S_r$.

Assumption (C3) allows for a general case of non-perfectly overlapping convenience and reference frames (from which the associated two samples are taken).

Our method requires Assumption (C4) on the independence of the reference and inclusion samples, but makes no assumptions about the degree of unit overlaps between the two samples.

It is typical to assume positive inclusion probabilities for all units as we do in Assumption (C5) for any rational sampling design in order to ensure that every unit in population U may be sampled, which in turn allows for unbiased inference about the population for the observed samples taken under this assumption.

3. Likelihood Based Estimation of Inclusion Probabilities Under Two-arm Samples

In this section we prove an identity that is central to our proposed approach for estimation of convenience sample inclusion probabilities. The proof is made from the first principals and under no requirement for disjointness among the sample arms. Namely, we derive the relationship between the propensity for the observed set of reference and convenience inclusion indicators and the associated inclusion probabilities in each sample.

Suppose, each frame is a subset of target population U^0, such that $U_c \subseteq U^0$ and $U_r \subseteq U^0$. Define probabilities $p_c(\mathbf{x}_i) = P\left\{i \in U_c | i \in U^0, \mathbf{x}_i\right\}$ and $p_r(\mathbf{x}_i) = P\left\{i \in U_r | i \in U^0, \mathbf{x}_i\right\}$. Quantities $p_r(\mathbf{x}_i)$ and $p_c(\mathbf{x}_i)$ are *known* coverage probabilities of population U^0 by frames U_c and U_r for a set of design variables $\mathbf{x}_i$. These probabilities depend on the same design variables, $\mathbf{x}_i$, though units will express differing values for the common design variables. For example, frame U_c could be the subset of individuals in U^0 with broadband internet access and frame U_r could be the subset of individuals in U^0 with mailable addresses.

While conditional inclusion probabilities $\pi_r(\mathbf{x}_i) = P\{i \in S_r | i \in U_r, \mathbf{x}_i\}$ for sample S_r are known, convenience sample conditional inclusion probabilities $\pi_c(\mathbf{x}_i) = P\{i \in S_c | i \in U_c, \mathbf{x}_i\}$ are unknown and can be inferred from combined sample $S = S_c + S_r$, where samples S_c and S_r are stacked together. As already mentioned in previous sections, samples S_c and S_r *may overlap*. The overlapping units appear in (stacked) set S *twice*: as units from S_c (with $z_i = 1$) and as units from S_r (with $z_i = 0$).

Proposition: Assume Conditions (C1)-(C5). Then, the following relationship between respective probabilities holds:

$$\pi_z(\mathbf{x}_i) = \frac{\pi_c(\mathbf{x}_i)\, p_c(\mathbf{x}_i)}{\pi_c(\mathbf{x}_i)\, p_c(\mathbf{x}_i) + \pi_r(\mathbf{x}_i)\, p_r(\mathbf{x}_i)}. \tag{1}$$

Proof: The combined set S emerges from the following scheme displayed in Figure 1 where we stack identical populations U^0 of units. The set of units in U^0 are *duplicated* from the

top-to-the-bottom stack. In the top layer we define U_r from which we draw sample, S_r and we do the same for the convenience population, U_c, and sample, S_c, in the bottom stack. We see that units in U_r and U_c may overlap in this scheme, which allows units in S_r and S_c to also overlap, though we don't know the identities of overlapping units because we have duplicated them in each stack, so our notation separately indexes the same unit in the reference and convenience frames and observed samples. This means that the sampling processes in each stack are independent from one another, but readily permit overlaps in (U_r, U_c) and (S_r, S_c). We next outline the scheme of Figure 1 in our proof.

To summarize, we consider two copies of target population U^0, where one copy of the population includes frame U_c, the other copy includes U_r. We *stack* the two copies of U^0 together and denote the result by U: $U = U^0 + U^0$.

For such a setup, by the Law of Total Probability (LTP), we have:

$$\begin{aligned} P\{i \in S_c | i \in U, \mathbf{x}_i\} &= P\left\{i \in S_c | i \in U_c, i \in U^0, \mathbf{x}_i\right\} P\left\{i \in U_c | i \in U^0, \mathbf{x}_i\right\} P\{i \in U^0 \mid i \in U\} \\ &= \frac{1}{2} \pi_c(\mathbf{x}_i)\, p_c(\mathbf{x}_i) \end{aligned} \tag{2}$$

We note that $i \in S_c$ implies that $i \in U_c$ since we draw the convenience sample from its associated frame, U_c. Similarly,

$$\begin{aligned} P\{i \in S_r | i \in U, \mathbf{x}_i\} &= P\left\{i \in S_r | i \in U_r, i \in U^0, \mathbf{x}_i\right\} P\left\{i \in U_r | i \in U^0, \mathbf{x}_i\right\} P\{i \in U^0 \mid i \in U\} \\ &= \frac{1}{2} \pi_r(\mathbf{x}_i)\, p_r(\mathbf{x}_i). \end{aligned} \tag{3}$$

Now, because we have stacked U^0 twice - once for the convenience sampling process and again for the reference sampling process - thus "shifted" sets S_c and S_r do not overlap (as illustrated in Figure 1), so we may sum them below to compute the total probability of being included into the pooled sample,

$$\begin{aligned} P\{i \in S | i \in U, \mathbf{x}_i\} &= P\{i \in S_c | i \in U, \mathbf{x}_i\} + P\{i \in S_r | i \in U, \mathbf{x}_i\} \\ &= \frac{1}{2} \pi_c(\mathbf{x}_i)\, p_c(\mathbf{x}_i) + \frac{1}{2} \pi_r(\mathbf{x}_i)\, p_r(\mathbf{x}_i). \end{aligned} \tag{4}$$

Finally, by the definition of conditional probability,

$$P\{i \in S_c | i \in S, i \in U, \mathbf{x}_i\} = \frac{P\{i \in S_c | i \in U, \mathbf{x}_i\}}{P\{i \in S | i \in U, \mathbf{x}_i\}}. \tag{5}$$

Equation 1 directly follows from Equations 2, 4, and 5.

We may now parameterize a likelihood for the observed indicator z_i using Equation 1:

$$z_i \mid \mathbf{x}_i, \beta \overset{\text{ind}}{\sim} \text{Bernoulli}(\pi_z(\mathbf{x}_i, \beta)). \tag{6}$$

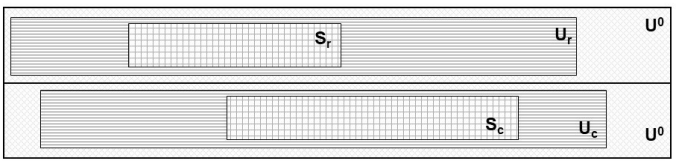

Figure 1 Gridded area represents observed convenience S_c and reference S_r samples stacked together to form combined sample S: $S = S_c + S_r$; under this scheme, if samples S_c and S_r overlap, the overlapping units are included in S twice. Convenience sample S_c is selected from population U_c, and reference sample S_r is selected from population U_r, where U_c and U_r are subsets of target population U^0: $U_c \subseteq U^0$ and $U_r \subseteq U^0$. In this setup, two identical copies of target population U^0 are stacked together, so that $U = U^0 + U^0$.

The likelihood of Equation 6 implicitly depends on parameter $\pi_c(\mathbf{x}_i)$ through Equation 1. We may specify a model for $\pi_c(\mathbf{x}_i) = f(\mathbf{x}_i, \beta)$ and fit parameters using either Frequentist or Bayesian approaches. We use a Bayesian approach in the sequel for its flexibility.

Remark 1: Our formulation for the propensity score does not rely on disjointness among the sampling arms. Our method explicitly allows for the *unknown* overlapping of units in S_r and S_c.

Remark 2: We can view the process as a two-phase selection. First, units are selected from target population U^0 to subpopulations U_c and U_r with probabilities $p_c(\mathbf{x}_i)$ and $p_r(\mathbf{x}_i)$, respectively. At the second phase, units are selected to respective samples with probabilities $\pi_c(\mathbf{x}_i)$ and $\pi_r(\mathbf{x}_i)$.

Remark 3: The equal frame scenario. If frames U_c and U_r coincide, we have $p_c(\mathbf{x}_i) = p_r(\mathbf{x}_i)$, and Equation 1 becomes

$$\pi_z(\mathbf{x}_i) = \frac{\pi_c(\mathbf{x}_i)}{\pi_c(\mathbf{x}_i) + \pi_r(\mathbf{x}_i)}. \tag{7}$$

A similar expression was derived by [**?**] under the assumption of non-overlapping convenience and reference samples. Our approach does not require this assumption.

Equation 7 holds even when the reference sample is the entire target population frame U. In this case, $\pi_r(\mathbf{x}_i) = 1$ for all units and Equation 7 reduces to

$$\pi_z(\mathbf{x}_i) = \frac{\pi_c(\mathbf{x}_i)}{\pi_c(\mathbf{x}_i) + 1}, \tag{8}$$

which is the same as that of 2021valliant before they approximate it on the observed sample. We label this as a “one-arm" case. One important simplification in the “one-arm" case is that π_r's are known (and equal to 1) for all units in combined set S.

We derive the same result as presented in Equation 1 under perfectly overlapping frames by extending a different result from the economics literature in Appendix A.

4. Hierarchical Estimation Model

We next specify a hierarchical probability model to estimate convenience sample inclusion probabilities for the units in the convenience sample.

We focus on the equal frame scenario where both the reference and convenience samples are assumed to be drawn from the same underlying frame to define our Bayesian hierarchical model and simulation setup. We do so for ease and clarity of explanation, with no loss of generality. In the common case where the frames do not perfectly overlap, we would use Equation 1 which inputs coverage probabilities $p_c(\mathbf{x}_i)$ and $p_r(\mathbf{x}_i)$ as *known* quantities.

We assume that our covariates $\mathbf{x}$ fully account for the sampling design. Thus, our goal is to formulate a model to estimate the inclusion probabilities of convenience sample units given covariates $\mathbf{x}$. We use them to formulate inverse probabilities based pseudo sampling weights to construct a survey expansion estimator using response variable of interest y.

4.1 Construction of unknown marginal inclusion probabilities, $(\pi_{\ell i})$

We parameterize our model using $\pi_{\ell i} = P\{i \in S_\ell \mid i \in U_\ell, \mathbf{x}_i\}$ to be the conditional inclusion probability for unit $i \in 1,\ldots,(n = n_r + n_c)$ in sampling arm $\ell \in (r,c)$; that is, ℓ indexes whether the conditional inclusion probability for unit i is specified for the reference $(\ell = r)$ or the convenience $(\ell = c)$ sampling arms. This modeling set-up only assumes that we observe $\pi_{\ell i}$ for $\ell = r$ and $i \in S_r$, the conditional sampling inclusion probabilities for the units observed in the reference sample.

Our model, however, will estimate $(\pi_{\ell i})$ for *all* units, $i \in (1,\ldots,n)$, for both $\ell = r$ and $\ell = c$ sampling arms. Of particular note, our model estimates π_{ri} for $i \in S_c$, the reference sample inclusion probabilities for the convenience units. So, estimation of the model does not require known π_{ri} for all units. The model will further simultaneously estimate π_{ci} for $i \in S_c$, the convenience sample inclusion probabilities for the convenience units (units in the convenience sampling arm), which is the primary goal of the model.

A Bayesian hierarchical model is able to be richly parameterized to estimate this matrix of only partially observed conditional inclusion probabilities through the borrowing of strength in the specifications of functional forms and prior distributions to follow.

4.2 Spline functional form for logit$(\pi_{\ell i})$

We input an $n \times K$ matrix of design variables, $X = (\mathbf{x}_1,\ldots,\mathbf{x}_K)$, where $\mathbf{x}_k$ denotes the $n \times 1$ vector for design variable k. We want our model specification to express a flexible functional form,

$$\text{logit}(\pi_{\ell i}) = f(x_{1i},\ldots,x_{Ki}), \tag{9}$$

where $f(\cdot)$ may be estimated as non-linear by the data. Complex sampling designs may utilize design variables with different emphases on different portions of the design space, which will induce such non-linearity. Two common examples are (i) scaling inclusion probabilities to exactly meet target sample sizes and (ii) thresholding size measures to create certainty units (with $\pi_{\ell i} = 1$). Both features induce non-linearity on the logit scale.

To accomplish the above non-linear formulation we utilize a B-spline basis due to its flexibility and computational tractability for illustration (of an implementation of our main result in Equation 1). We may also choose alternative non-linear formulations, such as a Gaussian process, to achieve similar results, but computation for the Gaussian process scales poorly in the number of data observations. The use of Bayesian adaptive regression trees is not easily purposed to our modeling set-up for estimating latent convenience sample inclusion probabilities. See Chipman et al. (2010).

A B-spline basis is specified for *each* predictor where $Q \times 1$, $g(x_{ki})$ is a B-spline basis vector with C denoting the number of bases set equal to the number of knots + number of spline degrees - 1. We use the vector of B-splines for each predictor k to formulate,

$$\text{logit}(\pi_{\ell i}) = \mu_{x,\ell i} = \mathbf{x}_i^\top \gamma_{x,\ell} + \sum_{k=1}^{K} g(x_{ki})^\top \beta_{\ell k}, \tag{10}$$

where $\mathbf{x}_i^\top \gamma_{x,\ell}$ is a linear component and $\beta_{\ell k}$ is a $Q \times 1$ vector of coefficients for the spline term for each predictor k (column of X) that parameterizes the possibility for a non-linear functional form for each of the K predictors. The spline term specifies distinct regression coefficients for each sampling arm, ℓ, and design variable, k, to allow estimation flexibility that makes few assumptions about the functional form for $\text{logit}(\pi_{\ell i})$. In this sense, even if we had only used the linear term, the use of distinct spline term regression coefficients for each predictor and sampling arm makes the model marginally non-linear across the data.

4.3 Random walk of order 1 (autoregressive) horseshoe prior on $\beta_{\ell k}$

We select a random walk of order 1 (based on first differences) formulation for the prior on each component of the $Q \times 1$, $\beta_{\ell k}$ of the spline term with,

$$\beta_{\ell kq} \mid \beta_{\ell kq-1}, \kappa_{\ell k}\tau_\ell \sim \mathcal{N}\left(\beta_{\ell kq-1}, \kappa_{\ell k}\tau_\ell\right),\ c = 2, \ldots, Q, \tag{11}$$

and $\beta_{\ell k1} \sim \mathcal{N}(0, \kappa_{\ell k}\tau_\ell)$ denotes a spline basis (used for each predictor $k \in 1, \ldots, K$). All to say, the random walk prior is constructed for the B-spline coefficients defined on each predictor, x_k. This random walk form for the prior enforces smoothness over the estimated regression coefficients such that the resulting estimated fit is less sensitive to the number of (spline) knots used and avoids overfitting. The overall mean intercept is identified by excluding an intercept from the linear term in Equation 10.

We also encourage sparsity in the number of estimated non-zero, $(\beta_{\ell k})_{k=1}^K$, as a group for predictor K, by using a set of K "local" scale (standard deviation) shrinkage parameters, $\kappa_{\ell k}$, where a value for $\kappa_{\ell k'}$ near 0 for some predictor k' will shrink all $Q \times 1$ coefficients, $\beta_{\ell k'}$, to 0. Similarly, global scale (standard deviation) shrinkage parameter, τ_ℓ, would shrink *all* $(\beta_{\ell k})_{k=1}^K$ to 0, which favors the linear model term in this limit. We place half Cauchy priors, $\kappa_{\ell k} \overset{\text{ind}}{\sim} C^+(0,1)$ and $\tau_\ell \sim C^+(0,1)$, respectively.

This use of local and global shrinkage parameters under a half Cauchy prior is known as the horseshoe prior. See Carvalho et al. (2009). If one marginalizes out the global and local scale shrinkage parameters under the half Cauchy priors, the marginal prior distribution

for $\beta_{\ell kq}$ will have a large spike at 0 (driving sparsity), but with very heavy tails allowing the coefficient values to "escape" the shrinkage where the data provide support. By tying together the priors for $(\beta_{\ell kq})_{q=1}^{Q}$ the spline coefficients for predictor k escaping shrinkage to 0 will be correlated and relatively smooth.

The vector of $K \times 1$ fixed effects parameters for sampling arm ℓ are each drawn as,

$$\gamma_{x,\ell k} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \tau_\gamma\right), \tag{12}$$

where $\tau_\gamma \sim$ student-t^{+}(d.f. $= 3, 0, 1$), where we use a relatively flat prior for τ_γ.

4.4 Joint likelihood for $(z_i)_{i \in S}$ and $(\pi_{ri})_{i \in S_r}$

We connect our parameters to the data with two likelihood terms. The first term constructs a Bernoulli likelihood for the observed sample,

$$z_i \mid \pi_{zi} \overset{\text{ind}}{\sim} \text{Bernoulli}(\pi_{zi}), \tag{13}$$

where we recall from Equation 7 that, $\pi_{zi} = \pi_{ci}/(\pi_{ci} + \pi_{ri})$ such that this likelihood provides information for estimation of π_{ci} for $i \in 1, \ldots, n$, as well as π_{ri} for $i \in S_c$.

We further borrow strength from the known reference sample conditional inclusion probabilities for the observed reference sample to estimate the unknown conditional inclusion probabilities by modeling the known reference sample inclusion probabilities for the observed reference sample units as a function of our parameters with,

$$\text{logit}(\pi_{ri}) \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{x,ri}, \phi), \tag{14}$$

only for units $i \in S_r$ such that observed π_{ri} is used to provide information about latent $\mu_{x,\ell i}$ for both sampling arms ($\ell \in (r, c)$ and all units ($i \in 1, \ldots, n$) based on their intercorrelations allowed by the prior distribution (and updated by the data). We recall from Equation 10 that each $\mu_{x,\ell i}$ is, in turn, connected with each $\pi_{\ell i}$.

The detailed Stan (Gelman et al., 2015) script that enumerates the likelihood and prior distributions for all parameters and hyper-parameters is included in Appendix B.

4.5 Bayesian hierarchical model implementations for pseudo likelihoods

We implement the pseudo likelihood formulations of Chen et al. (2020) and Wang et al. (2021) in the simulation study of Section 5 under our Bayesian hierarchical formulation. We accomplish these implementations by replacing the exact Bernoulli likelihood for the observed sample under our method of Section 3 with approximate likelihoods for the underlying population estimated on the sample. Both methods parameterize only π_{ci} for convenience units and use the inclusion probabilities for the reference sample as a plug-in. Let vector $\theta_c = \left(\gamma_{x,c}, (\beta_{ck})_{k=1}^{K}\right)$ denote the parameters in the non-linear logistic regression model for $\pi_{ci}(\mathbf{x}_i, \theta_c)$. Chen et al. (2020) specify the following pseudo log-likelihood,

$$\ell(\theta_c) = \sum_{i \in S_c} \log\left(\frac{\pi_{ci}(\mathbf{x}_i, \theta_c)}{1 - \pi_{ci}(\mathbf{x}_i, \theta_c)}\right) + \sum_{i \in S_r} d_{ri} \log\left(1 - \pi_{ci}\left(\mathbf{x}_i, \theta_c\right)\right), \tag{15}$$

where $d_r = 1/\pi_{ri}$. Equation 15 uses a survey approximation for the population in the second term by inverse probability weighting the reference sample contribution. This pseudo likelihood will tend to produce overly optimistic (narrow) credibility intervals because it uses π_{ri} as a plug-in (rather than co-modeling it). The first term will also induce a noisy estimator for unit with low values for π_{ci}, which will occur when there is a lot of separation in the covariate, $\mathbf{x}$, values between the convenience and reference samples.

As discussed in the introduction, Wang et al. (2021) develop a Bernoulli likelihood for the population augmented by the convenience sample. This approach specifies indicator $Z_i = 1$ if unit i is in the convenience sample, or 0 if it is the finite population and develops an associated propensity score, $\pi_{Zi} = \pi_{ci}/(\pi_{ci} + 1)$. This expression is a special case of our formula derived in Section 3 where one arm is the convenience sample and the other arm is the entire population. So the exact likelihood specified in Wang et al. (2021) is a special case of our method under a one-arm sample set-up. As with Chen et al. (2020), they approximate their log-likelihood on the observed sample with

$$\ell(\theta_c) = \sum_{i \in S_c} \log\left(\pi_{Zi}(\mathbf{x}_i, \theta_c)\right) + \sum_{i \in S_r} d_{ri} \log\left(1 - \pi_{Zi}(\mathbf{x}_i, \theta_c)\right). \tag{16}$$

This approximate likelihood will also tend to produce overly optimistic credibility intervals because it doesn't account for the uncertainty in the generation of samples (by plugging in the reference sample weights, d_{ri}, instead of modeling them).

Both comparator methods are implemented under our hierarchical Bayesian model such that they are benefited from our flexible, nonlinear formulation for the logit of the convenience sample inclusion probabilities and our autoregressive smoothing on spline coefficients. In this sense, these implementations are more robust than the estimating equation approaches used by the authors. In addition, in our implementation of Wang et al. (2021), we estimate convenience sample probabilities in a single step.

5. Simulation Study

We construct a finite population and two sets each of reference and convenience samples characterized by low and high overlaps in number of overlapping units between the two sampling arms. We perform this construction in each iteration of a Monte Carlo simulation study designed to compare the repeated sample (frequentist) properties of our two-arm exact likelihood approach with those of the pseudo likelihood approaches. We compare bias, root mean squared error and coverage of 90% credibility intervals.

5.1 Simulation Settings

To compare performance variability across multiple realized populations, we generate $M = 30$ distinct populations of size $N = 4000$. We chose a relatively small population size

and large sampling fractions to explore the full range of $\pi_c \in [0,1]$. A large sampling fraction and large inclusions probabilities is also reasonable in establishment surveys. We set the reference sample size at $n_r = 400$ using a proportion-to-size (PPS) sampling. We select convenience samples of size $n_c \approx 800$ using Poisson sampling. We recall our assumption that the convenience sample arises from a latent random sampling mechanism with unknown inclusion probabilities. We select two distinct independent convenience samples from each population, which we deem 'high' and 'low' overlap in comparison to the reference sample. High-overlap convenience samples have selection probabilities π_c with a similar relationship with population covariates X compared to the selection probabilities π_r for the reference sample. In constrast, low-overlap convenience sample probabilities π_c have the opposite relationship with covariates.

For each population, we let X have $K = 5$ columns, including an intercept, three independent binary variables (A,B,C) with $P(\mathbf{x}_i = 1) = 0.5$, and a continuous predictor drawn from a standard normal distribution $N(0,1)$. We do not explore the situation of correlated design variables in this simulation study. We generate the outcome y_i as a lognormal distribution with centrality parameter $\mu_i = \mathbf{x}_i\beta$ and scale parameter 2: $\log(y_i) \sim \mathcal{N}(\mu_i, 2)$. The generating parameters are $(\beta_{cont}, \beta_0, \beta_A, \beta_B, \beta_C) = (1.0, 0.5, 0.0, -0.5, -1.0)$. The inclusion probabilities for the reference sample are constructed by first setting size measure $s_{r_i} = \log(\exp(\mu_i) + 1)$. We then convert size to inclusion probabilities π_{r_i} via the *inclusionprobabilities()* function from the 'sampling' package in R. See Tillé and Matei (2021).. Most sizes of $s_{r_i} \propto \pi_{r_i}$, however the largest size values get mapped to $\pi_{r_i} = 1$, thus inducing a non-linear 'kink' in the mapping from $s_{r_i} \to \pi_{r_i}$. We note that our estimation model $\text{logit}(\pi_{r_i}) = \mu_i$ is misspecified leading to a non-linear relationship with the x_i. This motivates the use of splines to capture non-linear relationships and add robustness to the model estimation. It is common for the largest-sized units to be included in the sample with probability 1.

The inclusion probabilities for the convenience samples are inverse logit transformations of linear predictors with an offset adjustment to the intercept to approximately meet a target sample size: $\pi_{c_i} = \text{logit}^{-1}(\mathbf{x}_i\beta + \text{off})$. For the high-overlap sample:
$(\beta_{cont}, \beta_0, \beta_A, \beta_B, \beta_C, \text{off}) = (0.500, 0.175, -0.150, -0.475, -0.800, -0.900)$. For the low overlap sample: $(\beta_{cont}, \beta_0, \beta_A, \beta_B, \beta_C, \text{off}) = (-1.00, -0.50, 0.00, 0.50, 1.00, -2.23)$. It is generally more challenging to estimate convenience sample inclusion probabilities when there is a lower overlap of predictor values with the reference sample.

Each plot panel in Figure 2 compares the generated reference sample inclusion probabilities to the convenience sample inclusion probabilities for under a high-overlap size-based sampling design on the left and a low-overlap sampling design on the right. Each plot panel orders units by reference sample inclusion probabilities low-to-high along the x-axis. The degrees of similarity in the reference and convenience sample inclusion probabilities are achieved by manipulating the size and direction of the vector of coefficients β for the design variables.

Figure 3 compares the percent of the total combined sample (reference and convenience) units which overlap (e.g. is present in both samples) for realizations of 'high' and 'low' overlap convenience samples as well as a baseline expected overlap from two independent simple random samples. As indicated by their labels, 'high' overlap samples have a larger

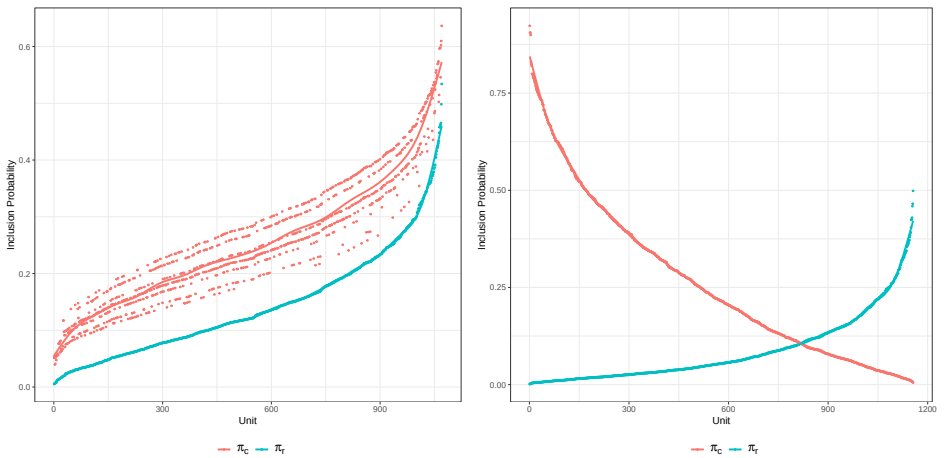

Figure 2: Comparison of inclusion probabilities for a single realization of reference and convenience samples for high overlap (left) and low overlap (right) designs. Units index the combined sample.

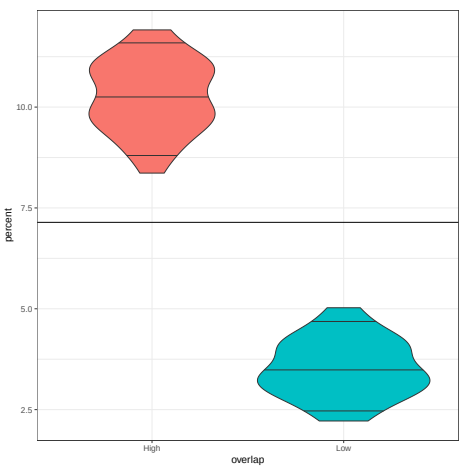

Figure 3: Percent of pooled sample present in both reference and convenience samples by type of convenience sample (High and Low). Distributions over 30 population and sample realizations. Expected percent for two independent simple random samples (solid horizontal line).

proportion of individuals in both the reference and convenience sample than each of a sample of the same sizes based on SRS and a 'low' overlap sample.

5.2 Results - Estimating convenience sample inclusion probabilities, π_{ci}

We begin our presentation of results by comparing the relative performances of exact (two-arm) and pseudo likelihood methods for the estimation of the convenience sample inclusion probabilities, π_{ci} $i \in S_c$ based on our known true values.

The plot panels of Figure 4 present the mean of bias (over the Monte Carlo iterations), the square root of the mean squared error and the (frequentist) coverage and average widths of 90% credibility intervals from left-to-right in the matrix of plot panels. These values are computed pointwise for increasing values of the true conditional inclusion probabilities from left-to-right in each plot panel. The top row of plot panels presents results for high-overlap (convenience and reference) sample datasets and the bottom row presents results for low-overlap sample datasets.

We compare 3 methods:

1. two-arm - constructs an exact likelihood from our main method of Equation 7 under a two-arm convenience and reference sample set-up.
2. CLW - The pseudo likelihood method of Chen et al. (2020).
3. WVL - The pseudo likelihood method of Wang et al. (2021).

For the two pseudo likehood methods, we implement each directly as pseudo posteriors and with a post-processing adjustment using a sandwich estimate of an asymptotic covariance matrix. See Williams and Savitsky (2021). The stability of estimation of the sandwich estimator for CLW was poor. In order to compensate, we first used a scalar down-weighting (or tempering) of all observations (both convenience and reference) such that the sum of the sum of the individual weights was equal to the total sample size. See Bhattacharya et al. (2019) for a detailed discussion on the stabilization of posterior estimation using such fractional weights.

We see that our two-arm method produces little mean bias for both small and large values of the true convenience sample inclusion probabilities and achieves nominal coverage of the 90% credibility intervals.

By contrast, both pseudo likelihood methods perform similarly to one another with high variability (RMSE) and severe undercoverage for medium-to-larger values of true convenience sample inclusion probabilities π_c. The collapse in coverage becomes worse for the low overlap dataset as the use of reference sample weights as a plug-in both under-estimates the uncertainty introduced by the reference sample design and induces noise over repeated samples. For high overlap, a post-processing adjustment for the pseudo likelihood methods improves coverage at the expense of increasing the width of the corresponding interval beyond that of the two-arm method. For low overlap, the post-processing adjustment can only adjust variance but not bias. In fact it may even amplify bias. Coverage is improved, but at the cost of very wide intervals.

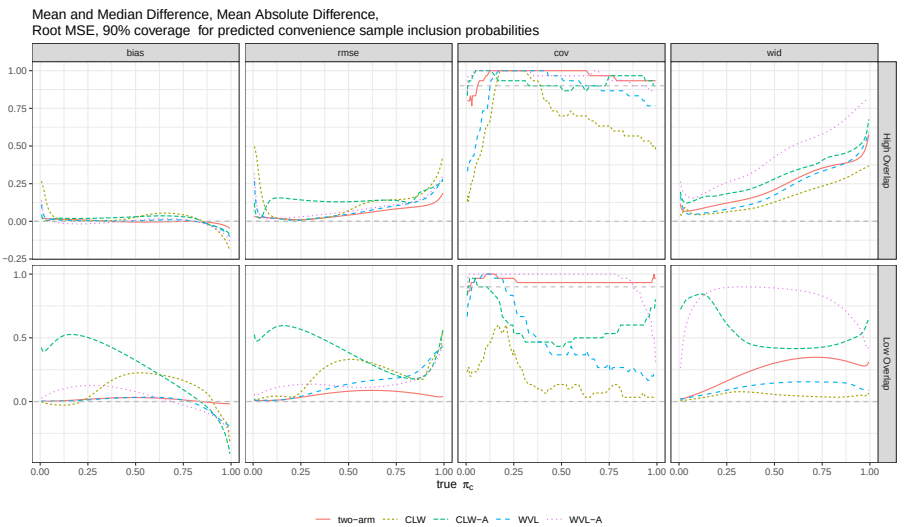

Figure 4: Performance of main approach for high overlap (top) and low overlap (bottom) samples across repeated simulations. Using informative reference sample for main approach (red), compare to pseudo-likehood based methods CLW (yellow) and WVL (blue). Adjusted versions of pseudo-likehood adjust based on an estimated sandwich covariance matrix: CLW-A (green) and WVL-A (purple). Left to Right: Mean Bias, Square Root Mean Squared Error, and Coverage and Interval Width for 90% intervals for predicting convenience sample inclusion probabilities π_c

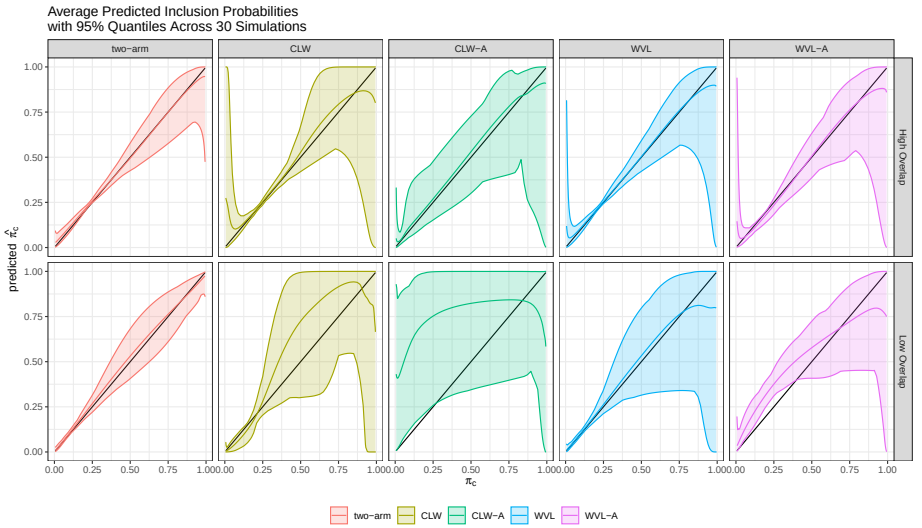

Figure 5: Average and pointwise 95% frequentist confidence intervals for the posterior mean estimator of π_c over the Monte Carlo iterations for high overlap (top) and low overlap (bottom) samples. Using informative reference sample for main approach (red), compare to pseudo-likehood based methods CLW (yellow) and WVL (blue). Adjusted versions of pseudo-likehood adjust based on an estimated sandwich covariance matrix: CLW-A (green) and WVL-A (purple).

Each plot panel in Figure 5 compares the average and pointwise 95% frequentist confidence intervals for the posterior mean estimator of π_c over the Monte Carlo iterations. The left-hand panel represents the results for the high-overlap datasets and the right-hand panel for the low-overlap datasets. We see that the two-arm exact likelihood method produces little-to-no bias.

By contrast, the pseudo likelihood methods produce an enormous amount of variability. While we would expect the performance of the pseudo likelihood methods to improve as the sample sizes increases since both methods produce consistent estimators, our chosen sample size is a very typical domain sample size for a survey such that the superior performance of our exact likelihood methods at these moderate sample sizes (for (n_r, n_c)) is an important result that demonstrates much faster convergence for our approach.

We use our method to combine convenience and reference sample inclusion probabilities (π_{ci}, π_{ri}) to construct a non-model-based survey direct estimator for the population mean, μ, of some response variable of interest, y, that is correlated with the survey design variables, $\mathbf{x}$ in Appendix C. We compare our resulting population mean estimator to that estimated from the two pseudo likelihood methods.

6. Application

We next present results from applying our proposed method to estimate pseudo weights for a quota sample of government employment collected in the Current Employment Statis-

tics (CES) survey administered by the U.S. Bureau of Labor Statistics (BLS). We subsequently apply the pseudo weights to estimate local government employment for the Metropolitan Statistical Areas (MSA) of California.

The CES uses probability-based sampling design for private industries. For government employers, however, the CES estimates are based on a non-probability sample. The employment coverage in government industries is generally high, so that the resulting unweighted estimates based on such a non-probability sample usually provide acceptable level of precision. For measuring employment of local governments, however, such an unweighted quota (convenience) sample based estimate may be biased.

We will use the quarterly census of employment and wages (QCEW), which is a census instrument administered by BLS that measures establishment employment, as our "reference sample" to estimate the pseudo weights for the CES government convenience sample. As a large census instrument, QCEW quality checking and reporting are lagged by many months, so the CES is used to provide the current month employment. The QCEW employment levels are maintained in an administrative source called the longitudinal database (LDB).

To estimate pseudo weights, we stack together the LDB and the CES sample and apply Equation 1 that links the propensity score for the pooled sample, π_{zi}, to the convenience (CES) inclusion probabilities, π_{ci}, and the reference sample inclusion probabilities, π_{ri}.

The LDB is designed to cover the target population; therefore, we set $\pi_r = 1$ for all units in LDB, regardless of $\mathbf{x}_i$. In addition, coverage probabilities are set $p_r = 1$ and $p_c = 1$ for all units. In the case that LDB frame were insufficient and didn't cover all of the CES sample we could set $p_r < 1$ in our set-up to account for it. We observe: $z = 1$ for units in the CES sample and $z = 0$ for units in the LDB. Note, even though CES units are a subset of LDB, we are not concerned with matching the CES to LDB. Instead, we stack the two sets together. Thus, CES units appear in the stacked set *twice*: once with $z = 1$ and again with $z = 0$.

We apply our model to estimate probabilities $\pi_c(\mathbf{x}_i)$ of inclusion into the CES sample, where $\mathbf{x}_i$ is employment level of unit i in September (the benchmark month). We formulate our model with domain level random effects u_d and use splines as described in Section 4.

The fit performance is assessed by comparing CES based estimates to QCEW-based employment levels that become available to researchers on a lagged basis. Due to different seasonality patterns between the employment series derived from QCEW data and CES, the most meaningful comparison of the two series is after 12 months of estimation. Mimicking the production setup, we obtain level estimates after 12 months of estimation from monthly ratio estimates, $\hat{R}_{d,\tau}$, for a set of domains $d \in 1,\ldots,N$ at month τ. The monthly ratio estimates are multiplied together and by the September starting level, $Y_{d,0}$, that is available to CES at the start of the estimation cycle,

$$\hat{Y}_{d,12} = Y_{d,0} \prod_{\tau=1}^{12} \hat{R}_{d,\tau}.$$

Monthly ratio estimates $\hat{R}_{d,\tau}$ are obtained using a link relative (LR) estimator, that is a ratio of the sum of the current month to the sum of previous month responses, over set $s_{d,\tau}$ of CES

respondents at a given month τ in domain d: $\hat{R}^{LR}_{d,\tau} = \sum_{i\in s_{d,\tau}} y_{i,\tau} / \sum_{i\in s_{d,\tau}} y_{i,\tau-1}$. Once we apply our approach to obtain pseudo weights w_i, we use them in the analogous formula to form a "pseudo" *weighted* link relative (WLR) estimator, $\hat{R}^{WLR}_{d,\tau} = \sum_{i\in s_{d,\tau}} w_i y_{i,\tau} / \sum_{i\in s_{d,\tau}} w_i y_{i,\tau-1}$.

We extract the posterior means of the pseudo weights and apply them to each month in the estimation cycle. Figure 6 displays examples of estimates of employment levels over the 12 months of the estimation cycle for California MSAs under both the LR and WLR estimators, both compared to the QCEW Historical (Hist) truth (that we obtain on a lagged basis). We readily see that our pseudo weighted WLR estimator generally does a better job of estimating the truth.

Figure 7 shows the distribution of *annual revisions* of the level estimates based on LR and WLR methods, respectively, over the set of MSAs in California. The annual revision, $rev_{d,12}$, is defined as the difference between the respective estimate, $\hat{Y}_{d,12}$, and "true" population level $Y_{d,12}$ that becomes available after the fact, at the 12th month after the benchmark month:

$$rev_{d,12} = \hat{Y}_{d,12} - Y_{d,12}.$$

Again, the WLR estimator demonstrates better fit performance than does the LR estimator in that the distribution of revision magnitudes is more compact.

To compute variance $v^{WLR}_{d,\tau}$ of the WLR estimate of relative change $R_{d,\tau}$, we extract 10 draws from the posterior distribution of the fitted pseudo weights, estimate sampling variance for each draw of the pseudo-weights and then use a multiple imputation procedure described in a Appendix C to compute the total variance of $R_{d,\tau}$ in a manner that accounts for the uncertainty in the estimation of weights. Coefficients of variations, $cv_{d,\tau} = \sqrt{v^{WLR}_{d,\tau}}/R_{d,\tau}$, are presented in Figure 8, where they are plotted against the employment level of respective domains. It can be observed that variances tend to be smaller in larger domains, as is expected.

7. Discussion

We introduced a novel approach that derived an exact relationship between the sample propensity score, π_{zi}, on the one hand, and the reference and convenience samples conditional inclusion probabilities, π_{ri}, and π_{ci}, on the other hand for an observed pooled sample. Our expression is valid for any size of the overlap between the reference and convenience samples. It allows us to specify a likelihood directly for the sample using π_{zi} and our specification of a Bayesian hierarchical probability model to simultaneously estimate all of them.

A. Estimation of Inclusion Probabilities Under Symmetric Two-arm Sampling

Our main method derives an expression connecting $(\pi_{zi}, \pi_{ci}, \pi_{ri})$ on the observed sample from first principles using the survey sampling literature. We proceed on an alternative path that also connects these quantities based on the economics literature. We will see in the sequel that this alternate path produces the same estimator, though they are derived from

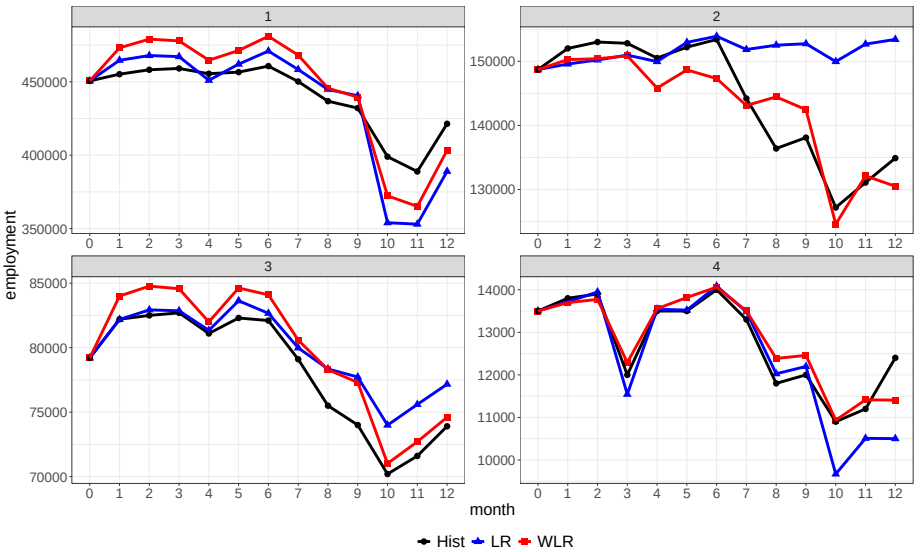

Figure 6: Examples of a 12-month CES estimation cycle for select MSAs in California, series starting from September 2019 true levels. The black solid line corresponds to the "true" monthly levels ("Hist", obtained after the fact from historical series), the blue line with triangles shows estimates based on unweighted monthly link relatives ("LR"), and the red line with squares shows estimates based on the weighted link relatives ("WLR").

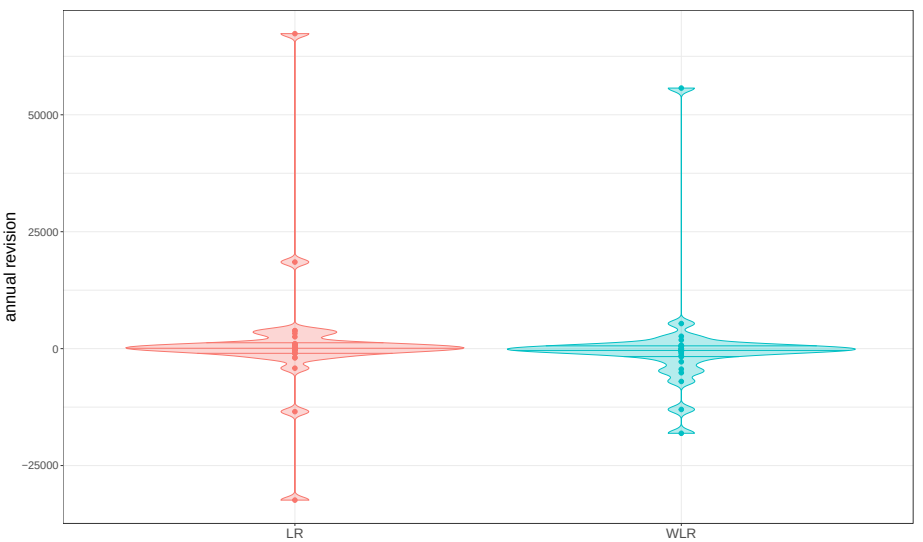

Figure 7: Distribution of annual revisions for MSAs in California. "Annual revisions" are differences between respective level estimates (LR or WLR based) and the true historical levels at the 12th month of the estimation cycle.

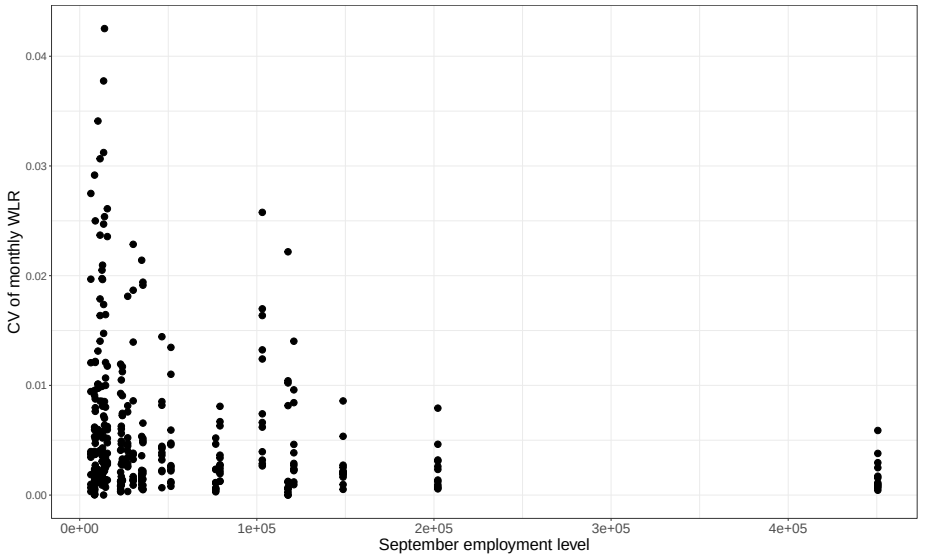

Figure 8: Coefficients of variation (CV) of monthly WLR estimates for 12 months of estimation versus respective domains employment levels at the start of the estimation period in September 2019, for California MSAs.

completely different approaches.

Lancaster and Imbens (1996) provide a modeling formulation for estimation of the conditional sample inclusion probabilities for case observations of interest under a two-arm experimental design with one-arm consisting of cases and the other consisting of an unknown collection of both case observations and control observations. Under an observed sample from each arm, they assume a $2-$step sample observation process where the first step is a Bernoulli draw for the observed sub-sample indicator into either a case sample arm or a mixed case and control sample arm, given the observed sample. The second step consists of the realization or appearance of units from the selected arm in the first stage. The process parameterizes an exact likelihood for the distribution for predictors, $\mathbf{x}$, conditioned on the sub-population of cases in that sampling arm and a marginal population distribution for $\mathbf{x}$ in the mixed arm. Using the distribution for $\mathbf{x}$ allows a clever and simple specification of the marginal distribution for $\mathbf{x}$ since they don't know the mix of cases and controls in the second arm. The conditional distribution in the case sampling arm is a function of the case sample conditional inclusion probability (by Bayes rule) parameterized by regression coefficients. This approach has the virtue of simultaneous estimation of conditional propensity scores and the conditional inclusion probabilities for cases.

We proceed to specialize and extend their $2-$step sample observation process and use of conditional distributions for $\mathbf{x}$ to our set-up of reference and control sampling arms and will specify a likelihood in each arm based on the sub-population of units linked to each type of sample.

Let $z_i \in \{0,1\}$ be the same binary inclusion indicator of selection into the convenience

sample for unit $i \in (1,\ldots,n)$ used in the previous section. When $z_i = 0$ unit i is drawn from the reference sample. We suppose the observed two-arm sample (with convenience and reference sample arms) arises from a Bernoulli draw into either arm with probability $P(z_i = 1) = P(i \in S_c \mid i \in S)$ and subsequently specify a conditional sub-population distribution for $\mathbf{x}_i$ whose form depends on the outcome of the Bernoulli draw for each unit, $i \in (1,\ldots,n)$. In particular, $p(\mathbf{x}_i \mid i \in S_c) = \pi_c(\mathbf{x}_i \mid \beta_c) \times f(\mathbf{x}_i)/P(i \in S_c \mid i \in S, i \in U_c)$ for the convenience sample by Bayes rule where we recall that $\pi_c(\mathbf{x}_i \mid \beta_c) = P(i \in S_c|\mathbf{x}_i, \beta_c)$. We drop the conditioning on U_c and U_r in the sequel where the context is obvious for readability. We have included regression parameters β_c that parameterizes a model for unknown $\pi_c(\mathbf{x}_i \mid \beta_c)$ that we wish to estimate. By a symmetric process for the reference sample we have, $p(\mathbf{x}_i \mid i \in S_r) = \pi_r(\mathbf{x}_i \mid \beta_r) \times f(\mathbf{x}_i)/P(i \in S_r \mid i \in S)$.

We note that *both* specifications for conditional distributions for $\mathbf{x}_i$ in each sampling arm use the same marginal distribution, $f(\mathbf{x}_i)$, because both samples are drawn from the same underlying population.

Let $q = P(i \in S_c) = \int \pi_c(\mathbf{x}_i \mid \beta_c) f(x) dx$ and $t = P(i \in S_r) = \int \pi_r(\mathbf{x}_i \mid \beta_r) f(x) dx$ denote the unknown marginal probabilities used above to specify the conditional distributions for $\mathbf{x}_i$ in each sampling arm. The marginal (over predictors, $\mathbf{x}$) probability for a unit to be selected into a sampling arm is denoted by $h = P(z_i = 1) = P(i \in S_c \mid i \in S)$. All of $(h, q, t, \beta_c, \beta_r)$ are unknown parameters that will receive prior distributions to be updated by the data.

The conditional distributions for $\mathbf{x}_i$ in each arm and the marginal probabilities for selection into each arm parameterize the likelihood for $(h, q, t, \beta_c, \beta_r)$,

$$L(h,q,t,\beta_c,\beta_r \mid \mathbf{z}, X) \times \prod_{i=1}^{n} f(\mathbf{x}_i) = \prod_{i=1}^{n} (h\pi_c(\mathbf{x}_i \mid \beta_c)/q)^{z_i} \times (((1-h)\pi_r(\mathbf{x}_i \mid \beta_r)/t)^{1-z_i} \times f(\mathbf{x}_i), \tag{17}$$

where we factor out the $f(\mathbf{x}_i)$ on both sides and subsequently propose to drop these marginal distributions for the covariates because we don't believe they are random. We use $f(\mathbf{x}_i)$ as a computation device to allow us to specify a likelihood with conditional distributions of $\mathbf{x}_i$ in each sampling arm.

We proceed to reparameterize Equation 17 by extending an approach of Johnson et al. (2021) from the case-control setting to our two-arm sampling set-up. We construct the following transformed parameters:

$$\begin{aligned} \psi &= q(1-h)/th \\ \pi_{zi} &= \pi_c(\mathbf{x}_i \mid \beta_c)/(\pi_c(\mathbf{x}_i \mid \beta_c) + \psi\pi_r(\mathbf{x}_i \mid \beta_r)) \\ 1 - \tilde{q}_i &= (1-h)\pi_r(\mathbf{x}_i \mid \beta_r)/t. \end{aligned} \tag{18}$$

Using the transformations of Equation 18 allows us to reparameterize the conditional likelihood (after dropping $f(\mathbf{x}_i)$ in Equation 17) to,

$$L(h,q,t,\beta_c,\beta_r \mid \mathbf{z}, X) = \prod_{i=1}^{n} \pi_{zi}^{z_i}(1-\pi_{zi})^{1-z_i} \times \frac{1-\tilde{q}_i}{1-\pi_{zi}}, \tag{19}$$

which is a product of a Bernoulli distributed term and a ratio of transformed parameters.

We examine the non-Bernoulli likelihood contribution, $\prod_{i=1}^{n} \frac{1-\tilde{q}_i}{1-\pi_{zi}}$ asymptotically as the reference sample size, $n_r \uparrow \infty$, under a fixed convenience sample size, n_c. We present a theoretical result in the following section that demonstrates the log of this ratio contribution to the likelihood limits to 0, asymptotically for n_r and n_r/n_c both sufficiently large to allow the ignoring or dropping of this term.

We may construct a model using the Bernoulli likelihood,

$$L(h,q,t,\beta_c,\beta_r \mid \mathbf{z},X) = \prod_{i=1}^{n} \pi_{zi}^{z_i}(1-\pi_{zi})^{1-z_i}, \tag{20}$$

with associated propensity,

$$\pi_{zi} = \pi_c(\mathbf{x}_i)/(\pi_c(\mathbf{x}_i) + \psi\pi_r(\mathbf{x}_i)) \tag{21}$$

where we have suppressed (β_c,β_r) to facilitate comparison with $\pi_z(\mathbf{x}_i) = \pi_c(\mathbf{x}_i)/(\pi_c(\mathbf{x}_i) + \pi_r(\mathbf{x}_i))$ from our main method.

We see that the propensity formulation here and under our main method are nearly identical, up to an inclusion of ψ in the denominator under the Symmetric two-arm approach, despite both being derived from different principles. We prove in the next section that under the above definitions for (h,q,t) that ψ must equal 1, which may be seen intuitively by noting that for a sample size, n, sufficiently large we may plug in modal quantities, $(h = n_c/n, q = n_c/N, t = n_r/N)$, for those marginal probabilities which produces $\psi = 1$. Our proof for $\psi = 1$ is true, however, for any sample size. The implication is that we have arrived at the very same result for the likelihood and conditional propensity, π_{zi}, as developed under our main approach. The reverse implication is that the classical setting for Lancaster and Imbens (1996) could be estimated more efficiently by setting $\psi = 1$. Investigating whether this simplification for $\psi = 1$ holds for more complex applications such as k-indexed simultaneous outcomes with uniques values for ψ_k (Johnson et al., 2021) is a subject for future work.

A.1 Proof that $\log\left(\frac{1-\tilde{q}_i}{1-\pi_{zi}}\right)$ asymptotically contracts on 0.

This proof performs an extension to the corresponding proof for stratified use-availability designs found in Johnson et al. (2021) to our case of a the two-arm sampling design under an arbitrary sampling design.

The pseudo log likelihood contribution $\sum_{i=1}^{n} \log\left(\frac{1-\tilde{q}_i}{1-p_{zi}}\right)$ contracts on 0 as the reference sample grows, $n_r \uparrow \infty$ and $h = n_c/n \downarrow 0$. We begin with some simple algebra to state the

likelihood term with marginal probabilities, (h, q, t),

$$\begin{aligned}\prod_{i=1}^{n} \frac{1-\tilde{q}_i}{1-\pi_{zi}} &= \prod_{i=1}^{n} \frac{(1-h)\pi_{ri}}{t} \times \frac{\pi_{ci} + \psi\pi_{ri}}{\psi\pi_{ri}} \\ &= \prod_{i=1}^{n} (1-h) \times \left(\frac{\pi_{ci}}{\psi t} + \frac{\pi_{ri}}{t}\right) \\ &= \prod_{i=1}^{n} (1-h) \times \left(\frac{\pi_{ci}}{\psi_c} + \frac{\pi_{ri}}{t}\right)\end{aligned} \tag{22}$$

where for readability we simplify the expression of $\pi_c(\mathbf{x}_i \mid \beta_c)$ with the short-hand, π_{ci}, and the same for π_{ri}. We plug in for $\psi = \frac{q(1-h)}{h} \times \frac{1}{t} = \frac{\psi_c}{t}$ into the last equation in the series where ψ_c is composed of quantities solely related to the convenience sample.

We take the logarithm of the last equation of Equation 22,

$$\log\left(\prod_{i=1}^{n} (1-h) \times \left(\frac{\pi_{ci}}{\psi_c} + \frac{\pi_{ri}}{t}\right)\right) = n\log(1-h) + \sum_{i=1}^{n} \log\left(\frac{\pi_{ci}}{\psi_c} + \frac{\pi_{ri}}{t}\right). \tag{23}$$

We proceed to take a Taylor series expansion of $\log\left(\frac{\pi_{ci}}{\psi_c} + \frac{\pi_{ri}}{t}\right)$ about $\frac{\pi_{ci}}{\psi_c} = 0$ and use the first term, which we may do since $\frac{\pi_{ci}}{\psi_c}$ grows vanishingly small in the limit as $n \uparrow \infty$ (since $h \downarrow 0$ such that $\psi_c \uparrow \infty$). This produces,

$$n\log(1-h) + \sum_{i=1}^{n} \log\left(\frac{\pi_{ci}}{\psi_c} + \frac{\pi_{ri}}{t}\right) = n\log(1-h) + \sum_{i=1}^{n} \frac{\pi_{ci}}{\psi_c}\frac{t}{\pi_{ri}} \tag{24}$$

$$= n\log(1-h) + \frac{t}{\psi_c}\sum_{i=1}^{n} \frac{\pi_{ci}}{\pi_{ri}} \tag{25}$$

$$= n\log(1-h) + \frac{t}{\psi_c} \times \overline{\left[\frac{\pi_c}{\pi_r}\right]} \tag{26}$$

$$= n\log(n-n_c) - n\log n + \frac{n_c}{n-n_c} \times \frac{nt}{q}\overline{\left[\frac{\pi_c}{\pi_r}\right]}, \tag{27}$$

where we have plugged in $h = n_c/n$ for n sufficiently large and $\overline{\left[\frac{\pi_c}{\pi_r}\right]}$ represents the mean of the ratio, $\frac{1}{n}\sum_{i=1}^{n} \frac{\pi_{ci}}{\pi_{ri}}$.

We next evaluate the limit of the above expression as $n_r \uparrow \infty$ under a constant value for

n_c,

$$\lim_{n_r\uparrow\infty} n\log(n-n_c) - n\log n + \frac{n_c}{n-n_c} \times \frac{nt}{q}\overline{\left[\frac{\pi_c}{\pi_r}\right]} = -n_c + n_c\frac{t}{q}\overline{\left[\frac{\pi_c}{\pi_r}\right]} \tag{28a}$$

$$= -n_c + n_c\frac{t}{q}\frac{\overline{\pi_c}}{\overline{\pi_r}} \tag{28b}$$

$$= -n_c + n_c\frac{t}{q}\frac{q}{t} \tag{28c}$$

$$= -n_c + n_c. \tag{28d}$$

In Equation 28b, the mean of the ratios contracts on the ratio of the means as $n \uparrow \infty$ because π_{ci} limits to 0 as n increases since n_c is fixed such that the $\lim_{n\uparrow\infty}\frac{\pi_{cn}}{\pi_{rn}}$ exists and is finite. Also required is that π_{rn} be non-decreasing as n increases, which we may achieve through reordering the terms. Next, we apply the Law of Large Numbers for the convergence of the sample mean under the assumption of absolutely bounded values in expectation for q and t.

A.2 Proof that $\psi = 1$ under the Symmetric Two-arm Method of Section A

Let marginal (over $\mathbf{x}_i$) probabilities be defined as, $h = P(i \in S_c \mid i \in S)$, $q = P(i \in S_c \mid i \in U)$, $t = P(i \in S_r \mid i \in U)$ and further define $\psi = q(1-h)/th$. Let $\mathscr{S}$ denote the space of all two-arm samples, (S_c, S_r) of size (n_c, n_r), respectively. Recall that U is the set of two stacked populations $\{U^0, U^0\}$ corresponding to each arm. Then if we assume strictly positive conditional inclusion probabilities for all units and that the convenience sample arises from an underlying latent random sampling design then,

$\psi = 1$ a.s. P_π, where P_π is the unknown true joint generating distribution for all $(S_c, S_r) \in U \subset \mathscr{S}$, given U. For any $S = S_c + S_r \in U \subset \mathscr{S}$,

$$\psi = \frac{q(1-h)}{th} \tag{29a}$$

$$\psi\frac{h}{q} = \frac{1-h}{t} \tag{29b}$$

$$\psi\frac{P(i \in S_c \mid i \in S)}{P(i \in S_c \mid i \in U)} = \frac{P(i \in S_r \mid i \in S)}{P(i \in S_r \mid i \in U)} \tag{29c}$$

$$\psi\frac{P(i \in S_c \mid i \in S)P(i \in S \mid i \in U)}{P(i \in S_c \mid i \in U)} = \frac{P(i \in S_r \mid i \in S)P(i \in S \mid i \in U)}{P(i \in S_r \mid i \in U)} \tag{29d}$$

$$\psi P(i \in S \mid i \in S_c) = P(i \in S \mid i \in S_r) \tag{29e}$$

$$\psi = 1, \tag{29f}$$

where in Equation 29d we multiply both left- and right-hand side of Equation 29c by $P(i \in S \mid i \in U) > 0$.

Remark: When the reference sample is the population $S_r = U^0$, the proof holds without modification.

B. Stan Model Estimation Script

We present the Stan estimation script (Gelman et al., 2015) for our two-arm exact likelihood method, below. The script is built around Stan's `partial_sum` function to allow within chain parallelization for computational scalability.

```
functions{

  vector build_b_spline(vector t, vector ext_knots, int ind, int order);
  matrix build_mux(int N, int start, int end, int K_sp, matrix X, matrix[] G,
                   matrix beta_x, matrix[] beta_w);
  row_vector build_muxi(int K_sp, int num_basis, row_vector x_i, vector[] g_i,
                        matrix beta_x, matrix[] beta_w);
  vector build_b_spline(vector t, vector ext_knots, int ind, int order) {
    // INPUTS:
    //    t:          the points at which the b_spline is calculated
    //    ext_knots:  the set of extended knots
    //    ind:        the index of the b_spline
    //    order:      the order of the b-spline
    vector[num_elements(t)] b_spline;
    vector[num_elements(t)] w1 = rep_vector(0, num_elements(t));
    vector[num_elements(t)] w2 = rep_vector(0, num_elements(t));
    if (order==1)
      for (i in 1:num_elements(t)) // B-splines of order 1 are piece-wise constant
        b_spline[i] = (ext_knots[ind] <= t[i]) && (t[i] < ext_knots[ind+1]);
    else {
      if (ext_knots[ind] != ext_knots[ind+order-1])
        w1 = (to_vector(t) - rep_vector(ext_knots[ind], num_elements(t))) /
             (ext_knots[ind+order-1] - ext_knots[ind]);
      if (ext_knots[ind+1] != ext_knots[ind+order])
        w2 = 1 - (to_vector(t) - rep_vector(ext_knots[ind+1], num_elements(t))) /
                 (ext_knots[ind+order] - ext_knots[ind+1]);
      // Calculating the B-spline recursively as linear interpolation of two lower-order splines
      b_spline = w1 .* build_b_spline(t, ext_knots, ind, order-1) +
                 w2 .* build_b_spline(t, ext_knots, ind+1, order-1);
    }
    return b_spline;
  }

  matrix build_mux(int N, int start, int end, int K_sp, matrix X, matrix[] G, matrix beta_x, matrix[] beta_w){
    matrix[N,2] mu_x;
    for( arm in 1:2 )
    {
      mu_x[1:N,arm]     = X[start:end,] * to_vector(beta_x[,arm]); /* N x 1 for each arm */
      // spline term
      for( k in 1:K_sp )
      {
        mu_x[1:N,arm]  += to_vector(beta_w[arm][,k]' * G[k][,start:end]); /* N x 1 */
      } /* end loop k over K predictors */

    }/* end loop arm over convenience and reference sample arms */

    return mu_x;
  }

  row_vector build_muxi(int K_sp, int num_basis,
     row_vector x_i, vector[] g_i, matrix beta_x, matrix[] beta_w){

     row_vector[2] mu_xi;

     for( arm in 1:2 )
     {
       mu_xi[arm]     = dot_product(x_i,beta_x[,arm]); /* scalar */
       // spline term
       for( k in 1:K_sp )
       {
         mu_xi[arm]   += dot_product(beta_w[arm][1:num_basis,k], g_i[k][1:num_basis]); /* scalar */
       } /* end loop k over K predictors */

     }/* end loop arm over convenience and reference sample arms */

    return mu_xi;

  } /* end function build_mu */

  real partial_sum(int[] s,
                   int start, int end, real[] logit_pw, int K_sp, int n_c, int n,
```

```
                int num_basis, matrix X, matrix[] G, matrix beta_x, matrix[] beta_w,
                real phi_w) {
    int N = end - start + 1;
    matrix[N,2] mu_x;
    matrix[N,2] p;
    vector[N] p_tilde; // this pseudoprobability must be in [0,1]
    real fred             = 0;

    // memo: slicing on all of mu_x[1i,arm], p[1i,arm], p_tilde[1i] for li in 1:(end-start+1)
    //       where p_tilde is the mean vector for binary data vector, s, and mu_x[,2]
    //       is the mean vector for data vector logit_pw.
    //       Also slicing data vectors s and logit_pw in their respective
    //       log-likelihood contributions.

    mu_x             = build_mux(N, start, end, K_sp, X, G, beta_x, beta_w);

    p                = inv_logit( mu_x );


   // 1. bernoulli likelihood contribution
   p_tilde          = p[1:N,1] ./ ( p[1:N,1] + p[1:N,2] );
   fred             += bernoulli_lpmf( s[1:N] | p_tilde );

   // 2. normal likelihood contribution
   // In non-threaded model, likelihood statement for n - n_c, logit_pw
   // logit_pw          ~ normal( mu_x[(n_c+1):n,2], phi_w );
   // slicing on n length vector mu_x[,2]
   // subsetting portion of mu_x[,2] linked to logit_pw
   if( start > n_c  ) ## all units in this chunk increment likelihood contribution for logit_pw
   {
     fred       += normal_lpdf( logit_pw[start-n_c:end-n_c] | mu_x[1:N,2], phi_w );
   }else{ /* start <= n_c */
     if( end > n_c ) /* some units in chunk < n_c and some > n_c; only those > n_c increment likelihood */
     {
       fred                     += normal_lpdf( logit_pw[1:end-n_c] | mu_x[n_c-start+2:N,2], phi_w );

     } /* end if statement on whether n_c \in (start,end ) */
   } /* end if-else statement on whether add logit_pw likelihood contributions */


  return fred;
  }/* end function partial_sum() */

} /* end function block */

data{
    int<lower=1> n_c; // observed convenience (non-probability) sample size
          int<lower=1> n_r; // observed reference (probability) sample size
          int<lower=1> N; // estimate of population size underlying reference and convenience samples
          int<lower=1> n; // total sample size, n = n_c + n_r
          int<lower=1> num_cores;
          int<lower=1> multiplier;
          int<lower=1> K; // number of fixed effects
          int<lower=0> K_sp; // number of predictors to model under a spline basis
          int<lower=1> num_knots;
    int<lower=1> spline_degree;
    matrix[num_knots,K_sp] knots;
    real<lower=0> weights[n_r]; // sampling weights for n_r observed reference sample units
    matrix[n_c, K] X_c; //  *All* predictors - continuous and categorical -  for the convenience units
    matrix[n_r, K] X_r; // *All* predictors - continuous and categorical -  for the reference units
    matrix[n_c, K_sp] Xsp_c; //  *Continuous* predictors under a spline basis for convenience units
    matrix[n_r, K_sp] Xsp_r; // *Continuous* predictors under a spline basis for convenience units
    int<lower=1> n_df;
} /* end data block */

transformed data{
  // create indicator variable of membership in convenience or reference samples
  // indicator of observation membership in the convenience sample
  int grainsize                     = ( n / num_cores ) / multiplier;
  real logit_pw[n_r]                = logit(inv(weights));
  int<lower=0, upper = 1> s[n]      = to_array_1d( append_array(rep_array(1,n_c),rep_array(0,n_r)) );
  matrix[n,K] X                     = append_row( X_c,X_r );
  matrix[n,K_sp] X_sp               = append_row( Xsp_c,Xsp_r );
  /* formulate spline basis matrix, B */
  int num_basis = num_knots + spline_degree - 1; // total number of B-splines
  matrix[spline_degree + num_knots,K_sp] ext_knots_temp;
  matrix[2*spline_degree + num_knots,K_sp] ext_knots;
  matrix[num_basis,n] G[K_sp]; /* basis for model on p_c */
  for(k in 1:K_sp)
  {
     ext_knots_temp[,k] = append_row(rep_vector(knots[1,k], spline_degree), knots[,k]);
    // set of extended knots
     ext_knots[,k] = append_row(ext_knots_temp[,k], rep_vector(knots[num_knots,k], spline_degree));
     for (ind in 1:num_basis)
     {
        G[k][ind,] = to_row_vector(build_b_spline(X_sp[,k], ext_knots[,k], ind, (spline_degree + 1)));
     }
```

```
    G[k][num_knots + spline_degree - 1, n] = 1;
  }

} /* end transformed data block */

parameters {
  matrix<lower=0>[K,2] sigma2_betax; /* standard deviations of K x 2, beta_x */
                                  /* first column is convenience sample, "c", and second column is "r" */

  matrix[K,2] betaraw_x; /* fixed effects coefficients - first colum for p_c; second column for p_r  */
  // spline coefficients
  vector<lower=0>[2] sigma2_global; /* set this equal to 1 if having estimation difficulties */
  matrix<lower=0>[2,K_sp] sigma2_w;
  matrix[num_basis,K_sp] betaraw_w[2];  // vector of B-spline regression coefficients for each predictor, k
                                        // and 2 sample arms
  real<lower=0> phi2_w; /* scale parameter in model for -log(weights) */
} /* end parameters block */

transformed parameters {
  matrix[K,2] beta_x;
  matrix[num_basis,K_sp] beta_w[2];
  matrix<lower=0>[K,2] sigma_betax;
  vector<lower=0>[2] sigma_global; /* set this equal to 1 if having estimation difficulties */
  matrix<lower=0>[2,K_sp] sigma_w;
  real<lower=0> phi_w;

  sigma_betax       = sqrt( sigma2_betax );
  sigma_global      = sqrt( sigma2_global );
  sigma_w           = sqrt( sigma2_w );
  phi_w             = sqrt( phi2_w );

  // for scale parameters for interaction effects from those for main effects to which they link
  for( arm in 1:2 )
  {
     beta_x[,arm]   = betaraw_x[,arm] .* sigma_betax[,arm]; /* Non-central parameterization */
  }/* end loop arm over convenience and reference sample arms */

  // spline regression coefficients
  for(arm in 1:2)
  {
    for( k in 1:K_sp )
    {
      beta_w[arm][,k]   = cumulative_sum(betaraw_w[arm][,k]);
      beta_w[arm][,k]  *= sigma_w[arm,k] * sigma_global[arm];
    } /* end loop k over K predictors */
  }

} /* end transformed parameters block */

 model {
  to_vector(sigma2_betax) ~ gamma(1,1);
  to_vector(sigma2_w)     ~ gamma(1,1);
  sigma_global            ~ gamma(1,1);
  phi2_w                  ~ gamma(1,1);

  to_vector(betaraw_x)   ~ std_normal();
  for(arm in 1:2)
    to_vector(betaraw_w[arm])   ~ std_normal();

  /* Model likelihood for y, logit_pw */
 // Sum terms 1 to n in the likelihood
 target            += reduce_sum(partial_sum, s, grainsize,
                                          logit_pw, K_sp, n_c, n, num_basis, X, G,
                                          beta_x, beta_w, phi_w);

} /* end model block */

generated quantities{
  matrix[n,2] p;
  matrix[n,2] mu_x;

  for( arm in 1:2 )
  {
    mu_x[,arm]     = X[,] * to_vector(beta_x[,arm]); /* n x 1 for each arm */
    // spline term
    for( k in 1:K_sp )
    {
      mu_x[,arm]  += to_vector(beta_w[arm][,k]' * G[k][,1:n]); /* n x 1 */
    } /* end loop k over K predictors */

   }/* end loop arm over convenience and reference sample arms */

  p        = inv_logit( mu_x );

  // smoothed sampling weights for convenience and reference units
  vector[n] weights_smooth_c = inv(p[,1]);
  vector[n] weights_smooth_r = inv(p[,2]);
```

```
  // inclusion probabilities in convenience and reference units for convenience units
  // use for soft thresholding
  vector[n_c] pi_c                  = p[1:n_c,1];
  vector[n_c] pi_r_c                = p[1:n_c,2];
  // normalized weights
  weights_smooth_c                  *= ((n_c+0.0)/(n+0.0)) * (sum(weights_smooth_r)/sum(weights_smooth_c));
  weights_smooth_r                  *= ((n_r+0.0)/(n+0.0));
} /* end generated quantities block */
```

C. Simulation Results for Estimating Population Mean, μ

We use the convenience sample inclusion probabilities, $\pi_{cmi},\ i \in S_c$, estimated from models on each Monte Carlo iteration, $m = 1,\ldots,(M = 30)$, to form a population mean estimator, μ_m. As discussed in the introduction, we use our Bayesian hierarchical model to estimate π_{cmi}, such these latent sampling probabilities may be used to construct survey estimators for focus response variables. We construct $\mu_m = \frac{\sum_{i\in S_c} y_i/\hat{\pi}_{cmi}+\sum_{j\in S_r} y_j/\pi_{rmj}}{\sum_{i\in S_c} 1/\hat{\pi}_{cmi}+\sum_{j\in S_r} 1/\pi_{rmj}}$ as a sample-weighted (Hajek) survey direct estimator, so there is no additional model specified; that is, the estimator each μ_m assumes the underlying population values for y_{mi} are fixed such the estimator is random with respect to the distribution that governs the taking of samples from that fixed population.

We propagate uncertainty in the model-based estimation of the convenience sample inclusion probabilities by taking multiple draws or imputes (e.g., $J = 10$) of each inclusion probability from its posterior distribution. We formulate the survey direct estimator using that draw of the inclusion probabilities. We compute the variance of the survey estimator for the population mean with respect to the survey sampling distribution. We repeat this procedure for each draw and then compute the between draws variance variance with respect to the modeling distribution. We put it together by using multiple imputation combining rules to construct a total variance for our survey direct estimate that now incorporates uncertainty with respect to both the model for estimating inclusion probabilities and the distribution governing the taking of samples.

More specifically, we construct a total, combined variance of the form $T = (1+1/J)B+\bar{U}$ based on multiple imputation rules of Reiter and Raghunathan (2007, See section 2.1.1), where T denotes the total variance of our μ estimator that accounts for both uncertainty with respect to drawing samples and with respect to the modeling of the inclusion probabilities used to form sampling weights. Let $j \in 1,\ldots,M$ index a randomly drawn imputation for $(\hat{\pi}_{cji})_{i\in S_c}$, $(\hat{\pi}_{rji'})_{i'\in S_r}$ from the set of MCMC samples for a model run. $\bar{U}$ denotes the within imputation sampling variance of μ_j and B denotes the between modeled variance of μ_j across the J imputations.

Once the total variance is computed, we then generate symmetric asymptotic intervals using the $t-$ distribution. The use of multiple imputation allows us to propagate the uncertainty in estimation of π_{cji} into the variance estimate for our direct estimator of μ_j. We next present details to construct the within impute variance, $\bar{U}$, and the between impute variance, B.

In what follows, we assume that we use the model-smoothed estimator, $\hat{\pi}_{rji} = \mu_{x,rji}$ (from Equation 14) for the reference sample inclusion probabilities to construct the mean statistic. We compare simulation study results for μ using both using the fixed π_{rji} and the

model-smoothed $\hat{\pi}_{rji}$ in the sequel.

Binder (1996) provides a general approach to Taylor linearization for computing the within impute variance. We fix an imputation $j \in (1,\ldots,M)$. For a simple weighted linear statistic such as μ_j, the approach simplifies to calculating the variance of the weighted residuals $w_{ji}(y_i - \mu_j)$ with weights $w_{ji} = \hat{\pi}_{cji} / \left(\sum_{i \in S_c} 1/\hat{\pi}_{cji} + \sum_{i' \in S_r} 1/\hat{\pi}_{rji'} \right)$ for convenience sample units or
$w_{ji'} = \hat{\pi}_{rji'} / \left(\sum_{i \in S_c} 1/\hat{\pi}_{cji} + \sum_{i' \in S_r} 1/\hat{\pi}_{rji'} \right)$ for reference sample units. We average over the J within-impute design (sample)-based variance estimates of μ_j (via Taylor linearization) to get $\bar{U}$.

We proceed to construct the model-based, between variance B by computing the variance over the J imputations for μ_j.

We first illustrate the benefit of incorporating the sample weighted convenience units with the reference sample units into the computation of μ. We then proceed to compare the pseudo likelihood methods for π_{ci} with our two-arm exact likelihood method under combined reference and convenience sample estimation of μ.

Finally, the two-arm method co-estimates $\pi_{ri},\ i \in S_c$ simultaneously with estimating π_{ci}. So, on each Monte Carlo iteration, m, we threshold or exclude those convenience sample units, $\{\ell \in S_c : \pi_{rm\ell} < \varepsilon\}$; that is, we exclude those convenience units that express small reference sample inclusion probabilities in order to reduce noisiness in our estimator. We experiment with setting $\varepsilon = (Q_1, Q_5, Q_{10})$, where Q_p is the p^{th} percent quantile of the distribution of smoothed $\pi_{ri},\ i \in S_r$.

Results are presented in the Figures 9 - 11. Each plot panel from left to right measures the bias, root mean squared error, mean absolute deviation and coverage for estimated μ.

We construct separate convenience samples under both the low- and high-overlap sampling designs used in the previous results for estimating the conditional convenience sample inclusion probabilities. In each row of every plot panel we present the result for the low-overlap sampling design, labeled “L" and the high-overlap sampling design, labeled "H" with those results connected by a horizontal bar. In practice, a convenience dataset will probably lie somewhere in between L and H.

Lastly, we lay in the result for the base case that constructs μ solely from the reference sample as a dashed black vertical line in each plot panel in all of the figures.

Figure 9 compares constructing μ solely from the convenience sample in the first row to using both the reference and convenience samples (both under true inclusion weights) in the second row. We see a dramatic improvement in the quality of estimated μ under the high-overlap convenience samples and a smaller, but still notable improvement under the low-overlap convenience samples, on the one hand, from use of solely the convenience samples, on the other hand.

Figure 10 compares the quality of estimated μ between our exact two-arm method (using published / known reference sample inclusion probabilities) with the pseudo likelihood methods. The first row presents the combined reference and convenience sample using the true values for the latent convenience sample weights as a comparator for all methods. The second row presents the combined reference and convenience samples now using the estimated convenience sample inclusion weights under our two-arm method. The performance

Figure 9: Performance of the weighted mean estimator between high (H) and low (L) overlapping samples using the convenience and reference sample with true weights across Monte Carlo Simulations for (top to bottom) Only Convenience, Convenience and Reference Sample. Left to right: Bias, root mean square error, mean absolute deviation, coverage of 90% intervals. Vertical reference line corresponds to using the reference sample only.

is very similar to using the true inclusion weights for the convenience sample. The third row presents the CLW method of Chen et al. (2020), which performs relatively poorly due to high estimation variation expressed by this method in estimation of π_{cmi} for our moderate sample sizes $(n_r, n_c) = (400, \sim 800)$. We achieve best performance for high-overlap convenience samples (labeled (H)). The last row presents the same, but using the WVL pseudo likelihood method of Wang et al. (2021) that expresses less variation in estimation of π_{cmi} than does CLW (though still substantially higher than our two-arm method). Yet, even though the estimated weights under WVL are biased under both low- and high-overlap samples, the method performs similarly in estimation of μ to our two-arm method because the bias for WVL is largest at high values for π_{ci} while most sampled units are assigned $\pi_{ci} < 0.75$. The low-overlap samples produce notably worse coverage under WVL due to the bias and failure to account for uncertainty in π_{rmi} by using them as plug-in.

It bears mention that even when using the true values for π_{ci} the coverage of the estimator for μ under the low-overlap datasets fails to achieve nominal coverage because of our moderate sample sizes. These moderate sample sizes realistically reflect the sampling of domains (e.g., geographic-by-industry for establishment surveys) used in practice. We render an estimator using the true sampling weights in each plot panel so that we may understand the performance of the methods in context of the best possible performance.

We conclude the exploration of methods for estimating π_{ci} on the quality of the resultant mean estimator, μ, by comparing versions or variations of the two-arm method for estimating π_{ci}. So far we have seen that the two-arm method outperforms the other methods and, in particular, the pseudo likelihood methods for estimation of μ in terms of bias, means squared error and converge (uncertainty quantification).

Since the two-arm method co-models π_{ri} to borrow strength in estimation of π_{ci}, we may use either fixed π_{ri} or modeled / smoothed values for π_{ri} to form our combined reference and convenience estimator for μ. The first row of Figure 11 presents the combined reference and convenience-based estimator for μ that uses true sample weights as the benchmark comparator. The second row uses fixed (and published) π_{ri} along with our two-arm method for estimating π_{ci} to produce the combined reference and convenience sample inclusion probability for μ. The third row is the same as the second except that we replace the fixed π_{ri} with modeled or smoothed values from our two-arm model. We observe that using smoothed weights improves the coverage for high-overlap datasets because co-modeling

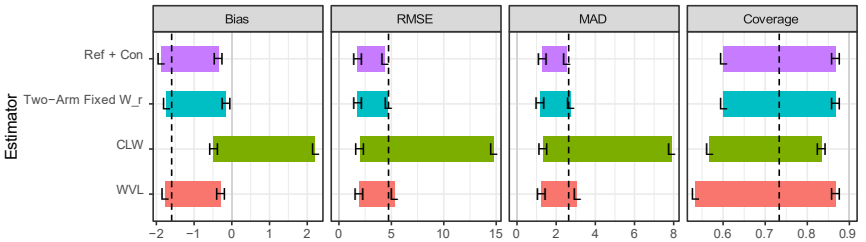

Figure 10: Performance of the weighted mean estimator between high (H) and low (L) overlapping samples using the convenience with modeled weights and reference sample with true weights across Monte Carlo Simulations for (top to bottom) True weights, Two-Arm weights, CLW weights, WVL weights. Left to right: Bias, root mean square error, mean absolute deviation, coverage of 90% intervals. Vertical reference line corresponds to using the reference sample only.

the π_{ri} accounts for uncertainty in the generation of samples (Leon-Novelo and Savitsky, 2019).

Lastly, we next leverage our co-estimation of π_{ri} for $i \in S_c$, which are typically *unknown* for non-overlapping units between the two sample arms, to threshold inclusion of units from the convenience sample. We seek to exclude those convenience sample units, $\ell \in S_c$ where the associated $\pi_{r\ell} < \varepsilon$; that is, we exclude units from the convenience sample that are estimated with very small values for $\pi_{r\ell}$ in order to remove units that would induce noise in our estimator for μ. We see that the estimator for μ that results from setting $\varepsilon = Q_1$ notably improves bias performance in the estimator for μ for low overlap samples while leading to only a slight increases in RMSE for high overlap. When we increase to $\varepsilon = Q_5$, bias increases slightly for the high-overlap datasets but further decreases for the low-overlap dataset. The coverage performance, however, notably improves under $\varepsilon = Q_5$ as compared to the non-thresholded two-arm-based estimator. The general pattern continues with $\varepsilon = Q_{10}$. This result suggests thresholding using $\varepsilon \approx Q_5$ would be advisable, particularly for lower overlapping samples.

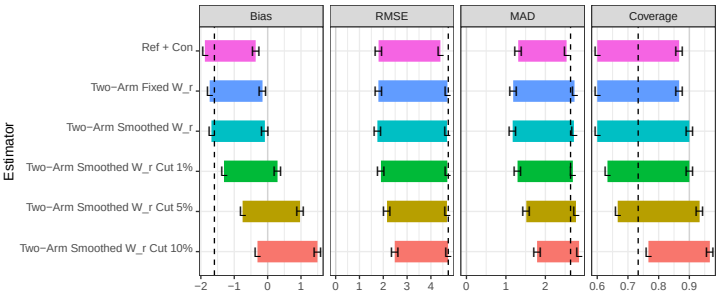

Figure 11: Performance of the weighted mean estimator between high (H) and low (L) overlapping samples using variations of the two-arm method across Monte Carlo Simulations for (top to bottom) True weights for both samples, Original weights for Reference Sample, Smoothed weights for reference sample, Subset of convenience sample meeting 1%, 5%, and 10% overlap threshold. Left to right: Bias, root mean square error, mean absolute deviation, coverage of 90% intervals. Vertical reference line corresponds to using the reference sample only.

References

Beaumont, J.-F., (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1–28.

Beresovsky, V., (2019). On application of a response propensity model to estimation from web samples. `https://www.researchgate.net/publication/333915871_On_application_of_a_response_propensity_model_to_estimation_from_web_samples`.

Bhattacharya, A., D. Pati, and Y. Yang, (2019). Bayesian fractional posteriors. *The Annals of Statistics*, 47(1), 39 – 66.

Binder, D. A., (1996). Taylor linearization for single phase and two phase samples: A cookbook approach. *Survey Methodology*, 17–26.

Carvalho, C. M., N. G., Polson, and J. G. Scott (2009, 16–18 Apr). Handling sparsity via the horseshoe. In D. van Dyk and M. Welling (Eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, Volume 5 of Proceedings of Machine Learning Research, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, pp. 73–80. PMLR.

Chen, Y., P. Li, and C. Wu, (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011–2021.

DiSogra, C., C. Cobb, E. Chan, and J. M. Dennis (2011). Calibrating nonprobability in-

ternet samples with probability samples using early adopter characteristics. *JSM Proceedings, Survey Research Methods Section*, Alexandria, VA: American Statistical Association., pp. 4501–4515.

Elliott, M. R., (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice* 2, 813–845.

Elliott, M. R. and R. Valliant, (2017). Inference for Nonprobability Samples. *Statistical Science*, 32(2), 249 – 264.

Gelman, A., D. Lee, and J. Guo, (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *In press. Journal of Educational and Behavior Science*.

Johnson, N. G., M. R. Williams, and E. C. Riordan, (2021). Generalized nonlinear models can solve the prediction problem for data from species-stratified use-availability designs. *Diversity and Distributions*, 27(11), 2077–2092.

Lancaster, T. and G. Imbens, (1996). Case-control studies with contaminated controls. *Journal of Econometrics*, 71(1-2), 145–160.

Leon-Novelo, L. G. and T. D. Savitsky, (2019). Fully Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 13(1), 1608 – 1645.

Reiter, J. P. and T. E. Raghunathan, (2007). The multiple adaptations of multiple imputation. Journal of the *American Statistical Association*, 102(480), 1462–1471.

Tillé, Y. and A. Matei, (2021). *sampling: Survey Sampling.* R package version 2.9.

Valliant, R., (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2), 231–263.

Valliant, R. and J. A. Dever, (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods and Research*, 40, 105–137.

Wang, L., R. Valliant, and Y. Li, (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med.*, 40(4), 5237–5250.

Williams, M. R. and T. D. Savitsky, (2021). Uncertainty Estimation for Pseudo-Bayesian Inference Under Complex Sampling. *International Statistical Review*, 89(1), 72–107.

Wu, C., (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2), 283–311.

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 35–44, https://doi.org/10.59170/stattrans-2023-062
Received – 31.08.2023; accepted – 27.06.2023

Power ratio cum median-based ratio estimator of finite population mean with known population median

Umar K. Abdullahi[1], Fidelis I. Ugwuowo[2], Nuanpan Lawson[3]

Abstract

The search for an efficient estimator of the finite population mean has been a critical problem to the sample survey research community. This study is motivated by the fact that the conducted literature review showed that no research has developed such an average ratio estimator of the population mean that would utilize both the population and the sample medians of study variable, as well as the Srivastava (1967) estimator at a time. In this paper we proposed the power ratio cum median-based ratio estimator of the finite population mean, which is a function of two ratio estimators in the form of an average. The estimator assumes the population to be homogeneous and skewed. The properties (i.e. the Bias and the Mean Squared Error – MSE) of the proposed estimator were derived alongside its asymptotically optimum MSE. We demonstrated the efficiency of the proposed estimator jointly with its efficiency conditions by comparing it to selected estimators described in the literature. Empirically, a real-life dataset from the literature and a simulation study from two skewed distributions (Gamma and Weibull) were used to examine the efficiency gain. The empirical analysis and simulation study demonstrated that the efficiency gain is significant. Hence, the practical application of the proposed estimator is recommended, especially in socio-economic surveys.

Key words: finite population mean, bias, mean squared error, power estimator, median-based, power ratio.

1. Introduction

Sampling is a technique for selecting a sample or subset of the population to make statistical inference on some characteristics of the whole population. The concept of utilizing means of auxiliary variable at estimation stage of a survey is due to Cochran (1940), the author expressed an estimator for population mean of study variable as

[1] Department of Statistics, University of Nigeria, Nsukka, Nigeria.
[2] Department of Statistics, University of Nigeria, Nsukka, Nigeria.
[3] Corresponding author. Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Wongsawang, Bangsue, Bangkok, Thailand. E-mail: nuanpan.n@sci.kmutnb.ac.th.

a function of mean per unit estimator of the study variable and ratio of population to sample means of an auxiliary variable, when the relationship between the study and auxiliary variables is positive. Srivastava (1967) defined yet another estimator of population mean of study variable in the form of ratio using a single auxiliary variable. In the developed estimator, the ratio of population to sample mean of the auxiliary variable is expressed in the form of power of a constant, where the constant is obtained in such a way that the mean squared error of the estimator is minimum. Olkin (1958) discussed the concept of ratio estimator with more than one auxiliary variables. The author defined a bivariate ratio estimator which utilizes two auxiliary variables, the estimator is expressed as a function of two ratio estimators in the form of average. The properties of the estimator were expressed and comparison was made with the estimator with one auxiliary variable. Gupta and Shabbir (2008) defined a general class of ratio-type estimator using weight function and some other known parameters of auxiliary variable, they used three real-life dataset to justify the efficiency gain due to the defined estimator, and observed that the developed estimator has the minimum MSE compared to linear regression estimator.

Recently Subramani (2016) defined an efficient median-based estimator of finite population mean using the median of the study variable. The estimator is a function of mean per unit estimator and ratio of the population to sample medians of the study variable. Subramani's estimator does not utilize any auxiliary parameter from auxiliary variable, but rather utilizes an auxiliary parameter from the same variable. Srija and Subramani (2018) defined a median-based estimators using mean, first and third quartiles of the auxiliary variable.

In the same vein, Abdullahi and Ugwuowo (2020) defined an efficient median-based linear regression estimator for population mean under simple random sampling scheme, assuming the population is homogeneous and skewed. Their estimator is expressed as a function of both mean per unit estimator of the study variable, population and sample medians of both study and auxiliary variables respectively. They discussed the properties of the estimator and justified the efficiency gain using both empirical and simulation studies. It is important to note that the difference between the estimator by Abdullahi and Ugwuowo (2020) and the estimator we proposed in this study is that the former is a regression estimator, which assumes that the regression line between the two variable passes through the origin, while the latter assumes that the regression line between the study variable and auxiliary variable does not pass through the origin, and the correlation between the two variables is positive. The strength of the positive correlation between the two variables determines the efficiency of auxiliary variable based estimators.

The search for efficient estimator of finite population mean has been a critical problem to sample survey research community. This study is motivated by the fact that

the conducted literature review showed that no research has developed such an average ratio estimator of the population mean that would utilize both the population and the sample medians of study variable, as well as the Srivastava (1967) estimator at a time.

2. Preliminaries

We assume that the population is finite of size N and a sample of size n is to be selected using simple random sampling scheme. Each unit of the population is identifiable by means of assigning the number to the population units from 1 to N, the numbers assigned are of nominal scale. We start by discussing some existing estimators to be considered in this study.

The existing estimators and their corresponding properties are presented in Table 1.

Table 1: Estimators

No.	Estimators	Bias	MSE
1.	$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$ Mean per unit	0	$\frac{(1-f)}{n}\bar{Y}^2\{C_Y^2\}$
2.	$\bar{y}_r = \bar{y}\left(\frac{\bar{X}}{\bar{x}}\right)$ Cochran (1940)	$\frac{(1-f)}{n}\bar{Y}\{C_X^2 - \rho C_Y C_X\}$	$\frac{(1-f)}{n}\bar{Y}^2\{C_Y^2 + C_X^2 - 2\rho C_Y C_X\}$
3.	$\bar{y}_{L-R} = \bar{y} + \beta(\bar{X} - \bar{x})$ Hansen, Hurwitz, and Madow (1953)	0	$\frac{(1-f)}{n}\bar{Y}^2 C_Y^2\{1-\rho^2\}$
4.	$\bar{y}_{Srivastava} = \bar{y}\left(\frac{\bar{x}}{\bar{X}}\right)^p$ Srivastava (1967)	$\left(\frac{1-f}{n}\right)\bar{Y}\left[\frac{p(p-1)}{2}C_X^2 + p\rho C_Y C_X\right]$	$\left(\frac{1-f}{n}\right)\bar{Y}^2 C_Y^2\left(1-\rho^2\right)$
5.	$\bar{y}_{S-med-based} = \bar{y}\left(\frac{M}{m}\right)$ Subramani (2016)	$\bar{Y}\left\{C_m^{!2} - C_{ym}^{!} - \frac{Bias(m)}{M}\right\}$	$V(\bar{y}) + R^{!2}V(m) - 2R^{!}c\,\mathrm{ov}(\bar{y}, m)$

Where

Notation

X and Y : auxiliary and study variables,

ρ : correlation coefficient between X and Y,

C_Y and C_X: population coefficient of variation of Y and X respectively,

$\bar{Y}$: population mean of X,

$f = n/N$: sampling fraction,

p : any chosen constant, which is defined in Srivastava (1967)

$R^{!} = \frac{\bar{Y}}{M}$, $\quad R = \frac{\bar{Y}}{\bar{X}}$

2.1. Description of the proposed estimator

We defined power ratio cum median-based ratio estimator for population mean under simple random sampling scheme as

$$\bar{y}_{\Pr opose} = \frac{\bar{y}}{2}\left\{\left(\frac{M}{m}\right)+\left(\frac{\bar{X}}{\bar{x}}\right)^{\theta}\right\} \tag{2.1}$$

Where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are mean per unit estimators of *Y* and *X*, $\bar{X}$ is the population mean *X*, M and m are the population and sample medians of *Y*, 'θ' is a real constant to be determined such that the mean squared error of $\bar{y}_{\Pr opose}$ is minimum.

2.2. Properties of the proposed estimator (bias and Mean Squared Error)

The bias and mean squared error (MSE) of the proposed estimator $\bar{y}_{\Pr opose}$ in (2.1) are presented as

Let $J_{\bar{y}} = \frac{(\bar{y}-\bar{Y})}{\bar{Y}}$, $J_{\bar{x}} = \frac{(\bar{x}-\bar{X})}{\bar{X}}$ and $J_m = \frac{(m-M)}{M}$

Such that $E(J_{\bar{y}}) = E(J_{\bar{x}}) = 0$ and $E(J_m) = \frac{bias(m)}{M}$

Where

$E(J_{\bar{y}}^2) = \frac{Var(\bar{y})}{\bar{Y}^2} = \left(\frac{1-f}{n}\right)C_Y^2 = C_y^{!2}$, $\quad E(J_{\bar{x}}^2) = \frac{Var(\bar{x})}{\bar{X}^2} = \left(\frac{1-f}{n}\right)C_X^2 = C_x^{!2}$

$E(J_m^2) = \frac{Var(m)}{M^2} = C_m^{!2}$, $\quad E(J_{\bar{x}}J_{\bar{y}}) = \frac{Cov(\bar{x},\bar{y})}{\bar{X}\bar{Y}} = \left(\frac{1-f}{n}\right)\rho_{XY}C_X C_Y = C_{xy}^{!}$,

$E(J_{\bar{y}}J_m) = \frac{Cov(\bar{y},m)}{\bar{Y}M} = C_{ym}^{!}$, $\quad E(J_m J_{\bar{x}}) = \frac{Cov(\bar{x},m)}{\bar{X}M} = C_{xm}^{!}$

And expanding (2.1) in terms of $J_{\bar{x}}$'s, we have

$$\bar{y}_{\Pr opose} = \frac{\bar{Y}}{2}(1+J_{\bar{y}})\left\{(1+J_m)^{-1}+(1+J_{\bar{x}})^{-\theta}\right\} \tag{2.2}$$

Note: $(1+J_m)^{-1} = 1-J_m+J_m^2$,

$$(1+J_{\bar{x}})^{-\theta} = 1-\theta J_{\bar{x}}+\theta(\theta+1)\frac{J_{\bar{x}}^2}{2}$$

We assume that $|J_m|<1$, $|J_{\bar{x}}|<1$ so that the expression, $(1+J_m)^{-1}$, $(1+J_{\bar{x}})^{-\alpha}$ can be expanded to a convergent infinite series using binomial theorem.

$$\bar{y}_{\Pr opose} = \bar{Y}(1+J_{\bar{y}})\left\{\left(1-J_m+J_m^2\right)+\left(1-\theta J_{\bar{x}}+\theta(\theta+1)\frac{J_{\bar{x}}^2}{2}\right)\right\} \tag{2.3}$$

We also assume that the contribution of terms involving powers in $J_m, J_{\bar{y}}, J_{\bar{x}}$ higher than the second is negligible, being of order $1/n^{v}$, where $v>1$. Thus, from the above expression we write to the first order of approximation.

$$= \bar{Y}\left\{\begin{array}{l}\left(1-J_m+J_m{}^2+J_{\bar{y}}-J_mJ_{\bar{y}}+J_m{}^2J_{\bar{y}}\right)\\ +\left(1-\theta J_{\bar{x}}+\theta(\theta+1)\dfrac{J_{\bar{x}}{}^2}{2}+J_{\bar{y}}-\theta J_{\bar{x}}J_{\bar{y}}+\theta(\theta+1)\dfrac{J_{\bar{x}}{}^2J_{\bar{y}}}{2}\right)\end{array}\right\} \tag{2.4}$$

$$\bar{y}_{\Pr opose}-\bar{Y}=\frac{\bar{Y}}{2}\left\{\begin{array}{l}\left(-J_m+J_m{}^2+J_{\bar{y}}-J_mJ_{\bar{y}}+J_m{}^2J_{\bar{y}}\right)\\ +\left(-\theta J_{\bar{x}}+\theta(\theta+1)\dfrac{J_{\bar{x}}{}^2}{2}+J_{\bar{y}}-\theta J_{\bar{x}}J_{\bar{y}}+\theta(\theta+1)\dfrac{J_{\bar{x}}{}^2J_{\bar{y}}}{2}\right)\end{array}\right\} \tag{2.5}$$

Taking the expectation of both sides of (2.5), we obtained the bias of $\left(\bar{y}_{\Pr opose}\right)$ to the first degree of approximation as

$$bias(\bar{y}_{\Pr opose})=\frac{\bar{Y}}{2}\left\{\left(-C_m^{\prime}+C_m^{\prime 2}-C_{my}^{\prime}\right)+\left(\theta(\theta+1)\frac{C_x^{\prime 2}}{2}-\theta C_{xy}^{\prime}\right)\right\} \tag{2.6}$$

Squaring both sides of the equation (2.5) and neglecting the terms of $J's$ having power greater than two we have

$$(\bar{y}_{\Pr opose}-\bar{Y})^2=\bar{Y}^2\left\{J_{\bar{y}}{}^2-J_{\bar{y}}J_m-\theta J_{\bar{x}}J_{\bar{y}}+\frac{\theta J_{\bar{x}}J_m}{2}+\frac{J_m{}^2}{4}+\frac{\theta^2J_{\bar{x}}{}^2}{4}\right\} \tag{2.7}$$

Taking the expectation of both sides of (2.7), we get the MSE of $\bar{y}_{\Pr opose}$ as

$$MSE(\bar{y}_{\Pr opose})=\bar{Y}^2\left\{\frac{Var(\bar{y})}{\bar{Y}}-\frac{Cov(m,\bar{y})}{M\bar{Y}}-\theta\frac{Cov(\bar{x},\bar{y})}{\bar{X}\bar{Y}}+\frac{\theta Cov(\bar{x},M)}{2\bar{X}M}+\frac{Var(m)}{4M}+\frac{\theta^2Var(\bar{x})}{4\bar{X}^2}\right\} \tag{2.8}$$

$$MSE(\bar{y}_{\Pr opose})=\bar{Y}^2\left\{C_y^{\prime 2}-C_{my}^{\prime}-\theta C_{xy}^{\prime}+\frac{\theta C_{xm}^{\prime}}{2}+\frac{C_m^{\prime 2}}{4}+\frac{\theta^2C_x^{\prime 2}}{4}\right\} \tag{2.9}$$

The minimum $MSE(\bar{y}_{\Pr opose})$ is obtained for the optimal value of θ which is

$$\frac{\partial MSE(\bar{y}_{\Pr opose})}{\partial\theta}=\left\{\frac{C_{xm}^{\prime}}{2}-C_{xy}^{\prime}+\frac{\theta C_{xx}^{\prime}}{2}\right\}=0 \tag{2.10}$$

$$\Rightarrow\theta=\frac{2C_{xy}^{\prime}-C_{xm}^{\prime}}{C_{xx}^{\prime}} \tag{2.11}$$

Therefore, the minimum MSE of the proposed estimator is obtained by substituting (2.11) into (2.9)

$$MSE(\overline{y}_{\Pr opose})_{\min} = \overline{Y}^2 \left\{ C_{yy}^{!} - C_{my}^{!} + \frac{C_{mm}^{!}}{4} + \frac{C_{xm}^{!} C_{xy}^{!}}{C_{xx}^{!}} - \frac{C_{xy}^{!\,2}}{C_{xx}^{!}} - \frac{\left(C_{xm}^{!}\right)^2}{4C_{xx}^{!}} \right\} \tag{2.12}$$

3. Efficiency comparison of proposed estimator with some selected estimators

3.1. The mean per unit unbiased estimator

Mean per unit estimator in SRSWOR is less efficient than the proposed estimator if $MSE(\overline{y}_{\Pr opose}) < MSE\left(\overline{y}\right)$, i.e.

$$\Rightarrow C_{my}^{!} + \theta C_{xy}^{!} \leq \frac{\theta C_{xm}^{!}}{2} + \frac{C_{mm}^{!}}{4} + \frac{\theta^2 C_{xx}^{!}}{4}$$

3.2. Cochran (1940) traditional ratio estimator

Traditional ratio estimator is less efficient than the proposed estimator if $MSE(\overline{y}_{\Pr opose}) < MSE\left(\overline{y}_r\right)$, i.e.

$$\frac{\theta R R^{'} Cov\left(m, \overline{x}\right)}{2} + \frac{R^{'2} Var(m)}{4} - R^{'} Cov\left(\overline{y}, m\right) \leq -2RCov\left(\overline{y}, \overline{x}\right) + R^2 Var\left(\overline{x}\right) + \theta RCov\left(\overline{y}, \overline{x}\right) - \frac{\theta^2 R^2 Var(\overline{x})}{4}$$

3.3. Hansen, Hurwitz and Madow (1953) linear regression estimator

Linear regression estimator is less efficient than the proposed estimator if $MSE(\overline{y}_{\Pr opose}) \leq MSE\left(\overline{y}_{L-R}\right)$, i.e.

$$\left\{ -R^{'} Cov\left(\overline{y}, m\right) - \theta RCov\left(\overline{y}, \overline{x}\right) + \frac{\theta R R^{'} Cov\left(m, \overline{x}\right)}{2} + \frac{Var\left(m\right) R^{'2}}{4} + \frac{\theta^2 Var\left(\overline{x}\right) R^2}{4} \right\} \geq \left(\frac{\left(Cov(\overline{y}, \overline{x})\right)^2}{Var\left(\overline{x}\right)} \right)$$

3.4. Subramani (2016) median ratio estimator

Median-based ratio estimator is less efficient than the proposed estimator if $MSE(\overline{y}_{\Pr opose}) < MSE\left(\overline{y}_{S-median-based}\right)$, i.e.

$$\frac{3C_{mm}^{!}}{4} - C_{ym}^{!} \geq -\theta C_{xy}^{!} + \frac{\theta C_{xm}^{!}}{2} + \frac{\theta^2 C_{xx}^{!}}{4}$$

4. Numerical Comparison

The merit of the proposed estimator $\overline{y}_{\Pr opose}$ over $\overline{y}$, $\overline{y}_r$ and $\overline{y}_{L-R}$ estimators is presented in this section.

Dataset: The populations considered in this study is a real-life dataset taken from Singh and Chaudhary (1986). The dataset is also used by Srija and Subramani (2018). The area under Wheat cultivation in 1971 is the auxiliary variable while area under

Wheat cultivation in 1974 is the study variable. Table 2 is the summary of the real-life dataset.

Table 2: Summary of the dataset

Parameter		Parameter		Parameter		Parameter	
N	34	R	4.0999	$Cov(\bar{y},m)$	90236.294	$C_{ym}^{'}$	0.1372841
n	3	$R^{!}$	1.1158	$Cov(\bar{y},\bar{x})$	15061.401	$C_{yx}^{'}$	0.0841917
$\bar{Y}$	856.42	$V(\bar{y})$	163356.41	$Cov(\bar{x},m)$	18342.18	$C_{xm}^{!}$	0.1144118
$\bar{X}$	208.88	$V(\bar{x})$	6884.45	$C_{yy}^{'}$	0.222726	$bias(m)$	-19.77774
$\bar{M}$	747.72	$V(m)$	101518.77	$C_{xx}^{!}$	0.1577848	$bias(m)/M$	-0.02576904
M	767.50	ρ	0.4491	$C_{mm}^{'}$	0.1723414		

Percentage Relative Efficiency (PRE)

The Percentage relative efficiency (PRE) of different estimators $T_{\cdot}$ in respect to $\bar{y}_{\Pr opose}$ is defined as $PRE\left(\bar{y}_{\Pr opose},T_{\cdot}\right)=\dfrac{MSE(T_{\cdot})}{MSE\left(\bar{y}_{\Pr opose}\right)}\text{X}100$

Table 3 gives the MSE and PRE of the existing and proposed estimators with respect to mean per unit ($\bar{y}$), usual ratio ($\bar{y}_r$) and linear regression ($\bar{y}_{L-R}$) estimators respectively.

Table 3: MSE/Variance of some selected existing estimators and that of proposed estimator

MSE/Variance of some selected existing estimators and that of proposed estimator		PRE with respect to ($\bar{y}$)	PRE with respect to ($\bar{y}_r$)	PRE with to respect ($\bar{y}_{L-R}$)
Estimator				
$\bar{y}$	163356.4	100	<100	<100
$\bar{y}_r$	155583	104.996	100	<100
$\bar{y}_{L-R}$	130408.93	<100	<100	100
$\bar{y}_{\Pr opose}$	90882.08	179.7455	171.19	143.49

The result from Table 3 reveals that the proposed estimator $\bar{y}_{\Pr opose}$ has the minimum mean square error compared to some existing estimators and it also shows significant efficiency gain in respect of percentage relative efficiency.

5. Simulation Study

Additionally, a simulation study is conducted to evaluate the effectiveness of the proposed estimator. The variables are created in accordance with Singh and Horn's (1998) definitions, which were also incorporated into Lamichhane, Singh, and

Diawara's work (2017). Firstly, Gamma distribution and secondly Weibull distribution, the population size is N =3003 with 7 varying sample sizes, while the number of trials is 500. Tables 4 and 5 present the PRE of the proposed estimator of population mean with respectto mean per unit estimator, traditional ratio and linear regression estimator for both Weibull and Gamma distributions respectively.

Table 4: PRE of the proposed and some existing estimators (Weibull Distribution)

PRE With respect to	Sample size (n)	Rho=0.30	Rho=0.50	Rho=0.60	Rho=0.75
$\bar{y}$	12	252.8685	249.8534	249.2091	250.5923
	33	249.5238	241.4079	250.3717	262.6104
	45	212.0155	231.8280	229.4620	237.4103
	79	238.2948	232.6957	238.2306	249.4543
	123	215.7753	195.5499	190.7657	199.4062
	202	238.2451	215.6402	209.2263	215.6172
	243	246.2527	232.7251	235.7246	237.7677
$\bar{y}_{L-R}$	12	240.7894	185.9880	153.2106	99.68960
	33	223.0672	176.8050	157.1801	114.94581
	45	194.9318	171.5238	142.4898	97.75636
	79	233.3049	194.2552	172.0632	124.68826
	123	195.1312	140.9331	114.3663	77.80536
	202	211.8818	155.1586	127.2922	88.57750
	243	215.1578	164.3203	141.0363	96.80333

Table 5: PRE of the proposed and some existing estimators (Gamma distribution)

PRE With respect to	Sample size (n)	Rho=0.30	Rho=0.45	Rho=0.75	Rho=0.9
$\bar{y}$	23	670.9549	548.6637	290.2100	114.34680
	43	542.1858	479.4620	228.1377	93.00426
	53	593.0156	495.9020	250.8398	104.42931
	93	500.1355	399.0709	190.4233	81.13561
	103	508.2268	395.2287	207.1991	85.57811
	203	532.8165	403.5352	215.9648	89.56156
	243	551.3976	443.8553	241.6027	98.21637
$\bar{y}_{L-R}$	(n)	Rho=0.30	Rho=0.45	Rho=0.60	Rho=0.75
	23	281.3494	230.0694	183.2459	121.69287
	43	239.7914	212.0506	153.7183	100.89799
	53	239.3716	200.1716	154.8750	101.25188
	93	225.2151	179.7048	138.7745	85.74912
	103	210.9156	164.0211	126.5875	85.98823
	203	236.4502	179.0785	147.3164	95.83958
	234	210.6481	169.5642	139.2650	92.29847

From the simulation study results in Tables 4 and 5, the proposed estimator $\bar{y}_{\text{Pr}opose}$ shows significant efficiency gain in respect to $\bar{y}$ and $\bar{y}_{L-R}$ compared to some existing estimators.

6. Discussion and Conclusion

In this study, an efficient median-based ratio estimator of population mean with known population median was proposed and named Power Ratio Cum Median-Based Ratio. The bias and MSE of the proposed estimator are derived. Comparing the proposed estimator with various other existing estimators in the literature, we showed that it meets the efficiency criteria. Results from both the real-life dataset and the simulation study show efficiency gain for the proposed estimator that incorporates median of study variable, while for the other estimators the result shows efficiency loss. With the significant performance of proposed estimator, which is function of both medians and mean per unit estimator of the study variable and ratio of population to sample means of auxiliary variable, it is revealed that there is a hidden significant relationship that exists between mean and median of the same variable. Hence, the proposed estimator is recommended for the use in practice when the efficiency conditions are satisfied.

References

Abdullahi, U. K., Ugwuowo, F. I., (2022). On efficient median-based linear regression estimator for population mean, *Communications in Statistics – Theory and Methods*, 51(15), pp. 5012–5024.

Cochran, W. G., (1947). Sampling Techniques, 3rd edition, New York: John Wiley and Sons.

Gupta, S., Shabbir, J., (2008). On improvement in estimating the population mean in simple random sampling. *Journal of Applied Statistics,* 35(5), pp. 559–566.

Lamichhane, R., Singh, S. and Diawara, N., (2017). Improved estimation of population mean using known median of auxiliary variable. *Communications in Statistics-Simulation and Computation,* 46(4), pp. 2821–2828.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G., (1953). *Sample survey methods and theory*. New York: John wiley.

Neyman, J., (1938). Contribution to the theory of sampling human populations. *Journal of American Statistical Association,* 33, pp. 101–116.

Olkin, I., (1958). Multivariate ratio estimation for finite populations. *Biometrika*, 45, pp. 154–165.

Singh, D., Chaudhary, F. S., (1986). *Theory and Analysis of Sample Survey Designs*, New Age International Publisher.

Singh, S., Horn S., (1998). An alternative estimator for multi-character surveys, *Metrika*, 48, pp. 99–107.

Srija, R., Subramani, J., (2018). Ratio cum median based modified ratio estimators with known first and third quartiles. *International Journal of Communication and Media Sciences,* 5(1), pp. 1–5.

Subramani, J., (2016). A New Median Based Ratio Estimator for Estimation of the Finite Population Mean. *Statistics in Transition*, 17 (4), pp. 591–604.

Srivstava, S. K., (1967). An estimator using auxiliary information in sample surveys. *Calcutta Statistical Association Bulletin*, 16, pp. 121–132.

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 45–61, https://doi.org/10.59170/stattrans-2023-063
Received – 25.06.2022; accepted – 04.05.2023

Ratio estimation of two population means in two-phase stratified random sampling under a scrambled response situation

Pitambar Das[1], **Garib Nath Singh**[2], **Arnab Bandyopadhyay**[3]

Abstract

In this paper, we have described the development of an effective two-phase stratified random sampling estimation procedure in a scrambled response situation. Two different exponential, regression-type estimators were formed separately for different structures of two-phase stratified sampling schemes. We have studied the properties of the suggested strategy. The performance of the proposed strategy has been demonstrated through numerical evidence based on a data set of a natural population and a population generated through simulation studies. Taking into consideration the encouraging findings, suitable recommendations for survey statisticians are prepared for the application of the proposed strategy in real-life conditions.

Key words: stratified random sampling, scrambled response, auxiliary variable, mean square error, simulation study.

2010 AMS Subject Classifications: 62D05

1. Introduction

In sample surveys, the population may be formed of heterogeneous units. For example, in socio-economic surveys, people may live in hospital, hostel, residential houses and jail, etc. The whole population is divided into certain internally homogeneous and externally heterogeneous groups, called strata, and then independent samples of different sizes are selected from each stratum. Stratified sampling is one of the most widely used sampling techniques as it increases the precision of the estimate of the survey variable when units of the population are from

[1] Department of Mathematics, Netaji Nagar Day College, Kolkata- 700092, India. E-mail: pitambardas.in@gmail.com.

[2] Department of Mathematics & Computing, Indian Institute of Technology (Indian School of Mines), Dhanbad-826004, India. E-mail: gnsingh@iitism.ac.in.

[3] Department of Mathematics, Asansol Engineering College, Asansol–713305, India. Email: arnabbandyopadhyay4@gmail.com. ORCID: https://orcid.org/0000-0002-0769-7491.

different portions of population. Many authors have discussed different types of estimators using the auxiliary information in stratified random sampling, e.g. Singh and Sukhatme (1973), Kadilar and Cingi (2000, 2003), Shabbir and Gupta (2005), Singh and Vishwakarma (2005), Koyuncu and Kadilar (2008, 2009), Singh *et al.* (2009), etc. It is noted that most of recently developed the estimation of ratio of population mean in simple random sampling only, limited attempts have been taken to estimate the ratio of population mean under stratified random sampling scheme.

In socio -economic surveys, an estimate of the population ratio of two characters for the stratified random sampling may be of considerable interest. For example, the ratio of per month total income and total expenditure of people of different classes in a locality.

In practice, it may also be observed that characteristics under study on sensitive data that someone wants to response but prefers to hide the true value for avoiding possible social stigma and harassment, etc. As many people prefers to hide their exact responses, available sample of returns is camouflaged. Basic idea behind scrambling the survey data is that the response given by the respondent against any query related to some sensitive character is camouflaged by adding or multiplying the data with any random number generated by the respondent by using any random device which will not be known by the surveyor although surveyor may know the mean and variance. This procedure is applied in such a way that the respondents feels that their privacy is protected. Commanding work was done by Greenberg *et al.* (1971), Pollok and Bek (1976), Eichhorn and Hayre (1983), Giancarlo and Pier (2010), Dianna and Perri (2010). In socio -economic surveys, people may live in a different economic zone. If we know the ratio of income and expenditure of people of different classes in a locality, then an investment may be planned suitably. In this case income and expenditure are highly sensitive variables in which scramble response situation may be found. We estimate the ratio of domestic violence reported in a society and the number of premarital abortions that took place in a society, then we reduce the rate of abortions.

It may be noted that no significant attempt has been taken to estimate the ratio of population parameters through two-phase stratified random sampling in the presence of scrambled response situation. Influenced and convinced with the points discussed above, we have recommended a general procedure to estimate population mean in stratified random sampling in presence of the above -mentioned situation. The findings are demonstrated through numerical illustrations carried over the data set of natural population and population generated through simulation studies using different types of correlations. Suitable recommendations are made to the survey statisticians for possible applications.

2. Sample structures and notations

Consider a finite population $U = (U_1, U_2, ..., U_N)$ of size N units and divided into L strata, each of size N_h units (h = 1, 2, ..., L) such that $\sum_{h=1}^{L} N_h = N$. Let y be the study variable and (x, z) be the auxiliary variables respectively taking values y_{hi} and (x_{hi}, z_{hi}) respectively, for the i^{th} unit (i = 1, 2, ..., N_h) of the h^{th} stratum (h = 1, 2, ..., L).

The first phase sample S_{nh} of size n_h is drawn at random without replacement from each stratum containing N_h units (h = 1, 2, …, L). Again the second phase sample S_{mh} of size m_h units is drawn using SRSWOR scheme from each first phase sample of size n_h (h = 1, 2, ..., L).

We consider the following notations for their further use:

$\bar{Y}_h$: Population mean of the study variable y_h (h = 1, 2, ... , L).

$\bar{X}_h, \bar{Z}_h$: Population means of the auxiliary variables x_h and z_h respectively (h = 1, 2, ... , L).

$R_h = \frac{\bar{Y}_h}{\bar{X}_h}$: Ratio of population means of the variables y_h and x_h of the h^{th} stratum (h = 1, 2, ... , L).

$R = \sum_{h=1}^{L} W_h R_h$: Total ratio of the population means.

$R_{nh} = \frac{\bar{y}_{nh}}{\bar{x}_{nh}}$: Ratio of sample means of the variables y_{nh} and x_{nh} based on the sample of size n_h (h = 1, 2, ... , L).

$R_{mh} = \frac{\bar{y}_{mh}}{\bar{x}_{mh}}$: Ratio of sample means of the variables y_{mh} and x_{mh} based on the sample of size m_h (h = 1, 2, ... , L).

$\bar{Y}_h = \sum_{i=1}^{N_h} \frac{y_{hi}}{N_h}$, $\bar{X}_h = \sum_{i=1}^{N_h} \frac{x_{hi}}{N_h}$, $\bar{Z}_h = \sum_{i=1}^{N_h} \frac{z_{hi}}{N_h}$: Population means of the respective variables on the stratum h (h = 1, 2, ..., L).

$\bar{Y} = \sum_{h=1}^{L} \bar{Y}_h W_h$, $\bar{X} = \sum_{h=1}^{L} \bar{X}_h W_h$, $\bar{Z} = \sum_{h=1}^{L} \bar{Z}_h W_h$: Total population means of the respective variables,

where $W_h = \frac{N_h}{N}$:Weight of the h^{th} stratum (h = 1, 2, ... , L).

$\bar{z}_{nh} = \frac{1}{n_h} \sum_{i=1}^{n_h} z_{hi}$: Mean of the variable z based on the sample S_{nh} of size n_h (h = 1, 2, ... , L).

$\bar{z}_{mh}$: Mean of the variable z based on the sample S_{mh} of size m_h (h = 1, 2, ... , L).

$\bar{x}_{nh} = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}$: Mean of the variable x based on the sample S_{nh} of size n_h (h = 1, 2, ... , L).

$\bar{x}_{mh}$: Mean of the variable x based on the sample S_{mh} of size m_h (h = 1, 2, ... , L).

$\bar{y}_{nh} = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$: Mean of the variable y based on the sample S_{nh} of size n_h (h = 1, 2, ... , L).

$\bar{y}_{mh}$: Mean of the variable y based on the sample S_{mh} of size m_h (h = 1, 2, ... , L).

$S_{yh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y}_h)^2$: Population mean square of the variable y based on the stratum h (h = 1, 2, ..., L).

S_{xh}^2, S_{zh}^2: Population mean squares of the variables based on the stratum h (h = 1, 2, ... , L).

$C_{yh} = \frac{S_{yh}}{\bar{Y}_h}$, $C_{xh} = \frac{S_{xh}}{\bar{X}_h}$, $C_{zh} = \frac{S_{zh}}{\bar{Z}_h}$: Coefficient of variations for the variables y, x and z for the h^{th} stratum (h = 1, 2, ..., L).

ρ_{yx_h}, ρ_{yz_h}, ρ_{xz_h}: Correlation coefficients between (y, x), (y, z), and (x, z) respectively in the h^{th} stratum (h = 1, 2, ... , L).

3. Scrambling technique

Eichhorn and Hayre (1983) studied a multiplicative randomized response method for obtaining responses to sensitive questions when the answers are quantitative. The method involves the respondent multiplying his sensitive answer by a random number from a known distribution, and giving the product to the interviewer, who does not know the value of the random number and thus receives a scrambled response. Some particular distributions for the random scrambling number are proposed and studied, and ways of generating the scrambling numbers are discussed.

We have denoted the scrambled variables as $y^* = yT$, $x^* = xT$, where T is the random number multiplied by the study variables y and x to yield y^* and x^*. The device is so selected that the mean of T, i.e. E (T) = 1 to minimize the effect of scrambling. Also, it is important to assume that any two random numbers created are mutually independent as well as the random number T and sensitive variables y and x are also mutually independent.

We use the following notation on scramble variables (i.e. y^* and x^*)

$R_h^* = \frac{\bar{Y}_h^*}{\bar{X}_h^*}$: Ratio of population means of the scramble variables y_h^* and x_h^* of the h^{th} stratum (h = 1, 2, ... , L).

$R^* = \sum_{h=1}^{L} W_h R_h^*$: Total ratio of the population means of the scramble variables.

$R_{nh}^* = \frac{\bar{y}_{nh}^*}{\bar{x}_{nh}^*}$: Ratio of sample means of the scramble variables y_{nh}^* and x_{nh}^* based on the sample of size n_h (h = 1, 2, ... , L).

$R_{mh}^* = \frac{\bar{y}_{mh}^*}{\bar{x}_{mh}^*}$: Ratio of sample means of the scramble variables y_{mh}^* and x_{mh}^* based on the sample of size m_h (h = 1, 2, ... , L).

$\bar{Y}_h^* = \sum_{i=1}^{N_h} \frac{y_{hi}^*}{N_h}$, $\bar{X}_h^* = \sum_{i=1}^{N_h} \frac{x_{hi}^*}{N_h}$: Population means of the respective scramble variables on the stratum h (h = 1, 2, ..., L).

$\bar{Y}^* = \sum_{h=1}^{L} \bar{Y}_h^* W_h$, $\bar{X}^* = \sum_{h=1}^{L} \bar{X}_h^* W_h$: Total population means of the respective scramble variables,

where $W_h = \frac{N_h}{N}$: Weight of the h^{th} stratum (h = 1, 2, ... , L).

$\bar{x}^*_{nh} = \sum_{i \in S_{nh}} \frac{x^*_{hi}}{n_h}$: Mean of the scramble variable x^* based on the sample S_{nh} of size n_h (h = 1, 2, ... , L).

$\bar{x}^*_{mh}$: Mean of the scramble variable x^* based on the sample S_{mh} of size m_h (h = 1, 2, ... , L).

$\bar{y}^*_{nh} = \sum_{i \in S_{nh}} \frac{y^*_{hi}}{n_h}$: Mean of the scramble variable y^* based on the sample S_{nh} of size n_h (h = 1, 2, ... , L).

$\bar{y}^*_{mh}$: Mean of the scramble variabley^* based on the sample S_{mh} of size m_h (h = 1, 2, ... , L).

$S^{2^*}_{yh} = \frac{1}{N_h\text{-}1}\sum_{i=1}^{N_h}(y^*_{hi}\text{-}\bar{Y}^*_h)^2$: Population mean square of the scramble variable y^* based on the stratum h (h = 1, 2, ..., L).

$S^{2^*}_{xh}$: Population mean square of the scramble variable based on the stratum h (h = 1, 2, ..., L).

$C^*_{yh} = \frac{S^*_{yh}}{\bar{Y}^*_h}$, $C^*_{xh} = \frac{S^*_{xh}}{\bar{X}^*_h}$: Coefficient of variations for the scramble variablesy^*, x^* for the h^{th} stratum (h = 1, 2, ..., L).

$\rho^*_{yx_h}$, $\rho^*_{yz_h}$, $\rho^*_{xz_h}$: Correlation coefficients between(y^*, x^*),(y^*, z), and (x^*, z)respectively in the h^{th} stratum (h = 1, 2, ..., L).

4. Formulations of the estimators

To estimate the ratio of population means $R^*\left(=\frac{\bar{Y}}{\bar{X}}\right)$on the two-phases stratified random sampling scheme. It is also noted that we are getting scrambled response only for the variables y and x in terms of the variables $y^*$and $x^*$from the respective samples.

Motivated with the previous work, we have proposed the following chain type exponential and regression type estimators for ratio of population means R^*on the two-phase stratified random sampling scheme as

$$\xi_1 = \sum_{h=1}^{L} W_h R^*_{mh} \exp\left(\frac{R'^*_{nh}\text{-}R'^*_{mh}}{R'^*_{nh}+R'^*_{mh}}\right) \quad (1)$$

$$\xi_2 = \sum_{h=1}^{L} W_h\left[R^*_{mh} + b_{2h}\left(R'^*_{nh}\text{-}R'^*_{mh}\right)\right] \quad (2)$$

where

$$R'^*_{nh} = R^*_{nh} + b_{1h}(\bar{Z}_h\text{-}\bar{z}_{nh}) \text{ and } R'^*_{mh} = R^*_{mh} + b'_{1h}(\bar{Z}_h\text{-}\bar{z}_{mh}).$$

We first estimate the ratio of the population means R^*_h on the stratum h (h = 1, 2, ..., L) using the proposed estimators and then estimate the total ratio of the population means R^*.

5.1. Mean square errors of the proposed estimators

The mean square errors (MSEs) of the proposed estimators ξ_1 and ξ_2 up to the first order of approximation are derived under large sample approximation using the following transformations:

$\bar{y}^*_{mh} = \bar{Y}_h(1+e_0)$, $\bar{x}^*_{mh} = \bar{X}_h(1+e_1)$,

$\bar{y}^*_{\mathrm{nh}} = \bar{\mathrm{Y}}_h(1+\mathrm{e}_2)$, $\bar{\mathrm{x}}^*_{\mathrm{nh}} = \bar{\mathrm{X}}_h(1+\mathrm{e}_3)$,

$\bar{z}_{\mathrm{mh}} = \bar{\mathrm{Z}}_h(1+\mathrm{e}_4)$, $\bar{\mathrm{z}}_{\mathrm{nh}} = \bar{\mathrm{Z}}_h(1+\mathrm{e}_5)$,

such that $\mathrm{E}(e_i) = 0$ and $|e_i| < 1$, for all i = 0, 1, ..., 5.

Under the above transformations the estimators take the following forms:

$$\xi_1 = \sum_{\mathrm{h}=1}^{L} W_h R_h^* (1+\mathrm{e}_0)(1+\mathrm{e}_1)^{-1} \exp\left[\begin{array}{l}\left\{(1+\mathrm{e}_2)(1+\mathrm{e}_3)^{-1} - (1+\mathrm{e}_0)(1+\mathrm{e}_1)^{-1} + \frac{b_{1\mathrm{h}}\bar{Z}_h}{R_h^*}(e_4 - \mathrm{e}_5)\right\} \\ \times \frac{1}{2}\left\{1 + \frac{1}{2}\left\langle (e_2 - \mathrm{e}_3) + (e_0 - \mathrm{e}_1) - \frac{b_{1\mathrm{h}}\bar{Z}_h}{R_h^*}(e_4 + \mathrm{e}_5)\right\rangle\right\}^{-1}\end{array}\right] \tag{3}$$

$$\xi_2 = \sum_{\mathrm{h}=1}^{L} W_h R_h^* \left[(1+\mathrm{e}_0)(1+\mathrm{e}_1)^{-1} + \mathrm{b}_{2\mathrm{h}}\{(1+\mathrm{e}_2)(1+\mathrm{e}_3)^{-1} - (1+\mathrm{e}_0)(1+\mathrm{e}_1)^{-1}\} + \frac{b_{3\mathrm{h}}\bar{Z}_h}{R_h^*}(e_4 - \mathrm{e}_5)\right] \tag{4}$$

Again, we obtain the following expression for expectations:

$E(e_0^2) = \mathrm{f}_1 C_{y_h}^{*2}$, $\mathrm{E}(e_1^2) = \mathrm{f}_1 C_{x_h}^{*2}$, $\mathrm{E}(e_2^2) = \mathrm{f}_2 C_{y_h}^{*2}$, $\mathrm{E}(e_3^2) = \mathrm{f}_2 C_{x_h}^{*2}$,

$E(e_4^2) = \mathrm{f}_1 C_{z_h}^2$, $\mathrm{E}(e_5^2) = \mathrm{f}_2 C_{z_h}^2$,

$E(e_0 e_1) = \mathrm{f}_1 \rho^*_{\mathrm{yx}_h} C^*_{y_h} C^*_{x_h}$, $\mathrm{E}(e_1 e_2) = \mathrm{f}_2 \rho^*_{\mathrm{yx}_h} C^*_{y_h} C^*_{x_h}$, $\mathrm{E}(e_2 e_3) = \mathrm{f}_2 \rho^*_{\mathrm{yx}_h} C^*_{y_h} C^*_{x_h}$,

$E(e_3 e_4) = \mathrm{f}_2 \rho^*_{\mathrm{xz}_h} C^*_{x_h} C_{z_h}$, $\mathrm{E}(e_3 e_5) = \mathrm{f}_2 \rho^*_{\mathrm{xz}_h} C^*_{x_h} C_{z_h}$, $\mathrm{E}(e_0 e_4) = \mathrm{f}_1 \rho_{\mathrm{yz}_h} C^*_{y_h} C_{z_h}$,

$E(e_0 e_5) = \mathrm{f}_2 \rho^*_{\mathrm{yz}_h} C^*_{y_h} C_{z_h}$, $\mathrm{E}(e_4 e_5) = \mathrm{f}_2 C^2_{z_h}$, $\mathrm{E}(e_1 e_4) = \mathrm{f}_1 \rho^*_{\mathrm{xz}_h} C^*_{x_h} C_{z_h}$,

$E(e_0 e_4) = \mathrm{f}_1 \rho^*_{\mathrm{yz}_h} C^*_{y_h} C_{z_h}$, $E(e_0 e_2) = \mathrm{f}_2 C_{y_h}^{*2}$, $\mathrm{E}(e_0 e_3) = \mathrm{f}_2 \rho^*_{\mathrm{yx}_h} C^*_{y_h} C^*_{x_h}$,

$E(e_1 e_3) = \mathrm{f}_2 C_{x_h}^{*2}$, $\mathrm{E}(e_0 e_5) = \mathrm{f}_2 \rho^*_{\mathrm{yz}_h} C^*_{y_h} C^*_{z_h}$, $\mathrm{E}(e_1 e_5) = \mathrm{f}_2 \rho^*_{\mathrm{xz}_h} C^*_{x_h} C_{z_h}$,

$E(e_2 e_4) = \mathrm{f}_2 \rho^*_{\mathrm{yz}_h} C^*_{y_h} C_{z_h}$, $\mathrm{E}(e_2 e_5) = \mathrm{f}_2 \rho^*_{\mathrm{xz}_h} C^*_{x_h} C_{z_h}$,

where $\mathrm{f}_1 = \left(\frac{1}{m_h} - \frac{1}{N_h}\right)$ and $\mathrm{f}_2 = \left(\frac{1}{n_h} - \frac{1}{N_h}\right)$.

Taking expectations on both sides of equations (3) and (4), we obtained the mean square error of the estimators ξ_1 and ξ_2 up to the first order of approximations as

$$\begin{aligned} M(\xi_1) &= \mathrm{E}(\xi_1 - \mathrm{R}^*)^2 \\ &= \frac{1}{4}\sum_{\mathrm{h}=1}^{L} W_h^2 R_h^{*2} \left[2(f_1 + \mathrm{f}_2)\left(C_{x_h}^{*2} + \mathrm{C}_{y_h}^{*2}\right) - 2(f_1 + 2\mathrm{f}_2)\rho^*_{\mathrm{yx}_h} C^*_{y_h} C^*_{x_h} \right. \\ &\qquad \left. + \frac{b_{1\mathrm{h}}\bar{Z}_h}{R_h^*}(f_1 - \mathrm{f}_2)\rho^*_{\mathrm{yz}_h} C^*_{y_h} C_{z_h} + \frac{b_{1\mathrm{h}}^2 \bar{Z}_h^2}{R_h^{*2}}(f_1 - \mathrm{f}_2) C^2_{z_h}\right] \\ &= \frac{1}{4}\sum_{\mathrm{h}=1}^{L} W_h^2 R_h^{*2} \left[A_1 + \frac{b_{1\mathrm{h}}\bar{Z}_h}{R_h^*} B_1 + \frac{b_{1\mathrm{h}}^2 \bar{Z}_h^2}{R_h^{*2}} C_1\right] \end{aligned} \tag{5}$$

where $\mathrm{A}_1 = 2(f_1 + \mathrm{f}_2)\left(C_{x_h}^{*2} + \mathrm{C}_{y_h}^{*2}\right) - 2(f_1 + 2\mathrm{f}_2)\rho^*_{\mathrm{yx}_h} C^*_{y_h} C^*_{x_h}$, $\mathrm{B}_1 = (f_1 - \mathrm{f}_2)\rho^*_{\mathrm{yz}_h} C^*_{y_h} C_{z_h}$,

$C_1 = (f_1 - \mathrm{f}_2) C^2_{z_h}$.

$$M(\xi_2)= \mathrm{E}(\xi_2 - \mathrm{R}^*)^2$$

$$\mathrm{M}(\xi_2) = \sum_{h=1}^{L} W_h^2 R_h^{*2} \left[\begin{array}{l} f_1 \left(C_{y_h}^{*2} + C_{x_h}^{*2} - 2\rho_{yx_h}^* C_{y_h}^* C_{x_h}^* \right) + b_{2h}^2 \left\{ (f_1+f_2)\left(C_{y_h}^{*2} + C_{x_h}^{*2} \right) + 2(f_2-f_1)\rho_{yx_h}^* C_{y_h}^* C_{x_h}^* \right\} + \\ \dfrac{b_{3h}^2 \bar{Z}_h}{R_h^{*2}} (f_1-f_2) C_{z_h}^2 + 2b_{2h} \left\{ (f_2-f_1)\left(C_{y_h}^{*2} + C_{x_h}^{*2} - 2\rho_{yx_h}^* C_{y_h}^* C_{x_h}^* \right) \right\} + \dfrac{2b_{3h}\bar{Z}_h}{R_h^*} (f_1-f_2) \\ \left(\rho_{yz_h}^* C_{y_h}^* C_{z_h} - \rho_{xz_h}^* C_{x_h}^* C_{z_h} \right) + \dfrac{2b_{4h}\bar{Z}_h}{R_h^*} \left\{ (2f_2-f_1)\left(\rho_{yz_h}^* C_{y_h}^* C_{z_h} - \rho_{xz_h}^* C_{x_h}^* C_{z_h} \right) \right\} \end{array} \right]$$

$$= \sum_{h=1}^{L} W_h^2 R_h^{*2} \left[A_2 + b_{2h}^2 B_2 + \frac{b_{3h}^2 \bar{Z}_h^2}{R_h^{*2}} C_2 + b_{2h} D_2 + \frac{b_{3h}\bar{Z}_h}{R_h^*} E_2 + \frac{b_{4h}\bar{Z}_h}{R_h^*} F_2 \right] \tag{6}$$

where $A_2 = f_1 \left(C_{y_h}^{*2} + C_{x_h}^{*2} - 2\rho_{yx_h}^* C_{y_h}^* C_{x_h}^* \right)$, $B_2 = \left\{ (f_1+f_2)\left(C_{y_h}^{*2} + C_{x_h}^{*2} \right) + 2(f_2-f_1)\rho_{yx_h}^* C_{y_h}^* C_{x_h}^* \right\}$,

$$C_2 = (f_1-f_2) C_{z_h}^2, \; D_2 = 2\left\{ (f_2-f_1)\left(C_{y_h}^{*2} + C_{x_h}^{*2} - 2\rho_{yx_h}^* C_{y_h}^* C_{x_h}^* \right) \right\},$$

$$E_2 = \frac{2\bar{Z}_h}{R_h^*} (f_1-f_2)\left(\rho_{yz_h}^* C_{y_h}^* C_{z_h} - \rho_{xz_h}^* C_{x_h}^* C_{z_h} \right), \; F_2 = \frac{2\bar{Z}_h}{R_h^*} \left\{ (2f_2-f_1)\left(\rho_{yz_h}^* C_{y_h}^* C_{z_h} - \rho_{xz_h}^* C_{x_h}^* C_{z_h} \right) \right\}.$$

5.2. Minimum mean square errors of the proposed estimators

It may be observed from equation (5) and (6) that the expressions for $M(\xi_1)$ and $M(\xi_2)$ depend on the values of b_{1h}, b_{2h}, b_{3h} and b_{4h}(h =1, 2, . . ., L), which are real constants. Therefore, we need to find the optimum values of b_{1h}, b_{2h}, b_{3h} and b_{4h} which can minimize the MSE of the estimators ξ_1 and ξ_2 respectively.The optimum values of b_{1h}, b_{2h}, b_{3h} and b_{4h}(h = 1, 2, . . ., L) are found as

$$b_{1h} = -\frac{R_h^* B_1}{2C_1 \bar{Z}_h} \tag{7}$$

$$b_{2h} = -\frac{D_2}{2B_2} \tag{8}$$

$$b_{3h} = -\frac{E_2 R_h^*}{2C_2 \bar{Z}_h} \tag{9}$$

$$b_{4h} = \frac{D_2 E_2 R_h^*}{4B_2 C_2 \bar{Z}_h} \tag{10}$$

Thus, substituting the values of b_{1h}, b_{2h}, b_{3h} and b_{4h} from equations (7), (8), (9) and (10) to the equations (5) and (6),we have derived the optimum mean square errors of the proposed estimators as

$$\mathrm{M}(\xi 1)\mathrm{opt} = \frac{1}{4}\sum_{\mathrm{h}=1}^{L} W_h^2 R_h^{*2}\left[A_1 - \frac{B_1^2}{4\mathrm{C}_1}\right] \tag{11}$$

$$M(\xi_2)_{\mathrm{opt}} = \sum_{\mathrm{h}=1}^{L} W_h^2 R_h^{*2}\left[A_2 - \frac{D_2^2}{4\mathrm{B}_2} - \frac{E_2^2}{4\mathrm{C}_2} + \frac{D_2 E_2 F_2}{4\mathrm{B}_2 C_2}\right] \tag{12}$$

Remark: 5.2.1. It is to be noted that the optimality condition of the estimators in equation (11) and (12) the optimum values b_{1h}, $b_{2h,}$ b_{3h} and b_{4h} depend on unknown population parameters such as R_h^{*2}, $C_{x_h}^{2*}$, $C_{y_h}^{2*}$, $\rho_{xy_h}^{*}$ and $\rho_{yz_h}^{*}$ (h = 1, 2, ..., L).

Thus, to make the classes of estimators practicable, these unknown population parameters may be estimated with their respective sample estimates or from past data or guessed from experience gathered over time. Such problems are also considered by Reddy (1978), Tracy *et al.* (1996) and Singh *et al.* (2007).

6. Efficiency comparison

It is important to investigate the situation under which our proposed estimators are more efficient than the conventional ones. Since no estimator has been improved yet for the ratio of population means in stratified random sampling under scrambled response. Therefore, we have consider the natural ratio of sample means estimator r^* under scramble response

situation, where $r^* = \sum_{\mathrm{h}=1}^{L} W_h \frac{\bar{y}_{\mathrm{mh}}^*}{\bar{x}_{\mathrm{mh}}^*}$.

Its variance may be derived under scrambled response situation as

$$V(r^*) = \sum_{\mathrm{h}=1}^{L} W_h f_1 R_h^{*2} C_{y_h}^{*2}.$$

The percent relative efficiency of the estimators with respect to ratio of sample means estimator r* (in the presence of scrambled response situation) is defined as

$$\mathrm{PRE}(E_1) = \frac{V(r^*)}{M(\xi_1)_{\mathrm{opt.}}} \times 100 \tag{13}$$

$$\mathrm{PRE}(E_2) = \frac{V(r^*)}{M(\xi_2)_{\mathrm{opt.}}} \times 100 \tag{14}$$

6.1. Numerical illustration

It is important to investigate the situation where our suggested methodologies are superior to the conventional ones. The performances of the suggested technique are demonstrated through empirical investigations carried over the data set of natural population and artificially generated population. For the sake of minimum impact of scrambling on the actual data we have considered random variable $T^* = 1$ and $S_T^{*2} = 0.16$. It is to be noted that the parametric values of the scrambled variables x* and y* for different populations may be obtained by using the statistical parameters of the variables x and y shown in sections.

6.2. Numerical illustration using known natural population

6.2.1. Natural population data set 1: We have selected natural population data sets on abortion rates form Statistical Abstract of the United States: 2011 to elucidate the efficacious performance of our proposed estimator. The nature of the variables y, x and z and the values of the various parameters are given below.

y, x, z: the number of abortions reported in the state of US during the years 2008, 2007 and 2005 respectively. The detailed calculations on various parameters are given in Table 1.

Table 1: Formation of 4 different strata (zone wise) out of 51 states of United States and corresponding parametric values

Strata	Constituent States	Statistical Parameters
Strata 1	Wyoming, Missouri, Mississippi, Kentucky, Oklahoma , Arkansas, Indiana, Nebraska, South Carolina, Wisconsin, Utah, South Dakota, Idaho, West Virginia.	N_h=14, n_h=13, m_h= 7, $\bar{X}_h$= 6.551, $\bar{Y}_h$= 6.59, $\bar{Z}_h$= 6.720, h = 1.
Strata 2	Alaska, Montana, New Hampshire, Minnesota, Vermont, Ohio, Arizona, New Mexico, North Dakota, Maine, Michigan, Massachusetts, Washington, Kansas, Virginia, North Carolina, Oregon, Pennsylvania, Texas, Louisiana, Colorado, Tennessee, Iowa, Alabama , Georgia.	N_h= 25, n_h= 22, m_h=15, $\bar{X}_h$=15.031, $\bar{Y}_h$=15.11, $\bar{Z}_h$=14.851, h = 2.
Strata 3	Hawaii, Rhode, Island, Connecticut, Nevada, Florida, California, Illinois.	N_h= 7, n_h= 6, m_h= 3, $\bar{X}_h$= 24.562, $\bar{Y}_h$= 24.48, $\bar{Z}_h$= 24.095, h = 3.
Strata 4	Maryland, District of Columbia, New Jersey, New York, Delaware.	N_h= 5, n_h= 4, m_h= 2, $\bar{X}_h$= 33.528, $\bar{Y}_h$= 33.55, $\bar{Z}_h$= 36.533, h = 4.

The computed values of population mean squares and correlation coefficients of the respective variables based on the strata h (h = 1, 2, . . ., 4) are shown in Table 2.

Table 2: Population mean squares and correlation coefficients of the respective variables

Strata	S_{yh}^2	S_{xh}^2	S_{zh}^2	ρ_{xy_h}	ρ_{yz_h}	ρ_{xz_h}
Strata 1	4.56	4.51	5.21	0.9784	0.9725	0.9484
Strata 2	6.91	6.93	8.87	0.9413	0.8780	0.8988
Strata 3	31.42	37.22	65.29	0.9885	0.9442	0.9545
Strata 4	19.31	29.03	53.75	0.9751	-0.3926	-0.5396

We have obtained the PRE of the proposed estimator with respect to natural ratio sample mean estimator for different values of sample sizes and outcomes are given in Table 3.

Table 3: PRE of the proposed strategy based on natural population data set 1

Sample Sizes		PRE (E_1)	PRE (E_2)	Sample Sizes		PRE (E_1)	PRE (E_2)
$n_h = 4$ (h = 1, 2, 3, 4)	$m_1 = 3$	546.7479	3356.626	$n_1 = 8$	$m_h = 3$ (h = 1, 2, 3, 4)	629.1048	4414.462
	$m_2 = 2$			$n_2 = 13$			
	$m_3 = 3$			$n_3 = 6$			
	$m_4 = 2$			$n_4 = 4$			
$n_h = 4$ (h = 1, 2, 3, 4)	$m_1 = 2$	585.2504	3862.067	$n_1 = 9$	$m_h = 3$ (h = 1, 2, 3, 4)	631.2996	4526.913
	$m_2 = 2$			$n_2 = 14$			
	$m_3 = 3$			$n_3 = 4$			
	$m_4 = 3$			$n_4 = 4$			
$n_h = 4$ (h = 1, 2, 3, 4)	$m_1 = 3$	606.8738	4256.116	$n_1 = 10$	$m_h = 3$ (h = 1, 2, 3, 4)	635.5419	4688.513
	$m_2 = 3$			$n_2 = 15$			
	$m_3 = 2$			$n_3 = 5$			
	$m_4 = 2$			$n_4 = 4$			

6.2.2. Natural population data set 2: We have taken another natural population data set on literacy rates in India based on the Census: 2011. The nature of the variables y, x and z and the values of the various parameters are given in Table 4.

y, x, z: the number of literates (persons) during the years 2001, 2011 and the female literacy rate (2011) respectively.

Table 4: Formation of 4 different strata out of 34 states of India (zone wise) and corresponding

Strata	Constituent States	Statistical Parameters
Strata 1	Andhra Pradesh (S), Karnataka (S), Kerala (S), Tamil Nadu (S), Chhattisgarh (C), Madhya Pradesh (C).	N_h= 6, n_h= 5, m_h= 3, $\bar{X}_h$= 68.06, $\bar{Y}_h$= 75.13, $\bar{Z}_h$= 65.18, h =1.
Strata 2	West Bengal(E), Jharkhand (E), Odisha (E), Bihar (E), Manipur (NE), Meghalaya (NE), Nagaland (NE), Arunachal Pradesh (NE), Sikkim (NE), Assam(NE), Tripura(NE).	N_h=11, n_h= 9, m_h= 6, $\bar{X}_h$= 62.2, $\bar{Y}_h$= 62.80, $\bar{Z}_h$= 59.7, h = 2.
Strata 3	Haryana(N), Himachal Pradesh (N), Jammu & Kashmir (N), Uttar Pradesh (N), Uttarakhand (N), Goa (W), Gujarat(W), Maharashtra (W), Punjab (W), Rajasthan (W).	N_h=10, n_h= 8, m_h= 5, $\bar{X}_h$= 68.62, $\bar{Y}_h$= 78.18, $\bar{Z}_h$= 68.12, h = 3.
Strata 4	A.&N. Islands, Chandigarh, D.&N. Haveli, Daman & Diu, Delhi, Lakshadweep, Pondicherry.	N_h= 7, n_h= 5, m_h= 3, $\bar{X}_h$= 78.07, $\bar{Y}_h$= 69.35, $\bar{Z}_h$= 79.54, h = 4.

The computed values of population mean squares and correlation coefficients of the respective variables based on the strata h (h = 1, 2, 3, 4) are shown in Table 5.

Table 5: Population mean squares and correlation coefficients of the respective variables

Strata	S^2_{yh}	S^2_{xh}	S^2_{zh}	ρ_{xy_h}	ρ_{yz_h}	ρ_{xz_h}
Strata 1	127.70	172.18	194.02	0.9929	0.9980	0.9970
Strata 2	56.01	80.53	107.77	0.9631	0.9631	0.9631
Strata 3	24.75	17.68	214.00	0.9633	0.9851	0.9988
Strata 4	21.76	89.89	52.67	0.9532	0.9822	0.9862

We have obtained the PRE of the proposed estimator with respect to natural ratio sample mean estimator for different values of sample sizes and outcomes are given in Table 6.

Table 6: PRE of the proposed strategy based on natural population data set 2

Sample Sizes		PRE (E1)	PRE (E2)	Sample Sizes		PRE (E1)	PRE (E2)
nh = 5 (h = 1, 2, 3, 4)	m1 = 3	560.256	3777.337	n1 = 5	mh = 4 (h = 1, 2, 3, 4)	587.743	4026.767
	m2 = 2			n2 = 8			
	m3 = 4			n3 = 9			
	m4 = 3			n4 = 6			
nh = 5 (h = 1, 2, 3, 4)	m1 = 4	603.106	4032.503	n1 = 5	mh = 4 (h = 1, 2, 3, 4)	592.549	4163.234
	m2 = 3			n2 = 9			
	m3 = 2			n3 = 8			
	m4 = 4			n4 = 6			
nh = 5 (h = 1, 2, 3, 4)	m1 = 2	669.103	5241.656	n1 = 5	mh = 4 (h = 1, 2, 3, 4)	595.730	4463.945
	m2 = 4			n2 = 10			
	m3 = 3			n3 = 7			
	m4 = 4			n4 = 6			

6.3. Numerical illustration using artificially generated population data set:

Efficiency comparison through artificial population generation technique helps in concluding whether a newly developed technique is better than the existing ones (see for instance the work of Singh *et al.* (2017)). Motivated by the artificial population generation techniques adopted by Singh and Deo (2003) and Singh *et al.* (2017).

We have generated the population artificially and taken 5 strata sequentially each of size 20. The PREs of the proposed estimators with respect to sample mean estimator are computed for different values of sample sizes (i.e. n and m) and correlation coefficients (i.e. rx1z1) as displayed in Table 7 and Table 8.

Table 7: PRE of the proposed strategy based on simulated population

Artificial data set -1 ryx = 0.7 ; rx1z1 = 0.5				**Artificial data set -2 ryx = 0.7; rx1z1 = 0.75**			
Sample Sizes		PRE (E_1)	PRE (E_2)	Sample Sizes		PRE (E_1)	PRE (E_2)
$n_h = 16$	$m_1 = 15$	573.1565	1193.307	$n_h = 16$	$m_1 = 15$	347.2386	515.4361
	$m_2 = 14$				$m_2 = 14$		
	$m_3 = 10$				$m_3 = 10$		
	$m_4 = 7$				$m_4 = 7$		
	$m_5 = 11$				$m_5 = 11$		
$n_h = 16$	$m_1 = 14$	585.493	1259.59	$n_h = 16$	$m_1 = 14$	359.9875	576.5938
	$m_2 = 13$				$m_2 = 13$		
	$m_3 = 9$				$m_3 = 9$		
	$m_4 = 6$				$m_4 = 6$		
	$m_5 = 10$				$m_5 = 10$		
$n_h = 16$	$m_1 = 13$	595.8184	1327.56	$n_h = 16$	$m_1 = 13$	371.6504	639.7454
	$m_2 = 12$				$m_2 = 12$		
	$m_3 = 8$				$m_3 = 8$		
	$m_4 = 5$				$m_4 = 5$		
	$m_5 = 9$				$m_5 = 9$		
$n_1 = 14$	$m_h = 10$	527.7561	1148.109	$n_1 = 14$	$m_h = 10$	295.8098	428.7791
$n_2 = 13$				$n_2 = 13$			
$n_3 = 12$				$n_3 = 12$			
$n_4 = 15$				$n_4 = 15$			
$n_5 = 16$				$n_5 = 16$			
$n_1 = 15$	$m_h = 10$	544.4047	1221.123	$n_1 = 15$	$m_h = 10$	312.4275	467.9867
$n_2 = 14$				$n_2 = 14$			
$n_3 = 13$				$n_3 = 13$			
$n_4 = 16$				$n_4 = 16$			
$n_5 = 17$				$n_5 = 17$			
$n_1 = 16$	$m_h = 10$	559.7598	1310.919	$n_1 = 16$	$m_h = 10$	328.579	504.79
$n_2 = 15$				$n_2 = 15$			
$n_3 = 14$				$n_3 = 14$			
$n_4 = 17$				$n_4 = 17$			
$n_5 = 18$				$n_5 = 18$			

Table 8: PRE of the proposed strategy based on simulated population

Artificial data set -3 ryx = 0.8 ; rx1z1 = 0.4				**Artificial data set -2 ryx = 0.7; rx1z1 = 0.75**			
Sample Sizes		PRE (E_1)	PRE (E_2)	Sample Sizes		PRE (E_1)	PRE (E_2)
$n_h = 15$	$m_1 = 8$	315.0862	484.5729	$n_h = 15$	$m_1 = 10$	331.263	523.5138
	$m_2 = 9$				$m_2 = 9$		
	$m_3 = 10$				$m_3 = 8$		
	$m_4 = 11$				$m_4 = 7$		
	$m_5 = 12$				$m_5 = 6$		
$n_h = 15$	$m_1 = 9$	305.04	447.2177	$n_h = 15$	$m_1 = 12$	313.6524	422.9525
	$m_2 = 10$				$m_2 = 11$		
	$m_3 = 11$				$m_3 = 10$		
	$m_4 = 12$				$m_4 = 13$		
	$m_5 = 13$				$m_5 = 14$		
$n_h = 15$	$m_1 = 10$	292.7033	422.6365	$n_h = 15$	$m_1 = 13$	291.9995	389.6614
	$m_2 = 11$				$m_2 = 14$		
	$m_3 = 12$				$m_3 = 11$		
	$m_4 = 13$				$m_4 = 13$		
	$m_5 = 14$				$m_5 = 12$		
$n_1 = 10$	$m_h = 9$	309.3693	427.3363	$n_1 = 11$	$m_h = 10$	274.0509	361.8313
$n_2 = 11$				$n_2 = 12$			
$n_3 = 13$				$n_3 = 13$			
$n_4 = 14$				$n_4 = 14$			
$n_5 = 15$				$n_5 = 15$			
$n_1 = 12$	$m_h = 9$	346.0258	528.5342	$n_1 = 12$	$m_h = 10$	307.0895	445.9498
$n_2 = 14$				$n_2 = 14$			
$n_3 = 15$				$n_3 = 15$			
$n_4 = 16$				$n_4 = 16$			
$n_5 = 18$				$n_5 = 18$			
$n_1 = 13$	$m_h = 9$	361.2879	610.2509	$n_1 = 15$	$m_h = 10$	340.2874	572.5906
$n_2 = 15$				$n_2 = 16$			
$n_3 = 17$				$n_3 = 17$			
$n_4 = 18$				$n_4 = 18$			
$n_5 = 19$				$n_5 = 19$			

7. Conclusion

We have noted the following observations from Tables 1 - 8.

Findings from natural population data sets (Tables 1 - 6)

a) It is clear that the selected natural data set 1 and data set 2 are heterogeneous as they have significantly different values of parameters. The values of the PREs, i.e. E_1 and E_2, are very high for different choices of the sample sizes, which indicates that

suggested estimators perform profoundly for the data sets belong to heterogeneous population. This phenomenon is a desirable one because most of the data sets we come across in practice belong to heterogeneous population. Thus, these performances of our estimators enhance their recommendation in practice.

b) For fixed values of the first phase sample sizes (i.e. n_h), the percent relative efficiencies of the proposed estimators in scrambled response situations, i.e. E_1 and E_2, are increasing when the values of the second phase sample sizes (i.e. m_h) are decreasing. This phenomenon indicates only a smaller fraction of the sample is to be drawn at the second phase but it may produce precise estimate. This reduces the cost of the survey.

Findings from artificially generated population data sets (Tables 7 - 8)

a) It may be observed that the data sets obtained from the artificially generated population are almost homogeneous as parametric values are close enough. It may be noted that our proposed strategies produce precise estimates as the values of PREs are high for different values of the sample sizes.
b) For the fixed values of n_h (first phase sample size), rx1z1 (correlation coefficient) and ryx (correlation coefficient), the percent relative efficiencies E_1 and E_2 are increasing when the values of the sample size at the second phase m_h is decreasing. This reduces the cost of the survey.
c) For the fixed values of n_h (first phase sample size), m_h (second phase sample size) and ryx (correlation coefficient) the percent relative efficiencies E_1 and E_2 are increasing when the values of rx1z1 (correlation coefficient) are decreasing. Thus, it is clear that for high values of the correlation coefficient between study variable and auxiliary variable our proposed strategy produces more precise estimates. This behaviour helps us in choosing a population for application of our strategy in real life.

Therefore, it is established that our proposed strategies may produce efficient estimators in comparison with the conventional ones and they are also applicable for homogeneous and heterogeneous population. Looking at the encouraging findings, our proposed methodologies are recommended to the survey practitioners for their applications in real life.

Acknowledgements

Authors are thankful to the editor and reviewers for their valuable and constructive suggestions, which motivated us to generate an improved version of the present manuscript.

References

Diana, G., Perri, P. F., (2010). New scrambled response models for estimating the mean of a sensitive quantitative character. *Journal of Applied Statistics*, Vol. 37, pp. 1875–1890.

Eichhorn, B., Hayre, L. S., (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, Vol. 7, pp. 307–316.

Giancarlo, D., Pier, P. F., (2010). New scrambled response models for estimating the mean of a sensitive quantitative character. *Journal of Applied Statistics*, Vol. 37, pp.1875–1890, DOI: 10.1080/02664760903186031.

Greenberg, B. G., Kuebler, R. R., Abernathy, J. R., Horvitz, D. G., (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of American Statistical Association*, Vol. 66, pp. 243–250.

Kadilar, C., Cingi, H., (2000). Ratio Estimator in stratified sampling. *Biometrical Journal*, Vol. 45, pp. 218–225.

Kadilar, C., Cingi, H., (2003). A new ratio Estimator in stratified sampling. *Communication in Statistics-Theory and Methods*, Vol. 34, pp. 597–602.

Pollock, K. H., Bek, Y., (1976). A comparison of three randomized response models for quantitative data. *Journal of American Statistical Association*, Vol. 71, pp. 884–886.

Koyuncu, N., Kadilar, C., (2008). Ratio and product estimators in stratified random sampling. *Journal of Statistical Planning and Inference*, Vol. 139, pp. 2552–2558.

Koyuncu, N., Kadilar, C., (2009). Family of estimators of population mean using two auxiliary variables in stratified random sampling. *Communications in Statistics-Theory and Methods*, Vol. 38, pp. 2398–2417.

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

Shabbir, J., Gupta, S., (2005). Improved ratio estimators in stratified sampling. *American Journal of Mathematical and Management Sciences*, Vol. 25, pp. 293–311.

Singh, S., Deo, B., (2003). Imputation by power transformation. *Statistical Papers*, Vol. 4, pp. 555–579.

Singh, H.P., Vishwakarma, G. K., (2005). Combined Ratio - product Estimator of Finite Population Mean in Stratified Sampling. *Metodologia de Encuestas*, Vol. 8, pp. 35–44.

Singh, R., Sukhatme, B. V., (1973). Optimum stratification with ratio and regression method of estimation. *Annals of the Institute of Statistical Mathematics*, Vol. 25, pp. 627–633.

Singh, R., Kumar, M., Chaudhary, M. K., Kadilar, C., (2009). Improved Exponential estimator in Stratified Random Sampling. *Pakistan Journal of Statistics and Operation Research*, Vol. 5, pp. 67–82.

Singh, H. P., Chandra, P., Joarder, A. H., Singh, S., (2007). Family of estimators of mean, ratio and product of a finite population using random non-response. *Test*, Vol. 16, pp. 565–597.

Singh, G. N., Sharma, A. K., Bandyopadhyay, A., (2017). Effectual Variance Estimation Strategy in Two Occasions Successive Sampling in Presence of Random Non-Response. *Communications in Statistics-Theory & Methods*, Vol. 46, pp. 7201–7224.

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*, Vol. 53, pp. 375–387.

Appendix

We have generated three sets of independent random numbers of size N (N = 100), namely $x[k]$, y$[k]$and z$[k]$(k $= 1, 2, 3, \ldots,$ N) from a standard normal distribution as presented below.

The following algorithm is used to generate the population artificially:

1. Generate there random variables x1 and z1 and a which are normally distributed with mean 0, S.D. = 1 and which are of size 100.
2. Define N = 100.
3. Define rx1z1 = 0.75, Sx1 $= \sqrt{50}$(s.d. of x1), Sz1 $= \sqrt{40}$ (s.d. of z1), mx1$= 20$ (i.e. mean of x1), mz1 $= 25$ (i.e. mean of z1) . [Note: x1, z1 are temporary variables]
4. a = sz1*sz1*(1-(rx1z1^2))
5. for (j in 1:N)
   {
   x[j] = 20.0+(sx1*x1[j])
   z[j] = 25+(√a)*z1[j])+(rx1z1*sz1*x1[j])
   }
6. Take output of the variables x and z
7. Generate the variable y with ryx = 0.7 from the variable
8. Take output of the variable y
9. Repeat the steps 1 to 8 with different values of rx1z1 (step 3), which will generate different population for different values of the correlation coefficients.

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 63–88, https://doi.org/10.59170/stattrans-2023-064
Received – 07.11.2022; accepted – 04.08.2023

Model for measuring the impact of good pharmacovigilance practices of COVID-19 patients on hcp reactivity: Morocco case study

Oumaima Oullada[1], **Mohamed Ben Ali**[2], **Ahmed Adri**[3], **Said Rifai**[4]

Abstract

This paper presents a conceptual model used to evaluate how the improvement of good pharmacovigilance practices by patients during COVID-19 period influences the reactivity of the healthcare professionals (HCPs) in the Draa Tafilalet region in Morocco, concerning the reporting of adverse drug reactions (ADRs) through barriers that influence the reporting from both patients and HCPs. The empirical study is based on a survey submitted to a sample of a total of 180 HCP and on the application of latent variable structural modelling through the partial least squares (PLS) method. The 2017 version of the XL-STAT software served to perform the statistical calculations. The study investigates the reliability and validity of the proposed model. Our conclusions show that the improvement of good pharmacovigilance practices impact positively the reactivity of HCP in terms of ADRs reporting. The reliability of the measurement was > 0.7, which allowed us to test the internal and external validity of our conceptual model. 11 hypotheses were validated against two invalid derivative hypotheses. Spontaneous ADRs reporting is the cornerstone of any pharmacovigilance system aiming to maintain patient safety. Our findings indicate the necessity firstly, to initiate a training program on reporting for all HCPs, and secondly, to inform the general public about the national pharmacovigilance center, where ADRs can be reported. Both initiatives aim to keep the culture of ADR reporting perennial.

Key words: pharmacovigilance, HCP reactivity, structural equations modelling, latent variable, PLS.

[1] National Higher School of Electricity and Mechanics - ENSEM- Hassan II University of Casablanca – B.P: 8118 Oasis – Casablanca – Morocco. Laboratory of Mechanics- Production and Industrial Engineering – LMPGI – Higher School of Technology of Casablanca – EST, Hassan II University of Casablanca – B.P 8112 Oasis – Casablanca, Morocco. E-mail: oumaimaoullada@gmail.com. https://orcid.org/0009-0004-6313-5532.

[2] Laboratory of Mechanics- Production and Industrial Engineering – LMPGI – Higher School of Technology of Casablanca – ESTC – Hassan II University of Casablanca – B.P 8112 Oasis – Casablanca, Morocco. E-mail: benali8mohamed@gmail.com. ORCID: https://orcid.org/0000-0002-8615-7935.

[3] Laboratory of Mechanics – Production and Industrial Engineering – LMPGI – Higher School of Technology of Casablanca – ESTC – Hassan II University of Casablanca – B.P 8112 Oasis- Casablanca, Morocco. E-mail: ahmedadri@gmail.com. ORCID: https://orcid.org/0000-0003-1355-5059

[4] Laboratory of Mechanics- Production and Industrial Engineering – LMPGI – Higher School of Technology of Casablanca – ESTC – Hassan II University of Casablanca – B.P 8112 Oasis – Casablanca, Morocco. E-mail: dptgmp@gmail.com. ORCID: https://orcid.org/0000-0002-2813-1606.

1. Introduction

The history of pharmacovigilance started 170 years ago, when a young girl died after receiving chloroform anesthetic before removal of an infected toenail (Routledge, 1998). The catalyst for the development of pharmacovigilance, was the thalidomide tragedy that occurred in the 1960s. Dr. McBride, an Australian doctor, observed that the incidence of congenital malformations of babies (1.5%) had increased up to 20% in women who had taken thalidomide during pregnancy (Yarrow, 1961). At the international level, the World Health Organization (WHO) began its pharmacovigilance operations after the discovery of the teratogenic effects of thalidomide, during the 16th WHO Assembly, the formation of the WHO Programme for International Drug Monitoring (PIDM) in 1968 (*Regulation and Prequalification*, n.d.). In Morocco, the National Pharmacovigilance Center (NCPV) was established in 1991. It gained WHO membership in 1992, becoming the first African, Arabic, and 34th international pharmacovigilance system (*CAPM Plateforme*, n.d.). According to the WHO, Pharmacovigilance is defined as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug related problems (Mirosevic Skvrce et al., 2020). The main scope of pharmacovigilance is to improve the safe and rational use of medicines. While the medicine brings a real benefit to the human being by saving his health and suffering, its use is never without risk. These risks defined by the term "Adverse Drug Reactions" (ADRs) are a significant cause of morbidity and mortality worldwide (Joubert & Naidoo, 2016). However, we have to admit that drugs can be responsible for adverse effects (AEs), several international studies have highlighted the harmful consequences of ADR, whether in terms of morbidity and mortality, hospitalization or medical costs. In the USA, it was estimated that ADR caused about 106,000 deaths a year, representing between the fourth and sixth cause of death (Starfield, 2000). According to the European Commission, ADRs cause 197,000 deaths a year, and represent the fifth cause of death in hospitalized patients (Montané et al., 2018). In Morocco, the hospitalization costs inherent to the management of patients with ADRs incur additional expenses and represent more than 15–20% of the hospital budget (El Hamdouni et al., 2020). Amid the global COVID-19 pandemic caused by the new SARS-CoV-2 Coronavirus, countries have recommended various protocols. In Morocco as soon as the therapeutic protocol was launched, the NCPV, set up a strong monitoring system for AEs to study drugs and vaccines used for new indications, assess their benefit/risk ratio, and improve patient safety. Healthcare professionals (HCPs) have access to reporting tools, while vaccinated individuals can report ADR (*Accueil*, n.d.), patients had an important contribution to signals for ADRs related to drug. Our research proposes a conceptual model to measure the impact of Good Pharmacovigilance Practices by patient over the

COVID-19 period on the HCP's responsiveness to ADRs. This study is conducted in a south-eastern region of Morocco. This model will be tested using partial least squares structural equation modelling (PLS-SEM). This choice is motivated by its ability to build models with several variables and complex interactions in order to approach the complexity of real situations and is also based on these research (Avkiran et al., 2018) (Ali et al., 2019) (Sahaf et al., 2018) (Sebtaoui et al., 2020)

2. Literature review

2.1. Construct 1: Good Pharmacovigilance Practices over the COVID-19 period by patients.

The patient is the main stakeholder in pharmacovigilance, which is the ultimate goal of ensuring the safe use of drugs. A patient reporting of ADRs could supplement the existing reporting system and contribute to early detection of ADRs (Weigmann, 2016). A growing number of countries are involving patients in the direct reporting of ADRs (e.g., European Union countries since 2012), but little is known about what the patient reporting adds to pharmacovigilance systems (Inácio et al., 2017). Patient-reported safety information leads to a better understanding of the patient's experiences of the ADRs (Härmark et al., 2016). In the UK, the patient reporting can significantly contribute to drug safety by detecting distinct signals of disproportionate reporting that may not be identified from HCP reports (Hazell et al., 2013). In the Netherlands patients' reporting ADR offer a valuable contribution to signal detection, complementing the reports from HCP (Maguire et al., 2007). Involving pharmacists and doctors to encourage patient participation in data reporting boosts awareness of ADR significance, motivating patients and potentially reducing mortality and morbidity rates (Naoual Nchinech et al., 2020) (Hadi et al., 2017) (Awodele et al., 2011)(Toklu & Uysal,2008). A study revealed that patients report symptoms earlier and more frequently than clinicians, with interesting information (Engla & Journal, 2010). The 2000 International Conference highlighted the importance of patient ADR reporting for pharmacovigilance, recognizing its insight-providing capability. Many comparative studies found that the patient ADR reporting frequently offers more comprehensive details than HCP' report (Assanee et al., 2021)(Avery et al., 2011). Here comes our aim through those studies carrying only for good pharmacovigilance practices of HCP (including knowledge, attitude, practice and perception). The research underscores the need to include patients in pharmacovigilance efforts, especially from the viewpoint of HCPs in DRAA TAFILALET, Morocco. Four key Good Pharmacovigilance Practices, often mentioned in the literature, were implemented during the COVID-19 period, extending beyond HCPs: *Patient's Knowledge of ADR; Patient's Attitude; Patient's Practice; Patient's Notification of ADR.*

2.2. The construct 2: Factors braking to report ADR

Factors affecting patient reporting of adverse drug reactions: ADR forms a significant problem, both from a medical point of view and as an economic burden. Spontaneous reporting of ADRs is one of the methods for post-marketing surveillance of drug safety. A systematic review was made from 1964 to December 2014 in the UK, the Netherlands, and Australia. It showed that from 15 studies, there is poor awareness, confusion about who should report the ADR, difficulties with reporting procedures, lack of feedback on submitted reports, mailing costs, ADRs resolved, and prior negative reporting experiences (Al Dweik et al., 2017). In 2012, a cross-sectional study was conducted in Saudi Arabia which revealed that the public lacked awareness about ADRs and had limited knowledge on how to report them (Sales et al., 2017). In a separate study conducted in Japan, 845 citizens were found unaware of the direct patient ADR reporting system (Kitabayashi & Inoue, 2022)**.** One other concern, patients may believe that public reporting of drug-related problems may affect the physician-patient relationship, which is proven by these studies (Kitabayashi & Inoue, 2022) (Inácio et al., 2017). On the other hand, in 2018, a cross-sectional survey among 360 patients in Nigeria demonstrated a low level of awareness of pharmacovigilance and ADR reporting (Adisa & Omitogun, 2019). A statistical study in Thailand utilizing PLS-SEM showed a notable link between instrumental attitude and patients' intention to report ADRs to community pharmacists (Assanee et al., 2021). Based on a literature review, we surveyed HCPs to gather their opinions on factors influencing patient ADR reporting, considering their frequent interactions with patients.

Factors inhibiting HCP reporting ADR: Factors that made the HCP refrain from reporting a suspected ADR were similar. In northern Sweden, a study aimed to explore attitudes and main factors that refrain from reporting ADR(s): lack of time and giving priority to other matters in medical care as well as confidence that no new information will be provided by reporting and unwillingness to write a report on just suspicion of cause and effect (Bäckström et al., 2000). Finland examined the reasons why HCP do not report all suspected ADRs. The COVID-19 pandemic might be one of the contributing factors explaining why the subgroup of Finnish physicians selected then the "Lack of time" as the primary reason. Other reasons include factors such as the suspected ADR already being known ("it is not clear what is worth reporting"), the belief that someone else will report the ADR and the perception that the patient's suspicion was not credible, confidence that no new information will be provided by reporting (Sandberg et al., 2022). Also, lack of time was only the most common answer from HCP on two different occasions pre-COVID-19 (Bäckström et al., 2000)(Sandberg et al., 2022). On the other side, the most important factors were: the reaction is already well known, never suspected any ADR, forgetfulness, difficulties in reporting only on suspicion, lack of time, and uncertain of how to report and the HCP stated that they

would like a feedback letter containing the causality assessment (Ekman & Bäckström, 2009). In the Norwegian healthcare system found that HCP often focused on patient-related information such as weight and height. HCP usually reported more serious reactions that lead to hospitalization, life-threatening conditions, or death (Vaismoradi et al., 2019). Whereas other studies confirmed the same main factors for the decision to report an ADR: lack of time, motivation (Biriell & Edwards, 1997) (Hazell & Shakir, 2006) (O'Callaghan et al., 2018) (Stergiopoulos et al., 2016) (Joubert & Naidoo, 2016) (Rabba & Ain, 2015).

2.3. The construct 3: Reactivity of HCP

Physicians, dentists, pharmacists, nurses, and midwives as well as all other paramedical professionals must collaborate in the safe use of health products in Morocco. They must report to the NCPV, as soon as possible of any suspected ADR related to the use of one or more products under normal conditions of use, whether expected, unexpected, serious, or not. Any ADR appearing outside the normal conditions of use, any other reaction they judge relevant to report (Moroccan good pharmacovigilance practice) (CAPM Plateforme, n.d.). Collaborative studies with the Moroccan Pharmacovigilance Center have assessed pharmacists' knowledge, revealing a moderate understanding of pharmacovigilance, with 11.5% encountering AEs requiring mandatory intervention in their practice (N Nchinech et al., 2019). A South African study found that community pharmacists exhibited positive knowledge, perception, and attitudes toward pharmacovigilance. A global analysis of 50 countries, including Morocco, revealed that direct patient reporting systems were present in 44 countries, contributing to 9% of total reports, while the majority came from HCP (Margraff & Bertram, 2014). A systematic review examined doctors' knowledge, attitude, and practice regarding ADR and pharmacovigilance. Knowledge refers to understanding, attitude is the predisposition to respond positively or negatively, and practice involves applying knowledge practically (Abubakar et al., 2014). The practice of doctors was based on four parameters in the majority of surveys conducted. These include: "encounter with ADRs", "number of ADRs ever reported", "training on ADR reporting" and "source of information" to the doctors (Abubakar et al., 2014). In developing countries, Knowledge (understanding), Attitude (emotional and cognitive beliefs), and Practice (observable healthcare actions) collectively define the behavior and decision-making of HCP, which can be influenced by internal and external factors (Thomas & Zachariah, 2018). In Kuwait, a study found that hospital pharmacists had strong knowledge and a positive attitude towards pharmacovigilance and ADR reporting, yet most had never reported an ADR (Alsaleh et al., 2017). In a Knowledge, attitude and practice KAP study, most pharmacists in South Africa

were aware of pharmacovigilance, but fewer than half had reported ADRs (Joubert & Naidoo, 2016). The KAP of pharmacovigilance among HCP was highlighted and studied in different countries, and developing countries need improvement. The relevant professionals have poor knowledge, a positive attitude, and poor practice (Thomas & Zachariah, 2018). In Istanbul the KAP of pharmacovigilance community pharmacists have insufficient knowledge about pharmacovigilance practices (Toklu & Uysal, 2008). On behalf of Moroccan pharmacy students' knowledge and perceptions about pharmacovigilance confirmed the utility of KAP among them even if they are future pharmacists to maintain the continuity of ADR reporting (N. Nchinech et al., 2020). A study in Bosnia and Herzegovina found a gap between positive perceptions and actual ADR reporting, recommending education and training to improve reporting and engagement with pharmacovigilance (Amrain & Bečić, 2014).

3. Research method:

3.1. Fundamentals of PLS-SEM modelling (theory)

Path analysis models were first developed by Sewall Wright (1921), a biostatistician, in the early 1920s. It was not until the 1970s that structural models started being used in the social sciences, as noted by Jöreskog (1973)(Joe F. Hair et al., 2011). The main function of modelling is to understand, test, analyze, and interpret a given (real) phenomenon by measuring the various causal links between its components. It is a simplification of the reality of a given phenomenon or problem in interaction. The aim is to understand and explain the complexity of a model (system) by measuring its observed variables. LISREL is the best-known technique for causal modelling (Joreskog and Sorbom,1989); (Hagedoorn & Schakenraad, 1994)). Nevertheless, LISREL's efficiency decreases when it is faced with small data samples (Fornell & Bookstein, 1982), an alternative causal modelling approach called partial least squares (PLS) has been developed to alleviate these problems (Wold, 1985). The PLS Path modelling (PLS-PM) approach is based on partial least-squares, it was initiated by (Wetzels et al., 2009). Its aim is to estimate the score of the various latent variables by an iterative procedure based on simple regressions using the ordinary least squares (OLS) method. Over the past two decades, the number of studies using the PLS-SEM method has increased. Which demonstrates its growing importance in research (Law & Fong, 2020). While Structural Equation Modelling (SEM) is a broad term that includes various statistical models, one of its specific approaches is covariance-based SEM (CB-SEM) (Jöreskog, 1978). Variance-based SEM techniques, like PLS-SEM (Avkiran et al., 2018) (Hair carole l. Hollingswoth, Chong, Jeo, 2017) (Cheah et al., 2018). As (Chin, 1998) points out "To many social science researchers, the covariance-based

procedure is tautologically synonymous with the term SEM". Furthermore, PLS-SEM presents advantageous attributes when handling intricate models, non-normal data, and small sample sizes (Joseph F. Hair et al., 2019). Indeed, PLS-SEM has gained widespread application in various social science disciplines: organizational management (Sosik et al., 2009), international management (Richter et al., 2016), human resource management (Ringle et al., 2019), supply chain management (Kaufmann & Gaeckler, 2015), the impact of quality practices on firm performance (Ali et al., 2019), the transmission of systemic risk (Avkiran et al., 2018), the relationship between future time perspective, wisdom, hospitality management discipline (Faizan et al., 2018), performance of fit indexes in Generalized structured component analysis (Cho et al., 2020), quality management (Magno et al., 2022), business marketing research (Guenther et al., 2023), IT research models (Robert & Brown, 2004). After reviewing the PLS-SEM literature, we are fortunate to find that the application of PLS-SEM in pharmacovigilance has been reported once in the existing literature reviews. It was used to determine the influencing factors with intention to report ADRs to community pharmacists in Thailand (Assanee et al., 2021). As a result, our article will be the first to present a PLS-SEM analysis in the field of pharmacovigilance.

3.2. Assessment of the PLS-SEM model

For the first time in the field of pharmacovigilance studies, we use the iterative OLS regression-based (PLS-SEM) ((Kroonenberg & Lohmoller, 1990); Wold, 1982).The goal of our research is predicting key target constructs. The research aims to explore the existing structural theory. The formative constructs are part of the structural model**.** Note that this approach is recommended when the theory is more approximate.

Reliability: Measurement reliability reflects the consistency in repeated measurements, crucial for obtaining consistent and close results. This process involves assessing internal consistency**,** we follow Hair et al.'s (2017a) recommendation of utilizing both Cronbach's alpha as the lower boundary and composite reliability as the upper boundary. The formulas for calculating Cronbach's alpha and composite reliability are provided in Hair et al.'s work.

Measurement models: In PLS-SEM, there are two types of measurement models: reflective indicators represent variations in the latent construct, while formative indicators are influenced by changes in the indicator variables, contributing to the formation of the latent construct (Joe F. Hair et al., 2011). The evaluation of (external) measurement models depends on the nature of the chosen diagram (formative, reflective or MIMEC) (Jacobowicz, 2007). The same author confirms that the reflective diagram (Figure 1) is the most suitable for most uses of latent variable structural

equation models and that this choice is based mainly on the researcher's subjectivity. Each manifest variable is related to its latent variable by a simple regression:

ξ: Latent variable

X : Manifest variable

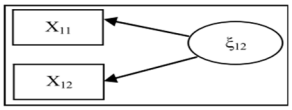

Figure 1: Reflective pattern (Joe F. Hair et al., 2011).

The relationship between the latent variable and the set of manifest variables that are associated with it can be written in the following form $\boldsymbol{X_{kj} = \pi_{kj} * \xi_{kj} + \varepsilon_{kj}}$

With: $\boldsymbol{X_{kj}}$: vector associated with the j[th] manifest variable of the latent variable ξ_k

$\boldsymbol{\xi}$: latent variable

K: index of latent variables

$\mathbf{k_j}$: index of manifest variables of the k bloc

π: loading associated with x_{kj}

$\boldsymbol{\varepsilon_{kj}}$: error term (measurement errors of manifest variables).

Convergent validity: For assessing validity, researchers should use the AVE (average variance extracted) to evaluate convergent validity. An AVE value of 0.50 or higher suggests adequate convergent validity, meaning a construct explains at least half of its items' variance, with the AVE of each latent construct exceeding the squared correlation with any other latent construct (Joe F. Hair et al., 2011) Chin et al. (2010). We calculate the AVE relating to each latent construct:

$$\boldsymbol{AVE = \frac{\sum[\gamma_i^2]var(VL)}{\sum[\gamma_i^2]var(VL) + \sum[var(\varepsilon_i)]}}$$

Along with: **VL:** latent variable

$\boldsymbol{\gamma_i^2}$ factorial contributions (loadings)

$\boldsymbol{\varepsilon_i}$: variance of errors.

Discriminating validity (divergent): Fornell and Larcker (1981) criterion: each construct's AVE should be higher than its squared correlation with any other construct. (e.g. Chin, 1998b).

Validation of the structural model:

The structural model explains connections between latent variables, and for PLS analysis, there are no fit adjustment indices. Model evaluation depends on the predictive relevance of measures, and validation of model adjustments is based on specific conditions :

- **Goodness of fit index (GoF):** This index takes into account both the performances of the structural model and of the measurement model (Wetzels et al., 2009). It is defined by the geometric mean of the average of the communities (or AVE) on all the latent variables $\overline{H^2}$ and the average of R^2 associated with the endogenous latent variables.

$$GOF = \sqrt{\overline{H^2} * \overline{R^2}}$$

- **The coefficient of determination (R^2):** used to judge the quality of a linear, single or multiple regression. It measures the adequacy between the model and the observed data. The value of $\mathbf{R^2}$ must be at least greater than 0.1 (Fricker et al., 2012).

$$R^2 = \frac{SCR}{SCT} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \Big/ \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

With: **SCR**: corresponds to the sum of the squares of the residues (residual variance);

SCT: corresponds to the sum of the total squares (total variance explained);

Y_i: the measurement values;

$\widehat{Y_i}$: the predicted values

$\overline{Y}$: The average of the measurements.

Similarly, referring to the guidelines of Croutsche (2002), and Falk and Miller (1992), the structural model can be retained ($R^2 > 0.1$). (Chin, 1998) articulated the values of 0.67, 0.33 and 0.19 are respectively considered as substantial, moderate and low.

- **Structural equations of the conceptual model:** The internal model is defined by linear equations connecting the latent variables between them. For all endogenous ξ_k, we have:

$$\xi_k = \sum_{i:\, \xi_i \to \xi_k} \beta_{ki}\, \xi_i + \zeta_k$$

where β_{ki} represents the coefficient associated with the relation between the variables ζ_k and ξ_i . $\zeta_k$in an error term and ξi → ξ_i . ζ_k explains ζ_k.

- **Hypothesis tests:** Confirmatory research aims to establish causal relationships using path models and fit indices. In PLS-SEM for explanation, the focus is on

understanding a dependent variable, achieved by maximizing explained variance (R^2) and analysing the significance, size, and direction of path coefficients to test model assumptions.

- **Effect size** (f^2): is an index that brings assessment of the effect size allowing the researchers to observe the effect of each exogenous construct on endogenous constructs. The f^2 can be evaluated using Cohen's f^2 (Cohen, 1988). f^2 0.02 represent small; 0.15 medium; 0.35 strong (large effect of the exogenous latent variable). The effect size is:

$$f^2 = \frac{R^2_{AB} - R^2_A}{1 - R^2_{AB}}$$

In PLS structural modelling, path coefficients standardized beta coefficients in ordinary least squares regressions and are evaluated for significance through bootstrapping. Non-significant or opposing paths undermine the initial hypothesis, while significant paths in the hypothesized direction offer empirical support for the proposed causal relationship. (Kroonenberg & Lohmoller, 1990). It explains whether the fact that a coefficient is significant depends on its standard errors that are obtained by bootstrapping to enable computing the empirical T values, P values. Most of researchers use P value to assess significance levels (Faizan et al., 2018).

3.3. Database and model specification

The methodology of this study consists of four steps dealt with in the following sections:

Preparation of the methodological framework of research: Our study, conducted from March to December 2021 in southeast Morocco, used a questionnaire distributed face-to-face and via Google Forms due to COVID-19 restrictions. We obtained 180 surveys with an impressive 90% response rate, despite the region's limited HCP and smaller cities.

Questionnaire design and judge validation: Based on the literature review to identify each latent variable, an item. Items related to good pharmacovigilance practices were collected on a Likert scale of 5 degrees ranging from very poor to very good, while factors breaking reporting ADRs were collected on a Likert scale of 5 degrees ranging from disagree to strongly agree, HCP practices were collected on a Likert scale of 5 degrees ranging from strongly satisfied to unsatisfied and HCP perception was collected on a Likert scale of 5 degrees ranging from strongly agree to disagree. Note that the total number of items is 53 (items/questions).

Research model: We seek, through our causal model, to measure the relationship between Patients' Good Pharmacovigilance Practices during the COVID-19 period toward ADR Reporting and HCP's reactivity toward ADR Reporting through barriers affecting ADR Reporting by Patients and HCP (Figure 2 and Table 1). To do this, an

overall hypothesis (OH) was formulated: '**The good pharmacovigilance practices of reporting ADR by patients positively influence the reactivity of HCP for reporting ADRs**'. For each causal relationship, we have formulated a derivative hypothesis (total: 13 derivative hypotheses), in Table 2.

Data gathering and validating of the instruments: After collecting survey data using SPSS 21, the data purification process follows Churchill's paradigm. It begins by defining the item list, setting boundaries for what to include in the measurement. Then, the proposed items are aligned with the model's dimensions. Purification steps are executed for pharmacovigilance practices, ADR reporting factors, and HCP reactivity, involving data summarization and potential item modification or deletion. Depending on results, adjustments may be made in steps 2 and 3. If items are retained, steps 6 and 7 evaluate scale reliability and validity. If these criteria are unmet, a revision of the item list is necessary (Churchill, 1979).

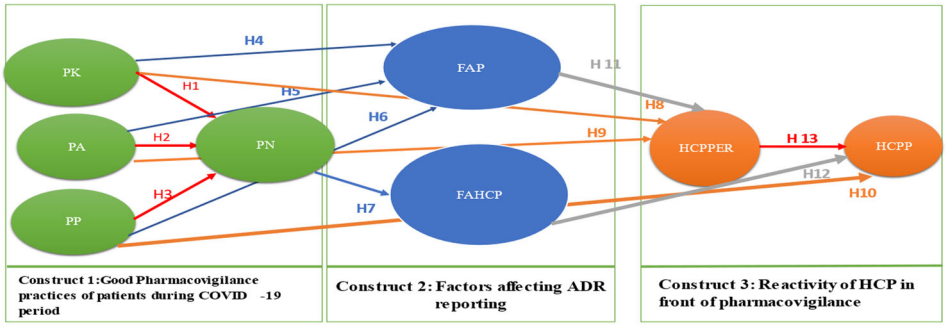

Figure 2: Proposed model

Table 1: Code used in the causal model

Constructs of the proposed model	Code of the construct's components	The components
Good Pharmacovigilance Practices over the COVID-19 period	PK	Patient's Knowledge of ADR
	PA	Patient's Attitude
	PP	Patients' practices
	PN	Patient's Notification of ADR
Factors affecting ADRs reporting	FAP	Factors affecting patients of reporting ADRs
	FAHCP	Factors affecting HCP of reporting ADRs
The reactivity of HCP	HCPP	HCP's Practices
	HCPPE	HCP's Perception

Table 2: List of hypotheses

Hypothesis Number	Causal relationship	Hypothesis Formulated
H1	PK >>> PN	We suppose that PK has a strong impact on PN
H2	PA >>>PN	We suppose that PA has a strong impact on PN
H3	PP>>>PN	We suppose that PP has a strong impact on PN
H4	PK>>>FAP	We suppose that PK has a strong impact on FAP
H5	PA>>>FAP	We suppose that PA has a strong impact on FAP
H6	PP>>>FAP	We suppose that PP has a strong impact on FAP
H7	PN>>>FAHCP	We suppose that PN has a strong impact on FAHCP
H8	PK>>>HCPPER	We suppose that PK has a strong impact on HCPPER
H9	PA>>>HCPPER	We suppose that PA has a strong impact on HCPPER
H10	PP>>>HCPP	We suppose that PP has a strong impact on HCPP
H11	FAP >HCPPER	We suppose that FAP has a strong impact on HCPPER
H12	FAHCP > HCPP	We suppose that FAHCP has a strong impact on HCPP
H13	HCPPER >HCPP	We suppose that HCPPER has a strong impact on HCPP

4. Results and discussion

4.1. Reliability

As mentioned in the methodology, the first step involves assessing **the reliability** of our measurements. Reliability ≥0.7 was considered acceptable. According to the results in Table 3, Cronbach's alpha and Rhô.D.G indexes calculated for each latent variable are above 0.7 and with reference to the recommendations of Nunnally and Bernstein (1994) (MERLEN, 2017), Fornell and Larker (1981) these results are satisfactory (reliable) according to (Kline, 1999).

4.2. Evaluation of external model (Measurement Model):

Note that the manifest variables form the blocks **around latent variables**. Since the measurement models are of the reflective type, the blocks must be one-dimensional to ensure that the obvious variables reflect their latent variable. The first eigenvalue for each block must represent at least 50% of the sum of all values in the same block. This is the case for the results depicted in **Table 4**. This confirms the one-dimensionality of the blocks.

Table 3: Reliability of measurements

Latent Variable	Items	Cronbach's alpha	Rho D. G
PK	5	0.6332	0.601
PA	5	0.7656	0.8434
PP	5	0.7737	0.8480
PN	4	0.8028	0.8715
FAP	8	0.784	0.826
FAHCP	10	0.9131	0.9781
HCPPER	8	0.9113	0.9293
HCPP	8	0.8023	0.8543

**Extracted from XL-stat software (2017 version).*

Table 4: Eigenvalues of the latent variables of the model

PK	PA	PP	PN	FAP	FAHCP	HCPPER	HCPP
2.0351	2.7389	2.6619	2.5190	3.1561	5.6848	5.0059	3.4516
1.5115	1.3235	0.8764	0.6099	1.4491	1.3188	1.1492	1.3622
0.8266	0.5445	0.7680	0.4870	1.0780	0.7598	0.5241	1.0797
0.4271	0.2193	0.4141	0.3841	0.7669	0.6556	0.3637	0.6448
0.1996	0.1739	0.2796		0.6920	0.5029	0.2943	0.5384
				0.4861	0.4129	0.2643	0.3857
				0.2460	0.2151	0.2526	0.3189
				0.1258	0.1900	0.1459	0.2187
					0.1558		
					0.1043		

Table 5: Quality index of measurement models

Latent Variable	AVE	Rho D. G
PK	0.5346	0.601
PA	0.5283	0.8434
PP	0.5109	0.8480
PN	0.6021	0.8715
FAP	0.5851	0.826
FAHCP	0.5171	0.9781
HCPPER	0.6172	0.9293
HCPP	0.6972	0.8543

**Extracted from XL-stat software (2017 version).*

Table 6: The discriminating validity (Extracted from XL-stat software (2017 version))

Latent Variable	PK	PA	PP	PN	FAP	FAHCP	HCPPER	HCPP	(AVE)
PK	0.7311*								0.5346
PA	0.4577	0.7268*							0.5283
PP	0.0861	0.1652	0.714*						0.5109
PN	0.0106	0.0874	0.183	0.7759*					0.6021
FAP	0.0732	0.0976	0.0147	0.0147	0.7649*				0.5851
FAHCP	0.0478	0.0444	0.0032	0.0315	0.3614	0.7190*			0.5171
HCPPER	0.0901	0.1414	0.0019	0.0014	0.0586	0.0340	0.7856*		0.6172
HCPP	0.0057	0.0612	0.0090	0.0446	0.0597	0.0099	0.0994	0.8349*	0.6972

** Square root of the average variance extracted (AVE).*

In our methodology, the second step involves calculating the AVE. **Table 5** demonstrates that our measurement model exhibits strong convergent validity, AVE exceed 0.5 for each latent variable, according to Fornell and Larcker's guidelines. The third phase of our methodology consists of calculating the **Square Root of AVE.** The results of **Table 6** show that the square root of the AVE of each latent variable exceeds the correlations between the latent variables (two by two). **Convergent validity and divergent validity confirm that our measurement model is valid**.

4.3. Evaluation of internal model (Structural model):

To test the internal model, we referred to Good of fit index (GoF). According to the results obtained in Table 7, the research model can be retained in terms of the threshold (GoF > 0.5), following the instructions of (Wetzels et al., 2009).

Table 7: Adjustment indices (*Extracted from XL-stat software, 2017 version)

Specification	GOF
Absolute	0.5337
Relative	0.8009
External model	0.9423
Internal model	0.8316

Chin (1998) articulated the values of 0.67, 0.33 and 0.19 are respectively considered as substantial, moderate and low. Similarly, referring to the guidelines of Croutsche (2002) and Falk and Miller (1992), the structural model can be retained ($R^2 > 0.1$). The results **of R^2 and R^2-adjusted** (**Table 8)** are substantial to moderate. The findings from our survey demonstrate the validity of both the external measurement model and the

internal structural model. This validation assures us of the credibility of the formulated hypotheses and measures the various causal links within our proposed causal model.

Table 8: R^2 Results and R^2-adjusted (*Extracted from XL-stat software, 2017 version)

Latent Variable	Type	R^2	R^2 adjusted
PK	Exogenous		
PA	Exogenous		
PP	Exogenous		
PN	Endogenous	0.2208	0.2119
FAP	Endogenous	0.2041	0.1939
FAHCP	Endogenous	0.3315	0.4153
HCPPER	Endogenous	0.1612	0.1517
HCPP	Endogenous	0.1075	0.6974

Our model holds a three **exogenous variable** , and has five endogenous variables. Each endogenous variable is explained by one or more variables and an error term. This model has five equations that were tested using the PLS approach through the XL-stat software (2017 version). The structural equations of the conceptual model are presented as follows:

$$PN = -0.1912 * PK + 0.27347 * PA + 0.37306 * PP \quad (1)$$

$$FAP = 0.10956 * PK + 0.24190 * PA - 0.00902 * PP \quad (2)$$

$$FAHCP = 0.17758 * PN \quad (3)$$

$$HCPPER = 0.06972 * PK + 0.28723 * PA + 0.13345 * FAP \quad (4)$$

$$HCPP = 0.07949 * PP + 0.03880 * FAHCP + 0.30475 * HCPPER \quad (5)$$

4.5. Hypothesis Test:

For each causal relationship, we have advanced a derived hypothesis and since we have 13 causal relationships, we have put in place 13 derived hypotheses. This assumption will also be subject to confirmation tests (Table 9).

Table 9: Research hypothesis tests

Causal relationship	Path coefficient	T*Student	Effect size	Signifi-cation	Validity	Association Degree
H1: PK > > PN	-0.194	-2.115	0.0255	0.0362	**Yes**	Small
H2:PA > > PN	0.2735	2.8834	0.0475	0.0044	**Yes**	Medium
H3: PP > >PN	0.3731	5.1062	0.159	00000	**Yes**	Strong
H4:PK >> >FAP	0.1098	1.1271	0.0073	0.2612	**Invalid**	Small
H5:PA >> >FAP	0.2419	2.3786	0.323	0.0185	Yes	Strong
H6: PP > >FAP	-0.0902	-0.1151	0.0001	0.9085	**Invalid**	Very weak
H7:PN > >FAHCP	0.1776	2.4007	0.361	0.0174	**Yes**	Strong
H8:PK > >HCPPER	0.0697	0.7390	0.0431	0.04609	**YES**	Medium
H9:PA > >HCPPER	0.2872	3.004	0.0516	0.0031	**Yes**	Small
H10: PP >> >HCPP	0.0794	1.1107	0.4570	0.0027	**Yes**	Strong
H11: FAP >HCPPER	0.13345	1.8246	0.019	0.0698	**Yes**	Small
H12: FAHCP > HCPP	0.0388	0.9333	0.216	0.0005	**Yes**	Medium
H13: HCPPER >HCPP	0.30475	4.1919	0.1004	0.0000	**Yes**	Medium

**Note: student test values are above |2.775| (|1.960|) which indicates significant parameters in the 1% (5%)*

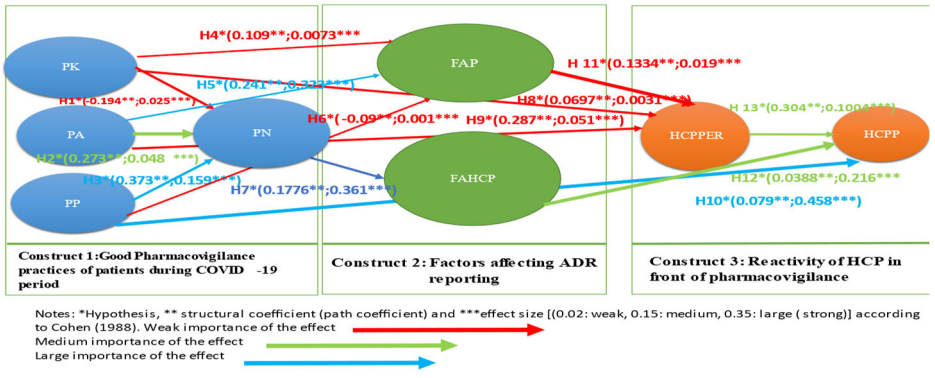

Figure 3: Final model estimated by PLS

4.6. Analysis of results:

The main goal of our empirical study is to test the impact of the good pharmacovigilance practice of patients over the COVID-19 period on the HCP's reactivity on pharmacovigilance. According to **Table 9,** we can confirm the validity of **Eleven hypotheses**: H1; H2; H3; H5; H7; H8; H9; H10; H11, and H13 (T>1.97), against two invalid derivative hypotheses H4 and H6. The final model can be represented in **Figure 3.** Based on the results in **Table 9** and **Figure 3**, we underline the following:
Direct effects: In the analysis, Patients' Knowledge negatively influences Patient's ADR notification with weak significance. Patients' attitude and Patients' practices positively influence Patient's ADR notification with strong and medium significance, respectively. Patient's attitude strongly affects factors influencing patients to report ADR. Patient's knowledge has a medium effect on HCP's perception. Patient's notification has a strong impact on factors influencing HCP to report ADR. Patient's attitude weakly influences HCP's perception to notify ADR. Patient's practices strongly affect HCP practices in pharmacovigilance. Factors influencing patient ADR reporting weakly affect HCP's perception. Factors influencing HCP reporting ADR have a medium impact on HCP's practices, and HCP's perception of reporting ADR moderately affects HCP's practices.

Table 10: Indirect effects (*Extracted from XL-stat software (2017 version))

Specification	PK	PA	PP	PN	FAP	FAHCP	HCPPER
PK							
PA	0.0000						
PP	0.0000	0.0000					
PN	0.0000	0.0000	0.0000				
FAP	0.0000	0.0000	0.0000	0.0000			
FAHCP	0.0340	0.0486	0.0662	0.0000	0.0000		
HCPPER	0.0146	0.0323	0.0012	0.0000	0.0000	0.0000	
HCPP	0.0244	0.0993	0.0022	0.0069	0.0407	0.0000	0.0000

We notice from **Table 10** that the Patient's Knowledge variable has positive and significant medium indirect effects on factors influencing HCP from reporting ADRs, "HCP's perception" and "HCP's Practices'. So, we have to improve the patient's knowledge about pharmacovigilance in order to initiate patient to report adverse drug

reaction and to maintain this culture and to improve the reactivity of HCP among reporting ADR's.

5. Conclusion

In a four-month exploratory study in the southeast region of Morocco involving HCP, it was found that spontaneous ADR reporting is an effective and low-cost method for detecting unknown AEs. The research confirmed the hypothesis that improving pharmacovigilance practices, particularly during the COVID-19 pandemic, enhances HCP reactivity in ADR reporting and reduces factors influencing reporting. However, HCP displayed varying levels of awareness and knowledge of pharmacovigilance and ADR reporting, emphasizing the need for ongoing education and training. Public awareness campaigns on ADR reporting were also recommended to boost reporting. The primary issue with spontaneous ADR reporting systems worldwide is under-reporting, acknowledged by national pharmacovigilance centers, with only 3 to 10% of ADRs being reported (OMS, 2004). Our study underscores the need for training, sustained awareness, and patient proximity to facilitate ADR reporting, suggesting awareness campaigns through social media, ADR reporting events, and maintaining post-graduate awareness for HCP via medical networks.According to this study among pharmacy student in Morocco, students expressed the desire to learn more about pharmacovigilance during their university education (N. Nchinech et al., 2020). This result led to the introduction of a system of pharmacovigilance work groups for third- and fourth-year pharmacy students for the 2018–2019 academic year. Also, from our experience in the NCPV, its proximity is of paramount importance in order to establish continuous communication with the HCP.

Acknowledgement

We are very grateful to the availability of pharmacists and doctors among the Draa-Tafilalet region southeast Morocco, thanks to Dr. NACHIT Nabil for his collaboration.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Abubakar, A. R., Simbak, N. Bin, & Haque, M., (2014). A Systematic Review of Knowledge, Attitude and Practice on Adverse Drug Reactions and Pharmacovigilance among Doctors. *Journal of Applied Pharmaceutical Science*, *4*(10), pp. 117–127. https://doi.org/10.7324/japs.2014.401021.

Accueil, (n.d.), Retrieved January 23, (2022). From: https://www.sante.gov.ma/Pages/Accueil.aspx.

Adisa, R., Omitogun, T. I., (2019). Awareness, knowledge, attitude and practice of adverse drug reaction reporting among health workers and patients in selected primary healthcare centres in Ibadan, southwestern Nigeria. *BMC Health Services Research*, *19*(1), pp. 1–14. https://doi.org/10.1186/s12913-019-4775-9.

Al Dweik, R., Stacey, D., Kohen, D., & Yaya, S., (2017). Factors affecting patient reporting of adverse drug reactions: a systematic review. *British Journal of Clinical Pharmacology*, *83*(4), pp. 875–883. https://doi.org/10.1111/bcp.13159.

Ali, M. Ben, Rifai, S., Bouksour, O., & Barrijal, S., (2019). Understanding the impact of quality practices on firm performance: insights from a structural equation modeling study of young manufacturing enterprises in Tetouan, Morocco. *International Journal of Quality and Innovation*, *4*(3/4), p. 189. https://doi.org/10.1504/ijqi.2019.105759.

Alsaleh, F. M., Alzaid, S. W., Abahussain, E. A., Bayoud, T., & Lemay, J., (2017). Knowledge, attitude and practices of pharmacovigilance and adverse drug reaction reporting among pharmacists working in secondary and tertiary governmental hospitals in Kuwait. *Saudi Pharmaceutical Journal*, *25*(6), pp. 830–837. https://doi.org/10.1016/j.jsps.2016.12.004.

Amrain, M., Bečić, F., (2014). Knowledge, perception, practices and barriers of healthcare professionals in Bosnia and Herzegovina towards adverse drug reaction reporting and pharmacovigilance. *Journal of Health Sciences*, *4*(2), pp. 120–125. https://doi.org/10.17532/jhsci.2014.183.

Assanee, J., Sorofman, B. A., Sirisinsuk, Y., & Kitisopee, T., (2021). Factors influencing patient intention to report adverse drug reaction to community pharmacists: A structural equation modeling approach. *Research in Social and Administrative Pharmacy*, *December 2020*. https://doi.org/10.1016/j.sapharm.2021.05.010.

Avery, A. J., Anderson, C., Bond, C. M., Fortnum, H., Gifford, A., Hannaford, P. C., Hazell, L., Krska, J., Lee, A. J., McLernon, D. J., Murphy, E., Shakir, S., & Watson, M. C., (2011). Evaluation of patient reporting of adverse drug reactions to the UK

"Yellow card scheme": Literature review, descriptive and qualitative analyses, and questionnaire surveys. In *Health Technology Assessment,* Vol. 15, Issue 20, pp. 1–234. https://doi.org/10.3310/hta15200.

Avkiran, N. K., Ringle, C. M., & Low, R., (2018). Monitoring transmission of systemic risk: Application of partial least squares structural equation modeling in financial stress testing. *Journal of Risk*, *20*(5), 83–115. https://doi.org/10.21314/JOR.2018.386.

Awodele, O., Akinyede, A., Adeyemi, O. A., & Awodele, D. F., (2011). Pharmacovigilance amongst doctors in private hospitals in Lagos West Senatorial District, Nigeria. *International Journal of Risk and Safety in Medicine*, *23*(4), pp. 217–226. https://doi.org/10.3233/JRS-2011-0541.

Bäckström, M., Mjörndal, T., Dahlqvist, R., & Nordkvist-Olsson, T., (2000). Attitudes to reporting adverse drug reactions in northern Sweden. *European Journal of Clinical Pharmacology*, *56*(9–10), pp. 729–732. https://doi.org/10.1007/s002280000202.

Biriell, C., Edwards, I. R., (1997). Reasons for reporting adverse drug reactions - Some thoughts based on an international review. *Pharmacoepidemiology and Drug Safety*, *6*(1), pp. 21–26. https://doi.org/10.1002/(SICI)1099-1557(199701)6:1<21::AID-PDS259>3.0.CO;2-I.

CAPM Plateforme, (n.d.), Retrieved June 24, (2022). From https://www.capmsante.ma/

Chaouali, W., Souiden, N., & Ringle, C. M., (2021). Elderly customers' reactions to service failures: the role of future time perspective, wisdom and emotional intelligence. *Journal of Services Marketing*, *35*(1), pp. 65–77. https://doi.org/10.1108/JSM-08-2019-0318.

Cheah, J. H., Sarstedt, M., Ringle, C. M., Ramayah, T., & Ting, H., (2018). Convergent validity assessment of formatively measured constructs in PLS-SEM: On using single-item versus multi-item measures in redundancy analyses. *International Journal of Contemporary Hospitality Management*, *30*(11), pp. 3192–3210. https://doi.org/10.1108/IJCHM-10-2017-0649.

Chin, W. W., (1998). The partial least squares approach for structural equation modeling. *Modern Methods for Business Research*, *April*, pp. 295–336.

Chin, W. W., Marcelin, B. L., & Newsted, P. R., (2003). A partial least squares latent variable modeling approach for measuring interaction effects: Results from a Monte Carlo simulation study and an electronic-mail emotion/adoption study. *Information Systems Research*, *14*(2). https://doi.org/10.1287/isre.14.2.189.16018.

Cho, G., Hwang, H., Sarstedt, M., & Ringle, C. M., (2020). Cutoff criteria for overall model fit indexes in generalized structured component analysis. *Journal of Marketing Analytics*, *8*(4), pp. 189–202. https://doi.org/10.1057/s41270-020-00089-1.

Ekman, E., Bäckström, M., (2009). Attitudes among hospital physicians to the reporting of adverse drug reactions in Sweden. *European Journal of Clinical Pharmacology*, *65*(1), pp. 43–46. https://doi.org/10.1007/s00228-008-0564-9.

El Hamdouni, M., Ahid, S., Bourkadi, J. E., Benamor, J., Hassar, M., & Cherrah, Y., (2020). Incidence of adverse reactions caused by first-line anti-tuberculosis drugs and treatment outcome of pulmonary tuberculosis patients in Morocco. *Infection*, *48*(1), pp. 43–50. https://doi.org/10.1007/s15010-019-01324-3.

Engla, N. E. W., Journal, N. D., (2010). *New engla nd journal*, pp. 865–869.

Faizan, A., Rasoolimanesh, M., Sarstedt, M., Ringle, C., & Ryu, K., (2018). An Assessment of The Use of Partial Least Squares Structural Equation Modeling. *International Journal of Contemporary Hospitality Management*, *34*(1), pp. 1–5. https://doi.org/10.1108/IJCHM-10-2016.

Fornell, C., Bookstein, F. L., (1982). Structural to Consumer. *Journal of Marketin Research*, *19*(4), pp. 440–452.

Guenther, P., Guenther, M., Ringle, C. M., Zaefarian, G., & Cartwright, S., (2023). Improving PLS-SEM use for business marketing research. *Industrial Marketing Management*, *111*(October 2020), pp. 127–142. https://doi.org/10.1016/j.indmarman.2023.03.010.

Hadi, M. A., Neoh, C. F., Zin, R. M., Elrggal, M., & Cheema, E., (2017). Pharmacovigilance: pharmacists’ perspective on spontaneous adverse drug reaction reporting. *Integrated Pharmacy Research and Practice*, *Vol. 6*, pp. 91–98. https://doi.org/10.2147/iprp.s105881.

Hagedoorn, J., Schakenraad, J. O. S., (1994). The Effect of Strategic Technology Alliances on Company Performance Author (s): John Hagedoorn and Jos Schakenraad Published by : Wiley Stable URL : http://www.jstor.org/stable/2486887 REFERENCES Linked references are available on JSTOR for this article. *Strategic Management Journal*, *15*(4), pp. 291–309.

Hair carole l. Hollingswoth, Chong, Jeo, A. B. R. A., (2017). Industrial Management & Data Systems. *Industrial Management & Data Systems Business Process Management Journal Iss Management Decision*, *110*(5), pp. 111–133. http://dx.doi.org/10.1108/02635571011008434%5Cnhttp://%5Cnhttp://dx.doi.org/10.1108/00251741211194903%5Cnhttp://dx.doi.org/10.1108/10878571111161507.

Hair, Joe F., Ringle, C. M., & Sarstedt, M., (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice*, *19*(2), pp. 139–152. https://doi.org/10.2753/MTP1069-6679190202.

Hair, Joseph F., Risher, J. J., Sarstedt, M., & Ringle, C. M., (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, *31*(1), pp. 2–24. https://doi.org/10.1108/EBR-11-2018-0203.

Härmark, L., Raine, J., Leufkens, H., Edwards, I. R., Moretti, U., Sarinic, V. M., & Kant, A., (2016). Patient-Reported Safety Information: A Renaissance of Pharmacovigilance? *Drug Safety*, *39*(10), pp. 883–890. https://doi.org/10.1007/ s40264-016-0441-x.

Hazell, L., Cornelius, V., Hannaford, P., Shakir, S., & Avery, A. J., (2013). How do patients contribute to signal detection?: A retrospective analysis of spontaneous reporting of adverse drug reactions in the UK's yellow card scheme. *Drug Safety*, *36*(3), pp. 199–206. https://doi.org/10.1007/s40264-013-0021-2.

Hazell, L., Shakir, S. A. W. (2006). Under-reporting of adverse drug reactions: A systematic review. *Drug Safety*, *29*(5), pp. 385–396. https://doi.org/10.2165/00002018-200629050-00003.

Hulland, J., (1999). of Four Recent Studies. *Strategic Management Journal*, *20*(2), pp. 195–204.

Inácio, P., Cavaco, A., & Airaksinen, M., (2017). The value of patient reporting to the pharmacovigilance system: a systematic review. *British Journal of Clinical Pharmacology*, *83*(2), pp. 227–246. https://doi.org/10.1111/bcp.13098.

Jöreskog, (1978). Structural Analysis of Covariance and Correlation Matrices Karl. *NEC Technical Journal*, *3*(2), pp. 27–32.

Joubert, M. C., Naidoo, P., (2016). Knowledge, perceptions and practices of pharmacovigilance amongst community and hospital pharmacists in a selected district of North West Province, South Africa. In *Health SA Gesondheid,* Vol. 21, pp. 238–244. https://doi.org/10.1016/j.hsag.2016.04.005.

Kaufmann, L., Gaeckler, J., (2015). A structured review of partial least squares in supply chain management research. *Journal of Purchasing and Supply Management*, *21*(4), pp. 259–272. https://doi.org/10.1016/j.pursup.2015.04.005.

Kitabayashi, A., Inoue, Y., (2022). Factors that Lead to Stagnation in Direct Patient Reporting of Adverse Drug Reactions: An Opinion Survey of the General Public and Physicians in Japan. *Therapeutic Innovation and Regulatory Science*, *56*(4), pp. 616–624. https://doi.org/10.1007/s43441-022-00397-x.

Kline, R. B., (1999). Book Review: Psychometric theory (3rd ed.). *Journal of Psychoeducational Assessment*, *17*(3), pp. 275–280. https://doi.org/10.1177/073428299901700307.

Kroonenberg, P. M., Lohmoller, J.-B., (1990). Latent Variable Path Modeling with Partial Least Squares. In *Journal of the American Statistical Association,* Vol. 85, Issue 411. https://doi.org/10.2307/2290049.

Law, L., Fong, N., (2020). Applying partial least squares structural equation modeling (PLS-SEM) in an investigation of undergraduate students' learning transfer of academic English. *Journal of English for Academic Purposes*, *46*, 100884. https://doi.org/10.1016/j.jeap.2020.100884.

Liu, Y. X., (2012). A new antihypertensive drug ameliorate insulin resistance. *Acta Pharmacologica Sinica*, *33*(4), pp. 429–430. https://doi.org/10.1038/aps.2012.31.

Magno, F., Cassia, F., & Ringle, C. M. M., (2022). A brief review of partial least squares structural equation modeling (PLS-SEM) use in quality management studies. *TQM Journal*. https://doi.org/10.1108/TQM-06-2022-0197.

Maguire, A., Douglas, I., Smeeth, L., & Thompson, M., (2007). Determinants of cholesterol and triglycerides recording in patients treated with lipid lowering therapy in UK primary care. *Pharmacoepidemiology and Drug Safety*, *16*(January), pp. 228–228. https://doi.org/10.1002/pds.

Margraff, F., Bertram, D., (2014). Adverse drug reaction reporting by patients: An overview of fifty countries. *Drug Safety*, *37*(6), pp. 409–419. https://doi.org/10.1007/s40264-014-0162-y.

MERLEN, C., (2017). *Gestion et maîtrise des risques des procédés de fabrication : le cas de l'industriepharmaceutiqueSous*. https://pepite-depot.univlille2.fr/nuxeo/site/esupversions/5443012c-4c8e-4c4d-bd12-1832dc0ff279.

Mirosevic Skvrce, N., Galic, I., Pacadi, C., Kandzija, N., & Mucalo, I., (2020). Adverse drug reactions that arise from the use of medicinal products outside the terms of the marketing authorisation. *Research in Social and Administrative Pharmacy*, *16*(7), pp. 928–934. https://doi.org/10.1016/j.sapharm.2019.10.003.

Montané, E., Arellano, A. L., Sanz, Y., Roca, J., & Farré, M., (2018). Drug-related deaths in hospital inpatients: A retrospective cohort study. *British Journal of Clinical Pharmacology*, *84*(3), pp. 542–552. https://doi.org/10.1111/bcp.13471.

Nchinech, N., Lachhab, Z., Obtel, M., Cherrah, Y., & Serragui, S., (2020). Moroccan pharmacy students' knowledge and perceptions about pharmacovigilance. *Annales Pharmaceutiques Francaises*. https://doi.org/10.1016/j.pharma.2020.10.005.

Nchinech, N, Lachhab, Z., Cherrah, Y., & Serragui, S., (2019). *6ER-023 Establishment of group work: what is the effect on the state of knowledge and perception of pharmacovigilance among our future moroccan pharmacists? March 2019*, A287.2-A288. https://doi.org/10.1136/ejhpharm-2019-eahpconf.620.

Nchinech, Naoual, Lachhab, Z., Obtel, M., Cherrah, Y., & Serragui, S., (2020). Connaissances et perception de la pharmacovigilance par les futurs pharmaciens marocains. *Annales Pharmaceutiques Françaises*. https://doi.org/10.1016/j.pharma.2020.10.005.

O'Callaghan, J., Griffin, B. T., Morris, J. M., & Bermingham, M., (2018). Knowledge of Adverse Drug Reaction Reporting and the Pharmacovigilance of Biological Medicines: A Survey of Healthcare Professionals in Ireland. *BioDrugs*, *32*(3), pp. 267–280. https://doi.org/10.1007/s40259-018-0281-6.

OMS, (2004). *Pharmacovigilance : assurer la sécurité d'emploi des médicaments*, pp. 1–6.

Rabba, A. K., Ain, M. R. (2015). Pharmacovigilance study: Exploring the role of community pharmacists in adverse drug reactions reporting in Alkharj city, Saudi Arabia. *Latin American Journal of Pharmacy*, *34*(5), 901–906.

Regulation and Prequalification, (n.d.), Retrieved January 22, (2022). From https://www.who.int/teams/regulation-prequalification/regulation-and-safety/pharmacovigilance.

Richter, N. F., Sinkovics, R. R., Ringle, C. M., & Schlägel, C., (2016). A critical look at the use of SEM in international business research. *International Marketing Review*, *33*(3), 376–404. https://doi.org/10.1108/IMR-04-2014-0148.

Robert, B., Brown, E. B., (2004). *A study of willigness to use chines hospital pharmacovigilance system based on structural equation modeling, 1*, pp. 1–14.

Ronald E. Walpole, Raymond H. Myers, S. L. M. Y. K. Y., (2012). Title. *BMC Public Health*, *5*(1), pp. 1–8.

Routledge, P., (1998). 150 Years of Pharmacovigilance. *Lancet*, *351*(9110), pp. 1200–1201. https://doi.org/10.1016/S0140-6736(98)03148-1.

Sahaf, K., Ali, M. Ben, Rifai, S., Bouksour, O., & Adri, A., (2018). Improvement of the hospital supply chain and its impact on reduction of patient waiting times. Case of the Oncology department of University Hospital IBN ROCHD. *2018 5th International Conference on Control, Decision and Information Technologies, CoDIT 2018*, pp. 222–227. https://doi.org/10.1109/CoDIT.2018.8394876.

Sales, I., Aljadhey, H., Albogami, Y., & Mahmoud, M. A., (2017). Public awareness and perception toward Adverse Drug Reactions reporting in Riyadh, Saudi Arabia. *Saudi Pharmaceutical Journal*, *25*(6), pp. 868–872. https://doi.org/10.1016/j.jsps.2017.01.004.

Sandberg, A., Salminen, V., Heinonen, S., & Sivén, M., (2022). Under-Reporting of Adverse Drug Reactions in Finland and Healthcare Professionals' Perspectives on How to Improve Reporting. *Healthcare (Switzerland)*, *10*(6). https://doi.org/10.3390/healthcare10061015.

Sebtaoui, F. E., Adri, A., Rifai, S., & Sahaf, K., (2020). How will the risk management impact the success of just-in-time implementation? *Journal of Industrial and Production Engineering*, *37*(7), pp. 333–344. https://doi.org/10.1080/21681015.2020.1806121.

Sosik, J. J., Kahai, S. S., & Piovoso, M. J., (2009). Silver bullet or voodoo statistics?: A primer for using the partial least squares data analytic technique in group and organization research. *Group and Organization Management*, *34*(1), pp. 5–36. https://doi.org/10.1177/1059601108329198.

Starfield, B., (2000). Is US Health Really the Best. *Jama*, *284*(4), 483. http://jama.ama-assn.org/cgi/content/abstract/284/4/483.

Stergiopoulos, S., Brown, C. A., Felix, T., Grampp, G., & Getz, K. A., (2016). A Survey of Adverse Event Reporting Practices Among US Healthcare Professionals. *Drug Safety*, *39*(11), 1117–1127. https://doi.org/10.1007/s40264-016-0455-4.

Thomas, D., Zachariah, S., (2018). Knowledge, Attitude, and Practice of Pharmacovigilance in Developing Countries. In *Social and Administrative Aspects of Pharmacy in Low-and Middle-Income Countries: Present Challenges and Future Solutions*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-811228-1.00011-X.

Toklu, H. Z., Uysal, M. K., (2008). The knowledge and attitude of the Turkish community pharmacists toward pharmacovigilance in the Kadikoy district of Istanbul. *Pharmacy World and Science*, *30*(5), pp. 556–562. https://doi.org/10.1007/s11096-008-9209-4.

Vaismoradi, Logan, Jordan, & Sletvold, (2019). Adverse Drug Reactions in Norway: A Systematic Review. *Pharmacy*, *7*(3), 102. https://doi.org/10.3390/pharmacy7030102.

Weigmann, K., (2016). Consumer reporting of adverse drug reactions. *EMBO Reports*, *17*(7), 949–952. https://doi.org/10.15252/embr.201642616.

Wetzels, M., Odekerken-Schröder, G., & Oppen, C. Van, (2009). Assessing Using PLS Path Modeling Hierarchical and Empirical Construct Models : Guidelines. *MIS Quarterly*, *33*(1).

Wold, H., (1985). Systems Analysis by Partial Least Squares. *Measuring the Unmeasurable*, *September*, pp. 221–251. https://doi.org/10.1007/978-94-009-5079-5_11.

Yarrow, A., (1961). Smoking By Schoolchildren. *The Lancet*, *278*(7216), p. 1358. https://doi.org/10.1016/S0140-6736(61)90926-6.

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 89–108, https://doi.org/10.59170/stattrans-2023-065
Received – 03.08.2021; accepted – 09.03.2023

Analysis for the xgamma distribution based on record values and inter-record times with application to prediction of rainfall and COVID-19 records

Zahra Khoshkhoo Amiri[1], **S. M. T. K. MirMostafaee**[2]

ABSTRACT

Recently, Sen et al. (2016) introduced a new lifetime distribution, called "xgamma distribution", which can be used as an alternative to other lifetime distributions, like the exponential one. In this paper, we study the problem of classical and Bayesian estimation of the unknown parameter of the xgamma distribution based on record values and inter-record times. The problem of Bayesian prediction of future record values based on record values and inter-record times is also discussed. A small simulation study has been performed to compare the performance of the proposed estimators and the approximate Bayes predictors. Two real data sets related to rainfall and COVID-19 records have been analysed. We considered four one-parameter lifetime distributions as the base models for each data set and compared the goodness-of-fit results. Then, the numerical results of estimation of the parameter and prediction of future records based on the xgamma and exponential records and inter-record times were presented. We observed that the record values and inter-record times from the xgamma distribution could predict future records in a relatively satisfactory way.

Key words: COVID-19 records, lower record values, Bayes predictive distribution, rainfall records, xgamma distribution.

1. Introduction

The xgamma distribution was first introduced by Sen et al. (2016) and its probability density function (PDF) is given by

$$f(x;\theta) = \frac{\theta^2}{1+\theta}\left(1+\frac{\theta}{2}x^2\right)\mathrm{e}^{-\theta x}, \quad x>0, \quad \theta>0. \tag{1}$$

The corresponding cumulative distribution function (CDF) is given by

$$F(x;\theta) = 1-\frac{1+\theta+\theta x+\frac{\theta^2x^2}{2}}{1+\theta}\mathrm{e}^{-\theta x}, \quad x>0, \quad \theta>0.$$

If the PDF of a random variable X is expressed by (1), then we write $X \sim \text{xgamma}(\theta)$. Indeed, the xgamma distribution is a special mixture of the exponential and gamma distributions. The hazard rate function (HRF) of the xgamma distribution can be bathtub-shaped,

[1]Department of Statistics, University of Mazandaran, Babolsar, Iran. E-mail: z.khoshkhoo@stu.umz.ac.ir. ORCID: https://orcid.org/0000-0001-8838-5870.
[2]Department of Statistics, University of Mazandaran, Babolsar, Iran. E-mail: m.mirmostafaee@umz.ac.ir. ORCID: https://orcid.org/0000-0003-2796-4427.

which makes it very suitable for many real lifetime phenomena. In recent years, several studies have been carried out on the inferential problems pertaining to the xgamma distribution; see for example Sen et al. (2018) and Yadav et al. (2019).

Let $\{X_n, n = 1, 2, \cdots\}$ be a sequence of identical and independent random variables. An observation X_j is called a lower record value if $X_j < X_i$ for all $i < j$. A similar definition can be given for upper record values. The sequence of lower record values along with the inter-record times can be given by $(\mathbf{R}, \mathbf{K}) = \{R_1, K_1, R_2, K_2, \cdots, R_{m-1}, K_{m-1}, R_m\}$ where R_i is the i-th record value and K_i is the i-th inter-record time, which is the number of observations after occurrence of R_i that are needed to obtain a new record value R_{i+1}. Record data arise in a wide variety of practical situations; see for example Arnold et al. (1998). Record values and the related subjects have been studied by many authors; see for example Ahmadi and MirMostafaee (2009), MirMostafaee et al. (2016) and Fallah et al. (2018). Record values, along with inter-record times, have become a favourite subject for many researchers in recent decades. Samaniego and Whitaker (1986) discussed the estimation problem of the mean parameter of the exponential distribution based on records and inter-record times. For other examples of recent studies in this regard, see Nadar and Kızılaslan (2015), Kızılaslan and Nadar (2016), Amini and MirMostafaee (2016), Pak and Dey (2019), Kumar et al. (2020) and Bastan and MirMostafaee (2022).

In this paper, first, we obtain the maximum likelihood (ML) estimate of the unknown parameter of the xgamma distribution based on lower record values and inter-record times, and then we construct an asymptotic confidence interval (ACI) for the xgamma parameter in Section 2. Next, we work on the Bayesian estimation of the parameter in Section 3. We approximate the Bayes estimates with the help of the Tierney and Kadane (TK) method, importance sampling (IS) method and Metropolis-Hastings (M-H) algorithm. We also discuss the Bayesian prediction problem of future record values arising from the xgamma distribution based on observed lower record values and inter-record times in Section 4. In Section 5, a small simulation study is conducted to compare the performances of the proposed estimators and approximate Bayes predictors. We analyse two real data sets that are related to rainfall and COVID-2019 phenomena. Section 6 concludes the paper with some remarks.

2. Maximum Likelihood Estimation

In this section, we focus on the ML estimate and an ACI for the parameter. Let $\{R_1, K_1, R_2, K_2, \cdots, R_{m-1}, K_{m-1}, R_m\}$ be a sequence of record data from xgamma(θ). Then, the likelihood function of θ given the observed lower records and inter-record times becomes

$$L(\theta|\mathbf{r},\mathbf{k}) = \prod_{i=1}^{m} f(r_i;\theta)\big[1-F(r_i;\theta)\big]^{k_i-1} = \frac{\mathrm{e}^{-\theta\sum_{i=1}^{m}k_i r_i}\theta^{2m}}{(1+\theta)^{\sum_{i=1}^{m}k_i}} \prod_{i=1}^{m}\left(1+\frac{\theta}{2}r_i^2\right)\big[\psi(\theta,r_i)\big]^{k_i-1}, \quad (2)$$

where $r_1 > \cdots > r_m$, $k_m = 1$, $\psi(\theta, r_i) = 1 + \theta + \theta r_i + \dfrac{\theta^2 r_i^2}{2}$ and $\mathbf{r} = \{r_1, \cdots, r_m\}$ and $\mathbf{k} = \{k_1, \cdots, k_{m-1}\}$ are the observed sets of $\mathbf{R} = \{R_1, \cdots, R_{m-1}, R_m\}$ and $\mathbf{K} = \{K_1, \cdots, K_{m-1}\}$, respectively. Note that we always take k_m equal to one for simplicity of the equations.

Consequently, the corresponding log-likelihood function is

$$\ell(\theta|\mathbf{r},\mathbf{k}) = 2m\ln(\theta) - \sum_{i=1}^{m} k_i \ln(1+\theta) - \theta\sum_{i=1}^{m} k_i r_i + \sum_{i=1}^{m} \ln\left(1+\frac{\theta r_i^2}{2}\right) + \sum_{i=1}^{m}(k_i-1)\ln\left(\psi(\theta,r_i)\right).$$

We take the first partial derivative of log-likelihood function with respect to (w.r.t.) θ and then equate it with zero. Thus, we have

$$\frac{\partial \ell(\theta|\mathbf{r},\mathbf{k})}{\partial\theta} = \frac{2m}{\theta} - \sum_{i=1}^{m}\frac{k_i}{1+\theta} - \sum_{i=1}^{m} k_i r_i + \sum_{i=1}^{m}\frac{r_i^2}{2+\theta r_i^2} + \sum_{i=1}^{m}\frac{(k_i-1)\left(1+r_i+\theta r_i^2\right)}{\psi(\theta,r_i)} = 0.$$

It seems that no explicit solution for the above equation exists, and we may use numerical techniques to calculate the ML estimate of θ.
Next, we aim at finding an ACI for the parameter θ. Here, the Fisher information is defined by $I(\theta) = -E\left\{\frac{\partial^2 \ln f_\theta(\mathbf{R},\mathbf{K})}{\partial\theta^2}\right\}$, where $f_\theta(\mathbf{r},\mathbf{k})$ is the joint probability function of $R_1, K_1, R_2, K_2, \cdots, R_{m-1}, K_{m-1}, R_m$, provided that the related integral exists. We have

$$\frac{\partial^2 \ell(\theta|\mathbf{r},\mathbf{k})}{\partial\theta^2} = -\frac{2m}{\theta^2} + \sum_{i=1}^{m}\frac{k_i}{(1+\theta)^2} - \sum_{i=1}^{m}\frac{r_i^4}{(2+\theta r_i^2)^2} + \sum_{i=1}^{m}\frac{(k_i-1)\left(\psi(\theta,r_i)r_i^2 - (1+r_i+\theta r_i^2)^2\right)}{\left[\psi(\theta,r_i)\right]^2}.$$

Let $\widehat{\theta}_{ML}$ denote the ML estimator (MLE) of θ and z_γ be the γ-th upper quantile of the standard normal distribution. Then, the $100(1-\alpha)\%$ modified asymptotic two-sided equi-tailed confidence interval (MATE CI) for θ is given by (see for example Lehmann and Casella, 1998)

$$\left(\max\left\{0, \widehat{\theta}_{ML} - \frac{z_{\frac{\alpha}{2}}}{\sqrt{\tilde{I}(\widehat{\theta}_{ML})}}\right\}, \widehat{\theta}_{ML} + \frac{z_{\frac{\alpha}{2}}}{\sqrt{\tilde{I}(\widehat{\theta}_{ML})}}\right),$$

where $\tilde{I}(\widehat{\theta}_{ML}) = -\frac{\partial^2 \ell(\theta|\mathbf{R},\mathbf{K})}{\partial\theta^2}\Big|_{\theta=\widehat{\theta}_{ML}}$.

3. Bayesian Estimation

In the context of Bayesian estimation, the information of the experimenter can be revealed in the form of a probability function for the parameter, which is called the prior distribution. Since the parameter of the xgamma distribution is positive, we consider the popular gamma prior for θ with the following PDF

$$\pi(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}\mathrm{e}^{-b\theta}, \qquad \theta > 0, \tag{3}$$

where a and b are positive hyperparameters that can be determined by the prior knowledge of the experimenter. From (2) and (3), the posterior density can be obtained as

$$\pi(\theta|\mathbf{r},\mathbf{k}) = \frac{1}{D}\frac{\theta^{2m+a-1}}{(1+\theta)^{\Sigma_{i=1}^m k_i}}\mathrm{e}^{-\theta(\Sigma_{i=1}^m k_i r_i + b)}\prod_{i=1}^{m}\left(1+\frac{\theta}{2}r_i^2\right)\left[\psi(\theta,r_i)\right]^{k_i-1},$$

where $D=\int_0^\infty \frac{\theta^{2m+a-1}}{(1+\theta)^{\sum_{i=1}^m k_i}} e^{-\theta(\sum_{i=1}^m k_i r_i+b)} \prod_{i=1}^m \left(1+\frac{\theta}{2}r_i^2\right)\left[\psi(\theta,r_i)\right]^{k_i-1} d\theta$. A very popular quadratic loss function is the squared error loss function (SELF). However, the SELF is not appropriate in many real situations, as it gives identical weights to underestimation and overestimation. One asymmetric loss function is the linear-exponential loss function (LELF), which was introduced by Varian (1975) and is defined by

$$L_{LE}(\theta,\hat{\theta}) = b\left[\exp\{c(\hat{\theta}-\theta)\}-c(\hat{\theta}-\theta)-1\right], \quad b>0, \quad c\neq 0,$$

where $\hat{\theta}$ is an estimator of θ. Without loss of generality, we assume $b=1$. The sign and magnitude of parameter c must be properly determined. If c is bigger than zero, then overestimation is more serious than underestimation and vice versa (Zellner, 1986). The Bayes estimates of θ under the SELF and LELF, become

$$\widehat{\theta}_{SE} = \int_0^\infty \theta\pi(\theta|\mathbf{r},\mathbf{k})d\theta = \frac{1}{D}\int_0^\infty \frac{\theta^{2m+a}}{(1+\theta)^{\sum_{i=1}^m k_i}} e^{-\theta(\sum_{i=1}^m k_i r_i+b)} \prod_{i=1}^m \left(1+\frac{\theta}{2}r_i^2\right)\left[\psi(\theta,r_i)\right]^{k_i-1} d\theta,$$

and

$$\begin{aligned}\widehat{\theta}_{LE} &= -\frac{1}{c}\ln M_\theta(-c|\mathbf{r},\mathbf{k}) = -\frac{1}{c}\ln\left(\int_0^\infty \exp(-c\theta)\,\pi(\theta\mid\mathbf{r},\mathbf{k})d\theta\right)\\ &= -\frac{1}{c}\ln\left(\frac{1}{D}\int_0^\infty \frac{\theta^{2m+a-1}}{(1+\theta)^{\sum_{i=1}^m k_i}} e^{-\theta(\sum_{i=1}^m k_i r_i+b+c)} \prod_{i=1}^m \left(1+\frac{\theta}{2}r_i^2\right)\left[\psi(\theta,r_i)\right]^{k_i-1} d\theta\right),\end{aligned}$$

respectively, provided that the integrals exist.

It seems that the above Bayes estimates of θ cannot be obtained in closed forms. So, we use three methods to find the approximate Bayes estimates of parameter θ.

3.1. Tierney and Kadane's Approximation

For a one-parameter model, Tierney and Kadane (1986) proposed a technique to approximate Bayes estimates. Let $\nu_0(\theta)=\frac{1}{n}\ln\pi(\theta\mid\mathbf{r},\mathbf{k})$ and $\nu^*(\theta)=\nu_0(\theta)+\frac{1}{n}\ln g(\theta)$. Then, according to the TK approximation method, the approximated Bayes estimate of θ is

$$\widehat{\theta}_{BT} = \sqrt{\frac{\tau^*}{\tau_0}}\exp\left\{n\left[\nu^*(\theta^*)-\nu_0(\theta_0)\right]\right\},$$

where θ^* and θ_0 maximize $\nu^*(\theta)$ and $\nu_0(\theta)$, respectively, and τ^* and τ_0 are minus the inverse of the second derivatives of $\nu^*(\theta)$ and $\nu_0(\theta)$ at the points θ^* and θ_0, respectively.

We have

$$\begin{aligned}\nu_0(\theta) = \frac{1}{n}\Bigg[&-\ln D+(2m+a-1)\ln\theta-\sum_{i=1}^m k_i\ln(1+\theta)-\theta\Big(\sum_{i=1}^m k_i r_i+b\Big)+\sum_{i=1}^m \ln\left(1+\frac{\theta}{2}r_i^2\right)\\ &+\sum_{i=1}^m (k_i-1)\ln\big(\psi(\theta,r_i)\big)\Bigg].\end{aligned}$$

Therefore, θ_0 can be derived from the following equation:

$$\frac{\partial v_0(\theta)}{\partial \theta} = \frac{1}{n}\left[\frac{2m+a-1}{\theta} - \frac{\sum_{i=1}^m k_i}{1+\theta} - \sum_{i=1}^m k_i r_i - b + \sum_{i=1}^m \frac{r_i^2}{2+\theta r_i^2} + \sum_{i=1}^m \frac{(k_i-1)(1+r_i+\theta r_i^2)}{\psi(\theta, r_i)}\right] = 0.$$

Let ξ_0 be the second order derivative of $v_0(\theta)$ at θ_0, namely

$$\xi_0 = \frac{\partial^2}{\partial \theta^2} v_0(\theta)\Big|_{\theta=\theta_0} = \frac{1}{n}\left[\frac{-(2m+a-1)}{\theta^2} + \frac{\sum_{i=1}^m k_i}{(1+\theta)^2} - \sum_{i=1}^m \frac{r_i^4}{(2+\theta r_i^2)^2} + \sum_{i=1}^m \frac{(k_i-1)\left[\psi(\theta, r_i) r_i^2 - (1+r_i+\theta r_i^2)^2\right]}{\left[\psi(\theta, r_i)\right]^2}\right]\Bigg|_{\theta=\theta_0}.$$

Then, set $\tau_0 = -\frac{1}{\xi_0}$. First, we derive the approximate Bayes estimate of θ under the SELF. Let $g(\theta) = \theta$ and θ_1^* be the maximum point of the following quantity:

$$v^{*SE}(\theta) = \frac{1}{n}\left[-\ln D + (2m+a)\ln\theta - \sum_{i=1}^m k_i \ln(1+\theta) - \theta\left(\sum_{i=1}^m k_i r_i + b\right) + \sum_{i=1}^m \ln\left(1+\frac{\theta}{2} r_i^2\right) + \sum_{i=1}^m (k_i - 1)\ln\left(\psi(\theta, r_i)\right)\right].$$

Then, θ_1^* can be derived from the following equation:

$$\frac{\partial v^{*SE}(\theta)}{\partial \theta} = \frac{1}{n}\left[\frac{2m+a}{\theta} - \frac{\sum_{i=1}^m k_i}{1+\theta} - \sum_{i=1}^m k_i r_i - b + \sum_{i=1}^m \frac{r_i^2}{2+\theta r_i^2} + \sum_{i=1}^m \frac{(k_i-1)(1+r_i+\theta r_i^2)}{\psi(\theta, r_i)}\right] = 0.$$

Let ξ_{SE}^* be the second order derivative of $v^{*SE}(\theta)$ at θ_1^*, namely

$$\xi_{SE}^* = \frac{\partial^2}{\partial \theta^2} v^{*SE}(\theta)\Big|_{\theta=\theta_1^*} = \frac{1}{n}\left[\frac{-(2m+a)}{\theta^2} + \frac{\sum_{i=1}^m k_i}{(1+\theta)^2} - \sum_{i=1}^m \frac{r_i^4}{(2+\theta r_i^2)^2} + \sum_{i=1}^m \frac{(k_i-1)\left[\psi(\theta, r_i) r_i^2 - \left(1+r_i+\theta r_i^2\right)^2\right]}{\left[\psi(\theta, r_i)\right]^2}\right]\Bigg|_{\theta=\theta_1^*}.$$

Then, set $\tau_{SE}^* = -\frac{1}{\xi_{SE}^*}$. So, the approximate Bayes estimate of θ under the SELF becomes

$$\widehat{\theta}_{ST} = \sqrt{\frac{\tau_{SE}^*}{\tau_0}} \exp\left\{n\left[v^{*SE}(\theta_1^*) - v_0(\theta_0)\right]\right\}.$$

Next, consider the LELF and let $g(\theta) = \mathrm{e}^{-c\theta}$. Then, we have

$$v^{*LE}(\theta) = \frac{1}{n}\left[-\ln D + (2m+a-1)\ln\theta - \sum_{i=1}^m k_i \ln(1+\theta) - \theta\left(\sum_{i=1}^m k_i r_i + b + c\right) + \sum_{i=1}^m \ln\left(1+\frac{\theta}{2} r_i^2\right) + \sum_{i=1}^m (k_i - 1)\ln\left(\psi(\theta, r_i)\right)\right].$$

The maximum point of $\nu^{*LE}(\theta)$, denoted by θ_2^*, can be derived from the following equation:

$$\frac{\partial}{\partial\theta}\nu^{*LE}(\theta)=\frac{1}{n}\left[\frac{2m+a-1}{\theta}-\frac{\sum_{i=1}^m k_i}{1+\theta}-\sum_{i=1}^m k_i r_i-b-c+\sum_{i=1}^m\frac{r_i^2}{2+\theta r_i^2}\right.$$
$$\left.+\sum_{i=1}^m\frac{(k_i-1)(1+r_i+\theta r_i^2)}{\psi(\theta,r_i)}\right]=0.$$

Let ξ_{LE}^* be the second order derivative of $\nu^{*LE}(\theta)$ at θ_2^*, namely

$$\xi_{LE}^*=\frac{\partial^2}{\partial\theta^2}\nu^{*LE}(\theta)\Big|_{\theta=\theta_2^*}=\frac{1}{n}\left[\frac{-(2m+a-1)}{\theta^2}+\frac{\sum_{i=1}^m k_i}{(1+\theta)^2}-\sum_{i=1}^m\frac{r_i^4}{(2+\theta r_i^2)^2}\right.$$
$$\left.+\sum_{i=1}^m\frac{(k_i-1)\left[\psi(\theta,r_i)r_i^2-(1+r_i+\theta r_i^2)^2\right]}{\left[\psi(\theta,r_i)\right]^2}\right]\Big|_{\theta=\theta_2^*}.$$

Then, set $\tau_{LE}^*=-\frac{1}{\xi_{LE}^*}$ and the approximate Bayes estimate of θ under the LELF becomes

$$\widehat{\theta}_{LT}=-\frac{1}{c}\ln\left(\sqrt{\frac{\tau_{LE}^*}{\tau_0}}\exp\left\{n\left[\nu^{*LE}(\theta_2^*)-\nu_0(\theta_0)\right]\right\}\right).$$

3.2. Importance Sampling Method

Another well-known method of approximating Bayes point estimates is the importance sampling method; see for example Albert (2009, Section 5.9). The posterior density function of θ given $\mathbf{r}$ and $\mathbf{k}$, can be rewritten as

$$\pi(\theta|\mathbf{r},\mathbf{k})=g(\theta|\mathbf{r},\mathbf{k})h(\theta,\mathbf{r},\mathbf{k})=\frac{\theta^{2m+a-1}}{D(1+\theta)^{\sum_{i=1}^m k_i}}\mathrm{e}^{-\theta(\sum_{i=1}^m k_i r_i+b)}\prod_{i=1}^m\left(1+\frac{\theta}{2}r_i^2\right)\left[\psi(\theta,r_i)\right]^{k_i-1},$$

where $g(\theta|\mathbf{r},\mathbf{k})$ is the gamma density with parameters $a+2m$ and $\sum_{i=1}^m k_i r_i+b$ and

$$h(\theta,\mathbf{r},\mathbf{k})=\frac{\Gamma(a+2m)}{D(\sum_{i=1}^m k_i r_i+b)^{a+2m}(1+\theta)^{\sum_{i=1}^m k_i}}\prod_{i=1}^m\left(1+\frac{\theta}{2}r_i^2\right)\left[\psi(\theta,r_i)\right]^{k_i-1}.$$

Algorithm 1:

Step 1. Generate θ_1 from $g(\theta|\mathbf{r},\mathbf{k})$.

Step 2. Repeat Step 1, N times to obtain $\theta_1,\ldots,\theta_N$, where N is a large number.

Step 3. The approximate Bayes estimates of θ under the SELF and LELF are given by

$$\widehat{\theta}_{SI}=\frac{\sum_{i=1}^N h(\theta_i,\mathbf{r},\mathbf{k})\theta_i}{\sum_{j=1}^N h(\theta_j,\mathbf{r},\mathbf{k})}=\sum_{i=1}^N\theta_i w_i,$$

and

$$\widehat{\theta}_{LI} = -\frac{1}{c}\ln\left(\frac{\sum_{i=1}^{N} \mathrm{e}^{-c\theta_i} h(\theta_i, \mathbf{r}, \mathbf{k})}{\sum_{j=1}^{N} h(\theta_j, \mathbf{r}, \mathbf{k})}\right) = -\frac{1}{c}\ln\left(\sum_{i=1}^{N} \exp(-c\theta_i) w_i\right),$$

respectively, where

$$w_i = \frac{h(\theta_i, \mathbf{r}, \mathbf{k})}{\sum_{j=1}^{N} h(\theta_j, \mathbf{r}, \mathbf{k})}, \quad i = 1, \cdots, N. \tag{4}$$

Let $\{\theta_1, \cdots, \theta_N\}$ be the generated sample using the IS method and $\theta_{(1)} \leq \cdots \leq \theta_{(N)}$ be the corresponding ordered values of $\{\theta_1, \cdots, \theta_N\}$. Let

$$w_i^* = \frac{h(\theta_{(i)}, \mathbf{r}, \mathbf{k})}{\sum_{j=1}^{N} h(\theta_j, \mathbf{r}, \mathbf{k})}, \quad i = 1, \cdots, N.$$

Consider the following intervals

$$L_j(N) = \left(\widehat{\theta}^{(\frac{j}{N})}, \widehat{\theta}^{(\frac{j+[(1-\alpha)N]}{N})}\right), \qquad j = 1, 2, ..., N - [(1-\alpha)N],$$

where $[x]$ denotes the integer part of x and $\widehat{\theta}^{(\alpha)} = \theta_{(i)}$ if $\sum_{j=1}^{i-1} w_j^* < \alpha \leq \sum_{j=1}^{i} w_j^*$. Then, the $100(1-\alpha)\%$ Chen and Shao shortest width credible interval (CSSW CrI) for θ is given by $L_q(N)$, where q is selected so that (Chen and Shao, 1999)

$$\widehat{\theta}^{(\frac{q+[(1-\alpha)N]}{N})} - \widehat{\theta}^{(\frac{q}{N})} = \min_{1 \leq j \leq N-[(1-\alpha)N]} \widehat{\theta}^{(\frac{j+[(1-\alpha)N]}{N})} - \widehat{\theta}^{(\frac{j}{N})}.$$

3.3. Metropolis-Hastings Method

The Metropolis-Hastings (M-H) method was originally proposed by Metropolis et al. (1953) and then was generalized by Hastings (1970). One M-H algorithm for our case can be summarized as follows.

Algorithm 2:
Step 1. Start with an initial guess $\theta_0 = \widehat{\theta}_{ML}$ and set $t = 1$.
Step 2. Given θ_{t-1}, generate θ^* from the truncated-normal distribution, $N(\theta_{t-1}, \sigma^2) I_{\{\theta > 0\}}$. Then, set $\theta_t = \theta^*$ with probability

$$P = \min\left\{\frac{\pi(\theta^* \mid \mathbf{r}, \mathbf{k}) q(\theta^{t-1} \mid \theta^*)}{\pi(\theta^{t-1} \mid \mathbf{r}, \mathbf{k}) q(\theta^* \mid \theta^{t-1})}, 1\right\},$$

where $q(x \mid b)$ is the density of $N(b, \sigma^2) I_{\{x>0\}}$, otherwise set $\theta_t = \theta_{t-1}$.
Step 3. Set $t = t + 1$ and repeat Step 2, T times, where T is a large number. Then, $\{\theta_{M+1}, \theta_{M+2}, \cdots, \theta_T\}$ is the generated sample, where M is a burn-in period. Now, the ap-

proximated Bayes point estimates of θ under the SELF and LELF are given by

$$\widehat{\theta}_{SM} = \frac{1}{M^*}\sum_{t=M+1}^{T}\theta_t, \qquad \text{and} \qquad \widehat{\theta}_{LM} = -\frac{1}{c}\ln\left(\frac{1}{M^*}\sum_{t=M+1}^{T}\mathrm{e}^{-c\theta_t}\right),$$

respectively, where $M^* = T - M$. We have taken $\sigma^2 = 1$ in Section 5.

Let $\theta_{(1)} \le \cdots \le \theta_{(M^*)}$ be the corresponding ordered values of $\{\theta_{M+1}, \cdots, \theta_T\}$. Consider the intervals $L_j(M^*) = \left(\theta_{(j)}, \theta_{(j+[(1-\alpha)M^*])}\right)$ for $j = 1, 2, ..., M^* - [(1-\alpha)M^*]$, then the $100(1-\alpha)\%$ CSSW CrI for θ can be reported as $L_q(M^*)$, where q is selected so that (Chen and Shao, 1999)

$$\theta_{(q+[(1-\alpha)M^*])} - \theta_{(q)} = \min_{1\le j\le M^*-[(1-\alpha)M^*]} \theta_{(j+[(1-\alpha)M^*])} - \theta_{(j)}.$$

4. Bayesian Prediction

Suppose that $R_1, K_1, R_2, K_2, \cdots, R_{m-1}, K_{m-1}, R_m$ are a sequence of available record data from xgamma(θ) and we wish to predict the s-th unobserved record value, denoted by $R_s (s > m)$. The conditional density function of R_s given $\mathbf{r}$ and $\mathbf{k}$ is given by

$$\begin{aligned} f(r_s|\theta,\mathbf{r},\mathbf{k}) &= \frac{f(r_s;\theta)[Q(r_s,\theta)-Q(r_m,\theta)]^{s-m-1}}{\Gamma(s-m)F(r_m;\theta)} \\ &= \frac{[Q(r_s,\theta)-Q(r_m,\theta)]^{s-m-1}}{\Gamma(s-m)}\left(1-\frac{\psi(\theta,r_m)}{(1+\theta)}\mathrm{e}^{-\theta r_m}\right)^{-1}\frac{\theta^2}{1+\theta}\left(1+\frac{\theta}{2}r_s^2\right)\mathrm{e}^{-\theta r_s}, \end{aligned} \tag{5}$$

where $0 < r_s < r_m$, $Q(r_s,\theta) = -\ln F(r_s;\theta)$ and $\psi(\theta, r_m) = 1 + \theta + \theta r_m + \dfrac{\theta^2 r_m^2}{2}$.

Then, from (5), the Bayes predictive density function of R_s is derived as

$$h(r_s \mid \mathbf{r},\mathbf{k}) = \int_0^\infty f(r_s \mid \theta,\mathbf{r},\mathbf{k})\pi(\theta \mid \mathbf{r},\mathbf{k})\mathrm{d}\theta.$$

The predictions of R_s under the SELF and LELF can be given by $\widehat{R}_s^S = \int_0^{r_m} r_s h(r_s \mid \mathbf{r},\mathbf{k})\mathrm{d}r_s$, and $\widehat{R}_s^L = -\frac{1}{c}\ln\int_0^{r_m}\mathrm{e}^{-cr_s}h(r_s \mid \mathbf{r},\mathbf{k})\mathrm{d}r_s$, respectively. It seems that the Bayes predictive density function of R_s cannot be obtained analytically. Therefore, we approximate $h(r_s \mid \mathbf{r},\mathbf{k})$ using the IS and M-H methods. Assume that $\{\theta_i, i = 1, ..., N\}$ and $\{\theta_t, t = M+1, ..., T\}$ are the generated samples using the IS and M-H procedures, respectively. Then, the estimates of $h(r_s \mid \mathbf{r},\mathbf{k})$ using these generated samples are given by

$$\widehat{h}_{IM}(r_s \mid \mathbf{r},\mathbf{k}) = \sum_{i=1}^{N} w_i f(r_s \mid \theta_i,\mathbf{r},\mathbf{k}), \qquad \text{and} \qquad \widehat{h}_{MH}(r_s \mid \mathbf{r},\mathbf{k}) = \frac{1}{M^*}\sum_{t=M+1}^{T} f(r_s \mid \theta_t,\mathbf{r},\mathbf{k}),$$

respectively, where w_i is defined in (4).

Now, using the generated sample obtained by the IS method, the approximate predic-

tions of R_s under the SELF and LELF are, respectively, given by (provided that they exist)

$$\widehat{R}_s^{SI} = \sum_{i=1}^{N} w_i \int_0^{r_m} r_s f(r_s \mid \theta_i, \mathbf{r}, \mathbf{k}) \mathrm{d}r_s, \quad \text{and} \quad \widehat{R}_s^{LI} = -\frac{1}{c} \ln \left\{ \sum_{i=1}^{N} w_i \int_0^{r_m} \mathrm{e}^{-cr_s} f(r_s | \theta_i, \mathbf{r}, \mathbf{k}) \mathrm{d}r_s \right\}.$$

Using the generated sample obtained by the M-H method, the approximate predictions of R_s under the SELF and LELF are, respectively, given by (provided that they exist)

$$\widehat{R}_s^{SM} = \frac{1}{M^*} \sum_{t=M+1}^{T} \int_0^{r_m} r_s f(r_s \mid \theta_t, \mathbf{r}, \mathbf{k}) \mathrm{d}r_s,$$

and

$$\widehat{R}_s^{LM} = -\frac{1}{c} \ln \left\{ \frac{1}{M^*} \sum_{t=M+1}^{T} \int_0^{r_m} \mathrm{e}^{-cr_s} f(r_s \mid \theta_t, \mathbf{r}, \mathbf{k}) \mathrm{d}r_s \right\}.$$

The $100(1-\alpha)\%$ Bayesian prediction interval (BPI) for R_s is given by $(L(\mathbf{R}, \mathbf{K}), U(\mathbf{R}, \mathbf{K}))$, where the prediction limits $L(\mathbf{r}, \mathbf{k})$ and $U(\mathbf{r}, \mathbf{k})$ can be obtained by solving the following nonlinear equations simultaneously (see for example Pak and Dey (2019))

$$\int_0^{L(\mathbf{r},\mathbf{k})} h(r_s \mid \mathbf{r}, \mathbf{k}) \mathrm{d}r_s = \frac{\alpha}{2}, \qquad \text{and} \qquad \int_0^{U(\mathbf{r},\mathbf{k})} h(r_s \mid \mathbf{r}, \mathbf{k}) \mathrm{d}r_s = 1 - \frac{\alpha}{2}.$$

Therefore, the $100(1-\alpha)\%$ approximate BPI (ABPI) for R_s, denoted by (L^*, U^*), based on the IS method, can be obtained by solving the following equations simultaneously:

$$\sum_{i=1}^{N} w_i \int_0^{L^*} f(r_s \mid \theta_i, \mathbf{r}, \mathbf{k}) \mathrm{d}r_s = \frac{\alpha}{2}, \qquad \text{and} \qquad \sum_{i=1}^{N} w_i \int_0^{U^*} f(r_s \mid \theta_i, \mathbf{r}, \mathbf{k}) \mathrm{d}r_s = 1 - \frac{\alpha}{2}.$$

Besides, the $100(1-\alpha)\%$ ABPI for R_s, denoted as (L^{**}, U^{**}), based on the M-H method, can be obtained by solving the following equations simultaneously (see for example AL-Hussaini and Al-Awadhi (2010)):

$$\frac{1}{M^*} \sum_{t=M+1}^{T} \int_0^{L^{**}} f(r_s \mid \theta_t, \mathbf{r}, \mathbf{k}) \mathrm{d}r_s = \frac{\alpha}{2}, \qquad \text{and} \qquad \frac{1}{M^*} \sum_{t=M+1}^{T} \int_0^{U^{**}} f(r_s \mid \theta_t, \mathbf{r}, \mathbf{k}) \mathrm{d}r_s = 1 - \frac{\alpha}{2}.$$

5. Numerical Illustration

In this section, we provide a simulation study and two real data examples.

5.1. Simulation Study

Here, we perform a Monte Carlo simulation to assess the point and interval estimators and approximate predictors that are developed in this paper. In this simulation study, the number of replications is taken to be $N^* = 5000$. We generate $(m+1)$ records and their corresponding inter-record times from xgamma(θ) in each replication, where two values are

considered for m, namely $m = 5$ and 7 and the parameter values are selected to be $\theta = 0.5$ and 1.5. In the context of the Bayesian inference, we use two priors: **Prior I**: In this prior, the hyperparameters are determined so that the prior mean equals the true value of the parameter and the prior variance equals 1. Thus, for $\theta = 0.5$, we have $(a,b) = (0.25, 0.5)$, and for $\theta = 1.5$, we have $(a,b) = (2.25, 1.5)$. **Prior II**: In this prior, the hyperparameters are determined so that the prior mean equals the true value of the parameter and the prior variance equals 100. Thus, for $\theta = 0.5$ we have $(a,b) = (0.0025, 0.005)$, and for $\theta = 1.5$ we have $(a,b) = (0.0225, 0.015)$. We compute the ML estimates and the approximate Bayes estimates under the TK, IS and M-H methods based on the generated first m records and $(m-1)$ record times. Besides, the Geweke test (see Geweke, 1992), Raftery and Lewis's diagnostic (see Raftery and Lewis, 1992, 1996) and Heidelberger and Welch's convergence diagnostic (see Heidelberger and Welch, 1983) are used to check the convergence of the generated M-H Markov chains. Note that Heidelberger and Welch (1983) used or hinted the results of Schruben et al. (1980), Heidelberger and Welch (1981a, 1981b), Schruben (1982) and Schruben et al. (1983). In some cases, we have taken every second or third sampled value (and increase the number of sampled values accordingly) to achieve a convergent M-H Markov chain. All the final chains have sizes equal to 10000. The performance of the competitive estimators has been compared in terms of their estimated risks (ERs). In addition, the average width (AW) and coverage probability (CP) criteria have been employed to evaluate the interval estimators and predictors. Let $\widehat{\theta}$ be an estimator of θ and $\widehat{\theta}_i$ be the corresponding estimate derived in the i-th replication. Then, the ERs of $\widehat{\theta}$ w.r.t. the SELF and LELF functions are, respectively, given by

$$ER_S(\widehat{\theta}) = \frac{1}{N^*}\sum_{i=1}^{N^*}(\widehat{\theta}_i - \theta)^2, \text{ and } ER_L(\widehat{\theta}) = \frac{1}{N^*}\sum_{i=1}^{N^*}\Big(\exp[c(\widehat{\theta}_i - \theta)] - c(\widehat{\theta}_i - \theta) - 1\Big). \quad (6)$$

The approximate point and interval predictions for the $(m+1)$-th record value, namely R_{m+1} are calculated as well. In the context of prediction, the evaluation is based on the estimated prediction risks (EPRs) w.r.t. to the SELF and LELF for the point predictors, which are formulated similarly to (6). The simulation results have been presented in Tables 1-3 for the point estimators and in Table 4 for the interval estimators. The results for the prediction have been presented in Tables 5-8. From Tables 1-8, we extract the following conclusions:

- The ERs of Prior I are less than those of Prior II in the most cases, as expected, since Prior I is more informative than Prior II, however, the EPRs of Prior I are very close to those of Prior II. Besides, the Bayesian credible intervals of Prior I have smaller AWs than those of Prior II, however, the AWs of the ABPIs of Prior I are very close to those of Prior II.

- The ERs of the point estimators and EPRs of the predictors are decreasing w.r.t. the number of records in the most cases, as expected. Besides, the AWs of the interval estimators and the ABPIs are decreasing w.r.t. the number of records.

Table 1: The ERs of the point estimators of θ when $\theta = 0.5$.

	Prior I			Prior II		
$m = 5$	ER_S	ER_L $c = 0.5$	ER_L $c = -0.5$	ER_S	ER_L $c = 0.5$	ER_L $c = -0.5$
$\widehat{\theta}_{ML}$	0.0357	0.0048	0.0042	0.0357	0.0048	0.0042
$\widehat{\theta}_{ST}$	0.0329	0.0044	0.0039	0.0358	0.0048	0.0042
$\widehat{\theta}_{LT}(c = 0.5)$	0.0303	0.0040	0.0036	0.0328	0.0044	0.0038
$\widehat{\theta}_{LT}(c = -0.5)$	0.0357	0.0048	0.0042	0.0390	0.0053	0.0045
$\widehat{\theta}_{SI}$	0.0330	0.0044	0.0039	0.0359	0.0049	0.0042
$\widehat{\theta}_{LI}(c = 0.5)$	0.0305	0.0041	0.0036	0.0330	0.0044	0.0039
$\widehat{\theta}_{LI}(c = -0.5)$	0.0359	0.0048	0.0042	0.0392	0.0053	0.0045
$\widehat{\theta}_{SM}$	0.0186	0.0024	0.0023	0.0356	0.0048	0.0041
$\widehat{\theta}_{LM}(c = 0.5)$	0.0176	0.0023	0.0021	0.0327	0.0044	0.0038
$\widehat{\theta}_{LM}(c = -0.5)$	0.0198	0.0026	0.0024	0.0388	0.0053	0.0045
$m = 7$						
$\widehat{\theta}_{ML}$	0.0220	0.0029	0.0026	0.0220	0.0029	0.0026
$\widehat{\theta}_{ST}$	0.0209	0.0027	0.0025	0.0220	0.0029	0.0026
$\widehat{\theta}_{LT}(c = 0.5)$	0.0198	0.0026	0.0024	0.0208	0.0027	0.0025
$\widehat{\theta}_{LT}(c = -0.5)$	0.0220	0.0029	0.0026	0.0232	0.0031	0.0028
$\widehat{\theta}_{SI}$	0.0212	0.0028	0.0025	0.0223	0.0029	0.0027
$\widehat{\theta}_{LI}(c = 0.5)$	0.0201	0.0026	0.0025	0.0211	0.0028	0.0025
$\widehat{\theta}_{LI}(c = -0.5)$	0.0224	0.0030	0.0027	0.0236	0.0031	0.0028
$\widehat{\theta}_{SM}$	0.0124	0.0016	0.0015	0.0218	0.0029	0.0026
$\widehat{\theta}_{LM}(c = 0.5)$	0.0121	0.0015	0.0015	0.0207	0.0027	0.0025
$\widehat{\theta}_{LM}(c = -0.5)$	0.0128	0.0016	0.0016	0.0231	0.0030	0.0027

- We see that the M-H method leads to the smaller ERs in comparison with the TK and IS methods in the most cases. Besides, the M-H method produces estimators that have ERs which are smaller than or close to those of the ML method in more than 50% of the cases.

5.2. Example 1 (Real Data Set 1)

Here, we consider the following data on the amount of rainfall (in inches) recorded at the Los Angeles Civic Center in February from 1998 to 2018; see the website of Los Angeles Almanac: www.laalmanac.com/weather/we08aa.php.

```
0.56, 5.54,  8.87, 0.29,  4.64, 4.89, 11.02, 2.37, 0.92,  1.64,
3.57, 4.27,  3.29, 0.16,  0.20, 3.58, 0.83,  0.79, 4.17,  0.03.
```

First, we compare the fit of the xgamma distribution with three other one-parameter lifetime distributions, listed as follows:
(i) The exponential distribution with a scale parameter θ.
(ii) The Lindley distribution with parameter θ whose PDF is given by (Lindley, 1958 and Ghitany et al., 2008)

$$f_{Lindley}(x) = \frac{\theta^2}{1+\theta}(1+x)\mathrm{e}^{-\theta x}, \qquad \theta > 0, \quad x > 0.$$

Table 2: The ERs of the point estimators of θ when $\theta = 1.5$.

	Prior I			Prior II		
$m = 5$	ER_S	ER_L $c = 0.5$	ER_L $c = -0.5$	ER_S	ER_L $c = 0.5$	ER_L $c = -0.5$
$\widehat{\theta}_{ML}$	0.5901	0.1251	0.0549	0.5901	0.1251	0.0549
$\widehat{\theta}_{ST}$	0.2312	0.0337	0.0255	0.6104	0.1293	0.0566
$\widehat{\theta}_{LT}(c = 0.5)$	0.1808	0.0253	0.0207	0.4152	0.0721	0.0418
$\widehat{\theta}_{LT}(c = -0.5)$	0.3026	0.0461	0.0323	1.0359	2.1959	0.0814
$\widehat{\theta}_{SI}$	0.2299	0.0335	0.0254	0.6078	0.1316	0.0563
$\widehat{\theta}_{LI}(c = 0.5)$	0.1810	0.0254	0.0207	0.4205	0.0751	0.0421
$\widehat{\theta}_{LI}(c = -0.5)$	0.3028	0.0462	0.0323	1.0289	1.4503	0.0814
$\widehat{\theta}_{SM}$	0.3965	0.0446	0.0555	0.5465	0.1078	0.0520
$\widehat{\theta}_{LM}(c = 0.5)$	0.4156	0.0466	0.0583	0.3810	0.0640	0.0390
$\widehat{\theta}_{LM}(c = -0.5)$	0.3771	0.0425	0.0526	0.8752	0.3042	0.0738
$m = 7$						
$\widehat{\theta}_{ML}$	0.3073	0.0510	0.0319	0.3073	0.0510	0.0319
$\widehat{\theta}_{ST}$	0.1741	0.0248	0.0196	0.3160	0.0526	0.0327
$\widehat{\theta}_{LT}(c = 0.5)$	0.1453	0.0201	0.0168	0.2481	0.0384	0.0268
$\widehat{\theta}_{LT}(c = -0.5)$	0.2119	0.0311	0.0233	0.4148	0.0794	0.0407
$\widehat{\theta}_{SI}$	0.1738	0.0247	0.0196	0.3146	0.0524	0.0326
$\widehat{\theta}_{LI}(c = 0.5)$	0.1456	0.0201	0.0168	0.2487	0.0385	0.0268
$\widehat{\theta}_{LI}(c = -0.5)$	0.2123	0.0311	0.0233	0.4152	0.0795	0.0408
$\widehat{\theta}_{SM}$	0.4584	0.0512	0.0645	0.2868	0.0463	0.0302
$\widehat{\theta}_{LM}(c = 0.5)$	0.4723	0.0527	0.0666	0.2289	0.0346	0.0250
$\widehat{\theta}_{LM}(c = -0.5)$	0.4443	0.0497	0.0625	0.3743	0.0675	0.0375

Table 3: The AWs and CPs of the 95% interval estimators of θ.

			Prior I		Prior II	
θ	m	MATE CI	IS method	M-H method	IS method	M-H method
0.5	5	0.6162 (0.9662)	0.5783 (0.9336)	0.5377 (0.9498)	0.5877 (0.9326)	0.5965 (0.9548)
	7	0.4987 (0.9566)	0.4652 (0.9140)	0.4421 (0.9408)	0.4691 (0.9120)	0.4860 (0.9486)
1.5	5	2.2977 (0.9650)	1.8428 (0.9746)	0.9582 (0.2540)	2.1983 (0.9542)	2.1713 (0.9542)
	7	1.8139 (0.9598)	1.5702 (0.9670)	0.7843 (0.0246)	1.7560(0.9516)	1.7359 (0.9518)

Table 4: The EPRs of approximate predictors of R_{m+1} when $\theta = 0.5$.

	Prior I			Prior II		
$m=5$	EPR_S	EPR_L $c=0.5$	EPR_L $c=-0.5$	EPR_S	EPR_L $c=0.5$	EPR_L $c=-0.5$
$\widehat{R}^{SI}_{m+1}$	0.0180	0.0023	0.0023	0.0180	0.0023	0.0023
$\widehat{R}^{LI}_{m+1}(c=0.5)$	0.0184	0.0022	0.0024	0.0184	0.0022	0.0024
$\widehat{R}^{LI}_{m+1}(c=-0.5)$	0.0184	0.0024	0.0022	0.0184	0.0024	0.0022
$\widehat{R}^{SM}_{m+1}$	0.0180	0.0023	0.0023	0.0180	0.0023	0.0023
$\widehat{R}^{LM}_{m+1}(c=0.5)$	0.0183	0.0022	0.0024	0.0184	0.0022	0.0024
$\widehat{R}^{LM}_{m+1}(c=-0.5)$	0.0184	0.0024	0.0022	0.0184	0.0024	0.0022
$m=7$						
$\widehat{R}^{SI}_{m+1}$	0.0018	0.0002	0.0002	0.0018	0.0002	0.0002
$\widehat{R}^{LI}_{m+1}(c=0.5)$	0.0017	0.0002	0.0002	0.0017	0.0002	0.0002
$\widehat{R}^{LI}_{m+1}(c=-0.5)$	0.0019	0.0003	0.0002	0.0019	0.0003	0.0002
$\widehat{R}^{SM}_{m+1}$	0.0018	0.0002	0.0002	0.0018	0.0002	0.0002
$\widehat{R}^{LM}_{m+1}(c=0.5)$	0.0017	0.0002	0.0002	0.0017	0.0002	0.0002
$\widehat{R}^{LM}_{m+1}(c=-0.5)$	0.0019	0.0003	0.0002	0.0019	0.0003	0.0002

Table 5: The EPRs of approximate predictors of R_{m+1} when $\theta = 1.5$.

	Prior I			Prior II		
$m=5$	EPR_S	EPR_L $c=0.5$	EPR_L $c=-0.5$	EPR_S	EPR_L $c=0.5$	EPR_L $c=-0.5$
$\widehat{R}^{SI}_{m+1}$	0.0005	0.0001	0.0001	0.0005	0.0001	0.0001
$\widehat{R}^{LI}_{m+1}(c=0.5)$	0.0005	0.0001	0.0001	0.0005	0.0001	0.0001
$\widehat{R}^{LI}_{m+1}(c=-0.5)$	0.0005	0.0001	0.0001	0.0005	0.0001	0.0001
$\widehat{R}^{SM}_{m+1}$	0.0005	0.0001	0.0001	0.0005	0.0001	0.0001
$\widehat{R}^{LM}_{m+1}(c=0.5)$	0.0005	0.0001	0.0001	0.0005	0.0001	0.0001
$\widehat{R}^{LM}_{m+1}(c=-0.5)$	0.0005	0.0001	0.0001	0.0005	0.0001	0.0001
$m=7$						
$\widehat{R}^{SI}_{m+1}$	0.00007	0.00001	0.00001	0.00007	0.00001	0.00001
$\widehat{R}^{LI}_{m+1}(c=0.5)$	0.00007	0.00001	0.00001	0.00007	0.00001	0.00001
$\widehat{R}^{LI}_{m+1}(c=-0.5)$	0.00007	0.00001	0.00001	0.00007	0.00001	0.00001
$\widehat{R}^{SM}_{m+1}$	0.00007	0.00001	0.00001	0.00007	0.00001	0.00001
$\widehat{R}^{LM}_{m+1}(c=0.5)$	0.00007	0.00001	0.00001	0.00007	0.00001	0.00001
$\widehat{R}^{LM}_{m+1}(c=-0.5)$	0.00007	0.00001	0.00001	0.00007	0.00001	0.00001

Table 6: The AWs and CPs of the 95% ABPIs of R_{m+1}.

		Prior I		Prior II	
θ	m	IS method	M-H method	IS method	M-H method
0.5	5	0.2001 (0.9514)	0.2001 (0.9514)	0.2001 (0.9514)	0.2001 (0.9514)
	7	0.0477 (0.9566)	0.0477 (0.9566)	0.0477 (0.9566)	0.0477 (0.9566)
1.5	5	0.0355 (0.9500)	0.0356 (0.9498)	0.0355 (0.9500)	0.0355 (0.9500)
	7	0.0087 (0.9500)	0.0087 (0.9500)	0.0087 (0.9500)	0.0087 (0.9500)

(*iii*) The Shanker distribution with the following PDF (Shanker, 2015):

$$f_{Shanker}(x) = \frac{\theta^2}{1+\theta^2}(\theta+x)\mathrm{e}^{-\theta x}, \qquad \theta > 0, \quad x > 0.$$

We use the ML method to obtain the parameter estimate. We use the Kolmogorov-Smirnov (K-S) test and the corresponding p-value, Akaike information criterion (AIC), Bayesian information criterion (BIC) and Hannan-Quinn information criterion (HQIC) to compare the fits of the considered distributions and the results have been given in Table 9. From Table 9, we observe that the xgamma and exponential models fit the data better than the Lindley and Shanker models, as they have smaller AICs, BICs, HQICs and K-S test statistics.

Table 7: MLEs and goodness-of-fit statistics for Example 1.

Distribution	MLE	AIC	BIC	HQIC	K-S	p-value
xgamma	0.6895	87.0868	88.0826	87.2812	0.1925	0.3978
exponential	0.3245	87.0166	88.0123	87.2110	0.1561	0.6575
Lindley	0.5358	89.0368	90.0325	89.2312	0.2068	0.3141
Shanker	0.5874	90.4794	91.4752	90.6738	0.2010	0.3465

We have extracted the lower records and the corresponding inter-record times as follows:

i	1	2	3	4
r_i	0.56	0.29	0.16	0.03
k_i	3	10	6	1

We have considered both the exponential and xgamma models, as we see that these models fit the data well. Here, we have used the approximate non-informative prior with the prior mean equal to 1.5 and the prior variance equal to 100, so we have $(a,b) = (0.0225, 0.0150)$. We have calculated the ML and approximate Bayes point estimates, as well as the 95% interval estimates of the parameter for both exponential and xgamma distributions. The point predictions and 95% ABPIs for the next future record, namely R_5, have been obtained as well. We have used the M-H method for the Bayesian estimation and prediction for the xgamma distribution. However, the Bayesian estimates of the unknown parameter for the exponential distribution have explicit forms, and we have also used the function `integrate` in R to evaluate predictions and ABPIs for the exponential distribution. The numerical results

of this example have been given in Table 10. From Table 10, we see that all the predictions and ABPIs are too close to each other in both exponential and xgamma distributions. Here, we predict that the next lowest amount of rainfall (after the year 2018) would be approximately 0.015 inches, which is the predicted 5-th lower record value since 1998.

Table 8: The numerical results of Example 1.

Estimation	$\widehat{\theta}_{ML}$	$\widehat{\theta}_{SE}$	$\widehat{\theta}_{LE}$ $(c=0.5)$	$\widehat{\theta}_{LE}$ $(c=-0.5)$	MATE CI
xgamma	1.3467	1.3293	1.2749	1.3904	(0.3999, 2.2935)
exponential	0.7181	0.7202	0.6898	0.7545	(0.0144, 1.4219)
Prediction		$\widehat{R}_5^S$	$\widehat{R}_5^L$ $(c=0.5)$	$\widehat{R}_5^L$ $(c=-0.5)$	ABPI
xgamma		0.0148	0.0149	0.0149	(0.0007, 0.0292)
exponential		0.0149	0.0149	0.0150	(0.0007, 0.0292)

5.3. Example 2 (Real Data Set 2)

The second data set includes daily numbers of deaths due to the COVID-19 virus in Poland from 1 September 2020 to 1 October 2020; see the website of COVID-19 data: https://ourwordindata. org/coronavirus-source-data. The data are:

```
19, 20, 14, 8, 13, 7, 4, 12, 11, 12, 10, 13, 6, 15, 24, 10,
16, 17, 12, 11, 5, 18, 28, 25, 23, 32, 8, 15, 36, 30, 30.
```

Despite the fact that the daily numbers of deaths possess a discrete nature, several authors have fitted continuous distributions to this type of data sets; see for example El-Monsef et al. (2021). Now, if we observe the number of deaths on a specified day, then we may ask what the next lower number would be and become interested in predicting the next future lower record. Once again, we fitted the xgamma, exponential, Lindley and Shanker models to the above data and the related results have been given in Table 11. We see that the xgamma distribution fits the data best among the four considered models. Note that the exponential model does not fit the data significantly based on the K-S test at the level $\alpha = 0.05$. However, proceeding the same line of Example 1, we consider both xgamma and exponential models to analyse the record data. We observe that the lowest record occurred on 7 September 2020, so we can consider only the following data:

```
19, 20, 14, 8, 13, 7, 4.
```

The extracted lower records and the corresponding inter-record times from the above data are as follows:

i	1	2	3	4	5
r_i	19	14	8	7	4
k_i	2	1	2	1	1

We intend to predict the 5-th lower record value based on the first 4 observed records. Here, we have used informative priors whose prior means are taken to be the corresponding ML estimates approximately and the prior variance is equal to a value close to 0.4 (see Yadav et al., 2019). So, we obtain a prior with $(a,b) = (0.0702, 0.4255)$ for the xgamma distribution and another prior with $(a,b) = (0.0094, 0.1537)$ for the exponential model. We have calculated the ML and approximate Bayes point estimates, as well as the 95% interval estimates of the parameter for both the exponential and xgamma distributions. The point predictions and 95% ABPIs for the next future record, namely R_5, have been obtained as well. The numerical results of Example 2 have been given in Table 12. From Table 12, we see that all the approximate predictions are somehow close to 4, which is the true value of R_5. Besides, the ABPIs contain the true value of R_5.

Table 9: MLEs and goodness-of-fit statistics for Example 2.

Distribution	MLE	AIC	BIC	HQIC	K-S	p-value
xgamma	0.1702	221.2816	222.7156	221.749	0.1334	0.6393
exponential	0.0615	236.8925	238.3265	237.36	0.2658	0.0249
Lindley	0.1165	223.8098	225.2438	224.2773	0.1694	0.3358
Shanker	0.1218	221.9618	223.3958	222.4293	0.2177	0.1057

Table 10: The numerical results of Example 2.

Estimation	$\widehat{\theta}_{ML}$	$\widehat{\theta}_{SE}$	$\widehat{\theta}_{LE}$ $(c = 0.5)$	$\widehat{\theta}_{LE}$ $(c = -0.5)$	MATE CI
xgamma	0.1750	0.1705	0.1699	0.1711	(0.0780, 0.2719)
exponential	0.0533	0.0554	0.0552	0.0556	(0.0011, 0.1056)
Prediction		$\widehat{R}_5^S$	$\widehat{R}_5^L$ $(c = 0.5)$	$\widehat{R}_5^L$ $(c = -0.5)$	ABPI
xgamma		3.8454	2.8450	4.7229	(0.2050, 6.8675)
exponential		3.2757	2.3813	4.2395	(0.1453, 6.7882)

6. Concluding Remarks

Recently, the xgamma distribution, which is a flexible distribution for lifetime phenomena, has been introduced by Sen et al. (2016). In this paper, first, we derived the ML estimate of the xgamma parameter based on record values and the corresponding inter-record times. Then, we focused on the Bayesian estimation of the parameter, and we used a symmetric loss function as well as an asymmetric one. The Bayesian point estimates involve complicated integrals that do not seem to have closed forms, so we have used the TK, IS and M-H methods, to evaluate them. We have also become involved in predicting future records, as the prediction of future records has attracted the researchers' attention in applied situations.

A simulation study has been conducted in order to assess the point and interval estimators of the unknown parameter of the xgamma distribution and the point and interval predictors of a future lower record value. From the simulation study, we conclude that the

number of observed records and the values of hyperparameters affect the performance of the estimators and predictors. Two real data sets have been analysed, where the first one includes the rainfall data. Here, a lower record value can be a warning about a future drought. The second data set involves the daily numbers of deaths in Poland due to the COVID-19 virus, where the lower records can show whether the virus can become under control or not. We compared the fits of the xgamma distribution with three other one-parameter lifetime distributions, and we observed the xgamma fitted both data sets well enough. Taking this information into account, we analysed the record data with the help of both the xgamma and exponential distributions through classical and Bayesian methods. We observe that the prediction results based on both the xgamma and exponential distributions are too close to each other for the rainfall data, and they are not much different from each other for the COVID-19 data. Besides, the obtained predicted values of the 5-th lower record are close to the true value of R_5 for the COVID-19 data, which confirms that the theoretical results of the paper may perform well in prediction. Summing up, we may conclude that the results of this paper may be useful in the estimation and prediction in real phenomena. All the computations of the paper were done using Maple 2016 and the statistical software R (R Core Team, 2020), and the packages coda (see Plummer et al., 2006, 2018), nleqslv (see Hasselman, 2018), truncnorm (see Mersmann et al., 2018) and AdequacyModel (see Marinho et al., 2013) therein.

Acknowledgements

We would like to express our sincere thanks to the Editor, Professor Patryk Barszcz, and the two anonymous referees for their valuable comments that highly improved the paper.

References

Ahmadi, J., MirMostafaee, S. M. T. K., (2009). Prediction intervals for future records and order statistics coming from two parameter exponential distribution. *Statistics & Probability Letters*, 79(7), pp. 977–983.

AL-Hussaini, E.K., Al-Awadhi, F., (2010). Bayes two-sample prediction of generalized order statistics with fixed and random sample size. *Journal of Statistical Computation and Simulation*, 80(1), pp. 13–28.

Albert, J., (2009). *Bayesian Computation with R*, 2nd Ed., LLC: Springer.

Amini, M., MirMostafaee, S. M. T. K., (2016). Interval prediction of order statistics based on records by employing inter-record times: A study under two parameter exponential distribution. *Metodološki Zvezki*, 13(1), pp. 1–15.

Arnold, B. C., Balakrishnan, N., Nagaraja, H. N., (1998). *Records*, New York: John Wiley & Sons.

Bastan, F., MirMostafaee, S. M. T. K., (2022). Estimation and prediction for the Poisson-exponential distribution based on records and inter-record times: A comparative study. *Journal of Statistical Sciences*, 15(2), 381–405.

Chen, M.-H., Shao, Q.-M., (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, 8(1), pp. 69–92.

El-Monsef, M. M. E. A., Sweilam, N.H., Sabry, M. A., (2021). The exponentiated power Lomax distribution and its applications. *Quality and Reliability Engineering International*, 37(3), pp. 1035–1058.

Fallah, A., Asgharzadeh, A., MirMostafaee, S. M. T. K., (2018). On the Lindley record values and associated inference. *Journal of Statistical Theory and Applications*, 17(4), pp. 686–702.

Geweke, J., (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Clarendon Press, Oxford, UK, pp. 169–193.

Ghitany, M. E., Atieh, B., Nadarajah, S., (2008). Lindley distribution and its application. *Mathematics and Computers in Simulation*, 78(4), pp. 493–506.

Hasselman, B., (2018). nleqslv: Solve systems of nonlinear equations, R package version 3.3.2, https://CRAN.R-project.org/package=nleqslv.

Hastings, W. K., (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), pp. 97–109.

Heidelberger, P., Welch, P. D., (1981a). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24(4), pp. 233–245.

Heidelberger, P., Welch, P. D., (1981b). Adaptive spectral methods for simulation output analysis. *IBM Journal of Research and Development*, 25(6), pp. 860–876.

Heidelberger, P., Welch, P. D., (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6), pp. 1109–1144.

Kızılaslan, F., Nadar, M., (2015). Estimation with the generalized exponential distribution based on record values and inter-record times. *Journal of Statistical Computation and Simulation*, 85(5), pp. 978–999.

Kumar, D., Dey, S., Ormoz, E., MirMostafaee, S. M. T. K., (2020). Inference for the unit-Gompertz model based on record values and inter-record times with an application.*Rendiconti del Circolo Matematico di Palermo Series 2*, 69(3), pp. 1295–1319.

Lehmann, E. L., Casella, G., (1998). *Theory of Point Estimation*, 2nd Ed., New York: Springer.

Lindley, D. V., (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society, Series B (Methodological)*, 20(1), pp. 102–107.

Marinho, P. R. D., Bourguignon, M., Dias, C. R. B., (2013). AdequacyModel: Adequacy of probabilistic models and generation of pseudo-random numbers, R package version 1.0.8., https://CRAN.R-project.org/package=AdequacyModel.

Mersmann, O., Trautmann, H., Steuer, D., Bornkamp, B., (2018). truncnorm: Truncated normal distribution, R package version 1.0-8, https://CRAN.R-project.org/package=truncnorm.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E., (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), pp. 1087–1092.

MirMostafaee, S. M. T. K., Asgharzadeh, A., Fallah, A., (2016). Record values from NH distribution and associated inference. *Metron*, 74(1), pp. 37–59.

Nadar, M., Kızılaslan, F., (2015). Estimation and prediction of the Burr type XII distribution based on record values and inter-record times. *Journal of Statistical Computation and Simulation*, 85(16), pp. 3297–3321.

Pak, A., Dey, S., (2019). Statistical inference for the power Lindley model based on record values and inter-record times. *Journal of Computational and Applied Mathematics*, 347, pp. 156–172.

Plummer, M., Best, N., Cowles, K., Vines, K., (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1), pp. 7–11.

Plummer, M., Best, N., Cowles, K., Vines, K., Sarkar, D., Bates, D., Almond, R., Magnusson, A., (2018). coda: Output analysis and diagnostics for MCMC, R package version 0.19-2, https://CRAN.R-project.org/package=coda.

R Core Team, (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Lewis, S. M., (1992). Comment: One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7(4), pp. 493–497.

Raftery, A. E., Lewis, S. M., (1996). Implementing MCMC. In *Markov Chain Monte Carlo in Practice*, Eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter, Chapman and Hall/CRC, Boca Raton, pp. 115–130.

Samaniego, F. J., Whitaker, L. R., (1986). On estimating population characteristics from record-breaking observations. i. parametric results. *Naval Research Logistics Quarterly*, 33(3), pp. 531–543.

Schruben, L. W., (1982). Detecting initialization bias in simulation output. *Operations Research*, 30(3), pp. 569–590.

Schruben, L., Singh, H., Tierney, L., (1980). A test of initialization bias hypotheses in simulation output. Technical Report 471, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York, 14853.

Schruben, L., Singh, H., Tierney, L., (1983). Optimal tests for initialization bias in simulation output. *Operations Research*, 31(6), pp. 1167–1178.

Sen, S., Chandra, N., Maiti, S. S., (2018). Survival estimation in xgamma distribution under progressively type-II right censored scheme. *Model Assisted Statistics and Applications*, 13(2), pp. 107–121.

Sen, S., Maiti, S. S., Chandra, N., (2016). The xgamma distribution: Statistical properties and application. *Journal of Modern Applied Statistical Methods*, 15(1), pp. 774–788.

Shanker, R., (2015). Shanker distribution and its applications. *International Journal of Statistics and Applications*, 5(6), pp. 338–348.

Tierney, L., Kadane, J. B., (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393), pp. 82–86.

Varian, H. R., (1975). Bayesian approach to real estate assessment. In *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, Eds. S.E. Fienberg and A. Zellner, North-Holland Pub. Co., Amsterdam, pp. 195–208.

Yadav, A. S., Saha, M., Singh, S. K., Singh, U., (2019). Bayesian estimation of the parameter and the reliability characteristics of xgamma distribution using type-II hybrid censored data. *Life Cycle Reliability and Safety Engineering*, 8(1), pp. 1–10.

Zellner, A., (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394), pp. 446–451.

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 109–127, https://doi.org/10.59170/stattrans-2023-066
Received – 08.03.2022; accepted – 13.02.2023

A study of a survival data using kernel estimates of hazard rate and aging intensity functions

Magdalena Szymkowiak[1], **Anasuya Roychowdhury**[2], **Satya Kr. Misra**[3], **Rajib Lochan Giri**[4], **Subarna Bhattacharjee**[5]

Abstract

Analyzing survival (life-testing) data and drawing inferences about them is a part of engineering and health sciences. So far, various statistical tools, e.g., survival (reliability) function (sf), probability density function (pdf), and hazard rate function (HR) were available among decision-making scientists to handle time-to-event data (complete or censored). But because functions (pdf) estimators were interval (window) based, they mostly gave qualitative ideas having pictorial representation resembling step functions, ordinate remain constant when abscissa vary over an interval, thereby giving incomplete information. However, it can be sorted out with the use of kernel estimates of the above mentioned functions, resulting into smooth estimators. Moreover, the metric based on aging intensity function (AI) gives an alternative way of studying lifetime or clinical datasets as it is a quantitative measure (not interval-based), thereby depicting a broader view of a given data. In our study, we primarily focus on AI and HR functions estimated using four different kernels. We apply them to a case study of patients with primary malignant tumors of sternum (cf. Daniel and Cross, 2014) with the right-censored data. Our result shows that kernel estimates of HR and AI functions for patients with high grade tumor (HGT) are higher than for patients with low grade tumor (LGT), as expected. Thus, the study opens up a new direction for applying AI and HR functions in health sciences and engineering studies.

Key words: hazard rate, aging intensity function, kernels, survival analysis, cancer statistics, clinical datasets.

1. Introduction

The comparison of two different products for two different brands has high importance in many fields including but not limited to reliability theory, biological sciences, and forensic sciences. In survival analysis, the remaining lifetimes of a component at different times of its life span needs to be compared to determine how the component is aging with time. Various stochastic orders between random variables, viz., classical stochastic (st) order, hazard rate (hr) order, likelihood ratio (lr) order, aging intensity (ai) order, etc. have been studied in the literature (cf. Shaked and Shanthikumar (2007)). In this regard, our article

[1]Institute of Automatic Control and Robotics, Poznan University of Technology, Poznań, Poland. ORCID: https://orcid.org/0000-0002-5066-8629.
[2]Biochemistry and Cell Biology Laboratory, School of Basic Sciences, IIT Bhubaneswar, India.
[3]Department of Mathematics, KIIT University, Bhubaneswar-751024, Odisha, India.
[4]Department of Mathematics, Ravenshaw University, Cuttack-753003, Odisha, India.
[5]Corresponding author :Department of Mathematics, Ravenshaw University, Cuttack-753003, Odisha, India. E-mail: subarna.bhatt@gmail.com. ORCID: https://orcid.org/0000-0002-3697-4216.

analyzes a case study of cancer data cited in Daniel and Cross (2014).

Emmerson and Brown (2021), Rosen et al. (2020) and others discuss the use of Kaplan-Meier survival analysis in evaluating the efficiency of onco-drugs in randomised controlled trials (RCTs). Further, Nayak et al. (2021) apply Kaplan-Meier analysis to assess a potential oncoprotein ATAD2 as a prognostic marker for stomach cancer. Inquisitive readers can further explore on Kaplan-Meier Plotter (KMP, http://kmplot.com/analysis/), which is a web-based meta-analysis biomarker validation tool used in medical research and also used by Nayak et al. (2021) in the analysis of stomach cancer.

The treatment management of cancer starts with the determination of stages (how big the tumor is and how far it is spread) and grades (how fast it grows) of the tumor. A higher stage and/or grade of tumor may grow and spread rapidly and may require immediate treatment action. For example, high-grade tumors (HGT) are more aggressive than the low grade tumors (LGT). Therefore, severity of the disease and treatment management could be more complicated for HGT. Moreover, every cancer sub-type is unique.

A statistical account on patients with the same cancer sub-type often help the public health community to estimate the prognosis better. Therefore, to improve the treatment spectrum of the complex disease like cancer, there is an immense importance of statistical analysis using incidence and survival data from the vast range of original datasets that are generally accumulated in authentic and authoritative public repository databases. However, to dig out meaningful information from those datasets, the statistical analytical tools are required to be robust and bias-free. So, we make an attempt to apply it for a particular data so as to follow its aging pattern.

Although aging intensity function AI, defined as the ratio of the instantaneous HR to its average, has already gathered some familiarity in recent literature of statistics, to the best of our knowledge its application in analysis of survival data is sparse (cf. Misra and Bhattacharjee (2018)). Here, in this study, we take up a strategy of implementing kernels estimates of hazard rate (HR) and aging intensity (AI) functions for the survival analysis of a particular censored data of patients suffering from malignant tumors of sternum (cf. Daniel and Cross (2014)). Censoring of data arises as lifetimes occur only within certain intervals. Censored data are useful when their survival time is truncated at a certain point of time.

The rest of our article is organized as follows. Portfolio of HR and the AI functions are presented in Section 2. A brief survey on kernels used in estimation follows. In Section 3, we cite a dataset which has been taken up in our present study and discuss the results so obtained. The significance of the paper is established in concluding remarks of Section 4. In Section 5, the Appendix speaks about the detailed calculations of this work connected with four presented kernels and gives a short comparison between them based on goodness-of-fit test. The notation *r.v.* is used in place of random variable.

2. Portfolio of HR and AI functions

The keywords of this work are hazard rate HR and aging intensity AI functions, which are being here used along with their kernel estimates for application in medical statistics, especially cancer analysis and in health sciences. To this end, we give a brief note of the

aforementioned concepts in the ensuing discussions.

2.1. Hazard rate interpretation

Let T be a random variable representing any lifetime of a system with a well-defined statistical distribution having probability density function (pdf) denoted by f, and survival function (or popularly known as reliability function) (sf) denoted by $\overline{F}$. Sf of a *r.v.* T at time t is given by $\overline{F}_T(t) = P(T > t)$ which represents the probability of surviving over time t. The cumulative distribution function (cdf) is $F_T(t) = 1 - \overline{F}_T(t)$, $t > 0$. Hazard rate (HR) function, known also in survival analysis as failure rate, is defined for a continuous random variable T as

$$h_T(t) = \lim_{\Delta t \to 0} \frac{P[t < T \leq t + \Delta t | T > t]}{\Delta t} = \frac{f_T(t)}{\overline{F}_T(t)}, \text{ where density is defined.} \quad (2.1)$$

If the hazard rate is high, then it implies that the corresponding unit with life-time T is aging faster. Other functions used in the study of aging analysis are reversed hazard rate, mean residual life (cf. Nanda et al. (2010)), reversed mean residual life (cf. Nanda et al. (2006), Shaked and Shanthikumar (2007)) functions, etc.

2.2. Significance of aging intensity function

Jiang et al. (2003) classifies a unimodal hazard rate as quasi-decreasing (anti-aging), quasi-increasing (aging) or quasi-constant (non-aging) depending on whether its mode t_c, (called critical time) is small, large or moderate, respectively. A distribution is classified as quasi-constant if the hazard rate curve is relatively flat. They claim that the representation of aging of a system by hazard rate is qualitative. Thereby, they introduced a notion, called aging intensity (AI), to quantitatively evaluate the aging property of a system. AI of a random variable T, denoted by $L_T(t)$, is defined as the ratio of the instantaneous hazard rate $h_T(t)$ given by (2.1) to the hazard rate average $\frac{1}{t}H_T(t)$, where $H_T(t) = \int_0^t h_T(u)du$ is the cumulative hazard rate, i.e.,

$$L_T(t) = \frac{h_T(t)}{\frac{1}{t}H_T(t)}. \quad (2.2)$$

It is easy to see that (2.2) can be also presented as

$$L_T(t) = \frac{-t f_T(t)}{\overline{F}_T(t) \ln \overline{F}_T(t)}, \text{ for } t > 0. \quad (2.3)$$

The concept of aging intensity (AI) function is found in Nanda et al. (2007), Bhattacharjee et al. (2013b), Bhattacharjee et al. (2022) for quantitative study of the aging process of a system. Bhattacharjee et al. (2013a), Misra and Bhattacharjee (2016), Swain et al. (2021) illustrate the properties of AI function and its usage on a complete dataset. Misra and Bhattacharjee (2018) gave a comparative role of HR, and AI functions on analysis of data (censored). Szymkowiak (2018, 2019, 2020) gave a detailed literature on AI function. Giri et al. (2023) studied HR, AI functions of different Weibull models and simulated data from Weibull distributions.

For a complete data, the empirical estimates of pdf, cdf and AI functions, denoted by $\hat{f}(t), \hat{F}(t)$ and $\hat{L}(t)$, respectively, are given by (cf. Bhattacharjee et al. (2013a), Szymkowiak (2018))

$$\hat{f}_{emp}(t) = \frac{N_s(t_j) - N_s(t_j + \Delta t_j)}{N \Delta t_j},$$

$$\hat{F}_{emp}(t) = \frac{N - N_s(t_j)}{N}, \tag{2.4}$$

$$\hat{L}_{emp}(t) = -\frac{t\hat{f}(t)}{\left[1 - \hat{F}(t)\right] \ln\left[1 - \hat{F}(t)\right]} = -t\left\{\frac{N_s(t_j) - N_s(t_j + \Delta t_j)}{\Delta t_j N_s(t_j) \ln \frac{N_s(t_j)}{N}}\right\}, \tag{2.5}$$

for $t_j \leq t \leq t_j + \Delta t_j$. Here, $N, N_s(t_j)$ and $N_s(t_j + \Delta t_j)$ refer to the total number of survivors at $t = 0$ (beginning of the life-testing), $t = t_j$ and $t = t_j + \Delta t_j$, respectively.

The above defined estimates $\hat{f}_{emp}(t)$ and $\hat{F}_{emp}(t)$ for HGT and LGT patients are presented in Figure 5.13 and Figure 5.29, respectively, as the blue step functions. One can justify the fact that AI function is a quantitative measure of aging as the factor t is involved in (2.5), which gives rise to a smooth (or not window) estimator.

Refer Klein and Moeschberger (2003) for detailed analysis on estimator of cumulative hazard rate $H(t) = \int_0^t h(u)du$ and hazard rate $h(t)$. $\hat{H}(t)$, is the estimator given by Nelson-Aalen for $H(t)$ and the slope of this estimator gives a rough estimate of the $\hat{h}(t)$. Clearly, the estimator of AI function (2.2) is given by $\hat{L}(t) = \frac{\hat{h}(t)}{\frac{1}{t}\hat{H}(t)}, t > 0$.

The following subsection gives a brief survey of kernels which helps us to give smooth (not window-based) estimators for functions used in statistics.

2.3. Kernels: a brief survey

If probability density function is unknown or difficult to obtain in parametric distributions, we can use kernel estimates of pdf and cdf functions for their applications in statistical inference. One can refer to DiNardo and Tobias (2001) to name a few. These problems are faced primarily by statisticians who are engaged in evaluating reliability. An important aspect of aforementioned kernel estimation is associated with selecting suitable kernel and the choice of its corresponding bandwidth (b). Readers can explore on some well-known literature (cf. Miladinovic (2008)) on ranking of seven crucial kernels on the basis of their (optimal) bandwidth.

The expressions of kernel estimates of pdf, cdf and sf are, respectively, given by the following definition (cf. Miladinovic (2008)).

Definition 2.1 *If $T_1, T_2, \ldots, T_n$ are i.i.d. random variables with the same $f_n(t)$, the kernel estimate of pdf is given by*

$$\hat{f}_n(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - T_i}{b}\right), \tag{2.6}$$

where b is the bandwidth and $K(u)$ is a kernel smoothing function so chosen. The kernel

estimate of the cdf and survival (reliability) function (sf) are, respectively, given by

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{t} K\left(\frac{u-T_i}{b}\right)du, \tag{2.7}$$

and $\widehat{\overline{F}}_n(t) = 1-\hat{F}_n(t)$.

We list four kernels as follows (cf. Bowman and Azzalini (1997), Miladinovic (2008)). Here, $I(A)$ refers to the fact that $I(x)=1$, if $x\in A$ and $I(x)=0$, if $x\notin A$, where $A\neq\emptyset$.

(i) Epanechnikov (EPA) kernel, $K(u)=\frac{3}{4}(1-u^2)I(|u|\leq 1)$

(ii) Normal (Gaussian) kernel, $K(u)=\frac{1}{\sqrt{2\pi}}e^{-0.5u^2}$

(iii) Triangle kernel, $K(u)=(1-|\,u\,|)I(|\,u\,|\leq 1)$

(iv) Box (Uniform) kernel, $K(u)=0.5I(|\,u\,|\leq 1)$

The kernels must satisfy a set of properties as given in the following remark.

Remark 2.1 *The kernels must satisfy the conditions* $\int_{-\infty}^{\infty}K(u)du=1$, $\int_{-\infty}^{\infty}uK(u)du=0$ *and* $\int_{-\infty}^{\infty}u^2K(u)du>0$.

The properties possessed by the kernels are partially the same as that of the kernel density estimates. Epanechnikov introduced the kernel after his name for density estimation in 1956. The bandwidth b plays a crucial role and is assigned a value in such a way that it minimizes mean-squared error or it helps in obtaining the required degree of smoothness. To obtain the aging intensity estimator, we use (2.6) and (2.7). These are available in *MATLAB* as `ksdensity` function.
As usual,

$$\hat{h}_n(t)=\frac{\hat{f}_n(t)}{1-\hat{F}_n(t)}=\frac{\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-T_i}{b}\right)}{1-\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{t}K\left(\frac{u-T_i}{b}\right)du}.$$

and by (2.3) aging intensity estimate is equal to

$$\hat{L}_n(t) \;=\; -\frac{t\,\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-T_i}{b}\right)}{\left[1-\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{t}K\left(\frac{u-T_i}{b}\right)du\right]\ln\left[1-\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{t}K\left(\frac{u-T_i}{b}\right)du\right]}.$$

Here, we apply only four kernels estimates of HR and AI: Box kernels, $\hat{h}_B$, $\hat{L}_B$, Epanechnikov (EPA) kernels, $\hat{h}_E$, $\hat{L}_E$, Normal kernels, $\hat{h}_N$, $\hat{L}_N$, and Triangle kernels, $\hat{h}_T$, $\hat{L}_T$, respectively.

3. Analysis of cancer data: results and discussion

We refer to data given in Martini et al. (1996) and Daniel and Cross (2014), displayed in Table 3.1. They noted primary malignant tumors of the sternum in patients with low-grade

Table 3.1: Data: Malignant Tumors of Sternum (cf. Daniel and Cross (2014))

Subject	Time t [month]	Vital Status	Tumor Grade	Subject	Time t [month]	Vital Status	Tumor Grade
1	29	*dod*	*LGT*	21	155	*ned*	*LGT*
2	129	*ned*	*LGT*	22	102	*dod*	*LGT*
3	79	*dod*	*LGT*	23	34	*ned*	*LGT*
4	138	*ned*	*LGT*	24	109	*ned*	*LGT*
5	21	*dod*	*LGT*	25	15	*dod*	*LGT*
6	95	*ned*	*LGT*	26	122	*ned*	*HGT*
7	137	*ned*	*LGT*	27	27	*dod*	*HGT*
8	6	*ned*	*LGT*	28	6	*dod*	*HGT*
9	212	*dod*	*LGT*	29	7	*dod*	*HGT*
10	11	*dod*	*LGT*	30	2	*dod*	*HGT*
11	15	*dod*	*LGT*	31	9	*dod*	*HGT*
12	337	*ned*	*LGT*	32	17	*dod*	*HGT*
13	82	*ned*	*LGT*	33	16	*dod*	*HGT*
14	33	*dod*	*LGT*	34	23	*dod*	*HGT*
15	75	*ned*	*LGT*	35	9	*dod*	*HGT*
16	109	*ned*	*LGT*	36	12	*dod*	*HGT*
17	26	*ned*	*LGT*	37	4	*dod*	*HGT*
18	117	*ned*	*LGT*	38	0	*dpo*	*HGT*
19	8	*ned*	*LGT*	39	3	*dod*	*HGT*
20	127	*ned*	*LGT*				

tumor *LGT* (25 patients) or high-grade tumor *HGT* (14 patients), respectively (source: data provided courtesy of Dr. Martini).

The notations used in Table 3.1 are depicted as *dod* for 'dead of disease' (treated as uncensored data); *ned* for 'no evidence of disease' (treated as censored data) and *dpo* for 'dead post operation' (treated as uncensored data). Throughout this paper, t is given in months. In this article we aim to study the aging phenomena among the disease groups

Table 3.2: Kernel estimates of *AI* for *HGT*

t	$\hat{L}_B(t)$	$\hat{L}_E(t)$	$\hat{L}_N(t)$	$\hat{L}_T(t)$
0	–	–	–	–
2	0.3306	0.3351	0.3660	0.3526
3	0.4413	0.4687	0.5132	0.4976
4	0.5328	0.5918	0.6392	0.6198
6	0.7542	0.7965	0.8350	0.8139
7	0.8995	0.8774	0.9082	0.8896
9	1.0286	1.0007	1.0137	1.0130
12	1.1289	1.1295	1.0942	1.1022
16	1.1802	1.1711	1.1230	1.1802
17	1.2682	1.1713	1.1257	1.1783
23	0.9968	1.1023	1.1513	1.1454
27	1.0251	1.0771	1.1057	1.1197

Table 3.3: Kernel estimates of *HR* for *HGT*

t	$\hat{h}_B(t)$	$\hat{h}_E(t)$	$\hat{h}_N(t)$	$\hat{h}_T(t)$
0	0.0318	0.0309	0.0305	0.0307
2	0.0383	0.0377	0.0395	0.0388
3	0.0398	0.0413	0.0441	0.0433
4	0.0415	0.0455	0.0486	0.0475
6	0.0504	0.0540	0.0570	0.0555
7	0.0585	0.0580	0.0608	0.0593
9	0.0663	0.0652	0.0673	0.0666
12	0.0744	0.0751	0.0738	0.0737
16	0.0816	0.0814	0.0783	0.0820
17	0.0890	0.0823	0.0791	0.0828
23	0.0734	0.0810	0.0843	0.0840
27	0.0771	0.0805	0.0828	0.0837

HGT and *LGT* implementing four kernels which are most commonly used in statistical estimation (available with *MATLAB* R2016a version). Our primary focus is on two different measures, one qualitative, i.e., *HR* function, and the other, quantitative , i.e., *AI* function. Here, it is worth mentioning that *HR* and *AI* bear two different dimensions, the former's unit is 1 per unit of time (for considered data $\left[\frac{1}{month}\right]$) and the latter is dimensionless. Their role in the study of system aging behaviour has been discussed in introduction section. However, to the best of our knowledge, not much work has been done where *HR* and *AI* functions

Table 3.4: Kernel estimates of *AI* for *LGT*

t	$\hat{L}_B(t)$	$\hat{L}_E(t)$	$\hat{L}_N(t)$	$\hat{L}_T(t)$
6	0.0430	0.0585	0.0623	0.0633
11	0.0765	0.1058	0.1119	0.1148
15	0.1020	0.1427	0.1501	0.1545
21	0.1383	0.1962	0.2049	0.2105
29	0.3133	0.2642	0.2739	0.2794
33	0.3448	0.2967	0.3065	0.3111
79	0.6124	0.6036	0.6087	0.5948
102	0.7100	0.7157	0.7177	0.6960
212	0.8208	0.9222	0.9834	1.0478

Table 3.5: Kernel estimates of *HR* for *LGT*

t	$\hat{h}_B(t)$	$\hat{h}_E(t)$	$\hat{h}_N(t)$	$\hat{h}_T(t)$
6	0.0013	0.0017	0.0018	0.0018
11	0.0013	0.0018	0.0019	0.0019
15	0.0013	0.0018	0.0019	0.0020
21	0.0013	0.0019	0.0020	0.0020
29	0.0023	0.0020	0.0021	0.0021
33	0.0023	0.0020	0.0021	0.0022
79	0.0026	0.0025	0.0026	0.0025
102	0.0027	0.0028	0.0028	0.0027
212	0.0029	0.0032	0.0035	0.0036

are implemented in assessing the survival of cancer patients. We use four different kernel estimators to address the issue. We find that all considered kernels exquisitely corroborated each other to infer the data.

Table 3.2 and Table 3.3, respectively, depict four different kernel estimates of *AI* and

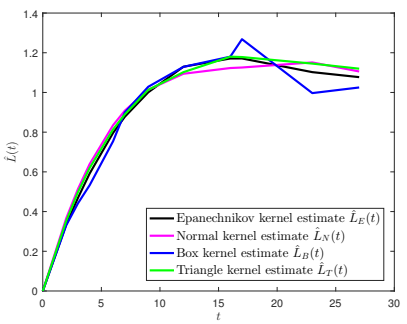

Figure 3.1: *AI* for *HGT*

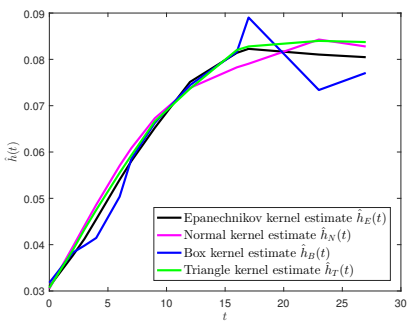

Figure 3.2: *HR* for *HGT*

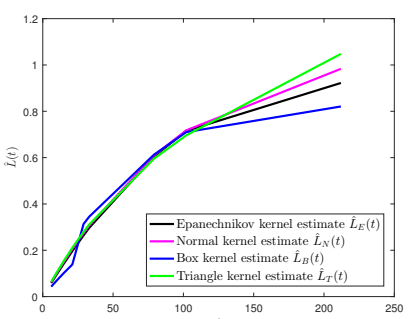

Figure 3.3: *AI* for *LGT*

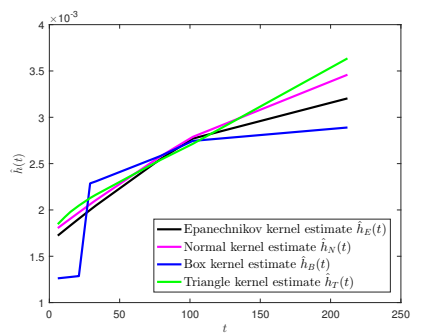

Figure 3.4: *HR* for *LGT*

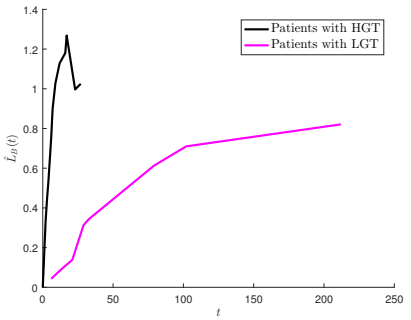

Figure 3.5: *AI*, Box kernel

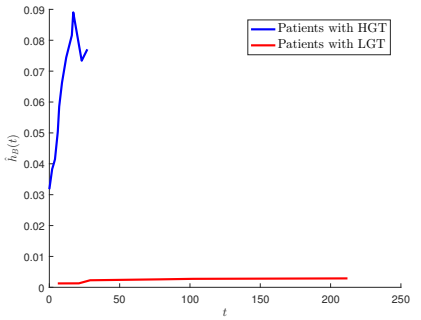

Figure 3.6: *HR*, Box kernel

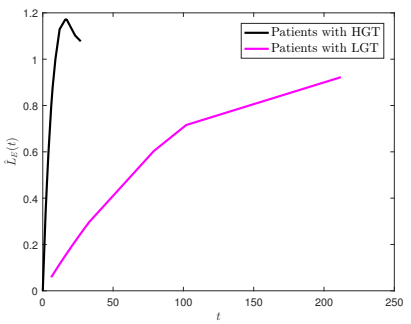

Figure 3.7: *AI*, EPA kernel

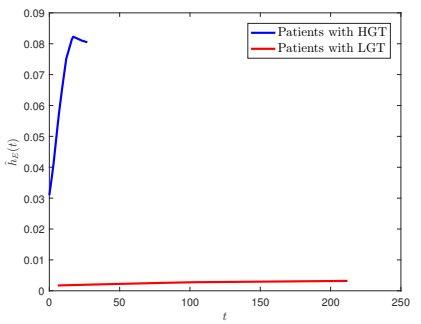

Figure 3.8: *HR*, EPA kernel

HR for patients with *HGT*. The respective four kernel estimates of *AI* and *HR* for patients with *LGT* are given in Table 3.4 and Table 3.5. Note, that for $t = 0$, *AI* estimates are not defined. Figures 3.1–3.12 are plotted for the purpose of drawing inference about the given data (with reference to Table 3.1) from Tables 3.2–3.5. In Figures 3.1–3.4, we have kernel estimates of *AI* for patients with *HGT*, kernel estimates of *HR* for patients with *HGT*, kernel estimates of *AI* for patients with *LGT*, and kernel estimates of *HR* for patients with *LGT*, respectively, which reveal the robustness of the kernels used to evaluate the values of instantaneous hazard rate (given by *HR*) and aging intensity (given by *AI*). While *AI* function is implemented in all four kernel estimators, we find all of them exquisitely corroborates with each other, both for *HGT* (Fig 3.1) and *LGT* (Fig. 3.3) datasets. The same observation is also found while *HR* function highlighting the qualitative aspect of aging is used for *HGT* (Fig 3.2) and *LGT* patients (Fig 3.4). This clearly indicates that both *AI* and *HR* functions could be efficiently implemented to the censored cancer data analysis. Figures 3.5-3.12 represent the differences in the impact of two different tumor grades on patients using *HR* and *AI* functions with respect to each of the four kernels. The sequel of the figures can be found in caption of each figure.

First, we implement *AI* (Fig 3.5) and *HR* (Fig 3.6) in Box kernel estimator for censored *HGT* and *LGT* datasets. As expected, *HGT* shows higher *AI* and *HR*, compared to *LGT*. However, due to the differential nature of these two functions (quantitative vs qualitative)

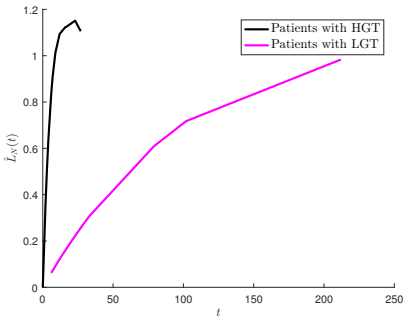

Figure 3.9: *AI*, Normal kernel

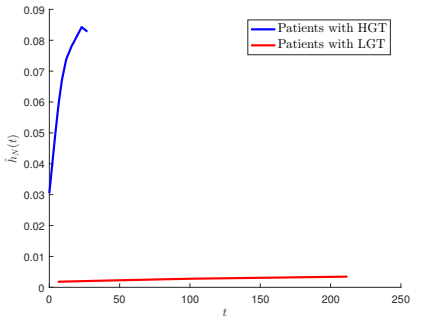

Figure 3.10: *HR*, Normal kernel

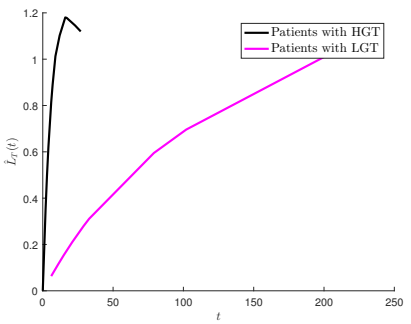

Figure 3.11: *AI*, Triangle kernel

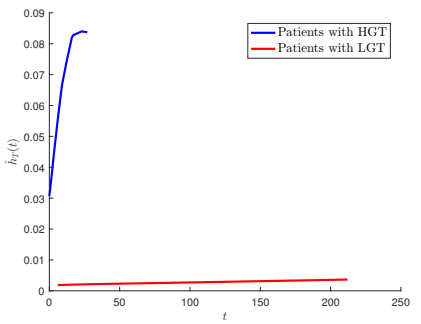

Figure 3.12: *HR*, Triangle kernel

the pattern of curve for *LGT* differs. Since, the representation of aging of a system by hazard rate is qualitative, the curve is relatively flat for *HR* compared to the *AI*. On the other hand, since *AI* quantitatively evaluates the aging property of a system, for *LGT* datasets, lifetime distribution is better represented by *AI* than by *HR* curve. This is worth mentioning here that since *HGT* patients do not survive for longer time period, qualitative (HR) and quantitative (AI) evaluation does not affect much.

Interestingly, similar observation is also found for other three estimators like Epanechnikov (EPA) kernel (Fig 3.7 and 3.8), normal kernel (Fig 3.9 and 3.10) and triangle kernel (Fig 3.11 and 3.12) indicating the importance of implementing *AI* and *HR* functions to the cancer data irrespective of the kernel estimator used. The calculations pertaining to each of the four kernels are given in detail in the Appendix section.

4. Conclusions

The concluding remarks of this paper are compiled as follows.

(i) Cancer statistics is an important domain of cancer treatment management. In this article, we focus on analysing robust statistical methods that can deal with cancer survival data effectively and it can be applied for any survival or life testing data.

(ii) To the best of our knowledge, there are no studies where *HR* and *AI* functions are applied in assessing the survival data from cancer patients. Reports from this work indicate that implementation of *HR* and *AI* functions in human diseases is promising. Therefore, we illustrate the same with detailed analysis using available censored data on cancer-survivals (Martini et al., 1996).

(iii) We use four different kernel estimators to apply *HR* and *AI* functions. Our analysis shows that *HR* and *AI* for patients with HGT are higher than for patients with LGT, as expected, showing a lower survival of HGT patients.

(iv) Since representation of aging in a system by *AI* is more quantitative, *AI* curves are able to provide more information than the *HR* (qualitative) curves (as depicted in their flattened nature) (see LGT curves of Fig 3.5 versus Fig 3.6; Fig 3.7 versus Fig 3.8; Fig 3.9 versus Fig 3.10 and Fig 3.11 versus Fig 3.12). The pattern is particularly prominent for all our LGT curves as we do not have data beyond 50 months for HGT patients (by that time all HGT patients die due to the severity of the disease). On the contrary, as LGT patients survived longer periods of time, we have data until 200 months (approx.). This allows us to visualize the full spectrum of the phenomena of *AI* (with more information) and appreciate the quantitative nature of the function as opposed to less informative flattened pattern for qualitative *HR* function.

(v) Our study strongly indicates that both *AI* and *HR* functions could be efficiently implemented to estimate the survival analysis for cancer patients. We believe, this new avenue of applying *AI* and *HR* functions will be adopted by researchers for implementation in any problem of health sciences or engineering studies.

5. Appendix

In this article, we intend to place the theme directly to the readers and as such we keep at bay the other statistical calculations for discussion in the Appendix section. Here, we give the details of the work done with reference to the aging metrics viz., cdf, pdf, *HR* and *AI* functions. First, we survey HGT patients followed by LGT patients.

5.1. HGT patients

First, we survey HGT patients.

5.1.1 HGT: Box kernel

For HGT patients, Box kernel with bandwidth $b = 6.1113$ (proposed by *MATLAB* R2016a) is used to determine function estimators. To be precise, we state that the estimates of cumulative distribution function (cdf), probability density function (pdf), hazard rate function (HR) and aging intensity function (AI) for patients with HGT using Box kernel are obtained. Accordingly, we receive plots for the following functions as mentioned here:

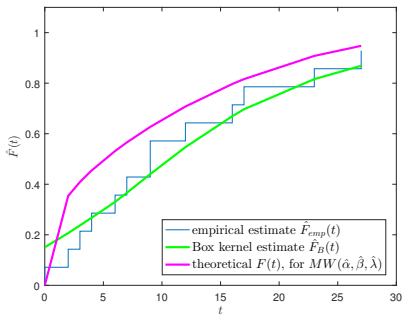

Figure 5.13: cdf for HGT, Box kernel

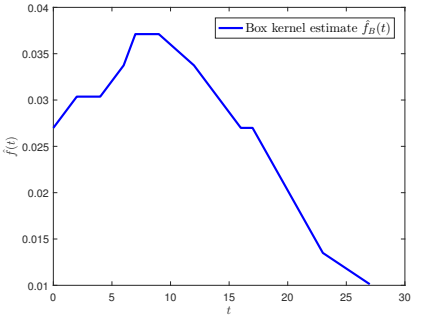

Figure 5.14: pdf for HGT, Box kernel

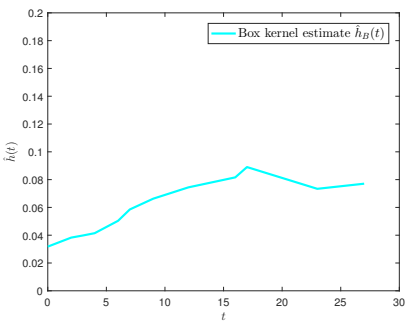

Figure 5.15: HR for HGT, Box kernel

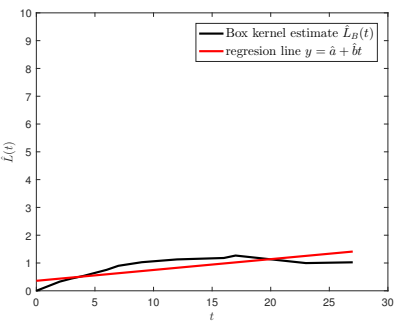

Figure 5.16: AI for HGT, Box kernel

(i) Figure 5.13, empirical cdf (blue step function) and Box kernel estimate of cdf (green function),

(ii) Box kernel estimate of pdf (Figure 5.14),

(iii) Box kernel estimate of HR (Figure 5.15),

(iv) Box kernel estimate of AI (Figure 5.16).

The function $\hat{L}(t)$ presented in Figure 5.16 is seen to oscillate around the linear function $y = a + bt$. So, the AI estimators of parameters of the Modified Weibull distribution $MW(\alpha,\beta,\lambda)$ (with linear AI, see, e.g., Szymkowiak (2020)) are $\hat{\alpha} = \hat{a} = 0.3598$, $\hat{\beta} = \hat{b} = 0.0390$, respectively, and the maximum likelihood estimate is $\hat{\lambda} = 0.3150$. Here, in this section, to obtain the desired value of $\hat{\lambda}$, we make use of the estimator

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} T_i^{\hat{\alpha}} \exp\left(\hat{\beta} T_i\right)}, \tag{5.8}$$

where n is a sample size. The theoretical $F(t)$ for $MW(\hat{\alpha},\hat{\beta},\hat{\lambda})$ with parameters received by Box kernel estimates is shown in Figure 5.13 (magenta function).

5.1.2 *HGT*: Epanechnikov (EPA) kernel

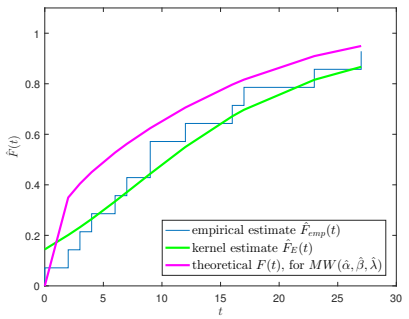

Figure 5.17: cdf for HGT, EPA kernel

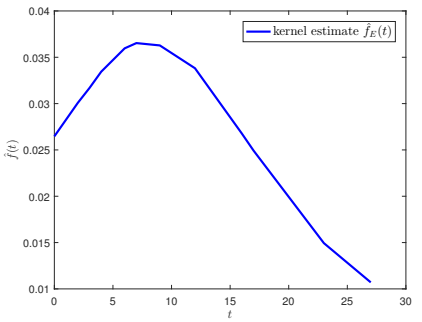

Figure 5.18: pdf for HGT, EPA kernel

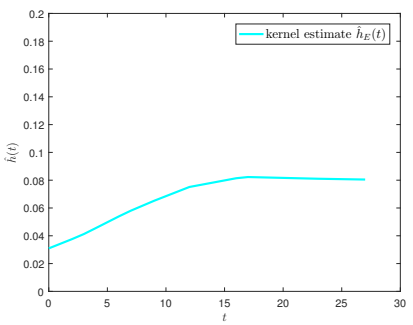

Figure 5.19: HR for HGT, EPA kernel

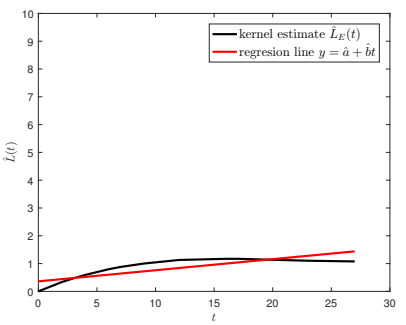

Figure 5.20: AI for HGT, EPA kernel

Next, for HGT patients, we determine the estimates of cdf, pdf, HR and AI by Epanechnikov (EPA) kernel with bandwidth $b = 6.1113$ (proposed by *MATLAB* R2016a). The corresponding plots are listed:

(i) Figure 5.17, empirical cdf (blue step function) and Epanechnikov kernel estimate of cdf (green function),

(ii) Epanechnikov kernel estimate of pdf (Figure 5.18),

(iii) Epanechnikov kernel estimate of HR (Figure 5.19),

(iv) Epanechnikov kernel estimate of AI (Figure 5.20).

The function $\hat{L}(t)$ in Figure 5.20 oscillates around the linear function $y = a + bt$. Then, the AI estimators of corresponding parameters of the Modified Weibull distribution $MW(\alpha,\beta,\lambda)$ (with linear AI) are $\hat{\alpha} = \hat{a} = 0.3606$, $\hat{\beta} = \hat{b} = 0.0400$, respectively, and the maximum likelihood estimate is $\hat{\lambda} = 0.3093$ (using (5.8)). The theoretical $F(t)$ for $MW(\hat{\alpha},\hat{\beta},\hat{\lambda})$ with parameters obtained by Epanechnikov kernel estimates is shown in Figure 5.17 (magenta function).

5.1.3 *HGT*: Normal kernel

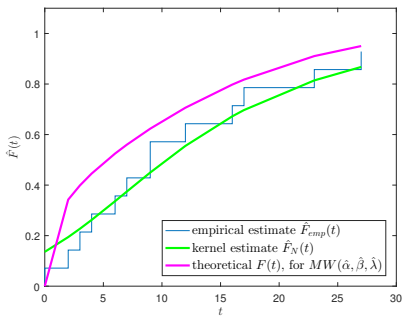

Figure 5.21: *cdf* for *HGT*, Normal kernel

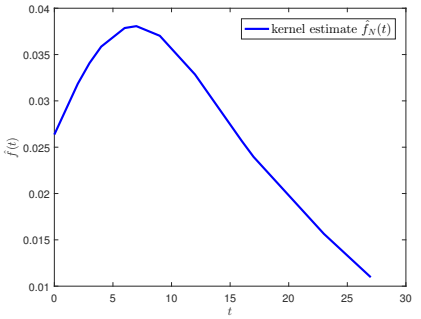

Figure 5.22: *pdf* for *HGT*, Normal kernel

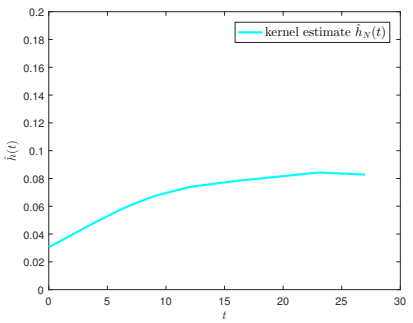

Figure 5.23: *HR* for *HGT*, Normal kernel

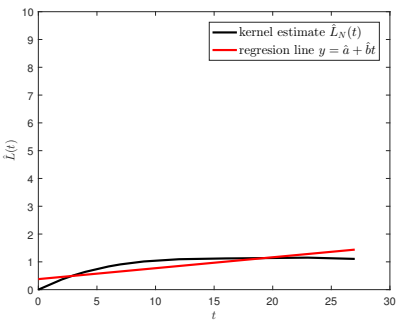

Figure 5.24: *AI* for *HGT*, Normal kernel

For *HGT* patients, Normal kernel with bandwidth $b = 6.1113$ (proposed by *MATLAB* R2016a) is used to find the estimators of *cdf*, *pdf*, *HR* and *AI*. The corresponding plots are listed:

(i) Figure 5.21, empirical *cdf* (blue step function) and Normal kernel estimate of *cdf* (green function),

(ii) Normal kernel estimate of *pdf* (Figure 5.22),

(iii) Normal kernel estimate of *HR* (Figure 5.23),

(iv) Normal kernel estimate of *AI* (Figure 5.24).

The function $\hat{L}(t)$ presented in Figure 5.24 oscillates around $y = a + bt$. Therefore, we can receive the *AI* estimators of parameters of the Modified Weibull distribution $MW(\alpha,\beta,\lambda)$ (with linear *AI*) $\hat{\alpha} = \hat{a} = 0.3789$, $\hat{\beta} = \hat{b} = 0.0393$, respectively, and the maximum likelihood estimate is $\hat{\lambda} = 0.2982$ (using (5.8)). The theoretical $F(t)$ for $MW(\hat{\alpha},\hat{\beta},\hat{\lambda})$ with parameters determined using Normal kernel estimates is shown in Figure 5.21 (magenta function).

5.1.4 *HGT*: Triangle kernel

Using Triangle kernel with bandwidth $b = 6.1113$ (proposed by *MATLAB* R2016a), we receive the estimators of cdf, pdf, HR and AI for patients with HGT. One can refer to the associated plots as given:

(i) Figure 5.25, empirical cdf (blue step function) and Triangle kernel estimate of cdf (green function),

(ii) Triangle kernel estimate of pdf (Figure 5.26),

(iii) Triangle kernel estimate of HR (Figure 5.27),

(iv) Triangle kernel estimate of AI (Figure 5.28).

The function $\hat{L}(t)$ presented in Figure 5.28 can be considered to oscillate around the linear function $y = a + bt$. So, we can determine the AI estimators of parameters of the Modified Weibull distribution $MW(\alpha,\beta,\lambda)$ (with linear AI) as $\hat{\alpha} = \hat{a} = 0.3672$, $\hat{\beta} = \hat{b} = 0.0408$, respectively, and the maximum likelihood estimate is $\hat{\lambda} = 0.2999$ (using (5.8)). The theoretical $F(t)$ for $MW(\hat{\alpha},\hat{\beta},\hat{\lambda})$ with parameters received by Triangle kernel estimates is shown in Figure 5.25 (magenta function).

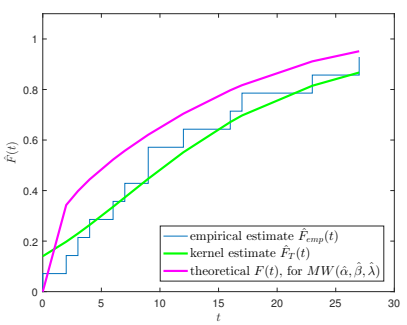

Figure 5.25: cdf for HGT, Triangle kernel

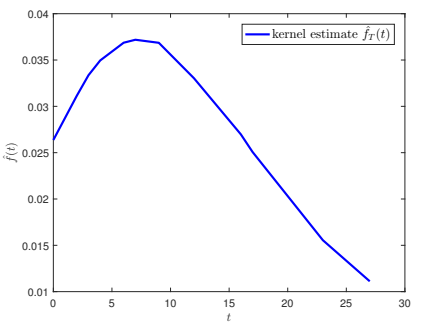

Figure 5.26: pdf for HGT, Triangle kernel

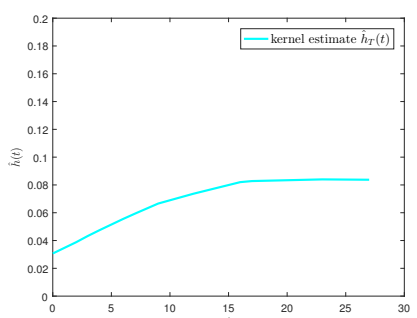

Figure 5.27: HR for HGT, Triangle kernel

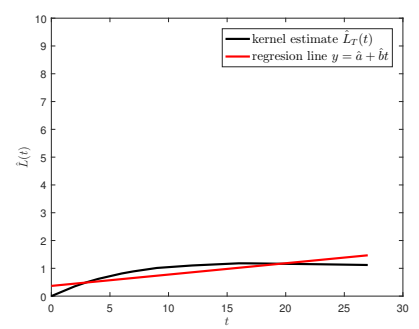

Figure 5.28: AI for HGT, Triangle kernel

5.1.5 Summary for HGT patients

We now summarize our analysis of data for HGT patients by four considered kernels. For each kernel and their AI estimators of parameters of the Modified Weibull distribution, using right type II censored data and Kolmogorov-Smirnov goodness-of-fit test (cf. Agostino and Stephens (1986)) we verify the hypothesis that the data really follow this distribution (with linear AI).

(i) Box kernel: statistics $D^* = 1.1216$ and p-value higher than 0.15

(ii) Epanechnikov kernel: statistics $D^* = 1.1051$ and p-value higher than 0.15

(iii) Normal kernel: statistics $D^* = 1.0787$ and p-value higher than 0.15

(iv) Triangle kernel: statistics $D^* = 1.0795$ and p-value higher than 0.15.

One can note that

$$D^* = \sqrt{n}D + \frac{0.24}{\sqrt{n}}$$

where n is a sample size and

$$D = \max_{1 \le i \le l} \left\{ \frac{i}{n} - F(i), F(i) - \frac{i-1}{n} \right\}.$$

Here, $F(t)$ is the theoretical cdf and l is the number of uncensored data. It means that at the significance level $\alpha < 0.15$, for all considered kernels, we do not reject the hypothesis that data follow the respective Modified Weibull distribution. Moreover, (although the differences are not large) we can notice that the value of statistics D^* is the smallest for Normal kernel. Therefore, we can claim that this kernel function is the best to use in our data analysis and so the Modified Weibull distribution $MW(\alpha, \beta, \lambda)$ with parameters $\hat{\alpha} = 0.3789$, $\hat{\beta} = 0.0393$ and $\hat{\lambda} = 0.2982$ fits the analyzed data best.

5.2. LGT: Normal kernel

For low grade tumor (LGT), we analyze the whole data from Table 3.1 using only Normal kernel with bandwidth $b = 109.2138$ (proposed by *MATLAB* R2016a) to get the function estimators. For patients with LGT we receive the associated plots as:

(i) Figure 5.29, empirical cdf (blue step function) and Normal kernel estimate of cdf (green function),

(ii) Normal kernel estimate of pdf (Figure 5.30),

(iii) Normal kernel estimate of HR (Figure 5.31),

(iv) Normal kernel estimate of AI (Figure 5.32).

The function $\hat{L}(t)$ presented in Figure 5.32 can be considered to oscillate around the linear function $y = a + bt$. Here, the AI estimators of parameters of the Modified Weibull

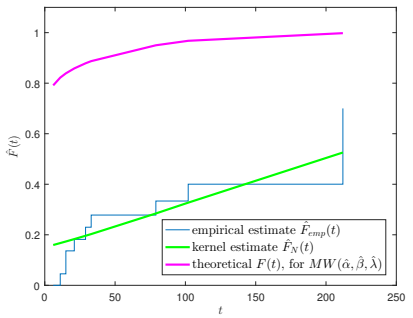

Figure 5.29: cdf for LGT, Normal kernel

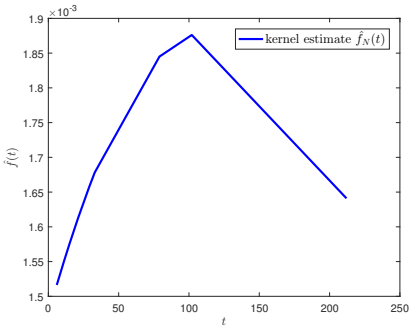

Figure 5.30: pdf for LGT, Normal kernel

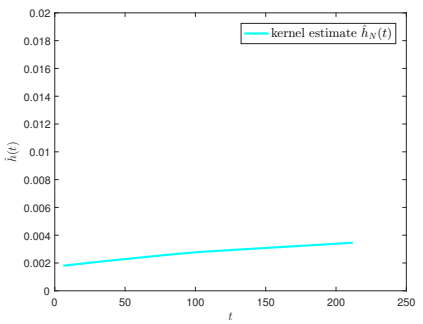

Figure 5.31: HR for LGT, Normal kernel

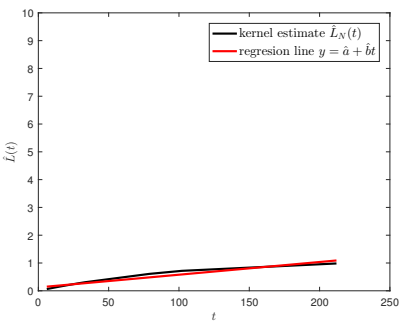

Figure 5.32: AI for LGT, Normal kernel

distribution $MW(\alpha,\beta,\lambda)$ (with linear AI) are $\hat{\alpha} = \hat{a} = 0.1220$, $\hat{\beta} = \hat{b} = 0.0046$, respectively, and the maximum likelihood estimate is $\hat{\lambda} = 1.2252$ (using (5.8)). But we can see in Figure 5.29 that the theoretical distribution function $F(t)$ of the determined (by Normal kernel estimates) Modified Weibull distribution (magenta function) is not close to the empirical estimate $\hat{F}_{emp}(t)$ (blue step function). So, we have to reject the hypothesis that the lifetime of LGT patients follows MW distribution (and this is also the case for the other Box, Epanechnikov, and Triangle kernels under consideration). Moreover, the usage of Kolmogorov-Smirnov goodness-of-fit (see D' Agostino and Stephens (1986)) to verify the hypothesis that data really follow Modified Weibull distribution is controversial because the percent of censored data is higher then 60%.

Acknowledgements

The authors thank the anonymous reviewers and the Editor for their constructive comments which led to the improved version of the manuscript. The first author was partially supported by PUT under grant 0211/SBAD/0123. The corresponding author would like to thank Higher Education Department, Government of Odisha under OHEPEE (Grant No. HE-PTC-WB-02017) and Odisha State Higher Education Council for providing support to carry out the research project under OURIIP, Odisha, India (Grant No. 22-SF-MT-073).

References

D'Agostino, R. B., Stephens, M. A., (1986). Goodness-of-Fit Techniques, New York, NY, USA: Marcel Dekker.

Bhattacharjee, S., Mohanty, I., Szymkowiak, M. and Nanda, A. K., (2022). Properties of aging functions and their means. *Communications in Statistics: Simulation and Computation*, https://doi.org/10.1080/03610918.2022.2141257

Bhattacharjee, S., Nanda, A. K., and Misra, S. K., (2013a). Reliability analysis using aging intensity function. *Statistics and Probability Letters*, 83 (5), pp. 1364–1371.

Bhattacharjee, S., Nanda, A. K., and Misra, S. K., (2013b). Inequalities involving expectations to characterize distributions. *Statistics and Probability Letters*, 83 (9), pp. 2113–2118.

Bowman, A. W., Azzalini, A., (1997). Applied smoothing techniques for data analysis, New York: Oxford University Press Inc.

Daniel, W. W., Cross, C. L., (2014). Biostatistics: Basic Concepts and Methodology for the Health Sciences, Tenth Edition, Wiley India Pvt Ltd.

DiNardo, J., Tobias, J., (2001). Nonparametric density and regression estimation, *The Journal of Economic Perspectives*, 15, pp. 11–28.

Emmerson, J., Brown, J. M., (2021). Understanding Survival Analysis in Clinical Trials, *Clinical Oncology*, 33 (1), pp. 12–14.

Giri, R. L., Nanda, A. K., Dasgupta, M., Misra, S. K., and Bhattacharjee, S., (2023). On aging intensity function of some Weibull models. *Communications in Statistics-Theory and Methods*, 52(01), pp. 227-262. DOI: 10.1080/03610926.2021.1910845

Jiang, R., Ji, P., and Xiao, X., (2003). Aging property of univariate failure rate models. *Reliability Engineering and System Safety*, 79, pp. 113–116.

Klein, J. P., Moeschberger, M. L., (2003). Survival Analysis Techniques for Censored and Truncated Data: Statistics for Biology and Health, Series Editors: Dietz, K., Gail, M., Krickeberg, K., Samet, J,. and Tsiatis, A., 2nd Edition, Springer-Verlag New York, Inc.

Martini, N., Huvos, A. G., Burt, M. E., Heelan, R. T., Bains, M. S., Mccrmack, P. M., Rusch, V. M., Weber, M., Downey, R. J., and Ginsberg, R. J., (1996). Predictions of Survival in Malignant Tumors of the Sternum, *Journal of Thoracic and Cardiovascular Surgery*, 111, pp. 96–106.

Miladinovic, B., (2008). Kernel density estimation of reliability with applications to extreme value distribution, Graduate Theses and Dissertations submitted in University of South Florida, http://scholarcommons.usf.edu/etd/408.

Misra, S. K., Bhattacharjee, S., (2016). Properties of Weibull models, *Far East Journal of Mathematical Sciences*, 100(12), pp. 1965–1979.

Misra, S. K., Bhattacharjee, S., (2018). A case study of aging intensity function on censored data, *Alexendria Engineering Journal*, 57, pp. 3931–3952.

Nanda, A. K., Bhattacharjee, S., and Alam, S. S., (2006). Properties of proportional mean residual life model, *Statistics and Probability Letters*, 76 (9), pp. 880–890.

Nanda, A. K., Bhattacharjee, S., and Alam, S. S., (2007). Properties of aging intensity function, *Statistics and Probability Letters*, 77, pp. 365–373.

Nanda, A. K., Bhattacharjee, S., and Balakrishnan, N., (2010). Mean residual life fuction, associated orderings and properties, *IEEE Transactions on Reliability*, 59(1), pp. 55–65.

Nayak, A., Kumar, S., Singh, S. P., Bhattacharyya A., Dixit A., and Roychowdhury A., (2021). Oncogenic potential of ATAD2 in stomach cancer and insights into the protein-protein interactions at its AAA+ATPase domain and bromodomain, *Journal of Biomolecular Structure and Dynamics*, pp. 1–17.

Rosen, K., Prasad, V., Chen, E. Y., (2020). Censored patients in Kaplan-Meier plots of cancer drugs: An empirical analysis of data sharing, *European Journal of Cancer*, 141, pp. 152-161. doi: 10.1016/j.ejca.2020.09.031.

Shaked, M., Shanthikumar, J. G., (2007). *Stochastic Orders (Springer Series in Statistics)*. San Diego.

Szymkowiak, M., (2018). Characterizations of Distributions Through Aging Intensity, *IEEE Transactions on Reliability*, 67, pp. 446–458.

Swain, P., Bhattacharjee S. and Misra, S. K., (2021). A Case Study to Analyze Ageing Phenomenon in Reliability Theory, *Reliability: Theory and Applications*, 16(4 (65)), pp. 275–285.

Szymkowiak, M., (2019). Measures of aging tendency, *Journal of Applied Probability*, 56(2), pp. 358–383.

Szymkowiak, M., (2020). Lifetime analysis by aging intensity functions studies in systems. In Decision and Control. Series editor. Vol. 196. Switzerland: Springer Nature.

Gamrot, W., (2012). Estimation of Finite Population Kurtosis under Two-Phase Sampling for Nonresponse. *Statistical Papers*, 53, pp. 887–894.

Gamrot, W., (2013). Maximum Likelihood Estimation for Ordered Expectations of Correlated Binary Variables. *Statistical Papers*, 54, pp. 727–739.

Gamrot, W., (2012). Estimation of Finite Population Kurtosis under Two-Phase Sampling for Nonresponse. *Statistical Papers*, 53, pp. 887–894.

Gamrot, W., (2013). Maximum Likelihood Estimation for Ordered Expectations of Correlated Binary Variables. *Statistical Papers*, 54, pp. 727–739.

Särndal, C-E., Swensson, B., Wretman, J., (1992). *Model Assisted Survey Sampling*, New York: Springer.

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 129–146, https://doi.org/10.59170/stattrans-2023-067
Received – 22.08.2021; accepted – 30.01.2023

Investigating the factors of blockchain technology influencing food retail supply chain management: a study using TOE framework

Manish Mohan Baral[1], Venkataiah Chittipaka[2], Surya Kant Pal[3], Subhodeep Mukherjee[4], Hari Shankar Shyam[5]

Abstract

In history, the food sector has remained the most vulnerable and is accountable for many crises and food scandals, so to avoid this in the near future, it is essential to have better control over the flow of products and the usage of blockchain technology can achieve this control in the supply chain, which can add trust, transparency, and traceability in the entire system. This paper investigates the factors affecting blockchain adoption in the food retail supply chain and creates awareness among retail managers for its adoption in their operations. A structured literature review is performed to identify the factors and a TOE framework is adopted for the research. Factors of technology, organization and environment are taken as independent variables; the intention to adopt the technology is taken as a mediating variable and blockchain adoption is considered a dependent variable. The findings contributed that TOE factors contribute to blockchain adoption by keeping the intention to adopt the technology as a mediating variable.

Key words: trust, transparency, traceability, blockchain, TOE framework.

[1] Department of Operations, GITAM School of Business, GITAM, Visakhapatnam, Andhra Pradesh, India. E-mail: manishmohanbaral.31@gmail.com.

[2] School of Management Studies, Indira Gandhi National Open University, Delhi, India. E-mail: venkatchitti@gmail.com.

[3] Corresponding Author. Department of Mathematics, Sharda School of Basic Sciences and Research, Sharda University, Uttar Pradesh, India. E-mail: suryakantpal6676@gmail.com. ORCID: https://orcid.org/0000-0001-8701-2095.

[4] Department of Operations, GITAM School of Business, GITAM, Visakhapatnam, Andhra Pradesh, India. E-mail: subhodeepmukherjee92@gmail.com.

[5] Sharda School of Business Studies, Sharda University, Greater Noida, Uttar Pradesh, India. E-mail: harishankar.shyam@sharda.ac.in.

1. Introduction

Food products travel a lot from agriculture farms to customers in different cities. In India, other states grow different food products, so it remains challenging for organizations to keep track of food safety (Zhang et al., 2020). Food industries are responsible for approximately 30% of energy consumption globally, leading to 22% greenhouse gas emissions. There has been increasing consumer awareness regarding food safety in recent times. Perishable food items are at a higher risk of getting wasted or contaminated. In 2013 in Serbia, a food scandal involved toxic elements found in milk items (Makhdoom et al., 2019). In 2018, another food scandal was revealed regarding baby milk powder in one of the largest milk manufacturing companies. The leading cause of all the scandals happening worldwide is the lack of transparency in the FSC. The customers purchase retail foods from different malls, departmental stores, local shops, etc. (Wong et al., 2020). The Indian retail market is approx. 500 billion USD and growing at a rate of 7% annually.

BT technology (BT) is a distributed information base shared among a distributed organization. It comprises blocks linked together in the organization networks. When a component is added to the BT, it cannot be modified, transforming a BT into a permanent record of past action (Mukherjee, Nagariya, Baral, et al., 2022). Each block of the information base is connected to the other. Each block of the chain comprises many exchanges and is checked by every individual inside (Antonucci et al., 2019). In this digital age, all plans of action have been going through changes because of the latest discoveries in information and communication technology (ICTs). Disruptive BT technology changes the organizations' business models and how they are done. BT developed as an innovation for performing exchanges in the cryptographic money sector (Mukherjee, Baral, & Chittipaka, 2022). BT found many applications in the finance and banking industry but has recently been used in other areas like the supply chain (SC), the education sector, and many more. BT is still developing, with many innovations coming forward in the near future. BT can be a challenge and an opportunity for industries. BT can improve the food supply chain (FSC) industry by creating trust, transparency, security, accountability, and efficiency (Mukherjee et al., 2021). BT can be a solution to SC traceability problems and can generate closer trust in the relationships. This trust will not be limited to suppliers but the entire SC (Gökalp et al., 2022).

This paper investigates the factors affecting the food retail supply chain (FRSC) while adopting BT. We have used the TOE framework, where the constructs of technology, organization, and environment are taken as independent variables. Intention to adopt the technology is taken as a mediating variable and blockchain adoption (BA) as a dependent variable. The research is carried out with retail stores in different stores in India.

The rest of the paper is organized as follows. Section 2 comprises a literature review followed by research methodology in Section 3. Section 4 provides results, Section 5 comprises discussions, Section 6 comprises managerial implications, and Section 7 contains a conclusion.

2. Literature review

2.1. Blockchain Technology in Food Supply Chain Management

Food industries can expect better outcomes after adopting BT in their process. There are numerous ongoing innovations in the field of FSC using BT. BT can alter SC supportability. Use cases show organizations trying to execute BT into their SC activities for traceability of items, as on account of Maersk, Provenance, and Walmart (Kamilaris et al., 2019). Limiting fake items has likewise been an objective of specific BT applications (Lin & Liao, 2017). Notwithstanding the potential of BT benefits for improving traceability in an organization, the quantity of utilization cases applying BT for supportability is restricted. In contrast, organizations keep battling with the more all-encompassing parts of sustainability (Atlam et al., 2018).

Tracking the shipment is a significant part of SC. BT will assist one with checking continuously if the shipment has been taken care of appropriately and has shown up on time at some random area. BT will help track the lost or tampered items in the flow of FRSC (Al-Jaroodi & Mohamed, 2019). Likewise, BT-based trades will help the FR purchase or sell from one another just as merchants through the BT-shared record (Hastig & Sodhi, 2020). BT can collect information identified from the customer purchasing behaviour, request situation pattern, etc. This information can be used to keep the actual product inventory, like the just-in-time inventory facility (Mukherjee, Baral, & Venkataiah, 2022). BT will decrease the danger of fake financial exchanges (Hew et al., 2020).

2.2. Conceptual Framework development

2.2.1. Technological factors (TF)

Perceived benefits (PB): Researchers and analysts believe that BT-engaged SC gains benefits, including progressing information sharing, cost decrease, adequacy, straightforwardness, tracking, tracing, and improving operational excellence (Roy, Babakerkhell, et al., 2022). BT-engaged SC aims to grow the association's effectiveness and reality in the market (Hassani et al., 2018; van Hoek, 2019). As required, the introduction saw an advantage as an essential factor for an association to grasp BT in its SC. The higher the SC execution advantage is seen by an organization, the more probable to accept BT.

H1: PB influences the manager's intention to adopt BT.

Cost (C): Costs for adopting the latest technology refer to all the expenses incurred in the adoption process (Clohessy et al., 2019). In adopting BT, the cost will be incurred, like modifying the technology systems, training employees, and buying the latest devices and software for the technology (Saberi et al., 2019). But earlier research shown that the companies are already using RFID or IoT-based technologies in their FSC, so it will add an advantage as the organization needs to upgrade their technologies, which will have less costs with more benefits (Roy, Chekuri, et al., 2022).

H2: C influences the manager's intention to adopt BT.

Relative advantage (RA): RA refers to the degree to which the latest innovations are perceived in the organizations' context and how they are adopting them (Mukherjee & Chittipaka, 2021). RA is measured in terms of time, effort, profits, cost reduction, and production increase. In this research, this construct refers to improving FSC professionals' performance using BT technologies. This can offer transparency to the existing system and improve the suppliers' performance (Mendling et al., 2018).

H3: RA influences the manager's intention to adopt BT.

Security (S): S remains the most concerning factor for organizations, as many feel that the data they share might get destroyed or tampered (Queiroz et al., 2020). This creates a lack of trust and confidence among the technology providers and the organization (Alsetoohy et al., 2019). But BT is the opposite of other technologies as it provides complete security for the organizations' data.

H4: S influences the manager's intention to adopt BT.

2.3.2. Organizational factors (OF)

Top management support (TMS): TMS means the level of support and resources the top management puts into technology adoption (Lengoatha & Seymour, 2020). Management encouragement and support are required in the adoption process (Mezquita et al., 2019). Top management helps create coordination and solve the conflicts between the technology providers and the organization (Mukherjee, Baral, Pal, et al., 2022). Supports that are required for the technology adoption, which can be provided by top management, are funds for the projects, training of the staff, motivation for the change resistance and creating a belief that with the adoption of the new technologies, there will be no job loss (Hassan, 2017). BT system adoption will improve organizational changes if the project gets complete top management support.

H5: TMS influences the manager's intention to adopt BT.

Organizational readiness (ORN): ORN means whether the organizations are ready to adopt new technologies. The things looked after in ORN are whether the organization can incur innovations cost (Tashkandi & Al-Jabri, 2015). Successful technology adoption can happen only if adequate resources, knowledge, and top management support exist.

H6: ORN influences the manager's intention to adopt BT.

Blockchain knowledge (BK): refers to the organizations' employees' experience with technology adoption (Kamble et al., 2019). Technical knowledge is vital for adopting new technologies in organizations (Chiu et al., 2017). Organizations need to provide adequate training for the employees in the area of BT so that they become habituated to working with it and face no problems.

H7: BK influences the manager's intention to adopt BT.

2.3.3. Environmental factors (EF)

Competitive pressure (CP): Companies need to change their technologies to remain growing in the market. CP refers to the organization's pressure from its competitors (Pateli et al., 2020). If the organizations do not adopt the latest innovations, they might go out of the market and incur a loss. Using innovative technology strengthens the organization's position in the market and with the customers (Yusof et al., 2018). They will be able to provide better service to the customers.

H8: CP influences the manager's intention to adopt BT.

Regulatory environment (RE): Some latest technologies come with regulations in different countries or markets (Zhu et al., 2006). This refers to the country's government's policies, rules, and rules for the latest technology. Organizations adopting this technology need to follow this (Xu et al., 2017).

H9: RE influences the manager's intention to adopt BT.

Government support (GS): Organizations need GS for any latest technology (Oliveira et al., 2014). This support can be in many forms like tax rebates, giving proper guidelines for the technology adoption, monitoring and advising in the adoption process (Puklavec et al., 2018). Without any GS no companies can adopt innovative technology in their operations.

H10: GS influences the manager's intention to adopt BT.

Intention to adopt the technology (I): Intention can be the willingness to embrace the organizations (Gangwar et al., 2015). Intention to adopt depends from person to

person and their knowledge towards technology. In this paper intention of the retail managers is being measured as a meditating variable. The intention to adopt the latest technology has been measured in earlier research (Gangwar, 2020).

H11: There is a positive relationship between the manager's intentions and the adoption of BT.

3. Research Methodology

Secondary data was collected through secondary sources like literature reviews and various other reports (Roy, Baral, et al., 2022). Primary data was collected through a structured questionnaire prepared with the consultant of the qualified persons. The questionnaire was checked and scrutinized by qualified professors and industry professionals in their respective academic fields before the survey began. The questionnaire utilized a seven-point Likert scale for measuring the constructs. Each sub-factor consisted of at least three indicators. A pilot survey was conducted by taking a sample size of 50. After that, the respondents, professors, and industry persons included the requirements and suggestions given by the respondents, professors, and industry persons included in the final questionnaire. The target crowd was retail stores in India, and a stratified random sampling approach was adopted. The retail stores mainly contacted retail managers, floor managers, and procurement managers for the surveys. The questionnaires were sent to 420 respondents from different retail stores through the mail, but only 303 respondents returned usable questionnaires, which were valid for analysis. Exploratory factor analysis (EFA) was done, and structural equation modeling (SEM) was performed to get the results. SPSS 20.0 was used to test reliability and EFA. AMOS version 22.0 was used for SEM.

Table 1: Demographics of the respondents

SL. NO	CHARACTERISTICS	PERCENTAGE
A	**Gender**	
1	Male	59
2	Female	41
B	**RESPONDENTS CURRENT POSITION**	
1	Department Manager	43
2	Retail Manager	32
3	Procurement Management	25
C	**BLOCKCHAIN ADOPTION**	
1	Already adopted	19
2	Not adopted	69
3	Adopted but not using it properly	12

4. Results

4.1. Reliability and Validity

The reliability test was performed for each construct based on Cronbach's alpha value, introducing Cronbach's alpha for the constructs (Fornell & Larcker, 1981). Table 2 shows the value of α, composite reliability, and average variance extracted.

4.2. Exploratory Factor Analysis (EFA)

The exploratory factor analysis (EFA) was performed at the initial stage to group the variables having similar properties, and each variable can be grouped under different factors during this process. Table 2 displays the KMO values for all the perspectives: TF (0.816), OF (0.807), and EF (0.729). This Rotated Component Matrix is important for interpreting the results of the analysis. Rotation helps group the items; each group contains more than one item, simplifying the structure. Hence, this is the aim of the goal of rotation. In this research, we have achieved this aim.

Table 2: Cronbach's alpha, Composite Reliability, Average variance extracted and KMO values

Construct	Latent Variables	No. of items	Measurement entry	Cronbach's alpha (α)	CR	AVE	KMO	Factor Loadings
Technological	Relative Advantage	4	RA1, RA2, RA3, RA4	0.8490	0.8662	0.6191	0.8160	0.761, 0.818, 0.842, 0.720
	Security	4	S1, S2, S3, S4	0.8300	0.8628	0.6117		0.813, 0.746, 0.756, 0.811
	Perceived Benefits	3	PB1, PB2, PB3	0.8400	0.8802	0.7106		0.848, 0.889, 0.789
	Cost	3	C1, C2, C3	0.7250	0.8297	0.6212		0.846, 0.832, 0.675
Organizational	Top management support	4	TMS1, TMS2, TMS3, TMS4	0.8460	0.8699	0.6269	0.8070	0.773, 0.823, 0.853, 0.711
	Organizational readiness	4	ORN1, ORN2, ORN3, ORN4	0.8290	0.8657	0.6174		0.815, 0.750, 0.759, 0.816
	Blockchain knowledge	3	BK1, BK2, BK3	0.7210	0.8334	0.6271		0.838, 0.839, 0.689

Table 2: Cronbach's alpha, Composite Reliability, Average variance extracted and KMO values (cont.)

Construct	Latent Variables	No. of items	Measurement entry	Cronbach's alpha (α)	CR	AVE	KMO	Factor Loadings
Environmental	Competitive pressure	4	CP1, CP2, CP3, CP4	0.8340	0.8844	0.6612	0.7290	0.839, 0.884, 0.888, 0.609
	Regulatory environment	3	RE1, RE2, RE3	0.7310	0.8513	0.6573		0.843, 0.858, 0.725
	Government support	4	GS1, GS2, GS3, GS4	0.8310	0.8837	0.6552		0.799, 0.805, 0.813, 0.821

4.3. Structural Equation Modeling

To test the hypothesis, SEM was used. AMOS 22.0 was utilized for this research because of its powerful graphic representations and user-friendly interfaces (Byrne, 2010). This section represents the outputs of hypothesis testing. The results of the significant paths of the model are shown here. Figure 1, 2, 3 represent the final model and the latent variables, along with their indicators, mediating, and dependent variable.

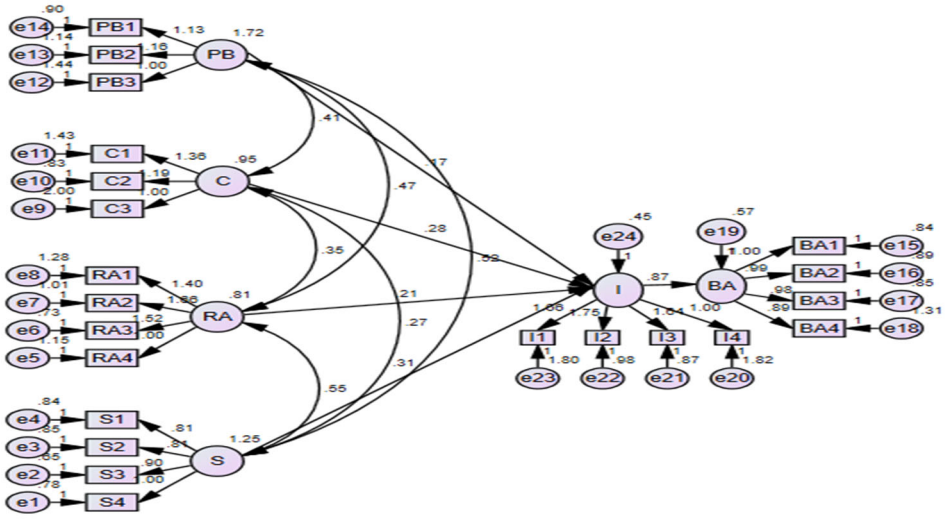

Figure 1: Final SEM for Technological Factor

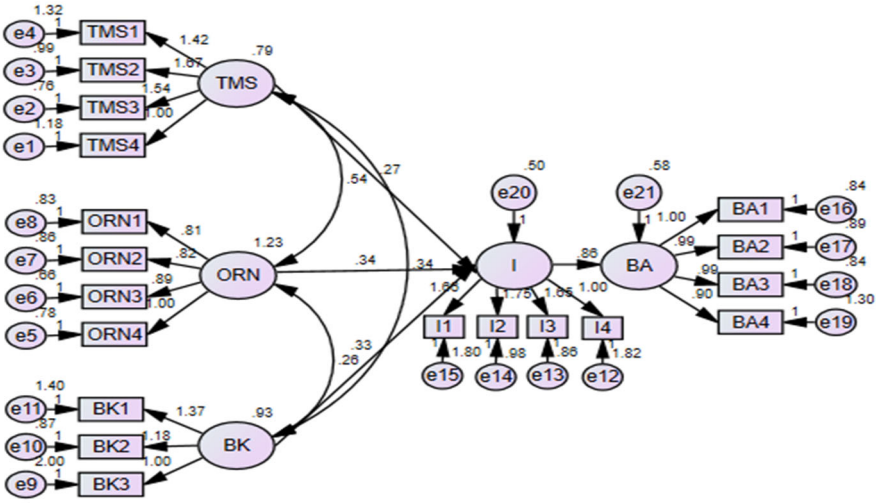

Figure 2: Final SEM for Organizational Factor

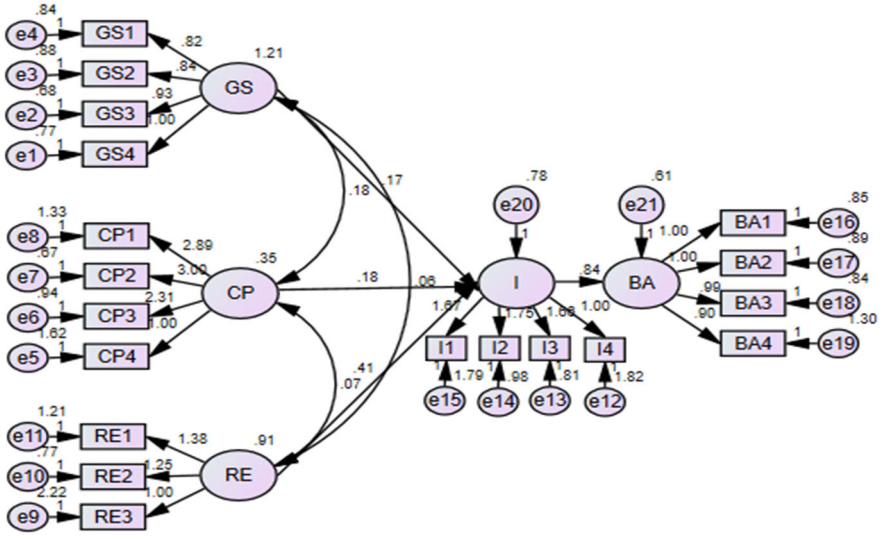

Figure 3: Final SEM for Environmental Factor

4.4. Structural Model Goodness of Fit

The above three models (Figure I, Figure II, and Figure III) show the latent variables and their indicators and a mediating variable with four indicators that contributed significantly to the dependent variable, which also had four indicators. The final output is shown in Table 3.

Table 3: SEM Output

Specification		Chi-square	DF	P-Level	CMIN/DF	RMSEA	CFI	NFI	IFI	GFI	AGFI
TP	Model	491.461	198	0	2.482	0.070	0.913	0.864	0.914	0.874	0.84
	Recommended Standard	-	-	<0.05	<3.0	<0.08	0-1.0	0-1.0	0-1.0	0-1.0	>0.80
OP	Model	399.983	145	0	2.759	0.076	0.911	0.869	0.912	0.88	0.842
	Recommended Standard	-	-	<0.05	<3.0	<0.08	0-1.0	0-1.0	0-1.0	0-1.0	>0.80
EP	Model	368.056	145	0	2.538	0.071	0.918	0.873	0.919	0.89	0.855
	Recommended Standard	-	-	<0.05	<3.0	<0.08	0-1.0	0-1.0	0-1.0	0-1.0	>0.80

5. Discussion

The primary aim of this research was to determine the factors affecting the FRSC in India's FR industry. Another objective was to create awareness among the managers towards BT and its advantages. In the semi-urban areas, knowledge about BT was less than in urban areas. For this, a structured literature review was done to identify the factors affecting BT. TOE framework was considered for the research; prior research was conducted using the TOE framework in many sectors like cloud computing in healthcare, intelligent agent technology in the hotel industry, internet of things adoption in the agriculture sector. BT adoption in the retail industry entirely depends on the retail managers' intention to implement this technology in their retail stores (Seebacher & Schüritz, 2017). The managers whose operations depend upon their decisions. The managers hold the sole responsibility for achieving technical excellence in retail operations. As found out in the research, many problems may arise with BT's adoption, like a lack of knowledgeable people in the area. Another reason may be different state laws compared to national laws.

The factor TF comprised latent variables like PB, RA, S, and C. Each of the constructs had three or more indicators. The recommended level was Cronbach's alpha, and composite reliability values were above 0.7. There were 14 indicators in measuring TF constructs' impact as an independent variable in the BA as a dependent variable, with I as a mediating variable. The KMO value of TF was 0.816, which is also above the recommended level of 0.6. The total variance explained was 69.298%, and in the rotated component matrix, the variables were grouped under four groups. Then, the SEM was performed in AMOS 22.0, CMIN/Df was 2.482, and all the fit indices were within the acceptance level. Hence, the model shows the goodness of fit. Prior research supports that factors like C, RA, and S were taken in BT for the banking industry (Önder & Treiblmaier, 2018). The factors PB, S, and C were found in the prior research using the

technology adoption model (Kumar, 2014). PB was found in another research for innovative education (Salam et al., 2016). So, all the TF constructs' hypotheses were accepted and fit the model well.

The factor OF comprised latent variables like TMS, ORN, and BK. Each of them had three or more indicators. The recommended level was Cronbach's alpha, and composite reliability values were above 0.7. A total of 11 indicators measured the impact OF constructs as an independent variable in the BA as a dependent variable, with I as a mediating variable. The KMO value of OP was 0.807, which is also above the recommended level of 0.6, which allows the data for factor analysis. The total variance explained was 66.969%, and in the rotated component matrix, the variables were grouped under three groups. Then, the SEM was performed in AMOS 22.0, CMIN/Df was 2.759, and all the fit indices were within the acceptance level. Hence, the model shows the goodness of fit. TMS and ORN constructs were found in the education sector's prior research (Clohessy & Acton, 2019). ORN factors were also found in the previous studies that supported these results. BK was found to be an essential factor as, without its knowledge, the adoption process cannot succeed, so the organizations should look into this so that the employees are given proper and adequate BT training (Teller et al., 2018).

The factor EF comprised latent variables like GS, CP, and RE. Each of them had three or more indicators. The recommended level was Cronbach's alpha, and composite reliability values were above 0.7. A total of 11 indicators measured the impact OF constructs as an independent variable in the BA as a dependent variable, with I as a mediating variable. The KMO value of TP was 0.729, which is also above the recommended level of 0.6. The total variance explained was 66.829%, and in the rotated component matrix, the variables were grouped under three groups. Then the SEM was performed in AMOS 22.0, CMIN/Df was 2.538, and all the fit indices were within the acceptance level. Hence, the model shows the goodness of fit. RE varies from country to country and state to state, so it is essential to go through the government's rules and regulations before adopting the technology. This research's results were also supported by the previous study using the same constructs like GS, RE, and CP (Queiroz et al., 2020).

6. Managerial implications

This research aimed to identify the factors affecting BT adoption in FRSC. The survey was conducted with the retail, department, and procurement managers. It was found that the awareness for adopting BT is much less as they do not want to change the technology they are using now. But after making them understand the benefits of BT in FR, they agreed to use it in their stores with the management's support. This

research used intention as a mediating variable for BT's adoption. The retail store is being entirely run under managers' supervision, so their willingness and attitude to adopt the latest technology are essential. For this, top management also needs to give a lot of support in training, funds allocation, and many more. Managers need to change their working styles with time as the latest technology makes the process faster and easier, but initial adoption will be difficult, which needs to be overcome. Only organizations can achieve customer satisfaction, reduce procurement timing, and have greater profits in the long run.

7. Conclusion

This research was conducted to determine the factors affecting BT adoption in the FRSC and to create awareness among the managers to adopt BT in their operations for better results. A structured literature review was conducted to identify the TOE factors used in the research. TOE factors were used in many previous studies on technology adoption, like RFID, IoT, cloud computing, intelligent agent technology, and many more. With these factors, a questionnaire was developed for the survey. The questionnaires were sent to retail stores across India through online mode. The results were analysed using EFA and SEM techniques. Eleven hypotheses supported this, and the three models fit well. Managers need to understand the latest technologies and implement them in their process with the management's support. This will create a significant impact on the way they deliver services to the customers.

7.1. Future research

This research was mainly concentrated on the FR sectors as the research can also be extended to other industries like fashion retail, e-commerce, and restaurants. This research was done for the only country to be extended to other countries. Comparative analysis can be performed between developing and developed countries across the world.

References

Al-Jaroodi, J., Mohamed, N., (2019). Blockchain in Industries: A Survey. *IEEE Access*, 7, pp. 36500–36515. https://doi.org/10.1109/ACCESS.2019.2903554.

Alsetoohy, O., Ayoun, B., Arous, S., Megahed, F. and Nabil, G., (2019). Intelligent agent technology: what affects its adoption in hotel food supply chain management? *Journal of Hospitality and Tourism Technology*, 10(3), pp. 317–341. https://doi.org/10.1108/JHTT-01-2018-0005.

Antonucci, F., Figorilli, S., Costa, C., Pallottino, F., Raso, L. and Menesatti, P., (2019). A review on blockchain applications in the agri-food sector. In *Journal of the Science of Food and Agriculture*, Vol. 99, Issue 14, pp. 6129–6138. John Wiley and Sons Ltd. https://doi.org/10.1002/jsfa.9912.

Atlam, H. F., Alenezi, A., Alassafi, M. O. and Wills, G. B., (2018). Blockchain with Internet of Things: Benefits, challenges, and future directions. *International Journal of Intelligent Systems and Applications*, 10(6), pp. 40–48. https://doi.org/10.5815/ijisa.2018.06.05.

Byrne, B. M., (2010). *Structural equation modeling with AMOS: basic concepts, applications, and programming (multivariate applications series)*. Taylor & Francis Group, 396, 7384.

Chen, G., Xu, B., Lu, M. and Chen, N.-S., (2018). Exploring blockchain technology and its potential applications for education. *Smart Learning Environments*, 5(1). https://doi.org/10.1186/s40561-017-0050-x.

Chiu, C.-Y., Chen, S. and Chen, C.-L., (2017). An integrated perspective of TOE framework and innovation diffusion in broadband mobile applications adoption by enterprises. *International Journal of Management, Economics and Social Sciences (IJMESS)*, 6(1), pp. 14–39.

Clohessy, T., Acton, T., (2019). Investigating the influence of organizational factors on blockchain adoption: An innovation theory perspective. *Industrial Management and Data Systems*, 119(7), pp. 1457–1491. https://doi.org/10.1108/IMDS-08-2018-0365.

Clohessy, T., Acton, T. and Rogers, N., (2019). Blockchain Adoption: Technological, Organisational and Environmental Considerations. In *Business Transformation through Blockchain*, pp. 47–76. Springer International Publishing. https://doi.org/10.1007/978-3-319-98911-2_2.

Fornell, C., Larcker, D. F., (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.1177/002224378101800104.

Gangwar, H., (2020). Big Data Analytics Usage and Business Performance: Integrating the Technology Acceptance Model (TAM) and Task Technology Fit (TTF) Model. *Electronic Journal of Information Systems Evaluation*, 23(1), pp. 45–64. https://doi.org/10.34190/ejise.20.23.1.004.

Gangwar, H., Date, H. and Ramaswamy, R., (2015). Understanding determinants of cloud computing adoption using an integrated TAM-TOE model. *Journal of*

Enterprise Information Management, 28(1), pp. 107–130. https://doi.org/10.1108/JEIM-08-2013-0065.

Ghosh, J., (2019). The Blockchain: Opportunities for Research in Information Systems and Information Technology. In *Journal of Global Information Technology Management*, Vol. 22, Issue 4, pp. 235–242. Taylor and Francis Inc. https://doi.org/10.1080/1097198X.2019.1679954.

Gökalp, E., Gökalp, M. O. and Çoban, S., (2022). Blockchain-Based Supply Chain Management: Understanding the Determinants of Adoption in the Context of Organizations. *Information Systems Management*, 39(2), pp. 100–121. https://doi.org/10.1080/10580530.2020.1812014.

Guo, Y., Liang, C., (2016). Blockchain application and outlook in the banking industry. In *Financial Innovation*, Vol. 2, Issue 1. Springer Open. https://doi.org/10.1186/s40854-016-0034-9.

Hassan, H., (2017). Organisational factors affecting cloud computing adoption in small and medium enterprises (SMEs) in service sector. *Procedia Computer Science*, 121, 976–981. https://doi.org/10.1016/j.procs.2017.11.126.

Hassani, H., Huang, X. and Silva, E., (2018). Big-crypto: Big data, blockchain and cryptocurrency. *Big Data and Cognitive Computing*, 2(4), pp. 1–15. https://doi.org/10.3390/bdcc2040034.

Hastig, G. M., Sodhi, M. M. S., (2020). Blockchain for Supply Chain Traceability: Business Requirements and Critical Success Factors. *Production and Operations Management*, 29(4), pp. 935–954. https://doi.org/10.1111/poms.13147.

Hew, J. J., Wong, L. W., Tan, G. W. H., Ooi, K. B. and Lin, B., (2020). The blockchain-based Halal traceability systems: a hype or reality? *Supply Chain Management*, 25(6), pp. 863–879. https://doi.org/10.1108/SCM-01-2020-0044.

Kamble, S., Gunasekaran, A. and Arha, H., (2019). Understanding the Blockchain technology adoption in supply chains-Indian context. *International Journal of Production Research*, 57(7), pp. 2009–2033. https://doi.org/10.1080/00207543.2018.1518610.

Kamilaris, A., Fonts, A. and Prenafeta-Boldύ, F. X., (2019). The rise of blockchain technology in agriculture and food supply chains. In *Trends in Food Science and Technology*, Vol. 91, pp. 640–652. Elsevier Ltd. https://doi.org/10.1016/j.tifs.2019.07.034.

Kim, J. S., Shin, N., (2019). The impact of blockchain technology application on supply chain partnership and performance. *Sustainability (Switzerland)*, 11(21). https://doi.org/10.3390/su11216181.

Kumar, S., (2014). Indian Consumer Attitudes Toward Food Safety: An Exploratory Study. *Journal of Food Products Marketing*, 20(3), pp. 229–243. https://doi.org/10.1080/10454446.2013.855992.

Lengoatha, L., Seymour, L. F., (2020). Determinant factors of intention to adopt blockchain technology across academic libraries. *ACM International Conference Proceeding Series*, pp. 244–250. https://doi.org/10.1145/3410886.3410905.

Lin, I. C., Liao, T. C., (2017). A survey of blockchain security issues and challenges. *International Journal of Network Security*, 19(5), pp. 653–659. https://doi.org/10.6633/IJNS.201709.19(5).01.

Makhdoom, I., Abolhasan, M., Abbas, H. and Ni, W., (2019). Blockchain's adoption in IoT: The challenges, and a way forward. In *Journal of Network and Computer Applications*, Vol. 125, pp. 251–279. Academic Press. https://doi.org/10.1016/j.jnca.2018.10.019.

Mendling, J., Weber, I., van der Aalst, W., Brocke, J. vom, Cabanillas, C., Daniel, F., Debois, S., di Ciccio, C., Dumas, M., Dustdar, S., Gal, A., García-Bañuelos, L., Governatori, G., Hull, R., la Rosa, M., Leopold, H., Leymann, F., Recker, J., Reichert, M., and Zhu, L., (2018). Blockchains for business process management - Challenges and opportunities. *ACM Transactions on Management Information Systems*, 9(1). https://doi.org/10.1145/3183367.

Mezquita, Y., Casado, R., Gonzalez-Briones, A., Prieto, J. and Corchado, J. M., (2019). Blockchain technology in IoT systems: Review of the challenges. In *Annals of Emerging Technologies in Computing*, Vol. 3, Issue 5, Special Issue, pp. 17–24. International Association for Educators and Researchers (IAER). https://doi.org/10.33166/AETiC.2019.05.003.

Mukherjee, S., Baral, M. M. and Chittipaka, V., (2022). Studying the Adoption of Blockchain Technology in the Manufacturing Firms. In *Utilizing Blockchain Technologies in Manufacturing and Logistics Management*, pp. 64–80. IGI Global. https://doi.org/10.4018/978-1-7998-8697-6.ch004.

Mukherjee, S., Baral, M. M., Pal, S. K., Chittipaka, V., Roy, R. and Alam, K., (2022). Humanoid robot in healthcare: A Systematic Review and Future Research Directions. *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, pp. 822–826. https://doi.org/10.1109/COM-IT-CON54601.2022.9850577.

Mukherjee, S., Baral, M. M. and Venkataiah, C., (2022). Supply Chain Strategies for Achieving Resilience in the MSMEs. In *External Events and Crises That Impact Firms and Other Entities*. pp. 158–183. IGI Global. https://doi.org/10.4018/978-1-7998-8346-3.ch004

Mukherjee, S., Chittipaka, V., (2021). Analysing the Adoption of Intelligent Agent Technology in Food Supply Chain Management: An Empirical Evidence. *FIIB Business Review*, 231971452110592. https://doi.org/10.1177/23197145211059243

Mukherjee, S., Chittipaka, V. and Baral, M. M., (2021). Developing a Model to Highlight the Relation of Digital Trust With Privacy and Security for the Blockchain Technology. In *Blockchain Technology and Applications for Digital Marketing* (pp. 110–125). IGI Global. https://doi.org/10.4018/978-1-7998-8081-3.ch007

Mukherjee, S., Nagariya, R., Baral, M. M., Patel, B. S., Chittipaka, V., Rao, K. S. and Rao, U. V. A., (2022). Blockchain-based circular economy for achieving environmental sustainability in the Indian electronic MSMEs. *Management of Environmental Quality: An International Journal*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/MEQ-03-2022-0045.

Oliveira, T., Thomas, M. and Espadanal, M., (2014). Assessing the determinants of cloud computing adoption: An analysis of the manufacturing and services sectors. *Information and Management*, 51(5), pp. 497–510. https://doi.org/10.1016/j.im.2014.03.006.

Önder, I. and Treiblmaier, H., (2018). Blockchain and tourism: Three research propositions. *Annals of Tourism Research*, 72, pp. 180–182. https://doi.org/10.1016/j.annals.2018.03.005.

Pateli, A., Mylonas, N. and Spyrou, A., (2020). Organizational Adoption of Social Media in the Hospitality Industry: An Integrated Approach Based on DIT and TOE Frameworks. *Sustainability*, 12(17), 7132. https://doi.org/10.3390/su12177132.

Puklavec, B., Oliveira, T. and Popovič, A., (2018). Understanding the determinants of business intelligence system adoption stages an empirical study of SMEs. *Industrial Management and Data Systems*, 118(1), pp. 236–261. https://doi.org/10.1108/IMDS-05-2017-0170.

Queiroz, M. M., Telles, R. and Bonilla, S. H., (2020). Blockchain and supply chain management integration: a systematic review of the literature. In *Supply Chain Management*, Vol. 25, Issue 2, pp. 241–254. Emerald Group Holdings Ltd. https://doi.org/10.1108/SCM-03-2018-0143.

Reimers, T., Leber, F. and Lechner, U., (2019). Integration of blockchain and internet of things in a car supply chain. *Proceedings - 2019 IEEE International Conference on Decentralized Applications and Infrastructures, DAPPCON 2019*, pp. 146–151. https://doi.org/10.1109/DAPPCON.2019.00028.

Roy, R., Babakerkhell, M. D., Mukherjee, S., Pal, D. and Funilkul, S., (2022). Evaluating the Intention for the Adoption of Artificial Intelligence-Based Robots in the University to Educate the Students. *IEEE Access*, 10, pp. 125666–125678. https://doi.org/10.1109/ACCESS.2022.3225555.

Roy, R., Baral, M. M., Pal, S. K., Kumar, S., Mukherjee, S. and Jana, B., (2022). Discussing the present, past, and future of Machine learning techniques in livestock farming: A systematic literature review. *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, pp. 179–183. https://doi.org/10.1109/COM-IT-CON54601.2022.9850749.

Roy, R., Chekuri, K., Sandhya, G., Pal, S. K., Mukherjee, S. and Marada, N., (2022). Exploring the blockchain for sustainable food supply chain. *Journal of Information and Optimization Sciences*, 43(7), pp. 1835–1847. https://doi.org/10.1080/02522667.2022.2128535.

Saberi, S., Kouhizadeh, M. and Sarkis, J., (2019). Blockchains and the Supply Chain: Findings from a Broad Study of Practitioners. *IEEE Engineering Management Review*, 47(3), pp. 95–103. https://doi.org/10.1109/EMR.2019.2928264.

Salam, A., Panahifar, F. and Byrne, P. J., (2016). Retail supply chain service levels: the role of inventory storage. *Journal of Enterprise Information Management*, 29(6), pp. 887–902. https://doi.org/10.1108/JEIM-01-2015-0008.

Seebacher, S. and Schüritz, R., (2017). Blockchain technology as an enabler of service systems: A structured literature review. *Lecture Notes in Business Information Processing*, 279, pp. 12–23. https://doi.org/10.1007/978-3-319-56925-3_2.

Subramanian, N., Chaudhuri, A. and Kayıkcı, Y., (2020). Blockchain and Supply Chain Logistics. In *Blockchain and Supply Chain Logistics*. Springer International Publishing. https://doi.org/10.1007/978-3-030-47531-4.

Tashkandi, A. N. and Al-Jabri, I. M., (2015). Cloud computing adoption by higher education institutions in Saudi Arabia: An exploratory study. *Cluster Computing*, 18(4), pp. 1527–1537. https://doi.org/10.1007/s10586-015-0490-4.

Teller, C., Holweg, C., Reiner, G. and Kotzab, H., (2018). Retail store operations and food waste. *Journal of Cleaner Production*, 185, pp. 981–997. https://doi.org/10.1016/j.jclepro.2018.02.280.

Tijan, E., Aksentijević, S., Ivanić, K. and Jardas, M., (2019). Blockchain technology implementation in logistics. In *Sustainability (Switzerland)*, Vol. 11, Issue 4. MDPI. https://doi.org/10.3390/su11041185.

van Hoek, R., (2019). Exploring blockchain implementation in the supply chain: Learning from pioneers and RFID research. *International Journal of Operations and Production Management*, 39, pp. 829–859. https://doi.org/10.1108/IJOPM-01-2019-0022.

Wang, H., Chen, K. and Xu, D., (2016). A maturity model for blockchain adoption. *Financial Innovation*, 2(1). https://doi.org/10.1186/s40854-016-0031-z.

Wang, Y. M., Wang, Y. S. and Yang, Y. F., (2010). Understanding the determinants of RFID adoption in the manufacturing industry. *Technological Forecasting and Social Change*, 77(5), pp. 803–815. https://doi.org/10.1016/j.techfore.2010.03.006.

Wong, L. W., Leong, L. Y., Hew, J. J., Tan, G. W. H. and Ooi, K. B., (2020). Time to seize the digital evolution: Adoption of blockchain in operations and supply chain management among Malaysian SMEs. *International Journal of Information Management*, 52. https://doi.org/10.1016/j.ijinfomgt.2019.08.005.

Xu, W., Ou, P. and Fan, W., (2017). Antecedents of ERP assimilation and its impact on ERP value: A TOE-based model and empirical test. *Information Systems Frontiers*, 19(1), pp. 13–30. https://doi.org/10.1007/s10796-015-9583-0.

Yusof, H., Farhana Mior Badrul Munir, M., Zolkaply, Z., Li Jing, C., Yu Hao, C., Swee Ying, D., Seang Zheng, L., Yuh Seng, L. and Kok Leong, T., (2018). Behavioral Intention to Adopt Blockchain Technology: Viewpoint of the Banking Institutions in Malaysia. *International Journal of Advanced Scientific Research and Management*, 3(10), pp. 1–6. www.ijasrm.com.

Zhang, X., Sun, P., Xu, J., Wang, X., Yu, J., Zhao, Z. and Dong, Y., (2020). Blockchain-based safety management system for the grain supply chain. *IEEE Access*, 8, pp. 36398–36410. https://doi.org/10.1109/ACCESS.2020.2975415.

Zhu, K., Kraemer, K. L. and Xu, S., (2006). The process of innovation assimilation by firms in different countries: A technology diffusion perspective on e-business. *Management Science*, 52(10), pp. 1557–1576. https://doi.org/10.1287/mnsc.1050.0487.

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 147–166, https://doi.org/10.59170/stattrans-2023-068
Received – 04.09.2022; accepted – 22.09.2023

Marketing segmentation of banks' corporate clients based on data mining technique

Rostyslav Klochko[1], **Olena Piskunova**[2]

Abstract

In recent years, the banking services market has been developing dynamically, experiencing a sharp increase in competition. Banks that provide maximum profitability for each client gain the most significant competitive advantage. The sales model in corporate banking is associated with personal interaction between bank employees and business owners, and the subsequent establishment of individual service conditions. However, this approach is often ineffective when a bank faces the issue of maximising the efficiency of business activities. This study aims to segment a bank's corporate client base and develop a pricing strategy for each of the groups that have been singled out in the process. The study sample consisted of 4,500 corporate clients of a Ukrainian bank who were active users of euro accounts. The k-means data mining algorithm was used to develop marketing segments. The optimal number of clusters was determined by weighing the results of calculating 26 indices from the NbClust package and the bank's business requirements. Six similarity groups were found during the calculation of the algorithm. The study found that clusters 1 and 2 were a concentration of unprofitable customers for whom an introduction of a service fee was urgently needed. Marketing segments 3 and 4 were customers who did not record net losses but with whom it was deemed necessary to work to improve their profitability. The remaining segments were 'healthy' users of euro accounts. With regard to these customers, it was recommended no additional service fees should be imposed. The proposed methodology makes it possible for a bank to remain attractive in a competitive environment while not incurring unnecessary costs.

Key words: clusterisation, k-means, pricing strategy, customer value, cost optimisation

1. Introduction

In recent years, the banking services market has been developing exceptionally dynamically, experiencing a sharp increase in competition due to the deregulation of

[1] Department of Mathematical Modeling and Statistics, Kyiv National Economic University named after Vadym Hetman, Ukraine. E-mail: rostislav.klochko@gmail.com. ORCID: https://orcid.org/0000-0003-2690-2785.
[2] Department of Mathematical Modeling and Statistics, Kyiv National Economic University named after Vadym Hetman, Ukraine. E-mail: EPiskunova@kneu.edu.ua. ORCID: https://orcid.org/0000-0002-0373-5358.

banking activities, the introduction of new technologies, and increased consumer demand. Banks are increasingly turning to the practice of cost optimization in order to achieve maximum business efficiency. Banks that provide maximum profitability for each client gain the most significant competitive advantage. Accordingly, the concept of customer value and its maximization acquires vital importance.

For most universal banks that provide a wide range of services for the retail and corporate segments, the share of profit received from the corporate is system-forming. Corporate banking refers to the part of banking that deals with corporate clients. It typically serves a wide range of clients, from small and medium-sized businesses with a few million in revenue to large corporations with billions in sales. Banks need more timely and adequate information to serve their corporate clients better. In contrast to the retail business model, the traditional sales model in corporate banking is associated with personal interaction between bank employees and business owners, with the subsequent establishment of individual service conditions. This approach is often ineffective when the bank faces the issue of maximizing the efficiency of business activities, especially when they have thousands of corporate clients in their portfolios. It is not easy to provide an individual approach for every client, not lose the effectiveness of clients' value management, and not spend many human resources simultaneously. On the other hand, the competitive environment does not allow to provide the same pricing policy for all corporate clients.

One of the main options for solving this problem is classifying corporate clients into appropriate similarity groups - marketing segments and developing a price strategy separately for each. The term "marketing segmentation of corporate clients" was first stated in this study. However, the proposed approaches to customer differentiation and further development of individual marketing proposals for each of the groups allow us to consider the usage of this concept appropriate.

2. Literature Review

The application of machine learning methods and data mining for customer segmentation has become one of the main topics in the literature that examines the banking market (He et al., 2022; Aghaei, 2021; Chawla & Joshi, 2021). This is primarily because banks, by their very nature, accumulate vast volumes of customer data, which is crucial in building an effective customer clustering algorithm.

Studies concerning the bank's corporate clients mainly aim to improve the operational efficiency of the business. Usually, it is about determining the riskiest customers unable to repay debts (Oleynik & Formánek 2020) or making banking operations that violate domestic legislation (Hamal & Senvar 2021). In these studies, the authors have a list of confirmed risky customers, based on the example of which

they try to classify all other unassessed clients. The basis of such studies is machine learning classification methods, such as support vector machine, naive Bayes, artificial neural network, k-nearest neighbour, random forest, and logistic regression.

However, if we analyse the research on all retail and corporate banking customer segments, customer segmentation is usually focused on achieving marketing goals (Rajaobelina et al., 2019; Firdaus & Utama, 2021; Fathima & Muthumani, 2019), where retail research has both a quantitative and a qualitative advantage. Two characteristics of the retail segment can explain this: this perimeter includes the largest number of customers and generates enormous volumes of data; the peculiarity of retail customers allows applying the same marketing policy to many customers, which is not usual if we are talking about corporate clients.

The range of problems covered in the literature on this topic is wide. Yanık & Elmorsy (2019) developed a clustering model based on the data of 40,000 credit card users, which made it possible to implement eight different marketing strategies according to the specifics of the use of banking products and socio-demographic characteristics of customers. The study is based on a self-organizing map and k-means clustering models. A similar study was conducted by Hung et al. (2019). The only difference is that they used the hierarchical agglomerative clustering algorithm to perform the analysis. In general, neural-network-based clustering approaches are gaining popularity among banking researchers. These algorithms are usually used in combination with hierarchical clustering (Kovacs et al., 2021) or minimum spanning tree (Barman & Chowdhury, 2019) to improve segmentation efficiency.

It is necessary to highlight studies on customer segmentation in the lending process. Vijayalakshmi et al. (2020) created a segmentation model based on random forest, logistic regression, and support vector machine learning algorithms to estimate the default of a future loan more accurately. At the same time, Umuhoza et al. (2020) used the k-means algorithm to better manage the portfolio of already issued loans.

Calvo-Porral & Lévy-Mangin (2020) proposed the algorithm for developing different communication policies based on the client's psycho type. They divide clients into groups using confirmatory factor analysis (CFA) and Anova test algorithms.

The presence of a large number of customer activity characteristics allowed Djurisic et al. (2020) to develop a segmentation model based on the combination of the customer significance assessment approach (Recency, Frequency, and Monetary) with clustering (k-means) and classification (Support Vector Machine) models. This algorithm helped to focus the bank's efforts on retaining only those customers who bring real value to the company. A similar study was proposed by Dang Tran et al. (2023) to address the problem of predicting bank customer churn.

In turn, due to mainly personal service and a small amount of structured data, there is no variety of research on the specifics of working with corporate clients, especially with the use of intelligent modelling technologies.

Formisano et al. (2020) made the first attempts at marketing research of the bank's corporate clients. With the help of the Kano model, the researchers tried to estimate the non-linear relationship between the level of customer satisfaction and the quality of banking services. The analysis results can be applied to identify groups of dissatisfied customers, the so-called irritants, and improve their customer experience.

Tungjitnob et al. (2021) focused on identifying corporate users (business owners) among mobile banking customers. The developed algorithms will allow companies to implement special offers specifically for the corporate segment of online banking users. The primary means of modelling were the Extreme Gradient Boosting algorithm and Convolutional Neural Network machine learning methods.

Osowski & Sierenski (2020) highlighted a vital research topic on corporate clients' activity. Based on the Neural Networks technology, they developed an algorithm that predicts the activity status of a corporate client. The proposed model will allow banks to identify groups of customers at risk of leaving the bank and influence them with marketing methods.

Studies of the bank's corporate clients from the marketing point of view have hardly found their development in the literature. Furthermore, the ones that describe the peculiarities of working with corporate customers' data mainly focus on efforts to predict future processes by applying complex machine learning or neural network models. Approaches to analysing the current behaviour characteristics of corporate customers are almost not revealed in the available research.

The establishment of approaches to individual pricing in the corporate perimeter reduced the number of publications on developing marketing proposals to zero. Existing studies often rely on data unique to a specific case study, which is tricky to apply to other banks. The described problems determined the relevance of this study - the development of marketing segmentation of bank's corporate clients with the subsequent implementation of the pricing policy for each similarity group. The value of our research is the transfer of retail customer segmentation approaches to the corporate perimeter. Most of these algorithms have been known for a long time, do not require unique data, and are easily calculated and scaled to other companies. Due to its simplicity in implementation, most studies used the k-means technique, which led to the choice of this clustering algorithm in our study. At the same time, the small number of customers for analysis and the limited available characteristics of their activity significantly complicate the segmentation process using the k-means algorithm. Nevertheless, the results of our research prove that the available volumes of data are pretty enough to talk about its effectiveness for solving the tasks.

3. Aim of The Study

This study aims to develop a marketing segmentation model for corporate clients of a universal bank that will help optimize profitability and efficiency in today's competitive banking landscape.

The following tasks were defined to achieve the goal:

1. Describe and systematize the data clustering methodology and algorithm hyperparameters' setting.
2. Following the developed methodology, implement the clustering algorithm of the bank's corporate clients.
3. Analyse the results and propose pricing strategies for each identified similarity group.

This segmentation approach, connected with recommendations, is expected to optimize operating costs without introducing a uniform fee structure for all bank corporate clients. As a result, banks can remain attractive in a competitive environment while avoiding unnecessary expenses. Furthermore, the practical research findings can serve as a basis for the development of automated marketing campaigns and personalized communication strategies tailored to each marketing segment of corporate clients, enhancing overall customer satisfaction.

4. Materials and Methods

4.1. Methodology of banks' corporate client segmentation.

Data mining is the process of analysing large data sets to identify patterns and relationships that can help solve business problems. Data mining technologies make it possible to synthesize valuable information that is implicitly contained in the data. The large volumes of accumulated data in companies make data mining approaches helpful in finding insights that managers can use in decision-making. Data mining methods and tools allow enterprises to predict future trends and make more informed business decisions.

Clustering is one of the most common data mining approaches to gaining an intuitive understanding of data structure. Cluster analysis is a method of analysing data to find and identify similar data. This process helps to understand the differences and similarities between the data.

In our study, we will use only the k-means clustering method, which, due to its simplicity, is considered one of the most used clustering algorithms. Moreover, we used RStudio software and the R programming language for all calculations.

k-means clustering aims to divide objects into k clusters so that data points in one cluster are similar and data points in different clusters are as distant as possible. The distance between them determines the similarity of the two points.

The stages of implementation of the algorithm are:

1) determination of the required number of clusters - k;
2) determination of the initial value of the centroid;
3) calculation of the distance from existing objects to the centroid;
4) assignment of each object to the corresponding cluster depending on the calculated distances.

The first hyperparameter of the k-means clustering model is the number of similar groups that must be obtained as a result of the algorithm implementation. There are many criteria by which the statistically necessary number of groups can be pre-estimated. Our study used the NbClust R library (Dimitriadou 2002), which allows us to estimate the required number of clusters according to 26 criteria (Table 1). The number for which most criteria "vote" should be considered optimal. However, the final number of clusters is always chosen considering the organization's business needs.

The second hyperparameter of the algorithm is the number of times the algorithm is calculated, where a different initial centroid is randomly assigned each time. The number depends on the computing capabilities of the system. In our study, we used 25 iterations of the algorithm.

Table 1: Characteristics of indexes in the NbClust R library

Full name	Short name	Selection criterion
Hubert index. Hubert and Arabie 1985	Hubert	Graphical method
Dindex. Lebart et al. (2000)	Dindex	Graphical method
KL index. Krzanowski and Lai (1988)	KL	The maximum value of the index
CH index. Calinski and Harabasz (1974)	CH	The maximum value of the index
Hartigan index. Hartigan (1975)	Hartigan	The maximum difference between hierarchy levels of the index
Cubic Clustering Criterion (CCC). Sarle (1983)	CCC	The maximum value of the index
Scott index. Scott and Symons (1971)	Scott	The maximum difference between hierarchy levels of the index
Marriot index. Marriot (1971)	Marriot	Max. value of second differences between levels of the index
TraceCovW index. Milligan and Cooper (1985)	TrCovW	The maximum difference between hierarchy levels of the index
TraceW index. Milligan and Cooper (1985)	TraceW	The maximum value of absolute second differences between levels of the index
Friedman index. Friedman and Rubin (1967)	Friedman	The maximum difference between hierarchy levels of the index

Table 1: Characteristics of indexes in the NbClust R library (cont.)

Full name	Short name	Selection criterion
Silhouette index. Kaufman and Rousseeuw (1990)	Silhouette	The maximum value of the index
Ratkowsky index. Ratkowsky and Lance (1978)	Ratkowsky	The maximum value of the index
Ball index. Ball and Hall (1965)	Ball	The maximum difference between hierarchy levels of the index
PtBiserial index. Examined by Milligan (1980,1981)	Ptbiserial	The maximum value of the index
Dunn index. Dunn (1974)	Dunn	The maximum value of the index
Rubin index. Friedman and Rubin (1967)	Rubin	The minimum value of second differences between levels of the index
C-index. Hubert and Levin (1976)	Cindex	The minimum value of the index
DB index. Davies and Bouldin (1979	DB	The minimum value of the index
Duda index. Duda and Hart (1973)	Duda	The smallest number of clusters such that index > criticalValue
Pseudot2 index. Duda and Hart (1973)	Pseudot2	The smallest number of clusters such that index < criticalValue
Beale index. Beale (1969)	Beale	The number of clusters such that the critical value of the index >= alpha
Frey index. Frey and Van Groenewoud (1972)	Frey	The cluster level before that index value < 1.00
Mcclain index. McClain and Rao (1975)	McClain	The minimum value of the index
SDindex. Halkidi et al. (2000)	SDindex	The minimum value of the index
SDbw. Halkidi et al. (2001)	SDbw	The minimum value of the index

The criterion for choosing the optimal initial centroid is the parameter Within Cluster Sum of Squares (W_total) (1). The algorithm assigns data points to a cluster so that the sum of the squares of the distances between the data points and the cluster's centroid is minimal. The less variation we have within the clusters, the more similar the data points in the same cluster.

$$W_{total} = \sum_{l=1}^{k} \frac{\rho(C^l)}{n^l}, \tag{1}$$

where $\rho(C^l)$ - the sum of Euclidean distances between points within the cluster l ; n^l – the number of points in cluster l; k– the number of clusters.

The final step of our research is the description of the constructed clusters. Each segment differs in its average indicators of activity, which must be calculated. Calculating the average values of the model's parameters makes it possible to provide recommendations regarding the need to carry out specific work with a particular group of clients to reduce bank operational costs.

4.2. Initial data set features

The basis of the study is data on the activity of euro current account users for the period from April 2020 to April 2021. The sample includes 4,500 corporate clients of the Ukrainian bank (Private entrepreneurs and Legal entities). The complete list of initial data fields is in Table 2.

In Ukraine, the economic crises of 2008-2010 and 2014-2015 resulted in stopping bank lending in foreign currency. As a result, euro deposits have become a loss-making type of service for banks (service cost is 1% per year). Accordingly, if a client has only one product - a euro account, and does not perform any banking operations for which the bank would receive commission income - such a client is entirely unprofitable. Given this, there is a need to minimize operating costs by the following:

- decrease in clients' euro account balances;
- increase in the number of operations with euro accounts;
- introduction of a new service fee for euro accounts.

Table 2: Characteristics of the initial data sample fields

Model Variables	Description
CL_ID	Client internal ID
NOM_nb	The number of months when the client's euro account had positive account balances
Oper_nb	The number of months when the client performed transactions with euros
NIM_nb	The number of months when the net interest income for this client was negative (i.e. net marginal loss)
NBI_nb	The number of months when the net banking income for this client was negative (i.e. net banking loss). Net banking income is the sum of net interest income and commission income.

5. Results

As mentioned above, the first step in implementing the bank's corporate clients clustering is determining the optimal number of similarity groups. Table 3 presents the results of calculating 26 evaluation criteria in the NbClust package (the optimal values for each criterion are highlighted). Table 4 shows the "voting" results for the optimal number of clusters.

Table 3: Values of calculated indices in the NbClust package

Index	Number of clusters								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
KL	1.293	0.318	1.624	**28.729**	0.030	20.489	3.749	0.189	1.322
CH	494	620	979	1238	1133	**1682**	1601	1510	1482
Hartigan	586	1006	765	190	1066	**169**	121	163	138
CCC	-19	-14	12	32	31	**71**	70	69	70
Scott	3272	4302	6233	**9692**	10997	13060	13561	14099	14956
Marriot	3.52E+12	4.47E+12	2.72E+12	**6.22E+11**	4.33E+11	1.88E+11	1.85E+11	1.74E+11	1.33E+11
TrCovW	1892585	1222424	**525480**	282585	256320	69466	62513	57201	49881
TraceW	5644	4257	**2730**	1914	1731	1085	992	929	852
Friedman	5	5	8	**17**	24	28	29	32	38
Rubin	1.275	1.690	2.636	3.759	4.158	**6.629**	7.256	7.745	8.449
Cindex	0.179	0.134	0.146	0.164	0.235	0.180	0.167	0.167	**0.134**
DB	1.721	1.471	1.050	0.901	0.911	**0.790**	0.904	0.990	1.055
Silhouette	0.328	0.382	0.375	0.435	0.436	**0.513**	0.511	0.511	0.446
Duda	**0.814**	4.857	2.629	0.428	5.324	1.175	1.486	2.098	1.780
Pseudot2	**293**	-1058	-442	76	-604	-90	-128	-274	-27
Beale	**0.551**	-1.915	-1.494	3.197	-1.905	-0.360	-0.788	-1.225	-1.053
Ratkowsky	0.213	0.347	**0.392**	0.381	0.353	0.348	0.328	0.311	0.297
Ball	2822	**1419**	682	383	288	155	124	103	85
Ptbiserial	0.266	0.413	0.457	0.509	0.513	**0.537**	0.528	0.527	0.476
Frey	-0.024	-0.198	-0.073	0.034	0.175	0.567	0.367	1.355	2.317
McClain	**0.570**	0.948	0.922	0.891	0.889	0.893	0.931	0.935	1.163
Dunn	0.019	0.018	0.022	0.030	0.044	0.044	0.044	**0.045**	0.040
Hubert	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SDindex	3.393	3.277	4.310	3.234	2.807	**2.501**	2.539	2.713	3.748
Dindex	1.398	1.152	1.005	0.871	0.847	0.655	0.634	0.619	0.560
SDbw	1.140	1.028	1.247	0.897	0.579	0.466	0.437	0.415	**0.400**

Table 4: "Voting" results for the optimal number of clusters

Number of clusters	**2**	**3**	**4**	**5**	**7**	**9**	**10**
The number of indices that voted for the optimum	4	1	3	4	8	1	2

Most of the indices voted for 7 clusters. Next is calculating the average model parameter values for the optimal number of clusters (Table 5).

Table 5: Average characteristics of clusters when clients are divided into seven marketing segments

Cluster	Number of customers	NOM_nb	Oper_nb	NIM_nb	NBI_nb
1	663	12	11	0	0
2	788	2	11	0	0
3	1073	11	1	0	0
4	**92**	**10**	**3**	**8**	**5**
5	**53**	**12**	**0**	**11**	**11**
6	1578	1	2	0	0
7	253	11	4	8	0

The determination of the initial centroid is performed randomly among all available objects. The number of repetitions of the algorithm calculation is pre-set to optimize the quality of the clustering algorithm, considering that a different centroid will be assigned each time. In our study, we used a 25-fold repetition of the calculation, which is sufficient to avoid the problem of an incorrectly chosen centroid.

The optimality of the initial centroid selection and the implementation of the entire algorithm depends on one more hyperparameter - the choice of the method of calculating the distance from existing objects to the centroid. Our study used the best-known calculation algorithm - the Euclidean distance between points. Accordingly, the optimal algorithm, and the optimal initial centroid, is the algorithm iteration in which the index of the total distance from the existing objects to the group's centroid is the smallest (equation 1).

Clusters 4 and 5 characterize the same group of clients. The only difference is the intensity of euro account usage. Because of this, we decided to implement an algorithm with only six groups. The characteristics of the distribution into six segments are presented in Table 6.

Table 6: Average characteristics of clusters when clients are divided into six marketing segments

Cluster	Number of customers	NOM_nb	Oper_nb	NIM_nb	NBI_nb
1	284	11	4	8	1
2	110	11	2	10	8
3	1 071	11	1	0	0
4	665	12	11	0	0
5	1 581	1	2	0	0
6	789	2	11	0	0

To facilitate the characterization of each defined marketing segment and to provide personalized recommendations for reducing the bank's operating costs, data on clients' activity in each segment were displayed as a boxplot (Fig. 1-6).

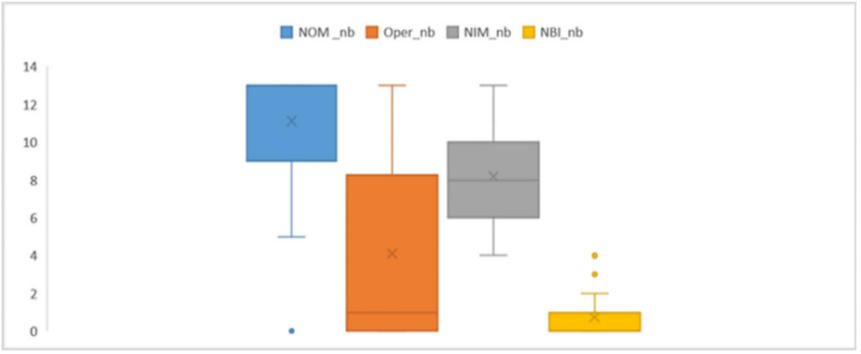

Figure 1: Activity in the first segment of corporate clients

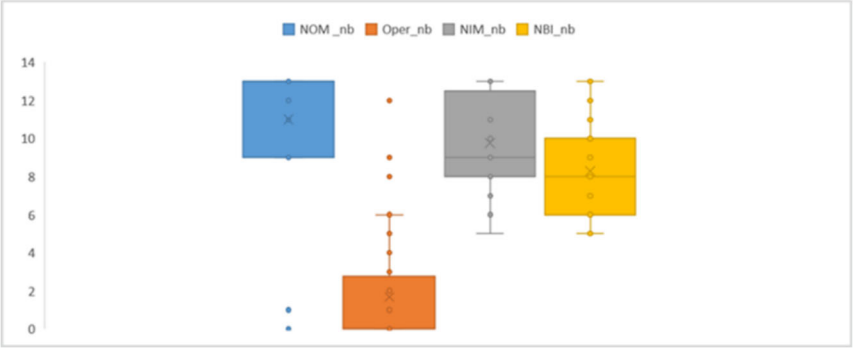

Figure 2: Activity in the second segment of corporate clients

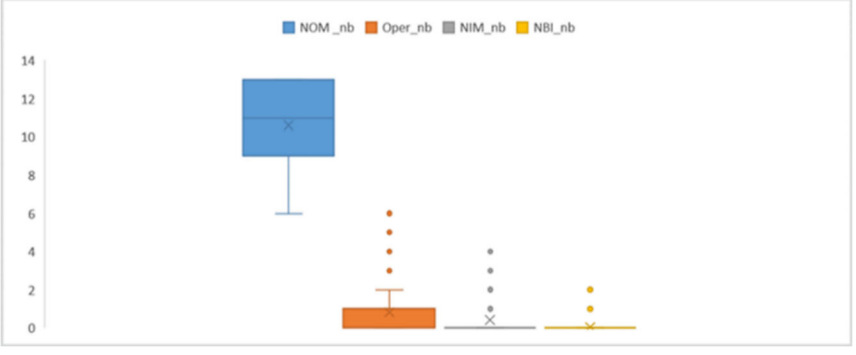

Figure 3: Activity in the third segment of corporate clients

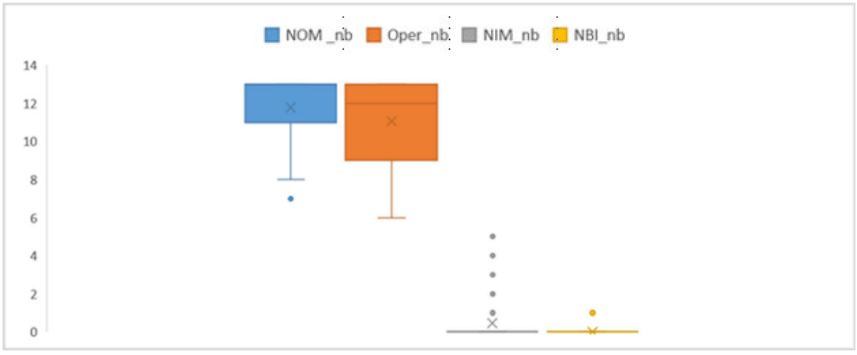

Figure 4: Activity in the fourth segment of corporate clients

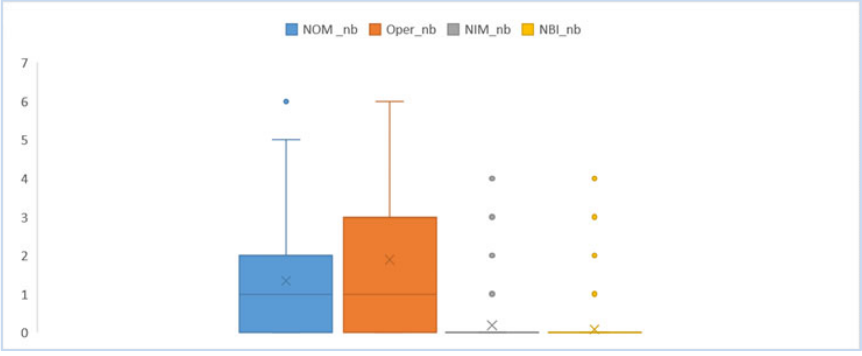

Figure 5: Activity in the fifth segment of corporate clients

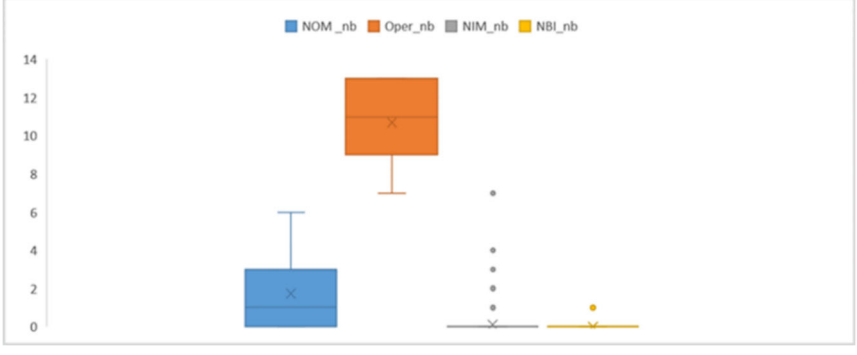

Figure 6: Activity in the sixth segment of corporate clients

6. Discussion

This study reveals the entire methodology of clustering the customer base, considering their behaviour's peculiarities. The object of the study was corporate clients who are active euro account users in one of the Ukrainian banks. Since, in Ukraine, there is no possibility to earn on foreign currency (due to insufficient lending), maintenance of euro accounts is a net operating loss for the bank.

After the customer activity analysis, each marketing segment can be characterized as follows (Table 6):

1) unprofitable transactors with euro accumulation;
2) unprofitable euro accumulation;
3) profitable euro accumulation;
4) profitable transactors with euro accumulation;
5) without euro accumulation;
6) profitable transactors without euro accumulation.

The last task is to develop a pricing strategy for the euro account service. Six graphs showing each parameter's average in a boxplot were created to understand better the customers' activity in each cluster (Figures 1-6).

Figure 1 displays the clients' activity in cluster 1:

- constant positive balance on euro accounts;
- euro transactions are performed only 4-6 times a year;
- these clients show a net marginal loss most of the time due to insufficient outstanding in the national currency or loans;
- euro transactions bring additional commission income, mainly covering net marginal loss.

Given the above characteristic and taking into account the small number of customers - 284 (Table 6), the following recommendations were developed to improve the profitability of the marketing segment:

- to communicate with clients and recommend purchasing euros at the time of need only;
- introduce a 1% service fee on euro accounts.

Figure 2 displays the clients' activity in cluster 2:

- constant positive balance on euro accounts;
- euro transactions are performed only a few times a year;
- these customers always show a net banking loss.

Given the above characteristic and taking into account the small number of customers - 110 (Table 6), the following recommendations were developed to improve the profitability of the marketing segment:

- introduce a 1% service fee on euro accounts;
- it is recommended to survey to determine the reason for the constant euro accumulation without further use.

Figure 3 displays the clients' activity in cluster 3:

- constant positive balance on euro accounts;
- euro transactions are performed only a few times a year;
- these clients, thanks to other banking products, are currently profitable for the bank.

Because of the above characteristic and taking into account a large number of customers - 1,071 (Table 6), the following recommendations were developed to improve the profitability of the marketing segment:

- it is recommended to survey to determine the reason for the constant euro accumulation without further use;
- to communicate with clients about introducing a 1% service fee on euro accounts in the second stage of the reduction operating costs exercise (to give time to prepare for this).

Figure 4 displays the clients' activity in cluster 4:

- constant positive balance on euro accounts;
- accumulation is constantly used for euros transactions;
- current activity allows clients to remain fully profitable for the bank.

Because of the above characteristic and taking into account a large number of customers - 665 (Table 6), the following recommendations were developed to improve the profitability of the marketing segment:

- to communicate with clients and recommend purchasing euros at the time of need only;
- to communicate with clients about introducing a 1% service fee on euro accounts in the second stage of the reduction operating costs exercise (to give time to prepare for this).

Figure 5 and Figure 6 display the clients' activity in clusters 5 and 6:

- several times a year show a positive balance on euro accounts;
- customers of the sixth segment usually buy currency at the time of the transaction, which allows them not to accumulate euro on current accounts;
- accumulation is constantly used for euros transactions (difference between clusters is in the number of transactions);
- current activity allows clients to remain fully profitable for the bank.

Given the above characteristic and taking into account the most significant number of customers - 1,581 and 789 (Table 6), the following recommendations were developed to improve the profitability of the marketing segment:

- do not introduce any commissions for maintenance of euro accounts.

As we can see, clusters 1 and 2 are a concentration of unprofitable customers for whom an introduction of a service fee is urgently needed. Marketing segments 3 and 4 are customers who do not show net losses but with whom it is necessary to work to improve their profitability. Furthermore, to prevent possible losses, introduce service fees for these customers in the future. The last segments are examples of the healthy use of euro accounts. Customers do not accumulate euro but buy it at the moment of need without bringing losses to the bank. These customers are recommended not to set additional service fees, which saves human resources since these customers make up more than half of all users of euro accounts.

The effectiveness of the developed algorithm can be evidenced by the case of description and characteristics of similarity groups, as each segment has clearly defined differences that allow for the development of individual marketing recommendations for pricing. This segmentation and recommendations will optimize operating costs without setting a fee for all bank corporate clients. In the first wave, only 9% of clients were processed (an individual service fee was introduced), which also affected the bank's human resources optimization. For the first time, segmentation marketing approaches inherent in retail were applied for a personalized approach to corporate customers. This methodology allows the bank to remain attractive in a competitive environment while not incurring unnecessary costs.

Research that concerns the bank's corporate clients mainly examines improving the operational efficiency of doing business. In contrast to works (Oleynik & Formánek 2020) and (Hamal & Senvar 2021), this study is one of the first attempts to consider the bank's corporate clients from the point of view of the marketing component, especially with intelligent modelling technologies. The approach's basis was practiced primarily developed for retail clients, which we projected for the bank's corporate clients.

k-means is the basis of most of these studies, but compared to the algorithms Yanık & Elmorsy (2019) and Djurisic et al. (2020), we have improved the process of selecting hyperparameters of the model. A 25-fold selection of the initial centroid was applied according to the Within Cluster Sum of Squares smallest value (1). Also, based on the calculation of 26 indices, the optimal number of clusters was selected (Table 3).

Also, in our study, compared to research (Hung et al. 2019) and (Calvo-Porral & Lévy-Mangin 2020), more attention is paid to characterizing marketing segments and the development of personalized marketing strategies.

If we discuss the research of corporate clients from the point of view of the marketing component, the advantage of our methodology over (Formisano et al. 2020)

is that we used data that are publicly available for each bank, and the modelling algorithm can be easily tested on the client base of other banks.

The current study is focused on the characteristics of the current activity of banks' corporate clients and the development of personalized recommendations that can improve business efficiency in the nearest time. In contrast, most studies that describe the features of working with corporate clients' behaviour data (Osowski & Sierenski (2020)) are mainly focused on efforts to predict future processes by applying complex models of machine learning or neural networks.

Attempts to develop personalized marketing activities are found in (Tungjitnob et al. 2021), but for the first time, we proposed an approach specifically to personalized pricing based on the bank's corporate clients' marketing segments.

Implementation of the developed methodology will be valuable only to those banks ready to partially move from personal services to retail approaches with their segmentation practices. Very often, key stakeholders are not ready to trust intelligent modelling technologies, so they make decisions based on their intuition and understanding of the business. Also, cooperation with corporate clients is often associated with previous long-term agreements, which makes it impossible to change specific price offers in the short term. The results of our research are not relevant for countries whose banks do not incur operational losses from servicing accounts in foreign currency. Also, for the algorithm's efficiency, the bank must have a large number of corporate clients as the algorithm is sensitive to the number of observations it must divide into similarity groups.

The main drawback of the study is the use of a small number of customer activity characteristics. Accordingly, further development may include additional criteria for evaluating customer activity to improve the quality of clustering of the customer base.

7. Conclusions

For most universal banks that provide a wide range of services for the retail and corporate segments, the share of profit received from the corporate is system-forming. At the same time, the traditional sales model in corporate banking is associated with personal interaction between bank employees and business owners. It is often ineffective when the bank faces the issue of maximizing the efficiency of business activities, especially when they have thousands of corporate clients in their portfolios. It is challenging to provide an individual approach for every client, not lose the effectiveness of clients' value management, and not spend many human resources simultaneously. Accordingly, marketing segmentation and subsequent personalized pricing are the main options for solving this problem in the current highly competitive banking environment.

The basis of the study was 4,500 corporate clients of the Ukrainian bank who are active users of euro accounts. The k-means data mining algorithm was used to develop marketing segments. Two hyperparameters were previously set before the implementation: the number of iterations of the algorithm to determine the initial centroid and the number of clusters to be selected. The optimal initial centroid was selected among 25 algorithm iterations based on the minimum value of the Within Cluster Sum of Squares. The optimal number of clusters was determined by weighing the results of calculating 26 indices from the NbClust package and the bank's business requirements. Most indices for evaluating the optimal number of clusters voted for seven similarity groups. Nevertheless, after conducting a preliminary analysis of the average activity of each segment, it was determined that clusters 4 and 5 are representatives of the same group. Accordingly, a decision was made to merge these segments. Six similarity groups were found during the calculation of the algorithm:

1) unprofitable transactors with euro accumulation;
2) unprofitable euro accumulation;
3) profitable euro accumulation;
4) profitable transactors with euro accumulation;
5) without euro accumulation;
6) profitable transactors without euro accumulation.

During the analysis of the average activity indicators of each cluster, it was found that clusters 1 and 2 are a concentration of unprofitable customers for whom an introduction of a service fee is urgently needed. Marketing segments 3 and 4 are customers who do not show net losses but with whom it is necessary to work to improve their profitability. Furthermore, to prevent possible losses, introduce service fees for these customers in the future. The last segments are examples of the healthy use of euro accounts. Customers do not accumulate euro but buy it at the moment of need without bringing losses to the bank. These customers are recommended not to set additional service fees, which saves human resources since these customers make up more than half of all users of euro accounts.

This segmentation and recommendations will optimize operating costs without setting a fee for all bank corporate clients. In the first wave, only 9% of clients were processed (an individual service fee was established), which also affected the optimization of the bank's human resources. Accordingly, for the first time, segmentation marketing approaches inherent in retail were applied for a personalized approach to corporate customers. This methodology allows the bank to remain attractive in a competitive environment while not incurring unnecessary costs.

Practical research results can also be the basis for building automatic marketing campaigns and communications with individual offers for each marketing segment separately. Further research could test an expanded number of customer activity

characteristics to improve the quality of the segmentation model. Also, based on the results of the development of marketing segmentation of the bank's corporate clients, it is possible to develop a model for forecasting client needs for a specific product or service.

References

Aghaei, M., (2021). Market segmentation in the banking industry based on customers' expected benefits: A study of Shahr Bank. *Interdisciplinary Journal of Management Studies*, 14(3), pp. 629–648.

Aysha Fathima, Y. and Muthumani, S., (2019). Client cluster identification of internet bank services. *Journal of Computational and Theoretical Nanoscience*, 16(8), pp. 3554–3559.

Barman, D., Chowdhury, N., (2019). A novel approach for the customer segmentation using clustering through self-organizing map. *International Journal of Business Analytics*, 6(2), pp. 23–45.

Calvo-Porral, C. and Lévy-Mangin, J., (2020). An emotion-based segmentation of bank service customers. *International Journal of Bank Marketing*, 38(7), pp. 1441–1463.

Chawla, D., Joshi, H., (2021). Segmenting mobile banking users based on the usage of mobile banking services. *Global Business Review*, 22(3), pp. 689–704.

Dang Tran, H., Le, N. and Nguyen, V.-H., (2023). Customer churn prediction in the banking sector using machine learning-based classification models. *Interdisciplinary Journal of Information, Knowledge, and Management*, 18, pp. 087–105.

Dimitriadou, E., Dolničar, S. and Weingessel, A., (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67(1), pp. 137–159.

Djurisic, V., Kascelan, L., Rogic, S. and Melovic, B., (2020). Bank CRM Optimization Using Predictive Classification Based on the Support Vector Machine Method. *Applied Artificial Intelligence*, 34(12), pp. 941–955.

Firdaus, U., Utama, D.-N., (2021). Development of bank's customer segmentation model based on rfm+b approach. ICIC Express Letters, *Part B: Applications*, 12(1), pp. 17–26.

Formisano, V., Moretta Tartaglione, A., Fedele, M. and Cavacece, Y., (2020). Banking services for SMEs' internationalization: evaluating customer satisfaction. *The TQM Journal*, 33(3), pp. 662–680.

Hamal, S., Senvar, O., (2021). Comparing performances and effectiveness of machine learning classifiers in detecting financial accounting fraud for Turkish SMEs. *International Journal of Computational Intelligence Systems*, 14(1), p. 769.

He, W., Hung, J.-L. and Liu, L., (2022). Impact of big data analytics on banking: A case study. *Journal of Enterprise Information Management*.

Hung, P., Lien, N. and Ngoc, N., (2019). Customer Segmentation Using Hierarchical Agglomerative Clustering. *Proceedings of the 2019 2nd International Conference on Information Science and Systems*.

Kovács, T., Ko, A. and Asemi, A., (2021). Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis. *Journal of Big Data*, 8(1).

Oleynik, M., Formánek, T., (2020). Predicting Default Probability of Bank's Corporate Clients in the Czech Republic. Comparison of Generalized Additive Models and Support Vector Machine Approaches. *Software Engineering Perspectives in Intelligent Systems*, pp. 709–722.

Osowski, S., Sierenski, L., (2020). Prediction of Customer Status in Corporate Banking Using Neural Networks. *2020 International Joint Conference on Neural Networks* (IJCNN).

Rajaobelina, L., Brun, I. and Ricard, L., (2019). A classification of live chat service users in the Banking Industry. *International Journal of Bank Marketing*, 37(3), pp. 838–857.

Tungjitnob, S., Pasupa, K. and Suntisrivaraporn, B., (2021). Identifying SME customers from click feedback on mobile banking apps: Supervised and semi-supervised approaches. *Heliyon*, 7(8).

Umuhoza, E., Ntirushwamaboko, D., Awuah, J. and Birir, B., (2020). Using unsupervised machine learning techniques for behavioral-based credit card users segmentation in Africa. *SAIEE Africa Research Journal*, 111(3), pp. 95–101.

Vijayalakshmi, M., Gupta, S.-S. and Gupta, A., (2020). Loan approval system through customer segmentation using big data analytics and machine learning. *International Journal of Advanced Science and Technology*, 29(6), pp. 2374–2380.

Yanık, S., Elmorsy, A., (2019). SOM approach for clustering customers using credit card transactions. *International Journal of Intelligent Computing and Cybernetics*, 12(3), pp. 372–388.

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 167–184, https://doi.org/10.59170/stattrans-2023-069
Received – 12.12.2022; accepted – 02.08.2023

Changepoint detection with the use of the RESPERM method – a Monte Carlo study

Grzegorz Kończak[1], **Katarzyna Stąpor**[2]

Abstract

RESPERM (residuals permutation-based method) is a single changepoint detection method based on regression residuals permutation, which can be applied to many physiological situations where the regression slope can change suddenly at a given point. This article presents the results of a Monte Carlo study on the properties of the RESPERM method for single changepoint detection in a linear regression model. We compared our method with a well-known segmented method for detection breakpoint in linear models. The Monte Carlo study showed that when the input data are very noisy, the RESPERM method outperforms the segmented approach in terms of variance, and in the case of bias, the results of the two methods are comparable.

Key words: changepoint detection, RESPERM, permutation methods.

1. Introduction

The changepoint analysis plays an important role in many fields including time series analysis, quality control, economy, finance, genome research, signal processing, medical research, and many others. The changepoint problem is generally referred to as identifying the changes at unknown times and of estimating the location of changes in stochastic processes. This problem was initially discussed by Quandt (1958, 1960) and also Chow (1960). The changepoint problem can be formulated in several models and numerous methodological approaches have been implemented in examining these models. Maximum-likelihood estimation, Bayesian estimation, piecewise regression, quasi-likelihood and non-parametric regression as well as grid-searching are among the methods which have been applied to resolving challenges in changepoint problems (Julious, 2001). Changepoint detection methods can be online (detecting in real-time

[1] Department of Statistics, Econometrics and Mathematics, University of Economics in Katowice, Poland. E-mail: grzegorz.konczak@ue.katowice.pl. ORCID: https://orcid.org/0000-0002-4696-8215.
[2] Department of Applied Informatics, Silesian University of Technology, Poland. E-mail: katarzyna.stapor@polsl.pl. ORCID: https://orcid.org/0000-0003-3003-6592.

setting) or offline, that retrospectively detect changes when all samples are received. The problem of estimation the location of changepoints has been intensively studied in the literature including significance testing of estimate as well (a pioneering work of Chow (1960) should be mentioned here, but there are also many newer ones). A detailed review, as well as the classification and evaluation of different changepoint detection methods based on the selected criteria can be found, for example, in (Aminikhanghahi and Cook, 2017; Truong et al., 2020).

In this article we consider locating changepoint in a linear regression model with one changepoint. There are different nomenclatures to describe the so-called changepoint regression, such as „segmented" (Lerman, 1980), „broken-line" (Ulm, 1991), „structural change" (Bacon and Watts, 1971) and some others, in which the relationship between the response and the explanatory variable (or variables) is piecewise linear. There are two possibilities in changepoint regression. The first one is a continuous piecewise model, in which regression lines with different slopes are connected at unknown changepoints. In the second, discontinuous model, the regression lines jump at the changepoint (Figure 1).

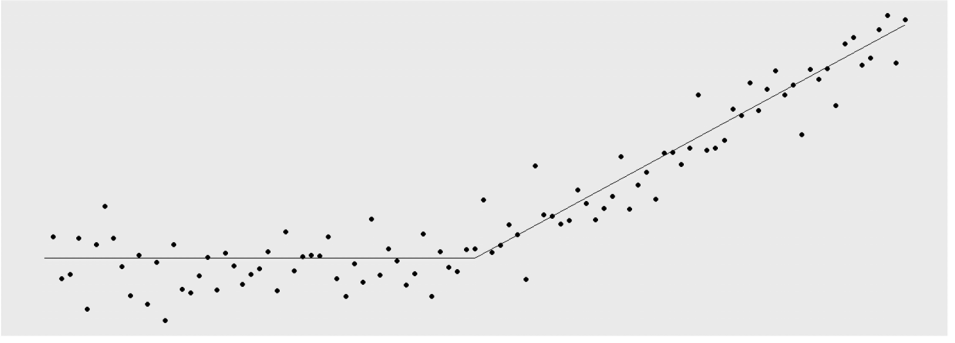

Figure 1. The illustrative example of one changepoint in linear regression model

The paper is presented as follows. After a brief introduction to the problem of changepoint detection in the literature, in Section 2 we present the general model for changepoint regression and describe, very shortly, two selected methods used in the comparison further. In Section 3 we present the concepts and the process of the Monte Carlo study. Sections 4 and 5 describe the Monte Carlo simulation study and the obtained results with discussion, respectively. The last section provides conclusion remarks.

2. The changepoint regression model and the selected methods

The general model of changepoint regression with J changepoints ($J+1$ regimes) can be written as:

$$y_i = \begin{cases} \beta_{0(1)} + \beta_{1(1)} x_i + \varepsilon_{i(1)} & x_i \le chp_1 \\ \beta_{0(2)} + \beta_{1(2)} x_i + \varepsilon_{i(2)} & chp_1 < x_i \le chp_2 \\ \vdots & \\ \beta_{0(j)} + \beta_{1(j)} x_i + \varepsilon_{i(j)} & chp_{j-1} < x_i \le chp_j \\ \vdots & \\ \beta_{0(J+1)} + \beta_{1(J+1)} x_i + \varepsilon_{i(J+1)} & chp_J < x_i \end{cases} \quad (1)$$

In this paper we will only consider the models where the changepoints are defined in terms of only one regressor denoted here as x. The model (1) considers only one regressor x along which the changepoints lie, however other regressors can also be included in the model. In the model (1) $i = 1, \ldots, n$ are the observation numbers, n is the total sample size, chp_j $j = 1, \ldots, J$ are the changepoint parameters for the regressor x which satisfy:

$$chp_1 < chp_2 < \ldots < chp_J$$

$\varepsilon_{i(j)}$ are independent, identically distributed random variables, having mean zero and possibly differing variances σ_j^2 $j = 1, \ldots, J$, respectively.

The changepoints locations given by chp_j are the unknown parameters to be estimated, but the number of changepoints in the observed sample is assumed to be known. The model above also assumes that regressor x can be ordered (that means partitioned by the changepoints chp_j), and sufficient number of observations can be placed in each intervals, for which the data came from different regimes (models data generation) for reliable estimation and inference. It is up to the user to determine what is „sufficient", but the rule of thumb may be to ensure at least 10 observations in each regime (Sheykhfard et al., 2020; Applied Regression Analysis, 2023).

The above model describes both continuous and discontinuous scenario, but to enforce the connected regression lines, the regression parameters must be constrained so that:

$$\beta_{0(j)} + \beta_{1(j)} \cdot chp_j = \beta_{0(j+1)} + \beta_{1(j+1)} \cdot chp_{j+1}. \quad (2)$$

There are many methods in the literature to detect the location of the unknown changepoints and estimate the regression model (1). We will consider two methods: the SEGMENTED (Muggeo, 2008), and RESPERM (Sommer et al., 2022) methods which will be shortly described below.

2.1. Segmented method

Muggeo (2003) proposed a method called *segmented regression*, which allows for multiple unknown changepoints but is restricted to continuous regression lines.

We briefly present this method for the single changepoint model ($J = 1$) with location in chp. The model (1) with constraints (2) for the segmented regression can be estimated iteratively via the following linear function of predictors:

$$\beta_0 + \beta_{1(1)}x_{i1} + (\beta_{1(2)} - \beta_{1(1)})(x_{i1} - chp_0)I(x_{i1} > chp_0) - \gamma \cdot I(x_{i1} > chp_0), \tag{3}$$

where $I(A)$ is an indicator function for an event A, chp_0 is an initial estimate for the changepoint, and γ is a re-parametrization of chp_0 that appears as a linear and continuously valued parameter which facilitates the estimation procedure. Muggeo (2003) recommends maximum likelihood (ML) under Gaussian errors with constant variance across regimes (homoscedasticity). The model enables for simultaneous ML inference on all model parameters, including the changepoint location. The procedure for segmented regression can be sketched as follows:

(1) choose an initial changepoint estimate chp_0,
(2) given the current estimated changepoint chp_0 estimate model (3) by Gaussian ML and update the changepoint via $\widehat{chp} = chp_0 + \hat{\gamma}/(\hat{\beta}_{1(2)} - \hat{\beta}_{1(1)})$,
(3) if $\hat{\gamma}$ is sufficiently closed to zero then stop, else set $chp_0 = \widehat{chp}$ and go to step (2),
(4) iterate steps (2) and (3) until termination.

In the above procedure $\hat{\gamma}$ measures the distance between the two fitted regression lines at the current estimate $\widehat{chp}$. It is not clear that this method can be extended to cover the discontinuous case. Muggeo also proposed the segmented package in R (Muggeo, 2008), which enables to estimate the parameters in GLM with segmented, continuous relationships via ML.

2.2. The residuals permutation-based method (RESPERM)

The RESPERM method was designed for detecting a changepoint in the EEG signal waveform in an experiment with showing a new brain-learned face. By definition, the EEG method is just for studying waveforms over time - hence this method uses only 1 regressor.

The RESPERM method considers a discrete set (i.e. a grid) of possible changepoint locations; for each possible changepoint an optimal set of parameter estimates for each of the two regimes is determined. The finally selected changepoint optimizes the chosen estimation criterion - Cohen's effect size (Cohen 1988), estimated based on the permutation method. Moreover, this method allows for different variances in each regime.

Let us consider n experimental data observations, which could be denoted by (x_i, y_i) for $i = 1, 2, \ldots, n$. We will consider two simple linear regression models with changepoint *chp*:

$$y = \beta_{01} + \beta_{11}x + \varepsilon_1 \quad \text{for } x \leq chp,$$

$$y = \beta_{02} + \beta_{12}x + \varepsilon_2 \quad \text{for } x \geq chp,$$

where y is the dependent variable, x is the regressorβ_{01}, β_{11}, β_{02} and β_{12} are parameters of linear models and ε_1, ε_2 are error terms.

The main goal of this method is to detect a change in the slope in the linear regression model. If $\beta_{11} \neq \beta_{12}$ then *chp* is a breakpoint in the considered linear model. To detect the breakpoint *chp* we use the Cohen effect size. Cohen (1988) defines an effect size *d* as follows:

$$d = \frac{m_A - m_B}{\sigma}, \tag{4}$$

where m_A and m_B are populations means under considerations expressed in raw (original) measurement units and σ is the standard deviation of either population of measurements. Instead of m_A, m_B and σ we use β_{11}, β_{12} and standard deviations of beta's σ_β.If the errors are very non-normal then the standard methods may not be reliable and a resampling method may offer some improvement (*Davidson* and Hinkley, 1997). The permutation of the residuals method is used to estimate the standard deviation σ_β. So, (4) could be rewritten in the form

$$d = \frac{\beta_{12} - \beta_{11}}{\sqrt{\frac{(k-1)S^2_{\beta 11} + (n-k-1)S^2_{\beta 12}}{n-2}}} \tag{5}$$

where *k* is the number of observations in the first group.

For the estimated two linear models the residuals are obtained separately for each one. Then, the residuals are permuted N_{perm}=1000 times and for each case, the coefficients β_{11}, β_{01} (the slope and the intercept of the first line) and β_{12} , β_{02} (the slope and the intercept of the second line) were estimated. Based on these estimates, the standard deviations $S_{\beta 11}$ and $S_{\beta 12}$ of the β_{11} and β_{12} coefficients are assessed.

Let us consider two sets $\mathbf{S_1} = \{(x_i, y_i): i = 1, 2, \ldots, k\}$ and $\mathbf{S_2} = \{(x_i, y_i): i = k + 1, k + 2, \ldots, n\}$ where $k = s, s+1, \ldots, n-s$ and s is the parameter of the method. Based on these two sets we get two regression lines with slopes $\hat{\beta}_{11}$and $\hat{\beta}_{12}$ and intercepts $\hat{\beta}_{01}$and $\hat{\beta}_{02}$. Let *d* now measure Cohen's effect (5) of the slope of the linear regression. We find k^* which maximizes Cohen's effect size *d* using the formula:

$$k^* = \min_{k \in K} \left\{ k: d(k) = \max_{k \in K} d(k) \right\} \tag{6}$$

where K = {s, s+1, …, *n*-s} and *s* is the parameter of the permutation-based method. So, the changepoint can be expressed as

$$chp = x_{k^*}.$$

3. Monte Carlo study

In the Monte Carlo study, the series of n = 100 observations with one changepoint were generated. The observations were generated according to the following model:

$$y = \begin{cases} 2 + p\varepsilon_j & \text{for } x \leq 50, \\ 2 + (x - 50) + pq\varepsilon_j & \text{for } x > 50, \end{cases} \tag{7}$$

where the covariate x = 1, 2, …, 100, coefficient p describes the level of the noise (p=1 for minor noise, p=3 for major noise, p = 5 for dominant noise), coefficient q = *1* for equal variances and q=2/3 for unequal variances, changepoint is established to chp = 50, ε_j (j = 1, 2, 3, 4) is the error term:

$\varepsilon_1 = \frac{1}{3}\varepsilon_N$, where ε_N has a standard normal distribution,

$\varepsilon_2 = \varepsilon_U - 0.5$, where ε_U has the uniform distribution on the [0, 1] interval,

$\varepsilon_3 = \varepsilon_{\beta 22} - 0.5$, where $\varepsilon_{\beta 22}$ has the beta distribution with shape parameters $s_1 = 2$, $s_2 = 2$ (symmetric distribution),

$\varepsilon_4 = \varepsilon_{\beta 26} - 0.25$, where $\varepsilon_{\beta 26}$ has the beta distribution with shape parameters $s_1 = 2$, $s_2 = 6$ (asymmetric distribution).

The main characteristics of these distributions are presented in Table 1.

Table 1. Expectations and variance of the distributions of errors in the present simulations

Error ε	E(ε)	D²(ε)	Error ε	E(ε)	D²(ε) Noise		
					Minor	Major	Dominant
ε_N	0	1	ε_1	0	$\frac{1}{9}$	1	$\frac{25}{9}$
ε_U	0	$\frac{1}{12}$	ε_2	0	$\frac{1}{12}$	$\frac{9}{12}$	$\frac{25}{12}$
$\varepsilon_{\beta 22}$	0.50	$\frac{1}{20}$	ε_3	0	$\frac{1}{20}$	$\frac{9}{20}$	$\frac{5}{4}$
$\varepsilon_{\beta 26}$	0.25	$\frac{1}{48}$	ε_4	0	$\frac{1}{48}$	$\frac{9}{48}$	$\frac{25}{48}$

Source: own calculations.

The expected values of errors in all models are 0 but differ in variance from 1/48 (for minor noise) up to 25/9 (dominant noise). The first part of Table 1 shows the four variants of the considered distributions (normal, uniform, beta symmetric and beta asymmetric) and their parameters. The second part of Table 1 shows the distributions and parameters used in the Monte Carlo study. Figure 2 shows the empirical density functions of errors for each model.

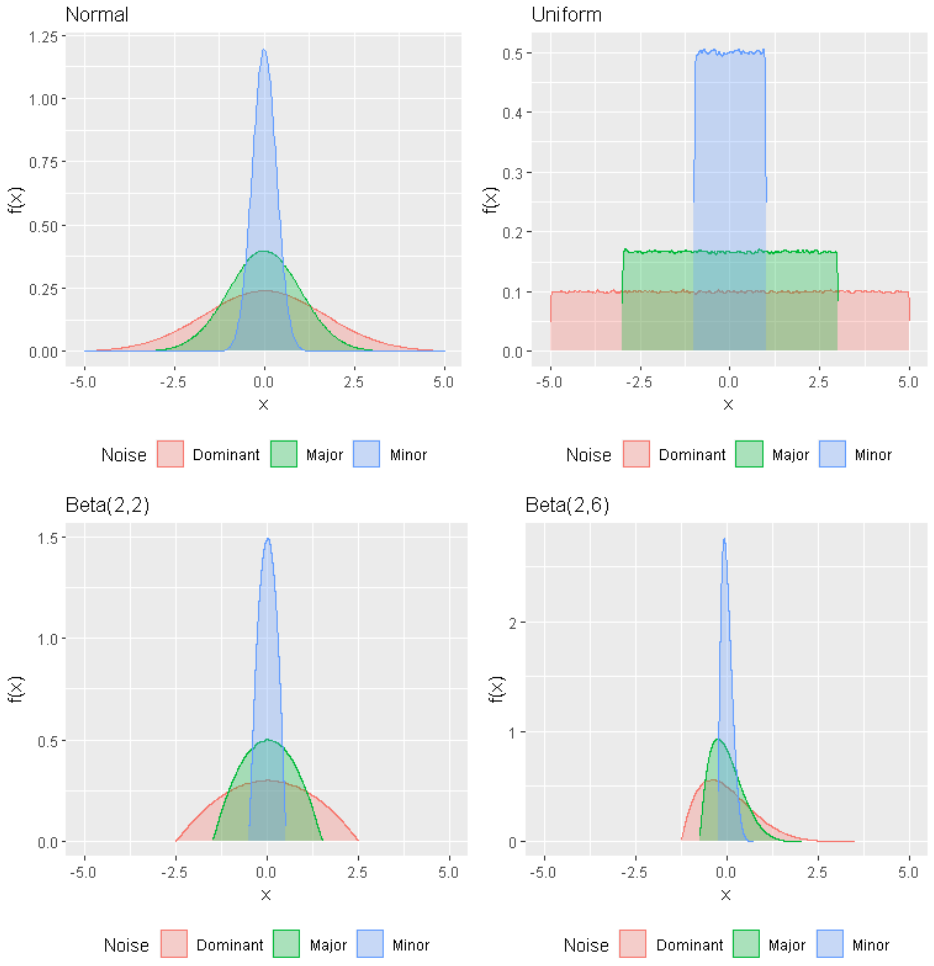

Figure 2. The empirical distributions of considered errors in computer simulations

Source: own calculations.

Six variants of error models are considered in the Monte Carlo study. The first one is the model with equal variances for the two parts considered and the second one for unequal variances. In the second case, the density function of errors from the first part were multiplied by the constant 2/3. In both of these variants, the noise is at a low level (minor noise). The next two variants of the Monte Carlo study are similar to the first two, but the density functions of errors were multiplied by 3 (major noise). The last two variants of the Monte Carlo study are similar to the first two, but the density functions of errors were multiplied by 5 (dominant noise). Typical random series of observations for equal variances and unequal variances are shown in Figure 2. The first part in model

(7) (from 0 to 50) is a linear model with a slope β_{11} equal to 0. The second part (from 50 to 100) has the slope β_{12} equal to 0.5.

The two cases presented at the top of Figure 3 relate to noise at the minor level. The next two cases in the middle row relate to noise at the major level. The last two cases at the bottom relate to noise at the dominant level. In the Monte Carlo study, the changepoint was set to *chp* = 50.

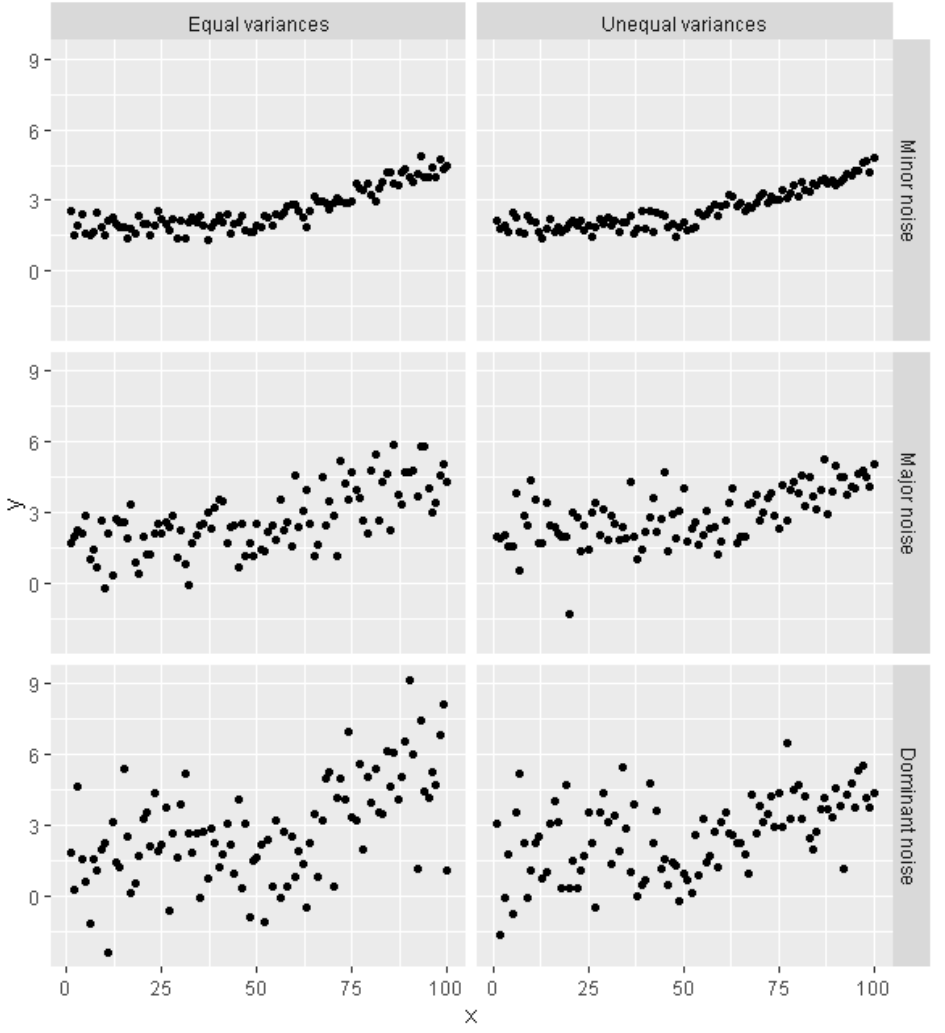

Figure 3. Typical time series with a single changepoint at *chp* = 50 with normal errors (noise: minor – top, major – middle, dominant – bottom, variance: equal – left, unequal – right)

Source: own calculations.

We considered the following steps in the series of computer simulations:

1. We generated $N = 100$ times the datasets according to model (7).
2. We estimated the changepoint with the use of the *permutation* methods for each dataset. For each point $k = 10, 11, \ldots, 90$ variances of parameters of the linear models are estimated based on $N_{perm} = 1000$ permutations of residuals.
3. The changepoint was estimated as a parameter k^* which maximizes Cohen's d size effect as in formula (6).
4. The estimated changepoint was obtained using the *segmented* methods.
5. To compare the two considered methods the standard deviation *SD* and *Bias* were calculated.

4. Results of the Monte Carlo study and discussion

In the Monte Carlo study, the changepoint has been established to $chp = 50$ and the number of replications to $N = 100$. The number of permutations for estimating the standard deviation of the slope coefficients of the regression function was assumed $N_{\mathbf{perm}} = 1000$. Let us denote the estimated changepoint in i-th replication of the model as chp_i. The estimation of standard deviation (*SD*) and *Bias* are also included in the results:

$$SD \approx \sqrt{VAR}, \tag{8}$$

$$Bias \approx \frac{1}{N}\sum_{i=1}^{N} chp_i - chp, \tag{9}$$

where $VAR = \frac{1}{N}\sum_{i=1}^{N}(chp_i - \overline{chp})^2$ and $\overline{chp} = \frac{1}{N}\sum_{i=1}^{N} chp_i$.

Taking into account the lack of change or the occurrence of a variance change and the noise level as minor or dominant, 6 simulation variants were considered (see Fig. 2)

- equal variances and minor noise,
- equal variances and major noise,
- equal variances and dominant noise,
- unequal variances and minor noise,
- unequal variances and major noise,
- unequal variances and dominant noise.

In each case, four types of error distribution were considered in the Monte Carlo study.

4.1. Minor noise

Table 2 presents the values of *SD* and *Bias* of the changepoint estimations by the segmented method and the permutation-based method for the first model considered with four types of error distributions. In all these cases, noticeably smaller (sometimes

by up to 50%) *SDs* were obtained by the permutation-based than with the segmented method. In all cases the values of *Bias* were similar for both methods.

Table 2. Estimated values of SD and Bias of changepoint estimating for minor noise

Errors distributions	Equal variances				Unequal variances			
	Segmented		RESPERM		Segmented		RESPERM	
	SD	Bias	SD	Bias	SD	Bias	SD	Bias
Normal	3.63	0.37	2.26	- 0.14	3.91	- 0.21	2.40	- 0.35
Uniform	2.32	0.16	1.84	- 0.27	2.20	0.00	1.65	- 0.45
Beta (2,2)	2.01	0.06	1.60	- 0.16	1.69	- 2.87	1.65	- 1.76
Beta (2,6)	2.46	0.02	1.76	0.07	1.00	- 1.36	1.41	- 1.13

Source: own calculation in R program.

The top row of Figure 4 shows box-whisker plots for the estimated changepoints for the residuals permutation-based method and the segmented method for the equal variances case. Both methods lead to similar results but the variance of changepoint estimates is much smaller for the permutation-based method than for the segmented method. Maximal errors of changepoint estimation are also greater for the segmented method than for the permutation-based method in each considered case. The advantage of the permutation method is that there are no requirements for the type of error distribution. The bottom row of Figure 4 shows box-whisker plots for the estimated breakpoints for each method for the unequal variance case. The breakpoint estimates from the *permutations-based* method and the *segmented* method led to similar results. For the first three considered error distributions, the variance of the changepoint assessment was smaller for the permutations-based than for the segmented method. Only for the asymmetric beta error distribution, the evaluation variance was smaller for the segmented method. In each case maximal errors of changepoint estimates were greater for the segmented than for the permutations-based method. There was a noticeable estimation bias for both methods in the symmetric beta error distribution. In this case, the bias was greater for the segmented method.

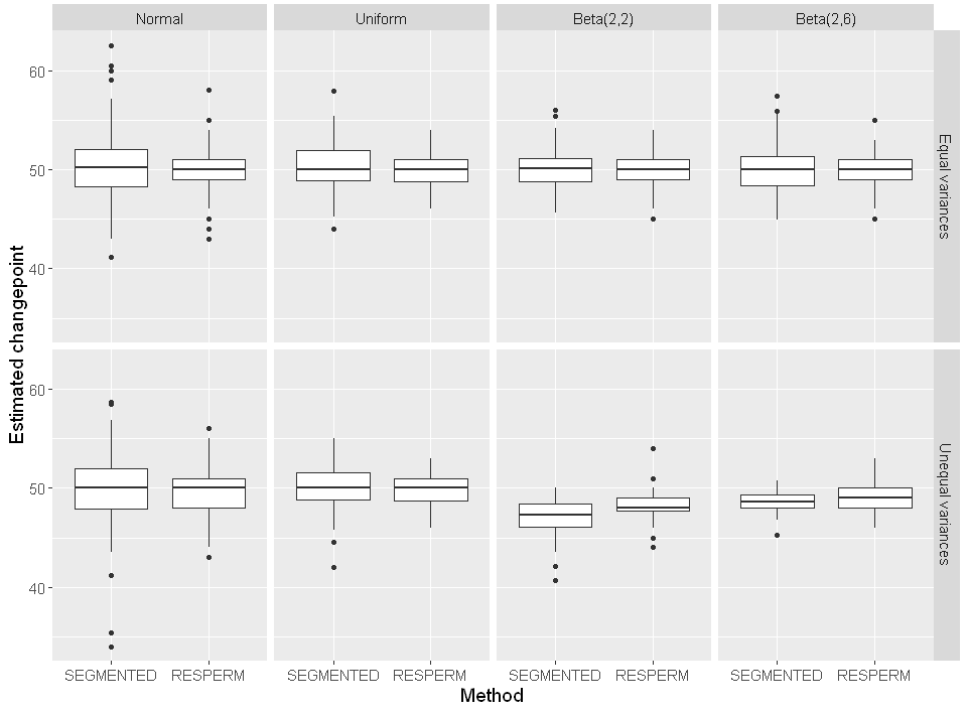

Figure 4. Changepoint estimates for the segmented and permutations -based methods for minor noise and four kinds of distributions of errors. Top: equal variances in both regimes. Bottom: unequal variances.

Source: own calculation in R program.

4.2. Major noise

Table 3 presents *SD* and *Bias* of the changepoint estimation for the segmented and the permutation methods in the cases of equal and unequal variances but at major levels. Four types of error distributions were taken into account as before. In all the error types considered, noticeably smaller errors in the estimation of the changepoint location were obtained using the permutation-based method.

The top row of Figure 5 shows the box-whisker plots for the estimated changepoints when there is major noise for the equal variances case. The mean changepoint estimates with the residuals permutation-based method and the segmented method lead to similar results but the variance is much smaller for the permutation-based method than for the segmented method. The bottom row of Figure 5 shows box-whisker plots for the estimated changepoints for the two methods for the unequal variances case. The variance of the changepoint estimates is consistently larger

for the segmented than for the permutation-based method. Maximal errors of changepoint estimates from the segmented method exceed those from the permutation-based method in each case.

Table 3. *SD* and *Bias* of changepoint estimates for major noise

Errors distributions	Equal variances				Unequal variances			
	Segmented		RESPERM		Segmented		RESPERM	
	SD	Bias	SD	Bias	SD	Bias	SD	Bias
Normal	12.96	0.06	7.86	- 0.63	9.13	- 0.80	6.80	- 1.04
Uniform	11.08	- 0.33	7.71	- 0.10	8.90	- 1.58	5.57	- 1.92
Beta (2,2)	8.07	0.83	4.63	0.16	6.09	- 0.66	3.11	- 1.10
Beta (2,6)	4.09	0.34	2.74	0.21	3.43	- 0.79	1.93	- 0.72

Source: own calculation in R program.

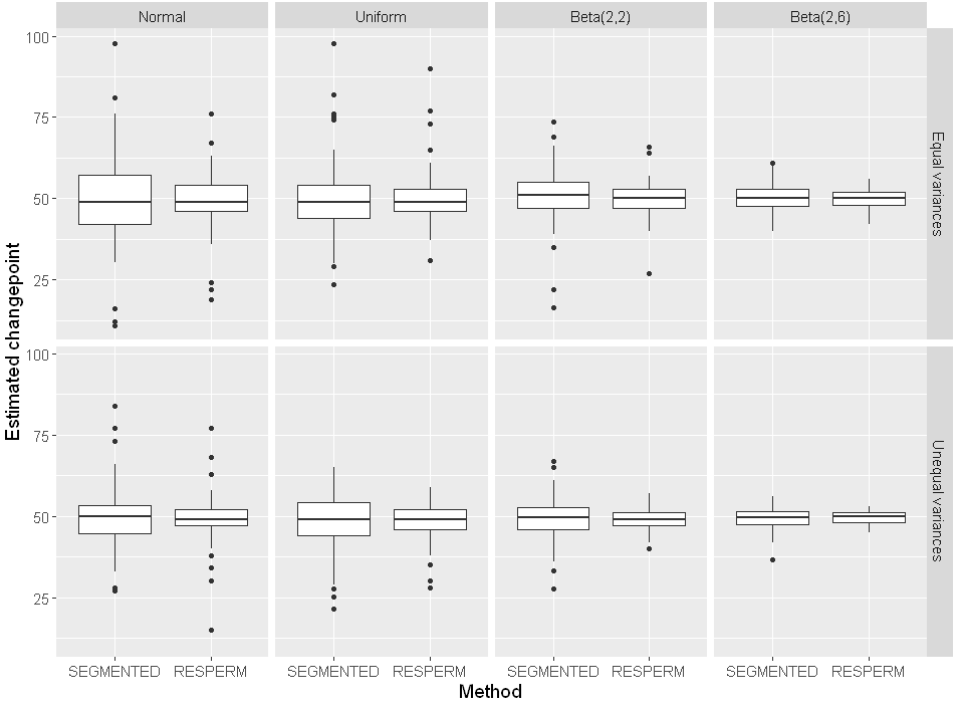

Figure 5. Changepoint estimates for the segmented and permutations-based methods for major noise and four kinds of distributions of errors. Top: equal variances in both regimes. Bottom: unequal variances.

Source: own calculation in R program.

4.3. Dominant noise

Table 4 presents *SD* and *Bias* of the changepoint estimates for the segmented and permutation methods for equal and unequal and dominant noise. In all four error distribution types, changepoint estimation errors were noticeably smaller for the permutation-based method.

Table 4. Estimated values of *SD* and *Bias* of changepoint for dominant noise

Errors distributions	Equal variances				Unequal variances			
	Segmented		RESPERM		Segmented		RESPERM	
	SD	Bias	SD	Bias	SD	Bias	SD	Bias
Normal	20.47	0.51	17.38	0.32	18.41	-1.92	14.66	-2.89
Uniform	19.54	-0.54	15.32	-0.86	16.21	-2.64	13.30	-4.84
Beta (2,2)	15.36	-1.51	10.35	-0.94	12.40	-1.65	8.51	-2.21
Beta (2,6)	8.05	0.90	4.16	-0.27	6.46	-1.28	3.51	-0.90

Source: own calculation in R program.

The top row of Figure 6 shows box-whisker plots for the estimated changepoints for both methods in the case of equal variances. The changepoint estimating with the residuals permutation-based method and the segmented method leads to similar results but the variance of changepoint assessment is much lower for the permutation-based method than for the segmented method. The relative bias of the changepoint assessment in most cases was much less than 2%. For the permutation method with normally distributed errors Average Relative Bias (*ARB*) was 0.64%. The bottom row of Figure 6 shows box-whisker plots for the estimated changepoints for the two methods in the case of unequal variances. The variance of the changepoint estimation was greater for the segmented than for the permutation-based method. Maximal errors of changepoint estimates were larger for the segmented than for the permutation-based method in each case.

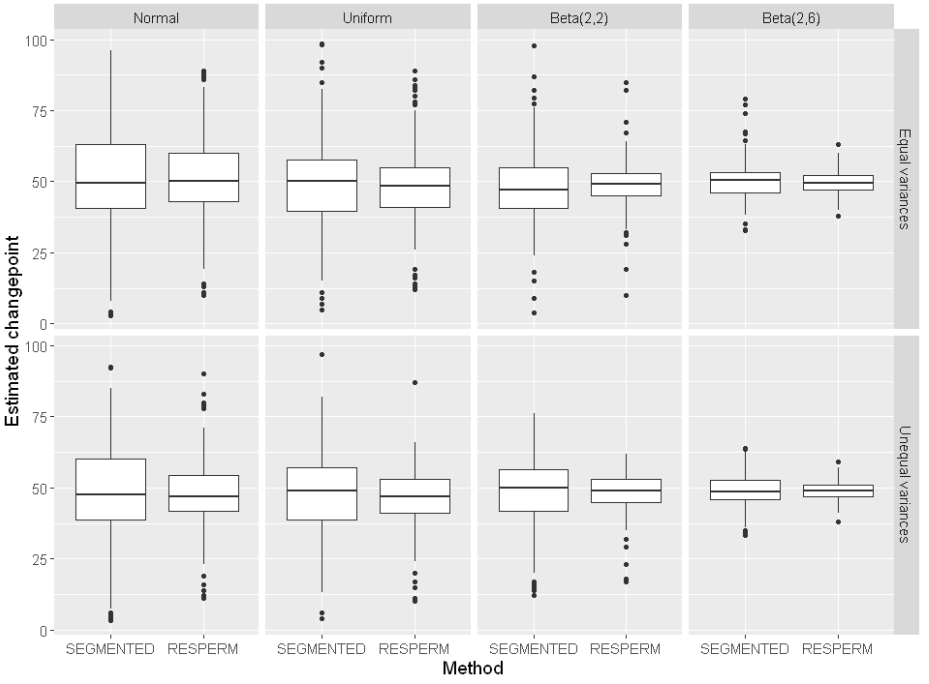

Figure 6. Changepoint estimates for the segmented and permutations-based methods for dominant noise and four kinds of distributions of errors. Top: equal variances in both regimes. Bottom: unequal variances.

Source: own calculation in R program.

5. Sensitivity analysis of the segmented and permutation-based methods

In the previous section, the standard deviation and *Bias* for the assessment of the changepoint location were presented for four types of error distribution: normal, uniform, Beta (2,2), and Beta (2,6). In each of these cases, the variance in the two regimes (before and after the changepoint) was either equal (constant) variances or unequal, that is, it was smaller after the changepoint. All simulations were performed for the changepoint $chp = 50$.

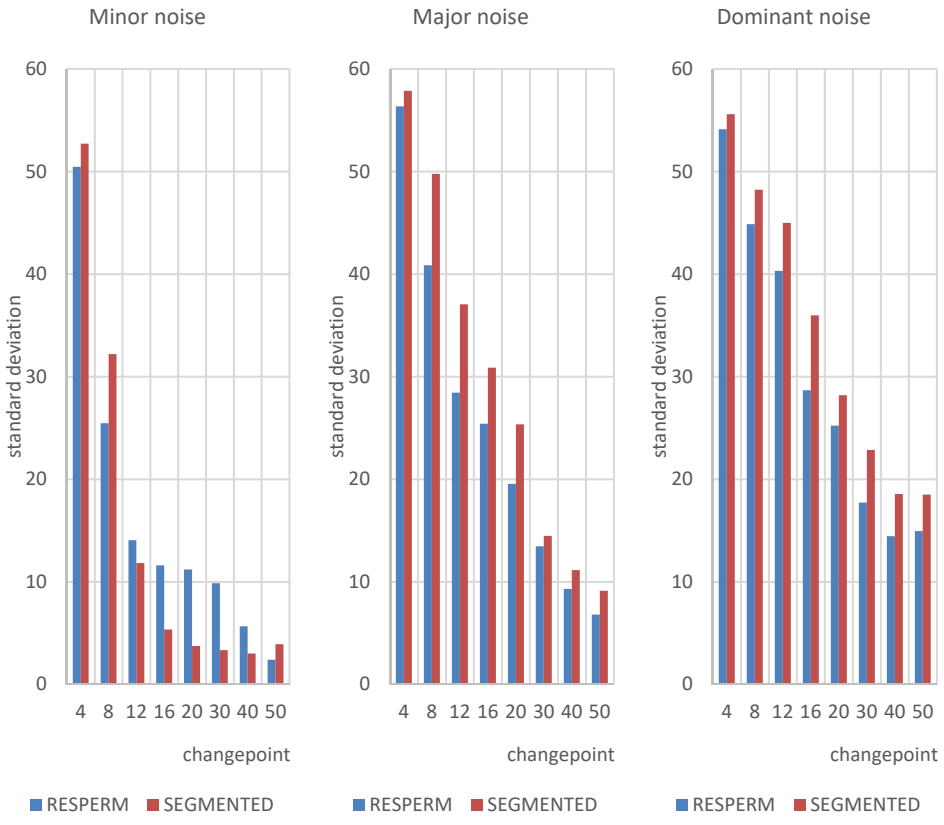

Figure 7. Standard deviation (SD) for RESPERM and SEGMENTED methods for minor, major and dominant noise in the case of normally distributed errors

Source: own calculation in R program.

The results of these simulations showed that for cases with different variance (i.e. reduced in the 2nd part), slightly more precise estimates (i.e. having lower bias) of *chp* were obtained, both for the segmented method and the permutation. Therefore, the following section focuses on time series only with unequal variance. Noticeable biases occurred only in a few cases in both methods. The further study thus considers only a comparative standard deviation (*SD*) analysis. The standard deviation analysis was performed for different changepoints: *chp* = 4, 8, 12, 20, 30, 40, 50. For these changepoints, only errors with normal distribution and diminished variance after the *chp* were considered. Three levels of variance in the distribution of random errors were taken into account: minor, major and dominant errors.

The purpose of this part of the computer simulations was to analyse the sensitivity of the RESPERM method. Sensitivity analysis determines how different values of the independent variable (changepoint) affect a particular dependent variable (estimated changepoint) under a given set of assumptions. In other words, sensitivity analyses examine how different sources of uncertainty in a mathematical model contribute to the overall uncertainty of the model.

It should be noted that in many cases (especially for *chp* = 4, 8, 12) the segmented method did not find the change point, so we had to run this method multiple times until a changepoint was found. The standard deviation of estimation (*SD*) of the changepoint (*chp*) for the above-mentioned cases is shown in Figure 7. In the case of minor error, the segmented method is mostly characterized by a smaller standard deviation than the permutation-based method. However, in the case of greater variance of random errors (minor and dominant), the permutation-based method is characterized by a consistently smaller standard error than the segmented method for each changepoint analysed. Overall, it can be seen that a higher level of random errors leads to a greater standard error in the changepoint assessment, but still with less error for the permutation method.

6. Conclusions

We considered the problem of changepoint detection in linear regression models based on noisy data. This residuals permutation-based method maximizes Cohen's effect size measure *d* with the parameters estimated by the permutation of residuals in the linear model. The residuals permutation-based method was compared in a number of computer simulations. In the simulation study six variants of noise were considered from normal, uniform and two variants of beta distributions together with two cases of equal and unequal variances. Three levels of variance in the distribution of random errors were taken into account: minor, major and dominant errors. The simulations were performed for different locations of changepoint in time series.

The results showed that for cases with different variance (i.e. reduced in the 2nd part), slightly less biased estimates of *chp* are obtained, both for the segmented method and the permutation. In the case of minor errors, that is for relatively clean data, the segmented method in most cases is characterized by a smaller standard deviation than the residuals permutation-based method. For more noisy data, that is in the case of major and dominant greater variance of random errors, the permutation method is superior and characterized by a smaller standard error than segmented for each of the analysed changepoints (i.e. independent of the location of *chp*). In these cases the biases of both methods are comparable. Although not systematically explored,

for early changepoints it was often hard to find such a changepoint with the regimented methods. Note that the RESPERM method was planned to work with 1 regressor only.

The current article presents the results of a simulation experiment designed to study the properties and behaviour of the RESPERM method designed and used earlier on for the purpose of detecting points of interest on the time course of the EEG signal. And it was in the paper Sommer et al. (2022) that the application of the method to real data from an EEG experiment conducted by the authors was presented.

In summary, the present results from the simulation study indicate that the proposed residuals permutation-based method shows a better performance in identifying a changepoint in noisy data and therefore may be recommended in such scenarios. The RESPERM method in such cases is more precise and the loadings of both methods are comparable. For data with minor noise, the results of the two methods are comparable.

References

Aminikhanghahi, S., Cook, D. J. (2017). A Survey of Methods for Time Series Changepoint Detection. *Knowl Inf Syst.,* 2017 May, 51(2), pp. 339–367. doi:10.1007/s10115-016-0987-z.

Applied Regression Analysis, (2023). https://online.stat.psu.edu/stat462/node/185/ (24.07.2023).

Bacon, D. W., Watts, D. G., (1971). Estimating the Transition Between Two Intersecting Straight Lines. *Biometrika*, 58, pp. 525–534.

Chow, G., (1960) Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28, pp. 591–05.

Cohen, J., (1988). Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, *Publishers*, New York.

Davidson, A. C., Hinkley, D. V., (1997). Bootstrap Methods and Their Application. *Cambridge University Press*, New York.

Julious, S. A., (2001) Inference and Estimation in a Changepoint Regression Problem. *The Statistician*, 50, pp. 51–61.

Lerman, P. M., (1980). Fitting Segmented Regression Models by Grid Search. *Applied Statistics*, 29, pp. 77–84.

Muggeo, V. M. R., (2003). Estimating Regression Models with Unknown Break-Points. *Statistics in Medicine*, 22, pp. 3055–3071.

Muggeo, V. M. R., (2008). Segmented: an R package to fit regression models with broken-line relationships. *R News,* 8/1, pp. 20–25.

Quandt, R. E., (1958). The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes. *Journal of American Statistical Association*, 53, pp. 873–880.

Quandt, R. E., (1960). Tests of the Hypotheses that a Linear Regression System Obeys Two Separate Regimes. *Journal of American Statistical Association* 55, p. 324.

Sheykhfard A., Haghighi F., Nordfiaern T., Soltaninejad M., (2020). Structural equation modelling of potential risk factors for pedestrian accidents in rural and urban roads, *International Journal of Injury Control and Safety Promotion*, 10.1080/17457300.2020.1835991, 28, 1, pp. 46–57.

Sommer, W., Stapor K., Kończak G., Kotowski K., Fabian P., Ochab J., Bereś A., Ślusarczyk G., (2022). Changepoint Detection in Noisy Data Using a Novel Residuals Permutation-Based Method (RESPERM). Benchmarking and Application to Single Trial ERPs, *Brain Sciences*, Vol. 12, iss. 5, nr art. 525.

Truong, Ch., Oudre L., Vayatis N., (2020). Selective Review of Offline Changepoint Detection Methods. *Signal Processing*, Vol. 167, February 2020, 107299.

Ulm, K., (1991). A Statistical Method for Assessing a Threshold in Epidemiological Studies. *Statistics in Medicine*, 10, pp. 341–349.

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 185–202, https://doi.org/10.59170/stattrans-2023-070
Received – 20.02.2023; accepted – 10.07.2023

Bayesian predictive probability design: theory and practical application in a prospective study

Adam Korczyński[1]

Abstract

In an experiment-based prospective study aiming to determine the efficiency of a treatment, the time by which it becomes clear whether a therapy is effective or not is critical. This applies specifically to clinical trials and refers to the same extent to both successful and futile therapies. This study seeks to answer the question when there is enough evidence allowing the trial to be finalised. The key is to find enough statistical signals, working on the smallest possible sample, to make a judgment whether to extend, continue or terminate the study. The Bayesian predictive design allows drawing conclusions about the prognosis of a study considering the actual results.

The article provides a theoretical background and presents a practical perspective, addressing the statistical properties and technical aspects of conducting a trial based on a predictive design. Additionally, the sensitivity of the design to the choice of prior distribution is discussed.

Key words: Prospective study analysis, adaptive design, predictive probability design, Bayesian statistics.

1. Bayesian adaptive design - overview

Bayesian predictive design[2] falls within the concept of adaptive design of a clinical trial allowing modifications of the trial conduct according to the intermediate results. One of the objectives is to end the trial as soon as the final outcome is apparent (George, Wang, & Pang, 2016, pp. 366–369). The predictive power within the clinical trial settings is described as "having a positive result from a trial based on the currently available data" (Heath, et al., 2020, p. 2). In fact, Bayesian design allows assessing positive or negative result and therefore the rules of stopping for either futility or efficacy.

[1] SGH Warsaw School of Economics, Poland. E-mail: akorczy@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-1533-7197.

[2] An overview of the Bayesian approach including comparison with the frequentist principles can be found in (Lesaffre & Lawson, 2012).

Additionally, it gives the ability to react to evidence on superior treatment factor allowing assignment of new subjects to the best dosing scheme under study. For clinical trials various incentives, primarily ethical (Zhou, Liu, Kim, Herbst, & Lee, 2008, p. 2; Yin, Chen, & Lee, 2012, p. 220) but also those related to safety, cost and time effectiveness (Heath, et al., 2020; Chen, Ibrahim, Lam, Yu, & Zhang, 2011) bring the adaptive design approach into the scope of methods attractive for the industry and regulatory agencies. In some settings of oncology, the studies require large number of patients and years of studies in order to gain the approval (Barker, et al., 2009). Therefore, timing and decision making is essential to the drug development.

The Bayesian design finds its application in non-inferiority trials which are with the aim to show similar effectiveness of a novel therapy compared to a standard but with additional benefits to patients. If a therapy turns to be inferior then it would ideally be stopped early. Clinical practice shows that in oncology interim analysis was limited in recent years and based on a review published in 2012 (Heath, et al., 2020, p. 2), only 36% of 72 non-inferiority oncology trials utilized a formal analysis.

The Bayesian framework can also be applied to targeted therapy, giving the opportunity of accounting for the variation in patients characteristics (biomarker profile) when assigning the therapy. An application of targeted therapy in patients with advanced non-small cell lung cancer with disease control rate as the primary endpoint is outlined in (Zhou, Liu, Kim, Herbst, & Lee, 2008). The disease control rate was monitored within the treatment and marker subgroups with the Bayesian design deciding about the randomization of patients to treatment arms, according to the ongoing assessment of the response. A simulation study showed higher disease control rate in the randomized patients in the adaptive design as compared to fixed randomization equivalent (Zhou, Liu, Kim, Herbst, & Lee, 2008, p. 11; Yin, Chen, & Lee, 2012, p. 231). Similarly, application is to be found in (Barker, et al., 2009) with the description of a trial aiming at identifying the biomarker profiles able to predict the response for each study treatment.

Another application is within the basket design comparing the results of the treatment in patients with tumors of various types (Simon, Geyer, Subramanian, & Roychowdhury, 2016). The goal is to detect the strata in which drug activity suggests promising future results, and those which should not be continued as the evidence is the opposite. In randomized settings adaptive design allows assigning patients to more promising therapies, and such a schedule can already be applied to phase II trials (Yin, Chen, & Lee, 2012; Harrington & Parmigiani, 2016). The drawback is though that adaptive design in randomized setting would require larger sample due to unbalanced patient allocation (Yin, Chen, & Lee, 2012, p. 234).

With appropriate design of the trial the ability to determine the final outcome could be very high. An overview of the sensitivity of the Bayesian predictive probability design

for non-informative priors is provided in (Mitchell, 2018). The conclusion from the study was that in about 93% simulated trials, the interim decision was in agreement with the final outcome.

Bayesian design offers a straightforward setting for implementing prior knowledge into the estimation. The gain from using the prior knowledge is in the possibility to reduce the final sample size ensuring sufficient power to detect the effect (Chen, Ibrahim, Lam, Yu, & Zhang, 2011, p. 1167). In the referenced study, the reduction of the sample size between an informative and non-informative approach was from n=1480 to 1080 patients. This finding would not be however applicable to randomized trials with adaptive design, as then the distribution of patients over the treatment arm is affected by the outcome and this introduces the lack of balance between the arms.

The following sections describe and discuss the special case of a Bayesian design with predictive probabilities at the end of the trial using a binary endpoint. An example application is provided for a response rate endpoint.

2. Theoretical background of Bayesian predictive probability for binary outcome

The parameter of interest is the number of responses which can be translated into a response rate π. The objective is to predict the final response rate observed when the maximum sample size is reached, based on the initial assumption on the distribution of potential response rates and the actual outcome at a given time point. Within Bayesian settings it requires defining the prior distribution $f(\theta)$, which reflects the primary expectation on the response rate and the likelihood $L(y_1, y_2, \ldots, y_n| \theta)$, which adjusts the initial belief by the empirical evidence. The posterior distribution of the response rate is derived using the Bayesian rule (Lesaffre & Lawson, 2012, p. 23):

$$f(\theta \mid y_1, y_2, \ldots, y_n) = \frac{L(y_1, y_2, \ldots, y_n \mid \theta) f(\theta)}{\int L(y_1, y_2, \ldots, y_n \mid \theta) f(\theta) d\theta}. \tag{1}$$

The response rate can be seen as a probability of an individual success in a Bernoulli experiment $\pi=P(Y=1)$. The likelihood function for a sample of Bernoulli random variables $y_1, y_2, \ldots, y_n$ is given by the formula (Jóźwiak & Podgórski, 2009, pp. 193-4.):

$$\begin{aligned} L(y_1, y_2, \ldots, y_n \mid \pi) &= \pi^{y_1}(1-\pi)^{1-y_1}\pi^{y_2}(1-\pi)^{1-y_2}\ldots\pi^{y_n}(1-\pi)^{1-y_n} \\ &= \pi^{\sum_i y_i}(1-\pi)^{n-\sum_i y_i}. \end{aligned} \tag{2}$$

It follows that the number of responses $R = \sum_{i=1}^{n} y_i$ is a random variable with a binomial distribution defined by the probability (Lesaffre & Lawson, 2012, p. 25):

$$P(R=r \mid \pi)=\binom{n}{r}\pi^{r}(1-\pi)^{n-r} \quad for\ r=1,2,\ldots,n. \tag{3}$$

The initial belief on the response rate can be expressed by the beta distribution, which has the following probability density function (Bolstad, 2007, p. 127):

$$f(\pi \mid a,b)=\begin{cases}\dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1} & for\ \pi\in[0,1],\\ 0 & for\ \pi\notin[0,1]\end{cases} \tag{4}$$

where $\Gamma(s)=\int_0^\infty t^{s-1}e^{-t}dt$, s>0 is the Gamma function, and a>0 and b>0.

The beta distribution associates probability with any value between 0 and 1 which covers all possible response rates. If there is no strong evidence for any particular set of values one can use Beta prior with parameters a=0.5 and b=0.5, which gives similar probability for the values in the central region of π distribution and with somewhat more probable observation of extreme values (see Figure 1). Alternative formulation with a=1 and b=1 would result in uniform distribution assuming the same probability for all the possible values of the parameter of interest.

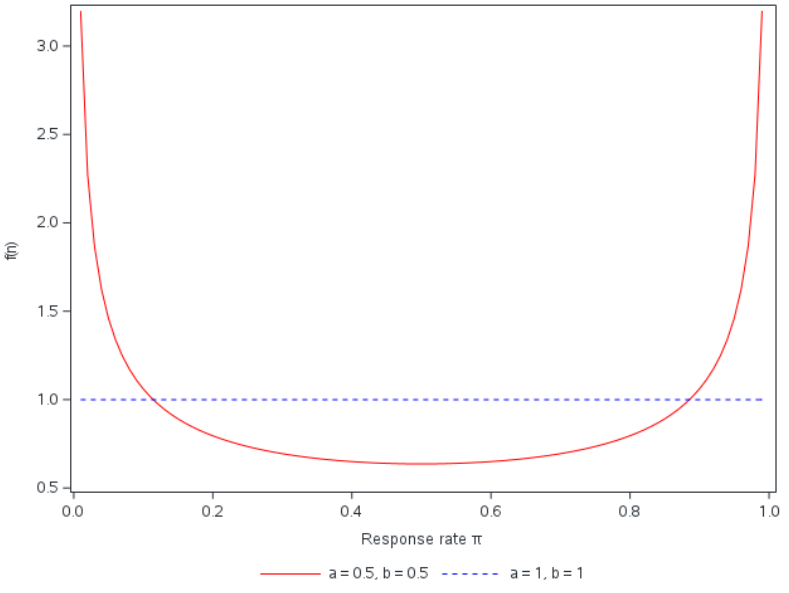

Figure 1: Beta probability density function with parameters a=0.5 and b=0.5

Source: own study.

The posterior distribution of the response rate is derived based on (1) using (2) and (4) (Lesaffre & Lawson, 2012, pp. 24-30)[3]:

$$f(\pi \mid y_1, y_2, \ldots, y_n) = \frac{L(y_1, y_2, \ldots, y_n \mid \pi) f(\pi \mid a, b)}{\int_0^1 L(y_1, y_2, \ldots, y_n \mid \pi) f(\pi \mid a, b) d\pi} = \frac{\Gamma(a+b+n)}{\Gamma(a+r)\Gamma(b+n-r)} \pi^{a+r-1}(1-\pi)^{b+n-r-1}, \tag{5}$$

which is a Beta distribution with parameters $a`=a+r$ and $b`=b+n-r$. The observed number of responses corrects the prior belief about the response rate distribution.

For example, if we observe r=10 responses in a sample of n=50 subjects, the prior beta distribution with parameters a=0.5 and b=0.5, would give the beta posterior with $a`$=10.5 and $b`$=40.5 shown in Figure 2. The empirical response rate equals π=0.2, and we can clearly see that the posterior distribution is heavily concentrated around that value.

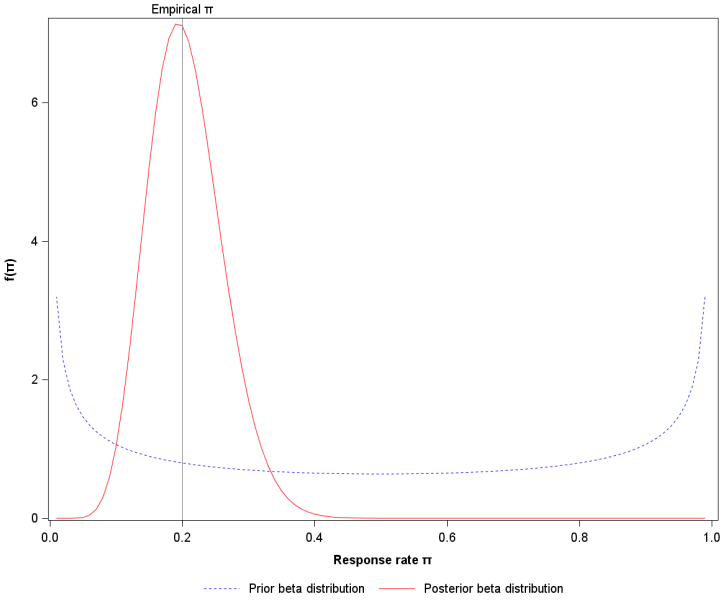

Figure 2: Beta prior distribution with a=0.5 and b=0.5 and posterior with a'=10.5 and b'=40.5

Source: own study.

The properties of Bayesian estimate of the response rate allows to make the judgement on the final outcome based on the actual number of responses at a given time point. We estimate the probability of any future outcome using posterior predictive distribution.

[3] Derivation of the posterior distribution of the response rate is included in the Appendix.

Firstly, let us note that the posterior distribution $f(\theta|\, y_1, y_2, \ldots, y_n)$ is our belief on the probability of potential values of θ given data. The posterior probability mass can be used to assess the future responses. The distribution of the future responses for continuous variable given data is (Lesaffre & Lawson, 2012, p. 53):

$$f(y_{t+1}, y_{t+2}, \ldots, y_n \mid y_1, y_2, \ldots, y_t) = \int f(y_{t+1}, y_{t+2}, \ldots, y_n \mid \theta) f(\theta \mid y_1, y_2, \ldots, y_t) d\theta. \quad (6)$$

In the settings of the response rate, the predictive posterior distribution given r_0 responses in n_0 subjects, defines the probability of observing any possible number of responses in the remaining subjects $m=n-n_0$. The number of responses in future m subjects is random binomial variable (3). Knowing that the posterior distribution of the response rate is given by (5), the predictive distribution for the response rate is:

$$\begin{aligned} P(X = x \mid r_0, n_0, a, b) &= \int_0^1 \binom{m}{x} \pi^x (1-\pi)^{m-x} \frac{\Gamma(a+b+n_0)}{\Gamma(a+r_0)\Gamma(b+n_0-r_0)} \pi^{a+r_0-1}(1-\pi)^{b+n_0-r_0-1} d\pi \\ &= \binom{m}{x} \frac{\Gamma(a+b+n_0)}{\Gamma(a+r_0)\Gamma(b+n_0-r_0)} \frac{\Gamma(a+r_0+x)\Gamma(b+n_0-r_0+m-x)}{\Gamma(a+b+n_0+m)} \quad for\ x = 0,1,\ldots, m. \end{aligned} \quad (7)$$

Let us assume that the maximum sample size in a study equals n=50. At a given timepoint we observed 21 subjects, 10 of which had a response. The current response rate equals π_0=10/21≈0.476. Figure 3 shows the probability of future responses in m remaining subjects. As we can see the mass of the probability concentrates around 14. The final response rate would then most likely be close to π = (10+14)/50=0.48.

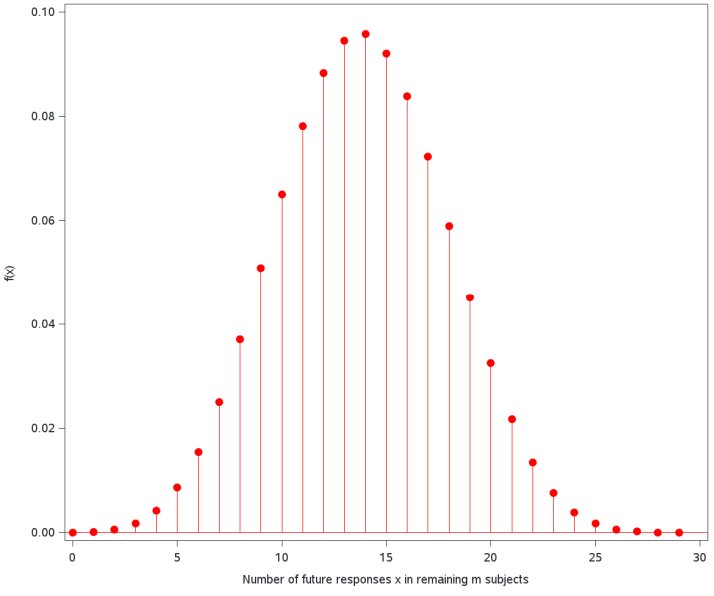

Figure 3: Posterior predictive distribution of responses in the remaining subjects with beta prior with parameters a=0.5 and b=0.5

Source: own study.

The procedure which enables to project the future results based on the current responses is built on the following steps defined in (Lee & Liu, 2008):

1. Find the posterior predictive distribution of future responses. The derivations involve the response rate π_0 computed for the number of responses r_0 and the number of subjects n_0 observed until given time point.
2. Calculate the probabilities p_x x=0,1,…,m of observing any potential number of responses x for the remaining m=n-n_0 subjects to assess how likely any possible future result is.
3. Assess the probability of observing the pre-specified response rate at each number of future responses (for example $\pi > 0.5$). This tells us how likely it is to reach the endpoint given the number of future responses. The individual probability of reaching the response rate is given by the posterior beta distribution (5) at each x.
4. Identify the number of responses for which the probability of reaching the pre-specified response rate is greater than $p_{\min}$ (e.g. $p_{\min} = 0.9$).
5. Sum the probabilities $p_x > p_{\min}$. The sum of these probabilities yields the predictive probability p of observing the pre-specified response rate when the maximum sample size is reached.

If the predictive probability is very low $p<p_L$ or very high $p>p_U$ the decision would be to stop the trial respectively for futility or efficacy. The lower bound p_L represents the probability of an event that is very unlikely whereas p_U represents the probability of an event which is very likely to be observed.

An example choice for the probability bounds is $p_L = 0.1$ and $p_U = 0.9$. In the context of the response rate assessment for $p<p_L$ the chance of reaching the requested response given the current results is very unlikely. On the contrary, when $p>p_U$ it is highly probable that the expected response rate will be reached and therefore we should consider extension of the study and moving into the large sample phase.

Continuing the example of the trial with n=50 subjects and 10 responses in 21 observed patients, we can calculate the probability of reaching the expected response rate π at the end of the trial for each number of future responses x.

Let us assume that the response is expected to occur among more than half of the patients. The probability that $P(\pi > 0.5|r_0, x)$ is calculated using the cumulative distribution function of the posterior beta distribution (5):

$$F(z \mid \mathbf{y}_0) = \begin{cases} 0 & z < 0, \\ \dfrac{\Gamma(a+b+n_0+m)}{\Gamma(a+r_0+x)\Gamma(b+n_0-r_0+m-x)} \displaystyle\int_0^z \pi^{a+r_0+x-1}(1-\pi)^{b+n_0-r_0+m-x-1} d\pi & 0 \le z \le 1, \\ 1 & z > 1. \end{cases} \quad (8)$$

where z represents the probability threshold determined by the expectation for the response rate and $\mathbf{y}_0$ reflects the characteristics of the process observed at time of the interim including r_0 and n_0.

The individual p_x along with the associated probabilities of reaching the expected response rate are shown in Table 1. Given that we observed 21 subjects the future number of responses can take a value of x=0,1,…,29, with the total number of subjects of n=50. The individual probabilities p_x are calculated based on (7) whereas probabilities $P(\pi > 0.5|r_0, x)$ are computed using (8). The table contains also the indicator function which takes the value of 1 for each number of responses for which it is highly likely that the pre-defined response threshold will be attained when the maximum sample size is reached.

Table 1. Posterior probabilities for each number of responses in m remaining subjects

x	p_x	$P(\pi > 0.5\|r_0, x)$	$I(x)$
0	1.816E-05	5.88E-06	0
1	0.00014	2.35E-05	0
2	0.0005853	8.42E-05	0
3	0.0017558	0.000271	0
4	0.0042212	0.00079	0
5	0.0086207	0.002095	0
6	0.0154922	0.005086	0
7	0.0250717	0.011339	0
8	0.0371253	0.023309	0
9	0.0508755	0.044338	0
10	0.0650539	0.078308	0
11	0.0780847	0.128848	0
12	0.088359	0.198193	0
13	0.094538	0.286031	0
14	0.0958121	0.38882	0
15	0.0920547	0.5	0
16	0.0838356	0.61118	0
17	0.0722937	0.713969	0
18	0.058906	0.801807	0
19	0.0452069	0.871152	0
20	0.0325269	0.921692	1
21	0.0218038	0.955662	1
22	0.0135001	0.976691	1
23	0.0076305	0.988661	1
24	0.0038731	0.994914	1
25	0.0017241	0.997905	1
26	0.0006494	0.99921	1
27	0.0001951	0.999729	1
28	4.181E-05	0.999916	1
29	4.826E-06	0.999976	1

Source: own study.

The probability of reaching the expected response rate is assessed given the current number of responses r_0 in n_0 subjects. In order to ensure that the goal of the study was met we would expect at least 20 responses in future patients. However, the probability

of observing these many success cases in the example setting is low. To be precise, the predicted probability of observing more than 50% of responses at the end of the trial equals $p = \sum_x I(x)p_x = 0.082$. If we took the lower bound for the predicted probability of p_L =0.1, then the conclusion would be to stop the study at the current stage for futility as $p < p_L$.

3. Empirical evidence – disease control at 12 weeks in patients with metastatic prostate cancer

The Bayesian predictive probability design has been applied to the data collected in metastatic prostate cancer clinical trial[4]. The aim of the trial was to assess the overall survival of patients with metastatic prostate cancer on standard and experimental combination therapies.

In order to exemplify the application and assess the usefulness of the Bayesian predictive probability design in clinical trial settings the following research problem has been undertaken. The purpose of the study is to find the number of patients with disease control at 12 weeks. The endpoint of the disease control rate at landmark time is based on binary outcome which simplifies the Bayesian settings. This type of endpoint finds application in phase II trials (Simon, Geyer, Subramanian & Roychowdhury, 2016, p. 18).

Disease control is defined as the number of subjects with complete response, partial response or stable disease as specified in the response criteria in the clinical study protocol for the prostate cancer clinical trial (Project Data Sphere, 2008, pp. CSP, p. 49)[5]. The disease control has been assessed at 12 weeks allowing for 2 week time window.

In our example the interim analyses have been carried out every 10 patients, until the maximum sample size has been reached. At each interim analysis the number of patients with disease control has been calculated and compared against the Bayesian predictive probability bounds, set up for the response rate at 12 weeks of 30% (π=0.3). In other words, we want to know what the likelihood of observing at least 30% of disease control patients at 12 weeks within all sampled patients is, given the actual data at time *t*.

The exercise has been carried out based on the n=203 patients with at least one target tumor lesion measurement. Bayesian predictive probability bounds have been computed based on the procedure outlined on page 191 using a Beta prior with parameters a=0.5 and b=0.5 for the response rate π, and the critical values for the

[4] The data are provided by CEO Roundtable on Cancer's Life Sciences Consortium, a free digital library with historical patient level data from cancer clinical trials: https://www.projectdatasphere.org/projectdatasphere/html/about, Accessed March 26, 2017.

[5] For overview of the RECIST criteria for efficacy endpoints in oncology, including disease control rate see for example: (George, Wang, & Pang, 2016, pp. 7-8).

probabilities $p_L = 0.1$ and $p_U = 0.9$. The results are presented in Table 2. In addition, the table provides the number of disease control patients along with the disease control rate and the dates expressing the recruitment process.

In the first row of Table 2 we see that the first interim disease control assessment was carried out for 10 patients in the study recruited within 11 weeks between 3rd Jan 2007 and 20th Mar 2007. Four out of ten patients reached disease control at 12 weeks, which gave disease control rate of 40% at that time point.

If we had observed only one patient with disease control at that time, then we would have had very low probability of reaching the target π=30% disease control rate in the maximum sample size. On the contrary, if we had observed six and more patients with disease control, it would have been a strong indication for efficacy.

A disease control between 2 and 5 would have positioned the investigator in the region where there is no strong indication for either of the two decisions. The indication of the model is therefore to continue the study because the current data do not provide enough statistical evidence about efficacy of the treatment.

From t=3 and n=30, which was 13 months prior to the enrolment of the last subject and until the end of the recruitment, the observed response rate was in the efficacy region. The data for 30 patients provided enough statistical evidence within the Bayesian design to conclude that the disease control rate at 12 weeks for the maximum sample size would be equal to at least 30%. The actual response rate for the maximum sample size was 90/203=44%.

Table 2. Bayesian predictive probability boundaries with actual disease control assessments at 12 weeks and recruitment time for each sample in prostate cancer clinical trial

	Recruitment			Bayesian predictive probability boundaries			Efficacy assessments	
Sample size at time t	Date first subject enrolled in each sample	Date last subject enrolled in each sample	Recruitment time in weeks	Futility region	Continuation region	Efficacy region	Number of disease control subjects	Disease control rate
10	2007-01-03	2007-03-20	11	1	2-5	6-10	**4**	40%
20	2007-03-26	2007-05-25	8	1-4	5-9	10-20	**9**	45%
30	2007-05-31	**2007-06-26**	4	1-7	8-13	14-30	**15**	50%
40	2007-07-02	2007-07-26	3	1-10	11-17	18-40	**18**	45%
50	2007-07-30	2007-09-03	5	1-13	14-20	21-50	**25**	50%
60	2007-09-05	2007-09-28	3	1-16	17-24	25-60	**29**	48%
70	2007-10-01	2007-10-17	2	1-19	20-28	29-70	**31**	44%
80	2007-10-23	2007-11-20	4	1-23	24-31	32-80	**34**	43%
90	2007-11-22	2007-12-13	3	1-26	27-35	36-90	**37**	41%
100	2007-12-14	2008-01-07	4	1-29	30-38	39-100	**44**	44%
110	2008-01-08	2008-02-08	4	1-33	34-41	42-110	**47**	43%

Table 3: Bayesian predictive probability boundaries with actual disease control assessments at 12 weeks and recruitment time for each sample in prostate cancer clinical trial (cont.)

Sample size at time *t*	Recruitment			Bayesian predictive probability boundaries			Efficacy assessments	
	Date first subject enrolled in each sample	Date last subject enrolled in each sample	Recruitment time in weeks	Futility region	Continuation region	Efficacy region	Number of disease control subjects	Disease control rate
120	2008-02-14	2008-03-11	4	1-36	37-45	46-120	**53**	44%
130	2008-03-14	2008-03-28	2	1-40	41-48	49-130	**56**	43%
140	2008-04-04	2008-04-21	3	1-43	44-51	52-140	**61**	44%
150	2008-04-24	2008-05-19	4	1-47	48-55	56-150	**66**	44%
160	2008-05-20	2008-06-11	3	1-51	52-58	59-160	**69**	43%
170	2008-06-11	2008-07-07	4	1-54	55-61	62-170	**76**	45%
180	2008-07-08	2008-08-04	4	1-58	59-64	65-180	**79**	44%
190	2008-08-06	2008-09-15	6	1-62	63-67	68-190	**83**	44%
200	2008-09-16	2008-10-08	3	1-67	68-69	70-200	**88**	44%
203	2008-10-10	**2008-10-17**	1				**90**	44%

Source: own study based on (Project Data Sphere, 2008).

The number of subjects with disease control at the subsequent interim assessments from Table 2 has been illustrated in Figure 4.

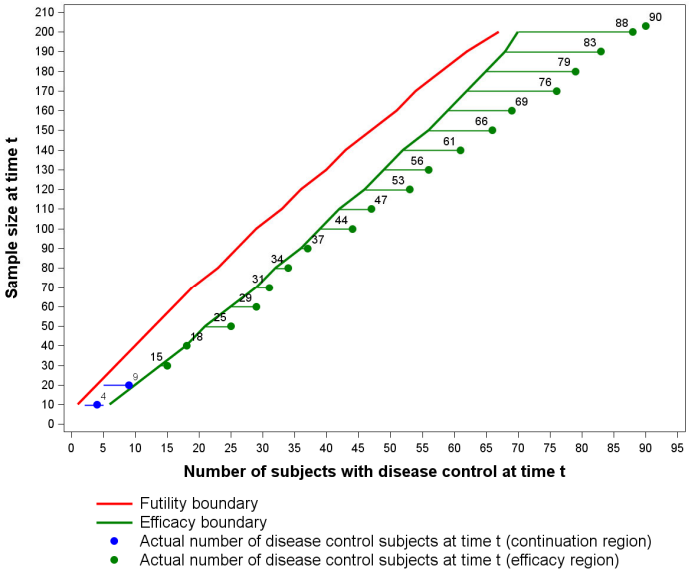

Figure 4: Disease control rate at week 12 in patients with metastatic prostate cancer and the Bayesian predictive boundaries assessed at every 10 patients in the study

Source: own study based on Table 2.

4. Sensitivity analysis – the choice of prior distribution

As noted earlier the Bayesian analysis allows bringing an external expertise into the estimation process. In our example we could look for evidence in the form of response rates from previous studies in metastatic prostate cancer. This evidence would serve as input for defining the parameters of the prior distribution.

In the example in Section 2 the target response rate is at least π=0.3 of disease control patients. In order to assess the sensitivity of the Bayesian design to the selection of prior distribution two extreme scenarios have been applied, additionally to the non-informative prior from the above example.

Figure 5 shows the posterior distribution resulting from prior assumption of poor performance of the treatment (red solid line for a=1 and b=10). As compared to the non-informative scenario (orange solid line for a=0.5 and b=0.5) the pessimistic posterior is shifted to the left, towards the lower response rate. We would then require a lot more evidence in the form of response from the new trial in order to claim efficacy (see Table 4, first section '*Low prior response rate*'). What is more, we would quicker consider the therapy futile, e.g. 3 responses out of 10 would result in futility decision. The opposite can be observed for the reverse selection of the prior (see Figure 6), however it must be noted that this scenario is far more extreme than the previous pessimistic one, and was applied only to provide a full overview over the resulting trial designs.

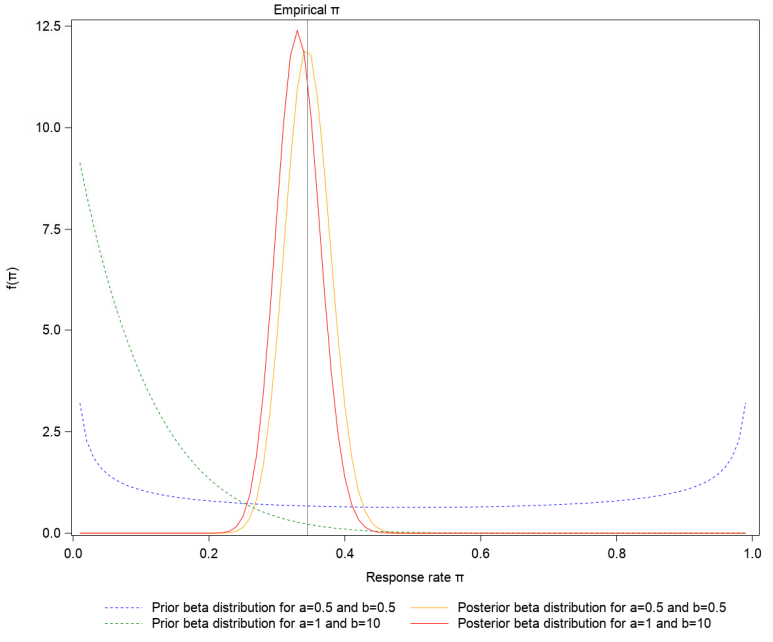

Figure 5: Beta prior distributions with a=0.5 and b=0.5, and a=1 and b=10 with resulting posterior distributions and assumption of empirical response rate of π=70/203≈0.34

Source: own study.

Figure 6 shows the posterior distribution resulting from prior assumption of very good performance of the treatment (red solid line). As noted this is an extreme case in the sense that we kept the target response rate the same at the level of approximately 30%. However, the prior would suggest far higher response rates in the former trials.

The resulting design would drive the efficacy lower boundary to the minimum around slightly above 30% (see Table 4, third column '*High prior response rate*'). We would relatively quickly consider the trial efficacious, e.g. for 3 responses out of 10 first patients.

Considering the above cases, the choice of the prior distribution is highly affecting the decision about the study. This observation is very clear for early stages of the study conduct and for low number of patients. Given the range of possible results the selection of informative prior for low number of patients would require very strong argument. On the other hand, the sensitivity to the choice of the prior distribution is to some extent diminishing for larger number of subjects, i.e. for later stages of the trial. For example, at the time of having 160 patients the lower efficacy bound is 62/160≈0.39 for conservative prior and 54/160≈0.34 for optimistic prior.

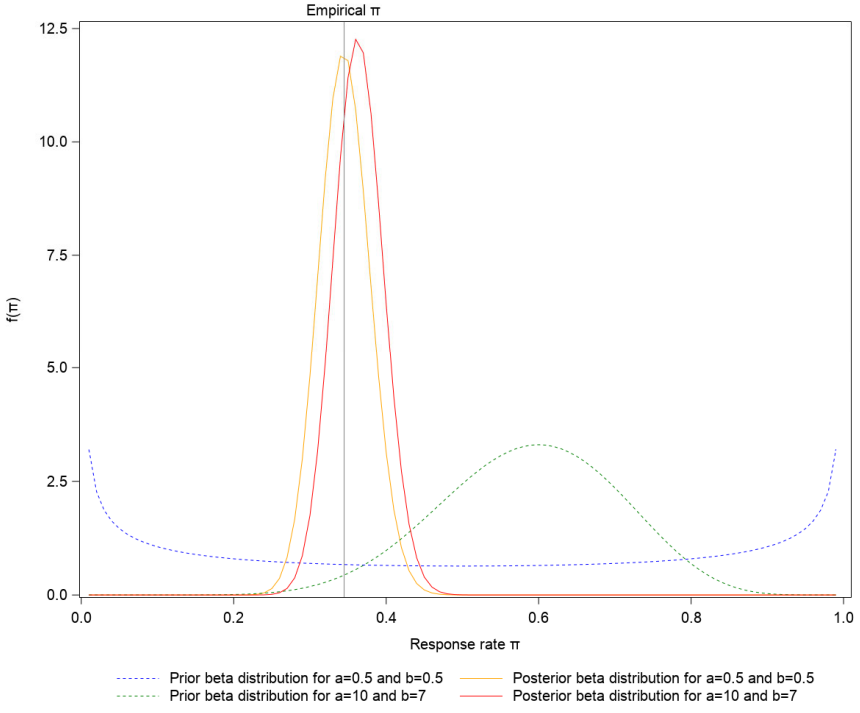

Figure 6: Beta prior distributions with a=0.5 and b=0.5, and a=10 and b=7 with resulting posterior distributions and assumption of empirical response rate of $\pi=70/203\approx0.34$

Source: own study.

Table 4: Bayesian predictive probability boundaries designed for prostate cancer clinical trial according to three types of prior distribution

Sample size at time *t*	Bayesian predictive probability boundaries ***Low prior response rate*** Beta prior with *a*=1 and *b*=10			Bayesian predictive probability boundaries ***Non-informative prior*** Beta prior with *a*=0.5 and *b*=0.5			Bayesian predictive probability boundaries ***High prior response rate*** Beta prior with *a*=10 and *b*=7		
	Futility region	Continuation region	Efficacy region	Futility region	Continuation region	Efficacy region	Futility region	Continuation region	Efficacy region
10	1-3	4-8	9-10	1	2-5	6-10		1-2	3-10
20	1-6	7-12	13-20	1-4	5-9	10-20		1-6	7-20
30	1-9	10-16	17-30	1-7	8-13	14-30	1-2	3-9	10-30
40	1-12	13-20	21-40	1-10	11-17	18-40	1-5	6-13	14-40
50	1-16	17-24	25-50	1-13	14-20	21-50	1-8	9-16	17-50
60	1-19	20-27	28-60	1-16	17-24	25-60	1-11	12-20	21-60
70	1-22	23-31	32-70	1-19	20-28	29-70	1-15	16-23	24-70
80	1-25	26-34	35-80	1-23	24-31	32-80	1-18	19-27	28-80
90	1-29	30-38	39-90	1-26	27-35	36-90	1-21	22-30	31-90
100	1-32	33-41	42-100	1-29	30-38	39-100	1-25	26-34	35-100
110	1-36	37-45	46-110	1-33	34-41	42-110	1-28	29-37	38-110
120	1-39	40-48	49-120	1-36	37-45	46-120	1-32	33-40	41-120
130	1-43	44-51	52-130	1-40	41-48	49-130	1-35	36-44	45-130
140	1-46	47-54	55-140	1-43	44-51	52-140	1-39	40-47	48-140
150	1-50	51-58	59-150	1-47	48-55	56-150	1-42	43-50	51-150
160	1-54	55-61	62-160	1-51	52-58	59-160	1-46	47-53	54-160
170	1-57	58-64	65-170	1-54	55-61	62-170	1-50	51-56	57-170
180	1-61	62-67	68-180	1-58	59-64	65-180	1-53	54-59	60-180
190	1-65	66-70	71-190	1-62	63-67	68-190	1-57	58-62	63-190
200	1-70	71-72	73-200	1-67	68-69	70-200	1-62	63-64	65-200
203									

Source: own study.

An example of the analysis incorporating the prior knowledge from findings in previous trials is provided in (Heath, et al., 2020; Chen, et al., 2019).

5. Summary

The theoretical properties of the Bayesian predictive probability design are appealing as a tool for detecting the treatment signal at earlier stages of studies with continuous recruitment of subjects to the sample. The practical application has shown the usefulness of the approach from the perspective of the timing of the decision. This goes along with the known argument for adaptive design allowing for reducing the overall sample size, cost of the study, drug development time length (George, Wang, & Pang, 2016, p. 367). The final decision would still require larger programs in terms of the sample size. It that sense the Bayesian design has a supportive role.

The results are affected by the level of the expected response rate and therefore the choice of that parameter is crucial for the analysis and conclusions. When discussing the application of Bayesian predictive design, the expected time to response on specific endpoint must also be considered (Zhou, Liu, Kim, Herbst, & Lee, 2008). The expectation is to use a conservative level of the response rate and refer to the findings in other clinical studies with similar indication in order to formulate the response rate that would be clinically beneficial. The other question arose in the analysis is how to define the appropriate frequency of the assessments in order to draw meaningful conclusions in possibly short time period.

From statistical perspective the interest is also in the choice of prior distribution, which affects the expected response rate and, as a consequence, the interim decision. The presented example showed the sensitivity of the prior assumption to the resulting predictive design. A very careful considerations would be required for the choice of informative prior distribution, especially in view of decision making at early stages of the recruitment in the trial. From broader perspective, the prior assumptions are vital for the trial assessment and it is important to consider how robust the design is in translating into phase III trial (Harrington & Parmigiani, 2016, p. 8).

Moreover, there are many practical problems with predictive design related to the conduct of a clinical trial and data collection associated with that process. Firstly, we need to decide on the subjects to be reviewed at each stage. Due to practical problems related to data collection the enrolment date might not be enough to decide about the assignments of the subjects to be included in the interim analysis. There would normally be operational reasons for specifying the enrolment stage at which the interim is planned (Heath, et al., 2020, p. 3). The practical aspects would have to be taken into account within the schedule determined through statistical considerations.

One can argue that by setting the sample size for early review too low could lead to a number of trials inappropriately stopped (Mitchell, 2018, p. 300). Practically, with a small number of patients the investigators would refrain from entering a large sample phase even with very promising results.

Lastly, the Bayesian predictive design requires specification of the criteria for inclusion in the interim analysis. Especially in cancer clinical trials we may expect deviations from the protocol-defined procedures which can bias the final result and therefore need to be carefully considered in the analysis.

References

Lee, J., Liu, D., (2008, 5). A predictive probability design for phase II cancer clinical trials. *Clinical Trials*, pp. 93–106.

Little, J. A., Rubin, D., (2002). *Statistical Analysis with Missing Data.* Hoboken: John Wiley & Sons.

Mittelhammer, R., (2013). *Mathematical Statistics for Economics and Business.* New York: Springer.

Lesaffre, E., Lawson, A., (2012). *Bayesian Biostatistics.* Chichester: John Wiley & Sons.

Jóźwiak, J., Podgórski, J., (2009). *Statystyka od podstaw.* Warszawa: Polskie Wydawnictwo Ekonomiczne.

Bolstad, W., (2007). *Introduction to Bayesian Statistics.* Hoboken: John Wiley & Sons.

Project Data Sphere, (2008, July 21). *A Randomized, Open Label Multi-Center Study of XRP6258 at 25 mg/m^2 in Combination With Prednisone Every 3 Weeks Compared to Mitoxantrone in Combination With Prednisone For The Treatment of Hormone Refractory Metastatic Prostate Cancer*. Retrieved March 7, 2017, from https://www.projectdatasphere.org/projectdatasphere/html/content/79.

George, S., Wang, X., & Pang, H., (2016). *Cancer Clinical Trials. Current and Controversial Issues in Design and Analysis.* Boca Raton: CRC Press.

Heath, A., Offringa, M., Pechlivanoglou, P., J. D., R., Klassen, T., Poonai, N., & Pullenayegum, E., (2020). Determining a Bayesian predictive power stopping rule for futility in a non-inferiority trial with binary outcomes. *Contemporary Clinical Trials Communications, 18*.

Zhou, X., Liu, S., Kim, E., Herbst, R., & Lee, J., (2008). Bayesian adaptive design for targeted therapy development in lung cancer--a step toward personalized medicine. *Clinical Trials, 5*(3), pp. 181–193.

Mitchell, P., (2018). A Bayesian single-arm design using predictive probability monitoring. *Biom Biostat Int J., 7*(4), pp. 299–309.

Barker, A., Sigman, C., Kelloff, G., Hylton, N., Berry, D., & Esserman, L., (2009). I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin Pharmacol Ther., 86*(1), pp. 97–100.

Simon, R., Geyer, S., Subramanian, J., & Roychowdhury, S., (2016). The Bayesian basket design for genomic variant-driven phase II trials. *Semin Oncol., 43*(1), pp. 13–18.

Chen, M., Ibrahim, J. G., Lam, P., Yu, A., & Zhang, Y., (2011). Bayesian Design of Noninferiority Trials for Medical Devices Using Historical Data. *Biometrics, 67*(3), pp. 1163–1170.

Yin, F., Chen, N., & Lee, J., (2012). Phase II trial design with Bayesian adaptive randomization and predictive probability. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 61*(2), pp. 219–235.

Harrington, D., Parmigiani, G., (2016). I-SPY 2 — A Glimpse of the Future of Phase 2 Drug Development? *The New England Journal of Medicine, 375*(1), pp. 7–9.

Chen, D., Schell, M., W. J., F., Pettersson, F., Kim, S., J. E., G., & E.B., H., (2019). Application of Bayesian predictive probability for interim futility analysis in single-arm phase II trial. *Transl Cancer Res., 8*, pp. 404–420.

Appendix A

Posterior distribution of the response rate with Beta distribution as prior:

$$
\begin{aligned}
f(\pi \mid y_1, y_2, \ldots, y_n) &= \frac{L(y_1, y_2, \ldots, y_n \mid \pi) f(\pi \mid a, b)}{\int_0^1 L(y_1, y_2, \ldots, y_n \mid \pi) f(\pi \mid a, b) d\pi} \\
&= \frac{\pi^r (1-\pi)^{n-r} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1} (1-\pi)^{b-1}}{\int_0^1 \pi^r (1-\pi)^{n-r} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1} (1-\pi)^{b-1} d\pi} \\
&= \frac{\pi^{a+r-1} (1-\pi)^{b+n-r-1}}{\int_0^1 \pi^{a+r-1} (1-\pi)^{b+n-r-1} d\pi} \\
&= \left(\frac{\Gamma(a+r)\Gamma(b+n-r)}{\Gamma(a+b+n)} \right)^{-1} \pi^{a+r-1} (1-\pi)^{b+n-r-1} \\
&= \frac{\Gamma(a+b+n)}{\Gamma(a+r)\Gamma(b+n-r)} \pi^{a+r-1} (1-\pi)^{b+n-r-1}.
\end{aligned}
\tag{9}
$$

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 203–219, https://doi.org/10.59170/stattrans-2023-071
Received – 22.03.2022; accepted – 16.06.2023

Another solution for some optimum allocation problem

Wojciech Wójciak[1]

Abstract

We derive optimality conditions for the optimum sample allocation problem in stratified sampling, formulated as the determination of the fixed strata sample sizes that minimize the total cost of the survey, under the assumed level of variance of the stratified π estimator of the population total (or mean) and one-sided upper bounds imposed on sample sizes in strata. In this context, we presume that the variance function is of some generic form that, in particular, covers the case of the simple random sampling without replacement design in strata. The optimality conditions mentioned above will be derived from the Karush-Kuhn-Tucker conditions. Based on the established optimality conditions, we provide a formal proof of the optimality of the existing procedure, termed here as *LRNA*, which solves the allocation problem considered. We formulate the *LRNA* in such a way that it also provides the solution to the classical optimum allocation problem (i.e. minimization of the estimator's variance under a fixed total cost) under one-sided lower bounds imposed on sample sizes in strata. In this context, the *LRNA* can be considered as a counterparty to the popular recursive Neyman allocation procedure that is used to solve the classical problem of an optimum sample allocation with added one-sided upper bounds. Ready-to-use R-implementation of the *LRNA* is available through our `stratallo` package, which is published on the Comprehensive R Archive Network (CRAN) package repository.

Key words: stratified sampling, optimum allocation, minimum cost allocation under upper bounds, optimum allocation constant variance, optimum allocation under lower bounds, recursive Neyman algorithm.

1. Introduction

Let us consider a finite population U consisting of N elements. Let the parameter of principal interest of a single study variable y in U be denoted by θ. This parameter is the population total (i.e. $\theta = \sum_{k\in U} y_k$, where y_k denotes the value of y for population element $k \in U$), or the population mean (i.e. $\theta = \frac{1}{N}\sum_{k\in U} y_k$). To estimate θ, we consider the *stratified π estimator*, i.e. the π *estimator* of Horvitz and Thompson (see, e.g. Särndal, Swensson and Wretman, 1992, Section 2.8, p. 42) in *stratified sampling*. Under this well-known sampling technique, population U is stratified, i.e. $U = \bigcup_{h\in\mathcal{H}} U_h$, where U_h, $h \in \mathcal{H}$, called strata, are pairwise disjoint and non-empty, and $\mathcal{H} = \{1, \ldots, H\}$ denotes a finite set of strata indices of size $H \geq 1$. The size of stratum U_h is denoted N_h, $h \in \mathcal{H}$ and clearly $\sum_{h\in\mathcal{H}} N_h = N$. Probability samples of size $n_h \leq N_h$, $h \in \mathcal{H}$ are selected independently from each stratum according to chosen sampling designs, which are often the same in all strata. The resulting total sample is of size $n = \sum_{h\in\mathcal{H}} n_h \leq N$. It is well know that the *stratified π estimator* $\hat{\theta}$ of θ and its variance $V_{\hat{\theta}}$ are expressed in terms of the first

[1]Warsaw University of Technology, Poland. E-mail: wojciech.wojciak.dokt@pw.edu.pl.
ORCID: https://orcid.org/0000-0002-5042-160X.

and second order inclusion probabilities (see, e.g. Särndal et al. (1992, Result 3.7.1, p. 102) for the case when θ is the population total). In particular, for several important sampling designs

$$V_{\hat{\theta}}(\mathbf{n}) = \sum_{h\in\mathcal{H}} \frac{A_h^2}{n_h} - A_0, \tag{1}$$

where $\mathbf{n} = (n_h,\, h \in \mathcal{H})$ and $A_0,\, A_h > 0,\, h \in \mathcal{H}$ do not depend on $\mathbf{n}$. Among the most basic and common examples that give rise to the variance of the form (1) is the *stratified* π *estimator* of the population total with *simple random sampling without replacement* design in strata. This case yields in (1): $A_h = N_h S_h,\, h \in \mathcal{H}$, and $A_0 = \sum_{h\in\mathcal{H}} N_h S_h^2$, where S_h denotes stratum standard deviation of study variable y (see, e.g. Särndal et al., 1992, equation 3.7.8, p. 103).

The values of the strata sample sizes $n_h,\, h \in \mathcal{H}$, are chosen by the sampler. They may be selected to minimize the variance (1) at the admissible level of the total cost of the survey or to minimize the total cost of the survey subject to a fixed precision (1). The simplest total cost function is of the form:

$$c(\mathbf{n}) = c_0 + \sum_{h\in\mathcal{H}} c_h n_h, \tag{2}$$

where c_0 is a fixed overhead cost and $c_h > 0$ is the cost of surveying one element in stratum $U_h,\, h \in \mathcal{H}$. For further references, see, e.g. Särndal et al. (1992, Section 3.7.3, p. 104) or Cochran (1977, Section 5.5, p. 96). In this paper, we are interested in the latter strategy, i.e. the determination of the sample allocation $\mathbf{n}$ that minimizes total cost (2) under assumed fixed level of the variance (1). We also impose one-sided upper bounds on sample sizes in strata. Such optimization problem can be conveniently written in the language of mathematical optimization as Problem 1.1, in the definition of which we intentionally omit fixed overhead cost c_0 as it has no impact on the optimal solution to this problem.

Problem 1.1. Given a finite set $\mathcal{H} \neq \emptyset$ and numbers $A_0,\, A_h > 0,\, c_h > 0,\, M_h > 0$, such that $M_h \leq N_h,\, h \in \mathcal{H}$, and $V \geq \sum_{h\in\mathcal{H}} \frac{A_h^2}{M_h} - A_0 \geq 0$,

$$\underset{\mathbf{x}\,=\,(x_h,\,h\in\mathcal{H})\,\in\,\mathbb{R}_+^{|\mathcal{H}|}}{\text{minimize}} \quad \sum_{h\in\mathcal{H}} c_h x_h \tag{3}$$

$$\text{subject to} \quad \sum_{h\in\mathcal{H}} \frac{A_h^2}{x_h} - A_0 = V \tag{4}$$

$$x_h \leq M_h, \qquad h \in \mathcal{H}. \tag{5}$$

To emphasize the fact that the optimal solution to Problem 1.1 may not be an integer one, we denote the optimization variable by $\mathbf{x}$, not by $\mathbf{n}$. Non-integer solution can be rounded up in practice with the resulting variance (1) being possibly near V, instead of the exact V. The upper bounds M_h imposed on $x_h,\, h \in \mathcal{H}$, are natural since for instance the allocation with $x_h > N_h$ for some $h \in \mathcal{H}$ is impossible. We assume that $V \geq \sum_{h\in\mathcal{H}} \frac{A_h^2}{M_h} - A_0$, since otherwise, if $V < \sum_{h\in\mathcal{H}} \frac{A_h^2}{M_h} - A_0$, the problem is infeasible. We also note that in the case when $V = \sum_{h\in\mathcal{H}} \frac{A_h^2}{M_h} - A_0$, the solution is trivial, i.e.: $\mathbf{x}^* = (M_h,\, h \in \mathcal{H})$.

It is worth noting that in the definition of Problem 1.1, we require (through (4)) that

the variance defined in (1) is equal to a certain fixed value, denoted as V, and not less than that, whilst, it might seem more favourable at first, to require the variance (1) to be less than or equal to V, especially given the practical context in which Problem 1.1 arises (see Section 2 below). It is easy to see, however, that the objective function (3) and the variance constraint (4) are of such a form that the minimum of (3) is achieved for a value that yields the allowable maximum of function (1), which is V. Thus, regardless of whether the variance constraint is an equality constraint or an inequality constraint, the optimal solution will be the same in both of these cases.

Our approach to the optimum allocation Problem 1.1 will be twofold. First, in Section 3, we make use of the Karush–Kuhn–Tucker conditions (see Appendix B) to establish necessary and sufficient conditions, the so-called optimality conditions, for a solution to slightly reformulated optimization Problem 1.1, defined as a separate Problem 3.1. This task is one of the main objectives of this paper. Optimality conditions, which are often given as closed-form expressions, are fundamental to the analysis and development of effective algorithms for an optimization problem. Namely, algorithms recognize solutions by checking whether they satisfy various optimality conditions and terminate when such conditions hold. This elegant strategy has evident advantages over some alternative ad-hoc approaches, commonly used in survey sampling, which are usually tailored for a specific allocation algorithm being proposed. Next, in Section 4, we precisely define the *LRNA* algorithm which solves Problem 3.1 (and in consequence Problem 1.1) and based on the established optimality conditions we provide the formal proof of its optimality, which is the second main objective of this paper. To complement our work on this subject, we provide user-end function in R (see R Core Team, 2023) that implements the *LRNA*. This function is included in our package `stratallo` (Wójciak, 2023), which is published on the Comprehensive R Archive Network (CRAN) package repository.

2. Motivation

Optimum sample allocation Problem 1.1 is not only a theoretical problem, but it is also an issue of substantial practical importance. Usually, an increase in the number of samples entails greater costs of the data collection process. Thus, it is often demanded that total cost (2) be somehow minimized. On the other hand, the minimization of the cost should not cause significant reduction of the quality of the estimation, which can be measured by the variance (1). Hence, Problem 1.1 arises very naturally in the planning of sample surveys, when it is necessary to obtain an estimator $\hat{\theta}$ with some predetermined precision V that ensures the required level of estimation quality, while keeping the overall cost as small as possible. Problem 1.1 appears also in the context of optimum stratification and sample allocation between subpopulations in Skibicki and Wywiał (2002) or Lednicki and Wieczorkowski (2003). The authors of the latter paper incorporate variance equality constraint into the objective function and then use numerical algorithms (for minimization of a non-linear multivariate function) to find the minimum of the objective function. If the solution found violates any of the inequality constraints, then the objective function is properly adjusted and the algorithm is re-run again. See also a related paper by Wright,

Noble and Bailer (2007), where the allocation under the constraint of the equal precision for estimation of the strata means was considered.

The problem of minimization of the total cost under constraint on stratified estimator's variance is well known in the domain literature. It was probably first formulated by Tore Dalenius in Dalenius (1949, 1953) and later in his Ph.D. thesis Dalenius (1957, Chapter 1.9, p. 19; Chapters 9.4 - 9.5, p. 199). Dalenius formed this allocation problem in the context of multicharacter (i.e. in the presence of several variables under study) stratified sampling (without replacement) and without taking into account upper-bounds constraints (5). He solved his problem with the use of simple geometric methods for the case of two strata and two estimated population means, indicating that the technique that was used is applicable also for the case with any number of strata and any number of variables. Among other resources that are worth mentioning are Yates (1960) and Chatterjee (1968).

Kokan and Khan (1967) considered a multicharacter generalization of Problem 1.1 and proposed a procedure that leads to the solution of this problem. The proof of the optimality of the obtained solution given by the authors is not strictly formal and, similarly to Dalenius' work, is based solely on geometrical methods. The *LRNA* algorithm presented in Section 4 below, can be viewed as a special case of that Kokan-Khan's procedure for a single study variable. It is this method that is generally accepted as the one that solves Problem 1.1 and is described in popular survey sampling textbooks such as, e.g. Särndal et al. (1992, Remark 12.7.1, p. 466) or Cochran (1977, Section 5.8, p. 104). For earlier references, see Hartley and Hocking (1963) or Kokan (1963), who discussed how to use non-linear programming technique to determine allocations in multicharacter generalization of Problem 1.1. More recent references can be made to Bethel (1989), who proposed a closed form expression (in terms of Lagrange multipliers) for a solution to relaxed Problem 1.1 without (5), as well as to Hughes and Rao (1979), who obtained the solution to Problem 1.1 by employing an extension of a result due to Thompson (1962). Eventually, for integer solution to Problem 1.1, we refer to Khan, Ahsan and Jahan (1997).

We would like to note that the form of the transformation (6) that we have chosen to convert Problem 1.1 into a convex optimization Problem 3.1 was not the only possible choice. An alternative transformation is for instance $z_h = \frac{1}{x_h}$, $h \in \mathcal{H}$, which was used by Kokan and Khan (1967) or Hughes and Rao (1979) in their approaches to (somewhat generalized) Problem 1.1. Nevertheless, as it turns out, transformation (6) causes that induced Problem 3.1 gains some interesting interpretation from the point of view of practical application. That is, if one treats z_h as stratum sample size x_h, $h \in \mathcal{H}$, then Problem 3.1 with $\widetilde{V} = n$ and $c_h = 1, h \in \mathcal{H}$, becomes a classical optimum allocation Problem A.1 with added one-sided lower-bounds constraints $z_h \geq m_h > 0$, $h \in \mathcal{H}$. Such allocation problem can be viewed as twinned to Problem A.2 and is itself interesting for practitioners. The lower bounds are necessary, e.g. for estimation of population strata variances S_h^2, $h \in \mathcal{H}$, which in practice are rarely known a priori. If they are to be estimated from the sample, it is required that at least $n_h \geq 2$, $h \in \mathcal{H}$. They also appear when one treats strata as domains and assigns upper bounds for variances of estimators of totals in domains. Such approach was considered, e.g. in Choudhry, Hidiroglou and Rao (2012), where the

additional constraints $(\frac{1}{n_h} - \frac{1}{N_h})N_h^2 S_h^2 \leq R_h$, $h \in \mathcal{H}$, where R_h, $h \in \mathcal{H}$ are given constants, have been imposed. Obviously, this system of inequalities can be rewritten as lower-bounds constraints on n_h, i.e. $n_h \geq m_h = \frac{N_h^2 S_h^2}{R_h + N_h S_h^2}$, $h \in \mathcal{H}$. The solution given in Choudhry et al. (2012) was obtained by the procedure based on the Newton-Raphson algorithm, a general-purpose root-finding numerical method. See also a related paper by Wright et al. (2007), where the allocation under the constraint of the equal precision for estimation of the strata means was considered. The affinity between allocation Problem 3.1 and Problem A.2 translates to significant similarities between the *LRNA* that solves Problem 3.1 and the popular recursive Neyman allocation procedure, *RNA*, that solves Problem A.2 (see Appendix A). To emphasize these similarities, the name *LRNA* was chosen for the former.

In summary, the *LRNA*, formulated as in this work, solves two different but related problems of optimum sample allocation that are of a significant practical importance, i.e. Problem 1.1 and Problem 3.1.

3. Optimality conditions

In this section, we establish a general form of the solution to (somewhat reformulated) Problem 1.1, the so-called optimality conditions. For this problem, the optimality conditions can be derived reliably from the Karush–Kuhn–Tucker (KKT) conditions, first derivative tests for a solution in nonlinear programming to be optimal (see Appendix B and the references given therein for more details). It is well known that for convex optimization problem (with some minor regularity conditions) the KKT conditions are not only necessary but also sufficient. Problem 1.1 is however not a convex optimization problem because the equality constraint function $\sum_{h\in\mathcal{H}} \frac{A_h^2}{x_h} - A_0 - V$ of $\mathbf{x} = (x_h,\, h \in \mathcal{H})$ is not affine and hence, the feasible set might not be convex. Nevertheless, it turns out that Problem 1.1 can be easily reformulated to a convex optimization Problem 3.1, by a simple change of its optimization variable from $\mathbf{x}$ to $\mathbf{z} = (z_h,\, \in \mathcal{H})$ with elements of the form:

$$z_h := \frac{A_h^2}{c_h x_h}, \qquad h \in \mathcal{H}. \tag{6}$$

Problem 3.1. Given a finite set $\mathcal{H} \neq \emptyset$ and numbers $A_h > 0$, $c_h > 0$, $m_h > 0$, $h \in \mathcal{H}$, $\widetilde{V} \geq \sum_{h\in\mathcal{H}} c_h m_h$,

$$\underset{\mathbf{z}\,=\,(z_h,\,h\in\mathcal{H})\,\in\,\mathbb{R}_+^{|\mathcal{H}|}}{\text{minimize}} \quad \sum_{h\in\mathcal{H}} \frac{A_h^2}{z_h} \tag{7}$$

$$\text{subject to} \quad \sum_{h\in\mathcal{H}} c_h z_h = \widetilde{V} \tag{8}$$

$$z_h \geq m_h, \qquad h \in \mathcal{H}. \tag{9}$$

For Problem 3.1 to be equivalent to Problem 1.1 under transformation (6), parameters

m_h, $h \in \mathcal{H}$, and $\widetilde{V}$ must be such that

$$\begin{aligned} m_h &:= \tfrac{A_h^2}{c_h M_h}, \qquad h \in \mathcal{H}, \\ \widetilde{V} &:= V + A_0 \geq A_0, \end{aligned} \tag{10}$$

where numbers V, A_0, M_h, $h \in \mathcal{H}$ are as in Problem 1.1. Nonetheless, as we explained at the end of Section 2 of this paper, Problem 3.1 can be considered as a separate allocation problem, unrelated to Problem 1.1; that is Problem A.1 with added one-sided lower-bounds constraints. For this reason, the only requirements imposed on these parameters are those given in the definition of Problem 3.1.

The auxiliary optimization Problem 3.1 is indeed a convex optimization problem as it is justified by Remark 3.1.

Remark 3.1. *Problem 3.1 is a convex optimization problem as its objective function* $f : \mathbb{R}_+^{|\mathcal{H}|} \to \mathbb{R}_+$,

$$f(\mathbf{z}) = \sum_{h \in \mathcal{H}} \tfrac{A_h^2}{z_h}, \tag{11}$$

and inequality constraint functions $g_h : \mathbb{R}_+^{|\mathcal{H}|} \to \mathbb{R}$,

$$g_h(\mathbf{z}) = m_h - z_h, \qquad h \in \mathcal{H}, \tag{12}$$

are convex functions, while the equality constraint function $w : \mathbb{R}_+^{|\mathcal{H}|} \to \mathbb{R}$,

$$w(\mathbf{z}) = \sum_{h \in \mathcal{H}} c_h z_h - \widetilde{V}$$

is affine. More specifically, Problem 3.1 is a convex optimization problem of a particular type in which inequality constraint functions (12) *are affine. See Appendix B for the definition of the convex optimization problem.*

As we shall see in Theorem 3.1, the optimization Problem 3.1 has a unique optimal solution. Consequently, due to transformation (6) and given (10), vector

$$\mathbf{x}^* = \left(\tfrac{A_h^2}{c_h z_h^*},\, h \in \mathcal{H} \right) \tag{13}$$

is a unique optimal solution of Problem 1.1, where $\mathbf{z}^* = (z_h^*,\, h \in \mathcal{H})$ is a solution to Problem 3.1. For this reason, for the remaining part of this work, our focus will be on a solution to Problem 3.1. We also note here that the solution to Problem 3.1 is trivial in the case of $\widetilde{V} = \sum_{h \in \mathcal{H}} c_h m_h$, i.e.: $\mathbf{z}^* = (m_h,\, h \in \mathcal{H})$.

Before we establish necessary and sufficient optimality conditions for a solution to convex optimization Problem 3.1, we first define a set function s, which considerably simplifies notation and many calculations that are carried out in this and subsequent section.

Definition 3.1. Let $\mathcal{H},\ A_h,\ c_h,\ m_h,\ h \in \mathcal{H}$, and $\widetilde{V}$ be as in Problem 3.1. Set function s is defined as:

$$s(\mathcal{L}) = \frac{\widetilde{V} - \sum_{h\in\mathcal{L}} c_h m_h}{\sum_{h\in\mathcal{H}\setminus\mathcal{L}} A_h\sqrt{c_h}}, \qquad \mathcal{L} \subsetneq \mathcal{H}. \tag{14}$$

Below, we will introduce the notation of vector $\mathbf{z}^{\mathcal{L}} = (z_h^{\mathcal{L}},\ h \in \mathcal{H})$. It turns out that the solution to Problem 3.1 is necessarily of the form (15) with the set $\mathcal{L} \subseteq \mathcal{H}$ defined implicitly through the inequality of a certain form given in Theorem 3.1.

Definition 3.2. Let $\mathcal{H},\ A_h,\ c_h,\ m_h,\ h \in \mathcal{H},\ \widetilde{V}$ be as in Problem 3.1 and let $\mathcal{L} \subseteq \mathcal{H}$. Vector $\mathbf{z}^{\mathcal{L}} = (z_h^{\mathcal{L}},\ h \in \mathcal{H})$ is defined as follows

$$z_h^{\mathcal{L}} = \begin{cases} m_h, & h \in \mathcal{L} \\ \frac{A_h}{\sqrt{c_h}}\, s(\mathcal{L}) & h \in \mathcal{H}\setminus\mathcal{L}. \end{cases} \tag{15}$$

The following Theorem 3.1 characterizes the form of the optimal solution to Problem 3.1 and therefore is the key theorem of this paper.

Theorem 3.1 (Optimality conditions)**.** *The optimization Problem 3.1 has a unique optimal solution. Point* $\mathbf{z}^* = (z_h^*,\ h \in \mathcal{H}) \in \mathbb{R}_+^{|\mathcal{H}|}$ *is a solution to optimization Problem 3.1 if and only if* $\mathbf{z}^* = \mathbf{z}^{\mathcal{L}^*}$ *with* $\mathcal{L}^* \subseteq \mathcal{H}$*, such that one of the following two cases holds:*

CASE I: $\mathcal{L}^* \subsetneq \mathcal{H}$ *and*

$$\mathcal{L}^* = \left\{ h \in \mathcal{H} :\ s(\mathcal{L}^*) \le \tfrac{\sqrt{c_h} m_h}{A_h} \right\}, \tag{16}$$

where set function s *is defined in* (14)*.*

CASE II: $\mathcal{L}^* = \mathcal{H}$ *and*

$$\widetilde{V} = \sum_{h\in\mathcal{H}} c_h m_h. \tag{17}$$

Proof. We first prove that the solution to Problem 3.1 exists and it is unique. In the optimization Problem 3.1, a feasible set $F := \{\mathbf{z} \in \mathbb{R}_+^{|\mathcal{H}|} :\ (8) \text{ and } (9) \text{ are satisfied}\}$ is non-empty as guaranteed by the requirement $\widetilde{V} \ge \sum_{h\in\mathcal{H}} c_h m_h$. The objective function in (7) attains its minimum on F since it is a continuous function on F and F is closed and bounded. Finally, the uniqueness of the solution is due to strict convexity of the objective function on the set F.

As mentioned at the beginning of Section 3, the form of the solution to Problem 3.1, can be derived from the KKT conditions (see Appendix B). Following the notation of Remark 3.1, gradients of the objective function f and constraint functions $w,\ g_h,\ h \in \mathcal{H}$, are as follows:

$$\nabla f(\mathbf{z}) = \left(-\tfrac{A_h^2}{z_h^2},\ h \in \mathcal{H} \right), \quad \nabla w(\mathbf{z}) = (c_h,\ h \in \mathcal{H}), \quad \nabla g_h(\mathbf{z}) = -\underline{1}_h, \quad \mathbf{z} \in \mathbb{R}_+^{|\mathcal{H}|},$$

where $\underline{1}_h$ is a vector with all entries 0 except the entry at index h, which is 1. Consequently, the KKT conditions (33) assume the following form for the optimization Problem 3.1:

$$-\frac{A_h^2}{z_h^{*2}} + \lambda c_h - \mu_h = 0, \qquad h \in \mathcal{H}, \tag{18}$$

$$\sum_{h\in\mathcal{H}} c_h z_h^* - \widetilde{V} = 0, \tag{19}$$

$$m_h - z_h^* \leq 0, \qquad h \in \mathcal{H}, \tag{20}$$

$$\mu_h(m_h - z_h^*) = 0, \qquad h \in \mathcal{H}. \tag{21}$$

Following Theorem B.1 and Remark 3.1, in order to prove Theorem 3.1, it suffices to show that there exist $\lambda \in \mathbb{R}$ and $\mu_h \geq 0$, $h \in \mathcal{H}$, such that (18) - (21) are met for $\mathbf{z}^* = \mathbf{z}^{\mathcal{L}^*}$ with $\mathcal{L}^* \subseteq \mathcal{H}$ satisfying conditions of CASE I or CASE II.

CASE I: Following (15) and (14), we get

$$\sum_{h\in\mathcal{H}} c_h z_h^* = \sum_{h\in\mathcal{L}^*} c_h m_h + \sum_{h\in\mathcal{H}\setminus\mathcal{L}^*} c_h \frac{A_h}{\sqrt{c_h}}\, s(\mathcal{L}^*) = \widetilde{V},$$

and hence, the condition (19) is always satisfied. Let $\lambda = \frac{1}{s^2(\mathcal{L}^*)}$, where $s(\mathcal{L}^*) > 0$ is defined in (14), and

$$\mu_h = \begin{cases} \lambda c_h - \frac{A_h^2}{m_h^2}, & h \in \mathcal{L}^* \\ 0, & h \in \mathcal{H} \setminus \mathcal{L}^*. \end{cases} \tag{22}$$

Note that $\mu_h \geq 0$, $h \in \mathcal{L}^*$, due to (16). Then, the condition (18) is clearly satisfied. Inequalities (20) and equalities (21) are trivial for $h \in \mathcal{L}^*$ since $z_h^* = m_h$. For $h \in \mathcal{H} \setminus \mathcal{L}^*$, inequalities (20) follow from (16), i.e. $\frac{A_h}{\sqrt{c_h}}\, s(\mathcal{L}^*) > m_h$, whilst (21) hold true due to $\mu_h = 0$.

CASE II: Take arbitrary $\lambda \geq \max_{h\in\mathcal{H}} \frac{A_h^2}{m_h^2 c_h}$ and $\mu_h = \lambda c_h - \frac{A_h^2}{m_h^2}$, $h \in \mathcal{H}$. Note that $\mu_h \geq 0$, $h \in \mathcal{H}$. Then, (18) - (21) are clearly satisfied for $(z_h^*,\, h \in \mathcal{H}) = (m_h,\, h \in \mathcal{H})$, whilst (19) follows after referring to (17).

□

Theorem 3.1 gives the general form of the optimum solution up to specification of the set $\mathcal{L}^* \subseteq \mathcal{H}$ that corresponds to the optimal solution $\mathbf{z}^* = \mathbf{z}^{\mathcal{L}^*}$. The issue of how to identify this set is the subject of the next section of this paper.

4. Recursive Neyman algorithm under lower-bounds constraints

In this section, we formalize the definition of the existing algorithm, termed here *LRNA*, solving Problem 3.1 and provide a formal proof of its optimality. The proof given is based on the optimality conditions formulated in Theorem 3.1.

Algorithm *LRNA*

Input: $\mathcal{H}$, $(A_h)_{h\in\mathcal{H}}$, $(c_h)_{h\in\mathcal{H}}$, $(m_h)_{h\in\mathcal{H}}$, $\widetilde{V}$.

Require: $A_h > 0$, $c_h > 0$, $m_h > 0$, $h \in \mathcal{H}$, $\widetilde{V} \geq \sum_{h\in\mathcal{H}} c_h m_h$.

Step 1: Let $\mathcal{L} = \emptyset$.

Step 2: Determine $\widetilde{\mathcal{L}} = \left\{h \in \mathcal{H} \setminus \mathcal{L} : \frac{A_h}{\sqrt{c_h}} s(\mathcal{L}) \leq m_h\right\}$, where function s is defined in (14).

Step 3: If $\widetilde{\mathcal{L}} = \emptyset$, go to Step 4 . Otherwise, update $\mathcal{L} \leftarrow \mathcal{L} \cup \widetilde{\mathcal{L}}$ and go to Step 2 .

Step 4: Return $\mathbf{z}^* = (z_h^*,\, h \in \mathcal{H})$ with $z_h^* = \begin{cases} m_h, & h \in \mathcal{L} \\ \frac{A_h}{\sqrt{c_h}}\, s(\mathcal{L}), & h \in \mathcal{H} \setminus \mathcal{L}. \end{cases}$

Theorem 4.1. *The* LRNA *provides an optimal solution to optimization Problem 3.1.*

Before we prove Theorem 4.1, we first reveal certain monotonicity property of set function s, defined in (14), that will be essential to the proof of this theorem.

Lemma 4.2. *Let $\mathcal{A} \subseteq \mathcal{B} \subsetneq \mathcal{H}$. Then*

$$s(\mathcal{A}) \geq s(\mathcal{B}) \quad \Leftrightarrow \quad s(\mathcal{A}) \sum_{h\in\mathcal{B}\setminus\mathcal{A}} A_h\sqrt{c_h} \leq \sum_{h\in\mathcal{B}\setminus\mathcal{A}} c_h m_h, \tag{23}$$

where set function s is defined in (14)*.*

Proof. Clearly, for any $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$, $\delta \in \mathbb{R}$, $\gamma \in \mathbb{R}_+$, such that $\gamma + \delta > 0$, we have

$$\tfrac{\alpha+\beta}{\gamma+\delta} \geq \tfrac{\alpha}{\gamma} \quad \Leftrightarrow \quad \tfrac{\alpha+\beta}{\gamma+\delta}\delta \leq \beta. \tag{24}$$

To prove (23), take

$$\alpha = \widetilde{V} - \sum_{h\in\mathcal{B}} c_h m_h \qquad \beta = \sum_{h\in\mathcal{B}\setminus\mathcal{A}} c_h m_h$$

$$\gamma = \sum_{h\in\mathcal{H}\setminus\mathcal{B}} A_h\sqrt{c_h} \qquad \delta = \sum_{h\in\mathcal{B}\setminus\mathcal{A}} A_h\sqrt{c_h}.$$

Then, $\frac{\alpha}{\gamma} = s(\mathcal{B})$, $\frac{\alpha+\beta}{\gamma+\delta} = s(\mathcal{A})$, and hence (23) holds as an immediate consequence of (24). □

We are now ready to give the proof of Theorem 4.1.

Proof of Theorem 4.1. Let $\mathcal{L}_r$, $\widetilde{\mathcal{L}}_r$ denote sets $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ respectively, as in the r-th iteration of the *LRNA* algorithm, at the moment after Step 2 and before Step 3. The iteration index r takes on values from set $\{1, \ldots, r^*\}$, where $r^* \geq 1$ indicates the final iteration of the algorithm. Under this notation, we have $\mathcal{L}_1 = \emptyset$ and in general for subsequent iterations, if any (i.e. if $r^* \geq 2$), we get

$$\mathcal{L}_r = \mathcal{L}_{r-1} \cup \widetilde{\mathcal{L}}_{r-1} = \bigcup_{i=1}^{r-1} \widetilde{\mathcal{L}}_i, \qquad r = 2, \ldots, r^*. \tag{25}$$

To prove Theorem 4.1, we have to show that:

(I) the algorithm terminates in a finite number of iterations, i.e. $r^* < \infty$,

(II) the solution computed at r^* is optimal.

The proof of (I) is relatively straightforward. In every iteration $r = 2, \ldots, r^* \geq 2$, the domain of discourse for $\widetilde{\mathcal{L}}_r$ at Step 2 is $\mathcal{H} \setminus \mathcal{L}_r = \mathcal{H} \setminus \bigcup_{i=1}^{r-1} \widetilde{\mathcal{L}}_i$, where $\widetilde{\mathcal{L}}_i \neq \emptyset$, $i = 1, \ldots, r-1$. Therefore, in view of Step 3, we have that $r^* \leq |\mathcal{H}| + 1 < \infty$, where $r^* = |\mathcal{H}| + 1$ if and only if $|\widetilde{\mathcal{L}}_r| = 1$ for each $r = 1, \ldots, r^* - 1$. In words, the algorithm terminates in at most $|\mathcal{H}| + 1$ iterations.

In order to prove (II), following Theorem 3.1, it suffices to show that for $\mathcal{L}_{r^*} \subsetneq \mathcal{H}$ (CASE I), for all $h \in \mathcal{H}$,

$$h \in \mathcal{L}_{r^*} \quad \Leftrightarrow \quad \tfrac{A_h}{\sqrt{c_h}}\, s(\mathcal{L}_{r^*}) \leq m_h, \tag{26}$$

and for $\mathcal{L}_{r^*} = \mathcal{H}$ (CASE II):

$$\widetilde{V} = \sum_{h \in \mathcal{H}} c_h m_h. \tag{27}$$

We first note that the construction of the algorithm ensures that $\mathcal{L}_r \subsetneq \mathcal{H}$ for $r = 1, \ldots, r^* - 1$, $r^* \geq 2$, and therefore $s(\mathcal{L}_r)$ for such r is well-defined.

CASE I. The $s(\mathcal{L}_{r^*})$ is well-defined since in this case $\mathcal{L}_{r^*} \subsetneq \mathcal{H}$.

Necessity: For $r^* = 1$, we have $\mathcal{L}_{r^*} = \emptyset$ and hence, the right-hand side of equivalence (26) is trivially met. Let $r^* \geq 2$. By Step 2 of the *LRNA*, we have

$$\tfrac{A_h}{\sqrt{c_h}}\, s(\mathcal{L}_r) \leq m_h, \qquad h \in \widetilde{\mathcal{L}}_r, \tag{28}$$

for every $r = 1, \ldots, r^* - 1$. Multiplying inequalities (28) sidewise by c_h and summing over $h \in \widetilde{\mathcal{L}}_r$, we get the right-hand side of equivalence (23) with $\mathcal{A} = \mathcal{L}_r$ and $\mathcal{B} = \mathcal{L}_r \cup \widetilde{\mathcal{L}}_r = \mathcal{L}_{r+1} \subsetneq \mathcal{H}$. Then, by Lemma 4.2, the first inequality in (23) follows. Consequently,

$$s(\mathcal{L}_1) \geq \ldots \geq s(\mathcal{L}_{r^*}). \tag{29}$$

Now, assume that $h \in \mathcal{L}_{r^*} = \bigcup_{r=1}^{r^*-1} \widetilde{\mathcal{L}}_r$. Thus, $h \in \widetilde{\mathcal{L}}_r$ for some $r \in \{1, \ldots, r^* - 1\}$, and again, using Step 2 of the *LRNA*, we get $\frac{A_h}{\sqrt{c_h}}\, s(\mathcal{L}_r) \leq m_h$. Consequently, (29) yields $\frac{A_h}{\sqrt{c_h}}\, s(\mathcal{L}_{r^*}) \leq m_h$.

Sufficiency: The proof is by establishing a contradiction. Assume that $\frac{A_h}{\sqrt{c_h}}\, s(\mathcal{L}_{r^*}) \leq m_h$ and $h \notin \mathcal{L}_{r^*}$. On the other hand, Step 3 of the *LRNA* yields $\frac{A_h}{\sqrt{c_h}}\, s(\mathcal{L}_{r^*}) > m_h$ for $h \in \mathcal{H} \setminus \mathcal{L}_{r^*}$ (i.e. $h \notin \mathcal{L}_{r^*}$), which contradicts the assumption.

CASE II. Note that in this case it must be that $r^* \geq 2$. Following Step 2 of the *LRNA*, the only possibility is that for all $h \in \mathcal{H} \setminus \mathcal{L}_{r^*-1}$,

$$\frac{A_h}{\sqrt{c_h}}\, s(\mathcal{L}_{r^*-1}) = \frac{A_h}{\sqrt{c_h}}\, \frac{\widetilde{V} - \sum_{i\in\mathcal{L}_{r^*-1}} c_i m_i}{\sum_{i\in\mathcal{H}\setminus\mathcal{L}_{r^*-1}} A_i\sqrt{c_i}} \leq m_h. \tag{30}$$

Multiplying both sides of inequality (30) by c_h, summing it sidewise over $h \in \mathcal{H} \setminus \mathcal{L}_{r^*-1}$, we get $\widetilde{V} \leq \sum_{i\in\mathcal{H}} c_i m_i$, which, when combined with the requirement $\widetilde{V} \geq \sum_{h\in\mathcal{H}} c_h m_h$, yields $\widetilde{V} = \sum_{h\in\mathcal{H}} c_h m_h$, i.e. (27).

□

Following Theorem 4.1 and given (13), the optimal solution to Problem 1.1 is vector $\mathbf{x}^* = (x_h^*,\, h \in \mathcal{H})$ with elements of the form:

$$x_h^* = \begin{cases} M_h, & h \in \mathcal{L}^* \\ \frac{A_h}{\sqrt{c_h}}\, \frac{\sum_{i\in\mathcal{H}\setminus\mathcal{L}^*} A_i\sqrt{c_i}}{V + A_0 - \sum_{i\in\mathcal{L}^*} \frac{A_i^2}{M_i}}, & h \in \mathcal{H} \setminus \mathcal{L}^*, \end{cases} \tag{31}$$

where $\mathcal{L}^* \subseteq \mathcal{H}$ is determined by the *LRNA*.

5. Final remarks and conclusions

Within this work we formulated the optimality conditions for an important problem of minimum cost allocation under constraints on stratified estimator's variance and maximum samples sizes in strata. This allocation problem was defined in this paper as Problem 1.1 and converted to Problem 3.1 through transformation (6) and under (10). Based on the established optimality conditions, we provided a formal and compact proof of the optimality of the *LRNA* algorithm that solves the allocation problem mentioned. As already outlined at the end of Section 2 of this paper, Problem 3.1 can be viewed in two ways, each of which is of a great practical importance. That is, apart from its primary interpretation as the problem of minimizing the total cost under given constraints, it can also be perceived as the problem of minimizing the stratified estimator's variance under constraint on total sample size (i.e. Problem A.1) and constraints imposed on minimum sample sizes in the strata. For this reason, all the results of this work established in relation to Problem 3.1 (i.e. optimality conditions and the *LRNA*) are of such a twofold nature.

For the reasons mentioned at the end of Section 2, the *LRNA* can be considered as a counterparty to the *RNA*. This resemblance is particularly desirable, given the popularity, simplicity as well as relatively high computational efficiency of the latter algorithm. Among the alternative approaches that could potentially be adapted to solve Problem 3.1 are the ideas that underlay the existing algorithms dedicated to Problem A.2, i.e.: *SGA* (Stenger and Gabler, 2005, Wesołowski, Wieczorkowski and Wójciak, 2021) and *COMA* (Wesołowski et al., 2021). For integer-valued algorithms dedicated to Problem 3.1 with added upper-bounds constraints, see Friedrich, Münnich, de Vries and Wagner (2015), Wright (2017,

2020). Nevertheless, it should be noted that integer-valued algorithms are typically relatively slow compared to not-necessarily integer-valued algorithms. As pointed out in Friedrich et al. (2015), computational efficiency of integer-valued allocation algorithms becomes an issue for cases with "many strata or when the optimal allocation has to be applied repeatedly, such as in iterative solutions of stratification problems".

Finally, we would like to emphasize that the optimality conditions established in Theorem 3.1 can be used as a baseline for the development of new algorithms that provide solution to the optimum allocation problem considered in this paper. For instance, such algorithms could be derived by exploiting the ideas embodied in *SGA* or *COMA*, dedicated to Problem A.2 as indicated above.

Theoretical results obtained in this paper are complemented by the R-implementation (R Core Team, 2023) of the *LRNA*, which we include in our publicly available package `stratallo` (Wójciak, 2023).

Acknowledgements

I am very grateful to Jacek Wesołowski from Warsaw University of Technology, my research supervisor, for his patient guidance on this research work. Many thanks to Robert Wieczorkowski from Statistics Poland for advice and explanations on the topic of optimum stratification. I would also like to thank Reviewers for taking the necessary time and effort to review the manuscript. In particular, I express my gratitude to the second of the Reviewers for his expertise, valuable suggestions and for pointing to the existing papers, particularly important from the point of view of the subject I am addressing in this work.

References

Bethel, J., (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15(1), pp. 47–57. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X198900114578

Boyd, S., and Vandenberghe, L., (2004). *Convex Optimization*, Cambridge University Press, Cambridge.

Chatterjee, S., (1968). Multivariate Stratified Surveys. *Journal of the American Statistical Association*, 63(322), pp. 530–534. https://doi.org/10.2307/2284023

Choudhry, G. H., Hidiroglou, M. and Rao, J., (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, 38(1), pp. 23–29. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X201200111682

Cochran, W. G., (1977). *Sampling Techniques*, 3rd edn, John Wiley & Sons, New York.

Dalenius, T., (1949). Den nyare utvecklingen inom teorin och metodiken för stickprovsundersökningar. *Förhandlingar vid Nordiska Statistikermötet i Helsingfors den 13 och 14 juni 1949*, Helsingfors, pp. 46–74.

Dalenius, T., (1953). The multivariate sampling problem. *Skandinavisk Aktuarietidskrift*, 36, pp. 92–102.

Dalenius, T., (1957). *Sampling in Sweden: Contributions to the Methods and Theories of Sample Survey Practice*, Almqvist & Wiksell, Stockholm.

Friedrich, U., Münnich, R., de Vries, S. and Wagner, M., (2015). Fast integer-valued algorithms for optimal allocations under constraints in stratified sampling. *Computational Statistics & Data Analysis*, 92, pp. 1–12. https://www.sciencedirect.com/science/article/pii/S0167947315001413

Hartley, H. O. and Hocking, R. R., (1963). Convex Programming by Tangential Approximation. *Management Science*, 9(4), pp. 600–612. https://doi.org/10.1287/mnsc.9.4.600

Hughes, E., and Rao, J. N. K., (1979). Some problems of optimal allocation in sample surveys involving inequality constraints. *Communications in Statistics-theory and Methods*, 8, pp. 1551–1574.

Khan, M. G. M., Ahsan, M. J. and Jahan, N., (1997). Compromise allocation in multivariate stratified sampling: An integer solution. *Naval Research Logistics (NRL)*, 44(1), pp. 69–79. https://doi.org/10.1002/(SICI)1520-6750(199702)44:1<69::AID-NAV4>3.0.CO;2-K

Kokan, A. R., (1963). Optimum Allocation in Multivariate Surveys. *Journal of the Royal Statistical Society. Series A (General)*, 126(4), pp. 557–565. https://doi.org/10.2307/2982579

Kokan, A. R. and Khan, S., (1967). Optimum Allocation in Multivariate Surveys: An Analytical Solution, *Journal of the Royal Statistical Society. Series B (Methodological)*, 29(1), pp. 115–125. https://doi.org/10.1111/j.2517-6161.1967.tb00679.x

Lednicki, B. and Wieczorkowski, R., (2003). Optimal Stratification and Sample Allocation between Subpopulations and Strata. *Statistics in Transition*, 6(2), pp. 287–305. https://stat.gov.pl/download/gfx/portalinformacyjny/en/defaultstronaopisowa/3432/1/1/sit_volume_4-7.zip

Neyman, J., (1934). On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4), pp. 558–625.

R Core Team, (2023). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Skibicki, M. and Wywiał, J., (2002). On optimal sample allocation in strata, *in* J. Paradysz (ed.). *Statystyka regionalna w służbie samorządu lokalnego i biznesu*, Internetowa Oficyna Wydawnicza Centrum Statystyki Regionalnej, Poznań, pp. 29–37.

Stenger, H. and Gabler, S., (2005). Combining random sampling and census strategies - Justification of inclusion probabilities equal to 1. *Metrika*, 61(2), pp. 137–156. https://doi.org/10.1007/s001840400328

Särndal, C.-E., Swensson, B. and Wretman, J., (1992). *Model Assisted Survey Sampling*, Springer, New York.

Tschuprow, A. A., (1923a). On the mathematical expectation of the moments of frequency distributions in the case of correlated observation (Chapters 1-3), *Metron*, 2(3), pp. 461–493.

Tschuprow, A. A., (1923b). On the mathematical expectation of the moments of frequency distributions in the case of correlated observation (Chapters 4-6), *Metron*, 2(4), pp. 636–680.

Thompson, W. A., (1962). The problem of negative estimates of variance components. *Annals of Mathematical Statistics*, 33, pp. 273–289.

Wesołowski, J., Wieczorkowski, R. and Wójciak, W., (2021). Optimality of the Recursive Neyman Allocation. *Journal of Survey Statistics and Methodology*, 10(5), pp. 1263–1275. https://academic.oup.com/jssam/article-pdf/10/5/1263/46878255/smab018.pdf

Wright, S. E., Noble, R. and Bailer, A. J., (2007). Equal-precision allocations and other constraints in stratified random sampling. *Journal of Statistical Computation and Simulation*, 77(12), pp. 1081–1089. https://doi.org/10.1080/10629360600897191

Wright, T., (2017). Exact optimal sample allocation: More efficient than Neyman. *Statistics & Probability Letters*, 129, pp. 50–57. https://www.sciencedirect.com/science/article/pii/S0167715217301657

Wright, T., (2020). A general exact optimal sample allocation algorithm: With bounded cost and bounded sample sizes. *Statistics & Probability Letters*, 165, pp. 108829. https://www.sciencedirect.com/science/article/pii/S0167715220301322

Wójciak, W., (2023). stratallo*: Optimum Sample Allocation in Stratified Sampling*. R package version 2.2.0. https://CRAN.R-project.org/package=stratallo

Yates, F., (1960). *Sampling Methods for Census and Surveys*, Griffin and Company, Ltd., London.

APPENDICES

A. Recursive Neyman allocation

The classical problem of optimum sample allocation is described e.g. in Särndal et al. (1992, Section 3.7.3, p. 104). It can be formulated in the language of mathematical optimization as Problem A.1.

Problem A.1. Given a finite set $\mathcal{H} \neq \emptyset$ and numbers $A_h > 0$, $h \in \mathcal{H}$, $0 < n \leq N$,

$$\begin{aligned} \underset{\mathbf{x}=(x_h,\, h\in\mathcal{H})\,\in\,\mathbb{R}_+^{|\mathcal{H}|}}{\text{minimize}} \quad & \sum_{h\in\mathcal{H}} \frac{A_h^2}{x_h} \\ \text{subject to} \quad & \sum_{h\in\mathcal{H}} x_h = n. \end{aligned}$$

The solution to Problem A.1 is $\mathbf{x}^* = (x_h^*,\, h \in \mathcal{H})$ with elements of the form:

$$x_h^* = A_h \frac{n}{\sum_{i\in\mathcal{H}} A_i}, \qquad h \in \mathcal{H}.$$

It was established by Tschuprow (1923a,b) and Neyman (1934) for *stratified* π *estimator* of the population total with *simple random sampling without replacement* design in strata, in the case of which $A_h = N_h S_h$, $h \in \mathcal{H}$, where S_h denotes stratum standard deviation of a given study variable. See, e.g. Särndal et al. (1992, Section 3.7.4.i., p. 106) for more details.

The recursive Neyman allocation algorithm, denoted here as *RNA*, is a well-established allocation procedure that finds a solution to the classical optimum sample allocation Problem A.1 with added one-sided upper-bounds constraints, defined here as Problem A.2.

Problem A.2. Given a finite set $\mathcal{H} \neq \emptyset$ and numbers $A_h > 0$, $M_h > 0$, such that $M_h \leq N_h$, $h \in \mathcal{H}$, and $0 < n \leq \sum_{h\in\mathcal{H}} M_h$,

$$\begin{aligned} \underset{\mathbf{x}=(x_h,\, h\in\mathcal{H})\,\in\,\mathbb{R}_+^{|\mathcal{H}|}}{\text{minimize}} \quad & \sum_{h\in\mathcal{H}} \frac{A_h^2}{x_h} \\ \text{subject to} \quad & \sum_{h\in\mathcal{H}} x_h = n \\ & x_h \leq M_h, \qquad h \in \mathcal{H}. \end{aligned}$$

Algorithm *RNA*

Input: $\mathcal{H}$, $(A_h)_{h\in\mathcal{H}}$, $(M_h)_{h\in\mathcal{H}}$, n.
Require: $A_h > 0$, $M_h > 0$, $h \in \mathcal{H}$, $0 < n \leq \sum_{h\in\mathcal{H}} M_h$.
Step 1: Let $\mathcal{U} = \emptyset$.
Step 2: Determine $\widetilde{\mathcal{U}} = \left\{h \in \mathcal{H} \setminus \mathcal{U} : A_h \frac{n-\sum_{i\in\mathcal{U}} M_i}{\sum_{i\in\mathcal{H}\setminus\mathcal{U}} A_i} \geq M_h\right\}$.
Step 3: If $\widetilde{\mathcal{U}} = \emptyset$, go to Step 4 . Otherwise, update $\mathcal{U} \leftarrow \mathcal{U} \cup \widetilde{\mathcal{U}}$, and go to Step 2 .
Step 4: Return $\mathbf{x}^* = (x_h^*,\, h \in \mathcal{H})$ with $x_h^* = \begin{cases} M_h, & h \in \mathcal{U} \\ A_h \frac{n-\sum_{i\in\mathcal{U}} M_i}{\sum_{i\in\mathcal{H}\setminus\mathcal{U}} A_i}, & h \in \mathcal{H}\setminus\mathcal{U}. \end{cases}$

For more information on this recursive procedure see Särndal et al. (1992, Remark 12.7.1, p. 466) and Wesołowski et al. (2021) for the proof of its optimality.

B. Convex optimization scheme and the KKT conditions

A convex optimization problem is an optimization problem in which the objective function is a convex function and the feasible set is a convex set. In standard form it is written as

$$\begin{aligned} \underset{\mathbf{x}\in\mathcal{D}}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & w_i(\mathbf{x}) = 0, \quad i = 1,\dots,k \\ & g_j(\mathbf{x}) \leq 0, \quad j = 1,\dots,\ell, \end{aligned} \tag{32}$$

where $\mathcal{D} \subseteq \mathbb{R}^p$, $p \in \mathbb{N}_+$, the objective function $f : \mathcal{D}_f \subseteq \mathbb{R}^p \to \mathbb{R}$ and inequality constraint functions $g_j : \mathcal{D}_{g_j} \subseteq \mathbb{R}^p \to \mathbb{R}$, $j = 1,\dots,\ell$, are convex, whilst equality constraint functions $w_i : \mathcal{D}_{w_i} \subseteq \mathbb{R}^p \to \mathbb{R}$, $i = 1,\dots,k$, are affine. Here, $\mathcal{D} = \mathcal{D}_f \cap \bigcap_{i=1}^k \mathcal{D}_{w_i} \cap \bigcap_{j=1}^\ell \mathcal{D}_{g_j}$ denotes a common domain of all the functions. Point $\mathbf{x} \in \mathcal{D}$ is called *feasible* if it satisfies all of the constraints, otherwise the point is called *infeasible*. An optimization problem is called *feasible* if there exists $\mathbf{x} \in \mathcal{D}$ that is *feasible*, otherwise the problem is called *infeasible*.

In the context of the optimum allocation Problem 3.1 discussed in this paper, we are interested in a particular type of the convex problem, i.e. (32) in which all inequality constraint functions g_j, $j = 1,\dots,\ell$, are affine. It is well known, see, e.g. the monograph Boyd and Vandenberghe (2004), that the solution for such an optimization problem can be identified through the set of equations and inequalities known as the Karush-Kuhn-Tucker (KKT) conditions, which in this case are not only necessary but also sufficient.

Theorem B.1 (KKT conditions for convex optimization problem with affine inequality constraints)**.** *A point* $\mathbf{x}^* \in \mathcal{D} \subseteq \mathbb{R}^p$ *is a solution to the convex optimization problem* (32)

in which functions g_j, $j = 1, \ldots, \ell$, *are affine if and only if there exist numbers* $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, k$, *and* $\mu_j \geq 0$, $j = 1, \ldots, \ell$, *called KKT multipliers, such that*

$$\begin{aligned} \nabla f(\mathbf{x}^*) + \sum_{i=1}^{k} \lambda_i \nabla w_i(\mathbf{x}^*) + \sum_{j=1}^{\ell} \mu_j \nabla g_j(\mathbf{x}^*) &= \mathbf{0} \\ w_i(\mathbf{x}^*) &= 0, \qquad i = 1, \ldots, k \\ g_j(\mathbf{x}^*) &\leq 0, \qquad j = 1, \ldots, \ell \\ \mu_j g_j(\mathbf{x}^*) &= 0, \qquad j = 1, \ldots, \ell. \end{aligned} \tag{33}$$

STATISTICS IN TRANSITION new series, December 2023
Vol. 23, No. 5, pp. 221–226

About the Authors

Adri Ahmed holds a PhD in Industrial Engineering from the Laboratory of Mechanical Productivity and Industrial Engineering, Computer-Integrated Manufacturing (LMPGI) at the High School of Technology and the Higher National School of Electricity and Mechanical, University Hassan II of Casablanca. He is also an Engineer in Design and Mechanical Manufacturing from the Engineers Mohammadia School in Rabat. Currently, he serves as a Professor of Higher Education at the High School of Technology, University Hassan II in Casablanca, and is a member of the Mechanical, Manufacturing, and Industrial Engineering Research Laboratory (LMPGI).

Bandyopadhyay Arnab is working as an Assistant Professor of the Department of Basic Science and Humanities (Mathematics section), Asansol Engineering College, Asansol–713305, West Bengal, India. He has more than fifteen years of teaching experience as an Assistant Professor in Mathematics. He has obtained his MSc and PhD in Applied Mathematics from the Indian Institute of Technology (Indian School of Mines) Dhanbad. His research field is Statistical Sample Surveys, Data Analysis and Statistical Inference. Dr. Bandyopadhyay is also a member of editorial board/ reviewers for many international journals. He has published 40 papers in international journals of repute and 4 international books on sampling theory.

Baral Manish Mohan is working as an Assistant Professor in the Department of Operations, GITAM School of Business, GITAM (Deemed to be University), Visakhapatnam, India. He is an engineering graduate from KIIT University, Bhubaneswar, Odisha, MBA in International Business from GITAM University, Visakhapatnam and pursued his PhD in Management from Birla Institute of Technology Mesra, Ranchi. He has published in International Journal of Logistics Management, Benchmarking: An International Journal, Management Decision, and Decision and other reputed journals and several Scopus indexed book chapters. His research areas include information technology, cloud computing, supply chain management, artificial intelligence, operations research, and quality management.

Ben Ali Mohamed holds a PhD in Industrial Engineering and is an Associate Researcher at the Mechanical Productivity and Industrial Engineering, Computer-Integrated Manufacturing Laboratory of the High School of Technology, University Hassan II of Casablanca. Formerly, he served as a Professor in the Bachelor's program in 'Quality Management and Productivity' at the Faculty of Science and Techniques of Tangier. His research interests encompass modelling, entrepreneurship, productivity,

quality management, safety and environmental practices, and logistics. Currently, he is a Professor at the High School of Technology, University Hassan II in Casablanca.

Beresovsky Vladislav is a Research Mathematical Statistician working for the Office of Survey Methods Research at the U.S. Bureau of Labor Statistics. Prior to that, he worked as a Mathematical Statistician at the Centers for Disease Control, National Center for Health Statistics. His research is focused on estimation from data collected by surveys with complex design, adjustment for survey nonresponse, estimation from nonprobability data sources and small area estimation. His work has been presented at multiple conferences and published in journals and conference proceedings.

Bhattacharjee Subarna received her MSc and PhD degree in Mathematics from Indian Institute of Technology, Kharagpur, India. Presently, she is a faculty member in the Department of Mathematics, Ravenshaw University, Cuttack, India. Her research areas include stochastic orders and aging classes.

Chittipaka Venkataiah is working as Associate Professor (Operations Management) at Indira Gandhi National Open University, Delhi, India. He has worked as a professor in the area of Operations, Quality and Project Management in GITAM Institute of Management, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India for 11 years. He is an engineering graduate and completed MBA from National Institute of Technology, Warangal and obtained his doctorate from the Department of Business Management, Osmania University and qualified in UGC-NET conducted by University Grants Commission, New Delhi. His area of interest includes operations, quality, marketing research, logistics, supply chain management and project management.

Das Pitambar is working as an Assistant Professor of the Department of Mathematics, Netaji Nagar Day College, Kolkata–700092, West Bengal, India. He has more than thirteen years of teaching experience as Assistant Professor in Mathematics. He has obtained his MSc, MPhil and PhD in Applied Mathematics from the Indian Institute of Technology (Indian School of Mines), Dhanbad. His research field is Statistical Sample Surveys and Statistical Inference. He has published 6 papers on sampling theory in international journals.

Gershunskaya Julie is a mathematical statistician with the Statistical Methods Staff of the Office of Employment and Unemployment Statistics at the U.S. Bureau of Labor Statistics. Her main areas of interest include statistical data integration, small area estimation, and treatment of influential observations, with application to the U.S. Current Employment Statistics Program.

Giri Rajib Lochan completed his MSc in Mathematics from Utkal University, Bhubaneswar. He did MPhil and PhD in Mathematics from Ravenshaw University, Cuttack, Odisha, India. Presently, he is working as an Assistant Professor

in Mathematics at the Sri Sri University, Cuttack. His research interests include reliability theory.

Khoshkhoo Zahra Amiri holds a PhD in Statistics at the Department of Statistics, University of Mazandaran, Iran. Her research interests are statistical inference, Bayesian estimation, records, prediction. Furthermore, she is committed to transferring statistical knowledge to students and enthusiasts in a clear and practical manner so that they can effectively apply statistics in their research. With educational background and relevant work experiences, she has robust research, teaching, and advisory skills in the field of statistics.

Klochko Rostyslav is a PhD Student at the Department of Mathematical Modeling and Statistics, Kyiv National Economic University named after Vadym Hetman. His research interests are the application of mathematical methods and models, including machine learning methods, to solve problems in marketing management. Rostyslav Klochko pays special attention to approaches of modelling the bank's customer behaviour. He is an active participant in scientific conferences and author of articles in both national and international scientific bodies.

Kończak Grzegorz is a Full Professor at the Department Statistics, Econometrics and Mathematics in University of Economics in Katowice, Poland. His research interests are statistical inference, data analysis, Monte Carlo study and permutation tests in particular. Professor Kończak has published more than 100 research papers in international/national journals and conferences. He has also published eleven books/monographs as an author or co-author. Professor Kończak is an active member of Polish Statistical Association.

Korczyński Adam is an Assistant Professor at the Institute of Statistics and Demography, Collegium of Economic Analyses, Warsaw School of Economics. He specializes in statistical modelling, in particular regression analysis including longitudinal data models, data quality aspects such as missing data and imputation techniques. He obtained a PhD degree in Economics in 2017 based on the thesis entitled: "Variance screen as a tool for detecting price collusion. Fundamentals and meaning of the imputation for missing data".

MirMostafaee S. M. T. K. is an Associate Professor at the Department of Statistics, University of Mazandaran, Babolsar, Iran. His research interests are distribution theory, Bayesian inference, ordered data, record data, reliability and prediction. Dr. MirMostafaee has published more than 50 research papers in international/national journals and conferences.

Mukherjee Subhodeep is working as an Assistant Professor in the Department of Operations, GITAM School of Business, GITAM (Deemed to be University), Visakhapatnam, India, Visakhapatnam, India. He obtained his Master's degree from

the Birla Institute of Technology, Mesra, Ranchi, India. He has publications in reputed journals and highly indexed book chapters. He has presented several papers in various conferences and has also received two best paper and best paper presented awards. His main research interests include food supply chain management, retail supply chain, cloud computing, artificial intelligent, and blockchain technologies. He has expertise in statistical techniques like SEM, etc.

Oullada Oumaima is a PhD candidate in industrial engineering at the Laboratoire de Productivité Mécanique et Génie Industriel, Fabrication Intégrée par Ordinateur (LMPGI) of the Ecole Supérieure de Technologie and the Ecole Nationale Supérieure d'Electricité et de Mécanique, Université Hassan II, Casablanca, Morocco.

Pal Surya Kant is currently working as an Assistant Professor in the Department of Mathematics, Sharda School of Basic Sciences and Research, Sharda University. He holds MPhil and PhD in Statistics from Vikram University Ujjain. He has worked with institution like BFIT-Groups of Institution, University of Petroleum and Energy Studies and Chandigarh University. His research area includes statistical techniques, sampling theory, and multivariate analysis. He has published more than 90 research papers in journals of national and international repute along with two book chapters. He is not only known in India but famous abroad too, serving as an editor and reviewer of more than 30 reputed international/national journals.

Piskunova Olena is a Professor at the Department of Mathematical Modeling and Statistics at Vadym Hetman Kyiv National Economic University. Professor Piskunova is a Doctor of Economics. She is a member of the editorial board of the periodical scientific journal "Modeling and Information Systems in Economics". Professor Piskunova has published 180 scientific works, including monographs, research articles, and conference papers. Her research interests include multivariate statistical analysis, econometric modelling methods, and machine learning.

Rifai Said, before his retirement, served as a Professor and Head of the Mechanical and Industrial Engineering Department at the High School of Technology in Casablanca, University Hassan II, Morocco. He earned his Ph.D. in Industrial Acoustics in 1988 from the University of Poitiers, France. He is a member of the Laboratory of Mechanical Engineering, Computer-Integrated Manufacturing, and Industrial Engineering (LMPGI), certified in Quality Management and Improvement from the High School of Technology and College Bois de Boulogne in Montreal, Canada. He is an expert in engineering and cross-disciplinary training, with a focus on quality, safety, environment, and logistics.

Roychowdhury Anasuya is an Associate Professor at the School of Basic Sciences of IIT Bhubaneswar, India. Her main areas of interest include oncogenic mechanisms of

gastrointestinal cancers, and cancer biomarkers. She has been a grant reviewer at Narodowe Centrum Nauk (National Science Centre), Poland.

Satya Kumar Misra has completed his MSc, MPhil degree from Utkal University and completed Ph.D. in 2014, from KIIT University, Bhubaneswar, India. Currently, he is working as an Associate Professor in the Mathematics at Department of Mathematics, KIIT University, Odisha. His area of research is reliability theory.

Savitsky Terrance is a Senior Research Mathematical Statistician at the U.S. Bureau of Labor Statistics. He obtained his Ph.D. in Statistics from Rice University. Dr. Savitsky's main research areas of interest include Bayesian modelling for survey data, including respondent and domain level data, nonparametric models, methods to construct privatizing mechanisms for survey data, and statistical data integration.

Shyam Hari Shankar is an eminent educationist, researcher and trainer in the areas of management education. He obtained his doctorate in the field of management and has done his Master's degree in Management in the area of Marketing. He has done a certificate course in SPSS & Cognos conducted by IBM & a Certificate program in Marketing by the marketing magnate Philip Kotler. He has a blend of industry academia experience of approximately 14 years. He is an expert in marketing planning & implementation, business analytics and strategy. His other areas of interest are brand, CRM, quality management & OB.

Singh Garib Nath is working as a Professor of the Department of Mathematics and Computing, Indian Institute of Technology (Indian School of Mines), Dhanbad – 826004, Jharkhand, India. He has more than 30 years of teaching experience in statistics. His research field is sample surveys, missing data analysis and data analysis. Professor Singh has published more than 200 papers in national and international journals of repute on sampling theory. He has worked as a PhD thesis examiner and paper setter of many other universities, such as BHU, Utkal University, BIT Mesra, IGNOU and Agra University.

Stąpor Katarzyna is a Full Professor at the Department of Automatic Control, Electronics and Computer Science in the Silesian Technical University in Gliwice, Poland. Her research interests are statistical pattern recognition, multivariate statistical analysis, protein bioinformatics, computational neuroscience based on EEG analysis in particular. Professor Stapor has published more than 120 research papers in international/national journals and conferences including one monograph on pattern recognition and the coursebook for a lecture on statistical methods. Professor Stapor is an active member of several scientific professional bodies.

Szymkowiak Magdalena received her PhD degree in Mathematics from Adam Mickiewicz University in Poznan and was an Associate Professor at the Institute of Mathematics at the Poznan University of Technology. She prepared her habilitation

in the discipline of Automation, Electronic and Electrical Engineering, and now she is an Associate Professor at the Institute of Automatic Control and Robotics at Poznan University of Technology. Her research concerns data mining, reliability theory, and stochastic orders.

Williams Matthew is a Senior Research Statistician at RTI International. Dr. Williams collaborates on both methodological research for and practical implementation of a variety of statistical topics: complex survey sample design and analysis, time-series and time-to-event data, combining multiple data sources, disclosure avoidance, optimization, Bayesian modelling and computation, and evaluation and evidence building. Prior to RTI, he worked with several US federal statistical and science agencies. He also collaborated on international projects related to agriculture and public health. He obtained a PhD in Statistics from Virginia Tech.

Wójciak Wojciech is a PhD student at the Faculty of Mathematics and Information Science, Warsaw University of Technology. He holds a Bachelor of Science in Engineering degree and Master of Science in Mathematics degree, both obtained from the Warsaw University of Technology. His main areas of interest include: optimal sample allocation in complex sampling, convex optimization, applied statistics. He has over 15 years of experience working in various technical and managerial roles in international companies in telecommunication, pharmaceutical and financial industries.

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 227–234

Acknowledgements to Reviewers

The Editor and Editorial Board of Statistics in Transition new series wish to thank the following persons who served as peer-reviewers of manuscripts for the *Statistics in Transition new series* – Volume 24, Numbers 1–5. The authors` work has benefited from their feedback.

Abdullah, Norli Anida A., Centre For Foundation Studies in Science, University of Malaya, Malesia

Adebisi, Ogunde, Department of Statistics, University of Ibadan, Nigeria

AlakuŞ, Kamil, Department of Statistics, Mayis University, Turkey

Aleksandrova, Petrova Simeonka, Department of Commerce, Tsenov Academy of Economics, Bulgaria

Al-Mofleh, Hazem, Department of Mathematics, Tafila Technical University, Jordan

Al-Qati, Ahmed B. J., Department of Mathematics, University of Thi-Qar, Iraq

Amin, Saqib, Department of Economics, University of Oulu, Finland

Andrzejczak, Karol, Departemnt of Mathematics, Poznan University of Technology, Poland

Ayhan, H. Öztaş, Department of Statistics, METU, Turkey

Barrijal, Said, Department of Statistics, Abdelmalek Essaadi University, Morocco

Bayoud, Housam A., Department of Mathematics, Fahad Ben Sultan University, Saudi Arabia

Bąk, Andrzej, Department of Econometrics and Computer Science, Wroclaw University of Economics and Business, Poland

Bdair, Omar, Department of Applied Sciences, Al Balqa Applied University, Amman, Jordan

Białek, Jacek, Department of Statistical Methods, University of Lodz, Poland

Bongiorno, Enea G., Department of Economics and Business Studies, University of Eastern Piedmont, Italy

Boubaker, Mechab, Department of Mathematics, University of Sidi-Bel-Abbes, Algeria

Bourazas, Kostas, KIOS Research and Innovation Center of Excellence, Nicosia, Cyprus

Bouza, Carlos, Department of Mathematics, University of Havana, Cuba

Brown, James J., Department of Statistics, University of Technology Sydney, Australia

Brzezińska, Justyna, Department of Statistics, University of Economics in Katowice, Poland

Cerqueti, Roy, Department of Social and Economic Sciences, Sapienza University of Rome, Italy

Chaturvedi, Anoop, Department of Statistics, University of Allahabad, India

Cheng, Hu Yu, Department of Applied Mathematics, Chung Yuan Christian University Taiwan

Civelek, Mustafa E., Department of Economics and International Trade, İstanbul Ticaret Üniversitesi, Turkey

Cyrek, Magdalena, Department of Economics and Finance, University of Rzeszow, Poland

Dehnel, Grażyna, Department of Statistics, Poznań University of Economics and Business, Poland

Domański, Czesław, Department of Statistical Methods, University of Lodz, Poland

Dubey, Vyas, Department of Statistics, Ravi Shankar Shukla University, India

Fesenko, Valeriia, Department of Accounting, University of Customs and Finance in Dnipro, Ukraine

Ganczarek-Gamrot, Alicja, Department of Demography and Economic Statistics, University of Economics in Katowice, Poland

Ganushchak-Efimenko, Lyudmyla, Department of Computer Engineering, Kyiv National University of Technologies and Design, Ukraine

Ghosh, Koushik, Department of Mathematics, University Of Burdwan, India

Gil-Alaña, Luis Alberiko, Department of Economy, University of Navarra, Spain

Kalton, Graham, USA

Kan, Kılınç Betül, Department of Statistics, Eskişehir Technical University, Turkey

Keser, İstem K., Deparment of Econometrics, Dokuz Eylul University, Turkey

Khan, Abdul N., Department of Community Medicine, University, Medical College, India

Khan, Bhola, Department of Economics, Yobe State University, Nigeria

Klein, Ingo, Department of Statistics and Econometrics, Friedrich Alexander University Erlangen-Nuremberg (FAU), Germany

Klotzke, Konrad, Department of Statistics, University of Twente, The Netherlands

Kordoš, Marcel, Department of Public Administration and Regional Economics, Alexander Dubček University in Trenčín, Slovakia

Kotlebová, Eva, Department of Statistics, University of Economics in Bratislava, Slovakia

Kott, Philipp, National Institue of Statistical Science, USA

Kostrzewska, Jadwiga, Department of Statistics, Krakow University of Economics, Poland

Kovtun, Natalia, Department of Statistics, Taras Shevchenko National University of Kyiv, Ukraine

Kowalczyk, Barbara, Department of Econometrics, Warsaw School of Economics (SGH), Poland

Krzyśko, Mirosław, Professor Emeritus, Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznan, Poland

Ladiray, Dominique, Institut National de la Statistique et des Études Économiques (INSEE), France

Lahiri, Partha, Department of Mathematics, University of Maryland, USA

Lunina, Inna, Department of Public Finances, National Academy of Sciences of Ukraine, Ukraine

Lupu, Radu, Institute for Economic Forecasting, Romanian Academy, Romania

Machiorlatti, Michael, Department of Population Health and Biostatistics, University of Texas, USA

Mahali Kamel, Faculty of Economics, Ferhat Abbas University Setif, Algeria

Majerowska, Ewa, Department of Econometrics, University of Gdansk, Poland

Osińska, Magdalena, Department of Economics, Nicolaus Copernicus University, Torun, Poland

Pawłyszyn, Irena, Institute of Logistics, Poznan University of Technology, Poland

Permpoonsinsup, Wachirapond, Faculty of Science and Technology, Pathumwan Institute of Technology, Thailand

Pietrzyk, Radosław, Department of Financial Investments and Risk Management, Wroclaw University of Economics and Business

Pipień, Mateusz, Department of Empirical Analyses of Economic Stability, Cracow University of Economics, Poland

Pitoňáková, Renáta, Faculty of Social and Economic Sciences, Comenius University in Bratislava, Slovakia

Przybysz, Dariusz, Polish Academy of Sciences, Poland

Rao, Prasada, School of Economics, University of Queensland, Australia

Rachwał, Tomasz, Department of Foreign Trade, Krakow University of Economics, Poland

Rębilas, Rafał, Department of Economics, WSB University, Poland

Riashchenko, Viktoriia, Department of Statistics, University of Applied Sciences (ISMA), Latvia

Rohr, Margarita, Department of Applied Economics, University of Valencia, Spain

Roths, Scott, Department of Statistics, Penn State University, USA

Rozkrut, Dominik, President of Statistics Poland, Poland

Sadova, Ulyana, Departement of Statistics, Dolishniy Institute of Regional Research of the National Academy of Sciences of Ukraine, Ukraine

Safari, Hadi K., Department of Mathematical Sciences, Stevens Institute of Technology, USA

Salamaga, Mariusz, Department of Statistics, Krakow University of Economics, Poland

Samb, Gane, Department of Statistics, Gaston Berger University, Senegal

Sarioglo, Volodymyr, Ptoukha Institute for Demography and Social Studies of the National Academy of Science of Ukraine, Ukraine

Wale-Orojo, O. A., Department of Statistics, College of Physical Sciences, Federal University of Agriculture, Nigeria

Walesiak, Marek, Department of Econometrics and Informatics, Wroclaw University of Economics and Business, Poland

Wanat, Stanisław, Department of Mathematics, Krakow University of Economics, Poland

Wesołowski, Jacek, Department of Probability and Stochastic Processes, Warsaw University of Technology, Poland

Wibowo, Wahyu, Department of Statistics, Institut Teknologi Sepuluh Nopember, Indonesia

Wieczorkowski, Robert, Statistics Poland, Poland

Wolniak, Radosław, Department of Economics and Computer Science, Silesian University of Technology, Poland

Wolny-Dominiak, Alicja, Department of Statistical and Mathematical Methods in Economics, University of Economics in Katowice, Poland

Wołyński, Waldemar, Department of Mathematical Statistics and Data Analysis, Collegium Mathematicum, Adam Mickiewicz University in Poznan, Poland

Wosiek, Małgorzata, Department of Economics and Finance, University of Rzeszow, Poland

Wyłomańska, Agnieszka, Faculty of Pure and Applied Mathematics, Wroclaw University of Science and Technology, Poland

Wywiał, Janusz, Department of Statistics, Econometrics and Mathematics, University of Economics in Katowice, Poland

Yadav, Rohini, Department of Statistics, University of Lucknow, India

Zhang, Ying-Ying, Department of Statistics and Actuarial Science, Chongqing University, China

Zieliński, Wojciech, Department of Econometrics and Statistics, Warsaw University of Life Sciences, Poland

STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 235–242

Index of Authors, Volume 24, 2023

Abdullahi, U. K., *Power ratio cum median – based ratio estimator of finite population mean with known population mean – SiTns, Vol. 24, No. 5*

Abimbola, O. V., *see under Adeboye N.O. – SiTns, Vol. 24, No. 2*

Adebiyi, A. A., *see under Olalude G. A. – SiTns, Vol. 24, No. 3*

Adeboye, N. O., *Modelling the Vol.atility of African capital markets in the presence of the COVID-19 pandemic: evidence from five emerging economies in Africa – SiTns, Vol. 24, No. 2*

Adesina, O. A., *see under Olalude G. A. – SiTns, Vol. 24, No. 3*

Adil, E., *see under El Moury I. – SiTns, Vol. 24, No. 2*

Adri, A., *see under Oullada O. – SiTns, Vol. 24, No. 5*

Ahmed, R., *see under Nadeem M. – SiTns, Vol. 24, No. 2*

Aidi, K., *see under Yousof H.M. – SiTns, Vol. 24, No. 4*

Akhtar, N., *Bayesian estimation of a geometric distribution using informative priors based on a Type – I censoring scheme – SiTns, Vol. 24, No. 3*

Akinbo, R. Y., *see under Adeboye N. O. – SiTns, Vol. 24, No. 2*

Ali, A., *see under Akhtar N. – SiTns, Vol. 24, No. 3*

Ali, M. M., *see under Yousof H. M. – SiTns, Vol. 24, No. 4*

Alizadeh, M., *see under Ranjbar V. – SiTns, Vol. 24, No. 4*

Amin, M., *see under Akhtar N. – SiTns, Vol. 24, No. 3*

Aslam, M., *see under Kiani S. K. – SiTns, Vol. 24, No. 4*

Bandyopadhyay, A., *see under Das P. – SiTns, Vol. 24, No. 5*

Baral, M. M., *see under Mukherjee S. – SiTns, Vol. 24, No. 5*

Ben, Ali M., *see under Oullada O. – SiTns, Vol. 24, No. 5*

Beresovsky, V., *see under Savitsky T. D. – SiTns, Vol. 24, No.5*

Bhattacharjee, A., *Bayesian modelling for semi – competing risks data in the presence of censoring – SiTns, Vol. 24, No. 3*

Bhattacharjee, S., *see under Szymkowiak M. – SiTns, Vol. 24, No. 5*

Bhatti, M. I., *see under Kiani S. K. – SiTns, Vol. 24, No. 4*

Bhushan, S., *On some efficient classes of estimators using auxiliary attribute – SiTns, Vol. 24, No. 2*

Białek, J., *Quality adjusted GEKS – type indices for price comparisons based on scanner data – SiTns, Vol. 24, No. 3*

Bondaruk, T., *Budgetary policy of Ukraine in times of challenges and its impact on financial security – SiTns, Vol. 24, No. 1, Special Issue*

Boumahdi, M., *Conditional density function for surrogate scalar response – SiTns, Vol. 24, No. 3*

Bulatova O., *see under Hrynchak N. – SiTns, Vol. 24, No. 1, Special Issue*

Campanelli, L., *Breaking Benford's law: a statistical analysis of COVID-19 data using the Euclidean distance statistic – SiTns, Vol. 24, No. 2*

Cembruch-Nowakowski, M., *see under Dorocki S. – SiTns, Vol. 24, No. 4*

Chala, T., *Statistical modeling and forecasting of wheat and meslin export from Ukraine using singular spectral analysis – SiTns, Vol. 24, No. 1, Special Issue*

Chaudhuri, A., *see under Patra D. – SiTn, 24, No. 4*

Chernenko, D., *see under Chala T. – SiTns, Vol. 24, No. 1, Special Issue*

Chittipaka, V., *see under Mukherjee S. – SiTns, Vol. 24, No. 5*

Chłoń-Domińczak, A., *see under Ptak-Chmielewska A. – SiTns, Vol. 24, No. 4*

Chugaievska, S., *Census administration in Ukraine: insight into the Polish experience in the context of international indicators Analysis – SiTns, Vol. 24, No. 3*

Chwila, A., *The prediction of new COVID-19 cases in Poland with machine learning Models – SiTns, Vol. 24, No. 2*

Chebir, A., *see under El Moury I. – SiTns, Vol. 24, No. 2*

Das, P., *Estimation of ratio of two population means in two – phase stratified random sampling under scrambled response situation – SiTns, Vol. 24, No. 5*

Dehnel, G., *see under Chugaievska, S. – SiTns, Vol. 24, No. 3*

Dey, R., *see under Bhattacharjee A. – SiTns, Vol. 24, No. 3*

Dorocki, S., *Application of statistical methods in socioeconomic geography and spatial management based on selected scientific journals listed in the Web of Sciences database – SiTns, Vol. 24, No. 4*

Dudek, H., *see under Landmesser-Dudek J. – SiTns, Vol. 24, No. 4*

Dwivedi, S. N., *see under Verma V. – SiTns, Vol. 24, No. 2*

Eftekharian, A., *see under Ranjbar V. – SiTns, Vol. 24, No. 4*

Eideh, A., *On representativeness, informative sampling, nonignorable nonresponse semiparametric prediction and calibration – SiTns Vol. 24. No. 2*

El Moury, I., *Proposal of a causal model measuring the impact of an ISO 9001 certified Quality Management System on financial performance of Moroccan service – based companies – SiTns, Vol. 24., No. 2*

Folorunso, S. O., *see under Adeboye No. – SiTns, Vol. 24, No. 2*

Gershunskaya, J., *Discussion – SiTns, Vol. 24, No. 3 & see under Savitsky T. D. – SiTns, Vol. 24, No.5*

Giri, R. L., *see under Szymkowiak M. – SiTns, Vol. 24, No. 5*

Gladun, O., *see under Puhachova M. – SiTns, Vol. 24, No. 1, Special Issue*

Grzenda, W., *Estimating the probability of leaving unemployment for older people in Poland using survival models with censored data – SiTns, Vol. 24, No. 3*

Hadini, M., *see under El Moury I. – SiTns, Vol. 24, No. 2*

Holubova, H., *Comparative analysis of the use of Principal Components Analysis and Parallel Analysis in working with official statistical data – SiTns, Vol. 24, No. 1, Special Issue*

Horobets, O., *see under Osaulenko O. – SiTns, Vol. 24, No. 1, Special Issue*

Hrynchak, N., *Problems of statistical survey of the national market of logistics services in war conditions – SiTns, Vol. 24, No. 1, Special Issue*

Ibrahim, M., *see under Yousof H. M. – SiTns, Vol. 24, No. 4*

Isamot, O. W., *see under Olanrewaju R. O. – SiTns, Vol. 24, No. 4*

Ivashchenko, O., *see under Reznikova N. – SiTns, Vol. 24, No. 1, Special Issue*

Jimoh, T. A., *see under Olalude G. A. – SiTns, Vol. 24, No. 3*

Kalton, G., *Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day & Rejoinder – SiTns, Vol. 24, No. 3*

Kashif Rasheed, H. M., *see under Nadeem M. – SiTns, Vol. 24, No. 2*

Kefelegn, E., *Determinants of livestock products export in Ethiopia – SiTns, Vol. 24, No. 2*

Keser Istem, K., *see under Kocakoç, Ipek D. – SiTns, Vol. 24, No. 2*

Khan, A. A. *see under Akhtar N. – SiTns, Vol. 24, No. 3*

Khan, S. A., *see under Akhtar N. – SiTns, Vol. 24, No. 3*

Kharazmi, O., *see under Ranjbar V. – SiTns, Vol. 24, No. 4*

Khoshkhoo Amiri, Z., *Analysis for the xgamma distribution based on record values and inter – record times with application to prediction of rainfall and COVID-19 records – SiTns, Vol. 24, No. 5*

Kiani, S. K., *Investigation of half– normal model using informative priors under Bayesian structure – SiTns Vol. 24. No. 4*

Klochko, R., *Marketing segmentation of banks corporate clients based on data mining technique – SiTns, Vol. 24, No. 5*

Kobylynska, T., *Statistical study of climate change in Ukraine under conditions of martial law – SiTns, Vol. 24, No. 1, Special Issue*

Kocakoç, I. D., *Outlier detection based on the functional coefficient of variation – SiTns, Vol. 24, No. 2*

Kończak, G., *Changepoint detection with the use of the resperm method – Monte Carlo study – SiTns, Vol. 24, No. 5*

Korczyński, A., *Bayesian predictive probability design – theory and practical example in a prospective study – SiTns, Vol. 24, No. 5*

Korepanov, G., *see under Chala T. – SiTns, Vol. 24, No. 1, Special Issue*

Korepanov, O., *see under Chala T. – SiTns, Vol. 24, No. 1, Special Issue*

Krekhivskyi, O., *A new industrial strategy for Europe – new indicators of implementation results – SiTns, Vol. 24, No. 1, Special Issue*

Kumar, A., *see under Bhushan S. – SiTns Vol. 24. No. 2*

Kuzmenko, O., *Assessing the maturity of the current system for combating financial and cyber fraud – SiTns, Vol. 24, No. 1, Special Issue*

Lahiri, P., *see under Gershunskaya J. – SiTns Vol. 24. No. 3 & Vogt M.,*

Landmesser-Rusek, J., *What explains the differences in material deprivation between rural and urban areas in Poland before and during the COVID-19 pandemic? – SiTns, Vol. 24, No. 4*

Lawson, N., *see under Abdullahi U. K. – SiTns, Vol. 24, No. 5*

Lazebnyk, J., *see under Chala T. – SiTns, Vol. 24, No. 1, Special Issue*

Legan, I., *see under Kobylynska T. – SiTns, Vol. 24, No. 1, Special Issue*

Lehtonen, R., *Comments – SiTns, Vol. 24, No. 3*

Libanova, E., *War wave of Ukrainian emigration: an attempt to evaluate the scale and consequences – SiTns, Vol. 24, No. 1, Special Issue*

Lubenchenko, O., *Method of auditing in conditions of martial law – SiTns, Vol. 24, No. 1, Special Issue*

Manzoor, S., *see under Akhtar N. – SiTns, Vol. 24, No. 3*

Melese, S. F., *see under Ndlovu B. D.*

Mir Mostafaee, S. M. T. K., *see under Khoshkhoo Amiri Z. – SiTns, Vol. 24, No. 5*

Misra, S. K., *see under Szymkowiak M. – SiTns, Vol. 24, No. 5*

Mohamed, B. A., *see under El Moury I. – SiTns, Vol. 24, No. 2*

Momotiuk, L., *see under Bondaruk T. – SiTns, Vol. 24, No. 1, Special Issue*

Motuzka, O., *see under Kobylynska T. – SiTns, Vol. 24, No. 1, Special Issue*

Mukherjee ,S., *Investigating the factors of Blockchain technology influencing food retail supply chain management: A study using TOE framework – SiTns, Vol. 24, No. 5*

Münnich, R., *Discussion – SiTns, Vol. 24, No. 3 & Vogt M.,*

Nadeem, M., *New generators for minimal circular generalised neighbour designs in blocks of two different sizes – SiTns, Vol. 24, No. 2*

Nath, D. C., *see under Verma V. – SiTns, Vol. 24, No. 2*

Ndlovu, B. D., *A nonparametric analysis of discrete time competing risks data: a comparison of the cause – specific– hazards approach and the vertical approach – SiTns Vol. 24. No. 3*

Noreen, K., *see under Nadeem M. – SiTns, Vol. 24, No. 2*

Ogay, M., *see under Sarioglo V. – SiTns, Vol. 24, No. 1, Special Issue*

Olalude, G. A., *Household expenditure in Africa: evidence of mean reversion – SiTns, Vol. 24, No. 3*

Olanrewaju, R. O, *Hyper – parametric Generalized Autoregressive Scores (GASs): an application to the price of United States cooking gas – SiTns, Vol. 24, No. 4*

Olanrewaju, S. A., *see under Olanrewaju R. O. – SiTns, Vol. 24, No. 4*

Olayinka, H. A., *see under Olalude G. A. – SiTns, Vol. 24, No. 3*

Osaulenko, O., *Using Big Data by the Ukrainian Official Statistics in the Conditions of Martial Law: Problems and Their Solutions – SiTns, Vol. 24, No. 1, Special Issue*

Ouassou, I., *see under Boumahdi M. – SiTns, Vol. 24, No. 2*

Oullada, O., *A model for measuring the impact of good pharmacovigilance practices by patients on the COVID-19 period on the reactivity of hcp: Case Study in Morocco – SiTns, Vol. 24, No. 5*

Öztaş Ayhan, H., *Models for survey nonresponse and bias adjustment techniques – SiTns, Vol. 24, No. 3*

Pal, S., *see under Patra D. – SiTn, 24, No. 4*

Pal, S. K., *see under Mukherjee S. – SiTns, Vol. 24, No. 5*

Panichkitkosolkul, W., *Testing the annual rainfall dispersion in Chaiyaphum, Thailand, by using confidence intervals for the coefficient of variation of an inverse gamma distribution – SiTns, Vol. 24, No. 4*

Patra, D., *Respondent– specific randomized response technique to estimate sensitive proportion – SiTns, Vol. 24, No. 4*

Pavlova, H., *see under Lubenchenko O. – SiTns, Vol. 24, No. 1, Special Issue*

Perkhun, L., *see under Kuzmenko O. – SiTns, Vol. 24, No. 1, Special Issue*

Pfeffermann, D., *Comments – SiTns, Vol. 24, No. 3*

Piskunova, O., *see under Klochko R. – SiTns, Vol. 24, No. 5*

Pozniak, O., *see under Libanova E. – SiTns, Vol. 24, No. 1, Special Issue*

Priyanka, K., *Advances in estimation by the item sum technique in two move successive sampling – SiTns, Vol. 24, No. 4*

Ptak-Chmielewska, A., *Analysis of social and economic conditions of microenterprises based on taxonomy methods – SiTns, Vol. 24, No. 4*

Ptashchenko, O., *see under Hrynchak N. – SiTns, Vol. 24, No. 1, Special Issue*

Puhachova, M., *Using Electronic Registries to Study the COVID-19 Pandemic and its Consequences – SiTns, Vol. 24, No. 1, Special Issue*

Rachdi, M. *see under Boumahdi M. – SiTns, Vol. 24, No. 2*

Ranjbar, V., *Odd log – logistic generalised Lindley distribution with properties and applications – SiTns, Vol. 24, No. 4*

Rajoriya, D., *Under Military War Weapon Support The Economic Bond Level Estimation Using Generalized Petersen Graph with Imputation – SiTns, Vol. 24, No. 1, Special Issue*

Reznikova, N., *The impact of the Russian– Ukrainian war on the green transition and the energy crisis: Ukrainian scenario of circular economy development – SiTns, Vol. 24, No. 1, Special Issue*

Rifai, S., *see under Oullada O. – SiTns, Vol. 24, No. 5*

Roychowdhury, A. *see under Szymkowiak M. – SiTns, Vol. 24, No. 5*

Salikhova, O., *see under Krekhivskyi O. – SiTns, Vol. 24, No. 1, Special Issue*

Sarioglo, V., *Approach to population estimation in Ukraine using mobile operators' data – SiTns, Vol. 24, No. 1, Special Issue*

Savitsky, T. D., *Methods for Combining Probability and Nonprobability Samples Under Unknown Overlaps – SiTns, Vol. 24, No. 5*

Sharma, A., *Does economic freedom promote financial development? Evidence from EU countries – SiTns, Vol. 24, No. 3*

Sharma, V., *see under Tiwari K. K. – SiTns, Vol. 24, No. 3*

Shukla, D., *see under Rajoriya D. – SiTns, Vol. 24, No. 1, Special Issue*

Shulga, S., *see under Lubenchenko O. – SiTns, Vol. 24, No. 1, Special Issue*

Singh, G. N. *see under Das P. – SiTns, Vol. 24, No. 5*

Stąpor, K., *see under Kończak G. – SiTns, Vol. 24, No. 5*

Szymkowiak, M., *A study of survival data using kernel estimates of hazard rate and aging intensity functions – SiTns, Vol. 24, No. 5*

Targonskii, A., *see under Chugaievska, S. – SiTns, Vol. 24, No. 3*

Tiwari, K. K., *Efficient estimation of population mean in the presence of non– response and measurement error – SiTns, Vol. 24, No. 3*

Tokas, S., *see under Sharma A. – SiTns, Vol. 24, No. 3*

Tomczyk, E., *Dynamics of survey responses before and during the pandemic: entropy and dissimilarity measures applied to business tendency survey data – SiTns, Vol. 24, No. 2*

Trisandhya, P., *see under Priyanka K.,*

Ugwuowo, F. I., *see under Abdullahi U. K. – SiTns, Vol. 24, No. 5*

ul Hassan, M., *see under Nadeem M. – SiTns, Vol. 24, No. 2*

Vasyechko, O., *Current Challenges of the Consumer Price Index (CPI) In Ukraine – SiTns, Vol. 24, No. 1, Special Issue*

Verma, V., *Bayesian estimation of fertility rates under imperfect age reporting – SiTns, Vol. 24, No. 2*

Vogt, M., *Spatial Prediction in Small Area Estimation – SiTns, Vol. 24, No. 3*

Wesołowski, J., *Rotation schemes and Chebyshev polynomials – SiTns, Vol. 24, No. 3*

Williams, M. R., *see under Savitsky T. D. – SiTns, Vol. 24, No.5*

Yarovenko, H., *see under Kuzmenko O. – SiTns, Vol. 24, No. 1, Special Issue*

Yatsenko, O., *see under Hrynchak N. – SiTns, Vol. 24, No. 1, Special Issue*

Yousof, H. M., *The modified Bagdonavičius– Nikulin goodness– of– fit test statistic for the right censored distributional validation with applications in medicine and reliability – SiTns, Vol. 24, No. 4*

Yaya, O. S., *see under Olalude G. A. – SiTns, Vol. 24, No. 3*

You, Y., *An empirical study of hierarchical Bayes small area estimators using different priors for model variances – SiTns, Vol. 24, No. 4*

Wójciak, W., *Another Solution for Some Optimum Allocation Problem – SiTns, Vol. 24, No. 5*

Zaichko, I., *see under Bondaruk T. – SiTns, Vol. 24, No. 1, Special Issue*

Zewotir, T., *see under Ndlovu B. D. – SiTns, Vol. 24, No. 3*

Zielińska-Kolasińska, Z., *A new confidence interval for the odds ratio – SiTns, Vol. 24, No. 2*

Zieliński, W., *see under Zielińska– Kolasińska Z. – SiTns, Vol. 24, No. 2*

Zvarych, I., *see under Reznikova N. – SiTns, Vol. 24, No. 1, Special Issue*

Zvarych, R., *see under Reznikova N. – SiTns, Vol. 24, No. 1, Special Issue*

Żebrowska-Suchodolska, D., *Elimination of characteristics concerning the performance of open– ended equity funds using PCA – SiTns, Vol. 24, No. 4*

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* https://sit.stat.gov.pl/ForAuthors.

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).