Volume 23, Number 4, December 2022 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

EDITORIAL OFFICE **ISSN 1234-7655**

Scientific Secretary
Marek Cierpiał-Wolan, *Statistical Office in Rzeszów, Rzeszów, Poland, e-mail: m.cierpial-wolan@stat.gov.pl*
Managing Editor
Adriana Nowakowska, *Statistics Poland, Warsaw, e-mail: a.nowakowska3@stat.gov.pl*
Secretary
Patryk Barszcz, *Statistics Poland, Warsaw, Poland, e-mail: p.barszcz@stat.gov.pl, phone number +48 22 – 608 33 66*
Technical Assistant
Rajmund Litkowiec, *Statistical Office in Rzeszów, Rzeszów, Poland, e-mail: r.litkowiec@stat.gov.pl*

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

CONTENTS

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. III

Submission information for Authors

Statistics in Transition new series (SiT) 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 2022
Vol. 23, No. 4, 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 2022
Vol. 23, No. 4, pp. VII–XII

From the Editor

Hereby, we present to our readers the last issue of Statistics in Transition new series in 2022, containing 12 articles on various topics and nature. As usual, articles are grouped into three conventionally defined categories: research papers, other articles and Communicates and Letters. The authors of these articles come from many countries: Czech Republic, USA, Egypt, Poland, Iran, Nigeria, India, Sri Lanka, Botswana, and Thailand.

As the last publication of the year, this issue gives us the opportunity to express our appreciation and thanks to all our contributors: authors, reviewers and all the participants of the editorial process. We consider it a great achievement to have such a large and growing community of internationally renowned experts among our collaborators and journal's stakeholders. I would like to express my special gratitude to the reviewers – also on behalf of the authors of the published articles – as their comments and suggestions had a positive impact on the overall quality of the submitted papers.

On behalf of the Editorial Board, Associate Editors and the journal's readers I sincerely thank all our partners and patrons.

Research articles

The first paper by **Richard Hindls, Lubos Marek** and **Stanislava Hronová** entitled ***Changes in the structure of household disposable income in selected countries as a reflection of the crises after 2000*** shows how the relationship between the shares of households' wages and final consumption expenditure in their gross disposable income has developed over the past 20 years. The presented analysis uses publicly available national accounts data for 30 countries for the period of 2000–2019. The studied indicators include the proportion of households' wages and salaries, and final consumption expenditure in their gross disposable income. The analysis of the newly constructed measure t has shown a decrease (i.e. an approach to the origin of the coordinates in the spatial map of the 30 countries) of these proportions in the years of financial crisis and economic recession and, on the contrary, an increase (i.e. a move away from the origin of the coordinates of the spatial map) of the examined proportions in the years of prosperity (economic growth). To confirm this assumption, along with the substantive reasoning, the authors have also used the original measure t, which not only quantifies these statements sensitively, but also defines the intensity of the

phenomenon (the degree of approach or departure from the origin of the coordinates). The aggregate analysis is then applicable without any limitation in terms of the number of countries (or entire territories) and years studied – the procedure can be applied, for example, to groups of countries according to their economic development, their geopolitical demarcation, etc.

Yang Liu and **Balgobin Nandram** in their article ***Sampling methods for the concentration parameter and discrete baseline of the Dirichlet process*** start with observation that there are many models in the current statistical literature for making inferences based on samples selected from a finite population. The authors review the current sampling methods for the concentration parameter, which use the continuous baseline distribution, and compare three different methods: the adaptive rejection method, the mixture of gammas method and the grid method. A new method based on the ratio of uniforms, and a discrete baseline approach to the DP prior and sample the unobserved responses from the finite population both using a Polya urn scheme and a multinomial distribution were proposed. The discrete baseline approach to a Phytophthora data set was applied. The ratio of uniforms is more accurate and it is faster considering the computational time. The authors have corrected the true number of distinct values in the sample by introducing a latent variable that indicates from which urn a new observation comes. Due to using this approach, the authors could give a more accurate estimation of the finite population mean when the observations are discrete.

The next paper ***Parameter estimation of exponentiated exponential distribution under selective ranked set sampling*** prepared by **Amal S. Hassan, Rasha S. Elshaarawy** and **Heba F. Nagy** describe the PRSS (Partial Ranked Set Sampling) method, which allows flexibility for the experimenter in selecting the sample when it is either difficult to rank the units within each set with full confidence or when experimental units are not available. The authors introduce and define the density and likelihood function for a random variable under the PRSS scheme. The suggested ranked schemes include the PRSS, RSS, neoteric RSS (NRSS), and extreme RSS (ERSS). An intensive simulation study was conducted to compare and explore the behaviour of the proposed estimators. The study demonstrated that the maximum likelihood estimators via PRSS, NRSS, ERSS, and RSS schemes are more efficient than the corresponding estimators under SRS. Also, PRSS is not the best method compared to the other ranked schemes, but it is important in some cases when selecting the sample.

Piotr Sulewski and **Magdalena Szymkowiak** present ***The Weibull lifetime model with randomised failure-free time.*** They indicated that treating failure-free time in the three-parameter Weibull distribution not a constant, but as a random variable, makes the resulting distribution much more flexible at the expense of only one additional

parameter. Four compound Weibull distributions with the location parameter having Uniform, Weibull, Gamma and Normal distribution were defined. Using these proposed models the analysis of three real lifetime data sets was performed. The received results showed that the new models fit better the data under consideration that the standard three-parameter Weibull distribution. However, anyone who will decide to use any of the proposed compound Weibull distributions in data analysis has to be equipped with a powerful computational environment – Excel, Mathcad, Mathematica, Matlab, Scilab, etc.

In the paper ***Robustness of randomisation tests as alternative analysis methods for repeated measures design*** **Abimibola Victoria Oladugba, Ajali John Obasi** and **Oluchukwu Chukwuemeka Asogwa** discuss the problem of using randomisation tests (*R*-tests) which are regularly proposed as an alternative method of hypothesis testing in a situation when assumptions of classical statistical methods are violated in data analysis. The authors describe the robustness in terms of the type-I-error and the power of the *R*-test, which were evaluated and compared with that of the *F*-test in the analysis of a single factor repeated measures design. The Monte Carlo approach was used in the simulation study. The results showed that when the data were normal, the *R*-test was approximately as sensitive and robust as the *F*-test, while being more sensitive than the *F*-test when data had skewed distributions. When the sphericity assumption was met, both the *R*-test and the *F*-test were approximately equally sensitive, whereas the *R*-test was more sensitive and robust than the *F*-test when the sphericity assumption was not met

Narendra Singh Thakur's and **Diwakar Shukla's** paper ***Missing data estimation by the technique of chaining in the survey sampling*** pointed out that the sample surveys are often affected by missing observations and non-response caused by the respondents' refusal or unwillingness to provide the requested information, or due to their memory failure. In order to substitute the missing data, a procedure called imputation is applied, which uses the available data as a tool for the replacement of the missing values. Two auxiliary variables create a chain which is used to substitute the missing part of the sample. The authors present the application of the chain-type factor estimator as a means of source imputation for the non-response units in an incomplete sample. The proposed strategies were found to be more efficient and bias-controllable than similar estimation procedures described in the relevant literature. These techniques could also be made nearly unbiased in relation to other selected parametric values. The findings are supported by a numerical study involving the use of a data set, proving that the proposed techniques outperform other similar ones.

The article ***Zero-modified Poisson-Modification of Quasi Lindley distribution and its application*** by **Ramajeyam Tharshan** and **Pushpakanthie Wijekoon** presents the

Poisson-Modification of Quasi Lindley (PMQL) distribution as a newly introduced mixed Poisson distribution for over-dispersed count data. The authors introduce the Zero-modified PMQL (ZMPMQL) distribution as an alternative to the PMQL distribution in order to accommodate zero inflation/deflation. The method of obtaining the ZMPMQL distribution jointly with some of its important properties, namely the probability mass and distribution functions, mean, variance, index of dispersion, and quantile function are presented. The maximum likelihood (ML) estimation method is used for the unknown parameter estimation, and the simulation study is conducted in order to evaluate the asymptotic theory of the ML estimation method and to show the superiority of the ML method over the method of moments estimation. The applicability of the introduced distribution is illustrated by using a real-world data set. In order to estimate its unknown parameters, the authors derived its log-likelihood function and score functions, which showed that the maximum likelihood estimation method is a suitable method to estimate its unknown parameters via a Monte Carlo simulation study. The results revealed its superiority over some other existing mixed Poisson and zero-modified mixed Poisson distributions.

Wilfred Molefe in the paper ***Optimal allocation for equal probability two-stage design*** examines the optimal designs when it is not feasible for every cluster to be represented in a sample as in stratified design, by assuming equal probability two-stage sampling where clusters are small areas. The paper develops allocation methods for two-stage sample surveys where small-area estimates are a priority. The author seeks efficient allocations where the aim is to minimize the linear combination of the mean squared errors of composite small area estimators and of an estimator of the overall mean. Several alternatives, including the area-only stratified design, are found to perform nearly as well as the optimal allocation but with better practical properties. Designs are evaluated numerically using Switzerland canton data as well as Botswana administrative districts data. This optimal design is less clustered than the usual classical two-stage optimal sample size $\bar{n}_{cl.}$ when more priority is given to larger clusters ($q > 0$). The area-only stratified optimum and the area-only simple two-stage optimum should always be the best designs in minimizing the objective function but they are not when there is equal priority for each cluster, that is when $q = 0$. These two designs have undesirable properties of allocating zero or even negative sample sizes to smaller clusters.

The next paper, by **Arisa Jiratampradab, Thidaporn Supapakorn** and **Jiraphan Suntornchost,** presents ***Comparison of confidence intervals for variance components in unbalanced one-way random effects model*** to study and compare the methods for constructing confidence intervals for variance components in an unbalanced one-way random effects model. The methods are based on a classical exact, generalised pivotal quantity, a fiducial inference and a fiducial generalised pivotal quantity.

The comparison of criteria involves the empirical coverage probability that is maintained at the nominal confidence level of 0.95 and the shortest average length of the confidence interval. The simulation results show that the method based on the generalised pivotal quantity and the fiducial inference perform very well in terms of both the empirical coverage probability and the average length of the confidence interval. The classical exact method performs well in some situations, while the fiducial generalised pivotal quantity performs well in a very unbalanced design. Therefore, the method based on the generalised pivotal quantity is recommended for all situations.

Arvind Pandey, David D. Hanagal and **Shikhar Tyagi** focus on ***Generalized Lindley shared additive frailty regression model for bivariate survival data.*** Frailty models are the possible choice to counter the problem of the unobserved heterogeneity in individual risks of disease and death. Based on earlier studies, shared frailty models can be utilised in the analysis of bivariate data related to survival times (e.g. matched pairs experiments, twin or family data). It was assumed that frailty acts additively to the hazard rate. A new class of shared frailty models based on generalised Lindley distribution is established. By assuming generalised Weibull and generalised log-logistic baseline distributions, the authors propose a new class of shared frailty models based on the additive hazard rate. The parameters in these frailty models and the use of the Bayesian paradigm of the Markov Chain Monte Carlo (MCMC) technique were estimated, and model selection criteria were applied for the comparison of models. The kidney infection data allowed to conclude that the best model was analysed. To fit the proposed model the hybrid M-H algorithm was applied.

Other articles

The XXXIX Multivariate Statistical Analysis 2021, Lodz, Poland. Conference Papers

In the paper by **Morteza Nazifi** and **Hamid Fadishei** ***Supsim: a Python package and a web-based JavaScript tool to address the theoretical complexities in two-predictor suppression situations*** the authors show that two-predictor suppression situations continue to produce uninterpretable conditions in linear regression. Their study introduces two different versions of software called suppression simulator (Supsim): a) the command-line Python package, and b) the web-based JavaScript tool, both of which are able to simulate numerous random two-predictor models (RTMs). Such a comparison suggests that the basic mathematical concepts of two-predictor suppression situations need to be reconsidered with regard to the important issue of the statistical control function.

The study depicts a clear picture of the performance of the statistical control function in different suppression and non-suppression situations, and provides a mathematical proof indicating that the statistical control function does not work

correctly in suppression situations. The study also introduces an algorithm that can generate numerous simulated data sets showing all different kinds of suppression and non-suppression situations known so far, and therefore they help resolve the theoretical complexities related to two-predictor suppression situations by expanding the pervious knowledge in this field.

Research Communicates and Letters

Barbara Wodecka and **Michał Stachura** discuss ***k*-th *record estimator of the scale parameter of the α-stable distribution***. The authors present an estimation technique that involves the *k*-th record theory. Several theoretical properties of the introduced scale parameter estimators are demonstrated. With the use of Monte Carlo methods, a comparative analysis is performed between the approach based on *k*-th records and approaches based on Hill's and Pickands' estimators. The research indicates several advantages of the *k*-th record approach over its other counterparts, especially when dealing with incomplete information about the underlying sample. It is also remarked that the insights, specified in the paper, should be perceived essentially as the advantages of the '*k*-th record' approach over the other ones presented, since Berred's estimator, and the scale parameter estimator based on it, may be employed in cases of incomplete information about an underlying sample. The authors show that, on the one hand, this incompleteness may be very useful if an analysed data base must stay undisclosed, even for a researcher/statistician working on it, or more, the data are only partially recorded (i.e. record values of a proper order or several orders). On the other hand, if an analysed data base is absolutely fulfilled and disclosed, the '*k*-th record' approach opens up opportunities to make use of permutation methods in order to make repeated estimation that leads to much more precise results.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 1–19, DOI 10.2478/stattrans-2022-0039
Received – 03.02.2022; accepted – 31.05.2022

Changes in the structure of household disposable income in selected countries as a reflection of crises after 2000

Richard Hindls[1], **Lubos Marek**[2], **Stanislava Hronová**[3]

ABSTRACT

Wages and salaries represent the most important component of household disposable income. The aim of the article is to examine how the relationship between the shares of households' wages and final consumption expenditure in their gross disposable income has developed over the past 20 years. The presented analysis uses publicly available national accounts data for 30 countries for the period of 2000–2019. The studied indicators include the proportion of households' wages and salaries, and final consumption expenditure in their gross disposable income. Using the proposed method based on the evaluation of changes in the spatial map, it is possible to observe any significant changes in these proportion values in the years of financial crisis and recession, as well as in the years of prosperity. The procedure can therefore serve as an indicator of appreciable changes in economic development.

Key words: gross domestic product, final consumption expenditure, disposable income, mutual change of two relative indicators in space and time, indicators of income and expenditure in households.

1. Introduction

Households (represented in the national accounts by the household sector) represent an entity with a specific main economic behaviour, namely, consumption. The final consumption expenditure is funded by households' disposable income, which is the result of the distribution and redistribution of income derived from productive activity and whose most important component is labour income, i.e. wages and salaries. Households enter into the distribution process as parties that get more than they pay;

[1] Department of Statistics and Probability, Faculty of Informatics and Statistics, Prague University of Economics and Business, Czech Republic. E-mail: hindls@vse.cz. ORCID: https://orcid.org/0000-0002-0887-3346.

[2] Department of Statistics and Probability, Faculty of Informatics and Statistics, Prague University of Economics and Business, Czech Republic. E-mail: marek@vse.cz. ORCID: https://orcid.org/0000-0003-4761-1936.

[3] Department of Economic Statistics, Faculty of Informatics and Statistics, Prague University of Economics and Business, Czech Republic. E-mail: hronova@vse.cz. ORCID: https://orcid.org/0000-0002-3568-9755.

households receive wages and salaries, social benefits and other income and have to pay taxes on production and imports, income taxes, social contributions, and other transfers. They thus generate resources sufficient for funding their consumption, and at the same time create savings. Moreover, households should be the source of most of the national saving. This role is particularly important in years of crisis, when government deficits and pressures on public budgets are growing.

From the household perspective, years of economic growth have brought not only rising income from wages (due to rising wage levels and falling unemployment) but also a rising level of confidence, which is undoubtedly important for their willingness to consume. Other factors that influence the level of household final consumption expenditure include the availability of consumer and mortgage credit (determined mainly by the level of interest rates), the inflation rate and the related development of the cost of living, housing prices, the tax burden, the unemployment rate, etc. However, wage levels and wage growth remain a key factor encouraging the appetite and courage to spend, which in turn increases the volume of final consumption expenditure. The other side of this coin, however, is that rapid growth in household consumption may result in households becoming more indebted in the form of loans. This fact, together with a declining savings rate and financial savings rate, may, despite a favourable economic climate, lead to households becoming over-indebted and jeopardise their ability to meet their obligations.

Periods of recession or even crisis accompanied by uncertainties (not only) on the labour market and stagnation of real income mean a change in household behaviour manifested by a cautious approach to consumption, reduced willingness to invest and take out long-term loans. However, this turnaround is not immediate. As a rule, the effects of the crisis will first hit governments and non-financial corporations or financial institutions, and households only after a delay. At the same time, households are reducing their non-financial investments and diversifying their financial investments, or trying to put their spare funds in less risky assets.

This paper should help answer the question to what extent the wage level is a determinant of changes in household final consumption expenditure, i.e. how household final consumption expenditure responds to changes in household income in the form of wages. For the analysis we have used publicly available data (see Eurostat) for 30 countries. The indicators monitored for the household sector are the proportion of wages and salaries received in gross disposable income and the proportion of household final consumption expenditure in their gross disposable income in the period 2000–2019. The method of analysis used is measure t, which describes changes in the values of variables in a spatial two-dimensional map, which is similar to the so-called perceptual map, known, for example, from marketing analysis.

2. Theoretical background

As mentioned above, the undisputed factor influencing household expenditure on durables and non-durables (the value of which is expressed by the final consumption expenditure indicator) is the level of income. The dominant item of household income is labour income, expressed as wages and salaries in the national accounts. The evolution of household income and its relationship to final consumption expenditure have been the subject of a number of theoretical papers. Understanding of and insight into the determinants of changes in household consumption are important aspects of economic policy, as household final consumption expenditure accounts for around half of gross domestic product (GDP) in developed countries and is an important factor in economic growth.

2.1. The relationship between income and consumption in economic theory

The basis for the discussion of the relationship between household income and consumption is undoubtedly Keynes's discussion of consumption (Keynes, 1936), where he states that as employment grows, labour income rises; this in turn leads to an increase in consumption, which, however, grows more slowly than labour income. The issue was already addressed by Modigliani and Brumberg (1954) and Friedman (1957) soon after World War II, during the post-war boom, probably as a result of reminiscences of the economic crisis of the late 1920s and early 1930s. They developed and described life-cycle models of permanent income in which they tried to show that households use savings to smooth changes in income, so that the effects of these changes on consumption levels are small.

Another model that focused on the evolution of consumption twenty years later was the Hall model, inspired by Friedman (1957, see Hall, 1978). Hall's work, to some extent, challenged the idea mentioned above – that households have only a weak propensity to consume and therefore their consumption is always closely linked to current income. On the contrary, he advocated the idea that, assuming useful and purposeful behaviour, households try to maintain a stable consumption trend in the long run. In his work, he also discussed the time lag between changes in income and changes in consumption expenditure with respect to the state of their assets. His work has had a very important impact on the further development of econometric models of consumption. The evolution of the income-consumption relationship on the background of the labour life cycle was the subject of Heckman (1974). He presented an alternative neoclassical model, in which he showed that as wages evolve over the labour life cycle, the level of consumption changes, or the level of consumption depends on the level of wages at each age. He thus confirmed the results arrived at by Thurow in the late 1960s (see Thurow, 1969).

An interesting empirical analysis of the relationship between permanent and current income and consumption can be found in Lusardi (1996), who shows in a panel data set that consumption is very sensitive to predictable income growth. Attanazio and Davis (1996) based their study on panel data on consumption, wage levels and employment in the US in the 1980s and showed that even small changes in the wage structure among different age and education groups of workers led to significant changes in household consumption expenditures. Jappelli and Pistaferri (2010) attempted to map studies dealing with the reflection of income changes in consumption levels in terms of whether the change is positive or negative and, moreover, expected or unexpected (the so-called income shock). Most models assume that consumption responds to an expected increase in income, significantly more than is assumed by permanent income models. When income is expected to fall (e.g., a transition from economic activity to inactivity), the impact on consumption is rather insignificant. However, the authors emphasise that, in such a situation, it is necessary to distinguish between higher and lower income households and hence easier or more difficult access to credit markets. Among the theoretical underpinnings of the income-consumption relationship, it is worth recalling Duesenberry's (1949) relative income hypothesis, which has, for many years, been unjustly neglected in economic theory. Its importance is presented and developed in Sanders (2010), where he shows the properties and empirical significance of this model.

The level and evolution of household final consumption expenditure provide important information for the direction of economic policy. It is clear that there are other factors besides households' income, such as various macroeconomic impulses and shocks, inflation rate, confidence in the economy, and consumer expectations. Changes in economic and non-economic conditions in the national economy (changes in interest rates, significant reversals in stock prices, natural disasters, corruption scandals, etc.) affect household economic behaviour, but their effect is usually short-lived and implemented through specific channels. For example, Aspergis et al. (2014) address the issue of the relationship between stock and house price movements on household consumption levels and conclude that a stock market slowdown may dampen households' willingness to spend. Hamburg et al. (2008) address the issue of the relationship between income, consumption and wealth in Germany and show that this relationship is dynamic and does not settle after a certain period of time. In general, it is a fact that households will not increase their consumption expenditure unless they consider their economic situation to be good and stable. Rising income, rising market prices of their financial and non-financial assets coupled with economic growth increase their willingness to spend and invest. Household investments in real estate (or financial assets), which are not taken for a part of the household final consumption expenditure indicator, give a strong signal of a satisfactory economic climate.

Conversely, a fall in consumer confidence is one of the signals of a coming recession or crisis. Campelo et al. (2020) investigated this relationship using data from Brazil, showing that indicators of consumer confidence and economic climate are better able to predict trends and changes in household final consumption expenditure, and that improvements in consumer confidence positively affect households' attitudes towards consumption.

The relationship between income shocks and consumption levels in the context of the business cycle is also addressed in Kovacz et al. (2019). Using data from the Netherlands, the authors show that income shocks observed during the years of the global crisis 2008-2009 are of a different nature than those observed during the years of the fiscal crisis 2011-2012, or that shocks induced by the fiscal crisis have a longer-term impact on consumption. This was consequently reflected in the fact that households as consumers reacted more cautiously in the fiscal crisis years than in the global financial crisis years, although the decline in their income in the fiscal crisis years was smaller. This empirical finding from the Dutch economy is very important in our consideration of households' different behaviour with respect to the nature of the economic crisis and raises the question whether this phenomenon is also observable in other, especially European, countries.

The relationship between wage levels and the level of consumption of an individual is unquestionable, although there are a number of other phenomena that influence consumer decision-making. From a macroeconomic point of view, individual consumer behaviour translates into the relationship between wages as remuneration for work and final consumption expenditure. The level of wages, or – from a macroeconomic point of view – the amount of wages received by households, is influenced by the phases of the economic cycle. In times of crisis the unemployment rate goes up, and wage growth slows down or stops. Conversely, in the boom phase, employment grows and the level of wages rises as demand for labour increases. This, of course, has an impact on the volume of household consumption expenditure. Can the relationship between wage developments and household final consumption expenditure be used to document the response of households to the phases of the business cycle? Is this relationship valid and can it be generalised to a larger set of countries? We have tried to answer this question by analysing the relationship between wages and household final consumption expenditure in a set of 30 countries over the last 20 years. For our analysis, we have used the original measure t describing the changes in the spatial map.

2.2. Statistical expression of income and consumption

However, it is useful to subsequently "put the relationships given by economic theory to the test", i.e. to verify the theoretical assumptions on statistical data. Here, we

encounter the first fundamental and ever-present problem of the discrepancy between economic theory and statistical practice, i.e. the discrepancy between the concepts of economic theory and the possibilities of their relevant quantification. This so-called adequacy gap lies at the heart of this problem – many of the concepts with which theoretical economics operates cannot be quantified to the full extent of the concept, and it is therefore necessary to resort to a certain quantitative approximation to these theoretical concepts. This mere "approximation" is therefore a necessary compromise between the needs for quantification of the concepts of theoretical economics and our real ability to carry out this quantification. The trade-off between the "necessary and the possible" is the structural content of the adequacy gap mentioned above[4].

If we are to examine the relationship between household income and consumption, it is necessary to define the data sources from which we will draw comparable data, to define the statistical population of households and to find appropriate indicators of income and consumption. The first two conditions are easy to fulfil – the basic source of internationally comparable data is given by the national accounts, whose standards[5] guarantee a common understanding and definition of indicators. The definition of the household population with respect to the national accounts data sources is also not a problem, since households form one of the five resident institutional sectors and the characteristics of the units belonging to this sector are clearly defined[6].

For indicators reporting household consumption, the national accounts offer two options – household final consumption expenditure and the actual household final consumption. Household final consumption expenditure includes the value of purchased (new and used) goods and services of short– and long-term consumption, excluding dwellings, houses and land, and the value of the so-called consumption in kind, i.e. subsistence, agricultural and food products from subsistence. The indicator also includes the so-called consumption of output for households' own final use, i.e. what households produce and consume themselves (in particular, provision of housing services to themselves, agricultural output from subsistence farming, services of employing domestic staff). This concept of household final consumption expenditure is traditional and, prior to the introduction of the ESA 1995 or SNA 1993[7], corresponded to the only indicator of household final consumption at that time.

The second indicator providing information on household final consumption is the household actual final consumption indicator. The concept of actual final consumption

[4] A simple example of this gap is, e.g. inflation as a theoretical economic category on the one hand and the consumer price index as a quantification of this theoretical concept on the other hand.

[5] See ESA 2010 (2013) and SNA 2008 (2013).

[6] See ESA 2010 (2013) and SNA 2008 (2013), paragraphs 2.118 – 2.128. The household sector according to this definition includes not only households as consumers (employees, recipients of social, property and other income) but also small producers (employers and self-employed).

[7] See ESA 1995 (1996) and SNA 1993 (1993).

was introduced as late as in the ESA 1995 and SNA 1993[8] standards in response to the requirements of international comparability of household consumption in terms of their living standards. Actual final consumption of households is equal to final consumption expenditure plus social transfers in kind[9] that are paid to households by general government and non-profit institutions serving households.

As can be seen from the definitions presented above, there is a dual concept of household final consumption; the first emphasises "what households spend"[10] and the second "what households actually consume". Therefore, when analysing the economic behaviour of the household sector, it is always necessary to choose the appropriate indicator for the purpose of the analysis. In our case, where the relationship between income and consumption is concerned, the indicator of final consumption expenditure is the obvious choice. The indicator of actual final consumption of households contains a part (social transfers in kind) which is mainly a reflection of the social policy of the state (the extent of non-market production of government institutions) and is therefore not a direct consequence of the economic behaviour of households[11].

For the choice of income indicator, the national accounts offer a number of indicators. The most general is undoubtedly disposable income (gross/net). Disposable income is the result of the primary and secondary distribution of income (value added) and its structure consists of business income (gross/net operating surplus and mixed income) + labour income (wages and salaries) + balance of property income + balance of social income (social benefits – social contributions)[12] + balance of other current transfers – current taxes. Disposable income is directly intended to cover final consumption expenditure. In analysing the economic behaviour of households as consumers, we are therefore interested in whether or not disposable income is sufficient to cover final consumption, which is monitored by indicators of the average propensity to consume[13], not whether changes in its level motivate households to change the nature and level of consumption. Disposable income is a macroeconomic statistical variable,

[8] See ESA 1995 (1996) and SNA 1993 (1993).

[9] Social transfers in kind correspond to the value of individual goods and services provided by non-profit institutions and government agencies to households free of charge or at economically insignificant prices, whether they are the result of non-market production (e.g. health, education, etc.) or purchased on the market for household use (housing transport services, etc.). For more details, see ESA 2010, paragraphs 4.108 through 4.111.

[10] Keeping in mind the consumption in kind. For a precise definition of the final consumption expenditure indicator, we refer to ESA 2010, paragraphs 3.94 through 3.99.

[11] For a precise definition of the indicator of actual final consumption, we refer to ESA 2010, paragraphs 3.100 through 3.109.

[12] Given the design of the compensation of employees, other investment income and net social contributions in the household sector account, social contributions are equal to households' actual social contributions + households' social contributions supplements – social insurance scheme service charges.

[13] It is the ratio of final consumption expenditure to (gross/net) disposable income.

i.e. it is not the income that an individual household may view as a certain limit to its consumption[14].

The notional limit of consumption is the wage or salary for employee households, the amount of their retirement income for pensioner households, and the amount of their so-called other income (i.e. social, property and other income) for other households. In the case of small producer households (employers and self-employed), this is undoubtedly part of their profits (mixed income), but their amount cannot be reasonably estimated.

Employee households represent the dominant group in the household sector, and their labour income (wages and salaries) provide the main component of disposable income[15]. Moreover, changes in wage levels can be viewed as a reflection of the economic situation in the national economy, i.e. they, to a certain extent, reflect the evolution of the short-term business cycle. Employee households are also understood as a crucial group in terms of the commodity structure and the volume of final consumption expenditure. Social or other income is independent of the phase of the business cycle; the demand of the corresponding households does not generally cover all commodities and is not a decisive component of household final consumption expenditure. However, ownership income is, to a certain extent, dependent on the business cycle, but the recipient households form only a small part of the units belonging to the household sector[16].

It follows from the above that if we want to analyse the evolution of changes in final consumption expenditure in response to income developments, and moreover in the context of the phases of the business cycle, then the best choice is the wages and salaries indicator. This indicator reflects both regular and irregular cash and in-kind income as remuneration for work performed under labour and other legislation[17].

Both indicators (household final consumption expenditure and wages and salaries) are indicators defined by the System of National Accounts, i.e. they are internationally comparable indicators.

[14] In general, disposable income can be understood as the upper limit of consumption that a household can realise without becoming poorer.

[15] The proportion of gross wages and salaries received as a proportion of gross disposable income is 66% on average in the 30 compared countries (see input data for this analysis) and in none of these countries has it fallen below 40% in recent years.

[16] This is also reflected in the proportion of the balance of proprietary income in gross disposable income, which does not, in the long term, exceed 10% in any of the countries compared.

[17] Wages and salaries represent basic wages and salaries, additional payments for overtime, night work, rest days, profit sharing, holiday pay, transport allowances to and from work, severance pay, remuneration for work under special regulations, professional fees, remuneration for the performance of public functions, compensation for paid time off on public holidays, holiday pay, benefits in kind, free shares distributed to employees, etc. Wages and salaries are gross, i.e. before deductions of income tax and social contributions paid by the employee – see ESA 2010, paragraphs 4.03 to 4.07 for details.

3. Our data and methodology of our analysis

The aim of this paper is to examine the evolution of household final consumption expenditure in relation to wage and salary developments in 30 countries over the period 2000–2019. The analysis is based on publicly available and internationally comparable Eurostat data for the European Union countries (excluding Malta) and selected other countries (UK, Norway, Switzerland, USA, and South Korea). While the methodological comparability of the content of the selected indicators is guaranteed when working with national accounts data, we encounter different currency units (and therefore different levels of values) in which the values of the indicators are expressed and the fact that national accounts data are always in current prices only. The solution is to use relative, i.e. dimensionless, indicators whose values are comparable over time and space.

We have chosen wages and salaries as the income indicator and household final consumption expenditure as the consumption indicator. The choice of appropriate relative indicators is clear in this case; both indicators are components of disposable income – wages and salaries in terms of its creation, final consumption expenditure in terms of its use. It is therefore logical to base our analysis on the indicator of the proportion of wages and salaries received by households in their disposable income on the one hand and the indicator of the proportion of household final consumption expenditure in their disposable income on the other hand.

It remains to resolve the question of whether to use net or gross disposable income in the denominator of these relative indicators. In theory, net disposable income is undoubtedly the more correct option, since the consumption of fixed capital, which makes up for the difference between gross and net disposable income, is meant not for consumption but for investment. However, the use of net disposable income in international comparisons is hampered by the incomparability of methods used for estimating consumption of fixed capital in different countries; for this reason, aggregates such as "gross" are generally used in cases of international comparison. Here, we therefore also use gross disposable household income in the denominator of the chosen indicators.

However, in view of the availability of comparable data on Eurostat's website, it should be noted that only total data for the household and non-profit institutions serving households sector are available. In our case, this concerns the indicators for final consumption expenditure and gross disposable income, but does not affect the value of the wages and salaries indicator, where households as consumers are the only beneficiaries. The combination of the household sector and the non-profit institutions serving the households sector will in principle not affect the values of the indicator for the proportion of final consumption expenditure in gross disposable income and will only slightly distort (downwards) the value of the indicator for the proportion of wages and salaries in gross disposable income. Given that the distortion applies approximately equally to all countries compared, it can be considered negligible. It is also insignificant

from the point of view that households account for between three-fifths and two-thirds of gross national disposable income in the countries surveyed, while non-profit institutions serving households most often account for around 1%, rarely 2%.

The data described above, i.e. the proportion of wages and salaries received by households in gross disposable income (households and non-profit institutions serving households) and the proportion of final consumption expenditure by households and non-profit institutions serving households in their gross disposable income for 30 countries over the period 2000–2019, are the inputs to the model.

The method used to analyse the relationship between the values of the selected indicators is the evaluation of changes in a two-dimensional spatial map. This procedure, published in Hindls, Hronova (2012), consists of an original development of a measure t for the situation where the data are arranged in a two-dimensional spatial map; over a given time period (here 2000–2019), we then observe how the individual values of the dot plot shift over time (i.e. over the years of observation). Measure t is to express the Euclidean distances in the spatial map (see below). This allows us to assess how the phases of the economic cycle (in the years in question) have affected the indicators analysed. Let us now describe the procedure.

To simplify the notations, we denote the proportion of wages and salaries received (hereinafter WSh) in gross disposable household income (hereinafter GDIh) in the i-th country as x_i, $i = 1, 2, ..., n$, where the symbol n denotes the number of countries. Analogously, the proportion of household final consumption expenditure (hereinafter FCEh) in GDIh as y_i, $i = 1, 2, ..., n$. We denote the year of the first observation by the symbol "1" (for the illustration below, let us choose, e.g., 2000 as the beginning year), the year of the second observation by "2" (let us choose 2001 for the illustration). Later, we will analyse all pairs of individual years, i.e. successive pairs of years over the whole period 2000–2019.

Specifically:

- By x_{1i} we mean the ratio of WSh/GDIh in the i-th country, $i = 1, 2, ..., n$, (in our case $n = 30$) in period 1 (year 2000);
- By x_{2i} we mean the ratio of WSh/GDIh in the i-th country, $i = 1, 2, ..., n$, ($n = 30$) in period 2 (year 2001);
- By y_{1i} we mean the ratio of FCEh/GDIh in the i-th country, $i = 1, 2, ..., n$, in period 1 (year 2000); and
- By y_{2i} we mean the ratio of FCEh/GDIh in the i-th country, $i = 1, 2, ..., n$, in the 2nd period (year 2001).

Each of the n countries is thus considered in the light of two different percentages of WSh/GDIh (variable x) and FCEh/GDIh (variable y), in two different periods (years). The baseline values of the relative indicators, i.e. variables x_{1i}, x_{2i}, y_{1i}, y_{2i}, are calculated on the basis of the values of the absolute indicators from the Eurostat database[18].

[18] See https://ec.europa.eu/eurostat/databrowser/view/nasa_10_nf_tr/default/table?lang=fr.

Next, let us denote:

- By K_1 the mean value (i.e. mean spatial localisation) of the two observed indices (i.e. WSh/GDIh and FCEh/GDIh ratios) in period 1 (here year 2000), i.e. the mean value of all points $[x_{1i}\ ;\ y_{1i}]$ located in the Cartesian coordinate space (x, y), see formulae (1) and (2) below; and
- By K_2 the mean value (i.e. mean spatial localisation) of the two observed indicators (i.e. WSh/GDIh and FCEh/GDIh ratios) in period 2 (year 2001), i.e. the mean value of all points $[x_{2i}\ ;\ y_{2i}]$ located in the Cartesian coordinate space (x, y), see formulae (1) and (2) below.

As a summary evaluation of changes in indicators, we propose – see Hindls, Hronová (2012) - a measure t, for which

$$t=\frac{\frac{\sum_{i=1}^{n} k_i}{n}}{\sqrt{\frac{\sum_{i=1}^{n}\left[k_i-\frac{\sum_{i=1}^{n} k_i}{n}\right]^2}{n(n-1)}}} \tag{1}$$

where

$$k_i=\text{sign}\{y_{2i}^2+x_{2i}^2-y_{1i}^2-x_{1i}^2\}\sqrt{(x_{2i}-x_{1i})^2+(y_{2i}-y_{1i})^2} \tag{2}$$

$$\frac{\sum_{i=1}^{n} k_i}{n} \tag{3}$$

is the estimate of the $K_2 - K_1$ value.

The sign{...} operator above is used to determine the sign of the "aggregate spatial" change ("±") of the level of the two-indicator assessment in the second (later) period (here, in the illustration, 2001) compared with the first period (2000). The sign{...} operator thus expresses whether the i-th spatial location (i.e. the location of the i-th country) in period 2 (i.e. 2001) has moved closer to ("–") or farther away from ("+") the centre [0;0] of the coordinates than in period 1 (i.e. 2000). If, for example, the i-th spatial location has moved farther away from the centre of the [0;0] coordinates, then the "+" sign indicates that the aggregate (i.e. for the two relative indicators observed together) position of the indicators in the i-th country has moved farther away from the centre [0;0] of the coordinates (i.e. it is a kind of "geometric" aggregation of the observed indicators). The "–" sign then, of course, represents the opposite situation, i.e. an approach to the centre [0;0] of the coordinates.

Based on the values of measure t, we formulate a conclusion about time changes in the values of the WSh/GDIh and FCEh/GDIh ratios for all n observed countries.

4. Relationship of household income and consumption to the growth rate of the economy

The above relationships show that WSh and FCEh play an extremely important role in the evolution of GDP. Therefore, let us first look at the graph of GDP evolution. The data used cover 30 economically important countries. Figure 1 presents the evolution of GDP (annual growth rates) in three key geopolitical territories between 2000 and 2019, namely Europe and the USA, and finally it shows the global evolution of GDP[19]. Logically, all three time series of annual growth rates are governed by a similar pattern, characterised by an upturn in the performance of the economies in the first 6–7 years of the new millennium, followed by a deep global economic crisis in 2008–2009, then a renewed but milder moderation of the economies around 2012, and then a global slowdown in economic growth rates after 2017.

In 2017, however, there was already open talk of the possible arrival of a recession. However, the subsequent onset of the global SARS-CoV-2 epidemic drowned out any further economic considerations about the real strength of the global economy at the end of the second decade of this century, and thus overshadowed how the economy would develop globally in 2018–2020. For this reason, we have not included 2020 in our considerations, because although it was marked by a severe crisis, it did not have primarily economic causes, but even more so economic consequences. This would have only clouded our purely economic considerations in a 20-year time series.

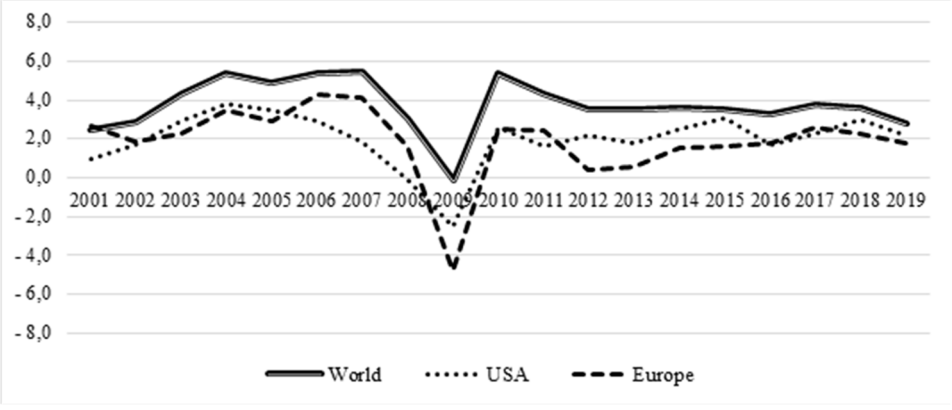


Figure 1. Annual GDP growth rates in selected territories 2000–2019 (percentage values)

Source: https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/OEMDC/ADVEC/WEOWORLD.

[19] See: https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/OEMDC/ADVEC/WEOWORLD

Figure 1 shows that one has to ask how the WSh/GDIh and FCEh/GDIh indicators respond to the aggregate data on the performance of economies (with quite significant periods of change, the 2000–2007 recovery, the crisis years 2008–2009, 2011–2013 and finally the increasing tendency towards recession after 2017). For this analysis, we use the method presented by formulae (1), (2) and (3) above. The inputs to monitoring of changes in the behaviour of the WSh/GDIh and FCEh/GDIh ratios are, naturally, WSh received by households, FCEh and GDIh, aggregated over 30 countries over the period 2000–2019.

From the input data, we determine the value of the measure t from formula (1). For the purpose of further analysis, we have evaluated the evolution of measure t in the years studied. We compute the value of the measure t for each pair of values of the indices $[x_{1i}\ ;\ y_{1i}]$ and $[x_{2i}\ ;\ y_{2i}]$, respectively, where the subscripts labelled "1" and "2" always denote a pair of years in the 20-year time series, i.e. 2000–2019.

This means that we will go for $\binom{20}{2} = 190$ pairs of years, where we always determine the measure t according to (1). The resulting matrix of dimension $20 \cdot 19$ values of the measure t is given in Table 1 in the Appendix. There are "x" symbols on the main diagonal of the matrix because it makes no sense to compare spatial changes in the measure t in the same year (logically, there cannot be a change in t in the same year). The matrix is symmetric due to the existence of relation (2), so we only report the values above the main diagonal of the matrix.

In terms of the objective of our analysis, we are interested in two sets of values:

1. Year-to-year changes in the measure t, i.e. the relationship between WSh/GDIh and FCEh/GDIh for the economic space of all 30 selected countries. That is, year-on-year changes in the relationship of these indicators, e.g. t_{2001}/t_{2000}, etc. (19 values in total). These year-to-year values are shown in Table 1 in the grey boxes diagonally directly above the main diagonal of the matrix (from left to right and simultaneously slanted from top to bottom), and are denoted as t_y-on-y in Figure 2; and
2. Changes in the measure t against the initial period, i.e. the year 2000, i.e. basically the baseline evolution of t-values against the year 2000 (again, 19 values in total), i.e. e.g. t_{2001}/t_{2000}, t_{2002}/t_{2000}, etc. (denoted as t_basic in Figure 2)

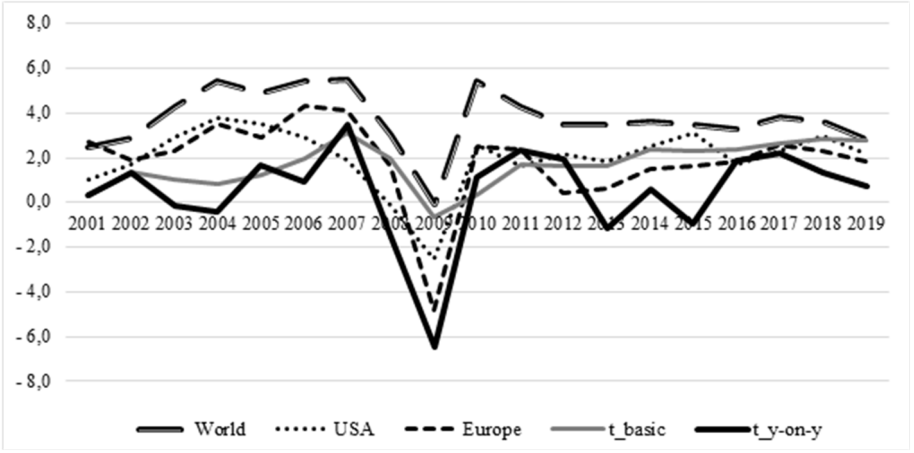


Figure 2. Annual GDP growth rates (%) and changes in measure t

Source: https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/OEMDC/ADVEC/WEOWORLD; own calculations.

Figure 2 and the calculations of the values of the measure t show some important substantive facts concerning the evolution of the key indicators WSh and FCEh. Comparing the evolution of GDP growth rates and the annual values of the measure *t* in Figure 2, it is quite evident that they follow the evolution of GDP with a slight lag. For example, it can be seen that the performance of the economies (US, Europe) has been growing since 2002, to which the evolution of household income and consumption, aggregated in the measure *t*, responds with a certain lag (this is about one year). This happens until 2007–2008, when the global economic recession arrives. Again, the response of the measure *t* to the economic recovery is delayed. The WSh/GDIh and FCEh/GDIh then respond with a similar delay to the 2012–2013 recession and similarly to the 2014–2017 recovery.

Hence, the *t*-values, capturing spatial and temporal changes in household behaviour in aggregate across 30 major economies confirm the well-known and well-described phenomenon of consumption smoothing, i.e. that households tend to stabilise their expenditure even at a certain level of income and gross disposable income, and postpone consumption from periods of higher income to periods when they gain a sense of greater stability and predictability in the economy.

Similarly, we could interpret the evolution of the basic values of the measure t, as can also be seen in Figure 2. Perhaps with a slight difference: the basic values do not reach such extreme values of the peaks and troughs in their evolution, so they are a bit "smoother".

It is also interesting to look at the graphical representation (see Figure 3) of all pairs of values of the measure t given by the matrix in Table 1 in the Appendix. Figure 3 shows that in the first decade of this century, the development of the relationship between the two relative indicators examined (the share of wages in household gross disposable income and the share of household final consumption expenditure in their gross disposable income) in the 30 major countries was quieter than in the second decade. While the first decade was characterised by a fairly calm development of this relationship (this period can be described as a "carpet", see Figure 3), the deep crisis towards the end of the decade (2008–2009) severely disrupted households' behaviour and there was no corresponding calming in the second decade. After a brief recovery in 2010–2011, there was another slowdown in 2012, and soon afterwards, after 2017, the tendency towards a looming recession started to float through the economic space again. Households naturally responded to this with unease, so that the "carpet" quickly became a "mountain range", expressing the increased unease in the economy in the second decade of the new millennium, as is evident in Figure 3.

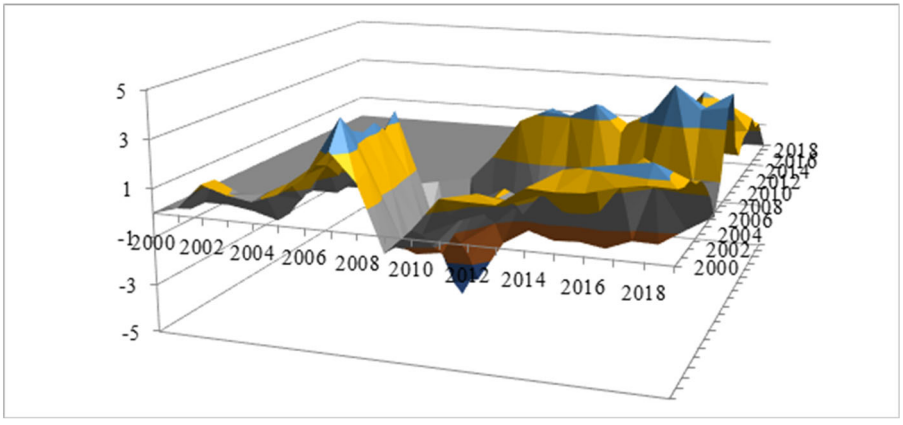


Figure 3. Summary expression of all measures t values given by the matrix in Table 1

Source: https://ec.europa.eu/eurostat/databrowser/view/nasa_10_nf_tr/default/table?lang=fr; own calculations

However, all such considerations were cut short by the arrival of the SARS-CoV-2 epidemic, so – as noted above – it makes sense to include the 2020 covariate in these considerations. It is confirmed that household behaviour is the key to the nature of the economy. It is a sensitive phenomenon, a litmus test of sorts, which we have included in the newly constructed measure t as reported in this paper.

5. Conclusions

The economic dynamics of all developed countries have been volatile in the two most recent decades. Naturally, this fact has also significantly affected the values of the indicators for the household sector, i.e. the proportion of WSh in GDIh and the proportion of FCEh in GDIh. The analysis of the newly constructed measure *t* has shown a decrease (i.e. an approach to the origin of the coordinates in the spatial map of the 30 countries) of these proportions in the years of financial crisis and economic recession and, on the contrary, an increase (i.e. a move away from the origin of the coordinates of the spatial map) of the examined proportions in the years of prosperity (economic growth).

To confirm this assumption, along with the substantive reasoning, we have also used the original measure *t*, which not only quantifies these statements sensitively, but also defines the intensity of the phenomenon (the degree of approach or departure from the origin of the coordinates). The aggregate analysis is then applicable without any limitation in terms of the number of countries (or entire territories) and years studied – the procedure can be applied, for example, to groups of countries according to their economic development, their geopolitical demarcation, etc.

Significant work for the future would of course be to extend the analysis to the crisis caused by the SARS-CoV-2 pandemic. However, this should only be done with some perspective, once there is sufficient certainty about the state of the epidemic in the world and sufficient quality and stability of the necessary data from the Systems of National Accounts. Of course, the current crisis does not primarily have economic causes, but it has strong economic consequences; it has fully exposed the fragility of the world economy. The generalisation of the analysis of household behaviour to the phenomenon of the impact of SARS-CoV-2 on the evolution of GDIh, WSh and FCEh will only be possible with the passage of a few years, when definitive reports for these specific years appear in the national accounts of the world's countries.

However, in such a post-Covid analysis, it should not be forgotten that the years 2017–2019 already signalled a certain tendency towards a slowdown in the world economy. This slowdown was quickly overshadowed by the viral epidemic. Therefore, after it has subsided, it will be necessary to revisit the phenomenon of 2017–2019, at least in part. And here the values of the measures *t* presented in this paper could provide some help to unravel the sensitive reactions and behaviour of households just before the SARS-CoV-2 epidemic and, of course, after it.

Acknowledgement

This paper has been prepared under the Institutional Support for Long-Term and Conceptual Development of Research at Faculty of Informatics and Statistics, Prague University of Economics and Business.

References

Alp, E., Seven, U., (2019). The dynamics of household final consumption: the role of wealth channel. *Central Bank Review*, 19(1), pp. 21–32.

Attanasio, O., Davis, S. J., (1996). Relative Wage Movements and the Distribution of Consumption. *Journal of Political Economy*, 104(6).

Apergis, N., Simo-Kengne, B., Gupta, R., (2014). The long-run relationship between consumption, house prices and stock prices in South Africa: evidence from province-level data. *J. Real Estate Lit*., 22, pp. 83–99.

Campelo, A., Bittencourt, V. S., Malgarini, M., (2020). Consumers Confidence and Households Consumption in Brazil: Evidence from the FGV Survey. *J Bus Cycle Res*, 16, pp. 19–34.

Duesenberry, J. S., (1949). Income, saving, and the theory of consumer behavior, Cambridge, MA: *Harvard University Press*.

European System of Accounts – ESA 1995, (1996). Luxembourg, *Eurostat*.

European System of Accounts – ESA 2010, (2013). Luxembourg, *Eurostat*.

Friedman, M., (1957). A Theory of the Consumption Function. Princeton, NJ: *Princeton University Press*.

Hall, R., (1978). Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence. *Journal of Political Economy*, 86(6), pp. 971–988.

Hamburg, B., Hoffmann, M., Keller, J., (2008). Consumption, wealth and business cycles in Germany. *Empir. Econ.*, 34(3), pp. 451–476.

Heckman, J., (1974). Life Cycle Consumption and Labor Supply: An Explanation of the Relationship between Income and Consumption Over the Life Cycle. *The American Economic Review*, 64(1), pp. 188–194.

Hindls, R., Hronova, S., (2012). Reflection of economic development of selected countries in the structure of final consumption expenditure. *Political Economy*, 60(4), pp. 425–441.

Jappelli, T., Pistaferri, L., (2010). The Consumption Response to Income Changes. *Annu. Rev. Econ.*, 2, pp. 479–506.

Keynes, J. M., (1936). The General Theory of Employment, *Interest and Money*. Macmillan, London.

Kovacs, A., Rondinelli, C., Trucchi, S., (2019). Permanent versus Transitory Income Shocks over the Business Cycle. *IFS Working Paper WP19/29*

Lusardi, A., (1996). Permanent Income, Current Income, and Consumption: Evidence From Two Panel Data Sets. *Journal of Business & Economic Statistics*, 14 (1), pp. 81–90.

Modigliani, F., Brumberg, R., (1954). Utility analysis and the consumption function: An interpretation of cross-section data. In Kurihara, K. K. *Post-Keynesian Economics*.

Sanders, S., (2010). A Model of the Relative Income Hypothesis. *The Journal of Economic Education*, 41(3), pp. 292–305.

System of National Accounts 1993 (SNA 1993), (1993). New York, United Nations, IMF, OECD. *Eurostat*, World Bank.

System of National Accounts 2008 (SNA 2008), (2013). New York, United Nations, IMF, OECD. *Eurostat*, World Bank.

Thurow, L., (1969). The Optimum Lifetime Distribution of Consumption Expenditure. *The American Economic Review*, 59, pp. 324–330.

Appendix

Table 1. Pairwise measures *t* for the years 2000-2019

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
2000	x	0,29826	1,36469	1,03038	0,85586	1,23529	1,94238	3,11858	1,99492	-0,67147	0,36242	1,67676	1,64421	1,63094	2,36837	2,30511	2,36378	2,65459	2,83169	2,79898
2001		x	1,32051	0,87759	0,92312	1,47194	2,25417	4,23289	2,24036	-0,72547	-0,10735	0,72862	1,70486	1,50706	1,88671	2,89976	3,09893	3,26749	3,42782	2,61164
2002			x	-0,16621	-0,44861	0,53632	2,00938	3,48987	1,22078	-1,11714	-0,86830	0,26249	0,66656	1,16887	1,59133	0,76849	1,14005	2,61591	2,14825	2,61344
2003				x	-0,43175	1,01307	1,34989	3,20145	1,31947	-1,62119	-1,45598	0,33844	1,04146	0,43846	1,77443	1,18063	1,90221	2,06993	2,14926	2,21027
2004					x	1,69658	2,06210	3,38423	1,04445	-1,14316	-0,89803	0,53386	1,11093	0,61031	1,10829	0,79988	1,81346	2,79617	2,04110	2,08277
2005						x	0,95426	2,84829	0,86976	-1,87395	-1,65090	-0,44867	0,29986	0,14053	0,61377	0,40104	1,27680	1,39891	1,77101	1,57827
2006							x	3,47029	0,61265	-2,73657	-2,62946	-0,99084	0,03560	-0,31000	0,05859	-0,38180	-0,00629	1,42087	0,90819	0,87869
2007								x	-1,59066	-4,15338	-3,87844	-2,15845	-1,49701	-1,82421	-1,77105	-1,95049	-1,51958	-1,54713	-0,83092	-0,17887
2008									x	-6,44116	-3,02429	-1,52542	-0,45541	-1,07984	0,05274	-0,43731	-0,18733	0,22655	0,80510	1,05173
2009										x	1,15913	2,70307	3,17315	3,07984	3,64697	2,49508	3,21307	4,67435	3,68942	3,85921
2010											x	2,30519	3,11964	2,00272	3,37410	2,68539	3,27242	3,97749	3,58220	4,26108
2011												x	1,93573	1,22865	1,89018	1,12360	1,34657	1,82101	2,48325	1,77755
2012													x	-1,14056	-0,05727	-0,13561	-0,06915	1,07676	1,21494	1,15788
2013														x	0,61248	0,15407	0,65426	1,68082	1,88253	1,70917
2014															x	-0,95459	-0,05808	2,02403	1,53216	1,79170
2015																x	1,88512	3,24154	2,74352	2,16121
2016																	x	2,19472	2,26735	1,54268
2017																		x	1,31332	1,49502
2018																			x	0,73300
2019																				x

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 3, pp. 21–36, DOI 10.2478/stattrans-2022-0040
Received – 26.01.2021; accepted – 14.08.2022

Sampling methods for the concentration parameter and discrete baseline of the Dirichlet Process

Yang Liu[1], **Balgobin Nandram**[2]

ABSTRACT

There are many models in the current statistical literature for making inferences based on samples selected from a finite population. Parametric models may be problematic because statistical inference is sensitive to parametric assumptions. The Dirichlet process (DP) prior is very flexible and determines the complexity of the model. It is indexed by two hyperparameters: the baseline distribution and concentration parameter. We address two distinct problems in the article. Firstly, we review the current sampling methods for the concentration parameter, which use the continuous baseline distribution. We compare three different methods: the adaptive rejection method, the mixture of Gammas method and the grid method. We also propose a new method based on the ratio of uniforms. Secondly, in practice, some survey responses are known to be discrete. If a continuous distribution is adopted as the baseline distribution, the model is misspecified and standard inference may be invalid. We propose a discrete baseline approach to the DP prior and sample the unobserved responses from the finite population both using a Polya urn scheme and a Multinomial distribution. We applied our discrete baseline approach to a Phytophthora data set.

Key words: concentration parameter, discrete baseline, empirical study, grid method, nonparametric Bayesian statistics.

1. Introduction

We often know very little about the specific parametric forms of the distributions, and it is also difficult to validate the parametric assumptions. The parametric Bayesian models, based on distributional assumptions, may be problematic because inferences are sensitive to such assumptions. It may be more appealing to use a nonparametric Bayesian approach. The existence of the Dirichlet Process (DP) was established by Ferguson (1973) and further developed by Blackwell and MacQueen (1973). It is a distribution over distributions, that is, each draw from a DP itself is a distribution (i.e. we are working on functional spaces). In this paper we provide an improved method to sample the concentration parameter and show that it is affected by a discrete baseline.

Another representation of the DP is the generalized Polya urn scheme (Blackwell and MacQueen, 1973). We consider two urns. Urn I is empty and Urn II contains an infinite number of balls, each with a different colour. Pick a ball from Urn II and put it in Urn I. For the next ball, we draw Urn I with probability $\frac{1}{\alpha+1}$. If Urn I is selected, we replace

[1]Worcester Polytechnic Institute, USA. E-mail: yliu22@wpi.edu. ORCID: https://orcid.org/0000-0003-1414-0349.
[2]Worcester Polytechnic Institute, USA. E-mail: balnan@wpi.edu. ORCID: https://orcid.org/0000-0002-3204-0301.

the selected ball with two balls of the same colour, and if Urn II is selected, we take a ball and place it into Urn I. This procedure is repeated until n balls are in Urn I; this is the sample. We observe that with positive probability draws from G (distribution of Urn I) can take the same value regardless of the smoothness of G_0 (distribution of Urn II). That is, G is a discrete distribution with probability one. Current literature have been using smooth functions such as Gaussian distribution as G_0; however this is not always reasonable. This paper will explore a different choice of G_0.

To sample the concentration parameter α of the DP is still an open topic. One can use Gilks' (1992) adaptive rejection sampling method, which relies on the logconcavity of the distribution of the logarithmic transformation of α. Nandram and Yin (2016 a, b) used a grid method to sample α from the posterior density of $\rho = 1/(1+\alpha)$; they have used a noninformative prior for α, different from the proper (informative) prior suggested by Escobar and West (1995). Antonelli, Trippa and Haneuse (2016) reviewed several methods and suggested a more complex method. The problem of sampling the posterior density of α is a difficult one. One of the reasons why it is difficult to estimate α is because it is based on a 'single' observation, k. There are no repeated sampling. So there will be computational instability. There is some research in which the authors set $\alpha = 1$ (e.g. Chaudhuri and Ghosh 2011) to overcome the difficulty in estimating α, thereby leading to an underestimation in variability. In this paper we will propose a new method based on the ratio of uniforms in random sampling.

Another concern that will be addressed is regarding the discreteness of the baseline distribution G_0. It is well known that inference is sensitive to the specification of baseline measure (e.g. McAuliffe, Blei and Jordan 2006 and Nandram and Yin 2016 a). So it is more robust if we have an unspecified distribution G_0. Camerlenghi et al. (2019) discussed ties across samples at the observed or latent level. In the discrete case we mention here, an observation can look like a tie, but it may not be. We are not actually talking about ties, although it is a part of what we are doing. The discreteness of G_0 means that the same value can come from either G_0 or from the balls already drawn in the Polya urn scheme. But it is mandatory to have G_0 discrete in this model if we have a strong belief that the observations are from a discrete family. In such a case, the number of distinct values in the sample, k, is no longer a sufficient statistic for α. This paper will correct this.

We demonstrate our discrete baseline approach to Phytophthora epidemics in bell pepper. The pathogen Phytophthora Capsici Leonian is a severe infectious disease and could rapidly cause death of the plant (Gumpertz 1997). Disease presence or absence was recorded for each cell in a 20×20 quadrats study field. We group the quadrats and count the number of diseased plants in each group so we know the response is guaranteed to be discrete. Now our goal is to obtain an estimator of the finite population mean provided by a nonparametric approach. It is apparent that this approach is more robust than the parametric models such as those based on normality. On the other hand, current nonparametric methods are all based on continuous baseline distribution (i.e. normal baseline). Our approach, with relaxation to the baseline distribution, gives a more realistic estimator when we know the response is discrete.

This paper is an extension of Nandram and Yin (2016 a,b), who studied the sensitivity of the baseline distribution to the finite population mean. They proposed the DP approach to

predict the nonsampled observations by using the Polya urn scheme. We choose a discrete baseline to the DP when the response is known to be from a discrete family. When G_0 is discrete, the number of distinct values in the sample is no longer a sufficient statistic of the concentration parameter α. We proposed a way to correct this by adding a latent variable to indicate which urn the observation is from.

When faced with a discrete baseline, researchers might resort to a DP-based mixture model (DPM) involving a continuous kernel density, however, this is not what we are trying to discuss here. Note that DPMs are often miscalled as "mixture of Dirichlet process model" (Neal 2000). There have been many computational methods to run the model over the past two decades (e.g. Escobar and West 1995, Neal 2000 and Kalli, Griffin and Walker 2011). The DPM is not appropriate in some applications like the example we discuss in this paper because we do not have well defined groups of data. For the DPM, we need different groups of data with different parameters and then a DP is assigned to these parameters. Of course, in applications the DPM is the workhorse of nonparametric Bayesian statistics, yet we need to solve the problem associated with discrete baseline distributions as they may be included as a step in a hierarchical Bayesian model.

The plan of this paper is as follows. In Section 2, we briefly review the Dirichlet process (DP), and different sampling algorithms for α, the concentration parameter. We also introduce our approach, the ratio of uniforms algorithm, and a simulation study to compare the different methods. In Section 3, we discuss one limitation that current literature has regarding the baseline distribution of the DP and how we resolve it. We also discuss the implementation of our method to the finite population mean. In Section 4, we discuss an illustrative example on Phytophthora data. We conclude this paper in Section 5. An appendix has technical details.

2. Dirichlet Process and Sampling the Concentration Parameter

In Section 2.1, we give a brief review of the Dirichlet process, and in Section 2.2, we review current methods to sample the posterior density of α. In Section 2.3 we present our new method based on the ratio of uniforms. In Section 2.4, we provide a small simulation study to compare our new method with few selected ones that we review.

2.1. Review of the Dirichlet Process

Let $(\Theta, \mathscr{B})$ be a measurable space, with G_0 the baseline measure (nonrandom) on the space, and let α be the concentration parameter. A Dirichlet process, $\mathrm{DP}(\alpha, G_0)$, is defined as the distribution of a random probability measure G over $(\Theta, \mathscr{B})$ such that, for any finite measurable partition of the measurable space, $(\Theta, \{A_i\}_{i=1}^n)$, with $A_i \bigcap A_j = \phi$, $\bigcup_{i=1}^n A_i = \Theta$,

$$\{G(A_1), \cdots, G(A_n)\} \sim \mathrm{Dirichlet}\{\alpha G_0(A_1), \cdots, \alpha G_0(A_n)\}.$$

We write $G \sim \mathrm{DP}(\alpha, G_0)$, if G is a random probability measure with a distribution given by the DP. For any measurable set, A, we have $\mathrm{E}[G(A)] = G_0(A)$, that is the mean of the DP is the baseline distribution G_0 and $\mathrm{Var}[G(A)] = G_0(A)[1 - G_0(A)]/(\alpha + 1)$. The larger α is,

the smaller the variance (i.e. the DP concentrates more of its mass around the baseline distribution). Here, G_0 and α are both parameters and they play intuitive roles in the definition of the DP. Here, G is constrained to be around G_0 and this is regulated by α.

Let $G \sim \text{DP}(\alpha, G_0)$ and $y_1, \cdots, y_n$ be a sequence of independent draws from G. The posterior distribution, $G|y_1, \cdots, y_n$ is

$$\text{DP}\left(\alpha+n, \frac{\alpha}{\alpha+n}G_0 + \frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{y_i}\right),$$

where δ_{y_i} is the cdf of a point mass at y_i. This conjugate property of the DP was motivated by Ferguson (1973), desirable for easy algebra and computations.

For a one-sample problem, one might take

$$Y_1, \cdots, Y_n|G \sim G,$$

$$G \sim DP(\alpha, G_0),$$

where G_0 is the baseline measure and α the concentration parameter. Assuming that there are k distinct values among $Y_1, \cdots, Y_n$, the baseline model is $Y_1^*, \cdots, Y_k^*|k \sim G_0$. Note that k is a random variable. The baseline measure G_0 is assumed continuous. Binder (1982) was the first to introduce this model to survey sampling; more recently, see Nandram and Yin (2016 a,b). Although G_0 can be discrete, it appears that this latter case was not discussed by Antoniak (1974).

Antoniak (1974) wrote down the distribution of the number of distinct values k given α and he proved that k is a sufficient statistic for α where G_0 is continuous. It is easy to write down the posterior density with an appropriate prior. The sampling methods being discussed in this section are all based on continuous baseline. We write here that

$$p(k|\alpha) = C \cdot \frac{\Gamma(\alpha)\alpha^k}{\Gamma(\alpha+n)}, \;\; k = 1, \cdots, n,$$

where C is a constant.

However, if G_0 is discrete, k is no longer a sufficient statistic; this result appears to be not so well known. Therefore, if the result is used, this is a violation of the sufficiency principle; we will discuss this issue in Section 3.

2.2. Current Sampling Methods

In this section, we will discuss three current sampling methods for the concentration parameter: the adaptive reject sampling method (ARS), the mixture of Gamma method and the grid method.

We first review the ARS method (Gilks 1992).

Theorem. Let $\phi = \ln(\alpha)$, where α is the concentration parameter. With a logconcave prior $\pi(\phi)$, the posterior density $\pi(\phi|k)$ is logconcave, (i.e. strongly unimodal with a unique mode).

Rasmussen (2000) first demonstrated the logconcavity of $\pi(\phi|k)$ but here we provide

our own proof in the appendix. More generally, we show that if prior $\pi(\phi)$ is logconcave, i.e. $\frac{d^2\ln(\pi(\phi))}{d\phi^2} < 0$, then the posterior density on the transformed scale is logconcave. We mention two useful priors in the appendix when $\phi = \ln(\alpha)$. The shrinkage prior, $f(2,2)$ distribution, is

$$\pi(\alpha) = \frac{1}{(1+\alpha)^2}, \alpha > 0.$$

Another example, the half-Cauchy prior, is

$$\pi(\alpha) = \frac{2}{\pi(1+\alpha^2)}, \alpha > 0.$$

Knowing that $\pi(\phi|k)$ is logconcave, we can use the adaptive reject sampling method (Gilks 1992) to draw ϕ. This sampling procedure was performed with the R package `ars`. Then we can compute α in the form $\alpha = e^{\phi}$. There is limitation to the ARS method due to tail problem, i.e. the sampling distribution for the two tails of the distribution is not accurate and this can be seen in the simulation section.

Nandram and Choi (2004) discussed the use of the gamma prior, which was introduced earlier by Escobar and West (1995). One concern is that the mix of Gamma method gives bimodal sampling distribution whereas a unimodal density of α is preferred. Another problem is that it requires informative Gamma prior and this remains to be validated.

Nandram and Yin (2016) transformed α according to $\rho = \frac{1}{1+\alpha}$, this is actually the correlation, $\text{Cor}(y_i, y_j)$, $i \neq j$, in the DP. The posterior density of ρ is

$$\pi(\rho|k) \propto \frac{(1-\rho)^{k-1}\rho^{n-k}}{\prod_{j=1}^{n-1}(1-\rho+\rho j)}, \ 0 \leq \rho \leq 1.$$

Note that $\pi(\rho|k)$ is well defined on $[0,1]$. However, we see that it is not in a simple form and a one-dimensional grid method was used to draw samples from it, thereby avoiding Markov chain Monte Carlo methods (e.g. Metropolis - Hastings sampler). The unit interval is simply divided into 100 sub-intervals of equal width, and the joint posterior density is approximated by a discrete distribution with probabilities proportional to the heights of the continuous distribution at the mid-points of these sub-intervals. Now, it is easy to draw a sample from this univariate discrete distribution of $\pi(\rho|k)$; the discreteness is removed by jittering. Nonetheless, there is a drawback of this method, because it may not perform well when ρ has substantial probability near 0 or 1.

2.3. Ratio of Uniforms Method

Liu and Nandram (2020) proposed to use the ratio of uniforms method to obtain posterior samples of α. Originally introduced by Kingderman and Monahan (1977), a point is generated uniformly over a certain region in the plane.

To achieve this, independent uniform random variables are simulated,

$$U, V \sim \text{Uniform}(0,1)$$

and those that fall outside some set are discarded. The ratio V/U is then calculated for those points inside the set. The ratio values obtained are used as observations from the required distribution.

There are other priors that can be used but here for illustration purpose, we use the posterior distribution of α with a noninformative prior, $\pi(\alpha) = \frac{1}{(1+\alpha)^2}$,

$$h(\alpha) = \pi(\alpha|k) \propto \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha+n)(\alpha+1)^2}, \ \ \alpha > 0.$$

A half Cauchy prior can be used for the prior of α, but there is very little difference between the two when transformed to $[0,1]$ a posteriori. This method can proceed using the following algorithm.

1. Generate u and v independently from $U(0,b)$ and $U(c,d)$.
2. Set $\alpha = v/u$ if $u^2 \leq h(v/u)$ and return to (i) otherwise.

Here, b, c and d are given by

$$b = \sup_{\alpha} \sqrt{h(\alpha)}, \quad c = -\sup_{\alpha} \alpha\sqrt{h(\alpha)}, \quad d = \sup_{\alpha} \alpha\sqrt{h(\alpha)}.$$

Because α is positive, we set $c = 0$. This algorithm is very easy to implement and it is very efficient to get samples.

2.4. Simulation Study

It is convenient to compare different sampling methods using simulations because we can obtain the true distribution of α and compare the theoretical values with the sampled values. Firstly we find the theoretical percentiles of α using fine grids of width 0.0025. Then we perform the four sampling methods to get 10,000 sample points. We can find the sample percentiles by ordering the sample values and find corresponding quantiles as the theoretical values. Lastly, we compare the theoretical value versus the sampled value using a quantile-quantile plot.

In order to compare the four sampling methods, we take the sample size $n = 12$, 25, 100 and the number of distinct values k to be roughly equal to $\ln n$, with $k = 2$, 3 and 5 respectively. We choose a common prior, the shrinkage prior,

$$\pi(\alpha) = \frac{1}{(1+\alpha)^2}, \alpha > 0.$$

to be used for all four sampling methods.

Results are shown in Figures 1, 2 and 3. All four methods provide reasonable sampling distributions for α. However, the ratio of uniforms method is most accurate. In all these figures, the points of ratio of uniforms method fall on almost a 45 degree straight line through the origin and there is some problem with the other plots at various places (e.g. not fitting exactly on the 45 degree straight line through the origin). As we mentioned in Section 2, the ARS and grid method have tail problems and mixture of gamma uses an informative prior

which remains to be validated. Our method does not require informative gamma prior and it is easy to implement. So we recommend using ratio of uniforms to get random samples of α.

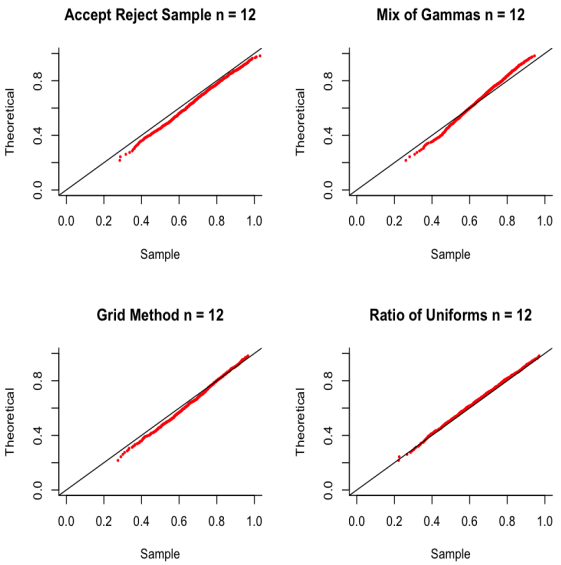


Figure 1: Comparison for the posterior distributions of the concentration parameter using the four sampling methods (n= 12, k = 2)

3. Discrete Baseline

Current literature on DP has been using continuous baseline distributions, see Antonelli, Trippa and Haneuse (2016). Teh, Jordan, Beal and Blei (2006) developed a hierarchical Dirichlet process model with a discrete baseline distribution. Apparently, they were not aware of the problem with the discrete baseline distribution when sampling the concentration parameter and they inadvertently attempted to "sweep the problem under the rug."

Here, we explore a possibility of using a discrete baseline. One problem is that the distinct values in the sample are no longer the true distinct ones because of discrete baseline. We allow observing a "new" value from the baseline distribution that is the same as one that is already in the sample. To solve this problem, we introduce latent variables,

$$Z_i = \begin{cases} 1, & \text{if a draw is made from the baseline,} \\ 0, & \text{if a draw is from the value that is already observed} \end{cases}$$

with $Z_i \overset{ind}{\sim} \text{Ber}(\frac{\alpha}{\alpha+i-1})$, $i = 1, \cdots, n$. Therefore, the true number of distinct values is $k = \sum_{i=1}^{n} z_i$.

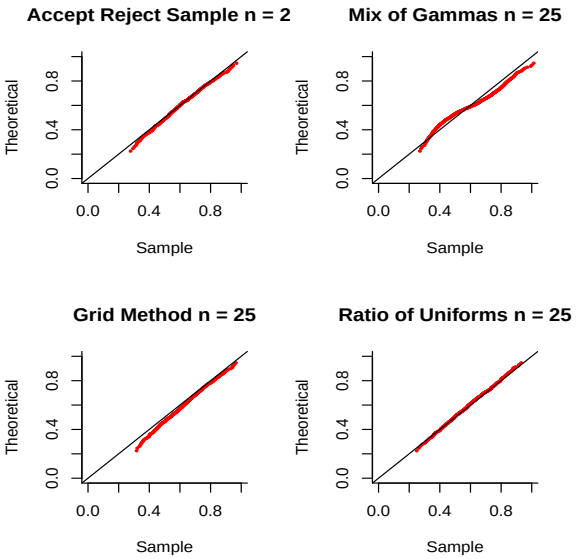


Figure 2: Comparison for the posterior distributions of the concentration parameter using the four sampling methods (n= 25, k = 3)

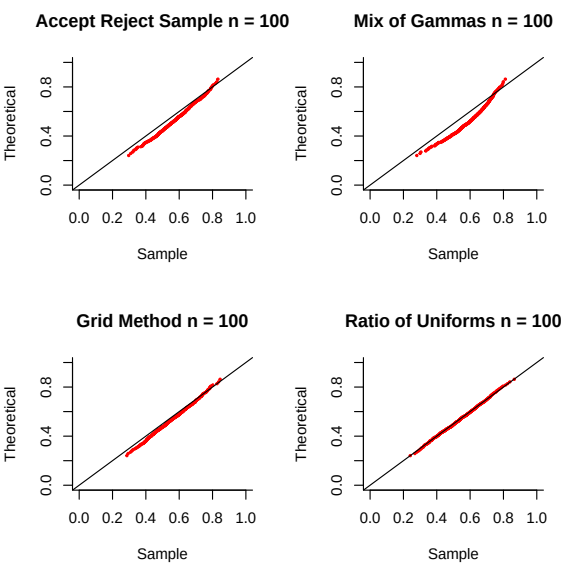


Figure 3: Comparison for the posterior distributions of the concentration parameter using the four sampling methods (n= 100, k = 5)

Our goal is to predict the finite population proportion for a given area based on a random sample from it. This could be applied to many areas of study, for example we want to predict the infectious rate (like a proportion where the denominator is fixed) of a given farmland for some disease and it is not feasible to observe all the plants on the farm, however, we could take a random sample and estimate the posterior mean using this sample. We have observed n of them and want to make predictions to the $N-n$ individuals. We consider three cases.

Case 1. We use the one-level DP model for the population values to make inference for a finite population mean. For this case, the baseline distribution is chosen to be normal. We assume that

$$\begin{aligned} y_1 \cdots, y_N | G &\sim G, \\ G &\sim DP(\alpha, G_0), \\ G_0 &\sim N(\mu, \sigma^2), \\ \pi(\mu, \sigma^2) &\propto \frac{1}{\sigma^2}, \\ \pi(\alpha) &= \frac{1}{(1+\alpha)^2}. \end{aligned}$$

Here, we observe the number of distinct values k and then sample α as discussed in Section 2. For each sampled α value, we predict the unobserved $Y_{n+1}, \cdots, Y_N$ using the Polya urn scheme,

$$Y_{n+i+1} | y_1, \cdots, y_n, y_{n+1}, \cdots, y_{n+i} \sim \frac{\alpha}{\alpha+n+i} G_0 + \frac{n+i}{\alpha+n+i} \sum_{j=1}^{n+i} \delta_{y_j},$$

for $i = 1, \cdots, N-n-1$, (Nandram and Yin 2016 a, b). So it is easy to draw the nonsampled values one by one using the Polya urn scheme.

Case 2a. We correct the true number of observations from the baseline distribution $\sum_{i=1}^{n} z_i$, where $z_i = 1$ when obervation i is a distinct value from G_0. The discrete model is

$$\begin{aligned} y_i | G &\overset{ind}{\sim} G, \; i = 1, \cdots, n, \\ G | p, \alpha &\sim \mathrm{DP}(\alpha, \mathrm{Bin}(m, p)), \\ z_i | \alpha, p &\overset{ind}{\sim} \mathrm{Ber}(\frac{\alpha}{\alpha+i-1}), \\ \pi(\alpha) &= \frac{1}{(1+\alpha)^2}, \pi(p) = 1. \end{aligned}$$

Where n is the sample size and m is the predefined number of the total trials in the Binomial distribution. The joint posterior density can be written as

$$\begin{aligned} \pi(z, p, \alpha | y) \propto & \frac{1}{(1+\alpha)^2} \times \\ & \prod_{i=1}^{n} [p^{y_i}(1-p)^{m-y_i}]^{z_i} [p^{y_i}(1-p)^{m-y_i}]^{1-z_i} \cdot \left[\frac{\alpha}{\alpha+i-1}\right]^{z_i} \left[\frac{1}{\alpha+i-1}\right]^{1-z_i}. \end{aligned}$$

We obtain the conditional distribution of the Gibbs sampler

$$z_i|\alpha,p,y \overset{ind}{\sim} \text{Ber}(Q_i),$$

where

$$Q_i = \frac{\frac{\alpha}{\alpha+i-1}p^{y_i}(1-p)^{m-y_i}}{\frac{\alpha}{\alpha+i-1}p^{y_i}(1-p)^{m-y_i}+\frac{i-1}{\alpha+i-1}p^{y_i}(1-p)^{m-y_i}} = \frac{\alpha}{\alpha+i-1},\ \ i=1,\cdots,n,$$

$$p|z,\alpha,y \sim \text{Beta}(\sum_{\{z_i=1\}} y_i+1,\ \sum_{\{z_i=1\}}(m-y_i)+1),$$

$$\pi(\alpha|z,p) \propto \frac{\alpha^{\sum_{i=1}^n z_i}\Gamma(\alpha)}{\Gamma(\alpha+n)(\alpha+1)^2}.$$

Here, α is drawn from its conditional posterior distribution using our preferred ratio of uniforms method. For each sampled α, we predict one set of unobserved values and compute the finite population mean using the Polya urn scheme. We used a burn in of 1000 and thinning of 10 to get a sample of 10,000. We diagnosed the Gibbs sampler after the chain is run. For the data example we use in this paper, the diagnostic result shows that the effective sample sizes are 4537, 5042 and 4978 for α, p and $\sum_{i=1}^n z_i$ respectively. P-values from the Geweke's tests are 0.384, 0.533 and 0.628 respectively. So at this setting the Gibbs sampler is mixing well. It took about 22 seconds to obtain 10,000 sample values on our computer (see Section 2).

Case 2b. It would be interesting to see the difference of the prediction between a Polya urn scheme and a stick breaking procedure with the idea borrowed from Sethuraman (1994), Ishwaran and James (2001), Kalli, Griffin and Walker (2011). Using the model in Case 2a, but suppose we have already observed $y_1^*,\cdots,y_d^*$, d distinct values ($1 \le d \le n$), with $n_1 \ge n_2,\cdots,\ge n_d$ being their corresponding counts. Here, we allow some values to be unobserved. Now we want to predict $N_1-n_1,\cdots,N_d-n_d$, for convenience, we write $N_1^*,\cdots,N_d^*$. Let $N^* = N-n$, so that we know $N^* = \sum_{i=1}^d N_i^*$. Now

$$N_1^*,\cdots,N_d^* \sim \text{Multinomial}\Big\{N^*,(w_1,\cdots,w_d)\Big\},$$

where $w_1,\cdots,w_d$ are the weights in stick-breaking algorithm with $\sum_{s=1}^\infty w_s = 1$, $w_1 = \nu_1$, $w_2 = \nu_2(1-\nu_1),\cdots,w_{d-1} = \nu_{d-1}\prod_{i=1}^{d-2}(1-\nu_i)$, $w_d = \prod_{i=1}^{d-1}(1-\nu_i)$, and

$$\nu_i \overset{iid}{\sim} \text{Beta}(1,\alpha).$$

Given α from the Gibbs sampler, we can draw ν_i and thus draw the predicted values from a Multinomial distribution.

With the posterior samples of α from the Gibbs sampler in Case 2a, the conditional

posterior distribution of ν is

$$\begin{aligned}\pi(\nu|\alpha,d) \propto\ & \nu_1^{n_1}[\nu_2(1-\nu_1)]^{n_2}\cdots[\nu_{d-1}(1-\nu_1)\cdots(1-\nu_{d-2})]^{n_{d-1}} \\ & [(1-\nu_1)\cdots(1-\nu_{d-2})(1-\nu_{d-1})]^{n_d}\times\prod_{i=1}^{d}(1-\nu_i)^{\alpha-1} \\ \propto\ & \nu_1^{n_1}(1-\nu_1)^{n_2+\cdots+n_d+\alpha-1}\nu_2^{n_2}(1-\nu_2)^{n_3+\cdots+n_d+\alpha-1}\cdots\nu_{d-1}^{n_{d-1}}(1-\nu_{d-1})^{n_d+\alpha-1}.\end{aligned}$$

Therefore,

$$\pi(\nu_i|\alpha,d) \overset{ind}{\sim} \text{Beta}(n_i+1, \sum_{j=i+1}^{d} n_j+\alpha).$$

Once samples of ν_i are obtained, we can predict the unobserved response values from a Multinomial distribution discussed above.

We will implement the three cases we discussed in the following example.

4. Real Data Analysis

The data we present here are about Phytophthora Epidemic in Bell Pepper from Gumpertz (1997). The pathogen Phytophthora Capsici Leonian causes lesions on the crown, stem, and leaves of bell pepper, and rapidly causes the plant to die. For their analyses, they took one field which was a square lattice of 20×20 quadrats with 2 to 3 bell pepper plants per quadrat as an example. The response variable within each quadrat was presence or absence of disease in a quadrat. If any plant was wilted, dead, or had lesions on stem, crown, or leaves, disease was considered to be present in the quadrat. Disease presence or absence was recorded for each quadrat on nine dates throughout the growing season, from 6/16/92 to 8/5/92. Figure 2 shows the disease incidence on 6/25/92.

We want to make this data set usable to mimic our discrete response scenario so we perform the following sampling procedure. We divide each row of the field by every fifth quadrats and then we take one random sample within each row of the field. We assume that the sampled value follows a binomial distribution with total number of trials being 5. Now, our goal is to predict the unobserved quadrats and estimate the infectious rate, which is really a finite population proportion (mean) in this application. We performed the estimation using both discrete baseline and continuous baseline approaches, as discussed in Section 3.2.

We report the posterior means (PM), posterior standard deviations (PSD) and the credible intervals (CI) in Table 1. Given the true infectious rate of 0.1525, we found that the continuous (Normal) baseline distribution provides an unbiased estimation to the infectious rate. However, the lower end of 95% credible interval is negative. This is because the posterior sample is taken over the whole real line as a nature of the Gaussian distribution. We know in reality, the infectious rate is a probability and should always be positive. Here, we naively use normal baseline because this is often chosen to be the G_0 in many practices. But for obvious reasons we now want to avoid it. PMs are roughly the same for Case 1 and Case 2a but Case 2b has some bias, with a larger estimation, however, the PSD is the smallest

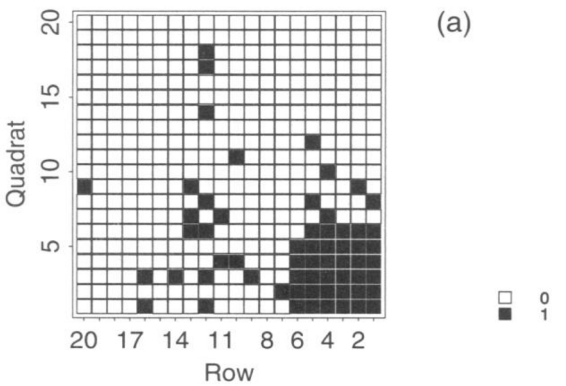


Figure 4: Map of Disease Incidence.

for Case 2b. Binomial baseline (Case 2a) has slightly more variability because it considers the uncertainty of which urn a new observation is drawn. Last but not least, the prediction using a Multinomial approach (Case 2b) can significantly reduce the standard deviation and it provides a realistic result because it is based on a discrete distribution. Figure 5 shows the posterior distribution of the three estimates of the finite population mean. The plots are similar for Cases 2a and 2b but Case 2a is more spread out. Note that the plots are in the same range (easier for us to visualize and compare). Similar to the table, the Multinomial approach gives most concentrated plot. The Normal approach exceeds zero to the negative side.

Table 1. Estimation of the Infectious Rate (True Rate*: 0.1525)

Baseline Distribution	PM	PSD	95% CI
Case 1. Normal (μ, σ^2)	0.1551	0.1728	(-0.2221,0.4735)
Case 2a. Binomial (n^{**}, p)	0.1588	0.219052	(0,0.6800)
Case 2b. Multinomial (N^{**}, w)	0.1698	0.0661	(0.0667,0.3267)

PM = Posterior Mean; PSD = Posterior Standard Deviation; CI = Credible Interval. * We can compute the true rate with data from the whole 20×20 study site. ** In our case $n = 5$ and $N = 80 - 20 = 60$.

5. Concluding Remarks

We have proposed a new sampling method for the standard concentration parameter of the Dirichlet Process and compared it with three methods. The Ratio of Uniforms is more accurate and it is faster considering the computational time. In the meantime, we pointed out a problem that current researchers have ignored regarding the baseline distribution of the DP. We have corrected the true number of distinct values in the sample by introducing a latent variable which indicated which urn a new observation is from. By using this approach,

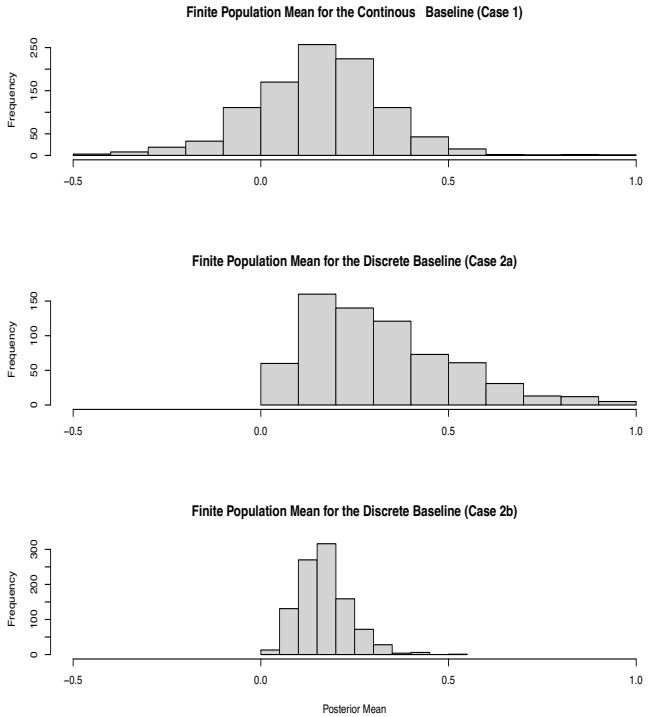


Figure 5: Posterior Distributions of the Finite Population Mean (Proportion) for the Three Cases

we are able to give a more accurate estimation of the finite population mean when the observations are discrete. We used a Phytophthora example to illustrate our approach. We concluded the discrete baseline method is more reasonable.

There are two directions we could proceed to extend our current work. First, we might consider a spatial model for the example provided in this paper. However, it is not the purpose of our paper to provide a complete analysis of these data.

Second, we could extend the one-level DP model to a two-level DP model, where there are groups naturally occur in the data. The two-level model is the Dirichlet process mixture (DPM) model with a DP on the second level. Recently, Yin and Nandram (2020 a,b) placed the DP on the first level but not on the second. They claimed that their approach is good for data with gaps, outliers and ties.

Third, the work we have done in this paper also inspired us to study sensitivity to the baseline. We may give a very weak assumption to the baseline, i.e. either logconcave or unimodal. These can be discretized nicely. For a logconcave density the slopes of the tangent lines decrease all the way or the chords joining any two points will have non-increasing slopes all the way from left to right on the real line. Also, a unimodal density has heights increasing to the mode and then decreasing (i.e. the cumulative distribution function is first

convex up to the mode and concave after the mode). So, essentially, we can use a discrete baseline distribution in the DPM.

Therefore, our work on discrete baseline distribution is an important start. However, although we have a good algorithm for the concentration parameter of the Dirichlet process, based on the ratio of uniforms, some improvement may be possible.

References

Antoniak, C. E., (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2 (6), pp. 1152–1174.

Antonelli, J., Trippa, L. and Haneuse, S., (2016). Mitigating bias in generalized linear mixed models: The case for Bayesian nonparametrics. *Statistical Science*, 31 (1), pp. 80–95.

Binder, D. A., (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society*, Series B, 44(3), pp. 388–393.

Blackwell, D., MacQueen, J. B., (1973). Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1 (2), pp. 353–355.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prunster, I. and Rodrigue,A., (2019). Latent nested nonparametric priors. *Bayesian Analysis*, 14, pp. 1303–1356.

Chaudhuri, S., Gosh, M., (2011). Empirical likelihood for small area estimation. *Biometrika*, 98, 2, pp. 473–480.

Escobar, M. D., West, M., (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90 (430), pp. 577–588.

Ferguson, T. S., (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1 (2), pp. 209–230.

Ishwaran, H., James, L. F., (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96 (453), pp. 161–173.

Gumpertz, M. L., Graham, J. M. and Ristaino, J. B., (1997). Autologistic Model of spatial pattern of Phytophthora Epidemic in bell pepper: Effects of Soil Variables on Disease Presence. *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 2, No. 2, pp. 131–156.

Kalli, M., Griffin, J. E. and Walker, S. G., (2011). Slice sampling mixture models. *Statistics and Computing*, 21 (1), pp. 83–105.

Kinderman, A. J., Monahan J. F., (1977). Computer generation of random variables using the ratio of uniform deviates. *Association for Computing and Machinery, Inc.*

Liu, Y., Nandram, B., (2020). Sampling methods for the concentration parameter of the Dirichlet process. In *JSM Proceedings, Nonparamtric Section.* Alexandria, VA: American Statistical Association. pp. 1121–1131.

Nandram, B., Choi, J. W., (2004). Nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse. *Journal of Nonparametric Statistics*, 16 (6), pp. 821–839.

Nandram, B., Yin, J., (2016a). Bayesian predictive inference under a Dirichlet process with sensitivity to the normal baseline. *Statistical Methodology*, 28, pp. 1–17.

Nandram, B., Yin, J., (2016b). A nonparametric Bayesian prediction interval for a finite population mean. *Journal of Statistical Computation and Simulation*, 86 (16), pp. 3141–3157.

Neal, R. M., (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9 (2), 249–265.

Rasmussen, C. E., (2000). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems* pp. 554–560.

Sethuraman, J., (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, pp. 639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M., (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101 (476), pp. 1566–1581.

Yin, J., Nandram, B., (2020a). A Bayesian small area model with Dirichlet processes on responses. *Statistics in Transition, New Series*, 21 (3), pp. 1–19.

Yin, J., Nandram, B., (2020b). A Nonparametric Bayesian Analysis of Response Data with Gaps, Outliers and Ties. *Statistics and Applications, New Series*, 18 (2), pp. 121–141.

Appendix

Logconcavity of the Posterior Density of α

Proof: It is easy to show that the likelihood function for α is

$$p(\alpha \mid k) \propto \frac{\alpha^k}{\prod_{j=1}^{n-1}(j+\alpha)}, \alpha > 0,$$

where k is the number of distinct values for a continuous baseline. For any prior $\pi(\alpha)$, using Bayes' theorem, the posterior density of α is

$$\pi(\alpha \mid k) \propto p(\alpha \mid k)\pi(\alpha).$$

If we make the transformation $\phi = \log(\alpha)$, $p(\alpha \mid k)$ will transform to $p_1(\phi \mid k) = \frac{e^{k\phi}}{\prod_{j=1}^{n-1}(j+e^{\phi})}$ and $\pi(\alpha)$ will transform to $\pi_1(\phi)$ and the Jacobian is e^{ϕ}. We show that if $\pi_1(\phi)$ is logconcave, i.e. $\frac{d^2\ln(\pi_1(\phi))}{d\phi^2} < 0$, then the posterior density on the transformed scale is logconcave. Let

$$\Delta(\phi) = (k+1)\phi - \sum_{j=1}^{n-1} \ln(j+e^{\phi}) + \ln(\pi_1(\phi)).$$

Then,

$$\frac{d\Delta(\phi)}{d\phi} = (k+1) - \sum_{j=1}^{n-1} \frac{e^{\phi}}{j+e^{\phi}} + \frac{d\ln(\pi_1(\phi))}{d\phi}$$

and

$$\frac{d^2\Delta(\phi)}{d\phi^2} = -\sum_{j=1}^{n-1} \frac{je^{\phi}}{(j+e^{\phi})^2} + \frac{d^2\ln(\pi_1(\phi))}{d\phi^2} < 0.$$

Therefore, under the assumption of logconcavity for $\pi_1(\phi)$, the posterior density of α is logconcave.
We mention two useful priors when $\phi = \ln(\alpha)$. The shrinkage prior, $f(2,2)$ distribution, is

$$\pi(\alpha) = \frac{1}{(1+\alpha)^2}, \alpha > 0.$$

Another example, the half-Cauchy prior, is

$$\pi(\alpha) = \frac{2}{\pi(1+\alpha^2)}, \alpha > 0.$$

Both priors after making the transformation $\phi = \ln(\alpha)$ are logconcave.

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 37–58, DOI 10.2478/stattrans-2022-0041
Received – 21.01.2021; accepted – 27.04.2022

Parameter estimation of exponentiated exponential distribution under selective ranked set sampling

Amal S. Hassan[1], **Rasha S. Elshaarawy**[2], **Heba F. Nagy**[3]

ABSTRACT

Partial ranked set sampling (PRSS) is a cost-effective sampling method. It is a combination of simple random sample (SRS) and ranked set sampling (RSS) designs. The PRSS method allows flexibility for the experimenter in selecting the sample when it is either difficult to rank the units within each set with full confidence or when experimental units are not available. In this article, we introduce and define the likelihood function of any probability distribution under the PRSS scheme. The performance of the maximum likelihood estimators is examined when the available data are assumed to have an exponentiated exponential (EE) distribution via some selective RSS schemes as well as SRS. The suggested ranked schemes include the PRSS, RSS, neoteric RSS (NRSS), and extreme RSS (ERSS). An intensive simulation study was conducted to compare and explore the behaviour of the proposed estimators. The study demonstrated that the maximum likelihood estimators via PRSS, NRSS, ERSS, and RSS schemes are more efficient than the corresponding estimators under SRS. A real data set is presented for illustrative purposes.

Key words: exponentiated exponential distribution, partial ranked set sampling, neoteric ranked set sampling, maximum likelihood method.

Mathematical Subject Classification: 62F10

1. Introduction

In many studies where sampling is used, such as environmental management, ecology, sociology, and agriculture, exact measurement of a selected unit is either difficult or costly and time-consuming. However, the ranking of a small set of selected units can be carried out easily either by visual inspection with respect to the study

[1] Department of Mathematical Statistics, Cairo University, Faculty of Graduate Studies for Statistical Research, Egypt. E-mail: amal52_soliman@cu.edu.eg. ORCID: https://orcid.org/0000-0003-4442-8458.

[2] Department of Mathematical Statistics, Cairo University, Faculty of Graduate Studies for Statistical Research, Egypt. E-mail: rashasaber123@hotmail.com. ORCID: https://orcid.org/0000-0001-7414-3950.

[3] Corresponding Author. Department of Mathematical Statistics, Cairo University, Faculty of Graduate Studies for Statistical Research, Egypt. E-mail: heba_nagy_84@cu.edu.eg. ORCID: https://orcid.org/0000-0003-0262-205X.

variable or on the basis of an auxiliary variable. The RSS scheme was first proposed by McIntyre (1952) to obtain a sample from a population in his study for estimating the yield of pastures. The RSS scheme outweighs the conventionally used SRS scheme in terms of the superior efficiency of the RSS estimators of population mean and variance (see Wolfe (2010)). Several studies have shown that the calculated estimators based on RSS are more efficient than their counterparts in SRS. For example, Bhoj and Ahsanullah (1996) used the RSS scheme to estimate the generalized geometric distribution parameters. Al-Odat and Al-Saleh (2001) considered estimation of the population mean using a variation of the RSS procedure. Mahdizadeh and Arghami (2010) discussed entropy estimation in RSS design and compared the results with those in SRS design. Hassan (2013) obtained a Bayesian estimator for the shape and scale parameters of the EE distribution using RSS. Abu-Dayyeh et al. (2013) used RSS to estimate the shape and scale parameters of the Pareto distribution. Samuh and Qtait (2015) used median RSS (MRSS) to estimate the shape and scale parameters of the EE distribution. Tahmasebi et al. (2017) provided Bayesian estimation for Rayleigh distribution based on SRS, RSS, and maximum RSS procedures with unequal samples in two cases: one cycle and *r*-cycles. Bantan et al. (2020) derived Zubair Lomax distribution parameter estimators under the RSS scheme. Al-Omari et al. (2020) considered stress-strength reliability estimator of the exponentiated Pareto model using MRSS and RSS designs. Almarashi et al. (2021) studied stress-strength reliability estimator for the Topp–Leone distribution using advanced sampling methods. Hassan et al. (2022) considered estimating system reliability using NRSS and MRSS data for generalized exponential distribution.

Some variations of the RSS scheme were proposed by several authors. The PRSS requires fewer sampling units and less ranking than the RSS and proves to be more efficient than the SRS (see Haq et al. (2013)). In the PRSS scheme, the experimenter selects (A) sample units using SRS and (B) sample units using RSS, producing a final sample of size M=A+B units. Thus, it requires fewer sampling units and fewer rankings than the RSS. The ERSS design has been suggested by Samawi et al. (1996) for estimating the population mean. Studies based on the ERSS scheme have been studied by several authors (see, for example, Hassan (2012), Hassan et al. (2014), (2015)). The NRSS scheme was suggested by Zamanzade and Al-Omari (2016) and it differs from the original RSS scheme by the composition of a single set of n^2 units instead of n sets of size n. This strategy has been shown to be effective, producing more efficient estimators for the population mean and variance than the SRS and RSS schemes. Several studies have been conducted based on the NRSS scheme by several authors (see, for example, Koyuncu and Karagöz (2018) and Sabry and Shaaban (2020)).

The EE distribution was introduced by Gupta and Kundu (1999) as a generalization of an exponential distribution. It is of great interest and is popularly used in analyzing

lifetime or survival data. The cumulative distribution function (cdf) and the probability density function (pdf) of the EE distribution are given, respectively, by:

$$F(x;\alpha,\lambda) = (1-e^{-\lambda x})^{\alpha}; \qquad x,\alpha,\lambda > 0, \tag{1}$$

and

$$f(x;\alpha,\lambda) = \alpha\lambda e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-1}; \qquad x,\alpha,\lambda > 0, \tag{2}$$

where α and λ are shape and scale parameters, respectively. Many authors have studied the properties and applications of the EE distribution, including Raqab and Ahsanullah (2001), Gupta and Kundu (2007), Nadarajah (2011), Ristić and Balakrishnan (2012), Abu-Youssef et al. (2015), de Andrade et al. (2016) and Chesneau et al. (2022).

In this study, we introduce, for the first time, the likelihood function for any random variable *X* based on the PRSS scheme, which has not been considered in the literature yet. Further, the population parameter estimators of the EE distribution are considered based on the maximum likelihood (ML) method. Simulation studies are carried out to compare the behaviour of the proposed estimators based on PRSS, RSS, NRSS, ERSS, and SRS designs. Finally, we present an application to real data. The rest of the article is organized as follows. Section 2 describes the RSS, ERSS, NRSS, and PRSS schemes. Section 3 provides the ML estimator of the EE model based on the suggested schemes. Section 4 gives a numerical study as well as application to real data. Finally, concluding remarks are handled in Section 5.

2. Some Ranked Set Sampling Schemes

This section provides the notion and a short description of the proposed RSS, ERSS, NRSS, and the PRSS schemes.

2.1. Ranked Set Sampling

The basic idea behind selecting a sample under RSS can be described as follows:

Step 1: Allocate n^2 randomly selected units from the target population into *n* sets, each of size *n*.

Step 2: Without knowing any values for the variable of interest, rank the units within each set in terms of the variable of interest using your professional judgment.

Step 3: Choose a sample for actual quantification by including the smallest ranked unit in the first set and the second smallest ranked unit in the second set. The process is continued in this way until the largest ranked unit is selected from the last set.

Step 4: Repeat Steps 1–3 for *r* cycles to obtain a sample of size $m = nr$ for measurement.

2.2. Partial Ranked Set Sampling

The PRSS scheme is used when the experimenter is unable to inspect the required number of units or when the inspection cost per unit is high. At the same time, the PRSS scheme requires fewer identified units as compared with a RSS, also it provides more precise estimates than the commonly used SRS scheme. Thus, the PRSS scheme helps in reducing the total cost and expenditure that are involved in sampling. In order to select a PRSS of size *m*, the following steps are carried out:

Step 1: Define a coefficient *k* such that $k = an$, where $0 \le a < 0.5$.

Step 2: Select 2*k* SRS each of size one from the parent population. In order to select the remaining $n-2k$ units, select $n-2k$ sets each of size *n* from the parent population. Rank the units within each set and select the i^{th} ranked unit of the i^{th} sample, for $i = k+1,\ldots,n-k$. This completes one cycle of a PRSS of size *n*.

Step 3: To obtain PRSS of size $m = nr$, we repeat steps 1 and 2 r times. The total number of units that are involved in selecting a PRSS of size $n^2 - 2k(n-1)$. Note that for $k = 0$, PRSS is equivalent to RSS.

2.3. Neoteric Ranked Set Sampling

The NRSS design is applied in situations where the ranking of sample observations is much easier than obtaining their precise values (Zamanzade and Al-Omari (2016)). The NRSS method can be described as follows:

Step 1: Allocate n^2 randomly selected units from the target population and rank the sample units based on the pre-established ordering criterion.

Step 2: If *n* is odd, then select the $[((n+1)/2)+(i-1)n]^{th}$ ranked unit for $i = 1,\ldots,n$. But if *n* is even, select the $[J+(i-1)n]^{th}$ ranked unit, where $J = (n/2)$ if *i* is an even and $J = ((n+2)/2)$ if *i* is an odd for $i = 1,\ldots,n$.

Step 3: Again, steps 1–2 can be repeated *r* times to obtain a final sample of size $m = nr$.

2.4. Extreme Ranked Set Sampling

The ERSS scheme is performed by quantifying the smallest and largest order statistics (Samawi et al. (1996)). The ERSS procedure is as follows:

Step 1: Allocate the n^2 selected units randomly from the target population into *n* sets, each of size *n*.

Step 2: Without yet knowing any values for the variable of interest, rank the units within each set with respect to a variable of interest.

Step 3: If the set size is odd, select the smallest unit from the first $(n-1)/2$ samples, from the other $(n-1)/2$ the largest unit and for the last sample select the median of the sample for actual measurement. If the set size is even, select the smallest unit from the first $n/2$ samples and from the other $n/2$ samples the largest unit for actual measurement.

Step 4: The steps 1 to 3 can be repeated *r* times to obtain a sample of size $m = nr$.

3. Parameter Estimation

In this section, the ML estimators of the EE distribution parameters are obtained based on SRS, RSS, ERSS, NRSS, and PRSS designs.

3.1. ML Estimator based on SRS

Let $X_1, X_2, \ldots, X_m$ be independent and identically distributed random variables from the EE distribution with pdf (2). The log-likelihood function of α and λ is specified by:

$$lnL_1 = m\ln\alpha + m\ln\lambda + (\alpha-1)\sum_{i=1}^{m} ln(1-e^{-\lambda x_i}) - \lambda\sum_{i=1}^{m} x_i.$$

The first partial derivatives of L_1 for each parameter are given by:

$$\frac{\partial ln\, L_1}{\partial\alpha} = \frac{m}{\alpha} + \sum_{i=1}^{m} ln(1-e^{-\lambda x_i}), \tag{3}$$

$$\frac{\partial lnL_1}{\partial\lambda} = \frac{m}{\lambda} + (\alpha-1)\sum_{i=1}^{m}\frac{x_i}{e^{\lambda x_i}-1} - \sum_{i=1}^{m} x_i. \tag{4}$$

Setting Equations (3) and (4) with zero and solving them numerically, we get the ML estimators of α and λ.

3.2. ML Estimator based on RSS

Here, we derive the ML estimators of the EE distribution parameters based on the RSS scheme. Assume that $X = \{X_{i(i)s}; i = 1,2,\ldots,n, s = 1,2,\ldots,r\}$ is a RSS observed from the EE distribution with sample size *nr*, *n* being the set size and *r* being the number of cycles. The likelihood function based on the RSS scheme is given by:

$$L_2 = \prod_{s=1}^{r}\prod_{i=1}^{n} C_1 f(x_{i(i)s})\left[F(x_{i(i)s}\right]^{i-1}\left[1-F(x_{i(i)s}\right]^{n-i}, \tag{5}$$

where $C_1 = n!/[(i-1)!(n-i)!]$. The log-likelihood function of (5), based on RSS, is yielded by substituting pdf (2) and cdf (1) in (5) as follows:

$$\ln L_2 \propto r n \ln\alpha + r n \ln\lambda + \sum_{s=1}^{r}\sum_{i=1}^{n}(\alpha i-1)\ln(T_{i(i)s}) - \lambda\sum_{s=1}^{r}\sum_{i=1}^{n}x_{i(i)s} + \sum_{s=1}^{r}\sum_{i=1}^{n}(n-i)\ln(1-(T_{i(i)s})^{\alpha}),$$

where $T_{i(i)s} = (1-e^{-\lambda x_{i(i)s}})$. The first derivatives of L_2 with respect to α and λ are given by:

$$\frac{\partial ln L_2}{\partial\alpha} = \frac{r n}{\alpha} + \sum_{s=1}^{r}\sum_{i=1}^{n} i\ lnT_{i(i)s} - \sum_{s=1}^{r}\sum_{i=1}^{n}\frac{(n-i)(T_{i(i)s})^{\alpha} lnT_{i(i)s}}{1-(T_{i(i)s})^{\alpha}}, \tag{6}$$

$$\frac{\partial \ln L_2}{\partial\lambda} = \frac{r n}{\lambda} + \sum_{s=1}^{r}\sum_{i=1}^{n}\frac{(\alpha i-1)x_{i(i)s}}{e^{\lambda x_{i(i)s}}-1} - \sum_{s=1}^{r}\sum_{i=1}^{n}x_{i(i)s} - \sum_{s=1}^{r}\sum_{i=1}^{n}\frac{\alpha(n-i)(T_{i(i)s})^{\alpha-1}x_{i(i)s}e^{-\lambda x_{i(i)s}}}{1-(T_{i(i)s})^{\alpha}}. \tag{7}$$

Differentiate (6) and (7) and equate by zero, the estimators of α and λ, say $\hat{\alpha}$ and $\hat{\lambda}$, are obtained through an appropriate numerical technique.

In the following, the pdf of a random variable *X* based on PRSS, as well as its likelihood function, are introduced in the case of any continuous probability distribution. Then, we obtain the pdf of the EE distribution, under PRSS, as well as we provide its likelihood function. Furthermore, based on the log-likelihood function, we obtain the ML estimator of the EE distribution via the PRSS scheme.

3.2.1. Likelihood Function via PRSS

Here, we will define the likelihood function for the PRSS scheme depending on Lemma 1 using the order statistics theory.

Lemma 1:

Let $X = (X_1, X_2, ..., X_k)$, and $X^* = (X_{n-k+1}, X_{n-k+2}, ..., X_n)$, be *k* independent simple random samples each of size *k*. Also, let $X^{**} = (X_{(k+1)n}, X_{(k+2)n}, ..., X_{(n-k)n})$, be the order statistics of size *n*–2*k*. We define the joint pdf of a random variable $X_{i(i)}$, under the PRSS scheme, as follows:

$$f_{X_{i(i)}}(x) = \begin{cases} f_{1X_i}(x) & ,i = 1,...,k, \\ f_{2X^{**}_{(i)}}(x) & ,i = k+1,...,n-k, \\ f_{3X^*_i}(x) & ,i = n-k+1,...,n, \end{cases} \tag{8}$$

where $f_{1X_i}(x)$ is the pdf of SRS, $X = (X_1, X_2, ..., X_k)$, and $f_{3X^*_i}(x)$ is the pdf of SRS $X^* = (X_{n-k+1}, X_{n-k+2}, ..., X_n)$, while $f_{2X^{**}_{(i)}}(x)$ is the pdf of *i*th order statistics of

sample $X^{**}=(X_{(k+1)n},X_{(k+2)n},...,X_{(n-k)n})$, where $(i=k+1,...,n-k)$. Hence, we define, for the first time, the pdf of $X_{i(i)}$ under the PRSS scheme as follows:

$$f_{X_{i(i)}}(x)=C^{*}f_1(x)f_2(x)\left[F_2(x)\right]^{i-k-1}\left[1-F_2(x)\right]^{n-i-k}f_3(x),\ -\infty<x<\infty, \tag{9}$$

where $C^{*}=\dfrac{(n-2k)!}{(i-k-1)!(n-i-k)!}$

Proposition 1:

Let $X_{i(i)s}=\{X_s\cup X_s^{**}\cup X_s^{*}\}=\{X_{is},i=1,...,k\}\cup\{X_{i(i)s},i=k+1,...,n-k\}\cup\{X_{is},i=n-k+1,...,n\}$, $s=1,...,r$ be a PRSS observed from continuous distribution, with a sample size $m=nr$, where *n* is the set size and *r* is the number of cycles. Based on pdf (9), the likelihood function of random variable $X_{i(i)s}$ based on the PRSS design is as follows:

$$L_3=\prod_{s=1}^{r}\left[\prod_{i=1}^{k}f_{1X_i}(x)\prod_{i=k+1}^{n-k}f_{2X^{**}_{(i)s}}(x)\prod_{i=n-k+1}^{n}f_{3X^{*}_{i}}(x)\right], \tag{10}$$

where $f_{1X_i}(x)$ is the pdf of SRS, $X=(X_1,X_2,...,X_k)$, and $f_{3X^{*}_{i}}(x)$ is the pdf of SRS $X^{*}=(X_{n-k+1},X_{n-k+2},...,X_n)$, while $f_{2X^{**}_{(i)s}}(x)$ is the pdf of *i*th order statistics of sample $X^{**}=(X_{(k+1)n},X_{(k+2)n},...,X_{(n-k)n})$, where $(i=k+1,...,n-k)$.

3.2.2. ML Estimator of EE Distribution

Here, the ML estimators of α and λ for the EE distribution are derived based on the PRSS scheme. Assume that $X_{i(i)s}=\{X_{is},i=1,...,k,s=1,...,r\}\cup\{X_{i(i)s},i=k+1,...,n-k,s=1,...,r\}\cup\{X_{is},i=n-k+1,...,n,s=1,...,r\}$ is a PRSS observed from the EE distribution with sample size *nr*, *n* being the set size and *r* being the number of cycles. The likelihood function, via RSS scheme, is obtained by inserting (1) and (2) in (10) as follows:

$$L_3=\prod_{s=1}^{r}\left[\prod_{i=1}^{k}\alpha\lambda e^{-\lambda x_{is}}(1-e^{-\lambda x_{is}})^{\alpha-1}\prod_{i=k+1}^{n-k}C^{*}\,\alpha\lambda e^{-\lambda x_{i(i)s}}(1-e^{-\lambda x_{i(i)s}})^{\alpha(i-k)-1}\left[1-(1-e^{-\lambda x_{i(i)s}})^{\alpha}\right]^{n-i-k}\times\prod_{i=n-k+1}^{n}\alpha\lambda e^{-\lambda x_{is}}(1-e^{-\lambda x_{is}})^{\alpha-1}\right]$$

.

Hence, the logarithm of L_3, under the PRSS design, is as follows:

$$ln\,L_3=r(n-2k)\,ln\,C^{*}+r\,n(ln\,\alpha+ln\,\lambda)+(\alpha-1)\sum_{s=1}^{r}\left(\sum_{i=1}^{k}ln\,(z_{is})+\sum_{i=n-k+1}^{n}ln\,(z_{is})\right)-\lambda\sum_{s=1}^{r}\sum_{i=n-k+1}^{n}x_{is}$$
$$-\lambda\sum_{s=1}^{r}\sum_{i=1}^{k}x_{is}+\sum_{s=1}^{r}\sum_{i=k+1}^{n-k}[\alpha(i-k)-1]ln\,(T_{i(i)s})-\lambda\sum_{s=1}^{r}\sum_{i=k+1}^{n-k}x_{i(i)s}+\sum_{s=1}^{r}\sum_{i=k+1}^{n-k}(n-i-k)ln\,(1-(T_{i(i)s})^{\alpha}),$$

where $Z_{is}=1-e^{-\lambda x_{is}}$ and $T_{i(i)s}=(1-e^{-\lambda x_{i(i)s}})$. The first partial derivatives of L_3, for each parameter are:

$$\frac{\partial ln\, L_3}{\partial \alpha}=\frac{r\,n}{\alpha}+\sum_{s=1}^{r}\left(\sum_{i=1}^{k}ln(Z_{is})+\sum_{i=n-k+1}^{n}ln(Z_{is})\right)+\sum_{s=1}^{r}\left(\sum_{i=k+1}^{n-k}(i-k)ln(T_{i(i)s})-\sum_{i=k+1}^{n-k}\frac{(n-i-k)ln(T_{i(i)s})}{(T_{i(i)s})^{-\alpha}-1}\right),$$

$$\frac{\partial lnL_3}{\partial \lambda}=\frac{r\,n}{\lambda}+\sum_{s=1}^{r}\left(\sum_{i=1}^{k}\frac{(\alpha-1)x_{i\,s}}{e^{\lambda x_{i\,s}}-1}+\sum_{i=n-k+1}^{n}\frac{(\alpha-1)x_{i\,s}}{e^{\lambda x_{i\,s}}-1}\right)+\sum_{s=1}^{r}\sum_{i=k+1}^{n-k}\frac{(\alpha(i-k)-1)x_{i(i)s}}{e^{\lambda x_{i(i)s}}-1}$$

$$-\sum_{s=1}^{r}\left(\sum_{i=1}^{k}x_{i\,s}+\sum_{i=k+1}^{n-k}x_{i(i)s}+\sum_{i=n-k+1}^{n}x_{i\,s}\right)-\sum_{s=1}^{r}\sum_{i=k+1}^{n-k}\frac{\alpha(n-i-k)(T_{i(i)s})^{\alpha-1}x_{i(i)s}e^{-\lambda x_{i(i)s}}}{1-(T_{i(i)s})^{\alpha}}.$$

Clearly, it is not easy to obtain a closed form solution for $\partial ln\, L_3/\partial\alpha, \partial ln\, L_3/\partial\lambda$ after setting them to zero. Therefore, an iterative technique must be applied to solve these equations numerically.

3.3. ML Estimator based on NRSS

Using the NRSS technique, we obtain the ML estimators of the EE distribution parameters. Let $\{X_{b(i)s}, i=1,2,...,n; s=1,2,...,r\}$ and $w=n^2$ be a NRSS where n is the set size, r is the number of cycles, and $b(i)$ is chosen as:

$$b(i)=\begin{cases}\frac{n+1}{2}+(i-1)n, & if\ n\ \ odd\\ \frac{n}{2}+(i-1)n, & if\ n\,even, i\ even\\ \frac{n+2}{2}+(i-1)n, & if\ n\,even, i\ odd\end{cases}$$

According to Sabry and Shaaban (2020), the likelihood function, under the NRSS scheme, is given by:

$$L_4=\prod_{s=1}^{r}C_3\left[\prod_{i=1}^{n}f(x_{b(i)s})\prod_{i=1}^{n+1}[F(x_{b(i)s})-F(x_{b(i-1)s})]^{b(i)-b(i-1)-1}\right], \tag{11}$$

where $C_3=\dfrac{w\,!}{\prod_{i=1}^{n+1}(b(i)-b(i-1)-1)!}$, $b(0)=0$, $b(n+1)=w+1$ $x_{(b(0))}=-\infty, w=n^2$ and $x_{(b(i+1))}=\infty$.

The logarithm of (11), based on the NRSS scheme, is obtained as follows:

$$ln\, L_4\propto r\,n\left(ln\,\alpha+ln\,\lambda\right)+\sum_{s=1}^{r}\sum_{i=1}^{n}(\alpha-1)ln(N_{b(i)s})-\sum_{s=1}^{r}\sum_{i=1}^{n}\lambda x_{b(i)s}$$
$$+\sum_{s=1}^{r}\sum_{i=1}^{n+1}[b(i)-b(i-1)-1]ln[(N_{b(i)s})^{\alpha}-(N_{b(i-1)s})^{\alpha}],$$

where $N_{b(i)s} = 1 - e^{-\lambda x_{b(i)s}}$ and $N_{b(i-1)s} = 1 - e^{-\lambda x_{b(i-1)s}}$.

The first partial derivatives of L_4 with respect to each parameter are given by:

$$\frac{\partial ln L_4}{\partial \alpha} = \frac{r\,n}{\alpha} + \sum_{s=1}^{r}\sum_{i=1}^{n} ln(N_{b(i)s}) + \sum_{s=1}^{r}\sum_{i=1}^{n+1} \frac{(b(i)-b(i-1)-1)\left[(N_{b(i)s})^{\alpha} ln(N_{b(i)s}) - (N_{b(i-1)s})^{\alpha} ln(N_{b(i-1)s})\right]}{[(N_{b(i)s})^{\alpha} - (N_{b(i-1)s})^{\alpha}]}, \tag{12}$$

$$\frac{\partial ln\, L_4}{\partial \lambda} = \frac{r\,n}{\lambda} + \sum_{s=1}^{r}\sum_{i=1}^{n} \frac{(\alpha-1)\,x_{b(i)s}}{e^{\lambda x_{b(i)s}} - 1} - \sum_{s=1}^{r}\sum_{i=1}^{n} x_{b(i)s} + \sum_{s=1}^{r}\sum_{i=1}^{n+1} \frac{(b(i)-b(i-1)-1)[\alpha(N_{b(i)s})^{\alpha-1} e^{-\lambda x_{b(i)s}} x_{b(i)s} - \alpha(N_{b(i-1)s})^{\alpha-1} e^{-\lambda x_{b(i-1)s}} x_{b(i-1)s}]}{[(N_{b(i)s})^{\alpha} - (N_{b(i-1)s})^{\alpha}]}. \tag{13}$$

There is no closed form solution to (12) and (13), so a numerical technique will be used to obtain the ML estimators for α and λ, represented by $\hat{\alpha}$, and $\hat{\lambda}$.

3.4. ML Estimator based on ERSS

In this section, the ML estimation approach will be used to estimate the EE distribution parameters on the basis of the ERSS scheme.

3.4.1 ML Estimator for Odd *Set Size*

Suppose that $X = \{X_{i(1)s}, i = 1,2,...g-1, s = 1,2,...,r\} \cup \{X_{i(n)s}, i = g, g+1,...,n-1, s = 1,2,...,r\} \cup \{X_{n(g)s}, , g = n+1/2, s = 1,...,r\}$ is an odd ERSS (ERSSO) design observed from the EE distribution, with sample size $m = nr$, where n is the set size, r is the number of cycles.

Then the likelihood function, under the ERSSO scheme, is given as follows:

$$L_5 \propto \prod_{s=1}^{r}\prod_{i=1}^{g-1} f_1(x_{i(1)s}) \prod_{s=1}^{r}\prod_{i=g}^{n-1} f_n(x_{i(n)s}) \prod_{s=1}^{r} f_g(x_{n(g)s}),$$

where $f_1(x_{i(1)s})$ and $f_n(x_{i(n)s})$ are the pdfs of the smallest and largest order statistics, respectively, and $f_g(x_{n(g)s})$ is the pdf of the median. Hence, the logarithm of L_5, based on ERSSO, is obtained as follows:

$$\begin{aligned} ln\, L_5 \propto\; & r\,n\left[ln\alpha + ln\lambda\right] - \lambda\sum_{s=1}^{r}\sum_{i=1}^{g-1} x_{i(1)s} + (n-1)\sum_{s=1}^{r}\sum_{i=1}^{g-1} ln[1-(W_{i(1)s})^{\alpha}] - \lambda\sum_{s=1}^{r}\sum_{i=g}^{n-1} x_{i(n)s} \\ & + (\alpha-1)\sum_{s=1}^{r}\sum_{i=1}^{g-1} ln(W_{i(1)s}) + (\alpha n-1)\sum_{s=1}^{r}\sum_{i=g}^{n-1} ln(V_{i(n)s}) - \lambda\sum_{s=1}^{r} x_{n(g)s} \\ & + (\alpha g-1)\sum_{s=1}^{r} ln(E_{n(g)s}) + (g-1)\sum_{s=1}^{r} ln(1-(E_{n(g)s})^{\alpha}), \end{aligned}$$

where $W_{i(1)s} = 1 - e^{-\lambda x_{i(1)s}}$, $V_{i(n)s} = 1 - e^{-\lambda x_{i(n)s}}$ and $E_{n(g)s} = 1 - e^{-\lambda x_{n(g)s}}$. The first partial derivatives of L_5 owing to α and λ are given, respectively, by:

$$\frac{\partial lnL_5}{\partial \alpha} = \frac{r\,n}{\alpha} - \sum_{s=1}^{r}\sum_{i=1}^{g-1}\frac{(n-1)(W_{i(1)s})^{\alpha} ln(W_{i(1)s})}{1-(W_{i(1)s})^{\alpha}} + \sum_{s=1}^{r}\sum_{i=1}^{g-1} ln(W_{i(1)s}) + \sum_{s=1}^{r}\sum_{i=g}^{n-1} n\, ln(V_{i(n)s}) + \sum_{s=1}^{r} g ln(E_{n(g)s}) - \sum_{s=1}^{r}\frac{(g-1)(E_{n(g)s})^{\alpha} ln(E_{n(g)s})}{1-(E_{n(g)s})^{\alpha}}, \tag{14}$$

$$\frac{\partial lnL_5}{\partial \lambda} = \frac{r\,n}{\lambda} - \sum_{s=1}^{r}\left[\sum_{i=1}^{g-1} x_{i(1)s} - \sum_{i=g}^{n-1} x_{i(n)s} - x_{n(g)s}\right] + \sum_{s=1}^{r}\sum_{i=1}^{g-1}\frac{(\alpha-1)x_{i(1)s}}{e^{\lambda x_{i(1)s}}-1} - \sum_{s=1}^{r}\sum_{i=1}^{g-1}\frac{\alpha(n-1)(W_{i(1)s})^{\alpha-1}e^{-\lambda x_{i(1)s}}x_{i(1)s}}{1-(W_{i(1)s})^{\alpha}} + \sum_{s=1}^{r}\sum_{i=g}^{n-1}\frac{(\alpha n-1)x_{i(n)s}}{e^{\lambda x_{i(n)s}}-1} + \sum_{s=1}^{r}\frac{(\alpha g-1)x_{n(g)s}}{e^{\lambda x_{n(g)s}}-1} - \sum_{s=1}^{r}\frac{\alpha(g-1)(E_{n(g)s})^{\alpha-1}e^{-\lambda x_{n(g)s}}x_{n(g)s}}{1-(E_{n(g)s})^{\alpha}}. \tag{15}$$

Using an iterative technique for (14) and (15) after setting them with zero to produce the ML estimators of α and λ.

3.4.2. ML Estimator for Even Set Size

Suppose that $X = \{X_{i(1)s}, i = 1,2,...g_1,\ s = 1,2,...,r\} \cup \{X_{i(n)s},\ i = g_1+1,\ g_1+2,...,n,\ s = 1,2,...,r\}$ is an even ERSS (ERSSE) scheme observed from an EE distribution, with a sample of size $m = nr$, where n is the set size, r is the number of cycles and $g_1 = n/2$. The likelihood function of the EE distribution from the ERSSE scheme is given by:

$$L_6 \propto \prod_{s=1}^{r}\prod_{i=1}^{g_1} f_1(x_{i(1)s}) \prod_{s=1}^{r}\prod_{i=g_1+1}^{n} f_n(x_{i(n)s}).$$

The logarithm of L_6 for the EE distribution, using the ERSSE scheme, is given by.

$$lnL_6 \propto r\,n\left(ln\,\alpha + ln\,\lambda\right) + (n-1)\sum_{s=1}^{r}\sum_{i=1}^{g_1} ln\left[1-(W_{i(1)s})^{\alpha}\right] + (\alpha-1)\sum_{s=1}^{r}\sum_{i=1}^{g_1} ln(W_{i(1)s}) - \lambda\sum_{s=1}^{r}\sum_{i=1}^{g_1} x_{i(1)s} + (\alpha n-1)\sum_{s=1}^{r}\sum_{i=g_1+1}^{n} ln(V_{i(n)s}) - \lambda\sum_{s=1}^{r}\sum_{i=g_1+1}^{n} x_{i(n)s}.$$

The first partial derivatives of L_6 owing to α and λ are given, respectively, by:

$$\frac{\partial lnL_6}{\partial \alpha} = \frac{r\,n}{\alpha} - \sum_{s=1}^{r}\sum_{i=1}^{g_1}\frac{(n-1)(W_{i(1)s})^{\alpha} ln(W_{i(1)s})}{1-(W_{i(1)s})^{\alpha}} + \sum_{s=1}^{r}\sum_{i=1}^{g_1} ln(1-e^{-\lambda x_{i(1)s}}) + \sum_{s=1}^{r}\sum_{i=g_1+1}^{n} n\, ln(V_{i(n)s}), \tag{16}$$

and,

$$\frac{\partial lnL_6}{\partial \lambda} = \frac{r\,n}{\lambda} - \sum_{s=1}^{r}\sum_{i=1}^{g_1} \frac{\alpha(n-1)(W_{i(1)s})^{\alpha-1} e^{-\lambda x_{i(1)s}}\, x_{i(1)s}}{1-(W_{i(1)s})^{\alpha}} + \sum_{s=1}^{r}\sum_{i=1}^{g_1} \frac{(\alpha-1)x_{i(1)s}}{e^{\lambda x_{i(1)s}}-1}$$
$$- \sum_{s=1}^{r}\left[\sum_{i=1}^{g_1} x_{i(1)s} + \sum_{i=g_1+1}^{n} x_{i(n)s}\right] + \sum_{s=1}^{r}\sum_{i=g_1+1}^{n} \frac{(\alpha n-1)x_{i(n)s}}{e^{\lambda x_{i(n)s}}-1}. \tag{17}$$

After setting (16) and (17) to zero, there is no closed form solution, hence the ML estimators α and λ are derived using a numerical technique.

4. Numerical Study and Application

In this section, a numerical study is provided to evaluate the behaviour of ML estimates (MLEs) of the EE distribution based on the SRS, RSS, PRSS, NRSS, and ERSS schemes. Also, an application to one real data set is provided.

4.1. Numerical Study

A numerical evaluation is carried out to examine the performance of the MLEs. The MLEs are evaluated based on absolute biases (ABs), mean squared errors (MSEs), and relative efficiencies (REs). The simulation procedure is achieved via the MATHEMATICA software. The simulation algorithm is performed as follows:

Step 1: An SRS scheme $X_1, X_2, \ldots, X_n$ of sample sizes; m = 20, 40, 60 and 100 are considered; and these random samples are generated from the EE distribution by using the inversion method.

Step 2: An RSS scheme is considered as: $X_{1(1)s}, X_{2(2)s}, \ldots, X_{n(n)s}$; $s = 1, \ldots, r$ having sample sizes; m = 20, 40, 60 and 100 with the number of cycles r = 5, 10, and 20 and set sizes n = 4, 5 and 6.

Step 3: A PRSS scheme is considered as:

$X_{1s}, X_{2s}, \ldots, X_{ks}, X_{k+1(k+1)s}, X_{k+2(k+2)s}, \ldots, X_{n-k(n-k)s}, X_{n-k+1s}, X_{n-k+2s}, \ldots, X_{ns}$; s=1,…, r of sample sizes; m = 20, 40, 60 and 100, where (n, r)= (4,5), (4,10), (6,10) and (5,20).

Step 4: An NRSS scheme is considered as $X_{b(1)s}, X_{b(2)s}, \ldots, X_{b(n)s}$; s =1,…, r of sample sizes; m = 20, 40, 60 and 100, where (n, r)= (4,5), (4,10), (6,10) and (5,20).

Step 5: An ERSSO scheme is considered as

$X_{1(1)s}, \ldots, X_{g-1(1)s}, X_{g(n)s}, \ldots, X_{n-1(n)s}, X_{n(g)s}; g = \frac{n+1}{2}$, $s = 1, \ldots, r$ of sample sizes; m = 20, 40, 60 and 100, where (n, r) = (5,4), (5,8), (5,12) and (5,20).

Step 6: An ERSSE scheme is considered as

$X_{1(1)s},...,X_{g_1(1)s},X_{g_1+1(n)s},...,X_{n(n)s};g_1=\frac{n}{2},s=1,...,r$ of sample sizes; m = 20, 40, 60 and 100, where $(n, r) = (4,5), (4,10), (6,10)$ and $(4,25)$.

Step 7: Parameters' values are selected as (α =0.5, λ =0.4), (α =1, λ =0.4), (α =2, λ =2) and (α =3, λ =2). The MSEs and ABs of $\hat{\alpha}$ and $\hat{\lambda}$ are evaluated for different sample sizes.

Step 8: The efficiencies of different estimates under selective schemes with respect to SRS are defined by $RE_{\zeta}(\hat{\theta})=\frac{MSE_{SRS}(\hat{\theta})}{MSE_{\zeta}(\hat{\theta})}$, where $\hat{\theta} = (\hat{\alpha}, \hat{\lambda}), \zeta$ = RSS, PRSS, NRSS, ERSSE, and ERSSO.

Step 9: The process is repeated 1000 times. The MLEs of $\hat{\alpha}$ and $\hat{\lambda}$ are inspected via ABs, MSEs, and their efficiencies.

Step 10: Empirical results are listed in Tables 1–3. Tables 1 and 2 list the observed results of ABs and MSEs of both estimates based on selective schemes. Also, Table 3 gives the efficiency of different schemes with respect to SRS.

Based on Tables 1–3 and Figures 1–11, we conclude the following:

1- For all sampling schemes, as m increases, the MSE and AB of $\hat{\alpha}$ and $\hat{\lambda}$ decreases (see Tables 1, 2).

2- The MLEs of $\hat{\alpha}$ and $\hat{\lambda}$ under the NRSS scheme provide more efficient estimates than the corresponding estimates in other schemes.

3- The MLEs of $\hat{\alpha}$ and $\hat{\lambda}$ under all modifications of the RSS schemes are more efficient than the corresponding estimates under the SRS scheme (see Figure 1 and Figure 2).

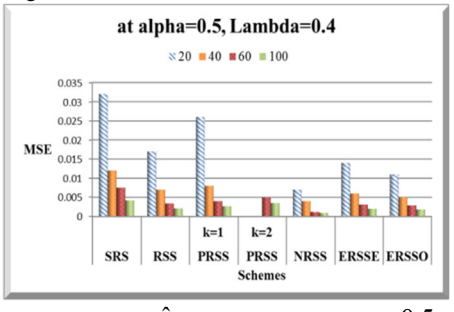


Figure 1. MSE of $\hat{\alpha}$ for all schemes at $\alpha = 0.5$ and $\lambda = 0.4$

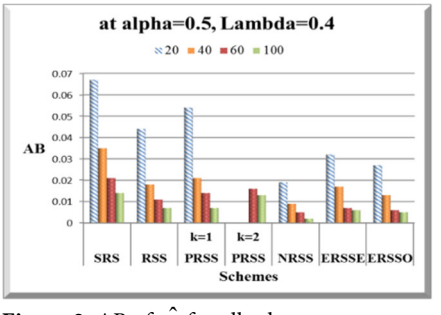


Figure 2. AB of $\hat{\alpha}$ for all schemes at $\alpha = 0.5$ and $\lambda = 0.4$

4- The MLEs of $\hat{\alpha}$ and $\hat{\lambda}$ under NRSS are more efficient than the others based on the RSS, PRSS (at $k = 1$ and $k = 2$) and ERSS schemes (see Figure 3 and Table 3).

5- The MLEs of $\hat{\alpha}$ and $\hat{\lambda}$ under the PRSS scheme at k =1, 2 are more efficient than the corresponding estimates under the SRS for all different values of m (see Figure 4).

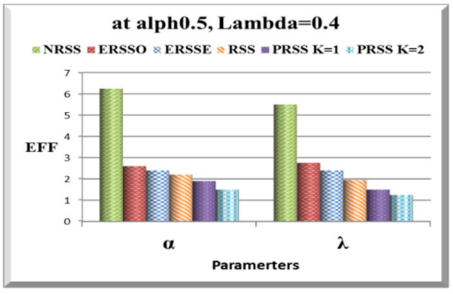


Figure 3. Efficiency of MLEs for all schemes at m = 60 at $\alpha = 0.5$, $\lambda = 0.4$

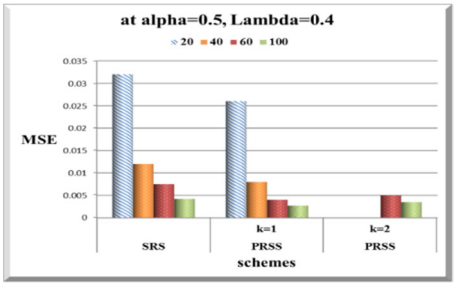


Figure 4. MSE of $\hat{\alpha}$ under SRS and PRSS schemes at $\alpha = 0.5$, $\lambda = 0.4$

6- The MSE of $\hat{\alpha}$ under PRSS increases as the value of k increases from k =1 to k =2, because the number of observations under SRS increases when selecting the PRSS. In this regard, we notice that as the value of k increases, the MSE of MLEs approaches the MSE of those under SRS (see Figures 4 and 5).

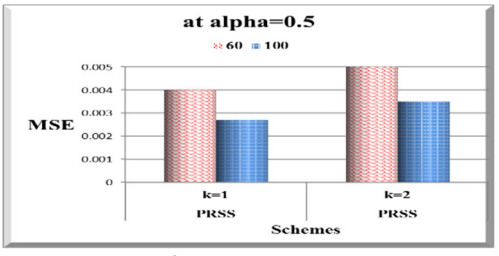


Figure 5. MSE of $\hat{\alpha}$ under PRSS for m = 60 and 100

7- As the value of α increases, the MSE of $\hat{\alpha}$ increases, while the MSE of $\hat{\lambda}$ decreases under different sampling schemes (see Figures 6, 7 and Tables 1, 2).

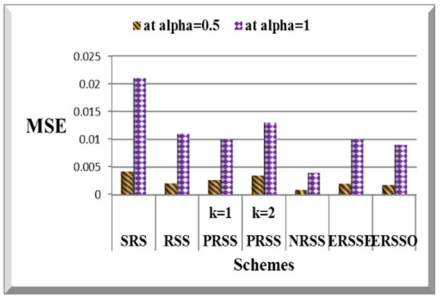


Figure 6. MSE of $\hat{\alpha}$ for all schemes at $m = 100$

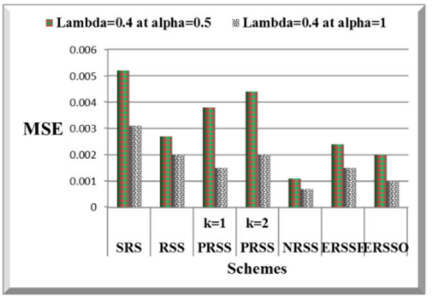


Figure 7. MSE of $\hat{\lambda}$ for all schemes at $m = 100$

8- As the value of α increases, from 0.5 to 1, the MSE and the AB of $\hat{\alpha}$ increase, while the MSE and the AB of $\hat{\lambda}$ decrease at $m = 100$ (see Figures 8, 9 and Tables 1, 2).

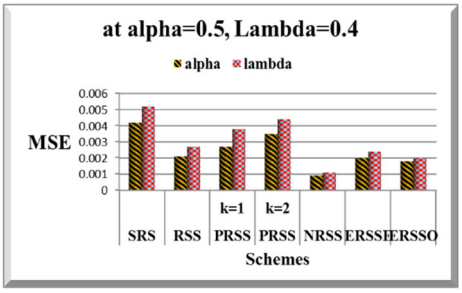


Figure 8. MSE of $\hat{\alpha}$ and $\hat{\lambda}$ for all schemes at $m = 100$

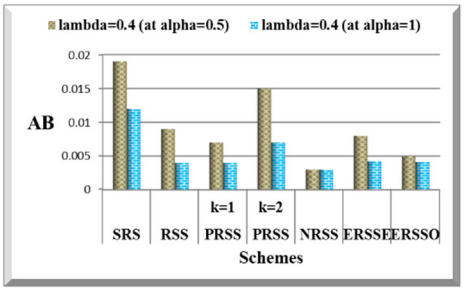


Figure 9. AB of $\hat{\lambda}$ when $\alpha = 0.5$ and 1 for all schemes at $m = 100$

9- The MLE of $\hat{\alpha}$ under the ERSSO scheme is more efficient than the others under the ERSSE for all *m* (see Figure 10 and Tables 1, 2).

10-As the sample size *m* increases, the efficiency of estimates also increases (see Figure 11 and Table 3).

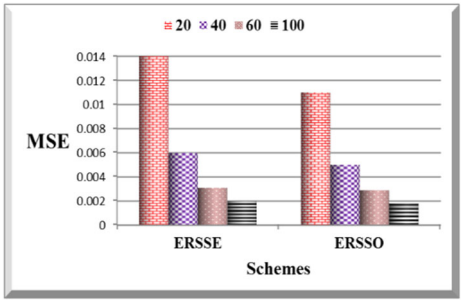


Figure 10. MSE of $\hat{\alpha}$ under ERSSE and ERSSO for all *m*

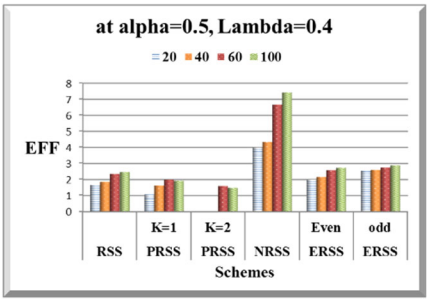


Figure 11. Efficiency of the MLEs for all schemes at all sample sizes

Table 1. The MSEs and ABs of the EE distribution based on different RSS schemes

m		scheme		$\alpha = 0.5,\ \lambda = 0.4$				$\alpha = 1,\ \lambda = 0.4$			
n	*r*			**MSE**		**AB**		**MSE**		**AB**	
				$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\lambda}$
20		SRS		0.028	0.039	0.057	0.074	0.201	0.021	0.169	0.051
4	5	RSS		0.017	0.022	0.044	0.047	0.102	0.012	0.097	0.032
4	5	PRSS	k=1	0.026	0.031	0.054	0.063	0.159	0.019	0.135	0.045
4	5	NRSS		0.007	0.010	0.019	0.025	0.043	0.006	0.052	0.016
4	5	ERSSE		0.014	0.017	0.032	0.038	0.081	0.010	0.092	0.029
5	4	ERSSO		0.011	0.014	0.027	0.037	0.077	0.009	0.088	0.028
40		SRS		0.013	0.017	0.032	0.041	0.078	0.009	0.088	0.030
4	10	RSS		0.007	0.009	0.018	0.022	0.038	0.005	0.047	0.015
4	10	PRSS	k=1	0.008	0.012	0.021	0.026	0.059	0.008	0.072	0.023
4	10	NRSS		0.003	0.004	0.007	0.012	0.016	0.002	0.026	0.009
4	10	ERSSE		0.006	0.007	0.017	0.019	0.03	0.004	0.028	0.011
5	8	ERSSO		0.005	0.006	0.013	0.013	0.028	0.003	0.021	0.010
60		SRS		0.0075	0.0088	0.021	0.024	0.037	0.006	0.047	0.016
6	10	RSS		0.0034	0.0045	0.011	0.012	0.017	0.003	0.021	0.007
6	10	PRSS	k=1	0.004	0.006	0.014	0.018	0.022	0.003	0.026	0.009
			k=2	0.005	0.007	0.016	0.020	0.033	0.004	0.041	0.013
6	10	NRSS		0.0012	0.0016	0.005	0.007	0.007	0.0011	0.013	0.005
6	10	ERSSE		0.0031	0.0036	0.007	0.008	0.014	0.0026	0.018	0.006
5	12	ERSSO		0.0029	0.0032	0.006	0.006	0.013	0.0018	0.015	0.0045
100		SRS		0.0042	0.0052	0.014	0.019	0.024	0.004	0.035	0.012
5	20	RSS		0.0021	0.0027	0.007	0.009	0.011	0.002	0.015	0.004
5	20	PRSS	k=1	0.0027	0.0038	0.007	0.007	0.010	0.0015	0.014	0.004
			k=2	0.0035	0.0044	0.013	0.015	0.013	0.0020	0.021	0.007
5	20	NRSS		0.0007	0.001	0.002	0.0002	0.004	0.0007	0.008	0.003
4	25	ERSSE		0.0019	0.0023	0.007	0.007	0.009	0.0015	0.014	0.0042
5	20	ERSSO		0.0018	0.002	0.005	0.005	0.008	0.0010	0.013	0.0041

Table 2. The MSEs and ABs of the EE distribution based on different RSS schemes

m	*r*	scheme		$\alpha=2,\ \lambda=2$				$\alpha=3,\ \lambda=2$			
n				MSE		AB		MSE		AB	
				$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\lambda}$
20		SRS		0.979	0.383	0.382	0.212	3.420	0.287	0.788	0.178
4	5	RSS		0.575	0.204	0.252	0.132	2.169	0.196	0.509	0.135
4	5	PRSS	k=1	0.958	0.282	0.369	0.160	2.565	0.260	0.768	0.168
4	5	NRSS		0.262	0.105	0.142	0.072	0.808	0.093	0.279	0.070
4	5	ERSSE		0.493	0.174	0.209	0.111	1.379	0.149	0.334	0.090
5	4	ERSSO		0.367	0.147	0.169	0.085	1.293	0.134	0.330	0.077
40		SRS		0.468	0.187	0.232	0.114	1.192	0.140	0.379	0.120
4	10	RSS		0.228	0.097	0.147	0.083	0.679	0.079	0.191	0.049
4	10	PRSS	k=1	0.340	0.127	0.180	0.084	0.885	0.112	0.327	0.079
4	10	NRSS		0.117	0.051	0.070	0.034	0.258	0.039	0.116	0.029
4	10	ERSSE		0.169	0.084	0.103	0.064	0.450	0.066	0.168	0.045
5	8	ERSSO		0.165	0.072	0.097	0.047	0.402	0.055	0.151	0.041
60		SRS		0.256	0.111	0.160	0.087	0.663	0.083	0.212	0.060
6	10	RSS		0.103	0.048	0.070	0.041	0.298	0.041	0.108	0.030
6	10	PRSS	k=1	0.156	0.069	0.079	0.040	0.459	0.058	0.161	0.039
			k=2	0.207	0.088	0.126	0.067	0.554	0.070	0.201	0.051
6	10	NRSS		0.033	0.017	0.019	0.013	0.091	0.019	0.051	0.017
6	10	ERSSE		0.098	0.048	0.072	0.042	0.240	0.038	0.102	0.025
5	12	ERSSO		0.089	0.041	0.067	0.033	0.206	0.032	0.099	0.031
100		SRS		0.222	0.078	0.092	0.050	0.314	0.050	0.139	0.040
5	20	RSS		0.072	0.033	0.058	0.029	0.172	0.025	0.061	0.019
5	20	PRSS	k=1	0.096	0.044	0.078	0.039	0.253	0.036	0.119	0.033
			k=2	0.120	0.053	0.089	0.048	0.303	0.042	0.133	0.034
5	20	NRSS		0.024	0.011	0.014	0.012	0.063	0.011	0.019	0.006
4	25	ERSSE		0.057	0.029	0.038	0.020	0.149	0.024	0.059	0.018
5	20	ERSSO		0.052	0.025	0.023	0.010	0.147	0.022	0.034	0.009

Table 3. Efficiency of the estimators based on RSS, PRSS (at k =1, 2), ERSSE, ERSSO, and NRSS

n	scheme		$\alpha = 0.5,$ $\lambda = 0.4$		$\alpha = 1,\ \lambda = 0.4$		$\alpha = 2,\ \lambda = 2$		$\alpha = 3,\ \lambda = 2$	
			$EFF(\hat{\alpha})$	$EFF(\hat{\lambda})$	$EFF(\hat{\alpha})$	$EFF(\hat{\lambda})$	$EFF(\hat{\alpha})$	$EFF(\hat{\lambda})$	$EFF(\hat{\alpha})$	$EFF(\hat{\lambda})$
20	RSS		1.65	1.77	1.97	1.75	1.7	1.87	1.57	1.46
	PRSS	k=1	1.07	1.25	1.26	1.11	1.02	1.35	1.33	1.10
	NRSS		4	3.9	4.67	3.5	3.73	3.65	4.23	3.08
	ERSSE		2	2.29	2.48	2.1	1.98	2.20	2.48	1.92
	ERSSO		2.55	2.78	2.61	2.33	2.66	2.60	2.64	2.14
40	RSS		1.85	1.88	2.05	1.8	2.05	1.93	1.75	1.77
	PRSS	k=1	1.63	1.42	1.32	1.13	1.37	1.47	1.34	1.25
	NRSS		4.33	4.25	4.87	4.5	4	3.67	4.62	3.58
	ERSSE		2.16	2.43	2.6	2.25	2.76	2.22	2.64	2.12
	ERSSO		2.6	2.8	2.78	3	2.83	2.59	2.96	2.54
60	RSS		2.35	2	2	2	2.48	2.31	2.22	2.02
	PRSS	k=1	2	1.5	1.7	2	1.64	1.61	1.44	1.43
		k=2	1.6	1.28	1.12	1.5	1.23	1.26	1.19	1.18
	NRSS		6.66	5.62	5.3	4.45	7.75	6.52	7.28	4.63
	ERSSE		2.58	2.5	2.64	2.3	2.61	2.31	2.76	2.18
	ERSSO		2.75	2.81	2.85	3.3	2.87	2.71	3.21	2.59
100	RSS		2.47	2.22	2.18	2	3.08	2.36	2.83	2.32
	PRSS	k=1	1.92	1.57	2.4	2.66	12.31	1.77	1.92	1.61
		k=2	1.48	1.36	1.85	2	1.85	1.47	1.60	1.38
	NRSS		7.42	6	6	5.7	9.25	7.09	7.73	5.27
	ERSSE		2.73	2.6	2.67	2.66	3.89	2.68	3.26	2.41
	ERSSO		2.88	3	3	4	2.26	3.12	3.31	2.63

4.2. Application to Real Data

Here, a real data set is considered, and all the details for illustrative purposes are described. The data represent the survival times (in days) of 72 guinea pigs infected with virulent tubercle bacilli, observed and reported by Bjerkedal (1960). To check the validity of the fitted model, the Kolmogorov-Smirnov (KS) goodness of fit test and its

P-value are obtained. It is observed that the KS distance is 0.0931 with a corresponding P-value of 0.561. Additionally, some criteria measurements including values of -2lnL = 188.472, Akaike information criterion (AIC) = 192.472, correct AIC (AICc) =192.646, Bayesian information criterion (BIC) = 192.187 and Hannan-Quinn information criterion (HQIC) =194.285 were used to acquire more information. These results show that the EE model fits the data reasonably well.

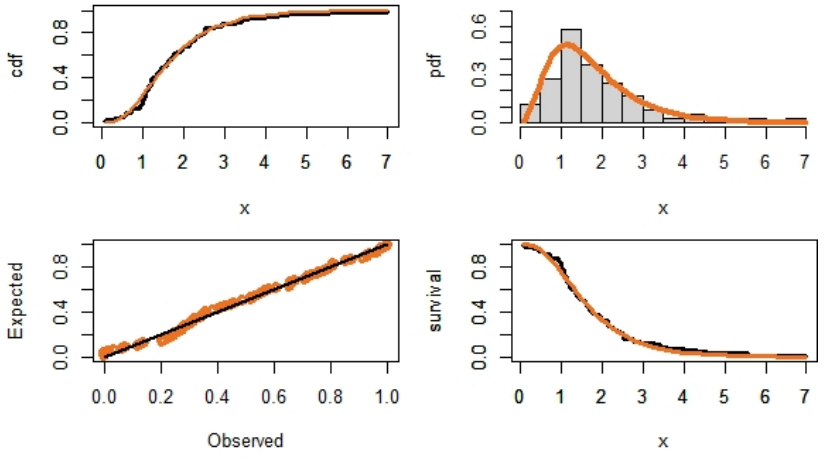


Figure 12. Plots of pdf, cdf, PP plots, and empirical survival function of the EE model

Table 4 gives the observed ranked values according to different sampling method techniques.

Table 4. The observation of different ranked sampling from real data set

Observation	Schemes						
	NRSS	RSS	PRSS, k=1	PRSS, k=2	SRS	ERSSE	ERSSO
1	0.10	0.10	0.10	0.10	0.10	0.56	0.10
2	0.74	0.77	0.44	0.33	0.33	0.92	0.72
3	1.00	1.05	0.39	0.59	0.44	1.07	0.77
4	1.15	1.12	1.07	1.00	0.56	1.09	0.93
5	1.24	1.22	1.15	1.05	0.59	1.22	1.05
6	1.46	1.46	1.20	1.07	0.72	1.36	1.07
7	1.53	1.53	1.21	1.07	0.74	1.63	1.08
8	1.71	1.72	1.22	1.08	0.77	1.76	1.15
9	1.97	2.13	1.46	1.09	0.92	2.15	1.20
10	2.53	2.45	1.71	1.22	0.93	2.40	1.22
11	3.42	3.27	2.02	1.30	0.96	2.93	1.36
12	5.55	5.55	2.15	1.34	1.00	4.02	1.44

Based on the theoretical study, we obtain the MLEs of α and λ under the PRSS, RSS, NRSS, ERSS, and SRS sampling from the considered data set. Table 5 gives the parameter estimators and their corresponding standard error (SE) of the EE model via the PRSS, RSS, NRSS, ERSS, and SRS schemes.

Table 5. Estimated parameters and SE of the EE distribution based on selective RSS schemes

Scheme		Estimators		SE		$RE_{\zeta}(\hat{\theta})$	
		$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\lambda}$
NRSS		1.759	0.747	0.730	0.243	2.99	3.10
RSS		1.789	0.755	0.744	0.244	2.94	3.08
ERSS (even)		2.948	1.890	1.269	0.517	1.80	1.20
ERSS (odd)		2.217	1.323	0.926	0.394	2.40	1.80
PRSS	k=1	3.809	1.963	1.440	0.781	1.40	1.19
	k=2	3.998	1.986	1.525	0.888	1.30	1.17
SRS		5.260	2.330	2.882	0.974	1	1

Table 5 shows that the SE of $\hat{\alpha}$ and $\hat{\lambda}$ based on NRSS, RSS, ERSSE, ERSSO, and PRSS (at $k=1$ and $k=2$) are smaller than the corresponding estimates based on SRS for the considered data.

5. Conclusion

This paper introduces and defines the density and likelihood function for a random variable under the PRSS scheme. The maximum likelihood estimators of exponentiated exponential distribution are discussed under selective RSS schemes and the SRS scheme. The proposed sampling schemes are SRS, RSS, PRSS, NRSS, and ERSS. An intensive numerical study was conducted to compare the performances of different estimators using some accuracy measures. Generally, based on a numerical study, we conclude that all ranked schemes (RSS, PRSS, NRSS, and ERSS) are more efficient than the SRS scheme as evidenced by the results in Table 3. Also, PRSS is not the best method compared to the other ranked schemes, but it is important in some cases, in selecting the sample, when it is either difficult to rank the units within each set with full confidence or due to non-availability of experimental units.

References

Abu-Dayyeh, W., Assrhani, A., Ibrahim, K., (2013). Estimation of the shape and scale parameters of Pareto distribution using ranked set sampling. *Statistical Papers,* 54(1), pp. 207–225.

Abu-Youssef, S. E., Mohammed, B. I., Sief, M. G., (2015). An extended exponentiated exponential distribution and its properties. *International Journal of Computer Applications,* 121(5), pp. 1–6.

Al-Odat, M. T., Al-Saleh, M. F., (2001). A variation of ranked set sampling. *Journal of Applied Statistical Science,* 10(2), pp. 137–146.

Al-Omari, A. I., Almanjahie, I. M., Hassan, A. S., Nagy, H. F., (2020). Estimation of the stress-strength reliability for exponentiated Pareto distribution using median and ranked set sampling methods. *CMC-Computers, Materials & Continua,* 64(2), pp. 835–857.

Almarashi, A. M., Algarni, A., Hassan, A. S., Elgarhy, M., Jamal, F., Chesneau, C., Alrashidi, K., Mashwani, W. K., Nagy, H. F., (2021). A new estimation study of the stress-strength reliability for the Topp–Leone distribution using advanced sampling methods. *Scientific Programming*, pp. 1–13, https://doi.org/10.1155/2021/2404997.

Bantan, R., Hassan, A. S., Elsehetry, M., (2020). Zubair Lomax distribution: properties and estimation based on ranked set sampling. *CMC-Computers, Materials & Continua,* 65(3), pp. 2169–2187.

Bhoj, D. S., Ahsanullah, M., (1996). Estimation of parameters of the generalized geometric distribution using ranked set sampling. *Biometrics*, pp. 685–694.

Bjerkedal, T., (1960). Acquisition of resistance in Guinea pigs infected with different doses of virulent tubercle bacilli. *American Journal of Hygiene,* 72(1), pp. 130–148.

Chesneau, C., Kumar, V., Khetan, M., Arshad, M., (2022). On a modified weighted exponential distribution with applications. *Mathematical and Computational Applications,* 27(1), https://doi.org/10.3390/mca27010017.

De Andrade, T. A., Bourguignon, M., Cordeiro, G. M., (2016). The exponentiated generalized extended exponential distribution. *Journal of Data Science,* 14(3), pp. 393–413.

Gupta, R. D., Kundu, D., (1999). Generalized exponential distributions. *Australian & New Zealand Journal of Statistics,* 41(2), pp. 173–188.

Gupta, R. D., Kundu, D., (2007). Generalized exponential distribution: Existing results and some recent developments. *Journal of Statistical Planning and Inference,* 137(11), pp. 3537–3547.

Haq, A., Brown, J., Moltchanova, E., Al-Omari, A. I., (2013). Partial ranked set sampling design. *Environmetrics,* 24(3), pp. 201–207.

Hassan, A. S., (2012). Modified goodness of fit tests for exponentiated Pareto distribution under selective ranked set sampling. *Australian Journal of Basic and Applied Sciences,* 6(1), pp. 173–189.

Hassan, A. S., (2013). Maximum likelihood and Bayes estimators of the unknown parameters for exponentiated exponential distribution using ranked set sampling. *International Journal of Engineering Research and Applications,* 3(1), pp. 720–725.

Hassan, A. S., Assar, S., Yahya, M., (2014). Estimation of R= P [Y< X] for Burr type XII distribution based on ranked set sampling. *International Journal of Basic and Applied Sciences,* 3(3), pp. 274–280.

Hassan, A. S., Assar, S., Yahya, M., (2015). Estimation of P (Y< X) for Burr distribution under several modifications for ranked set sampling. *Australian Journal of Basic and Applied Sciences,* 9(1), pp. 124–140.

Hassan, A. S., Elbagouri, R., Onyango, R., Nagy, H. F., (2022). Estimating system reliability using neoteric and median RSS data for generalized exponential distribution. *International Journal of Mathematics and Mathematical Sciences, 2608656,* https://doi.org/10.1155/2022/2608656.

Koyuncu, N., Karagöz, D., (2018). New mean charts for bivariate asymmetric distributions using different ranked set sampling designs. *Quality Technology & Quantitative Management,* 15(5), pp. 602–621.

Mahdizadeh, M., Arghami, N. R., (2010). Efficiency of ranked set sampling in entropy estimation and goodness-of-fit testing for the inverse Gaussian law. *Journal of Statistical Computation and Simulation,* 80(7), pp. 761–774.

Mcintyre, G., (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research,* 3(4), pp. 385–390.

Nadarajah, S., (2011). The exponentiated exponential distribution: a survey. *AStA Advances in Statistical Analysis,* 95, pp. 219–251.

Raqab, M. M., Ahsanullah, M., (2001). Estimation of the location and scale parameters of generalized exponential distribution based on order statistics. *Journal of Statistical Computation and Simulation,* 69(2), pp. 109–123.

Ristić, M. M., Balakrishnan, N., (2012). The gamma-exponentiated exponential distribution. *Journal of Statistical Computation and Simulation,* 82(8), pp. 1191–1206.

Sabry, M. A., Shaaban, M., (2020). Dependent ranked set sampling designs for parametric estimation with applications. *Annals of Data Science,* 7(2), pp. 357–371, https://doi.org/10.1007/s40745-020-00247-3.

Samawi, H. M., Ahmed, M. S., Abu-Dayyeh, W., (1996). Estimating the population mean using extreme ranked set sampling. *Biometrical Journal,* 38(5), pp. 577–586.

Samuh, M. H., Qtait, A., (2015). Estimation for the parameters of the exponentiated exponential distribution using a median ranked set sampling. *Journal of Modern Applied Statistical Methods,* 14(1), pp. 215–237.

Tahmasebi, S., Hosseini, E. H., Jafari, A. A., (2017). Bayesian estimation for Rayleigh distribution based on ranked set sampling. *New Trends in Mathematical Sciences,* 5(4), pp. 97–106.

Wolfe, D. A., (2010). Ranked set sampling. *Wiley Interdisciplinary Reviews: Computational Statistics,* 2(4), pp. 460–466.

Zamanzade, E., Al-Omari, A. I., (2016). New ranked set sampling for estimating the population mean and variance. *Hacettepe Journal of Mathematics and Statistics,* 45(6), pp. 1891–1905.

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 59–76, DOI 10.2478/stattrans-2022-0042
Received – 21.01.2022; accepted – 24.05.2022

The Weibull lifetime model with randomised failure-free time

Piotr Sulewski[1], **Magdalena Szymkowiak**[2]

ABSTRACT

The paper shows that treating failure-free time in the three-parameter Weibull distribution not a constant, but as a random variable makes the resulting distribution much more flexible at the expense of only one additional parameter.

Key words: Weibull lifetime model, randomised failure-free time, compound Weibull distributions.

1. Introduction

In the reliability domain we routinely treat time to failure of a particular technical product as a random variable. Finding the proper model that fits the reliability data is the main problem of reliability engineers and applied statisticians. The Weibull distribution, which has a particular significance in the reliability theory, is named after the Swedish physicist Waloddi Weibull, who was the first to promote the usefulness of the distribution to model reliability data sets of widely differing character (see, e.g. Weibull, 1951, and Murthy et al., 2004).

Recall that the two-parameter Weibull distribution ($2pW$) has the following cumulative distribution function (cdf)

$$F_{2pW}(t;a,b) = 1 - \exp\left[-\left(\frac{t}{a}\right)^b\right] \quad \text{for} \quad t > 0 \tag{1}$$

and the probability density function (pdf), based on (1), equal to

$$f_{2pW}(t;a,b) = \frac{b}{a}\left(\frac{t}{a}\right)^{b-1} \exp\left[-\left(\frac{t}{a}\right)^b\right] \quad \text{for} \quad t > 0, \tag{2}$$

where $a > 0$ and $b > 0$ are the scale and shape parameters, respectively. The hazard (failure) rate function (hrf) $h(t) = \frac{f(t)}{P(T>t)} = \frac{f(t)}{1-F(t)}$, interpreted as the instantaneous failure rate of a particular product occurring immediately after the time point t, given that the product has survived until the time point t, has for $2pW$, using (1) and (2), the following form

$$h_{2pW}(t;a,b) = ba^{-b}t^{b-1} \quad \text{for} \quad t > 0.$$

[1]Institute of Exact and Technical Sciences, Pomeranian University, Poland.
E-mail: piotr.sulewski@apsl.edu.pl.
[2]Institute of Automatic Control and Robotics, Poznan University of Technology, Poland.
E-mail: magdalena.szymkowiak@put.poznan.pl.

It can be increasing, decreasing or constant depending on $b > 1$, $b < 1$ or $b = 1$, respectively. It is easy to note that in the last mentioned case, when $b = 1$, we get the exponential distribution, the most standard distribution in the reliability theory, with a constant hazard rate. It is a well-known fact that the constant hazard rate function $h(t) = \frac{1}{a}$ characterizes the family of exponential distribution with scale parameter a.

Further, let us define another reliability function known as the aging intensity function (aif) in the form

$$L(t) = \frac{h(t)}{\frac{1}{t}\int_0^t h(u)\mathrm{d}u} = \frac{-tf(t)}{[1-F(t)]\ln[1-F(t)]} \quad \text{for} \quad t > 0, \tag{3}$$

being the ratio of the instantaneous hazard rate to its average and expressing the product average aging behaviour (see, e.g., Szymkowiak, 2018a). For $2pW$ it is constant

$$L_{2pW}(t;a,b) = b \quad \text{for} \quad t > 0. \tag{4}$$

This constant aging intensity $L(t) = b$ characterizes the subfamily of the family of $2pW$ with a fixed shape parameter b and varying scale parameter a (Szymkowiak, 2020). Moreover, the aif equal to 1, $L(t) = 1$, characterizes the family of exponential distributions.

However, certain lifetime data (i.a., human mortality, machine life-cycles and some biological studies) require non-monotonic shapes of the hazard rate, e.g., a bathtub shape or a unimodal (upside-down bathtub) shape. Therefore, many researchers have developed various modified forms of the Weibull distribution to achieve non-monotonic shapes of hazard function, i.a., Drapella, 1993, introduced the complementary Weibull distribution ($2pCW$), known also as the inverse Weibull distribution, with the following cdf

$$F_{2pCW}(t;a,b) = \exp\left[-\left(\frac{a}{t}\right)^b\right] \quad \text{for} \quad t > 0.$$

Further extensive literature is also available on modifications of the standard Weibull (see, e.g., Murthy et al., 2004, Almaki and Nadarajah, 2014, Lai, 2014), which in some cases involve one or more additional parameters. For example, the exponentiated Weibull distribution ($2pEW$) with a bathtub hazard rate function has the following cdf

$$F_{2pEW}(t;a,b,d) = \left\{1-\exp\left[-\left(\frac{t}{a}\right)^b\right]\right\}^d \quad \text{for} \quad t > 0$$

with $d > 0$ being a new shape parameter (Mudholkar and Srivastava, 1993).

To be precise, Waloddi Weibull introduced his distribution as a three-parameter model $3pW$ (known also as the shifted Weibull distribution) with an additional location parameter $\tau \geq 0$ and the following cdf

$$F_{3pW}(t;a,b,\tau) = 1-\exp\left[-\left(\frac{t-\tau}{a}\right)^b\right] \quad \text{for} \quad t > \tau, \tag{5}$$

pdf, based on (5), equal to

$$f_{3pW}(t;a,b,\tau)=\frac{b}{a}\left(\frac{t-\tau}{a}\right)^{b-1}\exp\left[-\left(\frac{t-\tau}{a}\right)^{b}\right] \quad \text{for} \quad t>\tau, \tag{6}$$

and hrf, based on (5) and (6), equal to

$$h_{3pW}(t;a,b,\tau)=ba^{-b}(t-\tau)^{b-1} \quad \text{for} \quad t>\tau.$$

Note that for a distribution with support $(\tau,+\infty)$ we use a modified version of aif, known as the support dependent aif (see Szymkowiak, 2018b)

$$L^{s}(t)=\frac{h(t)}{\frac{1}{t-\tau}\int_{\tau}^{t}h(u)\mathrm{d}u}=\frac{(\tau-t)f(t)}{[1-F(t)]\ln[1-F(t)]}, \tag{7}$$

which for $\tau=0$ corresponds to the classical definition (3). Then, the subfamily of $3pW$ distributions with fixed shape parameter b and location parameter τ, and varying scale parameter, a is characterized by constant support dependent aif

$$L^{s}_{3pW}(t;a,b,\tau)=b \quad \text{for} \quad t>\tau,$$

(compare formula (4)).
Figure 1 presents hrf (on the left) and support dependent aif (on the right) of the $3pW(a,b,\tau)$ for the scale parameter $a=1$ and the location parameter $\tau=1$, and different values of shape parameter b. As one can note hrf is decreasing for $b<1$ and increasing if $b>1$. When $b=1$, we get the shifted exponential distribution (see, e.g., Szymkowiak, 2020) with the constant hrf. On the other hand, the support dependent aging intensity functions, shown in the figure on the right, are constant, equal to b, for all $b>0$.

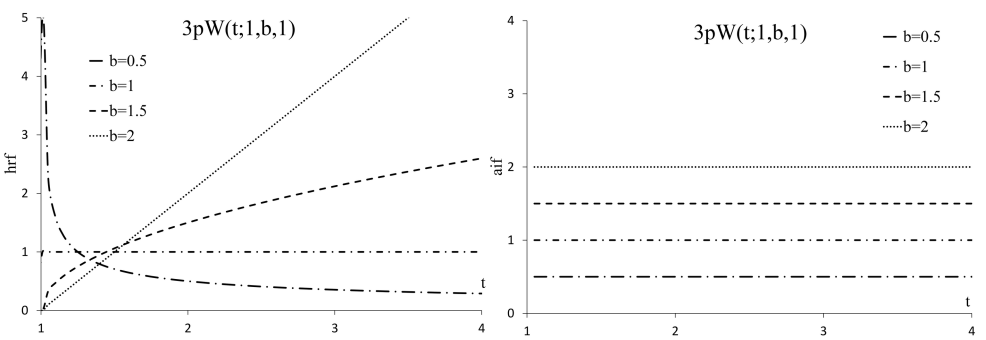


Figure 1: hrf and aif of the 3pW

Under the assumption of $3pW$ distribution, none failure of the analysed product can possibly occur prior to the time τ, therefore the location parameter τ is also referred to as failure-free time or minimum life.

In lifetime data analysis, first we try to find the proper model that best fits the data. Parameter estimation is the second step of our modeling process. In three-parameter Weibull distribution, the determination of a suitable location parameter τ is not a simple task.

The first and simplest τ estimate is $\hat{\tau} = t_{(1)}$, where $t_{(1)}$ is the smallest value in the order data set, or $\hat{\tau} = t_{(1)} - \frac{1}{n}$, where n is a sample size (see, e.g., Murthy et al., 2004). O'Connor, 2012, suggested an alternative procedure to estimate the location parameter given by

$$\hat{\tau} = t_{(2)} - \frac{\left(t_{(3)} - t_{(2)}\right)\left(t_{(2)} - t_{(1)}\right)}{\left(t_{(3)} - t_{(2)}\right) - \left(t_{(2)} - t_{(1)}\right)},$$

where further on $t_{(2)}$ and $t_{(3)}$ are the second and third value in the order data set. Drapella, 1999, improved this method and also Kececioglu, 1991, discussed two other methods to obtain estimates of τ.

In our paper, apart from the fact that we assume that the time to failure follows $3pW$ distribution, we also suggest that the failure-free time (location parameter) can be considered as a new random variable. It allows this parameter to vary and makes the estimation model more complex.

It is well-known fact (see, e.g., Qutb and Rajhi, 2016) that if X is a random variable following the known parametrized distribution with pdf f_X, and one of its parameters θ is considered as a new random variable Y with a specified pdf f_Y then a compound random variable T has a distribution with the following pdf

$$f_T(t) = \int_\theta f_X(t|\theta) f_Y(\theta) \mathrm{d}\theta \quad \text{for} \quad t > 0, \tag{8}$$

where $f_X(t|\theta)$ is a conditional density function depending on the parameter θ.

The compound Weibull distribution with random parameters was introduced earlier, e.g., by Dubey, 1968, and Qutb and Rajhi, 2016, but as far as we know, its location parameter has not yet been considered as being random.

The rest of our paper is organized as follows. In Section 2 the compound Weibull lifetime model with random failure free time is defined. In Section 3 four candidates for being the distribution of the random location parameter are presented. Section 4 contains analysis of three real lifetime data that compares the defined compound Weibull distributions with the standard three-parameter one. The conclusions are presented in Section 5.

2. Failure-free time as random variable

In this section the Weibull lifetime model with random failure-free time denoted as $4pWY$ is defined. Its pdf, according to formula (8), has a form of the convolution integral, namely

$$f_{4pWY}(t;a,b,c,d) = \int_0^t f_{3pW}(t;a,b,\tau) f_Y(\tau;c,d)\, d\tau \quad \text{for} \quad t > 0, \tag{9}$$

where f_Y is pdf of the failure-free time distribution with parameters c and d. For details related to the above formula please consult any advanced textbook on probability theory, e.g. Rossberg et al., 1985.

Cdf of the $4pWY$, based on (9), is given by

$$F_{4pWY}(t;a,b,c,d) = \int_0^t \left[1-\exp\left(-\left(\frac{t-\tau}{a}\right)^b\right)\right] F_Y(\tau,c,d)\,d\tau \quad \text{for} \quad t>0, \tag{10}$$

where F_Y is cdf of the failure-free time distribution with parameters c and d.

Hrf of the $4pWY$, using (9) and (10), is obviously defined as

$$h_{4pWY}(t;a,b,c,d) = \frac{f_{4pWY}(t;a,b,c,d)}{1-F_{4pWY}(t;a,b,c,d)} \quad \text{for} \quad t>0, \tag{11}$$

and aif of the $4pWY$ using (11) is given by

$$L_{4pWY}(t;a,b,c,d) = \frac{h_{4pWY}(t;a,b,c,d)}{\frac{1}{t}\int_0^t h_{4pWY}(u;a,b,c,d)\,du} \quad \text{for} \quad t>0. \tag{12}$$

Regarding the calculations, unfortunately, there will not always be analytical formulas to which (9)-(12) would be transformed. All applications of the $4pWY$ lifetime model will often be numerical. Fortunately, this is not an obstacle these days. Anyone who decides to use $4pWY$ to evaluate reliability must be equipped with a powerful computing environment. Fortunately, we have Excel, Mathcad, Matematica, Matlab, Scilab, and maybe a few other powerful, less known computing environments.

3. Four candidates for the failure-free time model

Now, four compound Weibull random variables with different distributions of the random location parameter will be presented. As the first distribution of the random location parameter we propose the Uniform distribution $U(c,d)$ giving a smooth transformation from $3pW$ to the compound Weibull-Uniform distribution $4pWU$. As the second very natural candidate – we propose the Weibull distribution $W(c,d)$. The next one will be the very popular two-parameter Gamma distribution $G(c,d)$ being the generalization of exponential, Erlang and chi-square distributions. As the last one we use the Normal distribution $N(c,d)$ with possibly positive support.

For all the proposed models, their hrf and aif for different parameters are determined and plotted (using formulas (11) and (12)). Determination of statistical measures of the presented random variables, such as their ordinary and central moments, quantiles, etc., because of their complex distribution forms, is possible only with the numerical calculations.

3.1. Compound Weibull-Uniform distribution

The first candidate for the failure-free time model is a Uniform distribution on interval $[c-d,c+d]$ denoted $U(c-d,c+d)$. For the purposes of the convolution integral (9) we can write pdf of the $U(c-d,c+d)$ using the Heaviside step function H in the form

$$f_U(t;c,d) = \frac{H(t-c+d)-H(t-c-d)}{2d} \quad \text{for} \quad t>0, c\geq d$$

where $c > 0$ and $d > 0$ are the position and scale parameters, respectively, as well as

$$H(x) = \begin{cases} 0 & \text{for } x < 0 \\ 0.5 & \text{for } x = 0 \\ 1 & \text{for } x > 0. \end{cases}$$

The Heaviside step function we can write as $H(x) = \int_{-\infty}^{x} \delta(s)\, ds$, where $\delta(x)$ is the Dirac function. Let us define the Dirac function $\delta(t;\tau)$ as

$$\delta(t;\tau) = \begin{cases} 0 & \text{for } t = \tau \\ \infty & \text{for } t \neq \tau. \end{cases}$$

The Dirac function fulfils, as any probability distribution, Normalization condition $\int_{-\infty}^{\infty} \delta(t;\tau)\, dt = 1$. This function can also be defined as the limit of the sequence of τ - centered Normal distributions, namely

$$\lim_{d\to 0} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -0.5 \left(\frac{t-\tau}{\sigma} \right)^2 \right\}.$$

Cdf of the $U(c-d, c+d)$ calculated in the Mathematica software is defined as

$$F_U(t;c,d) = I_1 + I_2,$$

where

$$I_1 = \frac{c+d-t-(c+d)H(-c-d)}{2b} H(t-c-d),$$
$$I_2 = \frac{-c+d+t+(c-d)H(d-c)}{2b} H(t-c+d).$$

Let $4pWU(a,b,c,d)$ denotes the four parameters compound Weibull-Uniform distribution. Then pdf of this distribution, based on (9), is given by

$$f_{4pWU}(t;a,b,c,d) = \frac{b}{2a^b d} \int_0^t \frac{H(\tau-c+d) - H(\tau-c-d)}{(t-\tau)^{1-b} \exp\left[\left(\frac{t-\tau}{a} \right)^b \right]} d\tau \quad \text{for} \quad t > 0. \tag{13}$$

The integral in (13) is possible to be calculated, namely

$$f_{4pWU}(t;a,b,c,d) = I_1 + I_2,$$

where

$$I_1 = \frac{b}{2a^b d} \int_0^t \frac{H(\tau-c+d)}{(t-\tau)^{1-b} \exp\left[\left(\frac{t-\tau}{b} \right)^b \right]} d\tau, \tag{14}$$

$$I_2 = \frac{b}{2a^b d} \int_0^t \frac{H(\tau-c-d)}{(t-\tau)^{1-b} \exp\left[\left(\frac{t-\tau}{b} \right)^b \right]} d\tau. \tag{15}$$

Using the Mathematica software by (14) we obtain

$$I_1 = \frac{H(t+d-c)\left[1+\exp\left[-\left(\frac{t+d-c}{a}\right)^b\right][H(d-c)-1]-\exp\left[-\left(\frac{t}{a}\right)^b\right]H(d-c)\right]}{2d}$$

and by (15), we get

$$I_2 = \frac{H(t-d-c)\left[1+\exp\left[-\left(\frac{t-d-c}{a}\right)^b\right][H(-c-d)-1]-\exp\left[-\left(\frac{t}{a}\right)^b\right]H(-c-d)\right]}{2d}.$$

It turns out that we can pass from $3pW$ (6) to $4pWU$ (13) smoothly. If $d \to 0$, then the $4pWU$ tends to the $3pW$ (see Figure 2).

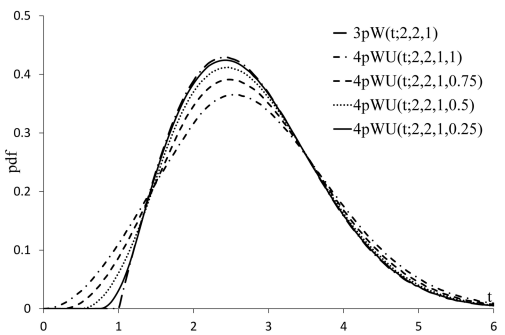


Figure 2: pdf of the $3pW$ and $4pWU$

Cdf of the $4pWU$, based on (13), is given by

$$\begin{aligned} F_{4pWU}(t;a,b,c,d) &= \frac{1}{2d}\int_0^t \left[1-\exp\left[-\left(\frac{t-\tau}{b}\right)^b\right]\right] \\ &\times \left[H(\tau-c+d)-H(\tau-c-d)\right]d\tau \quad \text{for} \quad t>0. \end{aligned}$$

Figure 3 presents hrf of the $4pWU(a,b,c,d)$ for various parameter values. For $b=0.5$ (Figure 3, left) hrf increases very quickly and then decreases very slowly. The maximum shifts to the right as d increases. For $d=0.75$ (Figure 3, right) hrf increases slowly and then decreases very quickly. The maximum shifts to the left as b increases.

Figure 4 presents aif of the $4pWU(a,b,c,d)$ for various parameter values. For $a=1$, $b=0.5$, $c=1$ regardless of the parameter d (Figure 4, left), aif (after some fluctuations for small t) tends to a constant function. For $a=1$, $c=1$, $d=0.75$ regardless of the parameter b (Figure 4, right), aif decreases for small t and then also tends to a constant function.

To generate data that follows a compound Weibull-Uniform distribution we provide that if $T_{4pWU} \sim 4pWU(a,b,c,d)$ and $T_U \sim U(c-d,c+d)$, $T_{2pW} \sim 2pW(a,b)$, $R \sim U(0,1)$ then generator of T_{4pWU} is given by the formula

$$T_{4pWU} = T_{2pW} + T_U = a\left[-\ln(1-R)\right]^{1/b} + (c-d) + 2dR.$$

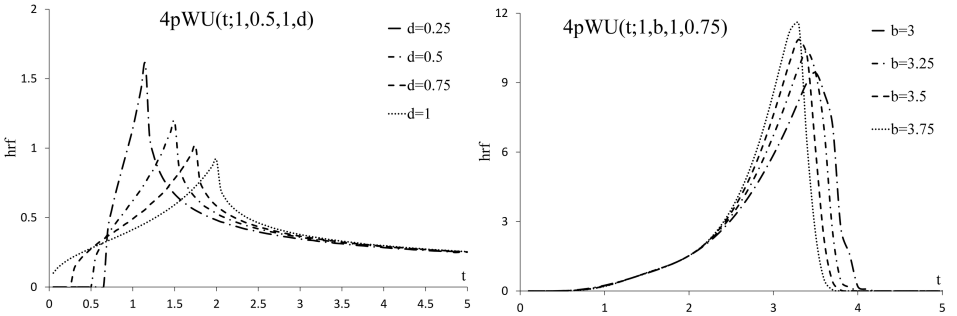


Figure 3: hrf of the $4pWU$

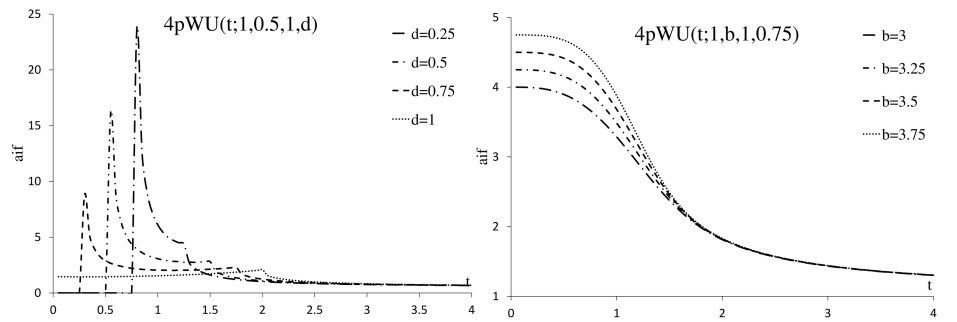


Figure 4: aif of the $4pWU$

3.2. Compound Weibull-Weibull distribution

The second candidate for the failure-free time model is the Weibull distribution with pdf given by formula (2). Let $4pWW(a,b,c,d)$ denotes the four parameters compound Weibull-Weibull distribution. Pdf of this distribution, based on (9), is given by

$$f_{4pWW}(t,a,b,c,d) = \frac{bd\int_0^t \left(\frac{t-\tau}{a}\right)^{b-1}\left(\frac{\tau}{c}\right)^{d-1}\exp\left[-\left(\frac{t-\tau}{a}\right)^{b}-\left(\frac{\tau}{c}\right)^{d}\right]d\tau}{ac} \quad t>0. \tag{16}$$

Cdf of the $4pWW$, based on (16), is given by

$$F_{4pWW}(t,a,b,c,d) = \frac{b\int_0^t \left[1-\exp\left[-\left(\frac{\tau}{c}\right)^{d}\right]\right]\left(\frac{t-\tau}{a}\right)^{b-1}\exp\left[-\left(\frac{t-\tau}{a}\right)^{b}\right]d\tau}{a} \quad t>0$$

Figure 5 presents hrf of the $4pWW(a,b,c,d)$ for various parameter values. For $d=0.5$ (see Figure 6, left) the hrf is a decreasing function. In other cases, the hrf increases strongly and then slowly decreases. The larger the d, the higher the maximum. For $b=2$ (see Figure 5, right) we obtain the inverse-bathtub hrf. For $b=1$ the hrf is initially an increasing function and then remains constant.

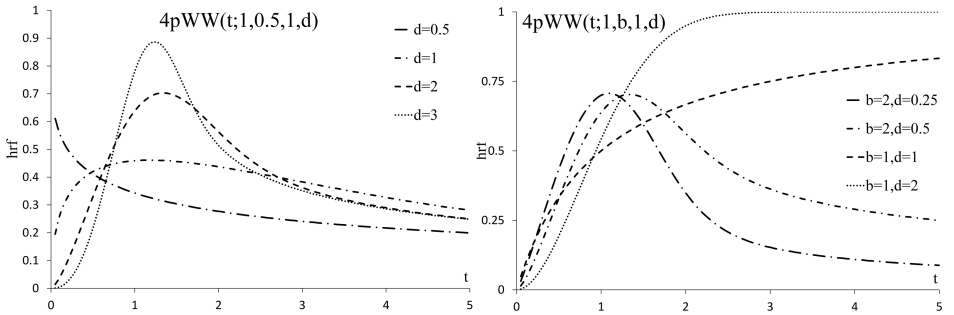


Figure 5: hrf of the $4pWW$

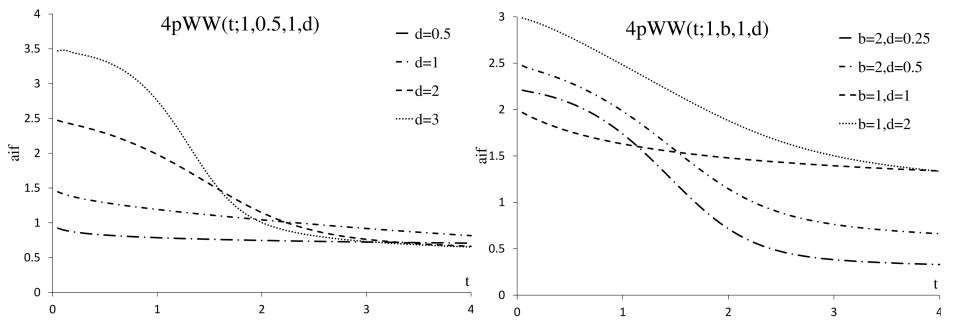


Figure 6: aif of the $4pWW$

Figure 6 presents aif of the $4pWW(a,b,c,d)$ for various parameter values. For $a=1$, $b=0.5$, $c=1$ regardless of the parameter d (Figure 6, left), aif decreases for small t and tends to a constant function. Also, for $a=1$, $b=2$ or $b=1$ and $c=1$, regardless of the parameter d (Figure 6, right), aif decreases for small t and then tends to a constant function.

To generate data that follows a compound Weibull-Weibull distribution we provide that if $T_{4pWW} \sim 4pWW(a,b,c,d)$, $T_{2pW1} \sim 2pW(a,b)$, $T_{2pW2} \sim 2pW(c,d)$, $R \sim U(0,1)$ then the generator of T_{4pWW} is given by the formula

$$T_{4pWW} = T_{2pW1} + T_{2pW2} = a\left[-\ln(1-R)\right]^{1/b} + c\left[-\ln(1-R)\right]^{1/d}.$$

3.3. Compound Weibull-Gamma distribution

The third candidate for the failure-free time model is the Gamma distribution with pdf

$$f_G(t;c,d) = \frac{1}{c^d \Gamma(d)} t^{d-1} \exp\left(-\frac{t}{c}\right) \quad \text{for} \quad t>0,$$

where $c>0, d>0$ are the scale and shape parameters, respectively.

Let $4pWG(a,b,c,d)$ denote the four-parameters compound Weibull-Gamma distribution then pdf of this distribution, based on (9), is given by

$$f_{4pWG}(t;a,b,c,d) = \frac{b\int_0^t (t-\tau)^{b-1}\,\tau^{d-1}\exp\left[-\frac{\tau}{c}-\left(\frac{t-\tau}{a}\right)^b\right]d\tau}{a^b c^d \Gamma(d)} \quad \text{for} \quad t>0. \qquad (17)$$

Cdf of the $4pWG$, based on (17), is given by

$$F_{4pWG}(t;a,b,c,d) = \frac{\int_0^t \tau^{d-1}\exp\left[-\frac{\tau}{c}\right]\left[1-\exp\left[-\left(\frac{t-\tau}{a}\right)^b\right]\right]d\tau}{c^d\Gamma(d)} \quad \text{for} \quad t>0.$$

Decades pass, but the Weibull plotting technique remains irreplaceable in failure data analysis. Therefore, Figure 7 shows appropriately transformed cdf of the $2pW$, $3pW$ and $4pWG$ plotted on the Weibull probability paper. Of course, the transformed cdf of $2pW$ appears as straight lines. In contrast, both transformed cdf of $3pW$ and of $4pWG$ appear as curves convex upward. As a rule the transformed cdf of $4pWG$ is less convex then the transformed cdf of $3pWG$.

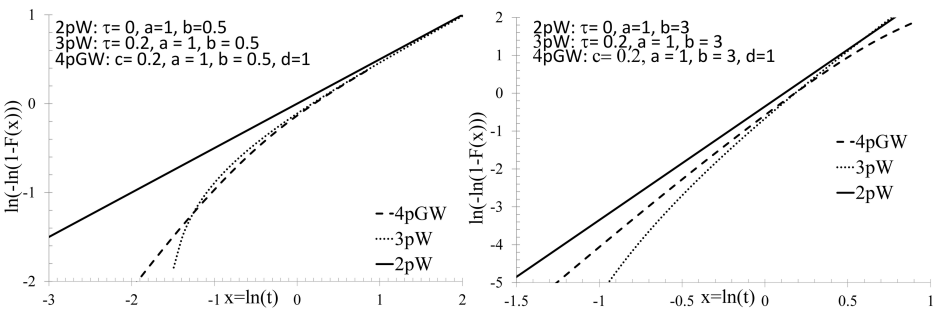


Figure 7: cdf of the $2pW$, $3pW$ and $4pWG$ plotted on the Weibull probability paper

Figure 8 has been prepared to express artificiality the stepwise $3pW$ (denoted as ps0) and compared it with the smooth $4pWG$. Table 1 contains parameter values of the Gamma component of the $4pWG$ used in Figure 8.

The $2pW$ and $3pW$ offer monotonic hazard rate functions, strictly decreasing or increasing ones, always convex downward. In contrast, $4pWG$ offers much more flexible hrf that may be non-monotonic and may have even two points of inflection. The main competitors of the $4pWG$ are the Complementary Weibull distribution (see Rossberg et al., 1985) and LogNormal distribution (see O'Connor, 2012). But they have two-parameter only. Let us look closely at Figure 8. One can pick up something resembling Moivre-Laplace limit process. Let us remember that when $p \to 0$, $n \to \infty$ and pn remains constant, the Binomial distribution tends to the Poisson distribution. By similarity, when $c \to 0$, $d \to \infty$ and cd remains constant, then $4pWG$ tends to $3pW$ (see Table 1).

Table 1. Sets of parameters of the Gamma component of the $4pWG$

Set of parameters	ps0*	ps1	ps2	ps3	ps4	ps5	ps6	ps7
c	0.2	0.2	0.1	0.05	0.033	0.025	0.02	0.01
d	-	1	2	4	6	8	10	20

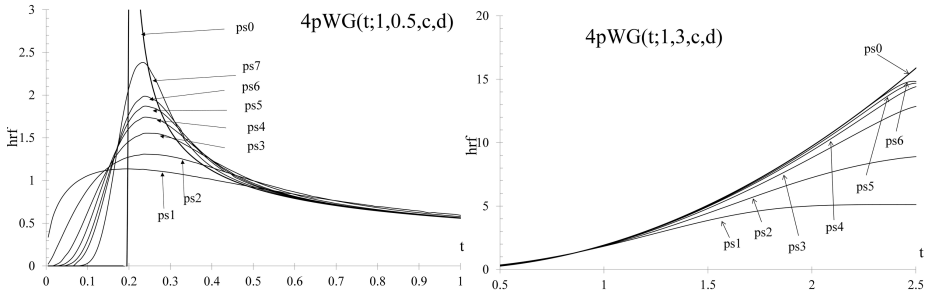


Figure 8: hrf of the $4pWG$

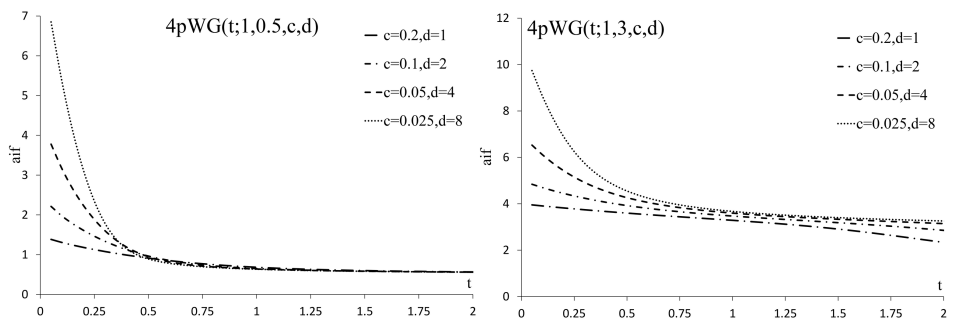


Figure 9: aif of the $4pWG$

Figure 9 presents aif of the $4pWG(a,b,c,d)$ for various parameter values. For $a=1$, $b=0.5$ (Figure 9, left) or $a=1$, $b=3$ (Figure 9, right), regardless of the parameters c and d, aif decreases for small t and tends to a constant function.

To generate data that follow a compound Weibull-Gamma distribution we provide that if $T_{4pWG} \sim 4pWG(a,b,c,d)$ and $T_G \sim G(c,d), T_{2pW} \sim 2pW(a,b)$ then the generator of T_{4pWG} is given by the formula

$$T_{4pWG} = T_{2pW} + T_G = a\left[-\ln(1-R)\right]^{1/b} + T_G$$

The generator of T_G, implemented in R software, for $d \geqslant 1$ and $0 < d < 1$ is described in Ahrens and Dieter, 1982, and Ahrens and Dieter, 1974, respectively.

3.4. Compound Weibull-Normal distribution

Many researchers may be tempted to replace the Gamma distribution with the Normal distribution with pdf

$$f_N(t;c,d) = \frac{1}{\sqrt{2\pi}d} \exp\left[-0.5\left(\frac{t-c}{d}\right)^2\right] \quad \text{for} \quad t>0,$$

where $c>0, d>0$ are the position and scale parameters, respectively (to ensure that Normal distribution has positive support – the failure-free time should not be negative – using the three-sigma rule we assume that $c>3.3d$). The argument was that when the shape parameter d increases, then the Gamma distribution tends to the Normal distribution. It directly proceeds from Lindeberg-Levy Limit theorem. Such a replacement of the Gamma distribution with the Normal one, has both advantages and disadvantages.

The advantages are:

- Firstly, one can skip this strong assumption that between-two-portions time interval follows the exponential distribution. Taking the Central Limit Theorem of Lapunov as a base, one can admit that particular intervals follow different distributions.

- Secondly, applying the Normal distribution as that between-two-portions time interval distribution, one makes $4pWN$ more flexible. It is because the mean value $E(t)$ and variance $D(t)$ of the Gamma distribution are strongly interrelated because $E(t)=cd$, $D(t)=cd^2$.

- Thirdly, the problem of interpretation of non-integer d value disappears.

The disadvantages are:

- Firstly, the moments mentioned above are to some extent interrelated. It is because the value of the coefficient of variation $\gamma_0 = D(t)/E(t)$ has to be carefully chosen for probability of negative t values to be negligible, for instance γ_0 has to be kept not greater than $1/3$.

- Secondly, it is true that the left-censored Normal distribution can alternatively be applied, and values of $\mathrm{E}(t)$ and $\mathrm{D}(t)$ can be freely set, but then the threshold function, that we wanted to eliminate, returns.

Let $4pWN(a,b,c,d)$ denotes the four-parameters compound Weibull-Normal distribution. Then, pdf of this distribution, based on (9), is given by

$$\begin{aligned} f_{4pWN}(t;a,b,c,d) &= \frac{b}{\sqrt{2\pi}ad}\int_0^t \left(\frac{t-\tau}{a}\right)^{b-1} \\ &\times \exp\left[-\frac{1}{2}\left(\frac{\tau-c}{d}\right)^2 - \left(\frac{t-\tau}{a}\right)^b\right] d\tau \quad \text{for} \quad t>0,\ c>3.3d. \end{aligned} \tag{18}$$

Cdf of the $4pWN$, using (18), is given by

$$F_{4pWN}(t;a,b,c,d) = \frac{b\int_0^t \Phi\left(\frac{\tau-c}{d}\right)\left(\frac{t-\tau}{a}\right)^{b-1}\exp\left[-\left(\frac{t-\tau}{a}\right)^{b}\right]d\tau}{a} \quad \text{for } t>0,\ c>3.3d,$$

where Φ is cdf of the standard Normal distribution.

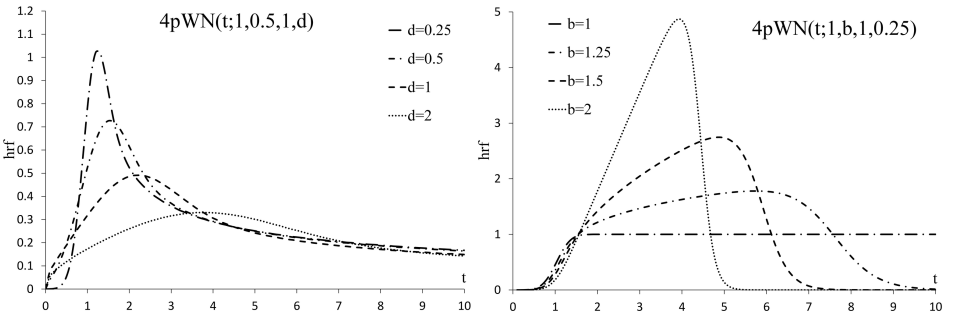


Figure 10: hrf of the 4pWN

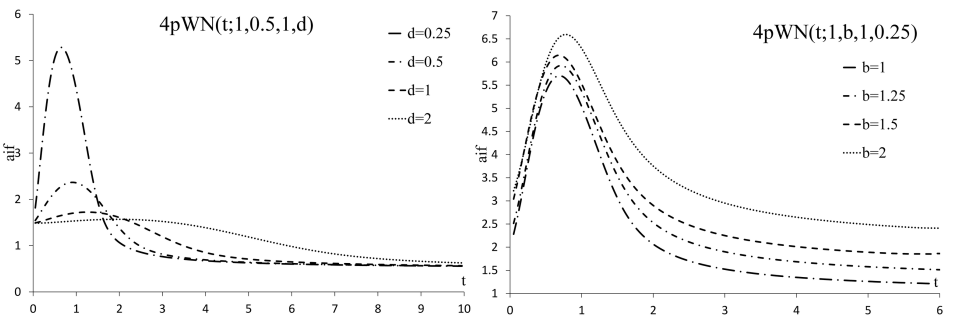


Figure 11: aif of the 4pWN

Figure 10 presents hrf of the $4pWN(a,b,c,d)$ for various parameter values. For $b=0.5$ (see Figure 10, left) hrf increases very quickly and then decreases very slowly. The maximum shifts to the right as d increases. Figure (10,right) presents an upside-down bathtub-shaped hrf except the case $b=1$. Then, initially hrf increases and then remains constant.

Figure 11 presents aif of the $4pWN(a,b,c,d)$ for various parameter values. For $a=1$, $b=0.5$, $c=1$, regardless of the parameter d (Figure 11, left), aif (after some fluctuations for small t) tends to the constant function. For $a=1$, $c=1$, $d=0.25$, regardless of the parameter b (Figure 11, right), aif decreases for small t and tends to the constant function.

To generate data that follows a compound Weibull-Normal distribution we provide that if $T_{4pWN} \sim 4pWN(a,b,c,d)$ and $T_N \sim N(c,d), T_{2pW} \sim 2pW(a,b), R \sim U(0,1)$ then the

generator of T_{4pWN} is given by the formula

$$T_{4pWN} = T_{2pW} + T_N = a\left[-\ln(1-R)\right]^{1/b} + \Phi^{-1}(R,c,d),$$

where Φ^{-1} is the inverse cdf of the Normal distribution implemented in R software (see Wichura, 1988).

At the end of this section, let us mention that any infinitely divisible (stable) probability distribution (see Seidel, 2010) with the positive support would be a good candidate for the failure-free time model. The only problem is to determine basic statistical functions for such a candidate.

4. Real lifetime data analysis

To demonstrate the flexibility and applicability of the new compound models in lifetime data analysis we consider three real data examples. Among the known random variables, the standard $3pW$ distribution seems to be the most natural choice for comparing the goodness-of-fit approach of the proposed models.

Example 1 is devoted to 33 operating times (in hours) of the first phase of the construction machine: 0.25, 9.25, 17.75, 22, 22.25, 22.5, 22.5, 23, 23.25, 30.25, 35.75, 41.75, 42.5, 48.75, 49.75, 51, 58.75, 63, 68.25, 72.5, 75.5, 127.5, 138.5, 140.5, 141, 146.5, 173, 193.5, 218.25, 237.5, 257.75, 312, 352 (see, e.g., Mahmood, 2021, Saffawy and Algmal, 2006).

Example 2 concerns leukaemia free survival times (in months) of 51 autologous transplant patients: 0.658, 0.822, 1.414, 2.500, 3.322, 3.816, 4.737, 4.836, 4.934, 5.033, 5.757, 5.855, 5.987, 6.151, 6.217, 6.447, 8.651, 8.717, 9.441, 10.329, 11.480, 12.007, 12.007, 12.237, 12.401, 13.059, 14.474, 15.000, 15.461, 15.757, 16.480, 16.711, 17.204, 17.237, 17.303, 17.644, 18.092, 18.092, 18.750, 20.625, 23.158, 27.730, 31.184, 32.434, 35.921, 42.237, 44.638, 46.480, 47.467, 48.322, 56.086 (see, e.g., LaiXie, 2006).

Example 3 refers to survival times (in days from diagnosis) of 43 patients suffering chronic granulocytic leukaemia: 7, 47, 58, 74, 177, 232, 273, 285, 317, 429, 440, 445, 455, 468, 495, 497, 532, 571, 579, 581, 650, 702, 715, 779, 881, 900, 930, 968, 1077, 1109, 1314, 1334, 1367, 1534, 1712, 1784, 1877, 1886, 2045, 2056, 2260, 2429, 2509 (see, e.g., Lai and Xie, 2006).

To estimate parameters of the considered models, in addition to commonly known estimation tools such as the maximum likelihood (not quite adequate in the case of three-parameter Weibull distribution, (see e.g. Murthy et al., 2004, Lam, 2010, Ramakrishnan, 2017, Park, 2018) or least squares methods, also goodness-of-fit tests can be used. For example, Kendall and Stuart, 1961 presented the minimum chi-square test statistic method in parameter estimation. Moreover, Weber, 2006, used the minimum Kolmogorov–Smirnov test statistic method to estimate the distribution parameters. In our analysis we apply the latter tool.

To avoid local maxima, the optimization routine was run with several different starting values that are widely scattered in the parameter space. The p-values for a given model were calculated as follows. Let Θ be the vector of model parameters. Having estimated parameters vector $\widehat{\Theta}$ for a given sample of size n, we calculate test statistics $T\left(\widehat{\Theta}, n\right)$. Next,

we generate 10^4 samples of size n for the given model with the estimated parameters vector $\widehat{\Theta}$. For each obtained sample s, we calculated the $T_i^s\left(\widehat{\Theta},n\right)$. Finally, the p-value is given by (see e.g. Balakrishnan and Ristic, 2016)

$$p \approx \#\left\{i : T_i^s\left(\widehat{\Theta},n\right) > T\left(\widehat{\Theta},n\right)\right\} 10^{-4}$$

Tables 2–4 present the estimated parameters of the analysed models, test statistics and p-values (in parentheses) calculated by the Kolmogorov-Smirnow (KS), Anderson-Darling (AD) and Cramer von Mises (CVM) tests. The lowest statistics values (the highest p-values) are noted in bold. The determined values show that for all the exemplary lifetime data sets, there are some compound Weibull distributions that fit better to the data then the standard three-parameter Weibull distribution $3pW$. For the first data set, two tests point to the compound Weibull-Gamma distribution, $4pWG$, as the distribution that best fits the data, and one test points to the compound Weibull-Weibull distribution, $4pWW$, as the best one (see Table 2). Further, for the second data set, all the tests point to the compound Weibull-Weibull distribution, $4pWW$, as the distribution that best fits the data (see Table 3). Finally, for the third data set, two tests point to the compound Weibull-Uniform distribution, $4pWU$, and one test points at the compound Weibull-Weibull distribution, $4pWW$, as the best models (see Table 4).

Table 2. Goodness-of-fit tests. Example 1

Model	Estimated parameters	KS	AD	CVM
$3pW$	$\widehat{a}=96.91, \widehat{b}=1.045, \widehat{\tau}=0.06$	0.1(0.864)	0.543(0.698)	0.082(0.673)
$4pWU$	$\widehat{a}=90.48, \widehat{b}=0.92, \widehat{c}=5.76, \widehat{d}=5.81$	0.094(0.910)	0.513(0.727)	0.062(0.810)
$4pWW$	$\widehat{a}=90.90, \widehat{b}=0.93, \widehat{c}=4.71, \widehat{d}=0.79$	0.095(0.901)	0.496(0.747)	0.067(0.764)
$4pWG$	$\widehat{a}=80.90, \widehat{b}=0.69, \widehat{c}=4.99, \widehat{d}=2.87$	0.083(0.962)	0.660(0.586)	0.045(0.901)
$4pWN$	$\widehat{a}=91.15, \widehat{b}=0.93, \widehat{c}=5.55, \widehat{d}=0.92$	0.094(0.913)	0.994(0.359)	0.064(0.793)

Table 3. Goodness-of-fit tests. Example 2

Model	Estimated parameters	KS	AD	CVM
$3pW$	$\widehat{a}=15.92, \widehat{b}=1.20, \widehat{\tau}=0.64$	0.076(0.910)	0.673(0.580)	0.071(0.746)
$4pWU$	$\widehat{a}=15.56, \widehat{b}=1.29, \widehat{c}=0.48, \widehat{d}=0.50$	0.071(0.945)	0.542(0.713)	0.065(0.797)
$4pWW$	$\widehat{a}=13.90, \widehat{b}=1.32, \widehat{c}=0.36, \widehat{d}=0.28$	0.070(0.946)	0.469(0.778)	0.062(0.802)
$4pWG$	$\widehat{a}=14.64, \widehat{b}=1.12, \widehat{c}=1.77, \widehat{d}=0.99$	0.074(0.928)	0.573(0.678)	0.066(0.776)
$4pWN$	$\widehat{a}=15.50, \widehat{b}=1.17, \widehat{c}=1.03, \widehat{d}=0.21$	0.075(0.921)	0.836(0.459)	0.070(0.759)

Table 4. Goodness-of-fit tests. Example 3

$3pW$	$\widehat{a}=993.19, \widehat{b}=1.18, \widehat{\tau}=4.37$	0.08(0.928)	0.435(0.812)	0.048(0.891)
$4pWU$	$\widehat{a}=984.62, \widehat{b}=1.18, \widehat{c}=4.85, \widehat{d}=4.5$	0.077(0.949)	0.431(0.834)	0.047(0.908)
$4pWW$	$\widehat{a}=993.61, \widehat{b}=1.18, \widehat{c}=0.96, \widehat{d}=1.5$	0.077(0.943)	0.392(0.866)	0.046(0.901)
$4pWG$	$\widehat{a}=957.54, \widehat{b}=1.18, \widehat{c}=3.55, \widehat{d}=0.4$	0.088(0.863)	0.403(0.847)	0.047(0.892)
$4pWN$	$\widehat{a}=991.98, \widehat{b}=1.17, \widehat{c}=4.94, \widehat{d}=0.6$	0.077(0.941)	0.428(0.817)	0.046(0.904)

5. Conclusions

In the paper, the failure-free time, being the location parameter of the shifted Weibull distribution, was proposed to be treated as a random variable. We defined four compound Weibull distributions with the location parameter having Uniform, Weibull, Gamma and Normal distribution, respectively. Using these proposed models the analysis of three real lifetime data sets were performed. The received results showed that the new models fit better the data under consideration that the standard three-parameter Weibull distribution.

However, anyone who will decide to use any of the proposed compound Weibull distributions in data analysis has to be equipped with a powerful computational environment. Luckily, nowadays it is not a problem since we have Excel, Mathcad, Mathematica, Matlab, Scilab and maybe some other not so widely known computational tools.

Acknowledgements

The authors thank Prof. Antoni Drapella, former Director of Institute of Mathematics at Pomeranian University in Słupsk, for providing assistance in formulating basics of the method presented. The second author was partially supported by PUT under grant 0211/SBAD/0121.

Conflict of interest. The authors declare that they have no conflict of interest.

References

Ahrens, J. H., Dieter, U. (1982). Generating gamma variates by a modified rejection technique. *Communications of the ACM*, 25, pp. 47–54.

Ahrens, J. H., Dieter, U. (1974). Computer methods for sampling from gamma, beta, poisson and binomial distributions. *Computing*, 12, pp. 223–246.

Alamlki, S. J.,Nadarajah, S., (2014). ,Modifications of the Weibull distribution: a review.*Reliability Engineering and System Safety*, 124, pp. 32–55.

Balakrishnan, N., Ristic, M. M., (2016). Multivariate families of gamma-generated distributions with finite or infinite support above or below the diagonal, *Journal of Multivariate Analysis*, 143, pp. 194–207.

Drapella, A., (1993). Complementary Weibull distribution: Unknown or just forgotten. *Quality and Reliability Engineering International*, 9, pp. 383–385.

Drapella, A., (1999). An improved failure-free time estimation method. *Quality and Reliability Engineering International*, 15, pp. 235–238.

Dubey, S. D. (1968). A compound Weibull distribution. *Naval Research Logistics Quarterly*, 15, pp. 179–188.

Gertsbakh, I. B., Kordonskiy, K. H. B., (1969). *Models of failure*, Verlag: Springer.

Kao, J. H. K. (1966). *Lifetime models with applications*, In: Reliability Handbook. W.G. Ireson Editor-in-chief. McGraw-Hill Company.

Kao, J. H. K. (1960). A summary of some new techniques on failure analysis, *Proc. Sixth Natl. Symp. on Reliability and Quality Control*.

Kececioglu, D. (1991). *Reliability Engineering Handbook*, New York: Prentice Hall, Eaglewood Cliffs.

Kendall, M. G., Stuart, A. (1961). *The advanced theory of statistics, Vol. 2*, Charles Griffin and Company.

Lai, C. D. (2014). *Generalized Weibull Distributions*, New York: Springer.

Lai, C. D., Xie, M. (2006). *Stochastic ageing and dependence for reliability*, Springer Science and Business Media.

Lam, S. W., Halim, T., Muthusamy, K. (2010). Models with failure-free life—Applied review and extensions, *IEEE Transactions on Device and Materials Reliability*, 10(2), pp. 263–270.

Mahmood, S. W., Algamal, Z. Y. (2021). Reliability Estimation of Three Parameters Gamma Distribution via Particle Swarm Optimization, *Thailand Statisticia*, 19(2), pp. 308–316.

Mudholkar, G. S., Srivastava, D. K., (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data *IEEE Transactions on Reliability*, 42, pp. 299–302.

Murthy, D. N. P., Xie, M., Jiang, R., (2004). *Weibull models*, Hoboken: Wiley.

O'Connor, P. D. T. (1985). *Practical reliability engineering*, New York: Wiley.

Park, C. (2018). A Note on the Existence of the Location Parameter Estimate of the Three-Parameter Weibull Model Using the Weibull Plot, *Mathematical Problems in Engineering*, 10(2), pp. 1–6.

Qutb, N., Rajhi, E., (2016). Estimation of the Parameters of Compound Weibull Distribution, *IOSR Journal of Mathematics*, 12, pp. 11–18.

Ramakrishnan, M., Viswanathan, N., (2017). Comparing the methods of estimation of three-parameter Weibull distribution, *IOSR Journal of Mathematics*, 13(1), pp. 42–45.

Rossberg, H. J., Jesiak, B., Siegel, G., (1985). *Analytic Methods of Probability Theory*, Berlin: Academie Verlag.

Saffawy, S. Y., Algmal, Z. Y (2006). The Use of Maximum Likelihood and Kaplan-Meir method to Estimate the Reliability Function An Application on Babylon Tires Factory, *Tanmiyat Al-Rafidain*, 82(28), pp. 9–20.

Seidel, W. (2010). *Mixture model*, In Lovric, M., International Encyclopedia of Statistical Science, Heidelberg: Springer.

Stacy, E.W., (1962). A generalization of the gamma distribution, *The Annals of Mathematical Statistics*, pp. 1187–1192.

Szymkowiak, M. (2018a). Characterizations of distributions through aging intensity, *IEEE Transactions on Reliability*, 67(2), pp. 446–296.

Szymkowiak, M. (2018b). Generalized aging intensity functions, *Reliability Engineering and System Safety*, 178, pp. 198–208.

Szymkowiak, M., (2020). *Lifetime analysis by aging intensity functions*, Cham: Springer.

Weber, M. D., Leemis, L. M., Kincaid, R. (2006). Minimum Kolmogorov–Smirnov test statistic parameter estimates, *Journal of Statistical Computation and Simulation*, 76(3), pp. 195–206.

Weibull, W. (1951). A statistical distribution function of wide applicability, *Journal of Applied Mechanics*, 18, pp. 29–296.

Wichura, M. J., (1988). Algorithm AS 241: The percentage points of the normal distribution, *Applied Statistics*, 37, pp. 477–484.

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 77–90, DOI 10.2478/stattrans-2022-0043
Received – 09.09.2021; accepted – 01.08.2022

Robustness of randomisation tests as alternative analysis methods for repeated measures design

Abimibola Victoria Oladugba[1], **Ajali John Obasi**[2], **Oluchukwu Chukwuemeka Asogwa**[3]

ABSTRACT

Randomisation tests (*R*-tests) are regularly proposed as an alternative method of hypothesis testing when assumptions of classical statistical methods are violated in data analysis. In this paper, the robustness in terms of the type-I-error and the power of the *R*-test were evaluated and compared with that of the *F*-test in the analysis of a single factor repeated measures design. The study took into account normal and non-normal data (skewed: exponential, lognormal, Chi-squared, and Weibull distributions), the presence and lack of outliers, and a situation in which the sphericity assumption was met or not under varied sample sizes and number of treatments. The Monte Carlo approach was used in the simulation study. The results showed that when the data were normal, the *R*-test was approximately as sensitive and robust as the *F*-test, while being more sensitive than the *F*-test when data had skewed distributions. The *R*-test was more sensitive and robust than the *F*-test in the presence of an outlier. When the sphericity assumption was met, both the *R*-test and the *F*-test were approximately equally sensitive, whereas the *R*-test was more sensitive and robust than the *F*-test when the sphericity assumption was not met.

Key words: randomisation test, repeated measures design, sensitivity, robustness, Monte Carlo.

1. Introduction

Research in many areas of application as affirmed by Ma et al. (2012) normally involves study plans in which measurements or responses are repeatedly obtained from an experimental unit (EU). According to Davis (2002), repeated measurements refer broadly to data in which the response of each experimental unit or subject is observed on multiple treatment conditions or time points. Repeated measures design (RMD)

[1] Corresponding Author. Department of Statistics, University of Nigeria, Nsukka, Nigeria. E-mail: abimibola.oladugba@unn.edu.ng. ORCID: https://orcid.org/0000-0002-6402-8833.

[2] Department of Statistics, University of Nigeria, Nigeria. ORCID: https://orcid.org/0000-0002-4761-9682.

[3] Department of Mathematics/Computer Science/Statistics and Informatics, Alex Ekwueme Federal University Ndufu Alike Ikwo, Nigeria. ORCID: https://orcid.org/0000-0001-7297-9201.

is an experimental design that involves multiple measures of the same variable(s) taken on the same EU either under different treatment conditions or over two or more time periods (Kreuger and Tian, 2004).The major advantage of RMD is that it uses exactly the same individuals or subjects in all treatment conditions thereby eliminating the influence of individual differences from the analysis and also being economical in the use of resources and enabling the subjects to be their own control as measurements are taken under both control and other experimental conditions (Reed III, 2003; Howitt and Cramer, 2011).

An approach to RMD data analysis is the repeated measures analysis of variance (RM ANOVA) that is based on the *F*-test statistic which has assumptions that must be met to ensure valid results are obtained from the analysis and therefore is limited in its application (Dragset, 2009). The assumptions include random sampling of EU from the population, normality of responses, and equality of all pairwise differences in variance between experimental conditions called sphericity (Girden, 1992, Lindman, 1992). The *F*-test is a statistical test in which the sampling distribution of the test statistic has an *F*-distribution when the null hypothesis is true (Oladugba et al., 2014). In statistical analysis, if the assumptions for any parametric test cannot be satisfied, there is risk of passing invalid inference if such test is deployed. So, researchers either transform the response data so that the resulting variable meets the conditions of the intended test to be used or resort to a different test such as the non-parametric test, which is not affected by the assumptions of the parametric test (Zimmerman and Zumbo, 1990) but transformation of data according to Sawilowsky et al. (1989) can have poor power properties. Also, the use of ranks in nonparametric tests leads to loss of information, thus the researchers cannot rely with high confidence level on ranking or transformation of data as an alternative to the *F*-test when its assumptions are not met (Gleason, 2013).

Randomization test (R-test) or permutation test can provide excellent solutions in the presence of unsuitable conditions for the use of the *F*-test or when the researchers want to maintain the use of the original data. The *R*-test is a way of hypothesis testing that can be deployed for analysis of experimental data when assumptions of parametric tests are not tenable (Edgington, 1995; Kherad-Pajouh and Renaudi, 2014). It provides an efficient approach to hypothesis testing. In other words, the *R*-test is perceived as an alternative method to data analysis in conditions when assumptions of parametric procedures are not met (Craig and Fisher, 2019; Berry et al., 2018). *R*-test performs well in conditions not favourably for the *F*-test and is as sensitive and robust as the *F*-test when parametric test assumptions are met (Mundry, 1999; Mewhort, 2005; Mewhort et al., 2010).

Since the validity of any statistical inference depends largely on satisfaction of the assumptions of the underlying model, researchers should not anticipate any statistical

test to be the most appropriate in any situation but rather subject proposed statistical test to scrutiny to ensure it is better than other alternatives in terms of sensitivity and robustness (Peres-Neto and Olden, 2001). The sensitivity of a test is the ability of a test to make right decision vis-à-vis rejection or acceptance of a hypothesis also known as power of a test; it is greatly influenced by sample size and presence of outliers (Cohen, 1988) and assumption of sphericity (constant variance) for RMD (Dragset, 2009) while robustness, on the other hand, refers to the ability of a test to yield correct conclusion or perform optimally in terms of controlling the type-I-error (p) that is not to falsely detect an effect when some of the distributional assumptions are not met or under unfavourable conditions (Vorapongsathorn et al., 2004).

Hence, this paper used the *R*-test to analyse the RMD and compared the results to that of the *F*-test in order to find out which was more sensitive and robust under the conditions that data are normal and non-normal (exponential, lognormal, Chi-square, and Weibull distributions), in the absence and presence of outliers, when sphericity assumption was met or not in variant number of treatments and sample sizes.

2. Materials and methods

2.1. Material

The data presented in Table 1 were obtained from Gravetter and Wallnau (2007). The responses generated from the study were based on the time (in seconds) lapsed until participants reported they felt nothing called latency when a stimulus (of 500-milligram weight) was gently placed on a region of the body. The study compared the adaptation for four regions of the body for a sample of 7 participants.

Table 1. Data on sensory adaption experiment

	Area of stimulation (Treatment)			
Subjects	**Back of hand**	**Lower back**	**Middle of Palm**	**Chin below lower Lip**
1	6.5	4.6	10.2	12.1
2	5.8	3.5	9.7	11.8
3	6.0	4.2	9.9	11.5
4	6.7	4.7	8.1	10.7
5	5.2	3.6	7.9	9.9
6	4.3	3.5	9.0	11.3
7	7.4	4.8	10.8	12.6

2.2. The *F*-test method for analysis of single factor RMD

The *F*-test procedure for hypothesis testing in analysis of RMD involves computing the *F*-statistic associated with the problem. In this section, the model, ANOVA table

presented in Table 2 and the *F*-test procedures for analysing single factor RMD are defined as follows.

The model for this design is defined as:

$$y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij} \quad i = 1, 2, \ldots, n; j = 1, 2, \ldots, t$$

$$\sum_{j=1}^{t} \tau_j = 0; \beta_i \sim N(0, \sigma_\beta^2); \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where, y_{ij} is the response from the i^{th} subject at treatment j; μ is the grand mean; τ_j is the fixed effect of the j^{th} treatment (the treatments are assumed to have fixed effects thus the zero-sum constraint); β_i is the random effect for i^{th} subject and ε_{ij} is a random error component specific to i^{th} subject at j^{th} treatment.

Table 2. ANOVA table for single factor RMD

Source of Variation	SS	df	Mean Square	F_0
Subject	SS_B	n -1	MS_S	
Treatments	SS_T	t – 1	MS_T	$\frac{MS_T}{MS_E}$
Error	SS_E	$(t - 1)(n - 1)$	MS_E	
Total	SS_T	tn – 1		

where $MS_S = \frac{SS_B}{n-1}$; $MS_T = \frac{SS_T}{t-1}$; $MS_E = \frac{SS_E}{(t-1)(n-1)}$. The sums of squares are then defined as follows:

$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{t}(\bar{y}_{.j} - \bar{y}_{..})^2$; $SS_S = \sum_{i=1}^{n}\sum_{j=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2$; $SS_E = \sum_{i=1}^{n}\sum_{j=1}^{t}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} - \bar{y}_{..})^2$

2.3. Randomization test procedure

The hypothesis to be tested is:

H_o: the different treatments had the same effect *vs* H_1: there is a differential effect of at least one treatment

$\alpha = 0.05$

Test statistic

Here, *F*-statistic was used as the test statistic. It summarizes the differences between means and eliminates the effects of between-subject variability.

Procedure

With repeated measures, we permute the data within subject. If there is no effect of treatments, then the set of scores from any subject can be exchanged across treatments.

The steps are as follows:

- Compute the *F*-statistic for the original data, and denote that as F_{cal}.
- Permute the data within each subject, and do it for every subject.

- Calculate an F-statistic for each of the permuted data.
- If this F-statistic is greater than F_{cal}, increment the counter.
- Repeat the preceding three steps B times, where $B \geq 10{,}000$.
- Divide the value in the counter by B to obtain the probability of obtaining an F-statistic as large as F_{cal} if the null hypothesis were true. Denote this value as empirical type-I-error (p-value).
- Reject the null hypothesis of no difference due to treatment if p-value is less than our chosen level of significance.

2.4. Randomization test procedure for RMD

The R-test for analysing single factor RMD involves the following procedures. Compute a test statistic that sufficiently explains the experimental data (the F-statistic in this case) for the data in Table 1. Afterwards, the data are rearranged within the subject repeatedly and the test statistic is recomputed for all resultant data permutations. Randomization test uses the obtained results from all data permutations and the original result of the experiment to form a reference set which is used to decide the significance of the test. The fraction of the data permutation in the reference set having test statistic values greater than or equal to the value obtained from the original results before data were permuted is the type-I-error (significance or probability value).

In permuting data in RMD, Edgington (1995) proposed two schemes, namely systematic and random permutation schemes. In this paper, the random permutation scheme was adopted and carried out in the following way. Firstly, the data are arranged in a table with k columns and n rows, where k is the number of treatments and n is the number of subjects. An index number 1 to n was assigned to the subjects and 1 to k to the treatments, so that each measurement has associated with it a compound index number, the first part which indicates the subject and the second indicates the treatments. Accordingly, index (2, 3) for instance referred to the measurement for the second person under the third treatment. Then a random number generation algorithm was used to randomly determine for each subject independently of the other subjects which of the k measurements is to be assigned to the first treatment, which of the remaining k-1 measurements to the second treatment, and so on. The random determination of order of measurements within each subject performed over all subjects constitutes a single permutation or arrangement of the data. The arrangement is repeated for a large number of times like 10,000 permutations, and for each permutation, the test statistic is computed. The p-value is computed as the number of the test statistic value, including the obtained test statistics values that are as large as the obtained test statistics value.

2.5. Outlier detection and sphericity assumption

Outliers were randomly injected into the dataset in Table 1, and Tukey's method of outlier detection as explained by Songwon (2006) was used in detecting them. One of the ways to test for sphericity in RMD is the use of Mauchly's test. Mauchly's test tests the hypothesis that the variances of the differences between any two conditions are equal. Thus, if the significance level of Mauchly's test is less than or equal to the alpha level, sphericity is violated. Mauchly's test of sphericity in SPSS version 22 was used to verify this condition.

2.6. Monte Carlo Simulation

In order to analyse RMD with the *R*-test and the *F*-test so as to check their robustness, a Monte Carlo simulation was conducted using RMD in Table 1 with $n = 7$ subjects and $t = 4$ treatments. Three variables were manipulated: (i) sample sizes (n); (ii) number of treatments (t); and (iii) distribution structure of the data (normal, exponential, lognormal, Chi-square and Weibull distributions). The performance of the two tests was investigated with three sample conditions $n = 5$, 7, and 9, and three treatment conditions $t = 3$, 4, and 5, under 5 distributional structures of the data in the presence and absence of outliers and when sphericity assumption is met or not, respectively.

The *R statistical package* was used to implement the Monte Carlo technique sampling of 10,000 permutations from the possible $(t!)^n$ permutations for the *R*-test. In the simulation, the experiment was repeated 1000 times for each distribution. In each repetition, the resulting tables of data set were analysed appropriately using the *F*-test and the *R*-test methods to obtain the rate of type-I-error and power. The percentage of significant tests out of 10,000 iterations was considered as the rejection rate.

The comparison procedures were considered in two scenarios. Firstly, in the scenario that the null hypothesis (H_0: $\mu_i = 0$) is true, the rejection rate of the null hypothesis was regarded as the type-I-error rate for each test. The test that had the closest type-I-error to the nominal $\alpha = 0.05$ was considered as the more robust of the two. Secondly, in the scenario that the alternative hypothesis (Ha: $\mu_i \neq 0$) was true, the rejection rate of the null hypothesis was considered as the power for each test. The test that had larger power was taken to be more sensitive than the other.

2.7. Distribution structure of data

Data were simulated from five theoretic distributions. The normal distribution was used to test condition under which normality assumption holds. The skewed distributions used include Chi-square, exponential, lognormal, and Weibull distributions; this represents condition under which the distribution assumption (normality) does not hold. The probability density function of the five distributions is defined as follows.

(a) Normal distribution

The normal distribution has probability density function (pdf) as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < \mu < \infty, \ \sigma > 0, \ x > 0$$

The parameters (μ and σ^2) of the normal distribution were estimated using the maximum likelihood estimators (MLE). For the normal distribution, data were simulated using mean, $\bar{x}$ = 7.7250 and variance, σ^2= 9.1180, of the experimental data.

(b) Exponential distribution

The exponential distribution has pdf with parameter θ is given by

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \quad , \ \theta > 0, x \geq 0$$

Data were simulated to follow the exponential distribution using the MLE of the exponential distribution parameters obtained as $\hat{\theta}$= 7.7250 as fitted using *fitdistrplus* package in *R* statistical computing.

(c) Chi-square distribution

The pdf of Chi-square distribution with parameter *n*, is given as

$$f(x) = \frac{x^{\frac{n}{2}-1} e^{-x/2}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})}, \ x > 0$$

Using *fitdistrplus* package in *R* statistical computing, the parameter of the Chi-square distribution,
n = 4.559 ~ 5, was used for simulation of data where n is the mean of Chi-square distribution.

(d) Lognormal distribution

The pdf for the two-parameter (μ and σ^2) lognormal distribution is

$$f(\mathrm{X}|\mu,\sigma^2) = \frac{1}{X\sqrt{(2\pi\sigma^2)}} e^{-\left[\frac{(\ln(X)-\mu)^2}{2\sigma^2}\right]}, \qquad X > 0, \ -\infty < \mu < \infty, \ \sigma > 0$$

The MLE of μ and σ^2 were obtained as:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \boldsymbol{ln}\,(\boldsymbol{X_i})}{n} = 7.7880 \text{ and } \hat{\sigma}^2 = \frac{\left(\sum_{i=1}^{n} (\boldsymbol{ln}\,(\boldsymbol{X_i}) - \frac{\sum_{i=1}^{n} \boldsymbol{ln}\,(\boldsymbol{X_i})}{n}\right)^2}{n} = 12.0120$$

(e) Weibull distribution

The two-parameter Weibull distribution has pdf given as

$$f(x/k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{(x/\lambda)^k}, \quad x \geq 0, \ k > 0, \lambda > 0$$

Data were simulated using the MLE of the Weibull distribution parameters obtained as $k = 7.020$ and $\lambda = 9.10$ as fitted using *fitdistrplus* package in *R* statistical computing.

3. Results

In Tables 3, 4, 5 and 6, the simulation results (type-I-error and power) for the *F*-test and the *R*-test based on the three manipulated variables (sample size, number of treatments and distribution structure of the data) are presented. Following from the methods mentioned in Section 2 as implemented in *R Statistical package*, for each sample size, the optimal values of the type-I-error and power were recorded. The sample size was denoted as n, the values in bracket indicate the number of treatment (t) that produced optimal type-I-error and highest power as the number of treatments were varied. The values in bold are either the optimal type-I-error or the highest power for each of the test.

Table 3 shows the type-I-error of the *F*-test and the *R*-test for the data in the absence of outliers. The results indicated that as n increased, type-I-error decreased for data with normal distribution, for Chi-square, lognormal, exponential and Weibull, it initially increased but afterwards decreased for the *F*-test while the *R*-test produced type-I-error that increased as n increased under the normal distribution but reduced as n increased for Chi-square, exponential and Weibull, while for data with lognormal distribution, the type-I-error decreased as n decreased. On the other hand, the power values for the normal data decreased initially but later increased as the sample size increased, it increased initially and subsequently decreased for Chi-square and Weibull distributions for the *F*-test and increased for exponential and lognormal data distributions. The *R*-test on the other hand had increasing power as n increased for exponential, Weibull, and Chi-square but had an increasing trend for lognormal although with a slight initial decrease at $n = 7$. When outliers were introduced, the type-I-error and power values are presented in Table 4. The results indicated that type-I-error for the *F*-test under all the data distributions had a decreasing trend as n increased but an increasing trend for Weibull distribution. Furthermore, the power for the *F*-test exhibited a slight decreasing trend for Chi-square and lognormal, while it increased for normal, exponential and Weibull as n increased. On the other hand, the power values of the *R*-test for all data distribution were increasing as sample size increased.

The results for when sphericity assumption was met are displayed in Table 5. The type-I-error for the *F*-test in this table revealed that as n increased, normal and exponential data distributions initially had a slight increasing trend but substantially increased afterwards for Weibull data distribution while a decreasing trend was observed for Chi-square and lognormal data distribution. The results of sphericity

assumption not met as displayed in Table 6 showed that type-I-error and power for both tests decreased for all distributions as the sample size increased.

Table 3. Simulation results (type-I-error and power) for *F*-test and *R*-test in the absence of outliers

		type-I-error		Power	
Distribution	***n (t)***	***F*-test**	***R*-Test**	***F*-test**	***R*-Test**
Normal	5(5)	0.**0500**	0.**0429**	0.9437	0.9226
	7(4)	0.0428	0.0546	0.9122	0.9179
	9(5)	0.0408	0.0595	0.**9914**	0.**9805**
Exponential	5(5)	0.0725	0.**0501**	0.7093	0.9032
	7(4)	0.1442	0.0613	0.7528	0.9211
	9(5)	0.**0611**	0.0581	0.**7828**	0.**9469**
Lognormal	5(5)	0.**0413**	0.0593	0.7229	0.8534
	7(5)	0.0662	0.0439	0.7237	0.8629
	9(5)	0.0599	0.**0490**	0.**8009**	0.**8979**
Chi-square	5(4)	0.0705	0.**0524**	0.6184	0.7528
	7(4)	0.1009	0.0687	0.**6729**	0.7367
	9(5)	0.**0704**	0.0591	0.5646	0.**8086**
Weibull	5(5)	0.0849	0.0640	0.5256	0.7439
	7(5)	0.1225	0.0580	0.**6804**	0.7811
	9(5)	0.0783	0.**0441**	0.5959	0.**8724**

Table 4. Simulation results (type-I-error and power) for *F*-test and *R*-test in the presence of outliers

		type-I-error		Power	
Distribution	***n (t)***	***F*-test**	***R*-Test**	***F*-test**	***R*-Test**
Normal	5(4)	0.1029	0.0699	0.5790	0.7498
	7(4)	0.**0824**	0.0601	0.5617	0.**8209**
	9(5)	0.0873	0.**0588**	0.**5869**	0.7998
Exponential	5(5)	0.2018	0.**0566**	0.5958	0.7909
	7(5)	0.**0755**	0.1003	0.**6963**	0.7304
	9(5)	0.1046	0.0708	0.5540	0.**8202**
Lognormal	5(5)	0.0815	0.0597	0.**6876**	0.7588
	7(5)	0.1174	0.0632	0.6011	0.6901
	9(5)	0.**0792**	0.**0512**	0.5906	0.**8094**
Chi-square	5(4)	0.2171	0.0696	0.5377	0.8132
	7(4)	0.1024	0.0741	0.5213	0.7995
	9(5)	0.**0843**	0.**0516**	0.**6628**	0.**8180**
Weibull	5(5)	0.**0818**	0.**0536**	0.5448	0.7800
	7(5)	0.1032	0.0684	0.**5994**	0.7468
	9(5)	0.0929	0.0684	0.5834	0.**7933**

Table 5. Simulation results (type-I-error and power) for *F*-test and *R*-test with sphericity assumption met

		type-I-error		Power	
Distribution	***n (t)***	***F*-test**	***R*-Test**	***F*-test**	***R*-Test**
Normal	5(4)	0.**0506**	0.0420	0.9637	0.9024
	7(5)	0.0431	0.0446	0.9202	0.9231
	9(5)	0.0521	0.**0511**	0.**9884**	0.**9531**
Exponential	5(4)	0.1011	0.0429	0.8663	0.9001
	7(5)	0.1502	0.0559	0.8818	0.8965
	9(5)	0.**0841**	0.**0523**	0.**9212**	0.**9045**
Lognormal	5(4)	0.0706	0.0462	0.8291	0.8088
	7(5)	0.0762	0.**0518**	0.8119	0.8321
	9(5)	0.**0699**	0.0442	0.**8921**	0.**8899**
Chi-square	5(5)	0.**0589**	0.**0493**	0.6610	0.8011
	7(5)	0.1209	0.0621	0.6690	0.7822
	9(5)	0.1022	0.0489	0.**6710**	0.**8399**
Weibull	5(4)	0.**0820**	0.0531	0.5006	0.7877
	7(4)	0.1015	0.**0429**	0.**5094**	0.8807
	9(5)	0.1183	0.0401	0.5009	0.**8991**

Table 6. Simulation results (type-I-error and power) for *F*-test and *R*-test with sphericity assumption not met

		type-I-error		Power	
Distribution	***n (t)***	***F*-test**	***R*-Test**	***F*-test**	***R*-Test**
Normal	5(5)	0.1112	0.0612	0.**6821**	0.8080
	7(5)	0.1230	0.0610	0.5417	0.8526
	9(5)	0.**0811**	0.**0588**	0.6809	0.**8595**
Exponential	5(4)	0.2074	0.0640	0.5958	0.8522
	7(5)	0.**0603**	0.0595	0.4993	0.8032
	9(5)	0.1032	0.**0467**	0.**6240**	0.**8704**
Lognormal	5(4)	0.1401	0.0531	0.4876	0.7863
	7(5)	0.0631	0.0699	0.4211	0.7902
	9(5)	0.**0503**	0.**0518**	0.**5906**	0.**8186**
Chi-square	5(5)	0.0813	0.4040	0.5307	0.7863
	7(5)	0.1109	0.0601	0.5213	0.**8039**
	9(5)	0.**0705**	0.**0517**	0.**6028**	0.7995
Weibull	5(4)	0.1207	0.0485	0.5448	0.7904
	7(4)	0.**0779**	0.0590	0.**5994**	0.**8002**
	9(5)	0.1052	0.**0508**	0.5891	0.7808

4. Discussion of results

The *R*-test and the *F*-test were used to analyse RMD with and without outlier and sphericity respectively. From the results, under the normal assumption, the type-I-error of both tests was within limits regarded as being robust with the *F*-test producing a better value at $n = 5$ ($p = 0.05$) while the power of both the *F*-test and *R*-test was very high (0.9914 and 0.9805 respectively) and it increased as the sample size and the number of treatments increased. This implies that both tests were approximately equally sensitive and robust under normal assumption. For the exponentially distributed data, as the sample size increased, the optimal type-I-error for the *F*-test was at $n = 9$, $t = 5$ ($p = 0.0611$) and $n = 5$, $t = 5$ for the *R*-test ($p = 0.051$), whereas the highest power for the *F*-test and the *R*-test was 0.7828 and 0.9469 respectively at $n = 9$, $t = 5$, which shows that the *R*-test was more powerful that the *F*-test and more robust too for exponential data. For lognormal distribution, the optimal type-I-error for both tests as the sample size increased was 0.0413 and 0.0490 for the *F*-test and the *R*-test respectively, while both tests exhibited power of 0.8009 and 0.8979 at $n = 9$ respectively for the *F*-test and the *R*-test. For the Chi-square distribution, the *F*-test had optimal type-I-error of 0.0704 at $n = 9$, $t = 5$ and 0.0524 for *R*-test at $n = 9$, $t = 5$. Also, the highest power of the *F*-test and the *R*-test was 0.6729 ($n = 7$) and 0.8076 ($n = 9$) respectively. For the Weibull distribution, the *R*-test was more robust with $p = 0.0441$ and more powerful with power = 0.8724.

When outliers are present, the *R*-test was more powerful and robust in all distributions: normal assumption ($p = 0.0588$, power = 0.8209), exponential distribution ($p = 0.0566$, power = 0.8202), lognormal ($p = 0.0512$, power=0.8094), Chi-square ($p = 0.0516$, power =0.8180), Weibull ($p = 0.0536$, power = 0.7933).

When sphericity condition was met, the *F*-test was more powerful and robust ($p = 0.0506$, power = 0.9531) for data with normal distribution while the *R*-test was more powerful and robust for lognormal data ($p = 0.0518$, power = 0.8899), Chi-square ($p = 0.0493$, power = 0.8399), Weibull distribution ($p = 0.0429$, power = 0.8991). For exponential data, the *F*-test was more robust for data ($p = 0.0523$) while the *R*-test was more powerful (0.9212). Furthermore, when sphericity assumption was not met, the *F*-test was only more robust for lognormal ($p = 0.0503$) while the *R*-test was more powerful (power = 0.8186). Meanwhile, the *R*-test was more robust and powerful for other distributions – normal ($p = 0.0588$, power = 0.8595), exponential ($p = 0.0467$, power = 0.8704), Chi-square ($p = 0.0517$, power = 0.8039), and Weibull ($p = 0.0508$, power = 0.8002).

5. Conclusion

In this paper, the *R*-test was used in analysing RMD with or without outlier and sphericity respectively. The test offers the freedom of choice of test statistic that sufficiently suits a particular statistical problem for researchers and is free from any distributional or test assumptions, but rather depends only on the randomization technique – thus the name the randomization test. The study also employed the classical test (*F*-test) for analysing RMD, which is hinged on a number of conditions for reliable valid inference. This paper compared both tests to ascertain which controlled the type-I-error better and had higher power than the other. These criteria of comparison were referred to as robustness and sensitivity respectively.

The results in Tables 3, 4, 5 and 6 showed that under the normal distribution when sphericity held, both tests were equally robust and approximately powerful with optimal values at $n = 5$, $t = 5$ ($p = 0.05$ power = 0.9914) for the *F*-test and at $n = 9$, $t = 5$ ($p = 0.0421$, power = 0.9805) for the *R*-test. When data had skewed distributions (exponential, Chi-square, lognormal and Weibull), the *R*-test was more robust and powerful. In the presence of an outlier and when sphericity condition was not met, the *F*-test was less robust and sensitive than the *R*-test. In the analysis of RMD when normality and sphericity conditions were met, the *R*-test was comparably as robust and sensitive as the *F*-test. When data had skewed distributions (exponential, lognormal, Chi-square and Weibull), the *F*-test was less robust and sensitive as the sample size and the number of treatments increased. Also, in the presence of an outlier and when sphericity condition was met or not, the *R*-test was more robust and sensitive than the *F*-test. In a nutshell, the *R*-test was approximately as sensitive as the *F*-test in RMD when data follow normal and sphericity conditions met but more sensitive when data were skewed (exponential, Chi-square, lognormal and Weibull).

In general, since the *R*-test is always as robust and sensitive and even more robust and sensitive than the *F*-test, to alleviate the burden of assessing parametric assumptions which is done before the use of the *F*-test, researchers are advised to go ahead with *R*-test which is not based on any assumption and is easily carried out with modern-day high-capacity computers.

References

Berry, K., Johnston, J., Mielke, P., (2018). *Permutation Statistical Methods: A Permutation Statistical Approach*, doi: 10.1007/978-3-319-98926-6_2.

Cohen, J., (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). New Jersey: *Lawrence Earlbaum Associates*.

Craig, A. R., Fisher, W. W., (2019). Randomization tests as alternative analysis methods for behavior-analytic data. *Journal of the Experimental Analysis of Behavior*, 111(2), pp. 309–328.

Davis, C. S., (2002). Statistical Methods for the Analysis of Repeated Measurements. New York, NY: *Springer Publishers*.

Dragset, I. G., (2009). Analysis of longitudinal data with missing values: Methods and Applications in Medical Statistics (Master's Thesis). Available from *Norwegian university of science and technology digital theses database*.

Edgington, E. S., (1995). Randomization Tests (3rd Ed). New York, NY: *Marcel Dekker*.

Girden, E. R., (1992). ANOVA: Repeated measures. *Sage Publications*, Newbury Park, CA.

Gleason, J., (2013). Comparative power of the ANOVA, randomization ANOVA, And Kruskal-Wallis test (Doctoral Dissertation). Available from *Wayne State University Digital Dissertations database*.

Gravetter, F. J., Wallnau, L. B., (2007). Statistics for the behavioral science. Canada: *Vicki Knight*.

Howitt, D., Cramer, D., (2011). Introduction to Research Methods in Psychology (3rd ed.). *Essex: Pearson Education Limited*.

Kherad-Pajouh, S., Renaud, O., (2014). A general permutation approach for analyzing repeated measures ANOVA and mixed-model designs. *Computational Statistics and Data Analysis*, 21 (5), pp. 42–59.

Krueger, C., Tian, L., (2004). A Comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological Research for Nursing*, 6, pp. 151–157.

Lindman, H. R., (1992). Analysis of Variance in Experimental Design. New York: *Springer-Verlag*.

Ma, Y., Mazumdar M., Memtsoudis, S. G., (2012). Beyond repeated-measures analysis of variance: advanced statistical methods for the analysis of longitudinal data in anesthesia research. *Regional Anesthesia and Pain Medicine*, 37, pp. 99–105.

Mewhort, D. J. K., (2005). A comparison of the randomization test with the F test when error is skewed. *Behavior Research Methods*, 37 (3), pp. 426–435.

Mewhort, D. J. K., Johns, B. T., Kelly, M., (2010). Applying the permutation test to factorial designs. *Behavior Research Methods*, 42 (2), pp. 366–372.

Mundry, R., (1999). Testing related samples with missing values: a permutation approach. *Animal Behaviour*, 58, pp. 1143–1153.

Oladugba, A. V., Udom, A. U., Ugah, T. E., Ukaegbu, E. C, Madukaife, M. S., Sanni, S. S., (2014). Principles of Applied Statistics. Nsukka: *University of Nigeria Press Ltd.*

Peres-Neto, P. R., Olden, J., (2001). Assessing the robustness of randomization tests: examples from behavioral Studies. *Animal Behaviour*, 61, pp. 79–86.

Reed III, J., (2003). Analysis of variance (ANOVA) models in emergency medicine. The *Journal of Emergency and Intensive Care Medicine*, 7(2), pp. 21–34.

Sawilowsky, S. S., Blair, R. C., Higgins, J. J., (1989). An investigation of the type-I-error and power properties of the rank transformation procedure in factorial ANOVA. *Journal of Educational Statistics*, 14 (3), pp. 255–267.

Songwon, S., (2006). A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets (Master's thesis) available from University of Pittsburgh, *Graduate School of Public Health database*.

Vorapongsathorn, T., Taejaroenkul, S., Viwatwongkasem, C., (2004). A comparison of type-I-error and power of Bartlett's test, Levene's test and Cochran's test under violation of assumptions. S*ongklanakarin Journal of Science and Technology*, 26(4), pp. 537–547.

Zimmerman, D. W., Zumbo, B. D., (1990). Effect of outliers on the relative power of parametric and nonparametric statistical tests. *Perceptual and Motor Skills*, 71, pp. 339–349.

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 91–111, DOI 10.2478/stattrans-2022-0044
Received – 15.01.2021; accepted – 28.02.2022

Missing data estimation based on the chaining technique in survey sampling

Narendra Singh Thakur[1], **Diwakar Shukla**[2]

ABSTRACT

Sample surveys are often affected by missing observations and non-response caused by the respondents' refusal or unwillingness to provide the requested information or due to their memory failure. In order to substitute the missing data, a procedure called imputation is applied, which uses the available data as a tool for the replacement of the missing values. Two auxiliary variables create a chain which is used to substitute the missing part of the sample. The aim of the paper is to present the application of the Chain-type factor estimator as a means of source imputation for the non-response units in an incomplete sample. The proposed strategies were found to be more efficient and bias-controllable than similar estimation procedures described in the relevant literature. These techniques could also be made nearly unbiased in relation to other selected parametric values. The findings are supported by a numerical study involving the use of a dataset, proving that the proposed techniques outperform other similar ones.

Key words: estimation, missing data, chaining, imputation, bias, mean squared error (MSE), factor type (F-T), chain type estimator, double sampling.

Mathematical Subject Code: 62D05

1. Introduction

In sample surveys, the auxiliary information is used to improve efficiency of the estimate [see, Cochran (2005), Sukhatme et al. (1984)]. The use of a ratio estimator is preferred when the population mean of auxiliary variate is known. However, when it is unknown then it is not possible to apply the ratio estimator directly and the concept of two-phase sampling is applied to get a sample-based estimate of population mean. Sometimes information on one more auxiliary variable highly correlated to earlier

[1] Govt. Adarsh Girls College, Sheopur (M.P.), India, Pin – 476337, Affiliation with Jiwaji University, Gwalior (M.P.), India. E-mail: nst_stats@yahoo.co.in. ORCID: https://orcid.org/0000-0001-9731-058X.

[2] Dr. Harisingh Gour Central University, Sagar (M.P.), India, Pin – 470003.
E-mail: diwakarshukla@rediffmail.com. ORCID: https://orcid.org/0000-0002-8694-0655.

auxiliary variate is available and easy to access at a lesser cost. This additional information could be intelligently utilized for obtaining efficient estimates. Chaining is one such technique, used by Chand (1975), Sukhatme and Chand (1977), which has a mechanism of combining wisely two auxiliary variates. Kiregyera (1980, 1984) proposed some chain type ratio and regression estimators whereas Singh et al. (1994) developed a class of chain type estimators under a double sample scheme. Al-Jararha and Ahmed (2002) discussed the class of chain type estimators for population variance using double a sampling scheme. Some other useful contributions are Kumar and Bahl (2006), Pradhan (2005), Rao and Sitter (1995), Sharma and Tailor (2010), Shukla (2002), Singh and Espejo (2007), Singh et al. (2009), Singh et al. (1993), Srivastava and Jhajj (1980), etc.

The use of auxiliary information in the estimation of population values of the study variate is a common phenomenon in sampling theory of surveys. Auxiliary information is successfully utilized either at the planning stage or at the design stage or at the information stage to arrive at improved estimator compared to those not utilizing auxiliary information. The use of ratio and product strategies in survey sampling solely depends upon the knowledge of population mean $\overline{X} = N^{-1}\sum_{i=1}^{N} X_i$ of the auxiliary character *X*. In many situations of practical importance, the population mean $\overline{X}$ is unknown before the start of a survey. In such a situation, the usual thing to do is to estimate it by the sample mean $\bar{x}_m = m^{-1}\sum_{i=1}^{m} x_i$ based on a preliminary sample of size m of which n is a sub-sample $(n < m)$. If the population mean $\overline{Z} = N^{-1}\sum_{i=1}^{N} Z_i$ of another auxiliary variate *Z*, closely related to auxiliary variate X but compared to X remotely related to study variate *Y* is known, it is advisable to estimate $\overline{X}$ by $\overline{X} = \bar{x}_m \frac{\overline{Z}}{z_m}$, where $\bar{z}_m = m^{-1}\sum_{i=1}^{m} z_i$, which would provide better estimate of $\overline{X}$ than $\bar{x}_m$ to the terms of order $o\left(n^{-1}\right)$ if $\rho_{XZ}\frac{C_X}{C_Z} > 0.5$ [see, Choudhury and Singh (2012)]. The symbol ρ_{XZ} is the coefficient of correlation between *X* and *Z* and C_X, C_Z are the coefficient of variation of X and Z respectively. Chand (1975) and Sukhatme and Chand (1977) proposed a technique of chaining of the available information on auxiliary characteristics with the main characteristic. Kiregyera (1980, 1984), Singh et al. (2006) also proposed some chain type ratio and regression estimators based on two auxiliary variables. Using prior information on parameters of auxiliary variate some useful contributions are Shukla et al. (1991), Bose (1943), Kadilar and Cingi (2003), Srivastava et al. (1990), Srivenkataramana (1980), etc.

According to Hietjan and Basu (1996), incompleteness in the form of missingness, censoring or grouping, is a troubling feature of several data sets. A key question is what one needs to assume to justify ignoring the incompleteness mechanism. Rubin (1976) addressed this question for Bayes/likelihood and frequentist inferences. Little and Rubin (1987) recognized for some time that failure to account for the stochastic nature of incompleteness can spoil inferences.

In brief, Rubin (1976) defined three key concepts: missing at random (MAR), observed at random (OAR) and Parameter Distinctness (PD). The data are MAR if the probability of the observed missingness pattern, given the observed and unobserved data, does not depend on the values of the unobserved data. The data are OAR if, for every possible value of the missing data, the probability of the observed missingness pattern, given the observed and unobserved data, does not depend on the values of the observed data. PD holds if there are no a priori ties between the parameters of the missingness model and those of the data model. For Bayesian inference this means that the parameters of the data model and missingness model are a priori independent. For direct likelihood inference it means that knowledge of one parameter's value does not place any constraints on the other parameter's value. Ignoring the missingness mechanism is justified for Bayes/likelihood inference if MAR and PD hold. The combination of MAR and OAR is called missing completely at random (MCAR). In what follows missing completely at random (MCAR) by Heitjan and Basu (1996) is used in this article. Some useful contributions available in the literature are Weeks (1999), Shukla et al. (2009), Seaman et al. (2013), Bhaskaran and Smeeth (2014), Pandey et al. (2015), Pandey et al. (2016), Doretti et al. (2018), etc. This manuscript presents the use of Chain-Type estimator as an imputation source for dealing with missing observations to estimate the population mean.

1.1. Some existing imputation strategies

A simple random sample S without replacement (SRSWOR), of size n is drawn from population $\Omega = \{1, 2, ..., N\}$ with Y_i as i^{th} unit of variable Y under study. Let $\overline{Y} = N^{-1} \sum_{i=1}^{N} Y_i$ be the mean of a finite population under estimation. The sample S of n units contains r responding units ($r < n$) forming a sub-space R and ($n - r$) non-responding with the sub-space ($n - r$) having symbol R^C in the space S. The sub-spaces R and R^C are disjoint and $R \cup R^C = S$. The variable Y is of main interest and X is auxiliary correlated with Y. For every unit $i \in R$, the value y_i is observed available. For units $i \in R^C$, the y_i values are missing and imputed values are to be derived. The i^{th} value x_i of X could be used as a source of imputation for y_i, $i \in R^C$. This is to assume for sample S, the data $x_s = \{x_i : i \in S\}$ is known and available completely. Responding units have missing data only for the study variable Y. Under this two variable set-up, some well-known imputation methods available in the literature are:

1.1.1. Ratio method of imputation

For y_i and x_i, define $y_{\bullet i}$ as

$$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \hat{b}x_i & if \quad i \in R^C \end{cases} \tag{1.1}$$

Where $\hat{b} = \dfrac{\sum_{i \in R} y_i}{\sum_{i \in R} x_i} = \dfrac{\bar{y}_r}{\bar{x}_r}$

Using the above, the imputation-based estimator is:

$$\bar{y}_S = \frac{1}{n}\sum_{i \in S} y_{\bullet i} = \frac{1}{n}\left[\sum_{i \in R} y_i + \hat{b}\sum_{i \in R^c} x_i\right] = \bar{y}_r\left(\frac{\bar{x}_n}{\bar{x}_r}\right) = \bar{y}_{RAT} \tag{1.2}$$

Where $\bar{y}_r = \dfrac{1}{r}\sum_{i \in R} y_i$, $\bar{x}_r = \dfrac{1}{r}\sum_{i \in R} x_i$ and $\bar{x}_n = \dfrac{1}{n}\sum_{i \in S} x_i$

1.1.2. Mean method of imputation

For y_i define $y_{\bullet i}$ as

$$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \bar{y}_r & if \quad i \in R^C \end{cases} \tag{1.3}$$

Using the above, the imputation-based estimator of population mean $\bar{Y}$ is:

$$\bar{y}_m = \frac{1}{r}\sum_{i \in R} y_i = \bar{y}_r \tag{1.4}$$

1.1.3. Compromised method of imputation

Singh and Horn (2000) proposed a compromised imputation procedure:

$$y_{\bullet i} = \begin{cases} (\alpha n/r)y_i + (1-\alpha)\hat{b}x_i & if \quad i \in R \\ (1-\alpha)\hat{b}x_i & if \quad i \in R^C \end{cases} \tag{1.5}$$

Where α is a suitably chosen constant, such that the resultant variance of the estimator is minimum. The imputation-based estimator, for this case, is

$$\bar{y}_{COMP} = \left[\alpha\bar{y}_r + (1-\alpha)\bar{y}_r\frac{\bar{x}_n}{\bar{x}_r}\right] \tag{1.6}$$

1.1.4. Ahmed methods of imputation

For the case where y_{ji} denotes the i^{th} available observation for the j^{th} imputation method Ahmed et al. (2006) suggested:

(A)

$$y_{1i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)}\left[n\bar{y}_r\left(\dfrac{\bar{X}}{\bar{x}_n}\right)^{\beta_1} - r\bar{y}_r\right] & if \quad i \in R^C \end{cases} \tag{1.7}$$

Under this, the point estimator is:

$$t_1 = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_n} \right)^{\beta_1} \tag{1.8}$$

(B)

$$y_{2i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\bar{y}_r \left(\dfrac{\bar{x}_n}{\bar{x}_r} \right)^{\beta_2} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \tag{1.9}$$

The point estimator is under this set-up:

$$t_2 = \bar{y}_r \left(\frac{\bar{x}_n}{\bar{x}_r} \right)^{\beta_2} \tag{1.10}$$

(C)

$$y_{3i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\bar{y}_r \left(\dfrac{\bar{X}}{\bar{x}_n} \right)^{\beta_3} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \tag{1.11}$$

The point estimator is:

$$t_3 = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_r} \right)^{\beta_3} \tag{1.12}$$

Terms β_1, β_2 and β_3 are suitably chosen constants, so as to keep the variance of the resultant estimator minimum. As special cases, when

$$\beta_3 = 1, \; t_{Ratio} = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_r} \right) \tag{1.13}$$

$$\text{and} \quad \beta_3 = -1, \; t_{\text{Pr}\,oduct} = \bar{y}_r \left(\frac{\bar{x}_r}{\bar{X}} \right) \tag{1.14}$$

The last one (1.14) is natural analogue of the ratio estimator called the product estimator used when an auxiliary variate X has negative correlation with Y.

1.1.5. Factor type methods of imputation

Shukla and Thakur (2008) suggested factor-type imputation procedures as:

(D)

$$\left(y_{FT1}\right)_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{\bar{y}_r}{(n-r)} \left[n\phi_1(k) - r \right] & if \quad i \in R^C \end{cases} \tag{1.15}$$

(E)

$$\left(y_{FT2}\right)_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{\bar{y}_r}{(n-r)} \left[n\phi_2(k) - r \right] & if \quad i \in R^C \end{cases} \tag{1.16}$$

(F)
$$(y_{FT3})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{\bar{y}_r}{(n-r)}\left[n\phi_3(k) - r\right] & if \quad i \in R^C \end{cases} \tag{1.17}$$

Where $\phi_1(k) = \dfrac{(A+C)\overline{X} + fB\bar{x}_n}{(A+fB)\overline{X} + C\bar{x}_n}$, $\phi_2(k) = \dfrac{(A+C)\bar{x}_n + fB\bar{x}_r}{(A+fB)\bar{x}_n + C\bar{x}_r}$, $\phi_3(k) = \dfrac{(A+C)\overline{X} + fB\bar{x}_r}{(A+fB)\overline{X} + C\bar{x}_r}$,

$A = (k-1)(k-2),\ B = (k-1)(k-4),\ C = (k-2)(k-3)(k-4),\ f = \dfrac{n}{N},\ 0 < k < \infty$

Under (1.15), (1.16) and (1.17) point estimators are:

$$\left.\begin{aligned} T_{FT1} &= \bar{y}_r\phi_1(k) \\ T_{FT2} &= \bar{y}_r\phi_2(k) \\ T_{FT3} &= \bar{y}_r\phi_3(k) \end{aligned}\right\} \tag{1.18}$$

As special cases, when $k = 1, \beta_l = 1$ then $T_{FTl} = t_l$ when $k = 2, \beta_l = -1$ then $T_{FTl} = t_l$
when $k = 4, \beta_l = 0$ then $T_{FTl} = t_l = \bar{y}_r$; $\quad (l = 1,2,3)$

2. Proposed imputation strategies

Consider a double sampling set-up with three variables *Y*, *X* and *Z* where *Y* is the main variable and *X*, *Z* are auxiliary variates. The correlation between *X* and *Z* is higher than other two. A specific way of combining *X* and *Z* is "chaining", which generates chain-type estimators in double sampling, and several authors have used this [see Singh and Singh (1991), Singh et al. (1994)] to get a series of alternative estimators for estimating population mean. Singh and Shukla (1987) discussed a family of factor-type ratio estimator for estimating population mean. In one more contribution, Singh and Shukla (1993) derived efficient factor-type estimator for estimating the same population parameter. Using the above contributions Singh et al. (1994) developed a factor-type-chain estimator, whose application as an imputation tool is the main source of motivation in this article.

2.1. Preliminaries

Typically, in double sampling, the population mean $\overline{X}$ of variable *X* is unknown. Hence, let S' be the preliminary sample drawn from $\Omega = \{1,2,....,N\}$ by *SRSWOR* containing *m* units with mean $\bar{x}_m$, $\bar{z}_m$ of *X* and *Z*. This implies $x_{s'} = \{x_j' : j \in S'\}$, $z_{s'} = \{z_j' : j \in S'\}$ are known data and at this stage data linked with variable *Y* are not known. A sub-sample *S* of *n* units $(n < m)$ is drawn from S' by *SRSWOR* having *r* responding units $(r < n)$ forming subspace *R*, having $(n-r)$ non-responding units with the sub-space R^C. Also, in *S*, $y_R = \{y_i, i \in R\}$, $x_s = \{x_i, i \in S\}$, $z_s = \{z_i, i \in S\}$ are available,

whereas $y_{R^C} = \{y_i, i \in R^C\}$ is missing and needs to be estimated by an appropriate imputation technique. As discussed in previous section the sub-spaces R and R^C are disjoint and $R \cup R^C = S$.

Let us consider Ahmed et al. (2006) point estimator from equation (1.10) t_2 with $\beta_2 = 1$:

$$t_2^* = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \qquad (*)$$

The term $\bar{x}_n$ could be improved by Chaining Technique as suggested by Chand (1975), Sukhatme and Chand (1977), Singh and Singh (1991) as:

$$t_2^{**} = \bar{y}_r \frac{\bar{x}_m}{\bar{x}_r} \frac{\bar{Z}}{z_m} \qquad \text{(With } \bar{z}_m \text{ and } \bar{Z} \text{ known)} \qquad (**)$$

Motivated from the above discussion, some proposed imputation strategies using Singh et al. (1994) are:

(G)
$$(y_{C1})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\psi_1(k) - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \qquad (2.1)$$

(H)
$$(y_{C2})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\psi_2(k) - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \qquad (2.2)$$

(I)
$$(y_{C3})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\psi_3(k) - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \qquad (2.3)$$

Where
$$\psi_1(k) = \bar{y}_r \frac{\bar{x}_m}{\bar{x}_r} \frac{(A+C)\bar{Z} + fB\bar{z}_m}{(A+fB)\bar{Z} + C\bar{z}_m} \qquad (2.4)$$

$$\psi_2(k) = \bar{y}_r \frac{\bar{x}_m}{\bar{x}_r} \frac{(A+C)\bar{z}_m + fB\bar{z}_r}{(A+fB)\bar{z}_m + C\bar{z}_r} \qquad (2.5)$$

$$\psi_3(k) = \bar{y}_r \frac{\bar{x}_m}{\bar{x}_r} \frac{(A+C)\bar{Z} + fB\bar{z}_r}{(A+fB)\bar{Z} + C\bar{z}_r} \qquad (2.6)$$

Where $A = (k-1)(k-2); B = (k-1)(k-4); C = (k-2)(k-3)(k-4)$ and $0 < k < \infty$, is a constant. Also, $\bar{y}_r = \frac{1}{r}\sum_{i \in R} y_i$, $\bar{x}_r = \frac{1}{r}\sum_{i \in R} x_i$, $\bar{z}_r = \frac{1}{r}\sum_{i \in R} z_i$, $\bar{x}_m = \frac{1}{m}\sum_{i \in S'} x_i$, $\bar{z}_m = \frac{1}{m}\sum_{i \in S'} z_i$, $\bar{Z} = \frac{1}{N}\sum_{i \in \Omega} Z_i$.

Under strategies (2.1), (2.2) and (2.3) the point estimators of population mean of study variable $\bar{Y}$ are like (2.4), (2.5) and (2.6) respectively.

2.2. Special Cases:

(i) At $k=1$; $A=0, B=0, C=-6$

$$\psi_1(1)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r}\frac{\bar{Z}}{\bar{z}_m}\ ; \qquad \psi_2(1)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r}\frac{\bar{z}_m}{\bar{z}_r}; \qquad \psi_3(1)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r}\frac{\bar{Z}}{\bar{z}_r} \tag{2.7}$$

(ii) At $k=2; A=0, B=-2, C=0$

$$\psi_1(2)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r}\frac{\bar{z}_m}{\bar{Z}}; \qquad \psi_2(2)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r}\frac{\bar{z}_r}{\bar{z}_m}; \qquad \psi_3(2)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r}\frac{\bar{z}_r}{\bar{Z}} \tag{2.8}$$

(iii) At $k=3; A=2, B=-2, C=0$

$$\psi_1(3)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r}\frac{Z-f\bar{z}_m}{(1-f)\bar{Z}}; \quad \psi_2(3)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r}\frac{\bar{z}_m-f\bar{z}_r}{(1-f)\bar{z}_m}; \quad \psi_3(3)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r}\frac{\bar{Z}-f\bar{z}_r}{(1-f)\bar{Z}} \tag{2.9}$$

(iv) At $k=4; A=6, B=0, C=0$

$$\psi_1(4)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r}; \qquad \psi_2(4)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r}; \qquad \psi_3(4)=\bar{y}_r\frac{\bar{x}_m}{\bar{x}_r} \tag{2.10}$$

3. Properties of the estimators under proposed strategies

Let $B(.)$ and $M(.)$ be the bias and mean squared error (*MSE*) of the estimators under a given sampling design respectively. Let the large sample approximations as $n \to N$ be: $\bar{y}_r=\bar{Y}(1+\delta_1)$; $\bar{x}_r=\bar{X}(1+\delta_2)$; $\bar{x}_m=\bar{X}(1+\delta_3)$; $\bar{z}_r=\bar{Z}(1+\delta_4)$ and $\bar{z}_m=\bar{Z}(1+\delta_5)$

Here, $|\delta_i|<1;\ i=1,2,3,4,5$.

Using the concept of two-phase sampling, following Rao and Sitter (1995) and using the mechanism of MCAR [Heitjan and Basu (1996)], for given r, n and m, we have:

$E(\delta_i)=0;\ i=1,2,3,4,5$; $E(\delta_1^2)=M_1C_Y^2$; $E(\delta_2^2)=M_1C_X^2$; $E(\delta_3^2)=M_2C_X^2$; $E(\delta_4^2)=M_1C_Z^2$;
$E(\delta_5^2)=M_2C_Z^2$; $E(\delta_1\delta_2)=M_1\rho_{YX}C_Y\,C_X$; $E(\delta_1\delta_3)=M_2\rho_{YX}C_YC_X$; $E(\delta_1\delta_4)=M_1\rho_{YZ}C_Y\,C_Z$;
$E(\delta_1\delta_5)=M_2\rho_{YZ}C_Y\,C_Z$; $E(\delta_2\delta_3)=M_2\,C_X^2$; $E(\delta_2\delta_4)=M_1\rho_{XZ}C_X\,C_Z$; $E(\delta_2\delta_5)=M_2\rho_{XZ}C_X\,C_Z$;
$E(\delta_3\delta_4)=M_2\rho_{XZ}C_X\,C_Z$; $E(\delta_3\delta_5)=M_2\rho_{XZ}C_X\,C_Z$; $E(\delta_4\delta_5)=M_2\,C_Z^2$

and $M_1=\frac{1}{r}-\frac{1}{N}$; $M_2=\frac{1}{m}-\frac{1}{N}$; $M_3=M_1-M_2=\frac{1}{r}-\frac{1}{m}$.

Remark 3.1: Define the symbols

$$\phi_1=\frac{fB}{A+fB+C};\phi_2=\frac{C}{A+fB+C};\phi_3=\frac{A+C}{A+fB+C};\phi_4=\frac{A+fB}{A+fB+C};\ (\phi_1+\phi_3)=(\phi_2+\phi_4)=1$$

$$\phi=(\phi_1-\phi_2)=-(\phi_3-\phi_4);K_{YX}=\rho_{YX}\frac{C_Y}{C_X};K_{YZ}=\rho_{YZ}\frac{C_Y}{C_Z};K_{XZ}=\rho_{XZ}\frac{C_X}{C_Z}$$

Theorem 3.1:

[a$_1$] The estimator $\psi_1(k)$ in terms of δ_i; $i=1,2,3,4,5$ up to the first order of approximation is:

$$\psi_1(k)=\overline{Y}\left[1+\delta_1-\delta_2+\delta_3-\delta_1\delta_2+\delta_1\delta_3-\delta_2\delta_3+\delta_2^2+\phi\left(\delta_5+\delta_1\delta_5-\delta_2\delta_5+\delta_3\delta_5-\phi_2\delta_5^2\right)\right] \quad (3.1)$$

[a$_2$] Bias of $\psi_1(k)$:

$$B[\psi_1(k)]=\overline{Y}\left[M_3C_X^2(1-K_{YX})-\phi M_2C_Z^2(\phi_2-K_{YZ})\right] \quad (3.2)$$

[a$_3$] Mean squared error of $\psi_1(k)$:

$$M[\psi_1(k)]=\overline{Y}^2\left[M_1C_Y^2+M_3C_X^2(1-2K_{YX})-\phi M_2C_Z^2(\phi+2K_{YZ})\right] \quad (3.3)$$

[a$_4$] Minimum MSE of the estimator $\psi_1(k)$ is when $\phi=-K_{YZ}$ holds and the expression is:

$$M[\psi_1(k)]_{\min}=\overline{Y}^2\left[M_1C_Y^2+M_3C_X^2(1-2K_{YX})+M_2K_{YZ}^2C_Z^2\right] \quad (3.4)$$

Proof:

[a$_1$] $\psi_1(k) = \overline{y}_r\dfrac{\overline{x}_m}{\overline{x}_r}\left[\dfrac{(A+C)\overline{Z}+fB\overline{z}_m}{(A+fB)\overline{Z}+C\overline{z}_m}\right]$

$$=\overline{Y}(1+\delta_1)(1+\delta_2)^{-1}(1+\delta_3)(1+\phi_1\delta_5)(1+\phi_2\delta_5)^{-1}$$

$$=\overline{Y}\left[1+\delta_1-\delta_2+\delta_3-\delta_1\delta_2+\delta_1\delta_3-\delta_2\delta_3+\delta_2^2+\phi\left(\delta_5+\delta_1\delta_5-\delta_2\delta_5+\delta_3\delta_5-\phi_2\delta_5^2\right)\right]$$

[a$_2$] $B[\psi_1(k)]=E\left[\psi_1(k)-\overline{Y}\right]=\left[E[\psi_1(k)]-\overline{Y}\right]$

Using (3.1) and taking expectation both sides

$$E[\psi_1(k)] = \overline{Y}E\left[1-\delta_1\delta_2+\delta_1\delta_3-\delta_2\delta_3+\delta_2^2+\phi\left(\delta_1\delta_5-\delta_2\delta_5+\delta_3\delta_5-\phi_2\delta_5^2\right)\right]$$

$$=\overline{Y}\left[1+M_3C_X^2(1-K_{YX})-\phi M_2C_Z^2(\phi_2-K_{YZ})\right]$$

$$B[\psi_1(k)]=\overline{Y}\left[M_3C_X^2(1-K_{YX})-\phi M_2C_Z^2(\phi_2-K_{YZ})\right]$$

[a$_3$] $M[\psi_1(k)]=E\left[\psi_1(k)-\overline{Y}\right]^2$

$$=E\left[\overline{Y}\left\{1+\delta_1-\delta_2+\delta_3-\delta_1\delta_2+\delta_1\delta_3-\delta_2\delta_3+\delta_2^2+\phi\left(\delta_5+\delta_1\delta_5-\delta_2\delta_5+\delta_3\delta_5-\phi_2\delta_5^2\right)\right\}-\overline{Y}\right]^2$$

[Using (3.1)]

$$=\overline{Y}^2\left[M_1C_Y^2+M_3C_X^2(1-2K_{YX})-\phi M_2C_Z^2(\phi+2K_{YZ})\right]$$

[a$_4$] To obtain minimum *MSE*, let

$$\frac{d}{d\phi}M[\psi_1(k)]=0 \quad \Rightarrow\overline{Y}^2\left[M_2C_Z^2(2\phi+2K_{YZ})\right]=0 \Rightarrow \phi=-K_{YZ}$$

$$M[\psi_1(k)]_{\min}=\overline{Y}^2\left[M_1C_Y^2+M_3C_X^2(1-2K_{YX})+M_2K_{YZ}^2C_Z^2\right]$$

Theorem 3.2:

[a5] The estimator $\psi_2(k)$ in terms of δ_i; $i = 1,2,3,4,5$ up to the first order of approximation is:

$$\psi_2(k) = \overline{Y}\Big[1 + \delta_1 - \delta_2 + \delta_3 - \delta_1\delta_2 + \delta_1\delta_3 - \delta_2\delta_3 + \delta_2^2 + \phi\big(\delta_4 - \delta_5 + \delta_1\delta_4 - \delta_1\delta_5 - \delta_2\delta_4 + \delta_2\delta_5 + \delta_3\delta_4 - \delta_3\delta_5 + (\phi_2 - \phi_4)\delta_4\delta_5 - \phi_2\delta_4^2 + \phi_4\delta_5^2\big)\Big] \quad (3.5)$$

[a6] Bias of the estimator $\psi_2(k)$:

$$B[\psi_2(k)] = \overline{Y}M_3\left[C_X^2(1 - K_{YX}) - \phi C_Z^2(\phi_2 - K_{YZ} + K_{XZ})\right] \quad (3.6)$$

[a7] Mean squared error of $\psi_2(k)$:

$$M[\psi_2(k)] = \overline{Y}^2\left[M_1C_Y^2 + M_3\left\{C_X^2(1 - 2K_{YX}) + \phi C_Z^2(\phi + 2K_{YZ} - 2K_{XZ})\right\}\right] \quad (3.7)$$

[a8] Minimum MSE of $\psi_2(k)$ is at $\phi = (-K_{YZ} + K_{XZ})$:

$$M[\psi_2(k)]_{\min} = \overline{Y}^2\left[M_1C_Y^2 + M_3\left\{(1 - 2K_{YX})C_X^2 - (K_{YZ} - K_{XZ})^2C_Z^2\right\}\right] \quad (3.8)$$

Proof:

[a5]
$$\psi_2(k) = \overline{y}_r\left(\frac{\overline{x}_m}{\overline{x}_r}\right)\left[\frac{(A + C)\overline{z}_m + fB\overline{z}_r}{(A + fB)\overline{z}_m + C\overline{z}_r}\right]$$

$$= \overline{Y}(1 + \delta_1)(1 + \delta_2)^{-1}(1 + \delta_3)(1 + \phi_1\delta_4 + \phi_3\delta_5)(1 + \phi_2\delta_4 + \phi_4\delta_5)^{-1}$$

$$= \overline{Y}\Big[1 + \delta_1 - \delta_2 + \delta_3 - \delta_1\delta_2 + \delta_1\delta_3 - \delta_2\delta_3 + \delta_2^2 + \phi\{\delta_4 - \delta_5 + \delta_1\delta_4 - \delta_1\delta_5 - \delta_2\delta_4 + \delta_2\delta_5 + \delta_3\delta_4 - \delta_3\delta_5 + (\phi_2 - \phi_4)\delta_4\delta_5 - \phi_2\delta_4^2 + \phi_4\delta_5^2\}\Big]$$

[a6]
$$B[\psi_2(k)] = E\left[\psi_2(k) - \overline{Y}\right] = E[\psi_2(k)] - \overline{Y}$$

Using (3.5) and taking the expectation both sides,

$$E[\psi_2(k)] = \overline{Y}E\Big[1 - \delta_1\delta_2 + \delta_1\delta_3 - \delta_2\delta_3 + \delta_2^2 + \phi\{\delta_1\delta_4 - \delta_1\delta_5 - \delta_2\delta_4 + \delta_2\delta_5 + \delta_3\delta_4 - \delta_3\delta_5 + (\phi_2 - \phi_4)\delta_4\delta_5 - \phi_2\delta_4^2 + \phi_4\delta_5^2\}\Big]$$

$$= \overline{Y}\left[1 + M_3\left\{C_X^2(1 - K_{YX}) - \phi C_Z^2(\phi_2 - K_{YZ} + K_{XZ})\right\}\right]$$

$$B[\psi_2(k)] = E[\psi_2(k)] - \overline{Y}$$

$$= M_3\overline{Y}\left[C_X^2(1 - K_{YX}) - \phi C_Z^2(\phi_2 - K_{YZ} + K_{XZ})\right]$$

[a7]
$$M[\psi_2(k)] = E\left[\psi_2(k) - \overline{Y}\right]^2$$

$$= \overline{Y}^2E\left[(\delta_1 - \delta_2 + \delta_3) + \phi(\delta_4 - \delta_5)\right]^2 \qquad \text{[Using (3.5)]}$$

$$M[\psi_2(k)] = \overline{Y}^2\left[M_1C_Y^2 + M_3\left\{C_X^2(1 - 2K_{YX}) + \phi C_Z^2(\phi + 2K_{YZ} - 2K_{XZ})\right\}\right]$$

[a8] To obtain minimum MSE, let

$$\frac{d}{d\phi} M\left[\psi_2(k)\right] = 0 \quad \Rightarrow \quad \phi = K_{XZ} - K_{YZ}$$

and substitution provides

$$M\left[\psi_2(k)\right]_{\min} = \overline{Y}^2\left[M_1 C_Y^2 + M_3\left\{(1-2K_{YX})C_X^2 - \left(K_{YZ} - K_{XZ}\right)^2 C_Z^2\right\}\right]$$

Theorem 3.3:

[a9] The estimator $\psi_3(k)$ in terms of $\delta_i;\ i = 1,2,3,4,5$ up to the first order of approximation could be expressed as:

$$\psi_3(k) = \overline{Y}\left[1 + \delta_1 - \delta_2 + \delta_3 - \delta_1\delta_2 + \delta_1\delta_3 - \delta_2\delta_3 + \delta_2^2 + \phi\left(\delta_4 + \delta_1\delta_4 - \delta_2\delta_4 + \delta_3\delta_4 - \phi_2\delta_4^2\right)\right] \tag{3.9}$$

[a10] Bias of $\psi_3(k)$:

$$B\left[\psi_3(k)\right] = \overline{Y}\left[M_3 C_X^2\left(1 - K_{YX}\right) + \phi C_Z^2\left(M_1 K_{YZ} - M_3 K_{XZ} - M_1\phi_2\right)\right] \tag{3.10}$$

[a11] Mean squared error of $\psi_3(k)$:

$$M\left[\psi_3(k)\right] = \overline{Y}^2\left[M_1 C_Y^2 + M_3 C_X^2(1-2K_{YX}) + \phi C_Z^2\left(\phi M_1 + 2M_1 K_{YZ} - 2M_3 K_{XZ}\right)\right] \tag{3.11}$$

[a12] Minimum MSE of $\psi_3(k)$ is when $\phi = M_1^{-1}\left(M_3 K_{XZ} - M_1 K_{YZ}\right)$ and the expression is:

$$M\left[\psi_3(k)\right]_{\min} = \overline{Y}^2\left[M_1 C_Y^2 + M_3 C_X^2\left(1 - 2K_{YX}\right) - \left(M_3 K_{XZ} - M_1 K_{YZ}\right)^2 M_1^{-1} C_Z^2\right] \tag{3.12}$$

Proof:

[a9] $\psi_3(k) = \overline{y}_r\left(\frac{\overline{x}_m}{\overline{x}_r}\right)\left[\frac{(A+C)\overline{Z} + fB\overline{z}_r}{(A+fB)\overline{Z} + C\overline{z}_r}\right] = \overline{Y}(1+\delta_1)(1+\delta_2)^{-1}(1+\delta_3)(1+\phi_1\delta_4)(1+\phi_2\delta_4)^{-1}$

$$= \overline{Y}(1+\delta_1)(1-\delta_2+\delta_2^2-\delta_3^3+.....)(1+\delta_3)(1+\phi_1\delta_4)\left(1-\phi_2\delta_4+\phi_2^2\delta_4^2-\phi_2^2\delta_4^3+.....\right)$$

$$= \overline{Y}\left[1 + \delta_1 - \delta_2 + \delta_3 - \delta_1\delta_2 + \delta_1\delta_3 - \delta_2\delta_3 + \delta_2^2 \quad + \phi\left(\delta_4 + \delta_1\delta_4 - \delta_2\delta_4 + \delta_3\delta_4 - \phi_2\delta_4^2\right)\right]$$

[a10] $B\left[\psi_3(k)\right] = E\left[\psi_3(k) - \overline{Y}\right]$

$$= \overline{Y}E\left[\delta_1 - \delta_2 + \delta_3 - \delta_1\delta_2 + \delta_1\delta_3 - \delta_2\delta_3 + \delta_2^2 \quad + \phi\left(\delta_4 + \delta_1\delta_4 - \delta_2\delta_4 + \delta_3\delta_4 - \phi_2\delta_4^2\right)\right]$$

$$= \overline{Y}\left[M_3 C_X^2\left(1 - K_{YX}\right) + \phi C_Z^2\left(M_1 K_{YZ} - M_3 K_{XZ} - M_1\phi_2\right)\right]$$

[a11] $M\left[\psi_1(k)\right] = E\left[\psi_1(k) - \overline{Y}\right]^2 = \overline{Y}^2 E\left[\delta_1 - \delta_2 + \delta_3 + \phi\delta_4\right]^2$

$$= \overline{Y}^2\left[M_1 C_Y^2 + M_3 C_X^2 - 2M_3\rho_{YX} C_Y C_X + \phi^2 M_1 C_Z^2 + 2\phi\left(M_1\rho_{YZ} C_Y C_Z - M_3\rho_{XZ} C_X C_Z\right)\right]$$

$$= \overline{Y}^2\left[M_1 C_Y^2 + M_3 C_X^2\left(1 - 2K_{YX}\right) \quad + \phi C_Z^2\left(\phi M_1 + 2M_1 K_{YZ} - 2M_3 K_{XZ}\right)\right]$$

[a12] To obtain minimum MSE, let

$$\frac{d}{d\phi} M\left[\psi_3(k)\right] = 0 \quad \Rightarrow \qquad \phi = M_1^{-1}\left(M_3 K_{XZ} - M_1 K_{YZ}\right)$$

and substitution provides

$$M\left[\psi_3(k)\right]_{\min} = \overline{Y}^2\left[M_1 C_Y^2 + M_3 C_X^2\left(1 - 2K_{YX}\right) - \left(M_3 K_{XZ} - M_1 K_{YZ}\right)^2 M_1^{-1} C_Z^2\right]$$

4. Comparison of the estimators under proposed imputation strategies

[b₁]
$$D_1 = M[\psi_1(k)]_{\min} - M[\psi_2(k)]_{\min}$$
$$= \overline{Y}^2 C_Z^2 \left[M_3 (K_{YZ} - K_{XZ})^2 - M_2 K_{YZ}^2 \right] \quad (4.1)$$
$\psi_2(k)$ is better over $\psi_1(k)$ if $D_1 > 0$

$$\Rightarrow \quad \frac{K_{YZ} - K_{XZ}}{K_{YZ}} > \sqrt{\frac{M_2}{M_3}} \Rightarrow \quad F_1 > F_2 \quad \text{(let)}$$

[b₂]
$$D_2 = M[\psi_1(k)]_{\min} - M[\psi_3(k)]_{\min}$$
$$= \overline{Y}^2 C_Z^2 \left[(M_3 K_{XZ} - M_1 K_{YZ})^2 M_1^{-1} - M_2 K_{YZ}^2 \right] \quad (4.2)$$
$\psi_3(k)$ is better over $\psi_1(k)$ if $D_2 > 0$

$$\Rightarrow \quad \frac{K_{XZ}}{K_{YZ}} > \frac{M_1 + \sqrt{M_1 M_2}}{M_3} \quad \Rightarrow \quad F_3 > F_4 \quad \text{(let)}$$

[b₃]
$$D_3 = M[\psi_2(k)]_{\min} - M[\psi_3(k)]_{\min}$$
$$= \overline{Y}^2 C_Z^2 \left[(M_3 K_{XZ} - M_1 K_{YZ})^2 M_1^{-1} - M_3 (K_{YZ} - K_{XZ})^2 \right] \quad (4.3)$$
$\psi_3(k)$ is better than $\psi_2(k)$ if $D_3 > 0$

$$\Rightarrow \quad \frac{K_{XZ}}{K_{YZ}} > \frac{M_1 + \sqrt{M_1 M_3}}{M_3 + \sqrt{M_1 M_3}} \quad \Rightarrow \quad F_3 > F_5 \quad \text{(let)}$$

5. Empirical study

For numerical study consider the data as attached in Appendix A, which is a generated artificial population of size $N = 200$ containing values of main variable Y and auxiliary variables X, Z. Parameters of this population are given below:

$\overline{Y} = 42.485$; $\overline{X} = 18.515$; $\overline{Z} = 20.52$; $S_Y^2 = 199.0598$; $S_X^2 = 48.5375$; $S_Z^2 = 45.7684$;
$\rho_{YX} = 0.8734$; $\rho_{YZ} = 0.8667$; $\rho_{XZ} = 0.9943$; $C_Y = 0.3287$; $C_X = 0.3755$; $C_Z = 0.3296$;
$K_{YZ} = 0.8643$; $K_{XZ} = 1.1326$; $K_{YX} = 0.7645$

Reddy (1978) proved that K_{YX}, K_{YZ}, K_{XZ} are ratio values and bear very small change over a span of time. It could be easily guessed or assumed to be known a priori. Using preliminary large sample of size $m = 80$ and sub-random sample of size $n = 30$ with the number of responding units $r = 22$ and $f = 0.15$ by *SRSWOR*. The optimum values of constants of different estimators at their optimal condition are $\alpha = 0.2354$, $\beta_1 = \beta_2 = \beta_3 = 0.7646$, $k_1' = 1.5206$, $k_2' = 2.4505$, $k_3' = 8.9456$ for compromised, Ahmed's methods and Factor Type F-T Estimators of imputation respectively. By simplifying optimum conditions of proposed estimators for minimum MSE, the cubic equations provide the values of constants k as shown in Table 5.1.

Table 5.1. Optimum *k*-values for minimum *MSE* of proposed estimators

Estimators	Condition for Optimum *MSE*	Three optimum Values of *k* on one condition		
$\psi_1(k)$	$\phi = -K_{YZ}$	k_1 = 1.3137	k_2 = 2.5180	k_3 = 13.5979
$\psi_2(k)$	$\phi = K_{XZ} - K_{YZ}$	k_4 = 1.9321	k_5 = -----	k_6 = -----
$\psi_3(k)$	$\phi = M_1^{-1}(M_3 K_{XZ} - M_1 K_{YZ})$	k_7 = 1.8759	k_8 = 3.2154	k_9 = 4.0919

Note: k_5, k_6 do not exist because the solution of cubic equations provided no real roots.

The formula for efficiency measurement is $e(\hat{T}) = \frac{MSE(\overline{y}_r)}{MSE(\hat{T})}$, where $\hat{T}$ is any estimator under consideration. The steps followed for the simulation procedure are:

Step 1: Draw a preliminary random sample *S*' of size m = 80 from the population of size 200.

Step 2: Again draw a random sub-sample of size n = 30 from *S*' drawn in step 1.

Step 3: Drop away 8 units randomly from each sample corresponding to variable *Y*.

Step 4: Compute and impute the dropped units of *Y* with the help of existing and proposed imputation methods.

Step 5: Obtain the estimates of the population mean for existing and proposed imputation methods.

Step 6: Repeat the above steps (1 to 5) 50,000 times, which provides multiple sample based estimates $\hat{T}_1, \hat{T}_2, \hat{T}_3, ..., \hat{T}_{50,000}$.

Step 7: The bias of $\hat{T}$ is obtained by $B(\hat{T}) = \frac{1}{50000}\sum_{i=1}^{50000}(\hat{T}_i - \overline{Y})$.

Step 8: The *MSE* of $\hat{T}$ is obtained by $MSE(\hat{T}) = \frac{1}{50000}\sum_{i=1}^{50000}(\hat{T}_i - \overline{Y})^2$.

Following the above procedure bias and *MSE* of the existing and proposed estimators are computed based on 50,000 repeated samples drawn by SRSWOR from population of N = 200. These computations and efficiencies with respect to $\overline{y}_r$ are given in Tables 5.2 and 5.3 respectively.

Table 5.2. Bias and *MSE* of existing estimators

Estimators	Optimum Value	Bias	*MSE*	Efficiency
$\overline{y}_r$	-----	-0.3123	9.7252	1
$\overline{y}_{RAT}$	-----	-0.0996	7.8457	1.2395
$\overline{y}_{COMP}$	$\alpha = 0.2354$	-0.0809	6.9649	1.3963
t_1	$\beta_1 = 0.7646$	-0.3983	5.8967	1.6492

Table 5.2. Bias and MSE of existing estimators (cont.)

Estimators	Optimum Value	Bias	*MSE*	Efficiency
t_2	$\beta_2 = 0.7646$	-0.1871	7.6655	1.2686
t_3	$\beta_3 = 0.7646$	-0.2151	3.2967	2.9499
T_{FT1}	$k_1^{'} = 1.5206$	-0.3878	4.8327	2.0123
	$k_2^{'} = 2.4505$	-0.3736	5.1655	1.8827
	$k_3^{'} = 8.9456$	-0.3961	4.9454	1.9665
T_{FT2}	$k_1^{'} = 1.5206$	-0.1071	6.3071	1.5419
	$k_2^{'} = 2.4505$	-0.0329	6.1072	1.5924
	$k_3^{'} = 8.9456$	-0.0980	6.0561	1.6058
T_{FT3}	$k_1^{'} = 1.5206$	-0.1826	1.8399	5.2857
	$k_2^{'} = 2.4505$	-0.1944	2.2685	4.2870
	$k_3^{'} = 8.9456$	-0.1818	1.9894	4.8885

5.1. Numerical computation of proposed estimators

From Section 4.0 we get computational values of conditions on the population given in Appendix A. $F_1 = \frac{K_{YZ} - K_{XZ}}{K_{YZ}} =$ - 0.3104; $F_2 = \sqrt{\frac{M_2}{M_3}} =$ 0.4774; $F_3 = \frac{K_{XZ}}{K_{YZ}} =$ 1.3104; $F_4 = \frac{M_1 + \sqrt{M_1 M_2}}{M_3} =$ 1.7570 and $F_5 = \frac{M_1 + \sqrt{M_1 M_3}}{M_3 + \sqrt{M_1 M_3}} =$ 1.1082

Since $F_1 < F_2$ holds, $\psi_1(k)$ is better than $\psi_2(k)$ for this data set.
Again, $F_3 < F_4$, which implies $\psi_1(k)$ is better than $\psi_3(k)$ for the data set, and $F_3 > F_5$, which implies $\psi_3(k)$ is better than $\psi_2(k)$ for this data set. Overall $\psi_1(k)$ is the best estimator.

Table 5.3. Bias and *MSE* of proposed chain type estimators

Estimator	*k-optimum*	Bias	*MSE*	Efficiency
$\psi_1(k)$	k_1 =1.3137	-0.0030	1.9169	5.0734
	k_2 =2.5180	0.0215	1.9328	5.0317
	k_3 =13.5979	-0.0038	1.9409	5.0106
$\psi_2(k)$	k_4 =1.9321	0.3534	9.0303	1.0769
	k_5 = -----	——	——	-----
	k_6 = -----	——	——	-----
$\psi_3(k)$	k_7 =1.8759	0.6036	8.6779	1.1206
	k_8 =3.2154	0.6215	8.6360	1.1261
	k_9 = 4.0919	0.5992	8.6621	1.1227

6. Almost unbiased imputation based chain type estimator

By expression (3.2), (3.6) and (3.10), bias of $\psi_i(k)$; i = 1, 2, 3 could be made zero up to the first order of approximation. This provides three equations:

$$M_3C_X^2(1-K_{YX})-\phi M_2C_Z^2(\phi_2-K_{YZ})=0 \tag{6.1}$$

$$C_X^2(1-K_{YX})-\phi C_Z^2(\phi_2-K_{YZ}+K_{XZ})=0 \tag{6.2}$$

and

$$M_3C_X^2(1-K_{YX})+\phi C_Z^2(M_1K_{YZ}-M_3K_{XZ}-M_1\phi_2)=0 \tag{6.3}$$

These equations are cubic or more function of k-values to provide multiple values of k on which bias is zero. The best choice is to have lowest mean squared error. So, the proposed estimators bear property of reducing *MSE* along with being almost unbiased also. Many similar estimators existing in the literature do not control both bias and *MSE* at their optimal level but the proposed estimators have this property. For equation (6.1), we get two real values k_1'' = 0.3829 and k_2'' =6.5038 and from (6.2) and (6.3) all values are imaginary, viz. there are no real roots. These results are obtained using the data set on which the empirical study was performed. The term almost unbiased is used because biases of proposed estimates $\psi_i(k)$ are obtained only up to the first order of approximation. The bias $B[\psi_2(k)]=0$ holds approximately not completely, therefore mentioned almost unbiased.

Table 6.1. Almost unbiased comparison of chain type estimators

***k*-values**	$\psi_1(k)$		$\psi_2(k)$		$\psi_3(k)$	
	Bias	***MSE***	**Bias**	***MSE***	**Bias**	***MSE***
k_1'' = 0.3829	0.0005	4.4522	0.0002	15.4062	0.0002	14.4033
k_2'' =6.5038	0.0004	2.4831	0.0001	7.4559	0.0011	6.4898

7. Discussion and conclusions

In the present article some imputation procedures and their estimators of population mean are suggested and the expression of their bias, mean squared error and minimum mean squared error have been derived under large sample approximations up to the first order. An empirical study has been done over a data set and the bias and mean squared error have been calculated. Among the existing and proposed estimators, under Chain-based imputation strategies, i.e. $\psi_i(k)$; (i = 1, 2, 3), the estimator $\psi_1(k)$ is found best. The general perception regarding imputation of missing data is that it increases the bias of the estimate when *MSE* is optimized. In contrary, a key feature of $\psi_i(k)$; (i = 1, 2, 3) is that there are many values of the parameter k on which *MSE* is optimal. One can choose the value with the lowest bias. Therefore, suggested strategies are bias-controlled at the optimum level of *MSE*. Apart from this, estimators are almost unbiased also over multiple choices of k-values. The

best selection is to have the lowest *MSE* by proposed strategies one can get almost unbiased estimator with lowest possible *MSE*. Thus, the suggested Chain-based imputation strategies $\psi_i(k)$; (i = 1, 2, 3) are useful and have advantage over other similar procedures.

Acknowledgement

Authors are thankful to the reviewers of this journal for their critical comments and useful suggestions, which has improved the quality of the manuscript.

References

Ahmed, M. S., Al-Titi, O., Al-Rawi, Z. and Abu-Dayyeh, W., (2006). Estimation of a population mean using different imputation methods. *Statistics in Transition*, 7, 6, pp. 1247-1264.

Al-Jararha, J., Ahmed, M. S., (2002). The class of chain estimators for a finite population variance using double sampling. *Information and Management Sciences*, 13(2), pp. 13–18.

Bhaskaran, K., Smeeth, L., (2014). What is the difference between missing completely at random and missing at random? I*nternational Journal of Epidemiology*, 43(4), pp. 1336–1339.

Bose, C., (1943). Note on the sampling error in the method of double sampling. *Sankhya*, 6, 330.

Chand, L., (1975). Some ratio-type estimators based on two or more auxiliary variables unpublished Ph.D. Thesis, *IOWA State University*, Ames, Iowa, U.S.A.

Choudhury, S., Singh, B. K., (2012). A class of chain ratio-cum-dual to ratio type estimator with two auxiliary characters under double sampling in sample surveys. *Statistics in Transition-new series*, 13(3), pp. 519–536.

Cochran, W. G., (2005). Sampling Techniques. *John Wiley and Sons*, New York.

Doretti, M., Geneletti, S. and Stanghellini, E., (2018). Missing data: A unified taxonomy guided by conditional independence. *International Statistical Review*, 86(2), pp. 189–204.

Heitjan, D. F., Basu, S., (1996). Distinguishing 'missing at random' and 'missing completely at random'. *The American Statistician*, 50, pp. 207–213.

Kadilar, C., Cingi, H., (2003). A study on the chain ratio-type estimator. *Hacettepe Journal of Mathematics and Statistics*, 32, pp. 105–108.

Kiregyera, B., (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. *Metrika*, 27 (1), pp. 217–223.

Kiregyera, B., (1984). Regression type estimators using two auxiliary variables and the model of double sampling from finite population. *Metrika*, 31, pp. 215–226.

Kumar, M., Bahl, S., (2006). Class of dual to ratio estimators for double sampling. *Statistical Papers*, 47, pp. 319–326.

Little, R. J. A., Rubin, D. B., (1987). Statistical analysis with missing data, New York: *John Wiley & Sons, Inc.*

Pandey, R., Thakur, N. S. and Yadav, K., (2016). *A*dapted factor-type imputation strategies. *Journal of Scientific Research*, J. Sci. Res., 8(3), pp. 321–339.

Pandey, R., Thakur, N. S. and Yadav, K., (2015). Estimation of population mean using exponential ratio type imputation method under survey non-response. *Journal of the Indian Statistical Association*, Vol.53 No. 1 & 2, pp. 89–107.

Pradhan, B. K., (2005). A chain regression estimator in two phase sampling using multi-auxiliary information. *Bulletin of the Malaysian Mathematical Sciences Society (2),* 28(1), pp. 81–86.

Rao, J. N. K., Sitter, R. R., (1995). Variance estimation under two-phase sampling with application to imputation for missing data*. Biometrika*, 82, pp. 453–460.

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

Rubin, D. B., (1976). Inference and missing data*. Biometrika*, 63, pp. 581–593.

Seaman, S., Galati, J., Jackson, D. and Carlin, J., (2013). *What is meant by "Missing at Random"? Statistical Science*, 28(2), pp. 257–268.

Sharma, B., Tailor, R., (2010). A new ratio-cum-dual to ratio estimator of finite population mean in simple random sampling. *Global Journal of Science Frontier Research*, 10(1), pp. 27–31.

Shukla, D., (2002). F-T estimator under two-phase sampling. *Metron*, 59, 1–2, pp. 253–263.

Shukla, D., Thakur, N. S., Pathak, S. and Rajput, D. S., (2009). Estimation of mean under imputation of missing data using factor type estimator in two-phase sampling. *Statistics in Transition*, Vol. 10, No. 3, pp. 397–414.

Shukla, D., Thakur, N. S., (2008). Estimation of mean with imputation of missing data Using Factor Type Estimator. *Statistics in Transition*, 9, 1, pp. 33–48

Shukla, D., Singh, V. K. and Singh, G. N., (1991). On the use of transformation in factor type estimator. *Metron*, 49(1-4), pp. 359–361.

Singh, H. P., Espejo, M. R., (2007). Double sampling ratio-product estimator of a finite population mean in sampling surveys. *Journal of Applied Statistics*, 34(1), pp. 71–85.

Singh, H. P., Mathur, N. and Chandra, P., (2009). A chain-type estimator for population variance using auxiliary variables in two-phase sampling. *Statistics in Transition-new series*, 10(1), pp. 75–84.

Singh, S., Horn, S., (2000). Compromised imputation in survey sampling. *Metrika*, 51, pp. 266–276.

Singh, S., Singh, H. P. and Upadhyaya, L. N., (2006). Chain ratio and regression type estimators for median estimation in survey sampling. *Statistical Papers*, 48, pp. 23–46.

Singh, V. K., Shukla, D., (1987). One parameter family of factor-type ratio estimator. *Metron*, 45, 1-2, pp. 273–283.

Singh, V. K., Shukla, D., (1993). An efficient one parameter family of factor – type estimator in sample survey. *Metron*, 51, 1–2, pp. 139–159.

Singh, V. K., Singh, G. N., (1991). Chain type estimator with two auxiliary variables under double sampling scheme. *Metron*, 49, pp. 279–289.

Singh, V. K., Singh, B. K. and Singh, G. N., (1993). An efficient class of dual to ratio estimators using two auxiliary characteristics. *Journal of Scientific Research*, 43, pp. 219–228.

Singh, V. K., Singh, G. N. and Shukla, D., (1994). A class of chain ratio estimator with two auxiliary variables under double sampling scheme. *Sankhya*, Ser. B., 46, 2, pp. 209–221.

Srivastava, S. K., Jhajj, H. S., (1980). A class of estimators using auxiliary information for estimating finite population variance. *Sankhya*, 42, pp. 87–96.

Srivastava, S. R., Khare, B. B. and Srivastava, S. R., (1990). A generalized chain ratio estimator for mean of finite population. *Journal of Indian Society of Agricultural Statistics,* 42(I), pp. 108–117.

Srivenkataramana, T., (1980). A dual to ratio estimator in sample surveys. *Biometrika*, 67(1), pp. 199–204.

Sukhatme, B. V., Chand, L., (1977). Multivariate ratio-type estimators, Proceeding of American Statistical Association. *Social Statistics Section,* pp. 927–931.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. and Ashok, C., (1984). Sampling Theory of Surveys with Applications. *Iowa State University Press*, I.S.A.S. Publication, New Delhi.

Weeks, M., (1999). Methods of imputation for missing data (fifth draft), Faculty of Economics and Politics and Department of Applied Econometrics. *University of Cambridge*, Cambridge, UK.

Appendix

A. Population (N = 200)

Y_i	45	50	39	60	42	38	28	42	38	35
X_i	15	20	23	35	18	12	8	15	17	13
Z_i	16	22	26	37	19	14	11	17	18	15
Y_i	40	55	45	36	40	58	56	62	58	46
X_i	29	35	20	14	18	25	28	21	19	18
Z_i	30	37	23	15	19	27	30	22	21	21
Y_i	36	43	68	70	50	56	45	32	30	38
X_i	15	20	38	42	23	25	18	11	09	17
Z_i	18	22	39	44	25	26	19	13	12	20
Y_i	35	41	45	65	30	28	32	38	61	58
X_i	13	15	18	25	09	08	11	13	23	21
Z_i	16	17	19	27	12	10	13	14	24	23
Y_i	65	62	68	85	40	32	60	57	47	55
X_i	27	25	30	45	15	12	22	19	17	21
Z_i	28	26	33	46	17	15	23	20	19	23
Y_i	67	70	60	40	35	30	25	38	23	55
X_i	25	30	27	21	15	17	09	15	11	21
Z_i	26	32	30	23	17	18	12	18	14	24
Y_i	50	69	53	55	71	74	55	39	43	45
X_i	15	23	29	30	33	31	17	14	17	19
Z_i	17	24	30	33	35	32	19	16	19	21
Y_i	61	72	65	39	43	57	37	71	71	70
X_i	25	31	30	19	21	23	15	30	32	29
Z_i	27	33	32	21	23	25	17	32	33	32
Y_i	73	63	67	47	53	51	54	57	59	39
X_i	28	23	23	17	19	17	18	21	23	20
Z_i	30	25	24	20	22	20	21	23	26	22
Y_i	23	25	35	30	38	60	60	40	47	30
X_i	07	09	15	11	13	25	27	15	17	11
Z_i	10	11	18	14	14	26	29	18	20	14
Y_i	57	54	60	51	26	32	30	45	55	54
X_i	31	23	25	17	09	11	13	19	25	27
Z_i	32	25	27	19	12	13	14	20	27	28
Y_i	33	33	20	25	28	40	33	38	41	33
X_i	13	11	07	09	13	15	13	17	15	13
Z_i	16	14	9	10	14	17	14	20	17	15
Y_i	30	35	20	18	20	27	23	42	37	45
X_i	11	15	08	07	09	13	12	25	21	22
Z_i	13	18	11	8	12	16	14	26	24	23

Y_i	37	37	37	34	41	35	39	45	24	27
X_i	15	16	17	13	20	15	21	25	11	13
Z_i	16	18	19	16	22	18	23	26	14	14
Y_i	23	20	26	26	40	56	41	47	43	33
X_i	09	08	11	12	15	25	15	25	21	15
Z_i	11	10	14	15	17	26	17	27	22	17
Y_i	37	27	21	23	24	21	39	33	25	35
X_i	17	13	11	11	09	08	15	17	11	19
Z_i	19	16	13	12	12	11	17	20	13	20
Y_i	45	40	31	20	40	50	45	35	30	35
X_i	21	23	15	11	20	25	23	17	16	18
Z_i	22	25	18	13	21	27	26	19	17	19
Y_i	32	27	30	33	31	47	43	35	30	40
X_i	15	13	14	17	15	25	23	17	16	19
Z_i	17	16	16	14	17	28	25	18	18	22
Y_i	35	35	46	39	35	30	31	53	63	41
X_i	19	19	23	15	17	13	19	25	35	21
Z_i	22	21	24	17	20	15	22	26	36	23
Y_i	52	43	39	37	20	23	35	39	45	37
X_i	25	19	18	17	11	09	15	17	19	19
Z_i	26	20	20	19	13	12	17	18	21	22

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 113–128, DOI 10.2478/stattrans-2022-0045
Received – 23.10.2021; accepted – 23.05.2022

Zero-modified Poisson-Modification of Quasi Lindley distribution and its application

Ramajeyam Tharshan[1], **Pushpakanthie Wijekoon**[2]

ABSTRACT

The Poisson-Modification of Quasi Lindley (PMQL) distribution is a newly introduced mixed Poisson distribution for over-dispersed count data. The aim of this article is to introduce the Zero-modified PMQL (ZMPMQL) distribution as an alternative to the PMQL distribution in order to accommodate zero inflation/deflation. The method of obtaining the ZMPMQL distribution jointly with some of its important properties, namely the probability mass and distribution functions, mean, variance, index of dispersion, and quantile function are presented. Furthermore, some of its special cases are discussed. The maximum likelihood (ML) estimation method is used for the unknown parameter estimation. A simulation study is conducted in order to evaluate the asymptotic theory of the ML estimation method and to show the superiority of the ML method over the method of moments estimation. The applicability of the introduced distribution is illustrated by using a real-world data set.

Key words: over-dispersion, mixed Poisson distribution, PMQL distribution, zero modification, maximum likelihood estimation

1. Introduction

The Poisson distribution is the most commonly used distribution for modelling count data. One of the important properties of the Poisson distribution is that the mean and variance of the random variable are equal. This property is commonly referred as to equidispersion. However, in some real-world applications, especially actuarial, biomedical, engineering, ecological sciences, and others, observed data do not obey the equidispersion property. Here, the variance of the observed data exceeds the mean. This phenomenon is called over-dispersion (Greenwood and Yule, 1920). In such a situation, the mixed Poisson distributions are often adopted for modelling the count data as an alternative to the Poisson distribution. The literature provides various mixed Poisson distributions as negative binomial/Poisson-gamma (Greenwood and Yule, 1920), Poisson-Gamma product ratio (Irwin, 1975), Poisson-Generalized gamma (Albrecht, 1984), Poisson-Lindley (Sankaran, 1970), Poisson-Sujatha (Shanker, 2016c), Poisson-Quasi Lindley (Grine et al., 2017) distributions, among others.

Even though such mixed Poisson distributions can accommodate the longer right-tails and observed over-dispersion by heterogeneous populations, they do not perform well for observed over-dispersion by zero-inflation/deflation. To tackle this problem, the researchers

[1]Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka & Department of Mathematics and Statistics, University of Jaffna, Jaffna, Sri Lanka. E-mail: tharshan10684@gmail.com. ORCID: https:/orcid.org/0000-0002-6112-2517.

[2]Department of Statistics and Computer Science, University of Peradeniya, Peradeniya, Sri Lanka. E-mail: pushpaw@pdn.ac.lk. ORCID: http://orcid.org/0000-0003-4242-1017.

have proposed several zero-modified mixed Poisson distributions. Here, we point out some notable examples for the zero-modified mixed Poisson distributions. Greenwood and Yule (1920) described the zero-inflated negative binomial (ZINB) distribution; Ghitany et al. (2008) proposed the Zero-truncated Poisson-Lindley distribution; Silva et al. (2018) introduced the Zero-modified Poisson-Sujatha distribution; Xavier et al. (2018) proposed the Zero-modified Poisson-Lindley (ZMPL) distribution.

Recently, Tharshan and Wijekoon (2021) introduced a lifetime distribution, namely the Modification of Quasi Lindley (MQL) distribution. Its probability density function (pdf) is given as

$$f_Y(y;\theta,\alpha,\delta) = \frac{\theta e^{-\theta y}}{(\alpha^3+1)\Gamma(\delta)}\left(\Gamma(\delta)\alpha^3 + (\theta y)^{\delta-1}\right);y>0,\theta>0,\alpha^3>-1,\delta>0, \quad (1)$$

where α and δ are shape parameters and θ is a scale parameter, and y is the respective random variable. Equation (1) presents the mixture of two non-identical distributions, exponential (θ), and gamma (δ,θ) with the mixing proportion, $p = \dfrac{\alpha^3}{\alpha^3+1}$. Then, the same authors (Tharshan and Wijekoon, 2022) obtained the Poisson-Modification of Quasi Lindley (PMQL) distribution by amalgamating the Poisson distribution and the MQL distribution. Its explicit form of the probability mass function (pmf) and some other important statistics are given in Section 2.

This paper aims to modify the PMQL distribution at zero probability to adopt the situation with an excessive number of zeros or a smaller number of zeros. The new distribution will be called the Zero-modified PMQL (ZMPMQL) distribution. The ZMPMQL distribution's unknown parameters will be estimated by the maximum likelihood estimation method. Further, the asymptotic property of the estimation method will be evaluated by a Monte Carlo simulation study.

This paper is structured as follows. Section 2 briefly presents the PMQL distribution and some of its statistical properties. In Section 3, we introduce the ZMPMQL distribution with some of its important structural properties. Its quantile function is discussed in Section 4. Section 5 covers the simulation of the random variables and the maximum likelihood estimator (MLE) for the ZMPMQL distribution. Section 6 studies the asymptotic property of the MLE and the applicability of the ZMPMQL distribution by designing a Monte Carlo simulation study and using a real-world data set, respectively.

2. PMQL distribution

Suppose the random variable $X|\Lambda$ is said to have the Poisson distribution with parameter λ. Then, its pmf can be written as

$$f_{X|\Lambda}(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!};\ x=0,1,2,...,\ \lambda>0. \quad (2)$$

As defined by Tharshan and Wijekoon (2022), the PMQL distribution is the resultant distribution of X by assuming the Poisson parameter λ to be followed the MQL distribution.

Its pmf is given as

$$f_X(x) = \frac{\theta\left(\Gamma(\delta)\Gamma(x+1)\alpha^3(1+\theta)^{\delta-1} + \theta^{\delta-1}\Gamma(x+\delta)\right)}{x!(\alpha^3+1)(1+\theta)^{x+\delta}\Gamma(\delta)}; \tag{3}$$

$$x = 0,1,2,...,\theta > 0, \delta > 0, \alpha^3 > -1.$$

It can be shown that equation (3) represents a two-component mixture of geometric $(\frac{\theta}{1+\theta})$ and negative binomial $(\delta, \frac{1}{1+\theta})$ with the mixing proportion $p = \frac{\alpha^3}{\alpha^3+1}$.

Its corresponding cumulative distribution function (cdf) is given as

$$F_X(x) = \sum_{t=0}^{x} f(x) =$$

$$\frac{\delta(1+\theta)^{\delta-1}\Gamma(\delta)\alpha^3\Gamma(x+1)((1+\theta)^{x+1}-1) + \theta^{\delta}\Gamma(x+\delta+1)_2F_1(1,x+\delta+1;\delta+1;\frac{\theta}{1+\theta})}{(\alpha^3+1)\Gamma(\delta)x!\delta(1+\theta)^{x+\delta+1}} \tag{4}$$

$$;x = 0,1,2,...,\theta > 0, \delta > 0, \alpha^3 > -1,$$

where ${}_2F_1(c,d;r;w)$ is the Gaussian hypergeometric function defined as

$${}_2F_1(c,d;r;w) = \sum_{i=0}^{\infty} \frac{(c)_i(d)_i w^i}{(r)_i i!},$$

which is a special case of the generalized hypergeometric function given by the expression

$${}_aF_b(p_1,p_2,...p_a;q_1,q_2,...q_b;w) = \sum_{i=0}^{\infty} \frac{(p_1)_i...(p_a)_i w^i}{(q_1)_i...(q_b)_i i!},$$

and $(p)_i = \frac{\Gamma(p+i)}{\Gamma(p)} = p(p+1)...(p+i+1)$ is the Pochhammer symbol.

Tharshan and Wijekoon (2022) showed that its r^{th} factorial moment is given as

$$\mu'_{(r)} = \frac{\Gamma(\delta)\Gamma(r+1)\alpha^3 + \Gamma(\delta+r)}{(\alpha^3+1)\Gamma(\delta)\theta^r}. \tag{5}$$

By using the following relationship

$$\mu'_r = E(X^r) = \sum_{i=0}^{r} S(r,i)\mu'_{(i)}\ ;\ \ r = 1,2,...,$$

where $S(r,i)$ is the Stirling numbers of the second kind, which is defined as

$$S(r,i) = \frac{1}{i!}\sum_{j=0}^{i}(-1)^{i-j}\binom{i}{j}j^r\ \ ,\ \ 0 < i < r,$$

they have obtained the raw moments of X. Then, they have shown that its mean and variance are

$$\mu_{(PMQL)} = \frac{\alpha^3+\delta}{(\alpha^3+1)\theta},\ \text{ and }\ \sigma^2_{(PMQL)} = \mu_{(PMQL)} + \mu^2_{(PMQL)}\left(\frac{\alpha^3(\alpha^3+2+\delta(\delta-1))+\delta}{(\alpha^3+\delta)^2}\right),$$

respectively. Its index of dispersion (ID) was derived as

$$ID_{(PMQL)} = \frac{\sigma^2_{(PMQL)}}{\mu_{(PMQL)}} = 1 + \frac{\alpha^3(\alpha^3+2+\delta(\delta-1))+\delta}{(\alpha^3+1)(\alpha^3+\delta)\theta}. \tag{6}$$

It is clear that the $ID_{(PMQL)} > 1$. Then, equation (6) implies that the PMQL distribution is an over-dispersed distribution. Further, the authors derived 2^{nd}, 3^{rd}, and 4^{th} raw moments of the PMQL distribution as

$$\mu_2^{'} = \frac{\theta(\alpha^3+\delta)+2\alpha^3+\delta(\delta+1)}{(\alpha^3+1)\theta^2},$$

$$\mu_3^{'} = \frac{\theta^2(\alpha^3+\delta)+3\theta(2\alpha^3+\delta(\delta+1))+6\alpha^3+\delta(\delta+1)(\delta+2)}{(\alpha^3+1)\theta^3},$$

$$\mu_4^{'} = \frac{1}{(\alpha^3+1)\theta^4}\Bigg(\theta^3(\alpha^3+\delta)+7\theta^2(2\alpha^3+\delta(\delta+1))+6\theta(6\alpha^3+\delta(\delta+1)(\delta+2))$$

$$+24\alpha^3+\delta(\delta+1)(\delta+2)(\delta+3)\Bigg).$$

3. Zero-modified PMQL distribution

The pmf of a zero-modified count distribution is given as

$$f_X(x) = \begin{cases} \phi+(1-\phi)g(0) & \text{for } x=0 \\ (1-\phi)g(x) & \text{for } x=1,2,..., \end{cases}$$

where $g(.)$ is the pmf of the parent count distribution and the parameter ϕ is the zero-modified parameter. Then, the random variable X is said to have the ZMPMQL $(\phi,\theta,\alpha,\delta)$ if its pmf is given as

$$f_X(x) = \begin{cases} \phi+(1-\phi)\dfrac{\theta((1+\theta)^{\delta-1}\alpha^3+\theta^{\delta-1})}{(\alpha^3+1)(1+\theta)^{\delta}} & \text{for } x=0 \\ (1-\phi)\dfrac{\theta\Big(\Gamma(\delta)\Gamma(x+1)\alpha^3(1+\theta)^{\delta-1}+\theta^{\delta-1}\Gamma(x+\delta)\Big)}{x!(\alpha^3+1)(1+\theta)^{x+\delta}\Gamma(\delta)} & \text{for } x=1,2,..., \end{cases} \tag{7}$$

where $\delta>0, \theta>0, \alpha^3>-1$, and $\frac{\theta((1+\theta)^{\delta-1}\alpha^3+\theta^{\delta-1})}{\theta^{\delta}-(1+\theta)^{\delta-1}(1+\theta+\alpha^3)} \leq \phi \leq 1$.

The corresponding cdf is given as

$$F_{(ZMPMQL)}(x) = \phi+(1-\phi)F_X(x), \tag{8}$$

where $F_X(x)$ is the cdf of the PMQL distribution, which is defined in equation (4).

Note that equation (7) is not a finite mixture model since ϕ can take negative values. Further, various ϕ values adopt various zero-modifications of the PMQL distribution.

Remarks:

(i) When $\phi = \frac{\theta((1+\theta)^{\delta-1}\alpha^3+\theta^{\delta-1})}{\theta^{\delta}-(1+\theta)^{\delta-1}(1+\theta+\alpha^3)}$, the ZMPMQL distribution reduces to the zero-truncat -ed PMQL distribution. Here, ϕ no longer appears. The zero-truncated models are commonly used to study the length of hospital stay.

(ii) For $\frac{\theta((1+\theta)^{\delta-1}\alpha^3+\theta^{\delta-1})}{\theta^{\delta}-(1+\theta)^{\delta-1}(1+\theta+\alpha^3)} < \phi < 0$, the ZMPMQL distribution reduces to the zero-deflated PMQL distribution, and zero-deflated models are very rare in practice.

(iii) When $\phi = 0$, the ZMPMQL distribution is the PMQL distribution.

(iv) For $0 < \phi < 1$, the ZMPMQL distribution reduces to the zero-inflated PMQL distribution. This can accommodate more zeros than the actual PMQL distribution.

(v) When $\phi = 1$, the ZMPMQL distribution is degenerated at zero, i.e. all probabilities of the distribution are concentrated at zero.

The pmf of the ZMPMQL distribution is shown in Figure 1. We can observe that the parameter ϕ controls the observed counts of zeros.

The mean, variance, and index of dispersion of the ZMPMQL distribution are given,

$$\mu_{(ZMPMQL)} = (1-\phi)\mu_{(PMQL)}, \quad \sigma^2_{(ZMPMQL)} = (1-\phi)\left(\sigma^2_{(PMQL)} + \phi\mu^2_{(PMQL)}\right)$$

and

$$ID_{(ZMPMQL)} = \frac{\sigma^2_{(PMQL)}}{\mu_{(PMQL)}} + \phi\mu_{(PMQL)} = ID_{(PMQL)} + \phi\mu_{(PMQL)},$$

respectively, where $\mu_{(PMQL)}$, $\sigma^2_{(PMQL)}$, and $ID_{(PMQL)}$ are mean, variance, and index of dispersion of the PMQL distribution, respectively. Further, the 2^{nd}, 3^{rd}, and 4^{th} raw moments of the ZMPMQL distribution are $(1-\phi)\mu_2^{'}$, $(1-\phi)\mu_3^{'}$, and $(1-\phi)\mu_4^{'}$, respectively, where $\mu_2^{'}$, $\mu_3^{'}$, and $\mu_4^{'}$ are 2^{nd}, 3^{rd}, and 4^{th} raw moments of the PMQL distribution, respectively discussed in Section 2.

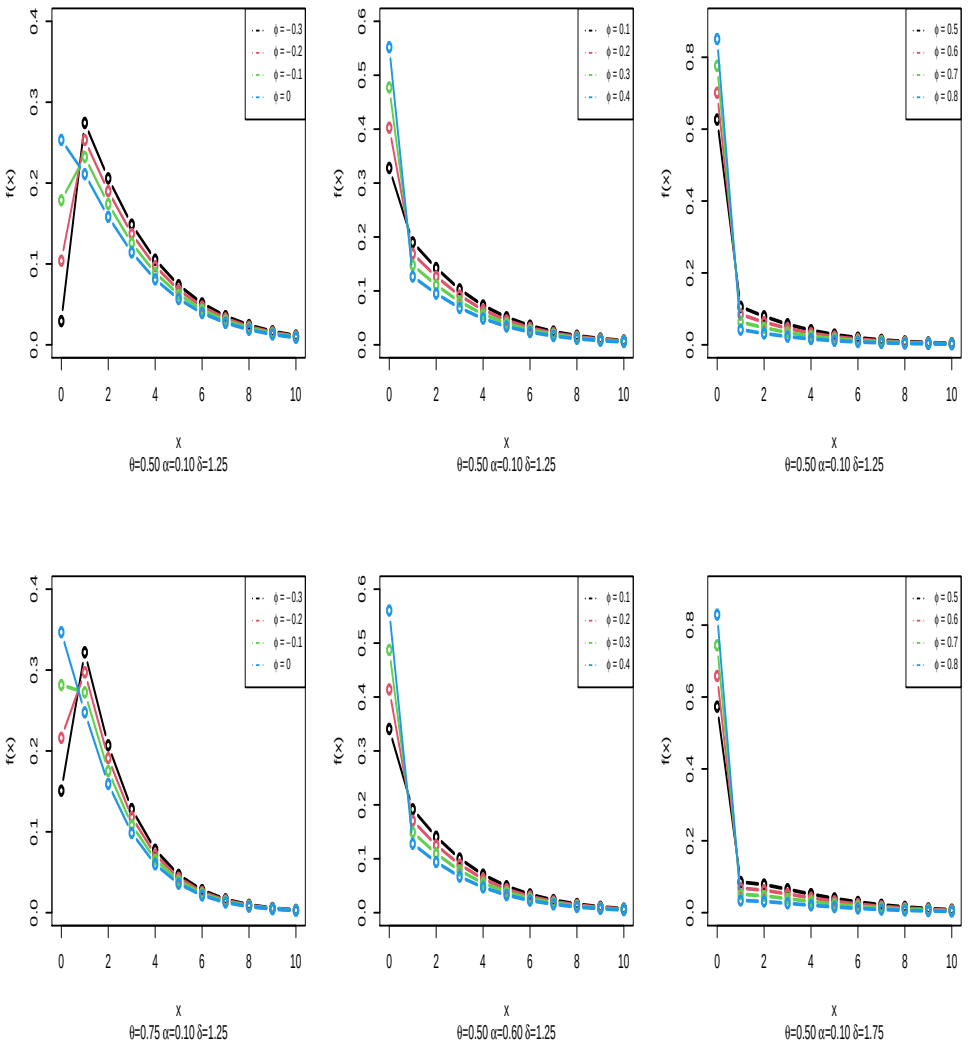


Figure 1: The pmf of the ZMPMQL distribution at different parameter values of ϕ, θ, α, and δ

4. Quantile function

The u^{th} quantile of the ZMPMQL distribution can be derived by solving $F(x_u) = u$ for $0 < u < 1$. It is defined as

$$\phi\beta_1(x_u) + (1-\phi)\left(\beta_2(x_u) + \theta^{\delta}\Gamma(x_u+\delta+1)_2F_1((1, x_u+\delta+1; \delta+1; \frac{\theta}{1+\theta}\right) - u\beta_1(x_u) = 0, \tag{9}$$

where

$$\beta_1(x_u) = (\alpha^3+1)\Gamma(\delta)x_u!\delta(1+\theta)^{x_u+\delta+1},$$

and

$$\beta_2(x_u) = \delta(1+\theta)^{\delta-1}\Gamma(\delta)\alpha^3\Gamma(x_u+1)((1+\theta)^{x_u+1}-1).$$

Since equation (9) is not a linear function with respect to x_u, the estimates of quantiles can be evaluated by using the Newton Raphson method. Further, the first three quantiles can be found by substituting $u = 0.25,\ 0.50$, and 0.75 in equation (9) and solving the respective non-linear equations.

5. Simulation and parameter estimation

5.1. Simulation of random variables

Here, we provide an algorithm to simulate the random variables $x_1, x_2, ..., x_n$ from the ZMPMQL $(\phi, \theta, \alpha, \delta)$ with size n based on the inverse transform method.

Algorithm

i Simulate random variables, $U_i \sim$uniform$(0,1);\ i = 1,2,...,n$.

ii Solve the non-linear equation for $[x_{u_i}]$;
$\phi\beta_1(x_{u_i}) + (1-\phi)\left(\beta_2(x_{u_i}) + \theta^{\delta}\Gamma(x_{u_i}+\delta+1)_2F_1((1, x_{u_i}+\delta+1; \delta+1; \frac{\theta}{1+\theta}\right) - u_i\beta_1(x_{u_i})$
$= 0$, where $\beta_1(x_{u_i})$ and $\beta_2(x_{u_i})$ are defined as in Section 4. Further, $[.]$ denotes the integer part.

5.2. Parameter estimation of the ZMPMQL distribution

This subsection presents the unknown parameter estimation of the ZMPMQL distribution based on the method of moments and the maximum likelihood estimation method.

5.2.1 Method of moments estimator (MME)

Let $x_1, x_2, ..., x_n$ be a random sample of size n from the ZMPMQL distribution. Then, the method of moments estimators of ϕ, θ, α, and δ, abbreviated as $\hat{\phi}_{MME}, \hat{\theta}_{MME}, \hat{\alpha}_{MME}$, and $\hat{\delta}_{MME}$ are found by equating the first four raw moments, $\mu_r^{'}\ (r = 1,2,3,4)$ to the sample moments, say $\frac{\sum_{i=1}^{n} x_i}{n}, r = 1,2,3,4$, and solving the system of non-linear equations. The system of non-linear equations are as follows:

$$n(1-\phi)(\alpha^3+\delta) - (\alpha^3+1)\theta\sum_{i=1}^{n} x_i = 0,$$

$$n(1-\phi)\bigg(\theta(\alpha^3+\delta)+2\alpha^3+\delta(\delta+1)\bigg)-(\alpha^3+1)\theta^2\sum_{i=1}^{n}x_i^2=0,$$

$$n(1-\phi)\bigg(\theta^2(\alpha^3+\delta)+3\theta(2\alpha^3+\delta(\delta+1))+6\alpha^3+\delta(\delta+1)(\delta+2)\bigg)$$

$$-(\alpha^3+1)\theta^3\sum_{i=1}^{n}x_i^3=0,$$

$$n(1-\phi)\bigg(\theta^3(\alpha^3+\delta)+7\theta^2(2\alpha^3+\delta(\delta+1))+6\theta(6\alpha^3+\delta(\delta+1)(\delta+2))$$

$$+24\alpha^3+\delta(\delta+1)(\delta+2)(\delta+3)\bigg)-(\alpha^3+1)\theta^4\sum_{i=1}^{n}x_i^4=0.$$

5.2.2 Maximum likelihood estimator (MLE)

Given a random sample $x_1,x_2,...,x_n$ with size n from the ZMPMQL$(\phi,\theta,\alpha,\delta)$, the likelihood function of the i^{th} sample value x_i is given as

$$L(\phi,\theta,\alpha,\delta|x_i)=\bigg(\phi+(1-\phi)\frac{\theta((1+\theta)^{\delta-1}\alpha^3+\theta^{\delta-1})}{(\alpha^3+1)(1+\theta)^{\delta}}\bigg)^{I_{(X=0)}(x_i)}\times$$

$$\bigg((1-\phi)\frac{\theta(\Gamma(\delta)\Gamma(x_i+1)\alpha^3(1+\theta)^{\delta-1}+\theta^{\delta-1}\Gamma(x_i+\delta)}{x_i!(\alpha^3+1)(1+\theta)^{x_i+\delta}\Gamma(\delta)}\bigg)^{(1-I_{(X=0)}(x_i))},$$

where $I_S(.)$ is the indicator function of subset S. Then, the log-likelihood function is given as

$$\ell(\phi,\theta,\alpha,\delta|x)=$$

$$n_0\log\bigg(\phi+(1-\phi)\frac{\theta((1+\theta)^{\delta-1}\alpha^3+\theta^{\delta-1})}{(\alpha^3+1)(1+\theta)^{\delta}}\bigg)+(n-n_0)\log\bigg(\frac{(1-\phi)\theta}{(\alpha^3+1)\Gamma(\delta)}\bigg)+$$

$$\sum_{i=1}^{n}(1-I_{(X=0)}(x_i))\bigg(\log\bigg(\Gamma(\delta)\Gamma(x_i+1)\alpha^3(1+\theta)^{\delta-1}+\theta^{\delta-1}\Gamma(x_i+\delta)\bigg)-$$

$$\log\bigg(x_i!(1+\theta)^{x_i+\delta}\bigg)\bigg),$$

where $n_0=\sum_{i=1}^{n}I_{(X=0)}(x_i)$, which is the zero counts of the sample.

The score functions are:

$$\frac{\partial\ell(\phi,\theta,\alpha,\delta|x)}{\partial\phi}=\frac{T_1}{T_2}-\frac{n-n_0}{1-\phi},$$

$$\frac{\partial \ell(\phi,\theta,\alpha,\delta|x)}{\partial \theta} =$$

$$\frac{n_0(1-\phi)\left(T_3-T_4\right)}{(1+\theta)^{\delta}T_2}+\frac{n-n_0}{\theta}+\sum_{i=1}^{n}(1-I_{(X=0)}(x_i))\frac{T_5}{T_6}-\sum_{i=1}^{n}(1-I_{(X=0)}(x_i))\frac{(x_i+\delta)}{1+\theta},$$

$$\frac{\partial \ell(\phi,\theta,\alpha,\delta|x)}{\partial \alpha} =$$

$$\frac{T_7}{(\alpha^3+1)T_2}-\frac{3\alpha^2(n-n_0)}{\alpha^3+1}+\sum_{i=1}^{n}(1-I_{(X=0)}(x_i))\frac{3\alpha^2\Gamma(\delta)\Gamma(x_i+1)(1+\theta)^{\delta-1}}{T_6},$$

and

$$\frac{\partial \ell(\phi,\theta,\alpha,\delta|x)}{\partial \delta} =$$

$$\frac{n_0(1-\phi)\theta\left(T_8-T_9\right)}{(1+\theta)^{\delta}T_2}-(n-n_0)(\psi(\delta)+\log(1+\theta))+\sum_{i=1}^{n}(1-I_{(X=0)}(x_i))\frac{T_{10}+T_{11}}{T_6},$$

where

$$T_1 = n_0((\alpha^3+1)(1+\theta)^{\delta}-\theta((1+\theta)^{\delta-1}\alpha^3+\theta^{\delta-1})),$$

$$T_2 = \phi(\alpha^3+1)(1+\theta)^{\delta}+(1-\phi)\theta((1+\theta)^{\delta-1}\alpha^3+\theta^{\delta-1}),$$

$$T_3 = (1+\theta)^{\delta}(\alpha^3(\theta(\delta-1)(1+\theta)^{\delta-2}+(1+\theta)^{\delta-1})+\delta\theta^{\delta-1}),$$

$$T_4 = \theta((1+\theta)^{\delta-1}\alpha^3+\theta^{\delta-1})\delta(1+\theta)^{\delta-1},$$

$$T_5 = \Gamma(\delta)\Gamma(x_i+1)\alpha^3(\delta-1)(1+\theta)^{\delta-2}+(\delta-1)\theta^{\delta-2}\Gamma(x_i+\delta),$$

$$T_6 = \Gamma(\delta)\Gamma(x_i+1)\alpha^3(1+\theta)^{\delta-1}+\theta^{\delta-1}\Gamma(x_i+\delta),$$

$$T_7 = n_0(1-\phi)\theta((\alpha^3+1)(3\alpha^2(1+\theta)^{\delta-1})-3\alpha^2((1+\theta)^{\delta-1}\alpha^3+\theta^{\delta-1})),$$

$$T_8 = (1+\theta)^{\delta}(\alpha^3(1+\theta)^{\delta-1}\log(1+\theta)+\theta^{\delta-1}\log(\theta)),$$

$$T_9 = ((1+\theta)^{\delta-1}\alpha^3+\theta^{\delta-1})(1+\theta)^{\delta}\log(1+\theta),$$

$$T_{10} = \Gamma(x_i+1)\alpha^3(\Gamma(\delta)(1+\theta)^{\delta-1}\log(1+\theta)+(1+\theta)^{\delta-1}\Gamma(\delta)\psi(\delta)),$$

and

$$T_{11} = \Gamma(x_i+\delta)\theta^{\delta-1}\log(\theta)+\theta^{\delta-1}\Gamma(x_i+\delta)\psi(x_i+\delta).$$

By setting the score functions equal to zero and solving the system of non-linear equations, the MLEs of ϕ, θ, α, and δ abbreviated as $\hat{\phi}_{MLE}$, $\hat{\theta}_{MLE}$, $\hat{\alpha}_{MLE}$, and $\hat{\delta}_{MLE}$ can be derived. The system of non-linear equations with respect to the parameters can be solved by the Newton Raphson method. Here, the solutions of the parameter estimates will be obtained by using the *optim* function in the R package *stats*.

The asymptotic confidence intervals for the parameters ϕ, θ, α, and δ are derived by the asymptotic theory. The estimates are asymptotic four-variate normal with mean $(\phi,\ \theta,\ \alpha,\ \delta)$ and the observed information matrix is

$$I(\phi,\theta,\alpha,\delta)=\begin{pmatrix} -\frac{\partial^2\ell}{\partial\phi^2} & -\frac{\partial^2\ell}{\partial\phi\partial\theta} & -\frac{\partial^2\ell}{\partial\phi\partial\alpha} & -\frac{\partial^2\ell}{\partial\phi\partial\delta} \\ -\frac{\partial^2\ell}{\partial\theta\partial\phi} & -\frac{\partial^2\ell}{\partial\theta^2} & -\frac{\partial^2\ell}{\partial\theta\partial\alpha} & -\frac{\partial^2\ell}{\partial\theta\partial\delta} \\ -\frac{\partial^2\ell}{\partial\alpha\partial\phi} & -\frac{\partial^2\ell}{\partial\alpha\partial\theta} & -\frac{\partial^2\ell}{\partial\alpha^2} & -\frac{\partial^2\ell}{\partial\alpha\partial\delta} \\ -\frac{\partial^2\ell}{\partial\delta\partial\phi} & -\frac{\partial^2\ell}{\partial\delta\partial\theta} & -\frac{\partial^2\ell}{\partial\delta\partial\alpha} & -\frac{\partial^2\ell}{\partial\delta^2} \end{pmatrix}$$

at $\phi=\hat{\phi}_{MLE},\theta=\hat{\theta}_{MLE},\alpha=\hat{\alpha}_{MLE}$, and $\delta=\hat{\delta}_{MLE}$.

That is $(\hat{\phi}_{MLE},\hat{\theta}_{MLE},\hat{\alpha}_{MLE},\hat{\delta}_{MLE})\sim N_4((\phi,\theta,\alpha,\delta),I^{-1}(\phi,\theta,\alpha,\delta))$. Since the mathematical expressions of the second order partial derivatives of the log-likelihood function are very long, we do not present the elements of the observed information matrix, $I(\phi,\theta,\alpha,\delta)$.

Therefore, $(1-a)100\%$ confidence intervals for the parameters ϕ,θ,α, and δ are given by

$$\hat{\phi}_{MLE}\pm z_{a/2}\sqrt{Var(\hat{\phi}_{MLE})},\qquad \hat{\phi}_{MLE}\pm z_{a/2}\sqrt{Var(\hat{\phi}_{MLE})},$$

$$\hat{\alpha}_{MLE}\pm z_{a/2}\sqrt{Var(\hat{\alpha}_{MLE})},\qquad \hat{\delta}_{MLE}\pm z_{a/2}\sqrt{Var(\hat{\delta}_{MLE})},$$

where the $Var(\hat{\phi}_{MLE}),Var(\hat{\theta}_{MLE}),Var(\hat{\alpha}_{MLE})$, and $Var(\hat{\delta}_{MLE})$ are the variance of $\hat{\phi}_{MLE}$, $\hat{\theta}_{MLE},\hat{\alpha}_{MLE}$, and $\hat{\delta}_{MLE}$, respectively. They can be derived by the diagonal elements of $I^{-1}(\phi,\theta,\alpha,\delta)$ and $z_{a/2}$ is the critical value at a level of significance.

6. Monte Carlo simulation study and real-world application

6.1. Monte Carlo simulation study

Here, we evaluate the performance of the MLEs $(\hat{\phi}_{MLE},\hat{\theta}_{MLE},\hat{\alpha}_{MLE}$, and $\hat{\delta}_{MLE})$ and MMEs $(\hat{\phi}_{MME},\hat{\theta}_{MME},\hat{\alpha}_{MME}$, and $\hat{\delta}_{MME})$ with respect to the sample size n by designing a simulation study. We consider sample sizes of 60, 100, 200, and 300, and the simulation is repeated 1000 times. The simulation study is designed as follows:

(i) Simulate 1000 samples of size n.

(ii) Compute the MLEs and MMEs for the 1000 samples, say $(\hat{\phi}_i,\hat{\theta}_i,\hat{\alpha}_i,\hat{\delta}_i),\ \ i=1,2,...,$ 1000.

(iii) Compute the average MLEs, MMEs, biases, and mean square errors (MSEs) by using the following equations:

$$\hat{S}(n)=\tfrac{1}{1000}\textstyle\sum_{i=1}^{1000}\hat{S}_i,\qquad bias_S(n)=\tfrac{1}{1000}\textstyle\sum_{i=1}^{1000}(\hat{S}_i-S),$$

$$MSE_S(n) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{S}_i - S)^2, \text{ for } S = \phi, \theta, \alpha, \delta, \text{ and } n = 60, 100, 200, 300.$$

Tables A1 and A2 present the performance of the MLEs and MMEs of ϕ, θ, α, and δ for different values of ϕ which are -0.4,-0.2,0.2, and 0.4. Here, the population values of θ, α, and δ are 0.10,0.50, and 2.50, respectively. Note that here the average MSEs are presented in parentheses in both tables. We can observe that in both estimation methods, the biases and MSEs decrease as n increases for all parameters in all given situations. This implies that the maximum likelihood estimation method and the method of moments estimation verify the asymptotic property for all given parameter estimates. Further, when comparing the performance of the MLEs and MMEs based on the estimators' biases and MSEs, it is clear that the maximum likelihood estimation method is better than the method of moment estimation.

6.2. Real-world application

This subsection is devoted to show the applicability of the ZMPMQL distribution over the negative binomial (NB), Zero-modified NB (ZMNB), Poisson-Lindley (PL), Zero-modifi -ed PL (ZMPL), and PMQL distributions. The best-fitted distribution is selected based on the negative log-likelihood $(-2logL)$, Akaike information criterion (AIC), and chi-square goodness of fit test statistic (χ^2). Further, the maximum likelihood estimation method is used to estimate the unknown parameters of the distributions.

The example data set presents the number of units of consumers' goods purchased by households over 26 weeks (Lindsey, 1995). The proportion of the zeros in the data set is 80.60%, which indicates that there exists inflation of zeros. The sample dispersion index of 4.761 shows that extreme over-dispersion is present. Further, the skewness and excess kurtosis of the data are 3.895 and 16.306, respectively. These results imply that the distribution of the data set is highly positively skewed having a very long right tail. Table 1 summarizes the comparability of the ZMPMQL distribution with the NB, ZINB, PL, ZMPL, and PMQL distributions. Based on the results, the ZMPMQL distribution having AIC=3419.95, $\chi^2 = 12.22$, p-value=0.06 gives a better fit than the other distributions. Further, when we compare the PMQL distribution and the ZMPMQL distribution, the likelihood ratio (LR) test statistic for the hypothesis testing $H_0 : \phi = 0$ vs $H_a : \phi \neq 0$ for this data set is 50.50, and it is greater than $\chi^2_{1,0.05} = 3.84$. Then, the parameter estimate $\hat{\phi}$ is significantly different from zero.

Table 1. Units of consumers good

Counts	Observed	Expected					
		NB	ZINB	PL	ZMPL	PMQL	ZMPMQL
0	1612	1239.62	1617.92	1272.56	1611.94	1585.72	1611.78
1	164	240.36	146.26	470.84	122.00	207.56	165.09
2	71	130.05	80.89	168.06	88.57	45.30	74.66
3	47	85.41	50.02	58.50	61.44	32.99	39.88
4	28	61.04	32.56	19.99	41.33	33.90	26.36
5	17	45.74	21.83	6.73	27.20	30.93	20.24
6	12	35.33	14.93	2.24	17.59	24.36	16.34
7	12	27.88	10.34	0.74	11.23	16.90	13.03
8	5	22.35	7.24	0.24	7.10	10.53	10.03
9	7	18.12	5.11	0.07	4.45	5.97	7.41
10	25	94.10	12.90	0.03	7.15	5.84	15.18
Total	2000	2000	2000	2000	2000	2000	2000
MLE		$\hat{\beta}=0.21$ (0.21) $\hat{\alpha}=0.12$ (0.01)	$\hat{\phi}=0.33$ (0.30) $\hat{\beta}=0.23$ (0.04) $\hat{\alpha}=0.20$ (0.13)	$\hat{\theta}=2.33$ (0.07)	$\hat{\phi}=0.73$ (0.01) $\hat{\theta}=0.75$ (0.04)	$\hat{\theta}=7.00$ (0.67) $\hat{\alpha}=2.12$ (0.07) $\hat{\delta}=32.88$ (2.81)	$\hat{\phi}=-0.61$ (0.03) $\hat{\theta}=1.45$ (0.25) $\hat{\alpha}=1.78$ (0.11) $\hat{\delta}=8.51$ (1.32)
χ^2		311.64	18.86	17992.93	77.89	110.99	12.22
p-value		0.00	0.01	0.00	0.00	0.00	0.06
$-2logL$		3800.90	3427.16	4216.21	3455.33	3462.45	**3411.95**
AIC		3804.90	3433.16	4218.21	3459.33	3468.45	**3419.95**

7. Conclusion

In this article, the zero-modified Poisson-Modification of Quasi Lindley distribution was introduced to model the over-dispersed count data having zero inflation/deflation. We derived some of its structural properties. Further, in order to estimate its unknown parameters, we derived its log-likelihood function and score functions. We showed that the maximum likelihood estimation method is a suitable method to estimate its unknown parameters via a Monte Carlo simulation study. The usefulness of the introduced distribution was illustrated by fitting it to a real-world data set. The results revealed its superiority over some other existing mixed Poisson and zero-modified mixed Poisson distributions.

Acknowledgements

We thank the Postgraduate Institute of Science, University of Peradeniya, Sri Lanka for providing all facilities to complete this research.

References

Albrecht, P., (1984). Laplace Transforms, Mellin Transforms and Mixed Poisson Processes. *Scandinavian Actuarial Journal*, 11, pp. 58–64.

Da Silva, W. B., Ribeiro, A. M. T., Conceicao, K. S., Andrade, M. G., Neto, F. L., (2018). On Zero-Modified Poisson-Sujatha Distribution to Model Overdispersed Count Data. *Austrian Journal of Statistics*, 47(3), pp. 1–19.

Ghitany, M.E., Atieh, B., Nadarajah, S., (2008). Zero-truncated Poisson-Lindley Distribution and Its Application. *Mathematics and Computers in Simulation*, 79, pp. 279–287. https://doi.org/10.1016/j.matcom.2007.11.021.

Greenwood, M., Yule, G. U., (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, 83, pp. 255–279. doi: 10.2307/2341080.

Grine, R., Zeghdoudi, H., (2017). On Poisson quasi-Lindley distribution and its applications. *J. of Modern Applied Statistical Methods*, 16, pp. 403–417.

Irwin, J., (1975). The Generalized Waring Distribution Parts I, II, III. *Journal of the Royal Statistical Society A*, 138, 18–31 (Part I), pp. 204–227 (Part II), pp. 374–384 (Part III). doi:10.2307/2345247.

Lindsey, J. K., (1995). Modelling Frequency and Count Data, Oxford science publications, Clarendon Press, UK.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sankaran, M., (1970). The discrete Poisson-Lindley distribution. *Biometrics*, 26, pp. 145–149. doi:10.2307/2529053.

Shanker, R., (2016b). Sujatha Distribution and Its Applications. *Statistics in Transition New series*, 17, pp. 1–20.

Shanker, R., (2016c). The Discrete Poisson-Sujatha Distribution. *International Journal of Probability and Statistics*, 5, pp. 1–9.

Tharshan, R., Wijekoon, P., (2021). A modification of the Quasi Lindley distribution, *Open Journal of Statistics*, 11, 369–392. doi:10.4236/ojs.2021.113022.

Tharshan, R., Wijekoon, P., (2022). A New Mixed Poisson Distribution for Over-dispersed Count Data: Theory and Applications, *Reliability: Theory and Applications*, 17, pp. 3351, doi: https://doi.org/10.24412/1932-2321-2022-167-33-51.

Xavier, D., Santos-Neto, M., Bourguignon, M., Tomazella, V., (2018). Zero-Modified Poisson-Lindley distribution with applications in zero-inflated and zero-deflated count data, *arXiv preprint*. arXiv:1712.04088.

Appendix

Table A1. Performance of MLEs for ZMPMQL($\phi, \theta = 0.10, \alpha = 0.50, \delta = 2.25$)

	$n = 60$		$n = 100$		$n = 200$		$n = 300$	
	MLE	Bias (MSE)	MLE	Bias (MSE)	MLE	Bias (MSE)	MLE	Bias (MSE)
$\phi = -0.40$								
ϕ	-0.0313	0.3686 (0.1402)	-0.0770	0.3229 (0.1092)	-0.0950	0.3049 (0.1043)	-0.1328	0.2671 (0.0809)
θ	0.0373	-0.0626 (0.0040)	0.0629	-0.0370 (0.0030)	0.0865	-0.0134 (0.0026)	0.0916	-0.0083 (0.0019)
α	0.5429	0.0429 (0.0933)	0.5359	0.0359 (0.0710)	0.4820	-0.0179 (0.0379)	0.4903	-0.0096 (0.0333)
δ	0.9333	-1.5666 (2.6203)	1.1108	-1.3891 (2.3570)	1.1776	-1.3223 (2.0278)	1.2210	-1.2789 (1.8462)
$\phi = -0.20$								
ϕ	-0.0566	0.1433 (0.0242)	-0.0818	0.1181 (0.0291)	-0.0975	0.1024 (0.0118)	-0.1129	0.0870 (0.0097)
θ	0.0451	-0.0548 (0.0035)	0.0848	-0.0151 (0.0025)	0.0912	-0.0087 (0.0018)	0.0961	-0.0038 (0.0009)
α	0.5629	0.0629 (0.1958)	0.5900	0.0900 (0.1790)	0.4547	-0.0452 (0.0805)	0.4701	-0.0298 (0.0412)
δ	1.0944	-1.4055 (2.2541)	1.2009	-1.2990 (1.9939)	1.3474	-1.1525 (1.6011)	1.4268	-1.0731 (1.2742)
$\phi = 0.20$								
ϕ	0.2582	0.0582 (0.0133)	0.2427	0.0427 (0.0053)	0.2257	0.0257 (0.0023)	0.2119	0.0119 (0.0004)
θ	0.1473	0.0473 (0.0195)	0.1379	0.0379 (0.0061)	0.1193	0.0193 (0.0022)	0.1063	0.0063 (0.0002)
α	0.4138	-0.0861 (0.0953)	0.4308	-0.0691 (0.0701)	0.4495	-0.0504 (0.0565)	0.4891	-0.0108 (0.0314)
δ	2.9704	0.4704 (1.2737)	2.7098	0.2098 (0.7320)	2.6435	0.1435 (0.3341)	2.5623	0.0623 (0.1134)
$\phi = 0.40$								
ϕ	0.4335	0.0335 (0.0105)	0.4213	0.0213 (0.0027)	0.4174	0.0174 (0.0016)	0.4099	0.0099 (0.0008)
θ	0.1447	0.0447 (0.0121)	0.1280	0.0280 (0.0022)	0.1177	0.0177 (0.0013)	0.1051	0.0051 (0.0002)
α	0.4109	-0.0890 (0.0988)	0.4245	-0.0754 (0.0722)	0.4309	-0.0690 (0.0650)	0.4467	-0.0532 (0.0424)
δ	3.0761	0.5761 (1.4439)	2.7313	0.2313 (0.7488)	2.6573	0.1573 (0.4785)	2.5932	0.0932 (0.2135)

Table A2. Performance of MMEs for ZMPMQL($\phi, \theta = 0.10, \alpha = 0.50, \delta = 2.25$)

	$n = 60$		$n = 100$		$n = 200$		$n = 300$	
	MME	Bias (MSE)	MME	Bias (MSE)	MME	Bias (MSE)	MME	Bias (MSE)
$\phi = -0.40$								
ϕ	-0.0223	0.3776 (0.1427)	-0.0547	0.3452 (0.1199)	-0.0769	0.3230 (0.1186)	-0.1062	0.2937 (0.1034)
θ	0.1832	0.0832 (0.0103)	0.1649	0.0649 (0.0076)	0.1376	0.0376 (0.0044)	0.1335	0.0335 (0.0028)
α	0.7600	0.2600 (0.1022)	0.6028	0.1028 (0.0843)	0.5921	0.0921 (0.0642)	0.5777	0.0777 (0.0440)
δ	4.5414	2.2914 (6.7310)	4.1311	1.8811 (4.2928)	3.6185	1.3685 (2.3961)	3.8331	1.3331 (2.2493)
$\phi = -0.20$								
ϕ	-0.3827	-0.1827 (0.0441)	-0.3612	-0.1612 (0.0301)	-0.3493	-0.1493 (0.0231)	-0.3243	-0.1243 (0.0173)
θ	0.1967	0.0967 (0.0131)	0.1650	0.0650 (0.0082)	0.1492	0.0492 (0.0034)	0.1407	0.0407 (0.0024)
α	0.9864	0.4864 (0.2694)	0.9691	0.4691 (0.3253)	0.7889	0.2889 (0.1238)	0.7324	0.2324 (0.0729)
δ	5.7357	3.4857 (15.7459)	5.1516	2.9016 (12.4246)	3.9256	1.6756 (3.7138)	3.6355	1.3855 (2.6906)
$\phi = 0.20$								
ϕ	-0.0349	-0.2349 (0.0706)	0.0166	-0.1833 (0.0530)	0.0700	-0.1299 (0.0216)	0.0993	-0.1006 (0.0139)
θ	0.2253	0.1253 (0.0246)	0.1759	0.0759 (0.0101)	0.1562	0.0562 (0.0047)	0.1459	0.0459 (0.0030)
α	1.0917	0.5917 (0.3969)	1.0101	0.5101 (0.3522)	1.0036	0.5036 (0.2782)	0.8582	0.3582 (0.1620)
δ	6.6309	4.3809 (27.6692)	5.5530	3.3030 (15.6071)	4.3311	2.0811 (5.8163)	4.1850	1.9350 (4.5527)
$\phi = 0.40$								
ϕ	0.2199	-0.1800 (0.0465)	0.2350	-0.1649 (0.0406)	0.3151	-0.0848 (0.0114)	0.3576	-0.0423 (0.0053)
θ	0.2581	0.1581 (0.0398)	0.1962	0.0962 (0.0141)	0.1670	0.0670 (0.0067)	0.1553	0.0553 (0.0048)
α	1.0907	0.5907 (0.4056)	1.0502	0.5502 (0.3492)	0.9728	0.4728 (0.2560)	0.8745	0.3745 (0.1674)
δ	7.5974	5.3474 (42.3990)	6.2256	3.9756 (20.4941)	4.6281	2.3781 (7.9752)	4.2158	1.9658 (5.2737)

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 129–148, DOI 10.2478/stattrans-2022-0046
Received – 14.04.2021; accepted – 09.06.2022

Optimal allocation for equal probability two-stage design

Wilford Molefe[1]

ABSTRACT

This paper develops optimal designs when it is not feasible for every cluster to be represented in a sample as in stratified design, by assuming equal probability two-stage sampling where clusters are small areas. The paper develops allocation methods for two-stage sample surveys where small-area estimates are a priority. We seek efficient allocations where the aim is to minimize the linear combination of the mean squared errors of composite small area estimators and of an estimator of the overall mean. We suggest some alternative allocations with a view to minimizing the same objective. Several alternatives, including the area-only stratified design, are found to perform nearly as well as the optimal allocation but with better practical properties. Designs are evaluated numerically using Switzerland canton data as well as Botswana administrative districts data.

Key words: sample designs, optimal allocation, composite estimation, mean squared error, two-stage sampling, simple random sampling without replacement

1. Introduction

In many situations it is not feasible for every small area to be represented in a sample. In practice, it is not possible to anticipate and plan for all possible areas (or domains) and uses of survey data as "the client will always require more than is specified at the design stage" (Fuller, 1999).

Longford (2006), Molefe (2011), Molefe and Clark (2015) and Molefe, Shangodoyin and Clark (2015) derive optimal allocations for stratified sampling, which minimize weighted sums of the MSEs of small area estimates and a grand mean estimate. In Longford (2006), the MSEs are design-based (that is, based on repeated probability sampling from a fixed population without reference to a model), and in Molefe (2011), Molefe and Clark (2015) and Molefe, Shangodoyin and Clark (2015) anticipated MSEs are used. In all the references above, stratified simple random sampling without replacement is assumed, where strata are small areas. All find that the optimal design could sometimes have zero sample size for the smallest areas. The authors establish numerically that simpler designs with positive stratum sample sizes give near optimal anticipated MSEs. Power allocation (Bankier, 1988) with stratum sample sizes proportional to a numerically optimized exponent of the stratum population performs particularly well.

In this paper we consider the case of equal probability two-stage sampling design where small areas are clusters or primary sampling units (PSUs). Two-stage sampling with equal probabilities of selection for all clusters (at least within broad regions) are used in many large scale sample surveys including the Australian and New Zealand labour force surveys.

[1]University of Botswana, Botswana. E-mail: molefewb@ub.ac.bw. ORCID: https://orcid.org/0000-0001-7674-2244.

It will be assumed that a sample of small areas is selected by SRSWOR, followed by a sample of the second stage units (units) from each selected small area, also by SRSWOR. There are several possible reasons for this approach. There may be a list of the small areas in the population, but not of the population units. Two stage surveys are also useful so that the sample can be made more geographically clustered, which often reduces enumeration costs (Cochran, Chapter 10 1977).

In optimizing this sampling design for small area estimation where small areas are clusters, the fundamental question is how to choose the number of clusters (m) and the number of subunits, referred to as just units (n_d) per cluster subject to fixed cost. One approach is to choose m and n_d to optimize some criteria subject to a cost constraint based on some model for cost.

We adopt the criterion of the weighted sum of the MSE for the small areas in-sample and the MSE of the estimator for the small areas out-of-sample.

A question within the above setup is when it is appropriate to have some sample in every small area. This would only be feasible when there are a relatively small number of small areas (M), or a very large survey budget, and would usually mean that the number of units $\{n_d\}$ in each small area would be fairly small. In this case the design will be a special case of stratified design considered by Longford (2006), Molefe (2011), Molefe and Clark (2015) and Molefe, Shangodoyin and Clark (2015).

In practice, it is not always feasible for every small area to be represented in the sample. This is clear from the fact that zero stratum sample sizes sometimes arise in Longford (2006), Molefe (2011) and Molefe and Clark (2015). In this paper, we explicitly allow for the sampling of small areas. It is assumed that a two-stage design is used, where clusters are small areas. A cluster d may be selected with equal probability $\pi_d = \frac{m}{M}$ and a different sample size n_d to be selected from each selected cluster. In Section 2 we state a two-level model and the resulting anticipated MSE of small area estimates. An objective function which is a linear combination of anticipated MSEs is defined. A linear cost model consisting of per-cluster and per-unit costs is assumed. The aim is then to minimize the objective function with respect to m and n_d subject to fixed expected cost for the survey. In Section 3 we develop an optimal analytical solution when only small area estimates are a priority. Section 4 suggests sensible but ad-hoc designs that include equal allocation, proportional allocation, classical optimal allocation and a combined design made up of the proportional allocation and the classical optimal design. Section 5 is a numerical study based on the Switzerland canton population sizes used by Longford (2006). Section 6 contains conclusions.

2. Methods

From a population of M small areas (clusters) indexed by d, denoted by U^1, a first stage sample of m small areas selected by SRSWOR is denoted s^1. In the second stage of the selection a sample of size n_d elements selected by SRSWOR from area d is denoted by s_d. The set of N_d population units in a particular cluster d is denoted by U_d. Let the sampling variances be $v_d = var_p(\bar{y}_d)$ and $v = var_p(\bar{y})$ respectively for the small area mean estimator and overall mean estimator. The composite estimator is denoted $\tilde{y}_d^{\mathscr{C}}[\phi_{d(opt)}]$.

Let Y_j be the value of the characteristic of interest for the jth unit in the population. The small area population mean is $\bar{Y}_d$ and the national mean is $\bar{Y}$. Auxiliary variables x_j are assumed to be available for the full population $j \in U^1$.

The following two-level linear mixed model ξ will be assumed:

$$\left.\begin{array}{l} Y_j = \beta^T x_j + u_d + \varepsilon_j \\ E_\xi[u_d] = E_\xi[\varepsilon_j] = 0 \\ var_\xi[u_d] = \sigma^2_{ud} \\ var_\xi[\varepsilon_j] = \sigma^2_{\varepsilon d} \end{array}\right\} \tag{1}$$

for $d \in U^1$ and $j \in U_d$ with mutual independence of u_d and ε_j for $d \in U^1$ and $j \in U_d$. This implies $var_\xi[Y_j] = \sigma^2_{ud} + \sigma^2_{\varepsilon d} = \sigma^2$ for all $j \in U$, and that the covariance $cov_\xi[Y_i, Y_j] = \rho_d \sigma^2_d$ for units $i \neq j$ in the same small area and 0 for units from different small areas, where $\rho_d = \sigma^2_{ud}/\sigma^2$. For simplicity it will be assumed that $\rho_d = \rho$.

Following Molefe and Clark (2015), we assume a small-area composite estimator which is a weighted mean of an approximately design unbiased estimator

$$\bar{y}_{dr} \quad = \quad \bar{y}_d + \hat{\beta}^T (\bar{X}_d - \bar{x}_d)$$

recommended by Hidiroglou and Patak (2004) for small domains, and a model-based synthetic estimator $\hat{\bar{Y}}_{d(syn)} = \hat{\beta}^T \bar{X}_d$.

The composite estimator which approximately minimizes the anticipated MSE is

$$\tilde{\bar{y}}^{\mathscr{C}}_d[\phi_{d(opt)}] \quad = \quad (1-\phi_d)\bar{y}_{dr} + \phi_d \hat{\bar{Y}}_{d(syn)} = \hat{\beta}^T \bar{X}_d + (1-\phi_d)\left(\bar{y}_d - \hat{\beta}^T \bar{x}_d\right)$$

where $\phi_{d(opt)} = (1-\rho)\left[1+(n^*_d-1)\rho\right]^{-1}$, assuming that n, N_d and M are all large (Molefe and Clark, 2015). Under the same assumptions, the approximate anticipated MSE of the optimal composite estimator of $\bar{Y}_d$ conditional on n^*_d is

$$\begin{aligned} & E_\xi MSE_p\left(\tilde{\bar{y}}^{\mathscr{C}}_d\left[\phi_{d(opt)}\right]; \bar{Y}_d | n^*_d\right) \\ & \approx \quad \left\{n^*_d \rho\left[1+(n^*_d-1)\rho\right]^{-1}\right\}^2 (n^*_d)^{-1} \sigma^2 (1-\rho) + \left\{(1-\rho)\left[1+(n^*_d-1)\rho\right]^{-1}\right\}^2 \sigma^2 \rho \\ & = \quad \sigma^2 \rho (1-\rho)/\left[1+(n^*_d-1)\rho\right] \end{aligned} \tag{2}$$

See Molefe (2011) for the derivation.

Small areas with no sample would have a direct estimate of zero. For these, a synthetic estimator is used. An indirect estimator, $\tilde{\bar{y}}^{\mathscr{C}}_d = \bar{y}$ is proposed, if cluster $d \notin s^1$. The MSE of $\bar{y}$ is given by $MSE_p\left(\bar{y}; \bar{Y}_d\right) = v + B^2_d$, where B_d is the design bias of using $\bar{y}$ to estimate $\bar{Y}_d$.

The population level mean estimator $\bar{y}$ and area mean $\bar{y}_d$ are assumed to be unbiased for $\bar{Y}$ and $\bar{Y}_d$ respectively. The design variance of the synthetic estimator will be small relative to the design variance of the direct estimator because it depends only on the precision of direct estimators at a larger area level. If the number of small areas in the sample is large, v is negligible and can be ignored. Therefore, we approximate $MSE_p\left(\bar{y}; \bar{Y}_d\right)$ by B^2_d.

For optimal allocation of sample sizes of clusters and subunits, we search for the area-

level sampling design that minimizes the weighted expected value of the sum of the sampling variances (MSEs) for a combination of small area composite estimates for clusters in-sample and out-of-sample and an overall estimator of the mean given by

$$F = \sum_{d\in U^1} \pi_d N_d^q AMSE_d\left\{\tilde{y}_d^{\mathscr{C}}\left[\phi_{d(opt)}\right];\bar{Y}_d\right\} + \sum_{d\in U^1}(1-\pi_d)N_d^q AMSE_d\left[\bar{y};\bar{Y}_d\right] \quad (3)$$

where the first component in (3) is due to the m clusters in-sample and the second component is due to the remaining $(M-m)$ clusters. The small-area population sizes N_d are weights, that is, N_d^q for $0 \leq q \leq 2$, where for $q=0$, inference is equally important for every area. With increasing q, relatively greater importance is ascribed to more populous areas, with $q=2$ corresponding to proportional allocation. $AMSE_d$ is the model assisted mean squared error, that is ξMSE_d.

We can then write the model expectation of the criterion function to be minimized, ignoring the goal of national estimation, as

$$F \approx \frac{m}{M}\sum_{d\in U^1} N_d^q \frac{\sigma^2\rho(1-\rho)}{[1+(n_d-1)\rho]} + \left(1-\frac{m}{M}\right)\sigma^2\rho\sum_{d\in U^1} N_d^q \quad (4)$$

2.1. Cost Models and Cost Estimates

In a two stage sampling scheme the sampling variance of the estimate of the overall population mean ($\bar{y}$) is minimized (for fixed sample size) when $\bar{n}=1$ since this is when the sample is most spread out. However, costs will be minimized when as few first stage units as possible are selected. Hence, some compromise between these two extremes has to be chosen and this is the optimal design problem in multistage sampling. As always costs and variances are pulling in opposite directions and the task of optimal design is to choose the optimal balance of these.

In a two-stage sample, several types of costs can be distinguished (Hansen, 1953; Cochran, 1977):

(a) Overhead costs - costs associated with planning, administration, setting up processing systems, etc. These costs do not depend on the sample sizes used at either stage;

(b) Costs associated with the selection of clusters - these arise from drawing maps, listing units within selected primary stage units, travel between selected primary stage units. These costs increase as the number of clusters selected increases;

(c) Costs associated with the selection of secondary stage units - these mainly arise from the enumeration of selected population units, e.g. the cost of time spent in interviewing people and the cost of processing an individual questionnaire. These costs increase as the number of selected units increases.

Linear cost models are commonly used by official statistics agencies (Hansen, 1953; Sukhatme, 1954; Cochran, 1977; Foreman, 1991; Clark, 2007). A linear cost model is often adequate for sample design, even though it cannot perfectly capture the real cost structure.

A simple cost model for a two-stage sample is given by

$$C_F = c_0 + c_1 m + c_2 \frac{m}{M} \sum_{d \in U^1} n_d$$

where m equals the number of primary sampling units (clusters) in the sample; n_d is the number of secondary sampling units (units), for example, households, in the sample from cluster d; the coefficient c_0 is the fixed costs of conducting the survey, independent of the number of sample clusters and subunits per cluster, including costs for survey planning, development of the survey design, preparatory work, survey management and data processing, analysis and presentation of results; the coefficient c_1 is the average cost of adding a cluster to the sample, consisting of travel by interviewers and supervisors between clusters and home base or between clusters (fuel costs, driver salaries) and interviewer salaries, including the cost of obtaining maps and other material for the cluster, the cost of establishing the survey in the local area, entailing, for example, meeting with and obtaining permission from local authorities, and the cost of listing and sampling of dwelling units within the cluster; the coefficient c_2 is the average cost of including an extra household in the sample, including the costs for locating, contacting and interviewing a household, where the costs consist of interviewer and supervisor salaries and allowance, and also costs for travel by interviewers and supervisors within clusters (Pettersson and Sisouphanthong, 2005).

Costs for the different components of a survey differ from survey to survey and from country to country. The survey manager often has a good idea of the time required for specific survey operations based on information from previous surveys of a similar nature. Experiences from prior surveys (or from pilot surveys) could often be used for reasonable estimates of time per household required for locating and interviewing the household. In these cases, reasonable estimates of c_2 could be compiled.

Computing a reasonable estimate of c_1 is often difficult because it involves determining the effect of additional interviewer travel when a cluster is added to the sample. The travel depends on the size of the area being covered, the number of clusters assigned to each interviewer, and the travel pattern of the interviewers. The travel includes between-cluster and within-cluster travel during a data collection trip.

Cost modelling is mainly used for budgetary purposes and for finding an efficient sample design. In this thesis, our interest is mainly in the use of cost models to find an efficient design. We do not consider the fixed costs (c_0) in trying to work out an efficient design; we only consider the fieldwork costs. The total sampling cost function has two components; the first part depends on how many small areas, $c_1 m$, and the other on the total number of units sampled, namely $c_2 \sum_{d \in s^1} n_d$. The second component will, however, vary from sample to sample of the clusters.

Therefore, the expected total sampling cost function will then be given by

$$C_E = c_1 m + c_2 \frac{m}{M} \sum_{d \in U^1} n_d \tag{5}$$

The aim is to minimise F with respect to n_d and m subject to a cost constraint $C_E = C_F$.

3. Area-Only Simple Two-Stage Optimal Design

The expected criterion function (4), eliminating the common σ^2 and ρ terms, reduces to

$$F \approx -\frac{m}{M}\sum_{d\in U^1}\frac{N_d^q n_d \rho}{1+(n_d-1)\rho} \tag{6}$$

plus constant terms which do not depend on m or n_d.

We minimize (6) subject to the cost constraint (5). The Lagrangian is:

$$L = F+\lambda\left(c_1 m + c_2\frac{m}{M}\sum_{d\in U^1} n_d - C_F\right) \tag{7}$$

To obtain an optimal number of clusters and subunits to take into the sample, we take partial derivatives of (7) with respect to n_d, λ and m.

We use the partial derivatives to derive the optimal design by firstly deriving $\bar{n}_{opt.}$, the optimal average within-cluster sample size. This result will then be used to derive the optimal values of n_d.

We use $\frac{\partial L}{\partial n_d}=0$ to obtain the optimal value for n_d as follows:

$$n_d = N_d^{\frac{q}{2}}\sqrt{(1-\rho)/(\lambda c_2\rho)}-(1-\rho)/\rho \tag{8}$$

This solution for n_d given implies that the average within-cluster sample size is

$$\bar{n} = \bar{N}_d^{\frac{q}{2}}\sqrt{(1-\rho)/(\lambda c_2\rho)}-(1-\rho)/\rho$$

Therefore, we can write

$$\sqrt{(1-\rho)/(\lambda c_2\rho)} = \left(\bar{N}_d^{\frac{q}{2}}\right)^{-1}\{\bar{n}+(1-\rho)/\rho\}$$

Then, the optimal cluster sample sizes can be expressed as

$$n_d = N_d^{\frac{q}{2}}\left(\bar{N}_d^{\frac{q}{2}}\right)^{-1}\bar{n}+(1-\rho)/\rho\left[N_d^{\frac{q}{2}}\left(\bar{N}_d^{\frac{q}{2}}\right)^{-1}-1\right]$$

We can also substitute for n_d given by (8) in $\frac{\partial L}{\partial \lambda}=0$ to obtain

$$c_1 m + c_2\frac{m}{M}\sum_{d\in U^1}\left(N_d^{\frac{q}{2}}\sqrt{(1-\rho)/(\lambda c_2\rho)}-(1-\rho)/\rho\right) = C_F$$

This simplifies to

$$C_F = \gamma m + \sqrt{\frac{c_2}{\lambda}}\frac{m}{M}\sum_{d\in U^1} N_d^{\frac{q}{2}}\sqrt{(1-\rho)/\rho}$$

where $\gamma = c_1 - c_2(1-\rho)/\rho$.

Similarly, we substitute for n_d in $\frac{\partial L}{\partial m} = 0$ and after simplifying we obtain

$$\frac{1}{M}\sum_{d\in U^1} N_d^q = \frac{1}{M}\sum_{d\in U^1} N_d^{\frac{q}{2}}\sqrt{\lambda c_2(1-\rho)/\rho} + \lambda\left(\gamma + \frac{1}{M}\sqrt{\frac{c_2}{\lambda}}\sum_{d\in U^1} N_d^{\frac{q}{2}}\sqrt{(1-\rho)/\rho}\right)$$

Removing the bracket on the right hand size, we obtain

$$\frac{1}{M}\sum_{d\in U^1} N_d^q = 2\frac{1}{M}\sum_{d\in U^1} N_d^{\frac{q}{2}}\sqrt{\lambda c_2(1-\rho)/\rho} + \lambda\gamma$$

The resulting two simultaneous equations in m and λ are:

$$\frac{m}{M}\sqrt{c_2/\lambda(1-\rho)/\rho}\sum_{d\in U^1} N_d^{\frac{q}{2}} + \gamma m = C_F \tag{9}$$

$$2\sqrt{\lambda c_2(1-\rho)/\rho}\sum_{d\in U^1} N_d^{\frac{q}{2}} + \lambda\gamma M = \sum_{d\in U^1} N_d^q \tag{10}$$

We use (9) to write λ in terms of m as follows:

$$\sqrt{\lambda} = \frac{1}{M\left(C_F/m - \gamma\right)}\sum_{d\in U^1} N_d^{\frac{q}{2}}\sqrt{c_2(1-\rho)/\rho}$$

Substituting for λ in (10) we obtain

$$\sum_{d\in U^1} N_d^q = \frac{2c_2(1-\rho)/\rho\left(\sum_{d\in U^1} N_d^{\frac{q}{2}}\right)^2}{M\left(C_F/m-\gamma\right)} + \frac{c_2(1-\rho)/\rho\left(\sum_{d\in U^1} N_d^{\frac{q}{2}}\right)^2}{M\left(C_F/m-\gamma\right)^2}\times\gamma$$

Cross-multiplying and further simplifying we obtain

$$\begin{aligned} 0 = {} & \gamma\left\{c_2(1-\rho)/\rho\left(\sum_{d\in U^1} N_d^{\frac{q}{2}}\right)^2 + \gamma M\sum_{d\in U^1} N_d^q\right\} + M\left(\frac{C_F}{m}\right)^2\sum_{d\in U^1} N_d^q - \\ & 2\frac{C_F}{m}\left\{c_2(1-\rho)/\rho\left(\sum_{d\in U^1} N_d^{\frac{q}{2}}\right)^2 + \gamma M\sum_{d\in U^1} N_d^q\right\} \end{aligned} \tag{11}$$

which is a quadratic in m^{-1} of the form $am^{-2} + bm^{-1} + c = 0$.

Define $C_{q/2}^2$ the relative population variance of $N_d^{\frac{q}{2}}$ given by

$$C_{q/2}^2 = M^{-1}\sum_{d\in U^1}\left(N_d^{\frac{q}{2}} - \bar{N}^{\frac{q}{2}}\right)^2/\left(\bar{N}^{\frac{q}{2}}\right)^2 \tag{12}$$

Then

$$\frac{M^{-1}\sum_{d\in U^1}(N_d^{\frac{q}{2}})^2}{(M^{-1}\sum_{d\in U^1}N_d^{\frac{q}{2}})^2} = \frac{M^{-1}\sum_{d\in U^1}N_d^q}{(M^{-1}\sum_{d\in U^1}N_d^{\frac{q}{2}})^2} = 1+C_{q/2}^2$$

Hence, we write

$$\sum_{d\in U^1}N_d^q = M^{-1}\left(\sum_{d\in U^1}N_d^{\frac{q}{2}}\right)^2\left(1+C_{q/2}^2\right) \tag{13}$$

We substitute for $\sum_{d\in U^1}N_d^q$ into (11) to obtain a reduced quadratic equation in m^{-1}:

$$0 = \left(\frac{C_F}{m}\right)^2(1+C_{q/2}^2)-2\frac{C_F}{m}\left[c_2(1-\rho)/\rho+\gamma\left(1+C_{q/2}^2\right)\right]+\gamma\left[c_2(1-\rho)/\rho+\gamma\left(1+C_{q/2}^2\right)\right] \tag{14}$$

Define $\bar{n}=E\left[\frac{n}{m}\right]=\frac{1}{M}\sum_{d\in U^1}n_d$. There is a one-to-one relationship between m and $\bar{n}$ because $C_F=c_1m+c_2m\bar{n}$ so that $m=C_F/(c_1+c_2\bar{n})$. Hence finding the optimal m is equivalent to finding $\bar{n}$. Substituting for m^{-1} into (14) we obtain

$$0 = (c_1+c_2\bar{n})^2(1+C_{q/2}^2)-2(c_1+c_2\bar{n})\left[c_2(1-\rho)/\rho+\gamma(1+C_{q/2}^2)\right]+\gamma\left[c_2(1-\rho)/\rho+\gamma(1+C_{q/2}^2)\right]$$

which is a quadratic in $\bar{n}$ of the form $a\bar{n}^2+b\bar{n}+c$.

Therefore, the optimum $\bar{n}$ is:

$$\bar{n}_{opt.} = \frac{-c_2(1-\rho)/\rho C_{q/2}^2 \pm \left[c_1c_2(1-\rho)/\rho+\left\{c_2(1-\rho)/\rho\gamma\right\}C_{q/2}^2\right]^{\frac{1}{2}}}{c_2\left(1+C_{q/2}^2\right)} \tag{15}$$

Of primary interest will be to compare the optimal sample size using composite estimation, $\bar{n}_{opt.}$, with the classical two-stage optimal design given by Hansen, Hurwitz and Madow (1953, page 173 equations 10.1 and 10.2) and Cochran (1977, page 281 equation 10.26) as $\bar{n}_{cl.}=\sqrt{c_1/c_2(1-\rho)/\rho}$ for the purpose of drawing general conclusions on whether the two-stage composite optimal is always more clustered or always less clustered than the standard or classical two-stage cluster optimal.

The classical optimal for the two-stage cluster design $\bar{n}_{cl.}$ coincides with $\bar{n}_{opt.}$ when $q=0$.

It is not obvious whether the two-stage general optimal $\bar{n}_{opt.}$ is larger or smaller than the classical two-stage cluster design optimal $\bar{n}_{cl.}$ when $q>0$. In fact, it is not clear that the stationary point for $\bar{n}_{opt.}$ exists at all. If ρ is small enough, then the contents of the square bracket in (15) will become negative, so that the square root will not exist.

Looking at (15) it appears that $\bar{n}_{opt.}$ will usually be less than in the classical design ($\bar{n}_{cl.}$), because in the square root of the discriminant, the coefficient of $C_{q/2}^2$ is $c_2(1-\rho)/\rho\gamma$. Usually, ρ is 0.05 or less, so that γ/c_2 becomes $\{c_1/c_2 - 19\}$. The cost of including a new PSU in the sample (c_1) will always be higher than the cost of including a new household in a selected PSU (c_2), hence the cost ratio will always be well above 1.0. The higher the cost ratio, the more costly it is to select a new PSU compared with selecting more households in selected PSUs; consequently, we should select more households in already selected PSUs. We assume that $c_1/c_2 < 19$, so the coefficient of $C_{q/2}^2$ is negative. In the term $-b$, the coefficient of $C_{q/2}^2$ is negative and in the denominator the coefficient of $C_{q/2}^2$ is positive. Hence, $\bar{n}_{opt.}$ is (usually) a decreasing function of $C_{q/2}^2$, so that for $C_{q/2}^2 > 0$, $\bar{n}_{opt.}$ will be less than the classical design. A sufficient condition for this is that $\gamma/c_2 < 0$, which would usually be satisfied, unless c_1 or ρ are unusually large. When $C_{q/2}^2 = 0$ as is the case when $N_d = \bar{N}$ the optimal sample size reduces to the standard optimal cluster size so that $\bar{n}_{opt.} = \bar{n}_{cl.}$.

Let $n_{tot} = \sum_{d\in U^1} n_d$ (note that $n_{tot} \neq n$, the sample size, since n_{tot} is the sum of n_d over all clusters in-sample and out-of-sample).

We now consider the solution of n_d given by n_d given by (8). Summing over all the clusters and dividing by the total number of clusters M we obtain

$$\frac{n_{tot}}{M} = \frac{1}{M\sqrt{\lambda c_2}} \sum_{d\in U^1} N_d^{\frac{q}{2}} \sqrt{(1-\rho)/\rho} - (1-\rho)/\rho \tag{16}$$

Solving for $\sqrt{\lambda}$ in (16) and substituting in (8) we obtain

$$n_d \quad = \quad n_{tot} P_d^{\frac{q}{2}} + (1-\rho)/\rho(MP_d^{\frac{q}{2}} - 1) \tag{17}$$

where $P_d^{\frac{q}{2}} = N_d^{\frac{q}{2}} / \sum_{d\in U^1} N_d^{\frac{q}{2}}$.

This solution for $\{n_d\}$ is identical to the area-only stratified formula for n_d given by Longford (2006), Molefe (2011) and Molefe and Clark (2015):

$$n_{h,opt.} \quad = \quad nP_h^{\frac{q}{2}} + (1-\rho)/\rho(HP_h^{\frac{q}{2}} - 1) \tag{18}$$

for stratified sampling design, with total sample size n replaced by n_{tot}. This shows that the two-stage allocation for n_d is the same as stratified allocation, given n_{tot}. We can then write the expected cost constraint (5) in terms of n_{tot} as

$$C_E \quad = \quad c_1 m + c_2 \frac{m}{M} n_{tot} \tag{19}$$

For $\frac{c_1}{c_2} = 10$, equation (14) gives a value of m which is greater than M when $q = 1$. When this happens, the optimal value for the number of clusters to take into the sample is $m = M$. As q approaches 2, the discriminant becomes negative so that there is no real-valued solution for m, implying that $m = M$ is optimal.

3.1. Numerical Example

The cost constraint for sampling is set at $C_F = 350$ cost units. The following per cluster to per subunit cost ratios are considered; $\frac{c_1}{c_2} = 10$, 4 and 2 where the cost per second stage unit $c_2 = 1$ cost units.

We used data on the 26 cantons (clusters) of Switzerland (Longford, 2006) to allocate the sample using the various simple two-stage designs. Throughout, we assume that $\rho = \frac{1}{40}$.

We compute the optimum sample sizes for each ratio $\frac{c_1}{c_2}$ by priority exponent q using (15) in Table 1. From the results, it is apparent that $\bar{n}_{opt.}$ is a decreasing function of q. As q increases the discriminant becomes small and eventually negative, resulting in the solution for $\bar{n}_{opt.}$ being negative or even a complex number. When this happens, the optimal sample size is $\bar{n}_{opt.} = 1$. We also observe that the optimum sample size decreases as $\frac{c_1}{c_2}$ decreases.

Therefore, the main finding here is that the general optimal gives a less clustered design when $q > 0$ than the classical two-stage optimal.

Table 1: Area-only simple two-stage optimum sample sizes

Priority	$\frac{c_1}{c_2} = 10$			$\frac{c_1}{c_2} = 4$			$\frac{c_1}{c_2} = 2$		
exponent	m_{opt}	$\bar{n}_{opt}$	$\bar{n}_{cl.}$	m_{opt}	$\bar{n}_{opt.}$	$\bar{n}_{cl.}$	$m_{opt.}$	$\bar{n}_{opt.}$	$\bar{n}_{cl.}$
$q = 0$	12	20	20	21	12	13	26	9	9
$q = \frac{1}{4}$	12	18	20	24	10	13	26	6	9
$q = \frac{1}{2}$	15	14	20	26	4	13	26	1	9
$q = \frac{3}{4}$	20	7	20	26	1	13	26	1	9
$q = 1$	26	1	20	26	1	13	26	1	9
$q = 2$	26	1	20	26	1	13	26	1	9

When $\frac{c_1}{c_2}$ is large, the sample is more clustered hence the CV's of the estimates of the cluster means are relatively smaller. However, the CV of the estimate of the grand mean will be large. When $\frac{c_1}{c_2}$ goes down, the sample becomes less clustered since we can take a larger number of clusters into the sample. When this happens the CV's of the estimates of the cluster means will be relatively larger since the within-cluster sample size is smaller, and the CV of the estimate of the grand mean will be smaller.

In the case of clusters of equal size, the within-cluster sample size is the same for all clusters selected into the sample. Hence, the optimization problem reduces to a singular problem of finding the optimal number of clusters to take into the sample.

The optimal number of clusters, $m_{opt.}$, and the optimal expected sample size of ultimate cluster, $\bar{n}_{opt.}$, subject to a fixed total expenditure, $C_F = c_1 m + c_2 m\bar{n}$, are $m_{opt.} = C_F/(c_1 + c_2\bar{n}_{opt.})$ where $\bar{n}_{opt.} = \sqrt{c_1/c_2(1-\rho)/\rho}$.

4. Other Designs

We consider several sensible but ad-hoc designs that include equal allocation, proportional allocation, classical optimal allocation and a combined design made up of the proportional allocation and the classical optimal design. We consider these ad-hoc designs

because sometimes the optimal design derived in Section 3 has undesirable properties such as negative or complex values for the analytical result for $\bar{n}_{cl.}$ or n_d (implying values of 1 in practice).

4.1. Equal Design

In the cluster equal design we consider the case in which a sample is taken from each and every small area (cluster). An equal number of secondary stage units is taken from each and every cluster. That is, $m = M$ and $n_d = n/M$ for $d = 1, \ldots, M$, where $n = (C_F - c_1 M)/c_2$ is the total sample size.

4.2. Proportional Design

In this design a sample is taken from each and every cluster. The within-cluster sample sizes are proportional to the population sizes of the clusters. The design is $m = M$ and $n_d = nP_d$ for $d = 1, \ldots, M$, where n is the same as in equal design and $P_d = N_d/N$.

4.3. Classical Optimal Design

The number of clusters taken into the sample is determined by the cost constraint. The within-cluster sample size is the standard optimal two-stage cluster design given by $m = C_F/(c_1 + c_2\bar{n}_{cl.})$ and $n_d = \bar{n}_{cl.}$ for $d = 1, \ldots, m$.

4.4. Proportional & Optimal Design

It may also be constructive to propose modifications of existing sampling designs. This design uses a combination of two designs. The within-cluster sample size is proportional to the cluster population size and also optimal for two-stage cluster design: $m = C_F/(c_1 + c_2\bar{n}_{cl.})$ and $n_d = P_d\bar{n}_{cl.}$ for $d = 1, \ldots, m$.

5. Numerical Evaluation

In this section, we compare the efficiency of the ad-hoc designs and the area-only optimum derived in Section 3. We consider the relative efficiency of these designs by calculating the ratios of F given by (6) of the designs using the equal design as the base design. A ratio less than one implies that a design is more efficient than the base design, whilst a ratio greater than one implies a design is less efficient than the base design.

In Table 2 we show the summary statistics of the CV's of the estimates of the cluster and national means for $\frac{c_1}{c_2} = 10$, 4 and 2 under the ad-hoc designs. The results show that for equal and classical optimum allocations, the CV's of the estimates of the small area means are narrowly dispersed by virtue of the design allocations being equal sample sizes. On the other hand, we see that the ranges of the CV's under proportional allocation and proportional & optimum allocation designs are widely dispersed since the clusters receive sample sizes that are proportionate to their population sizes.

Table 2: CV's of the ad-hoc designs

	Equal allocation	Proportional allocation	Classical optimum	Proportional & optimum	Area-only optimum[a]
$\frac{c_1}{c_2} = 10$ CV % (SAE's)					
Minimum	57.01	25.49	22.64	10.77	13.31
1st Quarter	57.01	44.16	22.65	20.50	20.16
Median	57.01	57.01	22.65	26.03	29.77
Mean	57.01	60.42	22.65	34.47	38.08
3rd Quarter	57.01	69.82	22.65	44.16	49.37
Maximum	57.01	98.74	22.65	98.74	98.74
CV % (National)	81.41	52.33	57.67	46.37	34.32
$\frac{c_1}{c_2} = 4$ CV % (SAE's)					
Minimum	32.91	15.24	27.37	13.19	12.34
1st Quarter	32.91	28.82	27.38	25.11	19.31
Median	32.91	37.61	27.39	33.07	27.39
Mean	32.91	49.54	27.38	41.87	36.05
3rd Quarter	32.91	69.82	27.39	57.01	38.82
Maximum	32.91	98.74	27.39	98.74	98.74
CV % (National)	46.75	31.75	45.71	33.75	31.24
$\frac{c_1}{c_2} = 2$ CV % (SAE's)					
Minimum	29.77	13.96	29.76	13.96	12.06
1st Quarter	29.77	26.64	29.77	26.64	18.60
Median	29.77	33.91	29.77	33.91	25.49
Mean	29.77	44.05	29.77	44.05	30.19
3rd Quarter	29.77	57.01	29.77	57.01	33.91
Maximum	29.77	98.74	29.77	98.74	69.82
CV % (National)	42.21	28.55	42.21	28.55	26.95

[a] Area-only optimum when $q = 1$

We observe that the classical optimum is relatively more efficient for estimating cluster means as shown by smaller CV's of the estimates of the cluster means compared to the other allocations. However, for estimating the national mean, the proportional & optimum allocation is relatively more efficient than the other designs. The CV of the estimate of the national mean is however considerably higher for the four ad-hoc designs, possibly showing that these two-stage cluster designs are not well suited for estimating the overall mean.

When $\frac{c_1}{c_2} = 4$, it implies more clusters and within-cluster samples for fixed C_F. The result of increased sample size is that the CV's of the estimates of the cluster means under

equal allocation and classical optimum are considerably lower than when $\frac{c_1}{c_2} = 10$. But proportional allocation and proportional & optimum allocation seem to be relatively better than equal and classical optimum allocations in estimating the national mean.

For $\frac{c_1}{c_2} = 2$ equal design give identical results to classical optimal design. Proportional design on the other hand also gives identical results to proportional & optimal design. These two designs perform better than equal and standard optimal designs for $q \geq \frac{1}{2}$ in terms of CV's and the criterion function F.

In Table 3 we see that proportional allocation and area-only stratified optimum given by (18) is less efficient than the base design. The area-only optimum designs should always be the best since the criterion function is minimized when the largest clusters are included in the sample but they are not when $q = 0$. Classical optimum and proportional & optimum are the only designs that are more efficient than equal allocation at $q = 0$. As the priority exponent q increases all the designs' efficiency against equal designs improve with the exception of classical optimum, whose efficiency is constant. At $q = 2$, the area-only stratified optimum and area-only optimum are nearly twice as efficient as equal design.

Table 3: Relative efficiency of two-stage designs for $\frac{c_1}{c_2} = 10$

		Priority Exponent (q)				
Designs	n_d	$q=0$	$q=\frac{1}{2}$	$q=1$	$q=\frac{3}{2}$	$q=2$
Equal allocation	$\frac{n}{M}$	1.00	1.00	1.00	1.00	1.00
Proportional allocation	$\frac{N_d}{N}n$	0.91	0.83	0.77	0.71	0.66
Classical optimum	$\bar{n}_{opt}$	0.92	0.92	0.92	0.92	0.92
Proportional & optimum	$\frac{N_d}{N}\bar{n}_{opt}$	0.91	0.83	0.77	0.71	0.61
Area-only stratified optimum	${n_d}^1$	1.00	0.87	0.76	0.63	0.53
Area-only optimum	${n_d}^2$	1.09	0.94	0.76	0.63	0.53

Table 4: Relative efficiency of two-stage designs for $\frac{c_1}{c_2} = 4$

		Priority Exponent (q)				
Designs	n_d	$q=0$	$q=\frac{1}{2}$	$q=1$	$q=\frac{3}{2}$	$q=2$
Equal allocation	$\frac{n}{M}$	1.00	1.00	1.00	1.00	1.00
Proportional allocation	$\frac{N_d}{N}n$	1.02	0.94	0.86	0.80	0.74
Classical optimum	$\bar{n}_{opt}$	0.98	0.98	0.98	0.98	0.98
Proportional & optimum	$\frac{N_d}{N}\bar{n}_{opt}$	1.02	0.94	0.87	0.82	0.77
Area-only stratified optimum	${n_d}^1$	1.00	0.92	0.81	0.69	0.58
Area-only optimum	${n_d}^2$	1.05	0.92	0.81	0.69	0.58

In Table 4 we present the relative efficiency for the two-stage designs when $\frac{c_1}{c_2} = 4$. We see that area-only stratified optimum is equally efficient as the base design whilst the area-only optimum is less efficient than the base design when $q = 0$. Also, proportional and proportional & optimum allocations are less efficient than equal allocation. The classical optimum design is the only design that is slightly more efficient than equal allocation at $q = 0$. As the priority exponent q increases all the designs' efficiency against equal designs improve with the exception of classical optimum, whose efficiency is marginal and constant. At $q = 2$, the area-only stratified optimum and the area-only optimum are almost twice as efficient as the equal design.

Table 5: Relative efficiency of two-stage designs for $\frac{c_1}{c_2} = 2$

		Priority Exponent (q)				
Designs	n_d	$q=0$	$q=\frac{1}{2}$	$q=1$	$q=\frac{3}{2}$	$q=2$
Equal allocation	$\frac{n}{M}$	1.00	1.00	1.00	1.00	1.00
Proportional allocation	$\frac{N_d}{N}n$	1.03	0.94	0.85	0.78	0.72
Classical optimum	$\bar{n}_{opt}$	1.00	1.00	1.00	1.00	1.00
Proportional & optimum	$\frac{N_d}{N}\bar{n}_{opt}$	1.03	0.94	0.85	0.78	0.72
Area-only stratified optimum	${n_d}^1$	1.00	0.92	0.81	0.70	0.59
Area-only optimum	${n_d}^2$	1.00	0.92	0.81	0.70	0.59

The area-only stratified optimum and the area-only optimum compare favorably to proportional allocation and proportional & optimum. Their relative efficiency improves as the priority exponent q approaches 2. At $q = 0$, equal allocation is as good as any of these designs, even better than, for example, proportional allocation and proportional & optimum. But at $q = 2$ area-only stratified optimum and the area-only optimum are twice as efficient as equal design, whilst proportional allocation and proportional & optimum are also more efficient than equal design but to a lesser extent.

In Table 5 one can observe that the relative performance of the area-only stratified optimum and the area-only optimum (relative to equal design) are only slightly superior to proportional allocation and proportional & optimum as q approaches 2.

When $\frac{c_1}{c_2} = 2$ the relative performance of the classical optimum is the same as the base design. We observe that the performance of proportional allocation is identical to proportional & optimum design. At $q = 0$ these two designs are less efficient than equal design. The area-only stratified optimum and the area-only optimum on the other hand are more efficient than the base design.

Overall we can see that the designs relative efficiencies improves as the ratio of $\frac{c_1}{c_2}$ goes up and q approaches 2.

[1] $n_d = nP_d^{\frac{q}{2}} + (1-\rho)/\rho(MP_d^{\frac{q}{2}} - 1)$
[2] $n_d = n_{tot}P_d^{\frac{q}{2}} + (1-\rho)/\rho(mP_d^{\frac{q}{2}} - 1)$
[1] $n_d = nP_d^{\frac{q}{2}} + (1-\rho)/\rho(MP_d^{\frac{q}{2}} - 1)$
[2] $n_d = n_{tot}P_d^{\frac{q}{2}} + (1-\rho)/\rho(mP_d^{\frac{q}{2}} - 1)$

6. Sensitivity Analysis

In section 5 the numerical evaluations of the sample designs were based on an assumed value of the intraclass correlation coefficient for the Switzerland canton data (Longford, 2006). In this section selected tables are replicated using Switzerland's cantons data for different values of ρ for the two-stage cluster designs, as well as for data on the population of the administrative districts of Botswana, to investigate how the optimal sample designs are altered as a result. For the two-stage designs we consider varying ρ, and C_F for $q = 1$.

6.1. Switzerland Canton Data

Here the interest is in finding out how the values of $\frac{c_1}{c_2}$, the cost ratio, C_F, the total fixed sampling cost, and ρ, the intraclass correlation coefficient, affect these designs. To investigate this we consider the relative efficiency of these designs by fixing one parameter and varying the others.

Table 6: Relative efficiency of simple two-stage designs for $\rho = \frac{1}{40}$, $\frac{c_1}{c_2} = 10$, $q = 1$

		Sampling cost (C_F)			
Designs	n_d	$C_F = 250$	$C_F = 300$	$C_F = 350$	$C_F = 400$
Equal allocation	$\frac{n}{M}$	1.00	1.00	1.00	1.00
Proportional allocation	$\frac{N_d}{N}n$	1.01	0.97	0.93	0.90
Classical optimum	$\bar{n}_{opt}$	0.88	0.90	0.92	0.94
Proportional & optimum	$\frac{N_d}{N}\bar{n}_{opt}$	1.01	0.97	0.93	0.90
Area-only stratified optimum	${n_d}^1$	0.73	0.74	0.76	0.76
Area-only optimum	${n_d}^2$	0.73	0.74	0.76	0.76

In Tables 6 - 7 we present the results of the numerical evaluation of the relative efficiency for the simple two-stage designs for $\rho = \frac{1}{4}$ and $q = 1$ when the sampling cost C_F is varied using data on the Switzerland's cantons. The results show that the area-only stratified optimum and the area-only optimum are the best designs and are identically efficient.

Table 7: Relative efficiency of simple two-stage designs for $\rho = \frac{1}{40}$, $\frac{c_1}{c_2} = 5$, $q = 1$

		Sampling cost (C_F)			
Designs	n_d	$C_F = 250$	$C_F = 300$	$C_F = 350$	$C_F = 400$
Equal allocation	$\frac{n}{M}$	1.00	1.00	1.00	1.00
Proportional allocation	$\frac{N_d}{N}n$	0.92	0.89	0.88	0.86
Classical optimum	$\bar{n}_{opt}$	0.97	0.98	0.99	1.00
Proportional & optimum	$\frac{N_d}{N}\bar{n}_{opt}$	0.92	0.89	0.88	0.86
Area-only stratified optimum	${n_d}^1$	0.80	0.80	0.80	0.80
Area-only optimum	${n_d}^2$	0.80	0.80	0.80	0.80

Table 8: Relative efficiency of simple two-stage designs for $C_F = 350$, $\frac{c_1}{c_2} = 10$, $q = 1$

		Intraclass Correlation (ρ)				
Designs	n_d	$\rho = \frac{1}{1000}$	$\rho = \frac{1}{100}$	$\rho = \frac{1}{4}$	$\rho = \frac{1}{20}$	$\rho = \frac{1}{10}$
Equal allocation	$\frac{n}{M}$	1.00	1.00	1.00	1.00	1.00
Proportional allocation	$\frac{N_d}{N}n$	1.00	0.97	0.93	0.89	0.84
Classical optimum	$\bar{n}_{opt}$	0.99	0.95	0.92	0.91	0.91
Proportional & optimum	$\frac{N_d}{N}\bar{n}_{opt}$	1.00	0.97	0.93	0.89	0.84
Area-only stratified optimum	${n_d}^1$	0.96	0.83	0.76	0.70	0.63
Area-only optimum	${n_d}^2$	0.96	0.83	0.76	0.76	0.76

Table 9: Relative efficiency of simple two-stage designs for $C_F = 350$, $\frac{c_1}{c_2} = 5$, $q = 1$

		Intraclass Correlation (ρ)				
Designs	n_d	$\rho = \frac{1}{1000}$	$\rho = \frac{1}{100}$	$\rho = \frac{1}{4}$	$\rho = \frac{1}{20}$	$\rho = \frac{1}{10}$
Equal allocation	$\frac{n}{M}$	1.00	1.00	1.00	1.00	1.00
Proportional allocation	$\frac{N_d}{N}n$	0.99	0.93	0.88	0.83	0.80
Classical optimum	$\bar{n}_{opt}$	1.00	0.99	0.99	0.99	1.03
Proportional & optimum	$\frac{N_d}{N}\bar{n}_{opt}$	0.99	0.93	0.88	0.83	0.80
Area-only stratified optimum	${n_d}^1$	0.96	0.85	0.80	0.76	0.74
Area-only optimum	${n_d}^2$	0.96	0.85	0.80	0.76	0.74

In Tables 8 - 9 we consider the relative efficiency of the designs for $C_F = 350$ cost units and $q = 1$ when ρ is varied. The results show that the area-only stratified optimum and the area-only optimum with partial coverage are the best designs for small values of ρ. As ρ increases, the area-only stratified optimum is the best design, with the area-only optimum nearly as good when $\rho = \frac{1}{40}$.

6.2. Botswana District Data

In this section we investigate the new sample designs for different data. We use data for the administrative districts of Botswana published by the Central Statistics Office (CSO). The population of Botswana is 1.67 million (Central Statistics Office, 2002). Botswana is divided into 16 administrative districts comprising major cities, towns and tribal territories. The smallest district is a mining town of Sowa with a population of almost 3,000 persons and the largest is Central district with a population of just over half a million inhabitants as per the 1991 population and housing census (Central Statistics Office, 2002).

For the simple two-stage designs we are interested in finding out whether the values of C_F and ρ has any effect on these designs. To investigate this we consider the relative

[1] $n_d = nP_d^{\frac{q}{2}} + (1-\rho)/\rho(MP_d^{\frac{q}{2}} - 1)$
[2] $n_d = n_{tot}P_d^{\frac{q}{2}} + (1-\rho)/\rho(mP_d^{\frac{q}{2}} - 1)$
[1] $n_d = nP_d^{\frac{q}{2}} + (1-\rho)/\rho(MP_d^{\frac{q}{2}} - 1)$
[2] $n_d = n_{tot}P_d^{\frac{q}{2}} + (1-\rho)/\rho(mP_d^{\frac{q}{2}} - 1)$

efficiency of these designs by fixing one parameter and varying the others.

In Tables 10 - 11 we present the results of the numerical evaluation of the relative efficiency for the simple two-stage designs for $\rho = \frac{1}{10}$ and $q = 1$ when the sampling cost C_F is varied using data on Botswana administrative data. The results show that the area-only stratified optimum given by (18) and the area-only optimum given by (17) are the best designs and are identical.

Table 10: Relative efficiency of simple two-stage designs for $\rho = \frac{1}{10}$, $\frac{c_1}{c_2} = 10$, $q = 1$

		Sampling cost (C_F)			
Designs	n_d	$C_F = 250$	$C_F = 300$	$C_F = 350$	$C_F = 400$
Equal allocation	$\frac{n}{M}$	1.00	1.00	1.00	1.00
Proportional allocation	$\frac{N_d}{N}n$	0.78	0.80	0.82	0.78
Classical optimum	$\bar{n}_{opt}$	0.96	1.00	1.00	1.00
Proportional & optimum	$\frac{N_d}{N}\bar{n}_{opt}$	0.90	0.83	0.77	0.78
Area-only stratified optimum	${n_d}^1$	0.68	0.69	0.71	0.72
Area-only optimum	${n_d}^2$	0.68	0.69	0.71	0.72

Table 11: Relative efficiency of simple two-stage designs for $\rho = \frac{1}{10}$, $\frac{c_1}{c_2} = 5$, $q = 1$

		Sampling cost (C_F)			
Designs	n_d	$C_F = 250$	$C_F = 300$	$C_F = 350$	$C_F = 400$
Equal allocation	$\frac{n}{M}$	1.00	1.00	1.00	1.00
Proportional allocation	$\frac{N_d}{N}n$	0.82	0.85	0.87	0.90
Classical optimum	$\bar{n}_{opt}$	1.00	1.00	1.00	1.00
Proportional & optimum	$\frac{N_d}{N}\bar{n}_{opt}$	0.77	0.77	0.78	0.78
Area-only stratified optimum	${n_d}^1$	0.72	0.74	0.74	0.75
Area-only optimum	${n_d}^2$	0.72	0.74	0.74	0.75

Table 12: Relative efficiency of simple two-stage designs for $C_F = 350$, $\frac{c_1}{c_2} = 10$, $q = 1$

		Intraclass Correlation (ρ)				
Designs	n_d	$\rho = \frac{1}{1000}$	$\rho = \frac{1}{100}$	$\rho = \frac{1}{4}$	$\rho = \frac{1}{20}$	$\rho = \frac{1}{10}$
Equal allocation	$\frac{n}{M}$	1.00	1.00	1.00	1.00	1.00
Proportional allocation	$\frac{N_d}{N}n$	0.98	0.89	0.83	0.80	0.82
Classical optimum	$\bar{n}_{opt}$	0.99	0.98	0.98	1.00	1.00
Proportional & optimum	$\frac{N_d}{N}\bar{n}_{opt}$	1.01	0.99	0.95	0.85	0.77
Area-only stratified optimum	${n_d}^1$	0.99	0.89	0.81	0.75	0.71
Area-only optimum	${n_d}^2$	0.99	0.89	0.81	0.75	0.71

In Tables 12 - 13 we consider the relative efficiency of the designs for $C_F = 350$ cost units and $q = 1$ when ρ is varied. The results show that the area-only stratified optimum and the area-only optimum with partial coverage are the best designs for all values of ρ. The proportional & optimum design is nearly as good.

Table 13: Relative efficiency of simple two-stage designs for $C_F = 350$, $\frac{c_1}{c_2} = 5$, $q = 1$

		Intraclass Correlation (ρ)				
Designs	n_d	$\rho = \frac{1}{1000}$	$\rho = \frac{1}{100}$	$\rho = \frac{1}{4}$	$\rho = \frac{1}{20}$	$\rho = \frac{1}{10}$
Equal allocation	$\frac{n}{M}$	1.00	1.00	1.00	1.00	1.00
Proportional allocation	$\frac{N_d}{N} n$	0.98	0.87	0.82	0.82	0.87
Classical optimum	$\bar{n}_{opt}$	1.00	1.00	1.00	1.00	1.00
Proportional & optimum	$\frac{N_d}{N} \bar{n}_{opt}$	0.99	0.93	0.80	0.78	0.78
Area-only stratified optimum	${n_d}^1$	0.98	0.89	0.81	0.76	0.74
Area-only optimum	${n_d}^2$	0.98	0.89	0.81	0.76	0.74

In this section we have used Switzerland canton data and Botswana district data. We have replicated the numerical evaluation of the various designs by considering the relative efficiencies of the designs, by computing the values of F for designs under consideration. We considered relative priority exponent $q = 1$ and selected values of the relative priority coefficient. Selected tables are replicated using Switzerland's cantons data for different values of ρ for the stratified designs, as well as for data on the population of the administrative districts for Botswana to investigate how the optimal sample designs are modified as a result. For the two-stage designs we consider varying $\frac{c_1}{c_2}$, ρ, and C_F for fixed q.

To investigate whether the value of ρ, the intraclass correlation, has an effect on the stratified allocations, we consider different values of ρ whilst keeping the priority coefficient and priority exponent fixed, for these designs. When $q = 1$ proportional allocation and optimal power allocation are the best designs when $\rho = \frac{1}{1000}$. As ρ increases, all designs are the best except for proportional allocation and equal allocation.

For the simple two-stage designs we are interested in finding out whether the values of C_F and ρ has any effect on the choice of the within-cluster sample size. The results as in section 5, show that the area-only stratified optimum given by equation 18 and the area-only optimum given by equation 17 are the best designs. When ρ is varied for fixed C_F and $q = 1$, the results show that the area-only stratified optimum and the area-only optimum with partial coverage are the best designs for all values of ρ.

7. Conclusions

An analytical solution for the stationary point exists when the only priority is small area estimation. This optimal design is less clustered than the usual classical two-stage optimal sample size $\bar{n}_{cl.}$ when more priority is given to larger clusters ($q > 0$). The optimal sample size depends on the cost per cluster relative to ($\frac{c_1}{c_2}$), intraclass correlation coefficient (ρ) and the relative variance of $N_d^{\frac{q}{2}}$ denoted by $C^2_{q/2}$. When the only priority is small area

estimation, that is, $q = 0$, or when the N_d's are constant, $C^2_{q/2} = 0$ and the general optimal coincides with the classical optimal. The area-only optimal average sample size is usually a decreasing function of $C^2_{q/2}$, so that when $C^2_{q/2} > 0$, $\bar{n}_{opt.}$ will be less than the classical optimum. A sufficient condition for this is that $\gamma/c_2 < 0$, which would usually be satisfied, unless $\frac{c_1}{c_2}$ or ρ are unusually large.

The area-only stratified optimum and the area-only simple two-stage optimum should always be the best designs in minimizing the objective function but they are not when there is equal priority for each cluster, that is when $q = 0$. These two designs have undesirable properties of allocating zero or even negative sample sizes to smaller clusters. Negative sample sizes are obviously not possible in practice and this anomaly is corrected by setting them to zero and reallocating again.

When the clusters are equally important ($q = 0$), classical optimum and proportional & optimum are the best designs especially when the cost ratio is high, in this case when $\frac{c_1}{c_2} = 10$. When $\frac{c_1}{c_2} = 2$, proportional design and proportional & optimum design are less efficient than equal allocation. Also, the classical optimum is as efficient as equal allocation. All the other designs are better as q approaches 2, with area-only stratified optimum and the area-only optimum being the best.

References

Bankier, M. D., (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42(3), pp. 174–177.

Clark, R. G., Steel, D. G., (2007). Sampling within households in households surveys. *Journal of the Royal Statistical Society A*, 170(Issue 1), pp. 63–82.

Cochran, W. G., (1977). Sampling Techniques. Wiley and Sons.

Foreman, E. K., (1991). Survey Sampling Principles. Marcel Dekker, Inc.

Fuller, W. A., (1999). Environmental Surveys Over Time. Journal of Agricultural, Biological and Environmental Statistics, 4, pp. 331–345.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G., (1953). Sample survey methods and theory, volume 1 & 2. Wiley, New York.

Hidiroglou, M. A., Patak, Z., (2004). Domain estimation using linear regression. *Survey Methodology*, 30, pp. 67–78.

Longford, N. T., (2006). Sample size calculation for small-area estimation. Survey Methodology, 32(1), pp. 87–96.

Molefe, W. B., (2011). Sample design for small area estimation. PhD thesis, University of Wollongong, http://ro.uow.edu.au/theses/3495.

Molefe, W. B., Clark, R. G., (2015). Model-assisted optimal allocation for planned domains using composite estimation. *Survey Methodology*, 41(2), pp. 377387.

Molefe, W. B., Shangodoyin, D. K., and Clark, R. G., (2015). An approximation to the optimal subsample allocation for small areas. *Statistics in Transition, new series*, 16(2), pp. 163–182.

Pettersson, H., Sisouphanthong, B., (2005). Household Sample Surveys in Developing and Transition Countries, chapter Cost Model for an Income and Expenditure Survey, pages 267–277. Number 96 in Series F. United Nations: Statistics Division, Department of Economic and Social Affairs.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., and Asok, C., (1954). Sampling theory of surveys with applications. Iowa State University Press, third edition.

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 149–160, DOI 10.2478/stattrans-2022-0047
Received – 23.12.2021; accepted – 22.08.2022

Comparison of confidence intervals for variance components in an unbalanced one-way random effects model

Arisa Jiratampradab[1], **Thidaporn Supapakorn** [2], **Jiraphan Suntornchost**[3]

ABSTRACT

The purpose of this paper is to study and compare the methods for constructing confidence intervals for variance components in an unbalanced one-way random effects model. The methods are based on a classical exact, generalised pivotal quantity, a fiducial inference and a fiducial generalised pivotal quantity. The comparison of criteria involves the empirical coverage probability that maintains at the nominal confidence level of 0.95 and the shortest average length of the confidence interval. The simulation results show that the method based on the generalised pivotal quantity and the fiducial inference perform very well in terms of both the empirical coverage probability and the average length of the confidence interval. The classical exact method performs well in some situations, while the fiducial generalised pivotal quantity performs well in a very unbalanced design. Therefore, the method based on the generalised pivotal quantity is recommended for all situations.

Key words: variance components, unbalanced one-way random effects model, pivotal quantity, fiducial inference, coverage probability.

1. Introduction

The one-way random effects model is studied in many applications, such as medical treatment, animal breeding studies, agricultural genetics and industrial process management, etc. The variance components of this model are used to consider the different sources of variation. For example, radiotherapy doses for cancer treatment are determined by process variation due to difference in area of organs of individual patients and diagnosis of individual physician (Demetrashvili et al., 2016). Thus, the inferences for variance components in the model is of interest. Consider the one-way random effects model

$$y_{ij} = \mu + a_i + e_{ij}, \;\; i = 1,\ldots,g, \;\; j = 1,\ldots,n_i, \tag{1}$$

where y_{ij} is the random observation, μ is the overall mean. The random group effects a_i and the random errors e_{ij} are mutually independent random variables, and distributed

[1]Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand. E-mail: arisajiratam@gmail.com. ORCID: https://orcid.org/0000-0001-6375-70922
[2]Corresponding author. Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand. E-mail: fscitdps@ku.ac.th. ORCID: https://orcid.org/0000-0003-0019-9884
[3]Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand. E-mail: jiraphan.s@chula.ac.th. ORCID: https://orcid.org/0000-0001-5410-9659

as $N(0,\sigma_a^2)$ and $N(0,\sigma_e^2)$, respectively. In addition, let $n=\sum_{i=1}^{g} n_i$ denote the number of the total observations. When the number of observations n_i of each group is equal, model (1) is called balanced model. Otherwise, it is called unbalanced model. The source of variation is known as the variance components, namely, σ_a^2 and σ_e^2. In general, σ_a^2 is called between-group variance component, and σ_e^2 is called within-group variance component. The proportion of the between-group variance component and the total variation can be written in the form $\rho=\sigma_a^2/(\sigma_a^2+\sigma_e^2)$, which measures the importance of one effect related to the other effect.

One very important property of an estimator is minimal sufficient statistics. A closed-form function of the minimal sufficient statistics is available in balanced random model. However, these functions are unavailable in unbalanced random model as described by Searle et al. (2006). Furthermore, solving the closed-form functions of the minimal sufficient statistics in the unbalanced case is computationally complicated for estimation of the variance components. There are several works in the literature that studied inferences for variance components in unbalanced model, such as Wald (1940), Thomas and Hultquist (1978), Park and Burdick (2003), and Arendacká (2005) which are based on a pivotal quantity approach. Ting et al. (1990) and Hartung and Knapp (2000) studied that by the classical exact method. Li and Li (2007) and Lidong et al. (2008) used the idea of a fiducial generalized confidence interval for variance components. Liu et al. (2016) proposed the concept of the fiducial generalized pivotal quantity for constructing the confidence interval for variance components in unbalanced model.

The aim of this paper is to compare five methods which are applicable to confidence intervals for between-group variance component in unbalanced one-way random effects model. These five methods are as follows: the Ting and others (TG) method (Ting et al., 1990), the Hartung-Knapp (HK) method (Hartung and Knapp, 2000), the Park-Burdick (PB) method (Park and Burdick, 2003), the Li-Li (LL) method (Li and Li, 2007), and the Liu-Xu-Hannig (LXH) method (Liu et al., 2016).

The paper is organized as follows. Section 2 describes the model and notation. Section 3 presents the methods for constructing a confidence interval for σ_a^2. Section 4 shows the results of a simulation study and compare the performance of the methods. Section 5 provides previously published data example. In the final Section 6, a conclusion is given.

2. Model and notation

A matrix formulation of the model (1) is given by

$$\mathbf{Y}=\mathbf{1}_n\mu+\mathbf{Z}\mathbf{A}+\mathbf{E}, \tag{2}$$

where $\mathbf{Y}=(\mathbf{Y}_1',\ldots,\mathbf{Y}_g')'$ with $\mathbf{Y}_i=(Y_{i1},\ldots,Y_{in_i})'$ for $i=1,\ldots,g$, $\mathbf{1}_n=(\mathbf{1}_{n_1}',\ldots,\mathbf{1}_{n_g}')'$ with $\mathbf{1}_{n_i}$ is a $n_i\times 1$ vector of ones, and $n=\sum_{i=1}^{g} n_i$. The matrix $\mathbf{Z}=\mathrm{diag}(\mathbf{1}_{n_1},\ldots,\mathbf{1}_{n_g})$ is known as incidence matrix of size $n\times g$. The random group effects vector $\mathbf{A}=(A_1,\ldots,A_g)'$ is distributed as $N(\mathbf{0}_g,\sigma_a^2\mathbf{I}_g)$ and the random errors vector $\mathbf{E}=(\mathbf{E}_1',\ldots,\mathbf{E}_g')'$ with $\mathbf{E}_i=(E_{i1},\ldots,E_{in_i})'$ is distributed as $N(\mathbf{0}_n,\sigma_e^2\mathbf{I}_n)$, where $\mathbf{0}_c$ is a $c\times 1$ vector of zeros, and $\mathbf{I}_c$ is a $c\times c$ identity matrix. The random vectors $\mathbf{A}$ and $\mathbf{E}$ are mutually independent. De-

note $r_1 = \text{rank}(\mathbf{X}) - \text{rank}(\mathbf{1}_n)$ and $r_2 = n - \text{rank}(\mathbf{X})$, where $\mathbf{X} = (\mathbf{1}_n, \mathbf{ZZ}')$ is the horizontal concatenation of matrices $\mathbf{1}_n$ and $\mathbf{ZZ}'$. Under model (2), the distribution function of $\mathbf{Y}$ is $\mathbf{Y} \sim N(\mu\mathbf{1}_n, \sigma_a^2\mathbf{ZZ}' + \sigma_e^2\mathbf{I}_n)$, then $\mathbf{H}'\mathbf{Y} \sim N(\mathbf{0}, \sigma_a^2\mathbf{W} + \sigma_e^2\mathbf{I})$, where $\mathbf{H}$ is matrix whose columns span the space orthogonal to the space spanned by the column vector of ones (Burch, 2011), $\mathbf{W}$ is the part of the variance-covariance matrix associated with σ_a^2, $\mathbf{0}$ is vector of zeros, and $\mathbf{I}$ is identity matrix. The quadratic form is denoted by $T = \mathbf{Y}'\mathbf{BY}$, where $\mathbf{B}$ is an appropriately chosen symmetric matrix of constants called the matrix of the quadratic form (Milliken and Johnson, 2009).

Graybill (1976) described the properties of quadratic forms for estimation of the variance components. The independently quadratic forms, denoted by $T_1, \ldots, T_d$, T_{d+1}, are minimal sufficient statistics for (σ_a^2, σ_e^2) under multivariate normal distribution of $\mathbf{Y}$. Burch (2011) showed that the sum of squares due to between groups SS_a and the sum of squares due to within groups SS_e can be expressed as quadratic forms $\mathbf{Y}'(\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}' - \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^-\mathbf{1}_n')\mathbf{Y} = T_1 + \cdots + T_d$ and $\mathbf{Y}'(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}')\mathbf{Y} = T_{d+1}$, respectively. The mean square for between groups and the mean square for within groups are denoted by $MS_a = SS_a/r_1$ and $MS_e = SS_e/r_2$, respectively. Furthermore, MS_a and MS_e are independent, and SS_e/σ_e^2 has a chi-squared distribution with r_2 degrees of freedom.

3. Approximate confidence intervals for σ_a^2

Several existing methods for constructing the confidence interval for σ_a^2 are reviewed in this section.

3.1. The TG method

Ting et al. (1990) suggested the method for constructing the confidence interval for the variance components in random effect model applying results provided by Howe (1974) and using cross-product terms in Ting et al. (1989). Let $\mathbf{W}_{\text{TG}} = \mathbf{H}_{\text{TG}}\mathbf{ZZ}'\mathbf{H}_{\text{TG}}$, where $\mathbf{H}_{\text{TG}}$ is a $n \times n$ matrix such that $\mathbf{H}_{\text{TG}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}' - \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^-\mathbf{1}_n'$. Let $\lambda_1 > \cdots > \lambda_d > 0$ be the distinct positive eigenvalues of $\mathbf{W}_{\text{TG}}$ having multiplicities $s_1, \ldots, s_d$. Define $SS_a = \mathbf{Y}'\mathbf{H}_{\text{TG}}'\mathbf{W}_{\text{TG}}^-\mathbf{H}_{\text{TG}}\mathbf{Y}$.

The approximate $100(1-\alpha)\%$ confidence interval for σ_a^2 is derived by

$$[MS_a - \frac{1}{b}MS_e - (G_1^2MS_a^2 + \frac{1}{b^2}C_2^2MS_e^2 + \frac{1}{b}G_{12}MS_aMS_e)^{1/2},$$
$$MS_a - \frac{1}{b}MS_e + (C_1^2MS_a^2 + \frac{1}{b^2}G_2^2MS_e^2 + \frac{1}{b}C_{12}MS_aMS_e)^{1/2}],$$

where $b = r_1(\sum_{\ell=1}^d s_\ell/\lambda_\ell)^{-1}$, $G_1 = 1 - 1/F_{1-\alpha,(r_1,\infty)}$, $C_2 = 1/F_{\alpha,(r_2,\infty)} - 1$,
$G_{12} = [(F_{1-\alpha,(r_1,r_2)} - 1)^2 - G_1^2F_{1-\alpha,(r_1,r_2)}^2 - C_2^2]/F_{1-\alpha,(r_1,r_2)}$, $C_1 = 1/F_{\alpha,(r_1,\infty)} - 1$,
$G_2 = 1 - 1/F_{1-\alpha,(r_2,\infty)}$, $C_{12} = [(1 - F_{\alpha,(r_1,r_2)})^2 - C_1^2F_{\alpha,(r_1,r_2)}^2 - G_2^2]/F_{\alpha,(r_1,r_2)}$.

Note that $F_{\alpha,(r_1,r_2)}$ and $F_{1-\alpha,(r_1,r_2)}$ are the α and $1-\alpha$ quantiles of the F-distribution with degrees of freedom r_1 and r_2, respectively. Furthermore, $F_{\alpha,(r_1,\infty)} = \chi^2_{\alpha,r_1}/r_1$, $F_{1-\alpha,(r_1,\infty)} = \chi^2_{1-\alpha,r_1}/r_1$, $F_{\alpha,(r_2,\infty)} = \chi^2_{\alpha,r_2}/r_2$, and $F_{1-\alpha,(r_2,\infty)} = \chi^2_{1-\alpha,r_2}/r_2$ (Milliken and Johnson, 2009).

3.2. The HK method

Hartung and Knapp (2000) developed the method for constructing the confidence interval for the between-group variance component using the concept of Wald (1940). The sufficient statistics of the HK method are defined as $T_{\mathrm{HK}_1},\ldots,T_{\mathrm{HK}_g}$, where $T_{\mathrm{HK}_i}=(\mathbf{1}'_{n_i}\mathbf{1}_{n_i})^{-}\mathbf{1}'_{n_i}\mathbf{Y}_i$, $i=1,\ldots,g$.

The approximate $100(1-\alpha)\%$ confidence interval for σ_a^2 is derived by

$$[MS_eR_1, MS_eR_2].$$

Note that R_1 and R_2 are the root of the equations as follows:

$$f(R_1)=\frac{\sum_{i=1}^{g}w_i(T_{\mathrm{HK}_i}-\sum_{i=1}^{g}w_iT_{\mathrm{HK}_i}/\sum_{i=1}^{g}w_i)^2}{r_1MS_e}\sim F_{1-\alpha/2,(r_1,r_2)} \quad \text{and}$$
$$f(R_2)=\frac{\sum_{i=1}^{g}v_i(T_{\mathrm{HK}_i}-\sum_{i=1}^{g}v_iT_{\mathrm{HK}_i}/\sum_{i=1}^{g}v_i)^2}{r_1MS_e}\sim F_{\alpha/2,(r_1,r_2)},$$

where $w_i=n_i/(1+n_iR_1)$ and $v_i=n_i/(1+n_iR_2)$.

3.3. The PB method

Park and Burdick (2003) proposed the generalized pivotal quantity for constructing the confidence interval for the between-group variance component using results provided by Olsen et al. (1976). Let $\mathbf{W}_{\mathrm{PB}}=\mathbf{H}_{\mathrm{PB}}\mathbf{Z}\mathbf{Z}'\mathbf{H}_{\mathrm{PB}}$, where $\mathbf{H}_{\mathrm{PB}}$ is a $n\times n$ matrix such that $\mathbf{H}_{\mathrm{PB}}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'-\mathbf{1}_n(\mathbf{1}'_n\mathbf{1}_n)^{-}\mathbf{1}'_n$. Let $\lambda_1>\cdots>\lambda_d>0$ be the distinct positive eigenvalues of $\mathbf{W}_{\mathrm{PB}}$ having multiplicities $s_1,\ldots,s_d$. Let $\mathbf{P}_{\mathrm{PB}}=[\mathbf{P}_{\mathrm{PB}_1},\ldots,\mathbf{P}_{\mathrm{PB}_d}]$ be $n\times n$ orthogonal matrix such that $\mathbf{P}'_{\mathrm{PB}}\mathbf{W}_{\mathrm{PB}}\mathbf{P}_{\mathrm{PB}}=\mathrm{diag}(\lambda_1\mathbf{1}'_{s_1},\ldots,\lambda_d\mathbf{1}'_{s_d})$, where $\mathbf{P}_{\mathrm{PB}_\ell}$, $\ell=1,\ldots,d$ corresponding to λ_ℓ is of dimension $n\times s_\ell$.

The minimal sufficient statistics of the PB method are defined as $T_{\mathrm{PB}_1},\ldots,T_{\mathrm{PB}_d}$, where $T_{\mathrm{PB}_\ell}=\mathbf{Y}'\mathbf{H}'_{\mathrm{PB}}\mathbf{P}_{\mathrm{PB}_\ell}(\mathbf{P}'_{\mathrm{PB}_\ell}\mathbf{P}_{\mathrm{PB}_\ell})^{-}\mathbf{P}'_{\mathrm{PB}_\ell}\mathbf{H}_{\mathrm{PB}}\mathbf{Y}$, $\ell=1,\ldots,d$. Lamotte (1976) showed that $SS_a=\sum_{\ell=1}^{d}T_{\mathrm{PB}_\ell}$, where $T_{\mathrm{PB}_\ell}/(\lambda_\ell\sigma_a^2+\sigma_e^2)$, $\ell=1,\ldots,d$ has the chi-squared distribution with s_ℓ degrees of freedom. The function of the generalized pivotal quantity is defined by R as the solution for σ_a^2 in the non-linear equation given by

$$U=\sum_{\ell=1}^{d}\frac{T_{\mathrm{PB}_\ell}}{\lambda_\ell R+r_2MS_e/K}, \tag{3}$$

where $U\sim\chi^2_{r_1}$ and $K\sim\chi^2_{r_2}$.

The approximate $100(1-\alpha)\%$ confidence interval for σ_a^2 is derived by

$$[\max(0,R_{\alpha/2}),\max(0,R_{1-\alpha/2})],$$

where $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of R in equation (3), respectively. Note that the solutions of $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are based on pivotal quantities.

3.4. The LL method

Li and Li (2007) presented the concept of the fiducial inference for constructing the confidence interval for the between-group variance component in random effect model applying results provided by Li and Li (2005). Let $\mathbf{W}_{\mathrm{LL}} = \mathbf{H}_{\mathrm{LL}}(\mathbf{Z}'\mathbf{Z})^{-}\mathbf{H}'_{\mathrm{LL}}$, where $\mathbf{H}_{\mathrm{LL}}$ is a $(g-1)\times g$ matrix such that $\mathbf{H}_{\mathrm{LL}}\mathbf{H}'_{\mathrm{LL}} = \mathbf{I}_{g-1}$ and $\mathbf{H}'_{\mathrm{LL}}\mathbf{H}_{\mathrm{LL}} = \mathbf{I}_g$. Let $\lambda_1 > \cdots > \lambda_d \geq 0$ be the distinct eigenvalues of $\mathbf{W}_{\mathrm{LL}}$ having multiplicities $s_1,\ldots,s_d$. Let $\mathbf{P}_{\mathrm{LL}} = [\mathbf{P}_{\mathrm{LL}_1},\ldots,\mathbf{P}_{\mathrm{LL}_d}]$ be $(g-1)\times(g-1)$ orthogonal matrix such that $\mathbf{P}'_{\mathrm{LL}}\mathbf{W}_{\mathrm{LL}}\mathbf{P}_{\mathrm{LL}} = \mathrm{diag}(\lambda_1\mathbf{1}'_{s_1},\ldots,\lambda_d\mathbf{1}'_{s_d})$, where $\mathbf{P}_{\mathrm{LL}_\ell}$, $\ell = 1,\ldots,d$ corresponding to λ_ℓ is of dimension $(g-1)\times s_\ell$.

The sufficient statistics of the LL method are defined as $\mathbf{T}_{\mathrm{LL}} = \mathbf{P}_{\mathrm{LL}}\mathbf{H}_{\mathrm{LL}}(\mathbf{Z}'\mathbf{Z})^{-}\mathbf{Z}'\mathbf{Y}$. The function of the fiducial inference is given by

$$R = \frac{\mathbf{T}'_{\mathrm{LL}}\mathbf{T}_{\mathrm{LL}} - \mathbf{Q}'\mathbf{C}\mathbf{Q}SS_e/K}{\mathbf{Q}'\mathbf{Q}}, \tag{4}$$

where $\mathbf{C} = \mathbf{P}_{\mathrm{LL}}\mathbf{W}_{\mathrm{LL}}\mathbf{P}'_{\mathrm{LL}}$, $\mathbf{Q} \sim N(\mathbf{0},\mathbf{I}_{r_1})$, and $K \sim \chi^2_{r_2}$.

The approximate $100(1-\alpha)\%$ confidence interval for σ_a^2 is derived by

$$[\max(0,R_{\alpha/2}),\max(0,R_{1-\alpha/2})],$$

where $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of R in equation (4), respectively. Note that the solutions of $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are based on pivotal quantities.

3.5. The LXH method

Liu et al. (2016) proposed the least squares idea of the fiducial generalized pi-votal quantity for constructing the confidence interval for the variance components in random effect model. Let $\mathbf{W}_{\mathrm{LXH}} = \mathbf{H}'_{\mathrm{LXH}}\mathbf{Z}\mathbf{Z}'\mathbf{H}_{\mathrm{LXH}}$, where $\mathbf{H}_{\mathrm{LXH}}$ is a $n\times(n-1)$ matrix such that $\mathbf{H}_{\mathrm{LXH}}\mathbf{H}'_{\mathrm{LXH}} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}'_n$ and $\mathbf{H}'_{\mathrm{LXH}}\mathbf{H}_{\mathrm{LXH}} = \mathbf{I}_{n-1}$. Let $\lambda_1 > \cdots > \lambda_d \geq 0$ be the distinct eigenvalues of $\mathbf{W}_{\mathrm{LXH}}$ having multiplicities $s_1,\ldots,s_d$. Let $\mathbf{P}_{\mathrm{LXH}} = [\mathbf{P}_{\mathrm{LXH}_1},\ldots,\mathbf{P}_{\mathrm{LXH}_d}]$ be $(n-1)\times(n-1)$ orthogonal matrix such that $\mathbf{P}'_{\mathrm{LXH}}\mathbf{W}_{\mathrm{LXH}}\mathbf{P}_{\mathrm{LXH}} = \mathrm{diag}(\lambda_1\mathbf{1}'_{s_1},\ldots,\lambda_d\mathbf{1}'_{s_d})$, where $\mathbf{P}_{\mathrm{LXH}_\ell}$, $\ell = 1,\ldots,d$ corresponding to λ_ℓ is of dimension $(n-1)\times s_\ell$.

The minimal sufficient statistics of the LXH method are defined as $T_{\mathrm{LXH}_1},\ldots,T_{\mathrm{LXH}_d}$, where $T_{\mathrm{LXH}_\ell} = \mathbf{Y}'\mathbf{H}_{\mathrm{LXH}}\mathbf{P}_{\mathrm{LXH}_\ell}\mathbf{P}'_{\mathrm{LXH}_\ell}\mathbf{H}'_{\mathrm{LXH}}\mathbf{Y}$, $\ell = 1,\ldots,d$. The variables U_ℓ, $\ell = 1,\ldots,d$ are mutually independent and $U_\ell = T_{\mathrm{LXH}_\ell}/(\lambda_\ell\sigma_a^2 + \sigma_e^2)$, $\ell = 1,\ldots,d$ has the chi-squared distribution with s_ℓ degrees of freedom. The function of the least squares fiducial inference is given by

$$R = \frac{\sum_{\ell=1}^d U_\ell^2\sum_{\ell=1}^d \lambda_\ell T_{\mathrm{LXH}_\ell}U_\ell - \sum_{\ell=1}^d \lambda_\ell U_\ell^2\sum_{\ell=1}^d T_{\mathrm{LXH}_\ell}U_\ell}{\sum_{\ell=1}^d U_\ell^2\sum_{\ell=1}^d \lambda_\ell^2 U_\ell^2 - (\sum_{\ell=1}^d \lambda_\ell U_\ell^2)^2}. \tag{5}$$

The approximate $100(1-\alpha)\%$ confidence interval for σ_a^2 is derived by

$$[\max(0,R_{\alpha/2}),\max(0,R_{1-\alpha/2})],$$

where $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of R in equation (5), respectively. Note that the solutions of $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are based on pivotal quantities.

4. Simulation study

In this section, a comparison of the methods for constructing the confidence interval for σ_a^2 with the methods described in Section 3 is studied by the Monte Carlo simulation. Without loss of generality, it is assumed that $\mu = 0$ in model (2). The values chosen for (σ_a^2, σ_e^2) are $(0.001, 0.999)$, $(0.1, 0.9)$, $(0.2, 0.8)$, $(0.3, 0.7)$, $(0.4, 0.6)$, $(0.5, 0.5)$, $(0.6, 0.4)$, $(0.7, 0.3)$, $(0.8, 0.2)$, $(0.9, 0.1)$, and $(0.999, 0.001)$. The ratio of va-riance components, $\rho = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$ varies from small to large. The nominal confidence level of 0.95 is considered. The simulation study is based on $5,000$ iterations for each setting of the values (σ_a^2, σ_e^2) and the sample size pattern $(n_i,\ i = 1, \ldots, g)$.

The criteria for analysing the performance of the methods are the empirical coverage probability that maintains at the nominal confidence level, and the shortest average length of the confidence interval. The empirical coverage prob-ability is firstly considered, and the average length of the confidence interval is later compared. The degree of imbalance is $\Phi = (g/\sum_{i=1}^{g} n_i)(g/\sum_{i=1}^{g} 1/n_i)$, which is used to measure imbalance in one-way model (Ahrens and Pincus, 1981). Note that $0 < \Phi \leq 1$ is equal to 1 if and only if the model is balanced, and Φ is close to 0 when the model is very unbalanced. The coverage probability of confidence interval for σ_a^2 depends on the degree of imbalance and the design $(n_1, \ldots, n_g)$. The simulation patterns are shown in Table 1.

Table 1. Unbalanced patterns used in simulations

Pattern	Φ	g	n_i									
1	0.044	3	1	1	100							
2	0.570	3	3	7	20							
3	0.818	3	5	10	15							
4	0.068	6	1	1	1	1	1	100				
5	0.700	6	5	10	15	20	25	30				
6	0.957	6	6	6	8	8	10	10				
7	0.525	10	1	1	4	4	6	6	8	8	10	10
8	0.835	10	3	3	4	5	6	6	8	8	10	10

The simulation results are represented in the boxplots of Figures 1 and 2. The empirical coverage probabilities of the confidence interval for σ_a^2 with the number of groups $g = 3$, 6, and 10, where $\rho < 0.5$ and $\rho \geq 0.5$, are shown in Figure 1. The relative difference of the average length of the confidence interval for σ_a^2 with the number of groups $g = 3$, 6, and 10, where $\rho < 0.5$ and $\rho \geq 0.5$ is shown in Figure 2. The relative length is defined as $(L_{\mathrm{M}} - L_{\mathrm{PB}})/L_{\mathrm{PB}}$, where L_{M} denotes the average interval length of competing methods and

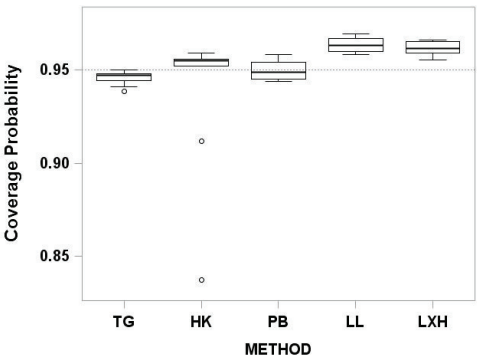
L_{PB} denotes the average interval length of the PB method. Clearly, the positive value of the relative length implies that L_{PB} is shorter than L_{M}. On the contrary, the negative value of the relative length implies that L_{M} is shorter than L_{PB}. Moreover, the relative length equal to 0 implies that L_{M} and L_{PB} are equal.

Regarding the empirical coverage probabilities, from Figure 1, the PB procedure maintains the nominal confidence level for all situations. The LL procedure provides a larger than the nominal confidence level for all situations. The TG procedure maintains the nominal confidence level for all situations except for $g = 10$. However, the TG procedure provides a smaller than the nominal confidence level when $\rho < 0.5$ for $g = 10$. The HK procedure mostly maintains the nominal confidence level when $\rho < 0.5$ and it provides a larger than the nominal confidence level when $\rho \geq 0.5$ for $g = 3$. The HK procedure provides a smaller than the nominal confidence level when $\rho < 0.5$ and it provides a larger than the nominal confidence level when $\rho \geq 0.5$ for $g = 6$ and 10. The LXH procedure provides a larger than the nominal confidence level for all ρ for $g = 3$. The LXH procedure provides a smaller than the nominal confidence level for all ρ for $g = 6$ and 10 except in a very unbalanced design (pattern 4), that is, the LXH procedure maintains the nominal confidence level in a very unbalanced design.

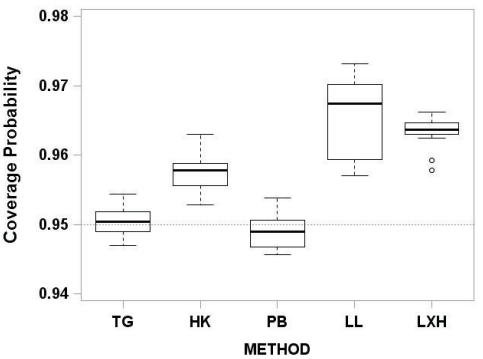
Comparing the average length of the confidence interval, Figure 2 clearly indicates that the average lengths of the TG, LL, and PB intervals behave very similar. The average length of the LXH interval is the shortest. For the number of groups $g = 6$ and 10, the average length of the HK interval is shorter than the average length of the PB interval when $\rho < 0.5$. Conversely, the average length of the PB interval is shorter than the average length of the HK interval when $\rho \geq 0.5$.

5. Application

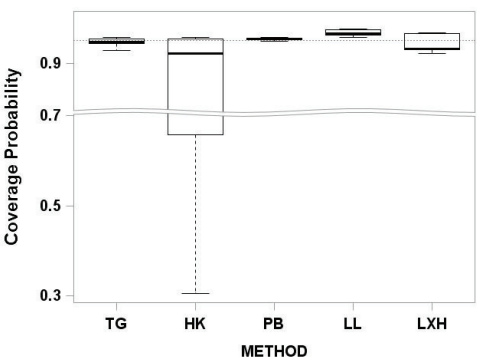
The numerical example from Brownlee (1965) is a study of the effects of environmental conditions on the measure of the ratio of electromagnetic and electrostatic units of electricity. The data set is shown in Table 2. Model (1) is used to des-cribe this data set, that is, $g = 5$, $n_i = (11, 8, 6, 24, 15)$, and $\Phi = 0.796$. Furthermore, a_i denote the random group effects of the environmental conditions and assume $a_i \sim N(0, \sigma_a^2)$, e_{ij} represent the random effect of the jth measure of electricity on the ith environmental condition and assume $e_{ij} \sim N(0, \sigma_e^2)$. Independence among a_i and e_{ij} is also assumed. The five confidence intervals for σ_a^2 based on the five methods in Section 3 are presented in Table 3. Table 3 shows that the PB method provides the shortest confidence interval for this data set.



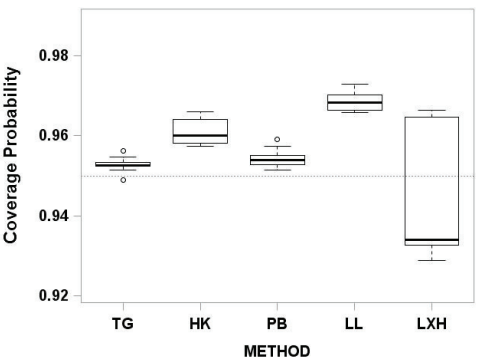
(a) $g = 3$, $\rho < 0.5$



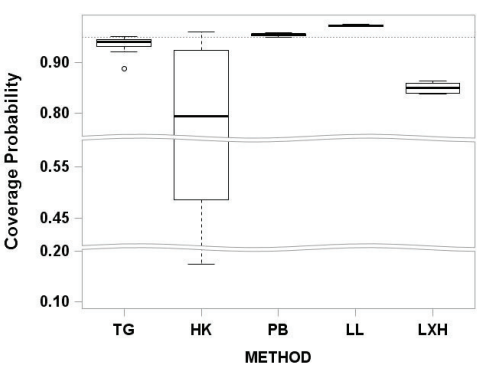
(b) $g = 3$, $\rho \geq 0.5$



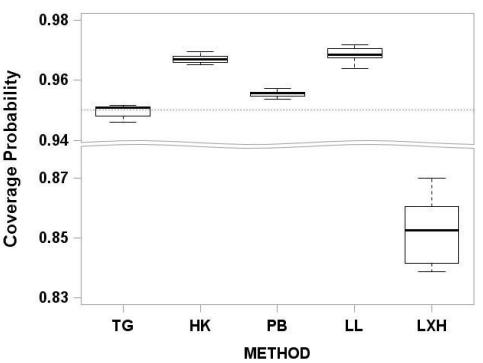
(c) $g = 6$, $\rho < 0.5$



(d) $g = 6$, $\rho \geq 0.5$

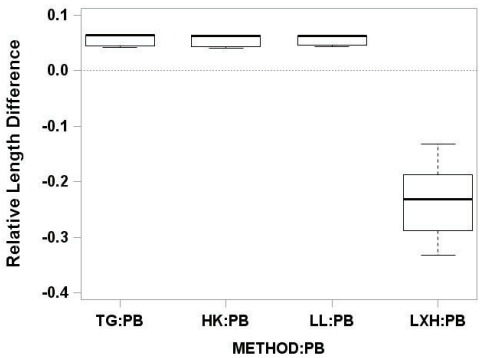


(e) $g = 10$, $\rho < 0.5$

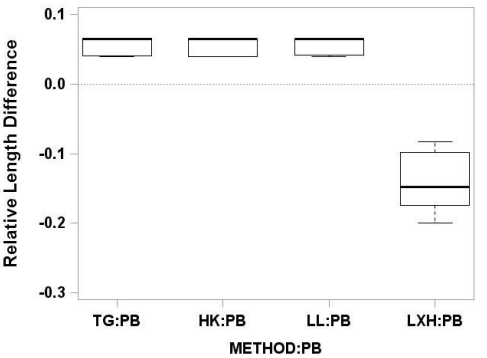


(f) $g = 10$, $\rho \geq 0.5$

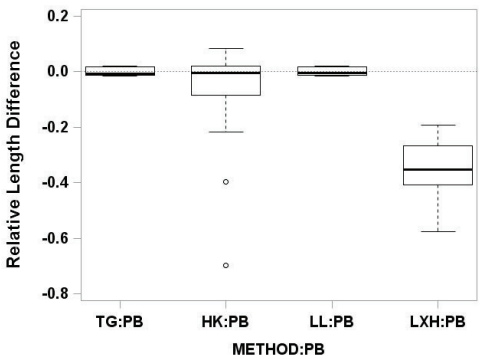
Figure 1: The empirical coverage probabilities of 95% confidence interval for σ_a^2



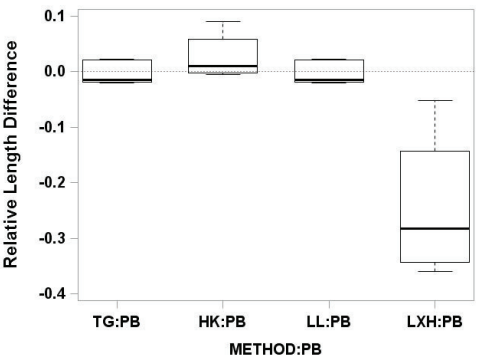
(a) $g = 3$, $\rho < 0.5$



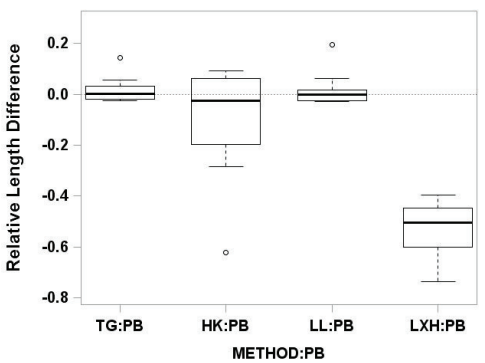
(b) $g = 3$, $\rho \geq 0.5$



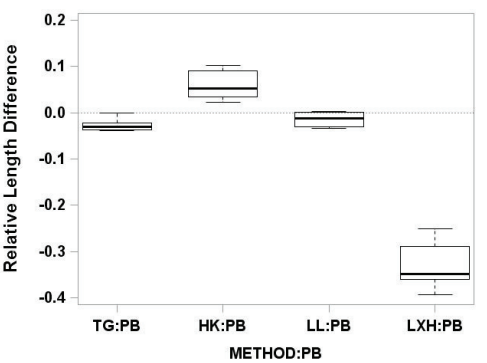
(c) $g = 6$, $\rho < 0.5$



(d) $g = 6$, $\rho \geq 0.5$



(e) $g = 10$, $\rho < 0.5$



(f) $g = 10$, $\rho \geq 0.5$

Figure 2: Relative difference of the average length of 95% confidence interval for σ_a^2

Table 2. The ratio of the electromagnetic to electrostatic units of electricity

Groups	Observations											
1	62	64	62	62	65	64	65	62	62	63	64	
2	65	64	63	62	65	63	64	63				
3	65	64	67	62	65	62						
4	62	66	64	64	63	62	64	64	66	64	66	63
	65	63	63	63	61	56	64	64	65	64	64	65
5	66	65	65	66	67	66	69	70	68	69	63	65
	64	65	64									

Table 3. Nominally 95% confidence interval for the data

Method	TG	HK	PB	LL	LXH
confidence interval	(0, 10.901)	(0, 11.311)	(0, 9.595)	(0, 10.836)	(0, 9.728)

6. Conclusion

This article studies the methods for constructing 95% confidence intervals for variance components in an unbalanced one-way random effects model. Simulation studies indicate that the TG procedure maintains the nominal confidence level for all situations except for the number of group $g = 10$, which is liberal when ρ is small. The HK procedure is conservative when ρ is large. On the contrary, when ρ is small, the HK procedure mostly maintains the nominal confidence level for the number group $g = 3$ and is liberal for the number of groups $g = 6$ and 10. The PB procedure maintains the nominal confidence level for all situations. The LL procedure is conservative for all situations. The LXH procedure is conservative for all ρ in the number of group $g = 3$. Nevertheless, for the number of groups $g = 6$ and 10, the LXH procedure does not adequately maintain the nominal confidence level. All of the average lengths of the confidence intervals behave similarly, but the average length of the LXH interval always has the shortest. Notice that the relative length values of the LXH method is negative.

In summary, the PB and LXH methods are recommended for the number of group $g = 3$. The PB and LL methods are recommended for the number of groups $g = 6$ and 10. The TG and HK methods are useful when ρ is large. Furthermore, the LXH method is preferred in a very unbalanced design.

Acknowledgements

The authors thank the reviewers for their valuable comments and suggestions which substantially improved the quality of the article. This work is supported by the Science Achievement Scholarship of Thailand (SAST).

References

Ahrens, H., Pincus, R., (1981). On two measures of unbalancedness in a one-way model and their relation to efficiency. *Biometrical Journal*, Vol. 23, pp. 227–235.

Arendacká, B., (2005). Generalized confidence intervals on the variance component in mixed linear models with two variance components. *Statistics*, Vol. 39, pp. 275–286.

Brownlee, K. A., (1965). *Statistical theory and methodology in science and engineering*. New York: John Wiley & Sons.

Burch, B. D., (2011). Confidence intervals for variance components in unbalanced one-way random effects model using non-normal distributions. *Journal of Statistical Planning and Inference*, Vol. 141, pp. 3793–3807.

Demetrashvili, N., Wit, E. C., Van Den Heuvel, E. R., (2016). Confidence intervals for intraclass correlation coefficients in variance components models. *Statistical Methods in Medical Research*, Vol. 25, pp. 2359–2376.

Graybill, F. A., (1976). *Theory and application of the linear model*. California: Wadsworth.

Hartung, J., Knapp, G., (2000). Confidence intervals for the between group variance in the unbalanced one-way random effects model of anaylsis of variance. *Journal of Statistical Computation and Simulation*, Vol. 65, pp. 311–323.

Howe, W. G., (1974). Approximate confidence limits on the mean of $X+Y$ where X and Y are two tabled independent random variables. *Journal of the American Statistical Association*, Vol. 69, pp. 789–794.

Lamotte, L. R., (1976). Invariant quadratic estimators in the random, one-way anova model. *Biometrics*, Vol. 32, pp. 793–804.

Li, X., Li, G., (2005). Confidence intervals on sum of variance components with unbalanced designs. *Communications in Statistics—Theory and Methods*, Vol. 34, pp. 833–845.

Li, X., Li, G., (2007). Comparison of confidence intervals on the among group variance in the unbalanced variance component model. *Journal of Statistical Computation and Simulation*, Vol. 77, pp. 477–486.

Lidong, E., Hannig, J., Iyer, H., (2008). Fiducial intervals for variance components in an unbalanced two-component normal mixed linear model. *Journal of the American Statistical Association*, Vol. 103, pp. 854–865.

Liu, X., Xu, X., Hannig, J., (2016). Least squares generalized inferences in unbalanced two-component normal mixed linear model. *Computational Statistics*, Vol. 31, pp. 973–988.

Milliken, G. A., Johnson, D. E., (2009). *Analysis of messy data, volume I: designed experiments*. Boca Raton: CRC Press.

Olsen, A., Seely, J., Birkes, D., (1976). Invariant quadratic unbiased estimation for two variance components. *The Annals of Statistics*, Vol. 4, pp. 878–890.

Park, D. J., Burdick, R. K., (2003). Performance of confidence intervals in regression models with unbalanced one-fold nested error structures. *Communications in Statistics-Simulation and Computation*, Vol. 32, pp. 717–732.

Searle, S. R., Casella, G., Mcculloch, C. E., (2006). *Variance components*. New Jersey: John Wiley & Sons.

Thomas, J. D., Hultquist R. A., (1978). Interval estimation for the unbalanced case of the one-way random effects model. *The Annals of Statistics*, Vol. 6, pp. 582–587.

Ting, N., Burdick, R. K., Graybill, F. A., Gui, R., (1989). One-sided confidence intervals on nonnegative sums of variance components. *Statistics & Probability Letters*, Vol. 8, pp. 129–135.

Ting, N., Burdick, R. K., Graybill, F. A., Jeyaratnam, S., Lu, T. F. C., (1990). Confidence intervals on linear combinations of variance components that are unrestricted in sign. *Journal of Statistical Computation and Simulation*, Vol. 35, pp. 135–143.

Wald, A., (1940). A note on the analysis of variance with unequal class frequencies. *The Annals of Mathematical Statistics*, Vol. 11, pp. 96–100.

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 161–176, DOI 10.2478/stattrans-2022-0048
Received – 22.02.2021; accepted – 10.01.2022

Generalised Lindley shared additive frailty regression model for bivariate survival data

Arvind Pandey [1], **David D. Hanagal** [2], **Shikhar Tyagi**[3]

ABSTRACT

Frailty models are the possible choice to counter the problem of the unobserved heterogeneity in individual risks of disease and death. Based on earlier studies, shared frailty models can be utilised in the analysis of bivariate data related to survival times (e.g. matched pairs experiments, twin or family data). In this article, we assume that frailty acts additively to the hazard rate. A new class of shared frailty models based on generalised Lindley distribution is established. By assuming generalised Weibull and generalised log-logistic baseline distributions, we propose a new class of shared frailty models based on the additive hazard rate. We estimate the parameters in these frailty models and use the Bayesian paradigm of the Markov Chain Monte Carlo (MCMC) technique. Model selection criteria have been applied for the comparison of models. We analyse kidney infection data and suggest the best model.

Key words: Bayesian estimation, frailty, generalised Lindley frailty, generalised log-logistic distribution, generalised Weibull distribution, hazard rate, MCMC, random censoring.

1. Introduction

To analyse the survival data in biological, epidemiological, and medical studies, a common approach is that subjects are supposed to have the same risk of occurrence of an event of interest, which acts multiplicatively. However, this assumption rarely occurs because neither all the covariates can be measured nor can be included in the study due to technical difficulties, time limitations, or financial implications. In real-life situations risk (hazard rate) changes from one family to another family, one group to another group, one cluster to another cluster. Heterogeneity in the population exists, because of the mixture of groups of individuals with different risk factors. This heterogeneity is called frailty. Ignoring frailty may have adverse consequences. A random impact that is an unobservable risk shared by the subject is characterized as frailty, which was introduced by Vaupel et al. (1979). To handle such kinds of problems, many models have been derived in survival analysis. Since the establishment of the proportional hazard model given by Cox (1972), the survival function has been dominated by hazard rate models. The reason behind the popularity of this model is the significance of known covariates that can be tested, also a relationship between lifetimes and covariates can be incorporated.

[1]Department of Statistics, Central University of Rajasthan, Rajasthan, India. E-mail: arvindmzu@gmail.com. ORCID: https://orcid.org/0000-0003-1324-6985.
[2]Department of Statistics, Savitribai Phule Pune University, Pune-411007, India. E-mail: david.hanagal@gmail.com. ORCID: https://orcid.org/0000-0002-8918-6934.
[3]Department of Statistics, Central University of Rajasthan, Rajasthan, India. E-mail: shikhar1093tyagi@gmail.com. ORCID: https://orcid.org/0000-0003-1606-0844.

Research on the bivariate survival models has grown rapidly over the past few years. Clayton's (1978) random effect model of the bivariate survival was a key innovation. He introduced the notion of the shared relative risk. This model was further developed by Oakes (1982) to analyze the association between two non-negative random variables. Hougaard (1985, 1991, 2000) discussed the different aspects of frailty on a broad scale. In the last decade, frailty regression models in mixture distribution were discussed by Hanagal (2008). Modelling kidney infection data for inverse Gaussian shared frailty was done by Hanagal and Pandey (2014a). Gamma frailty models for bivariate survival data were given by Hanagal and Pandey (2015a). Hanagal and Pandey (2017a) used the shared inverse Gaussian frailty models based on additive hazard. Hanagal (2019) gave an extensive literature review on different shared frailty models. Pandey et al. (2020a) presented shared inverse Gaussian frailty models for bivariate findings. Pandey et al. (2020b) looked at generalised inverse Gaussian shared frailty models based on reversed hazard rates. Pandey et al. (2021a, 2022) and Tyagi et al. (2021a) developed distinct Generalised Lindley (GL) shared frailty models based on the reversed hazard rate. Tyagi et al. (2021b, 2022a, 2022b), Gupta et al. (2022), Pandey et al. (2021b), and Pandey and Tyagi (2021) developed inverse weighted Lindley, and GL shared frailty models, respectively. In this article, we assume that frailty acts additively to the hazard rate. The additive hazard models characterize a different facet of the association between covariates and the failure time than the proportional hazard model and are more plausible than the latter for many applications (Lin and Ying, 1994; Bin, 2010). The additive hazard models can be authentically a better alternative to proportional hazard or other nonlinear hazard regression models to narrate the consequences of covariates on survival time (Hosmer and Royston, 2002). When the absolute change in risk, instead of the risk ratio, is of primary interest or when the proportional hazard assumption for the Cox proportional hazard model is violated, an additive hazard regression model may be more appropriate (Xie et al., 2013). Let a continuous random variable T be a lifetime of an individual and the random variable W be frailty variable. The conditional hazard function for a given frailty variable, $W = w$ at time $t \in \mathbb{R}^+$ is

$$\phi(t \mid w) = \phi_0(t) + e^{\underline{K}\underline{\beta}+\underline{V}\underline{\beta}_1} = \phi_0(t) + we^{\underline{K}\underline{\beta}}, w \in \mathbb{R}^+, V \in \mathbb{R}, \tag{1}$$

where $w = e^{\underline{V}\underline{\beta}_1}$ and $\phi_0(t)$ is a baseline hazard function at time $t \in \mathbb{R}^+$, $\underline{K}$ is a row vector of covariates, and $\underline{\beta}$ is a column vector of regression coefficients. The cumulative hazard rate function is given by

$$\Phi(t \mid z) \quad = \quad \Phi_0(t) + wte^{\underline{K}\underline{\beta}}. \tag{2}$$

The conditional survival function for given frailty at time $t \in \mathbb{R}^+$ is

$$S(t \mid w) \quad = \quad e^{-\int_0^t \phi(x|w)dx} = e^{-\left[\Phi_0(t)+wte^{\underline{K}\underline{\beta}}\right]}, \tag{3}$$

where $\Phi_0(t)$ is the cumulative baseline hazard function at time $t \in \mathbb{R}^+$. Integrating over the

range of frailty variable W having density $f(w)$, we get the marginal survival function as

$$S(t) = \int_{w\in\mathbb{R}^+} S(t \mid w) f(w) dw = \int_{w\in\mathbb{R}^+} e^{-\left[\Phi_0(t)+wte^{\underline{K}\underline{\beta}}\right]} f(w) dw, = S_0(t) L_w(te^{\underline{K}\underline{\beta}}), \quad (4)$$

where $L_Z(.)$ is the Laplace transformation of the distribution of Z and $S_0(t)$ is the baseline survival function of T. Once we get the survival function at time $t \in \mathbb{R}^+$, of lifetime random variable for an individual, we can obtain the probability structure and make its inferences based on it.

The main objective of this article is threefold. First, generalised Lindley (GL) shared frailty models for additive hazard rate with generalised Weibull and generalised log-logistic as baseline distributions have been introduced. Second, the Bayesian approach of estimation has been employed to estimate the unknown parameters under random censoring. Third, a simulation study and data analysis have been done for the Kidney infection data set.

2. General Shared Frailty Model

The shared frailty models are applicable to event time of the related individuals, similar organs, and repeated measurements. In this model individuals from a group shares common covariates. It has been considered that survival times are conditionally independent, for a given shared frailty. Shared frailty indicates dependence between survival times is only because of unobservable covariates (frailty). Frailty variable W has a degenerate distribution in the absence of variability. If the dependence is positive, the distribution of W is not degenerate.

Assume n individuals are considered under the study. Bivariate random variables (T_{1j}, T_{2j}) are postulated as the first and the second survival times of the j^{th} individual $(j = 1,2,3,\ldots,n)$. Also m known covariates are supposed to be collected in a vector $\underline{K}_j = (K_{1j},\ldots,K_{mj})$ for the j^{th} individual where K_{aj} $(a = 1,2,3,\ldots,m)$ represents the value of the a^{th} observed covariate for the j^{th} individual. Under shared frailty model, it has been presupposed that both survival times for everyone share the similar value of the covariates. Let W_j be shared frailty for the j^{th} individual. Assuming that the frailties are acting additively on the baseline hazard function and both the survival times of individuals are conditionally independent for given frailty, the conditional hazard function for the j^{th} individual at the i^{th} $(i = 1,2)$ survival time $t_{ij} \in \mathbb{R}^+$ for given frailty $W_j = w_j$ has the form

$$\phi(t_{ij} \mid W_j, \underline{K}_j) \quad = \quad \phi_0(t_{ij}) + w_j e^{\underline{K}_j\underline{\beta}},$$

where $\phi_0(t_{ij})$ is the baseline hazard at time $t_{ij} \in \mathbb{R}^+$ and $\underline{\beta}$ is a vector of order m, of the regression coefficients. The conditional cumulative hazard function for the j^{th} individual at the i^{th} survival time $t_{ij} \in \mathbb{R}^+$ for a given frailty $W_j = w_j$ is

$$\Phi(t_{ij} \mid w_j, \underline{X}_j) \quad = \quad \Phi_0(t_{ij}) + w_j t_{ij} \rho_j,$$

where $\rho_j = e^{\underline{K}_j\underline{\beta}}$ and $\Phi_0(t_{ij})$ is the cumulative baseline hazard function at time $t_{ij} \in \mathbb{R}^+$. The conditional survival function for the j^{th} individual at the i^{th} survival time $t_{ij} \in \mathbb{R}^+$ for

a given frailty $W_j = w_j$ is

$$S(t_{ij} \mid w_j, \underline{K}_j) = e^{-\Phi(t_{ij}|w_j,\underline{K}_j)} = e^{-\left[\Phi_0(t_{ij})+w_j t_{ij}\rho_j\right]}.$$

Under the assumption of independence, the bivariate conditional survival function for a given frailty, $W_j = w_j$ at time $t_{1j} \in \mathbb{R}^+$ and $t_{2j} \in \mathbb{R}^+$ is

$$S(t_{1j},t_{2j} \mid w_j, \underline{K}_j) = S(t_{1j} \mid w_j, \underline{K}_j)S(t_{2j} \mid w_j, \underline{K}_j) = e^{-\left[(\Phi_{01}(t_{1j})+\Phi_{02}(t_{2j}))+w_j(t_{1j}+t_{2j})\rho_j\right]}.$$

The unconditional bivariate survival function at time $t_{1j} \in \mathbb{R}^+$ and $t_{2j} \in \mathbb{R}^+$ can be obtained by integrating over the frailty variable W_j having the probability function $f_W(w_j)$, for the j^{th} individual

$$S(t_{1j},t_{2j} \mid \underline{K}_j) = \int_{W_j \in \mathbb{R}^+} S(t_{1j},t_{2j} \mid w_j) f_W(w_j) dw_j = e^{-(\Phi_{01}(t_{1j})+\Phi_{02}(t_{2j}))} L_{W_j}[(t_{1j}+t_{2j})\rho_j], \tag{5}$$

where $L_{Z_j}(.)$ is the Laplace transform of the frailty variable of W_j for the j^{th} individual. Here onwards we represent $S(t_{1j},t_{2j} \mid \underline{K}_j)$ as $S(t_{1j},t_{2j})$.

3. Generalised Lindley Frailty Model

Lindley (1958) proposed a distribution with one parameter. Because of having only one parameter, the Lindley distribution does not provide enough flexibility for modelling purposes. It will be useful to consider further alternatives of this distribution. For a frailty distribution, a new generalised Lindley distribution has been considered in this paper. This distribution is the mixture of two gamma distributions G(θ,μ) and G(θ,η) with mixing coefficient $\theta/(\theta+1)$ (Elbatal, et al. (2013)). Because of the mixture of two gamma densities, a slight suppleness can be seen during analysis of time to event data. That is the reason why the GL frailty model is more adaptable in comparison with the gamma frailty model. the probability density function of GL distribution has been specified below:

$$f_W(w) = \begin{cases} \frac{1}{(1+\theta)}\left[\frac{\theta^{\mu+1}w^{\mu-1}}{\Gamma\mu} + \frac{\theta^{\eta}w^{\eta-1}}{\Gamma\eta}\right]e^{-\theta w} & ;w \in \mathbb{R}^+, \mu, \eta, \theta \in \mathbb{R}^+ \\ 0 & ;otherwise, \end{cases}$$

with mean $E[W] = \frac{1}{1+\theta}\left[\mu + \frac{\eta}{\theta}\right]$. And corresponding variance is,

$$V(W) = \frac{1}{(1+\theta)}\left[\left(\mu^2 + \frac{\eta^2}{\theta}\right)\left(\frac{1}{\theta(1+\theta)}\right) + \left(\frac{\mu+\eta}{\theta}\right) - \left(\frac{2\mu\eta}{\theta(1+\theta)}\right)\right],$$

after applying identifiability property, i.e., $E[W] = 1$ we get a relation between parameters $\eta = \theta(1+\theta-\mu)$. Consequently, the density function, the Laplace transformation and variance for GL are reduced to

$$f_W(w) = \begin{cases} \frac{1}{(1+\theta)}\left[\frac{\theta^{\mu+1}w^{\mu-1}}{\Gamma\mu} + \frac{\theta^{\theta(1+\theta-\mu)}w^{\theta(1+\theta-\mu)-1}}{\Gamma\theta(1+\theta-\mu)}\right]e^{-\theta w} & ;\ w, \theta \in \mathbb{R}^+, \mu \in (0, 1+\theta) \\ 0 & ;\ otherwise. \end{cases}$$

$$L_W(s) = \frac{1}{(1+\theta)}\left[\frac{\theta^{\mu+1}}{(s+\theta)^{\mu}} + \frac{\theta^{\theta(1+\theta-\mu)}}{(s+\theta)^{\theta(1+\theta-\mu)}}\right], \tag{6}$$

$$V(W) = \frac{\theta^4 - \theta^3\mu + 3\theta^2(1+\theta) - 4\theta^2\mu + 3\theta\mu(\mu-1) + \mu^2}{\theta(1+\theta)^2}. \tag{7}$$

Let n be the number of observations under study. Let (T_{1j}, T_{2j}) be the first and second survival times of pairs of components of j^{th} $(1,2,...,n)$ objects. The unconditional bivariate survival function at time $t_{1j} \in \mathbb{R}^+$ and $t_{2j} \in \mathbb{R}^+$ using equations (5) and (6) can be written as

$$S(t_{1j}, t_{2j}) = \frac{e^{-(\Phi_{01}(t_{1j})+\Phi_{02}(t_{2j}))}}{(1+\theta)}\left[\frac{\theta^{\mu+1}}{(\theta+\rho(t_{1j}+t_{2j}))^{\mu}} + \frac{\theta^{\theta(1+\theta-\mu)}}{(\theta+\rho(t_{1j}+t_{2j}))^{\theta(1+\theta-\mu)}}\right], \tag{8}$$

where $\Phi_{01}(t_{1j})$, $\Phi_{02}(t_{2j})$ are the cumulative baseline hazard rate functions of the lifetime T_{1j} and T_{2j}, respectively. One can have different baseline distributions for T_1 and T_2. After substituting different cumulative hazard functions in (8), we get different generalised Lindley frailty distributions.

4. Dependence Measure

Sometimes due to complex form of frailty models, it is difficult to compare the degree of dependence between different frailty models. Kendall's τ can be used to quantify dependence because it is independent of transformations on the time scale and the frailty model used. It is a rank-based dependence measure.

$$\tau = \int_{s\in\mathbb{R}^+} 4sL_W^{"}(s)L_W(s)ds - 1. \tag{9}$$

After using equation (8) and (9), we get,

$$\tau = \int_{s\in\mathbb{R}^+} R(s \mid \theta,\mu)ds - 1, \tag{10}$$

where $R(s \mid \theta,\mu) = \frac{4\theta s\left(\theta^{\mu+1}A^{-\mu}+\theta^{\theta B}A^{\theta(\mu-\theta-1)}\right)\left(\mu(\mu+1)\theta^{\mu}A^{-\mu}+\theta^{\theta B}B\left(-\mu\theta+\theta^2+\theta+1\right)A^{\theta(\mu-\theta-1)}\right)}{(\theta+1)^2A^2}$.
$A = (\theta+s)$, $B = (1+\theta-\mu)$.
Kendall's τ cannot be found in closed form for GL frailty. Some numerical approaches can be utilized to obtain Kendall's τ dependence measure.

5. Baseline Distributions

5.1. Generalised Weibull Distribution

Here, the generalised Weibull distribution has been postulated as a baseline distribution. If a continuous random variable $T \in \mathbb{R}^+$ follows the generalised Weibull distribution then the survival and cumulative hazard function, are respectively,

$$S(t) = \begin{cases} 1-\left(1-e^{-\delta t^{\xi}}\right)^{\zeta} & ;t \in \mathbb{R}^+, \delta, \zeta, \xi \in \mathbb{R}^+ \\ 1 & ;otherwise \end{cases} \tag{11}$$

$$\Phi_0(t) = \begin{cases} -\log\left(1-\left(1-e^{-\delta t^{\xi}}\right)^{\zeta}\right) & ;t \in \mathbb{R}^+, \delta, \zeta, \xi \in \mathbb{R}^+ \\ 0 & ;otherwise \end{cases} \tag{12}$$

5.2. Generalised log-logistic distribution

Bacon (1993) used the log-logistic distribution for modelling saturation effects. The survival function of the log-logistic distribution is given by,

$$S(t) = (1+\delta t^{\xi})^{-1} \tag{13}$$

Due to having heavier tail in camparison to the gamma distribution, the log-logistic distribution can be more beneficial to be used for finance and insurance variables. The log-logistic distribution provides two parametric models for the survival analysis. Unlike the more commonly used Weibull distribution, it can have a non-monotonic hazard function: when $\xi > 1$ the hazard function is unimodal (when $\xi \leq 1$, the hazard decreases monotonically). The fact that the cumulative distribution function can be written in the closed form is particularly useful for the analysis of the survival data with censoring.

Lehmann family (Deshpande and Purohit, 2005) is a very useful family of life distributions generated from a given survival function and extensively used to model the effect of covariates. Let $S_0(t)$ be an arbitrary known survival function. If ζ is positive, then

$$S(t) = (S_0(t))^{\zeta} \tag{14}$$

is also a survival function. If, in particular, ζ is the positive integer n, then it represents the survival function of $min(X_1, ..., X_n)$ where X_i's are i.i.d. random variables with $S_0(t)$ as the common survival function. The hazards are proportional ζ times. Lehmann family is also known as the proportional hazards family. We use the same property and obtain, the survival function and the cumulative hazard rate as follows.

$$S(t) = \begin{cases} (1+\delta t^{\xi})^{-\zeta} & ;t \in \mathbb{R}^+, \delta, \zeta, \xi \in \mathbb{R}^+ \\ 1 & ;otherwise \end{cases} \tag{15}$$

$$\Phi_0(t) = \begin{cases} \zeta \log(1+\delta t^{\xi}) & ;t \in \mathbb{R}^+, \delta, \zeta, \xi \in \mathbb{R}^+ \\ 0 & ;otherwise \end{cases} \quad (16)$$

6. Proposed model

Due to group variation or frailty and individual variation described by the hazard function, a shared frailty model can be considered as a mixture model in survival analysis. Distribution of W is convergent, if dependence is positive. After substituting a cumulative hazard function for generalised Weibull and generalised log-logistic baseline distributions in equation (8)

$$S(t_{1j},t_{2j}) = \frac{1}{(1+\theta)}\left[\frac{\theta^{\mu+1}}{(\theta+\rho(t_{1j}+t_{2j}))^{\mu}} + \frac{\theta^{\theta(1+\theta-\mu)}}{(\theta+\rho(t_{1j}+t_{2j}))^{\theta(1+\theta-\mu)}}\right] \prod_{i=1}^{2}\left(1-\left(1-e^{-\delta_i t_{ij}^{\xi_i}}\right)^{\zeta_i}\right), \quad (17)$$

$$S(t_{1j},t_{2j}) = \frac{1}{(1+\theta)}\left[\frac{\theta^{\mu+1}}{(\theta+\rho(t_{1j}+t_{2j}))^{\mu}} + \frac{\theta^{\theta(1+\theta-\mu)}}{(\theta+\rho(t_{1j}+t_{2j}))^{\theta(1+\theta-\mu)}}\right] \prod_{i=1}^{2}(1+\delta_i t_{ij}^{\xi_i})^{-\zeta_i}, \quad (18)$$

here, equations (17), (18) can be called Model-I, Model-II respectively, which have been established for generalised Weibull and generalised log-logistic baseline distributions.

7. Likelihood Design and Bayesian Paradigm

For the study, n individuals have been considered. Observed failure times have been indicated by (t_{1j},t_{2j}). We are using the random censoring scheme. Let censoring time is indicated by c_{1j} and c_{2j} for j^{th} individual $(j = 1,2,3,...,n)$. Independence between censoring schemes and lifetimes of individuals has been presumed. The probability density function can be described for bivariate lifetime random variable of the j^{th} individual as

$$f_j(t_{1j},t_{2j}) = \begin{cases} f_1(t_{1j},t_{2j}), & ; t_{1j} < c_{1j}, t_{2j} < c_{2j}, \\ f_2(t_{1j},c_{2j}), & ; t_{1j} < c_{1j}, t_{2j} > c_{2j}, \\ f_3(c_{1j},t_{2j}), & ; t_{1j} > c_{1j}, t_{2j} < c_{2j}, \\ f_4(c_{1j},c_{2j}), & ; t_{1j} > c_{1j}, t_{2j} > c_{2j}. \end{cases}$$

The likelihood function will be

$$L(\underline{\Theta},\underline{\beta},\theta,\mu) = \prod_{j=1}^{n_1} f_1(t_{1j},t_{2j}) \prod_{j=1}^{n_2} f_2(t_{1j},c_{2j}) \prod_{j=1}^{n_3} f_3(c_{1j},t_{2j}) \prod_{j=1}^{n_4} f_4(c_{1j},c_{2j}), \quad (19)$$

where $\underline{\Theta}$, $\underline{\beta}$, θ and μ are the vector of baseline parameters and the vector of regression coefficients and frailty parameters respectively. Let n_1, n_2, n_3, and n_4 be the number of pairs for which the first and the second failure times (t_{1j}, t_{2j}) lie in the ranges $t_{1j} < c_{1j}, t_{2j} < c_{2j}$; $t_{1j} < c_{1j}, t_{2j} > c_{2j}$; $t_{1j} > c_{1j}, t_{2j} < c_{2j}$, and $t_{1j} > c_{1j}, t_{2j} > c_{2j}$ respectively and let

$$f_1(t_{1j}, t_{2j}) = \frac{\partial^2 S(t_{1j}, t_{2j})}{\partial t_{1j} \partial t_{2j}}; f_2(t_{1j}, c_{2j}) = -\frac{\partial S(t_{1j}, c_{2j})}{\partial t_{1j}},$$
$$f_3(c_{1j}, t_{2j}) = -\frac{\partial S(c_{1j}, t_{2j})}{\partial t_{2j}}; f_4(c_{1j}, c_{2j}) = S(c_{1j}, c_{2j}). \tag{20}$$

Substituting cumulative hazard rates $\Phi_{01}(t_{1j})$ and $\Phi_{02}(t_{2j})$ and survival function $S(t_{1j}, t_{2j})$ in equation (29) for Model-I and Model-II and by differentiating we get the likelihood function. The maximum likelihood method has crucial importance in computing efficient estimators. Inappropriately, due to a convergence problem, maximum likelihood failed to estimate the parameters, because Model-I and Model-II have thirteen-dimensional optimization problems. The Bayesian scenario has been discussed by several researchers for estimating parameters of the frailty models. For gamma and log-normal frailty models, the Bayesian paradigm has been contemplated by Santos and Achcar (2010). Weibull and piecewise exponential models have been discussed by Ibrahim et al. (2001) with gamma frailty. The joint posterior density function of parameters for given failure times is obtained as

$$\pi(\Theta, \theta, \mu, \underline{\beta_0}) \propto L(\Theta, \mu, \underline{\beta_0}) g_1(\zeta) g_2(\xi) g_3(\delta) g_4(\theta) g_5(\mu) \prod_{i=1}^{5} p_i(\beta_{0i\times 1})$$

where $g_i(.)$ indicates the prior density function with known hyperparameters of the corresponding argument for baseline parameters and frailty variance; $p_i(.)$ is prior density function for regression coefficient β_{0i} and the likelihood function is $L(.)$. An important assumption here is that all the parameters are independently distributed. In a similar way, the joint posterior density function can be written for without frailty models. To estimate the parameters of the models, hybrid Metropolis-Hastings algorithms have been used. The Geweke test (see Geweke, 1992) and Gelman-Rubin (see Gelman and Rubin, 1992) statistics have been used to monitor the convergence of a Markov chain to a stationary distribution.
Due to the high dimensions of conditional distributions, it is difficult to integrate out. Thus, it has been considered that full conditional distributions can be obtained as they are proportional to the joint distribution of the parameter of the model. The conditional distribution for single parameter δ with frailty is obtained as

$$\psi_1(\delta \mid \xi, \zeta, \theta, \mu, \underline{\beta_0}) \quad \propto \quad L(\delta, \xi, \zeta, \theta, \mu, \underline{\beta_0}) \cdot g_1(\delta) \tag{21}$$

the conditional distribution for single parameter δ without frailty is obtained as

$$\psi_1(\delta \mid \xi, \zeta, \underline{\beta_0}) \quad \propto \quad L(\delta, \xi, \zeta, \underline{\beta_0}) \cdot g_1(\delta).$$

Similarly, the conditional distributions for other parameters can be obtained.

8. Simulation Study

A simulation study has been executed to appraise the Bayesian estimation paradigm for Model-I and Model-II. Single covariate K_1 has been considered folliwng normal distribution. The frailty variable W is assumed to follow generalised Lindley distribution. Independence between lifetimes of individuals has been considered. Samples are generated using the subsequent mechanism,

1. From the binomial distribution with probability 0.6, 25 values for K_1 has been generated.

2. For known covariates, compute $\rho = e^{K_1\beta_1}$.

3. The distribution of lifetimes follow generalised Weibull and generalised log-logistic baseline distributions for given frailty W_j. 25 values of lifetimes have been generated. The conditional survival function for lifetime t_j $(j = 1,2,...,n)$ for given frailty $W_j = w_j$ and covariate K_1 is

$$S(t_j \mid w_j, K_1) \quad = \quad e^{-(\Phi_0(t_j)+w_j t\rho)}$$

Equating $S(t_j \mid w_j, K_1)$ to random number, say $v_j (0 < v_j < 1)$ generated from $U(0,1)$ over $t_j \in \mathbb{R}^+$ we get:
for Model-I and Model-II

$$v_j \quad = \quad \left(1 - \left(1 - e^{-\delta t_j^{\xi}}\right)^{\zeta}\right) * e^{-wt_j\rho},$$

$$v_j \quad = \quad (1+\delta t_j^{\xi})^{-\zeta} * e^{-wt_j\rho} \quad \text{respectively.}$$

4. Censoring time c_j has been generated from $G(0.9, 0.01)$ for Model-I.

5. Observe the j^{th} survival time $t_j^* = min(t_j, c_j)$ and the censoring indicator χ_j for the j^{th} individual $(j = 1,2,...,25)$ where

$$\chi_j = \begin{cases} 1, & ; t_j < c_j \\ 0, & ; t_j > c_j \end{cases}$$

thus we have data consisting of 25 pairs of survival times t_j^* and the censoring indicator χ_j.

Concurrently, with different priors and starting points, two chains have been operated. Both chains were recapitulated 100,000 times. Gelman-Rubin test values are very close to one. Due to small values of Geweke test statistic and corresponding p-values, the chains reach stationary distribution for both prior sets. The estimates of parameters were the same for both the priors, no impact of prior distributions has been found on posterior summaries. Here, the analysis for one chain has been exhibited because both the chains have shown the

same results. Tables 1 and 2 present the estimates and the credible intervals of the parameters for Models I and II based on the simulation study. The Gelman-Rubin convergence statistic values are nearly equal to one and also the Geweke test values are quite small, and the corresponding p-values are large enough to say that the chain attains stationary distribution.

9. Applicability on Kidney Infection Data

To elucidate the Bayesian estimation paradigm, kidney infection data of McGilchrist and Aisbett (1991) have been considered. This data consists of 38 patients and recurrence times (in days) of infection are given. Table 3 gives the p-values of goodness of fit test for Model I and Model II. Consequently, on the basis of p-values of K-S test it is clear that there is no statistical evidence to reject the hypothesis that data are from Model I and Model II in the marginal case and it can be assumed that they also fit for bivariate case. For frailty parameters, gamma distribution with very small shape and scale parameters (say, 0.0001) has been used. Additionally, it can be considered that regression coefficients are normally distributed with mean zero and high variance (say 1000). A similar type of prior was used in Ibrahim et al. (2001) and Santos and Achcar (2010). Thus for frailty parameters θ, μ and regression coefficients $\beta_{0i}, i = 1, ..., 5$, vague priors have been used. Because of no information about the baseline parameter, the prior distribution corresponding to baseline parameters is also considered flat. We considered two different vague prior distributions for baseline parameters, one is gamma distribution with shape and scale hyperparameters $\varepsilon_1, \varepsilon_2$ respectively and another is uniform distribution with interval (ν_1, ν_2). All the hyperparameters are known. Under the Bayesian paradigm, for both models, two parallel chains have been run. Also, two sets of prior distributions have been used with different starting points using the hybrid Metropolis-Hastings algorithm based on normal transition kernels. It can be said that estimates are independent of the different prior distributions because, for both sets of priors, estimates of parameters are approximately similar. We got an almost similar convergence rate of the Gibbs sampler for both sets of priors. Here, the analysis for one chain has been exhibited because both the chains have shown the same results. The Gelman-Rubin convergence statistic values are closely equal to one. The Geweke test statistic values are somewhat small, and the corresponding p-values are large enough to say that the chains reach stationary distribution. Tables 4-5 contained the values of posterior mean and the standard error with 95% credible intervals, the Gelman-Rubin statistics values, and the Geweke test with p-values for Model I and II. The AIC, BIC, and DIC values, given in Table 7, have been used to compare both models. Model-I holds the lowest possible values of AIC, BIC, and DIC. For Model-I and Model-II, the credible interval of all regression coefficients does not contain zero. It indicates that all covariates have a significant effect on both models. With a negative value, it is being indicated that age, sex, disease PKD are significant factors for kidney infection, having negative effects. But, with positive value diseases, GN and AN have a positive significant effect with a higher chance of infection. It is observed that female patients have a lower risk of kidney infection as compared to male counterparts.

10. Conclusions

A generalised Lindley additive frailty model under generalised Weibull and generalised log-logistic baseline distributions has been proposed. To fit the proposed model the hybrid M-H algorithm has been applied. Analysis has been done in R statistical software with self-written programs. The value of both frailty parameters for Model-I ($\theta = 2.29258, \mu = 1.38391$) and Model-II ($\theta = 2.12060, \mu = 1.28878$) is very high and corresponding variances are 1.434811 and 1.36565 by using equation (3.2). In Table 6, we calculate Kendall's τ measure of dependence by using equation (4.2). All these values are large enough to exhibit that there is a strong indication of heterogeneity among the patients in the population for the data set. To take the decision about all models, different tools have been utilized. With the lowest value of AIC, BIC, and DIC, from Table 7, and the value of Bayes factor for Model-I against Model-II (1.122368), it can be said that Model-I is better than the Model-II to analyze kidney infection data. For kidney infection data, all the covariates have been found statistically significant factors for both models (see Tables 4-5). Our proposed frailty models, Model-I and Model-II, are better as compared to the frailty models by Hanagal et al. (2017) and Hanagal and Pandey (2017a) with baseline generalised log-logistic distribution. In a similar way, with a minimum value of AIC, our proposed frailty models are better as compared to the frailty models by Pandey et al. (2018).

References

Bacon, R. W., (1993). A note on the use of the log-logistic functional form for modeling saturation effects. *Oxford Bulletin of Economics and Statistics*, 55, pp. 355–361.

Bin, H., (2010). Additive hazards model with time-varying regression coefficients. *Acta Math. Sci.*, 30B(4), pp. 1318–1326.

Clayton's, D. G., (1978). A model for association in bivariate life tables and its applications to epidemiological studies of familial tendency in chronic disease incidence. *Biometrica*, 65, pp. 141–151.

Cox, D. R., (1972). Regression Models and Life Tables (with Discussion). *Journal of Royal Statistical Society*, Series B, 34, pp. 187–220.

Deshpande, J. V., Purohit, S. G., (2005). Life Time Data: Statistical Models and Methods. *World Scientific*, New Jersey.

Elbatal, I., Merovci, F., & Elgarhy, M., (2013). A new generalized Lindley distribution. *Mathematical theory and Modeling*, 3(13), pp. 30–47.

Gelman, A., Rubin, D. B., (1992). A single series from the Gibbs sampler provides a false sense of security. In Bayesian Statistics 4 (J. M. Bernardo, J. 0. Berger, A. P. Dawid

and A. F. M. Smith, eds.). *Oxford University Press*. pp. 625–632.

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. Berger, A. P. Dawid and A. F. M. Smith), *Oxford University Press*, pp. 169–193.

Gupta, P., Pandey, A., and Tyagi, S., (2022). Comparison of Multiplicative Frailty Models under Generalized Log-Logistic-II Baseline Distribution for Counter Heterogeneous Left Censored Data, 1, pp. 97–114.

Hanagal, D. D., (2008a). Frailty regression models in mixture distributions. *Journal of Statistical Planning and Inference*, 138(8), pp. 2462–68.

Hanagal, D. D. (2019). Modeling Survival Data Using Frailty Models. 2nd Edition. *Springer*, Singapore.

Hanagal, D. D., & Pandey, A., (2014a). Inverse Gaussian shared frailty for modeling kidney infection data. *Advances in Reliability*, 1, pp. 1–14.

Hanagal, D. D., & Pandey, A., (2015a). Gamma frailty models for bivariate survival data. *Journal of Statistical Computation and Simulation*, 85(15), pp. 3172–3189.

Hanagal, D. D., Pandey, A., & Ganguly, A., (2017). Correlated gamma frailty models for bivariate survival data. *Communications in Statistics-Simulation and Computation*, 46(5), pp. 3627–3644.

Hanagal, D. D., & Pandey, A., (2017a). Shared inverse Gaussian frailty models based on additive hazards. *Communications in Statistics-Theory and Methods*, 46(22), pp. 11143–11162.

Hosmer, D. W., Royston, P., (2002). Using Aalen's linear hazards model to investigate time varying effects in the proportional hazards regression model. *Stat Journal*, 2(4), pp. 331–350.

Hougaard, P., (1985). Discussion of the paper by D.G. Clayton and J. Cuzick. *Journal of the Royal Statistical Society*, A, 148, pp. 113–14.

Hougaard, P., (1991). Modeling heterogeneity in survival data. *Journal of Applied Probability*, 28, pp. 695–701.

Hougaard, P., (2000). Analysis of Multivariate Survival Data. *Springer*, New York.

Ibrahim, J. G., Ming-Hui C. and Sinha, D., (2001). Bayesian Survival Analysis. *Springer*, Verlag.

Johnson, N. L., Kotz, S., (1975). A vector valued multivariate hazard rate. *Journal of Multivariate Analysis*, 5 (1), pp. 53–66. doi:10.1016/0047-259X(75)90055-X.

Lin, D. Y., Ying, Z., (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81(1), pp. 61–71.

Lindley, D. V., (1958). Fiducial distributions and Bayes's theorem. *Journal of the Royal Statistical Society*, B, 20, pp. 102–107.

McGilchrist, C. A., Aisbett, C. W.,(1991): Regression with frailty in survival analysis. *Biometrics*, 47, pp. 461–466.

Oakes, D.,(1982). Bivariate Survival Models Induced by Frailties. Journal of the American Statistical Association, 84(406), pp. 487–493.

Pandey, A., Lalpawimawha, L., & Bhushan, S., (2018). Additive shared inverse Gaussian frailty model. *Pak. J. Statist*, 34(4), pp. 311–330.

Pandey, A., Bhushan, S., Pawimawha, L., and Tyagi, S., (2020a).Analysis of Bivariate Survival Data using Shared Inverse Gaussian Frailty Models: A Bayesian Approach, Predictive Analytics Using Statistics and Big Data: Concepts and Modeling, *Bentham Books*, 14, pp. 75–88.

Pandey, A., Hanagal, D. D., Gupta, P., & Tyagi, S., (2020b). Analysis of Australian Twin Data Using Generalized Inverse Gaussian Shared Frailty Models Based on Reversed Hazard Rate. *International Journal of Statistics and Reliability Engineering*, 7(2), pp. 219–235.

Pandey, A., & Tyagi, S., (2021). Comparison of Multiplicative Frailty Models Under Weibull Baseline Distribution. *Lobachevskii Journal of Mathematics*, 42(13), pp. 3184–3195.

Pandey, A., Hanagal, D. D., Tyagi, S., and Gupta, P., (2021a). Generalized Lindley Shared Frailty Based on Reversed Hazard Rate. *International Journal of Reliability*, Quality and Safety Engineering, 2150040.

Pandey, A., Hanagal, D. D., and Tyagi, S., (2021b). Generalized Lindley Shared Frailty Models. *Statistics and Applications*, 19(2), pp. 41–62.

Pandey, A., Hanagal, D. D., Tyagi, S., & Gupta, P., (2022). Modeling Australian Twin Data Using Generalized Lindley Shared Frailty Models. *In Annual Conference of the Society of Statistics, Computer and Applications*, pp. 143–169. *Springer*, Singapore.

Santos, C. A., Achcar, J. A.,(2010). A Bayesian analysis for multivariate survival data in the presence of covariates. *Journal of Statistical Theory and Applications*, 9, pp. 233–253.

Shaked, M., Shantikumar, J. G. (1994). Stochastic Orders and Their Applications. *Academic Press*, New York.

Tyagi, S., Pandey, A., Hanagal, D. D., and Gupta, P.,(2021a). Bayesian inferences in generalized Lindley shared frailty model with left censored bivariate data. *Advance Research Trends in Statistics and Data Science*, pp. 137–157.

Tyagi, S., Pandey, A., Agiwal, V., and Chesneau, C.,(2021b). Weighted Lindley multiplicative regression frailty models under random censored data. *Computational and Applied Mathematics*, 40(8), pp. 1-24.

Tyagi, S., Pandey, A. & Chesneau, C., (2022a).Identifying the Effects of Observed and Unobserved Risk Factors Using Weighted Lindley Shared Regression Model. *J Stat Theory Pract 16*, 16, https://doi.org/10.1007/s42519-021-00241-9.

Tyagi, S., Pandey, A. & Chesneau, C.,(2022b). Weighted Lindley Shared Regression Model for Bivariate Left Censored Data. *Sankhya B.*, https://doi.org/10.1007/s13571-022-00278-1.

Vaupel, J. W., Manton, K. G. and Stallaed, E., (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, pp. 439–454.

Xie, X., Strickler,H.D., Xue, X., (2013). Additive hazard regression models: An application to the natural history of human papillomavirus. *Comput. Math. Methods Med.*, pp. 1–7.

Appendix

Tables and Figures

Table 1: Posterior Summary of Generalised Lindley Frailty with Baseline Generalised Weibull (Simulation Study: Model-I)

Parameter	*Estimate*	*s.e.*	*L.C.L.*	*U.C.L.*	*Geweke test*	$p-value$	*Gelman Rubin test*
burn-in period = 6900;		autocorrelation lag = 400					
$\zeta_1(22.5)$	23.07359	1.34778	20.32034	25.68016	0.00151	0.50060	1.01140
$\delta_1(0.013)$	0.01418	0.00291	0.00919	0.01884	-0.00694	0.49723	1.00058
$\xi_1(0.35)$	0.35559	0.03256	0.29407	0.42018	0.00562	0.50224	1.00174
$\zeta_2(22.5)$	22.77481	3.04714	17.77885	27.43331	-0.00067	0.49973	1.00054
$\delta_2(0.013)$	0.01395	0.00287	0.00916	0.01865	-0.00405	0.49839	0.99996
$\xi_2(0.33)$	0.33549	0.03187	0.27242	0.40088	-0.00716	0.49714	0.99996
$\theta(2.8)$	2.57442	0.50458	1.83115	3.55436	0.00639	0.50255	1.00289
$\mu(1.5)$	1.51839	0.18151	1.21484	1.93333	0.00094	0.50037	1.00154
$\beta_1(0.15)$	0.12616	0.06989	-0.00552	0.26851	-0.00119	0.49953	0.99996

Table 2: Posterior Summary of Generalised Lindley Frailty with Baseline Generalised Log-Logistic-II (Simulation Study: Model-II)

Parameter	*Estimate*	*s.e.*	*L.C.L.*	*U.C.L.*	*Geweke test*	$p-value$	*Gelman Rubin test*
burn-in period = 6900;		autocorrelation lag = 400					
$\zeta_1(4.5)$	4.42991	0.32462	3.78131	5.04944	0.00185	0.50074	1.00044
$\delta_1(0.02)$	0.02039	0.00291	0.01514	0.02481	-0.00352	0.49860	1.06288
$\xi_1(0.75)$	0.76296	0.06621	0.63513	0.87928	-0.00773	0.49692	1.01162
$\zeta_2(7.5)$	7.36200	0.55461	6.51050	8.40194	0.00984	0.50392	1.00001
$\delta_2(0.05)$	0.04829	0.00577	0.04045	0.05902	0.00597	0.50238	0.99996
$\xi_2(0.65)$	0.64983	0.06029	0.53887	0.77669	-0.00359	0.49857	0.99996
$\theta(4.8)$	4.65038	0.52427	3.85500	5.62394	0.01400	0.50559	1.00018
$\mu(2.5)$	2.50104	0.32257	1.90090	3.14192	0.01360	0.50543	0.99997
$\beta_1(0.15)$	0.13734	0.06798	-0.00299	0.27876	0.01239	0.50494	0.99996

Table 3: p-value of K-S statistics for goodness of fit test for Kidney Infection data set

Model	T_1 p-value	T_2 p-value
$Model-I$	0.7912	0.4490
$Model-II$	0.5722	0.6860

Table 4: Posterior Summary of Generalised Lindley Frailty with Baseline Generalised Weibull for Kidney Infection Data (Model-I)

Parameter	*Estimate*	*s.e.*	*L.C.L.*	*U.C.L.*	*Geweke test*	$p-value$	*Gelman Rubin test*
burn-in period = 6900;		autocorrelation lag = 400					
ζ_1	1.99061	0.06654	1.84351	2.10812	0.00268	0.50107	1.00003
δ_1	0.05530	0.00310	0.04954	0.06115	-0.00126	0.49950	1.00001
ξ_1	0.66315	0.02125	0.61705	0.70345	-0.00648	0.49741	1.00006
ζ_2	2.71016	0.06297	2.59528	2.83639	-0.00474	0.49811	0.99998
δ_2	0.06205	0.00311	0.05563	0.06818	0.00865	0.50345	1.00034
ξ_2	0.67052	0.02313	0.62961	0.71617	-0.00296	0.49882	0.99998
θ	2.29258	0.09757	2.11305	2.47659	0.00077	0.50031	0.99998
μ	1.38391	0.09709	1.20062	1.58699	-0.00608	0.49757	1.00025
β_1	-0.10576	0.01289	-0.13073	-0.08109	-0.00220	0.49912	0.99997
β_2	-8.88412	1.46382	-11.50565	-6.16638	0.00658	0.49912	1.00032
β_3	2.44371	0.33770	1.84343	3.11420	0.00775	0.50309	1.00075
β_4	1.61045	0.29506	1.08735	2.19532	-0.00539	0.49785	1.00093
β_5	-52.67579	27.25061	-101.04670	-4.02850	0.00950	0.50379	0.99996

Table 5: Posterior Summary of Generalised Lindley Frailty with Baseline Generalised Log-Logistic-II for Kidney Infection Data (Model-II)

Parameter	*Estimate*	*s.e.*	*L.C.L.*	*U.C.L.*	*Geweke test*	$p-value$	*Gelman Rubin test*
burn-in period = 6900;		autocorrelation lag = 400					
ζ_1	3.79268	0.10452	3.59085	4.01330	0.00174	0.50070	1.00009
δ_1	0.00160	0.00006	0.00148	0.00173	-0.00034	0.49987	1.00000
ξ_1	1.04034	0.04412	0.95236	1.11628	0.00441	0.50176	1.00096
ζ_2	4.30595	0.09619	4.11335	4.49439	-0.00039	0.49984	0.99997
δ_2	0.00043	0.00001	0.00041	0.00045	0.00431	0.50172	1.00010
ξ_2	1.25850	0.04593	1.16443	1.34386	0.00485	0.50194	0.99997
θ	2.12060	0.10701	1.92252	2.34629	0.00182	0.50072	0.99997
μ	1.28878	0.09992	1.10590	1.49725	-0.00126	0.49950	1.00025
β_1	-0.10630	0.01145	-0.12756	-0.08297	-0.00153	0.49939	0.99997
β_2	-67.94356	33.76583	-132.77210	-7.62369	0.00755	0.49939	1.00175
β_3	2.51987	0.26359	2.04163	2.97889	0.00187	0.50075	1.00056
β_4	1.51014	0.20139	1.14758	1.87047	-0.00018	0.49993	1.00104
β_5	-54.94150	31.09431	-111.91510	-3.56763	0.00406	0.50162	0.99997

Table 6: Kendall's τ Measure of Dependence

Model	Kendall's τ value
$Model-I$	0.297939
$Model-II$	0.303226

Table 7: AIC, BIC and DIC Comparison

Model	AIC	BIC	DIC
Model-I	685.3974	706.6861	664.4514
Model-II	686.2751	707.5637	665.7128

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 177–202, DOI 10.2478/stattrans-2022-0049
Received – 22.09.2021; accepted – 26.09.2022

Supsim: a Python package and a web-based JavaScript tool to address the theoretical complexities in two-predictor suppression situations

Morteza Nazifi[1], **Hamid Fadishei**[2]

ABSTRACT

Two-predictor suppression situations continue to produce uninterpretable conditions in linear regression. In an attempt to address the theoretical complexities related to suppression situations, the current study introduces two different versions of a software called suppression simulator (Supsim): a) the command-line Python package, and b) the web-based JavaScript tool, both of which are able to simulate numerous random two-predictor models (RTMs). RTMs are randomly generated, normally distributed data vectors x_1, x_2, and y simulated in such a way that regressing y on both x_1 and x_2 results in the occurrence of numerous suppression and non-suppression situations. The web-based Supsim requires no coding skills and additionally, it provides users with 3D scatterplots of the simulated RTMs. This study shows that comparing 3D scatterplots of different suppression and non-suppression situations provides important new insights into the underlying mechanisms of two-predictor suppression situations. An important focus is on the comparison of 3D scatterplots of certain enhancement situations called Hamilton's extreme example with those of redundancy situations. Such a comparison suggests that the basic mathematical concepts of two-predictor suppression situations need to be reconsidered with regard to the important issue of the statistical control function.

Key words: Supsim, multicollinearity, suppression effects, statistical control function.

1. Introduction

Two-predictor suppression effects remain among complex and confusing situations in linear regression (eg. Holling, 1983, Ludlow and Klein, 2014, McFatter, 1979, Friedman and Wall, 2005). When the inclusion of a second predictor, say x_2, which is relatively highly correlated with x_1, in the regression equation leads to some kind of two-predictor suppression effect, possible contradictory results include: calculating a negative part of the explained variance in y when partitioning R^2

[1] University of Bojnord, Iran. E-mail: nazifi@ub.ac.ir. ORCID: https://orcid.org/0000-0002-0155-6743.
[2] University of Bojnord, Iran. ORCID: https://orcid.org/0000-0002-6207-369X.

(Cohen et al., 2003), finding opposite signs between the second predictor's zero-order correlation with *y* and its regression coefficient in the equation, observing situations in which one of the two predictors or both of them get a large regression coefficient in the equation despite showing "no or low" zero-order correlation with *y*, and finally finding situations in which $R^2 > r_{y1}^2 + r_{y2}^2$ (Hamilton, 1987), where r_{y1} and r_{y2} are the zero-order correlations between the outcome variable *y* and x_1 or x_2. Suppression situations have attracted attention for several decades because it is generally believed that such situations can increase the predictive validity especially in the context of psychological testing (Conger and Jackson, 1972, Horst, 1941, Pedhazur, 1997, Tzelgov and Henik, 1991, Watson et al., 2013, Friedman and Wall, 2005, Darlington and Hayes, 2017, Cohen et al., 2003). Under the condition of $R^2 > r_{y1}^2 + r_{y2}^2$, Hamilton (1987) describes an even more challenging two-predictor suppression effect, in which r_{y1} and r_{y2} are both close to 0 but R^2 and $|r_{12}|$ are both near 1, where r_{12} is the correlation between x_1 and x_2. Given that research on these challenging two-predictor suppression effects requires access to some simulation algorithm that can generate three-variable datasets showing different suppression and non-suppression situations, the authors develop and introduce a computerized algorithm called suppression simulator (Supsim), some open-source software (Nazifi and Fadishei, 2021a), made available in two different versions: a) the command-line Python package of Supsim, and b) the web-based JavaScript tool (see screenshots from the user-interface of the web-based Supsim in panel B of Figure 1). This algorithm enables researchers to easily generate numerous series of random data vectors x_1, x_2, and *y* so that one can generate numerous regression models with or without suppression by regressing *y* on both x_1 and x_2. The web-based Supsim is more user-friendly in that it does not require any coding skills and in addition it allows investigators to automatically produce 3D scatterplots of the simulated random two-predictor models (RTM's). Elsewhere, the authors explain in a video how to install and work with both the command-line Python package and the web-based, JavaScript versions of Supsim (Nazifi and Fadishei, 2021b). Before proceeding, a comprehensive definition of two-predictor suppression effects is needed to be used as a frame of reference.

Friedman and Wall (2005) provide a comprehensive review of two-predictor suppression effects, which incorporates different definitions of suppression situations that have been presented so far. Holding arbitrary selected r_{y1} and r_{y2} constant and letting r_{12} vary over its possible limits (see inequality (1) below), Friedman and Wall (2005) show that for each fixed pair of r_{y1} and r_{y2}, letting r_{12} vary, different suppression and non-suppression situations can occur. They are illustrated with some graphical views showing the variations in R^2, $\hat{\beta}_1$ or $\hat{\beta}_2$ in response to the variations in r_{12}. In such graphical views the vertical axis represents either R^2, $\hat{\beta}_1$ or $\hat{\beta}_2$ and the horizontal axis represents r_{12}. Each of the regions in Friedman and Wall's systematic graphs corresponds to some suppression or non-suppression situations defined previously by

other leading researchers in this field (e.g. see Horst, 1941, Lynn, 2003, Conger, 1974, Cohen and Cohen, 1975, Currie and Korabinski, 1984, Shieh, 2001, Sharpe and Roberts, 1997, Velicer, 1978, Hamilton, 1987, Darlington, 1968). According to Friedman and Wall (2005) as long as r_{y1} and r_{y2} *are both positive,* and $r_{y1} > r_{y2}$, as it is common in the linear regression research, the regions on the graph, from left to right, are defined according to Table 1 (Note that in Table 1, Friedman and Wall's definitions are subtly altered to also include situations where r_{y1} and r_{y2} are both *negative* and $|r_{y1}| > |r_{y2}|$). It should be noted that in Friedman and Wall's graphs, when r_{y1} and r_{y2} *are of opposite sign*s the order of the regions described above becomes reverse (see Table 2 for more details). When the reverse graph is the case, region I covers any positive values of r_{12} (all r_{12}'s > 0), and regions II, III, and IV all are shifted to the negative side of the r_{12} axis. In addition, when $r_{y2} = 0$, a situation called "classical suppression", Friedman and Wall's graph has only two regions including, from left to right, region I (enhancement), and region IV (enhancement) (see Figure 2 below; also see the application by Brown (2005) to be able to generate the graphs).

Table 1. Definitions of the Different Suppression and Non-Suppression Situations As Long As $\boldsymbol{r_{y1}}$ and $\boldsymbol{r_{y2}}$ are of Similar Signs, and $|\boldsymbol{r_{y1}}| > |\boldsymbol{r_{y2}}|$

Regions	**Region I: enhancement**	**Region II: Redundancy (Non-Suppression Situations)**	**Region III: Suppression**	**Region IV: enhancement**
Definitions of Suppression and Non-Suppression Situations	▪ All r_{12}'s < 0 ▪ $\|\hat{\beta}_1\| > \|r_{y1}\|$ ▪ $R^2 > r_{y1}^2 + r_{y2}^2$ ▪ And the signs of $\hat{\beta}_1$ and $\hat{\beta}_2$ are always similar to the signs of r_{y1} and r_{y2}, respectively.	▪ $0 \leq r_{12} \leq \gamma$ ▪ $\|\hat{\beta}_1\| \leq \|r_{y1}\|$ ▪ $R^2 \leq r_{y1}^2 + r_{y2}^2$	▪ $\gamma < r_{12} \leq \frac{2\gamma}{1+\gamma^2}$ ▪ $\|\hat{\beta}_1\| > \|r_{y1}\|$ ▪ $R^2 \leq r_{y1}^2 + r_{y2}^2$ ▪ And in which r_{y2} and $\hat{\beta}_2$ are always of the opposite signs.	▪ All $r_{12}'s > \frac{2\gamma}{1+\gamma^2}$ ▪ $\|\hat{\beta}_1\| > \|r_{y1}\|$ ▪ $R^2 > r_{y1}^2 + r_{y2}^2$ ▪ And in which r_{y2} and $\hat{\beta}_2$ are always of the opposite signs.

Note: $\gamma = \frac{r_{y2}}{r_{y1}}$; and $\frac{2\gamma}{1+\gamma^2} = \frac{2\,(r_{y1} \times r_{y2})}{r_{y1}^2 + r_{y2}^2}$

It should be noted that it is also possible to provide simplified, practical definitions of suppression situations. According to such simplified definitions, suppression situations occur when each of the following conditions are met: 1) the absolute value of the collinearity between the two predictors, x_1 and x_2, exceeds the ratio of $|\gamma|$, which

means $|r_{12}| > \left|\frac{r_{y2}}{r_{y1}}\right|$ (negative suppression); 2) r_{y1} and r_{y2} are of similar signs, while the sign of the collinearity between x_1 and x_2 is negative (i.e. $r_{12} < 0$) (reciprocal suppression); and finally 3) r_{y1} and r_{y2} are of opposite signs, while the sign of the collinearity between x_1 and x_2 is positive (i.e. $r_{12} > 0$) (reciprocal suppression).

Table 2. Definitions of the Different Suppression and Non-Suppression Situations As Long As $\boldsymbol{r_{y1}}$ and $\boldsymbol{r_{y2}}$ are of Opposite Signs, and $\boldsymbol{|r_{y1}| > |r_{y2}|}$

Regions	Region IV: enhancement	Region III: Suppression	Region II: Redundancy (Non-Suppression Situations)	Region I: enhancement
Definitions of Suppression and Non-suppression Situations	▪ $r_{12} < \frac{2\gamma}{1+\gamma^2}$ ▪ $\lvert\hat{\beta}_1\rvert > \lvert r_{y1}\rvert$ ▪ $R^2 > r_{y1}^2 + r_{y2}^2$ ▪ And r_{y2} and $\hat{\beta}_2$ are always of the opposite signs.	▪ $\gamma > r_{12} \geq \frac{2\gamma}{1+\gamma^2}$ ▪ $\lvert\hat{\beta}_1\rvert > \lvert r_{y1}\rvert$ ▪ $R^2 \leq r_{y1}^2 + r_{y2}^2$ ▪ And r_{y2} and $\hat{\beta}_2$ are always of the opposite signs.	▪ $0 \geq r_{12} \geq \gamma$ ▪ $\lvert\hat{\beta}_1\rvert \leq \lvert r_{y1}\rvert$ ▪ $R^2 \leq r_{y1}^2 + r_{y2}^2$	▪ All r_{12}'s > 0 ▪ $\lvert\hat{\beta}_1\rvert > \lvert r_{y1}\rvert$ ▪ $R^2 > r_{y1}^2 + r_{y2}^2$ ▪ And the signs of $\hat{\beta}_1$ and $\hat{\beta}_2$ are always similar to the signs of r_{y1} and r_{y2}, respectively.

Note: $\gamma = \frac{r_{y2}}{r_{y1}}$; and $\frac{2\gamma}{1+\gamma^2} = \frac{2\,(r_{y1} \times r_{y2})}{r_{y1}^2 + r_{y2}^2}$

Friedman and Wall (2005) believe that in order to get an accurate picture of two-predictor suppression effects each fixed pair of r_{y1} and r_{y2} should be considered separately allowing r_{12} vary over its possible limit. They state that it is not the r_{12} per se but the combination of the three correlations (i.e. r_{y1}, r_{y2} and r_{12}) that affects the sign change in $\hat{\beta}_2$. The possibility limit of r_{12}, when a fixed pair of r_{y1} and r_{y2} is given, is defined by the following inequality (e.g. Neill, 1973, Sharpe and Roberts, 1997):

$$r_{y1} \times r_{y2} - \sqrt{(1 - r_{y1}^2)(1 - r_{y2}^2)} \leq r_{12} \leq r_{y1} \times r_{y2} + \sqrt{(1 - r_{y1}^2)(1 - r_{y2}^2)} \qquad (1)$$

The limits were imposed by the fact that the correlation matrix which r_{y1}, r_{y2}, and r_{12} come from must be nonnegative, definite (Neill, 1973, Sharpe and Roberts, 1997, Friedman and Wall, 2005). The limits defined by inequality (1) imply that the possible interval of r_{12} can become very wide when both $|r_{y1}|$ and $|r_{y2}|$ are close to 0 and it can also become very narrow when both $|r_{y1}|$ and $|r_{y2}|$ are near 1. Concentrating on the possible limits of r_{12} is extremely important in understanding two-predictor

suppression effects, because formulas of both R^2 and $\hat{\beta}_2$ (and $\hat{\beta}_1$ as well) are sensitive to the values of r_{12} as it is evident from formula (2) (Cohen et al., 2003) and formula (3) below (Cohen et al., 2003, Hamilton, 1987):

$$\hat{\beta}_2 = \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2} \tag{2}$$

$$R^2 = \frac{r_{y1}^2 + r_{y2}^2 - 2 r_{y1} r_{y2} r_{12}}{1 - r_{12}^2} \tag{3}$$

Friedman and Wall's approach, beside its strengths, has an important limitation because in their method only arbitrary selected pairs of correlations are used and therefore one is completely unaware of the data vectors x_1, x_2, and y and what the 3D scatterplots of each particular regression model looks like. Hamilton (1987) does explain a method for generating artificial data vectors x_1, x_2, and y that are used in building regression models in which $R^2 > r_{y1}^2 + r_{y2}^2$, but he uses the data vectors x_1, x_2, and y only in drawing two-dimensional scatterplots and fails to explore 3D scatterplots of the resulting two-predictor models. This study shows that comparing 3D scatterplots of two-predictor regression models with or without suppression bear important new insights into the effects of multicollinearity on the results of linear regression models. In addition, in the previous research, little attention has been paid to the mechanisms of statistical control in redundancy situations compared to suppression situations. Objectives of this study are as follows:

1- Describing the Supsim and showing how simulation with Supsim works (see Section 2).
2- Generating several examples of RTM's, falling within different suppression and non-suppression regions including enhancement, suppression, and redundancy (see Section 3)
3- Generating 3D scatterplots for each simulated RTM to be able to compare them with each other. To make this comparison more meaningful, RTM's from enhancement or suppression regions are matched with those of redundancy regions in terms of either R^2 values or zero-order correlations with y (see Section 3).
4- Making new mathematical reasoning with respect to statistical control mechanisms (see Section 4).
5- Discussing the significance and implications of the findings (see Section 5).
6- Concluding and describing the strengths and weaknesses of this study (see Section 6).

2. Supsim or the RTM Generation Algorithm

The idea behind the RTM generation algorithm or "Supsim" is to facilitate the study of two-predictor suppression effects by generating numerous random functions (i.e. $y_o = f(x_1, x_2)$) and inserting errors into the outputs of those functions and then fitting an OLS regression surface to the resulting noisy data (y). The proposed algorithm is illustrated by panel A of Figure (1). This iterative process starts by choosing two random vectors x_1 and x_2 so that the correlation between x_1 and x_2 (r_{12}) is set to a desired amount. Next, a random function is generated to produce y_o as a function of both x_1 and x_2 and then a normally distributed noise vector, e, is added to y_o in order to generate a noisy data vector y (i.e. $y = y_o + e$). It should be noted that, before running the algorithm, the distribution of the noise vector, $e = N(\mu_e, \sigma_e)$, is arbitrarily determined by the user through selecting an A coefficient where $\mu_e = A\mu_{y_o}$ and $\sigma_e = A\sigma_{y_o}$. Also other arbitrary, user-provided constraints can be set to constrain r_{y1}, r_{y2}, r_{12}, and the amount of R^2 enhancement before running the Supsim. Otherwise all the required constraints are met, the current RTM shall be discarded and the current iteration shall be started again. When designing the Supsim algorithm, an important technical problem was meeting the constraint imposed on r_{12} range. If this problem is left unresolved, the algorithm would be trapped in an exhaustive search over a very large space of all possible RTM's to find those meeting the desired r_{12} range. In order to overcome this limitation and speed up the simulation process, a specific random number generation method is used, which can generate a data vector (x_1) that not only is random, but also shows a desired amount of correlation with another random data vector (x_2) (Whuber, 2017).

The first two steps of the algorithm shown in Figure 1 are designed according to the method described by Whuber (2017). The algorithm first chooses a normal random vector x_1 and then another normal random vector a with the same length, mean, and standard deviation as x_1 and then applies a transformation to a to calculate b in a way that the correlation between b and x_1 is set to the desired amount (r_{12}). Such a transformation is described in Equation (4) where d is the vector of residuals resulted from regressing a on x_1, σ_d represents the standard deviation of d, and σ_{x_1} represents the standard deviation of x_1, and r is the desired amount of correlation between b and x_1. It should be noted that such a transformation changes the initial distribution of the b vector. Therefore, in order to return b to a mean and a standard deviation equal to those of x_1, the b vector again is transformed into x_2 vector by using $x_2 = mb+n$, where $m = \sigma_{x_1} / \sigma_b$ and $n = \mu_{x_1} - m.\mu_b$. Now x_2 is a random, normal vector, with the same length, mean, and standard deviation as x_1 , which shows specific amount of correlation with x_1.

$$b = r.\sigma_d.x_1 + d.\sigma_{x_1}.\sqrt{1 - r^2} \tag{4}$$

2.1. The Simulation Process in Supsim

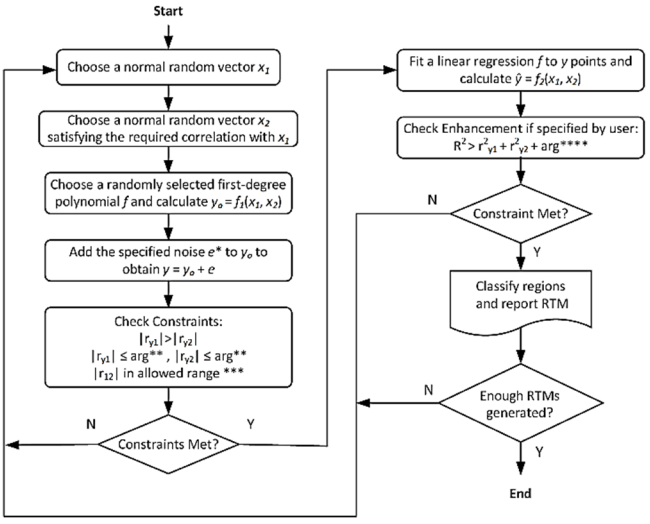
After generating RTM's, the regression parameters R^2, $\hat{\beta}_1$, and $\hat{\beta}_2$ for each of the simulated RTM's are automatically estimated, the simulated RTM's are classified according to the definitions by Friedman and Wall, and then the regression parameters for each of the simulated RTM's are scattered over four different regions on Friedman and Wall's graphs (see Figure 2, panels A through C). Generated by the Python package of Supsim, Figure 2 shows the distribution of the regression parameters of 10,000 simulated RTM's. As it is shown in Figure 2, the regression parameters of the majority of the simulated RTM's fall within the four regions of either the regular graph (in which r_{y1} and r_{y2} are of similar signs, and $|r_{y1}| > |r_{y2}|$) or the reverse graph (in which r_{y1} and r_{y2} are of opposite signs, and $|r_{y1}| > |r_{y2}|$), and only a few of them fall within the two regions of the classical graph (representing classical suppression situations). To avoid overcrowding, in Figure 2, before running Supsim, the algorithm is constrained to plot only the R^2 parameters (and not $\hat{\beta}_1$, and $\hat{\beta}_2$) for each of the simulated RTM's (see Figure 2 below) (For more details about the Supsim algorithm please see user's guide for Supsim (Nazifi and Fadishei, 2021c)).

3. Case Studies on Unique RTM's

Supsim allows users to constrain the magnitudes of r_{y1}, r_{y2}, r_{12}, noise, and the amount of R^2 enhancement to facilitate the production of unique cases of RTM's with desired characteristics that are useful for specific purposes like case studies on unique RTM's. This section is devoted to case studies on unique RTM's with fixed pairs of r_{y1} and r_{y2}. The authors primarily focus on the most challenging situation defined by Hamilton (1987) in which r_{y1} and r_{y2} are both close to 0 but R^2 and $|r_{12}|$ are both near 1 and then extend the discussion to other suppression situations.

3.1. Comparing 3D Scatterplots of Different Regions

After running several simulations by using Supsim, with predetermined constraints, resulting in several sets of large number of RTM's, the authors searched among numerous simulated RTM's to find matched examples of RTM's belonging to different suppression or non-suppression regions. The selected RTM's were then plotted in Figures 3 and 4. It should be noted that in Figure 3, R^2 values are matched between the following pairs: panels A and B, panels C and D, panels E and F. In panels A, C, and E of Figure 3, RTM's are selected in such a way that x_1 and x_2 are not correlated with y (i.e. the y vectors are orthogonal to both x_1 and x_2 vectors). In Figure 4, the R^2 values are matched between the following pairs: panels A and B, and panels C and D. In Figure 4, the absolute values of the zero-order correlations with y also are matched between panels A and C (interested readers can contact the authors to reach datasets for Figures 3 and 4).



Notes for Panel A:

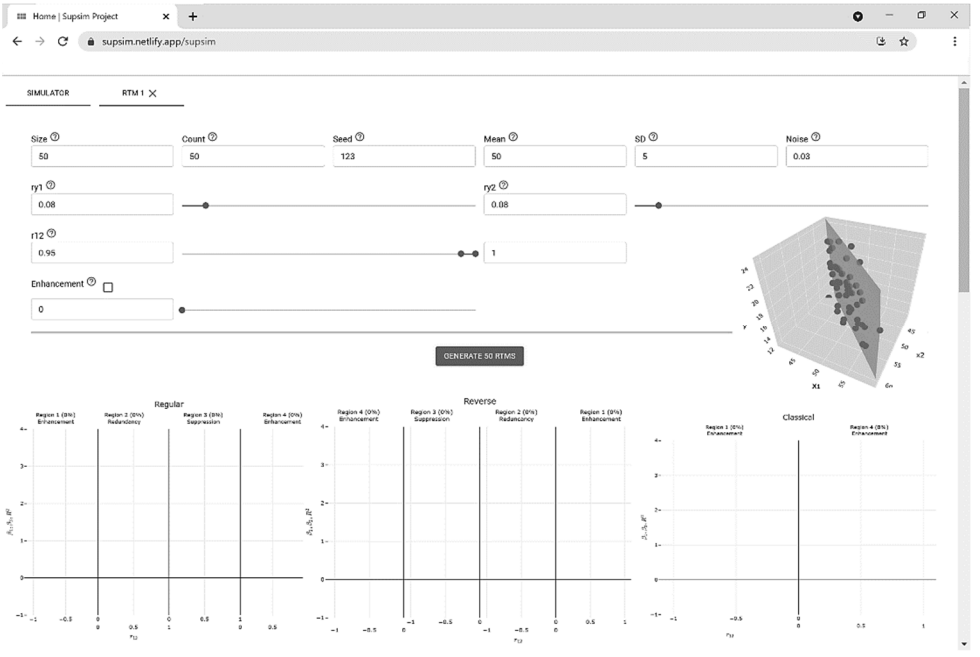
* "e" is a distribution of errors of the same length as *Yo* (or original *Y*), while mean and standard deviation of "e" is determined arbitrarily by the user as a proportion of mean and standard deviation of *Yo*. "e" enables users to control the fit levels of the RTM's.

** arguments (or arg's) are arbitrarily selected by the users to limit the magnitude of r_{y1} and r_{y2}. By using arg's, users control the amount of r_{y1} and r_{y2}.

*** There are two kinds of "allowed range" for r_{12} in Supsim: first, the default allowed range is defined by "$r_{y1} \times r_{y2} - \sqrt{(1 - r^2_{y1})(1 - r^2_{y2})} \leq r_{12} \leq r_{y1} \times r_{y2} + \sqrt{(1 - r^2_{y1})(1 - r^2_{y2})}$"; Second, users are allowed to further limit the magnitude of r_{12} by selecting an arbitrary range between 0 and 1.

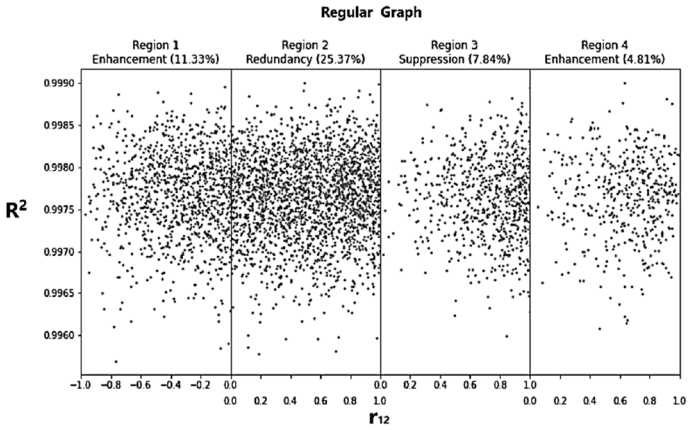
**** arg's about the amount of R^2 enhancement enable users to arbitrarily control the levels of R^2 enhancement by selecting a proportion between 0 and 1.

A: The Iterative Process of Python Package of Supsim

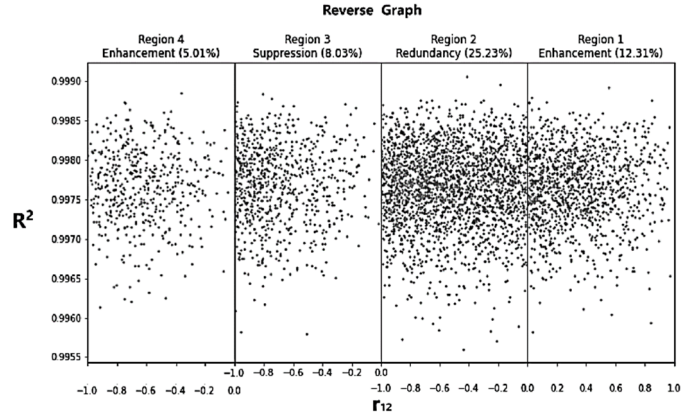


B: Screenshots from the user-interface of the web-based JavaScript version of Supsim

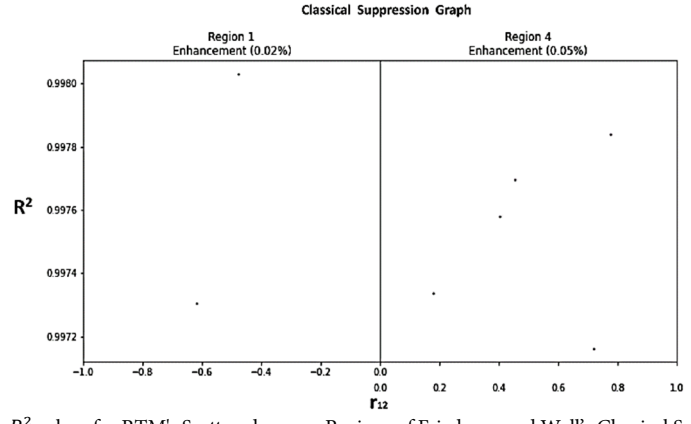
Figure 1. Flowchart of the Python package of Supsim and Screenshots from the JavaScript version of Supsim



A: The R^2 values for thousands of RTM's Scattered among Regions of Friedman and Wall's Regular Graph

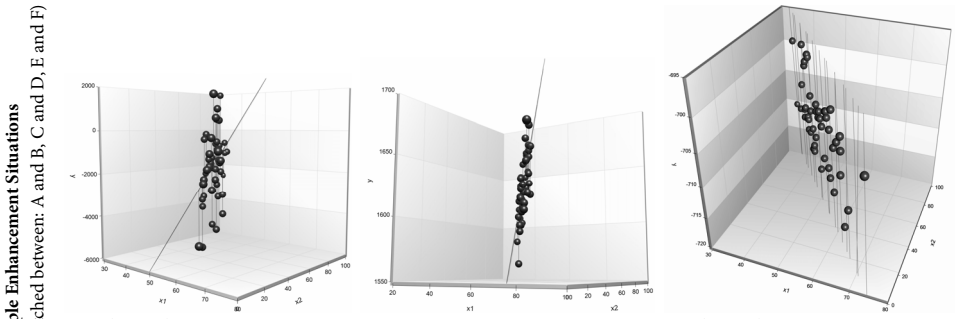


B: The R^2 values for thousands of RTM's Scattered among Regions of Friedman and Wall's Reverse Graph



C: The R^2 values for RTM's Scattered among Regions of Friedman and Wall's Classical Suppression Graph

Figure 2. Distribution of a Large-Scale Sample of RTM's (N = 10,000) among The Regions of Friedman and Wall's Graph

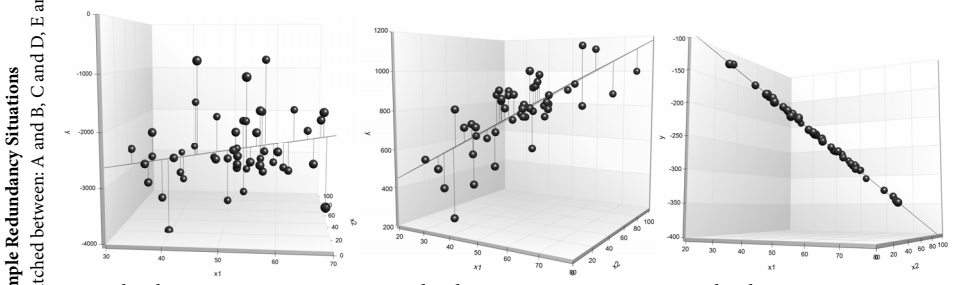


Sample Enhancement Situations
(R^2 values are matched between: A and B, C and D, E and F)

A: Classical Suppression with 0.11 Enhancement
$\boldsymbol{R^2 = 0.119}$, $r_{y1} = 0.08$, $r_{y2} = 0.008$, $r_{12} = -0.965$, $\beta_1 = 1.322$, $\beta_2 = 1.284$;
noise magnitude = 2.00

C: Region I Situation with 0.483 Enhancement
$\boldsymbol{R^2 = 0.492}$, $r_{y1} = 0.07$, $r_{y2} = 0.065$, $r_{12} = -0.981$, $\beta_1 = 3.635$, $\beta_2 = 3.632$;
noise magnitude = 1.00

E: Classical Suppression with 0.995 Enhancement
$\boldsymbol{R^2 = 0.999}$, $r_{y1} = -0.056$, $r_{y2} = -0.00036$, $r_{12} = -0.996$, $\beta_1 = -17.674$ $\beta_2 = -17.647$;
noise magnitude = 0.04



Sample Redundancy Situations
(R^2 values are matched between: A and B, C and D, E and F)

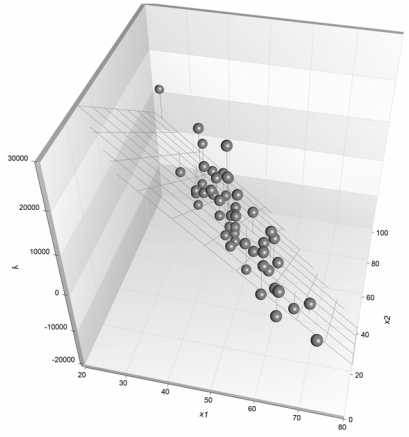
B: Redundancy (RTM without Suppression)
$\boldsymbol{R^2 = 0.115}$, $r_{y1} = 0.27$, $r_{y2} = -0.21$, $r_{12} = -0.212$, $\beta_1 = 0.227$, $\beta_2 = -0.209$;
noise magnitude = 2.00

D: Redundancy (RTM without Suppression)
$\boldsymbol{R^2 = 0.49}$, $r_{y1} = 0.688$, $r_{y2} = 0.657$, $r_{12} = 0.86$, $\beta_1 = 0.47$, $\beta_2 = 0.253$;
noise magnitude = 1.00

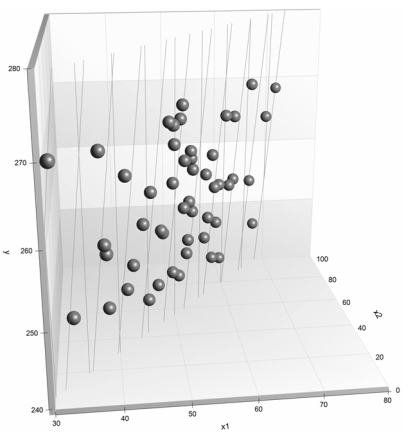
F: Redundancy (RTM without Suppression)
$\boldsymbol{R^2 = 0.998}$, $r_{y1} = -0.856$, $r_{y2} = -0.548$, $r_{12} = 0.056$, $\beta_1 = -0.837$, $\beta_2 = -0.501$;
noise magnitude = 0.04

Figure 3. Matched Scatterplots from Enhancement Regions Compared to Redundancy Regions (Matched for $\boldsymbol{R^2}$)

RTM's from Enhancement Regions
(R^2 values are matched between: A and B, C and D)
(Correlations are Matched between A and C)

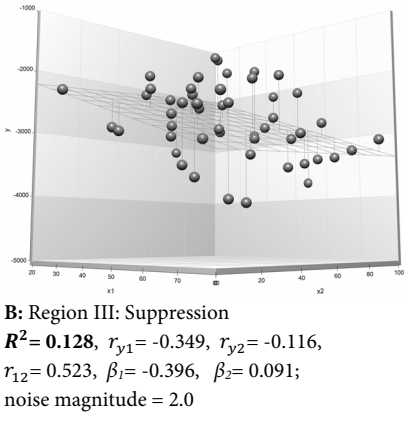


A: Region I Situation with 0.121 Enhancement
$\boldsymbol{R^2}$**= 0.128**, r_{y1}= -0.07, r_{y2}= -0.03,
r_{12}= -0.956, β_1= -1.215, β_2= -1.194;
noise magnitude = 2.0

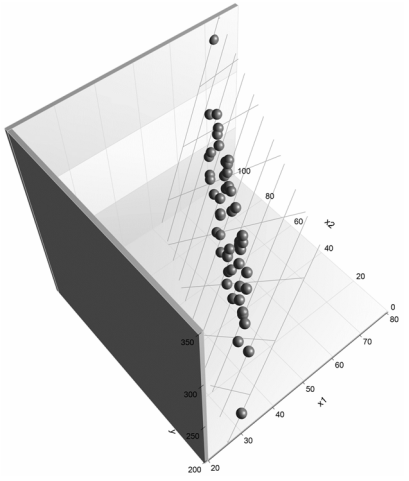


C: Region I Situation with 0.99 Enhancement
$\boldsymbol{R^2}$**= 0.997**, r_{y1}= 0.07, r_{y2}= -0.03,
r_{12}= 0.994, β_1= 9.48, β_2= -9.46;
noise magnitude = 0.05

RTM's from Region III Suppression
(R^2 values are matched between: A and B, C and D)



B: Region III: Suppression
$\boldsymbol{R^2}$**= 0.128**, r_{y1}= -0.349, r_{y2}= -0.116,
r_{12}= 0.523, β_1= -0.396, β_2= 0.091;
noise magnitude = 2.0



D: Region III: Suppression
$\boldsymbol{R^2}$**= 0.997**, r_{y1}= 0.901, r_{y2}= 0.801,
r_{12}= 0.981, β_1= 3.07, β_2= -2.211;
noise magnitude = 0.05

Figure 4. Matched Scatterplots of Enhancement Situations Compared to Region III Suppression (Matched for $\boldsymbol{R^2}$or Zero-Order Correlations)

To obtain the best image quality, the 3D scatterplots in Figure 3 and Figure 4 are generated manually by entering x_1, x_2 and y vectors into the NCSS software and then drawing the 3D scatterplots. However, the entire process of drawing 3D scatterplots

like those in Figure 3 and Figure 4 can be performed automatically by a few clicks using the web-based version of Supsim (Nazifi and Fadishei, 2021c).

For all three enhancement situations in panels A, C, and E of Figure 3, the values of x_1 and x_2 are almost independent from the values of y, which is evident from the scattered dots being almost orthogonal to the plane spanned by x_1 and x_2 in all the three scatterplots (it is also evident from the zero-order correlations with y in Figure 3 panels A, C and E that all of them are smaller than $|0.08|$). Indeed, for panels A, C, and E while x_1 and x_2 are highly sensitive to each other's variability (i.e. all $|r_{12}|$'s ≥ 0.965) they are almost indifferent to the variability in y. Surprisingly, however, not only the three R^2 parameters in panels A, C, and E of Figure 3 are not near 0 but also they are considerably different from each other *as a function of different* $|r_{12}|$ *values* (estimated R^2 values are 0.119, 0.492, and 0.997 respectively for panels A, C, and E of Figure 3). Consider, for example, the scatter plot in Figure 3, panel E, where the possibility interval of r_{12} is -0.99841 to 0.99845, and the regression surface is almost parallel to the y axis and orthogonal to the plane spanned by x_1 and x_2. However, again the estimated value of R^2 is 0.999 (i.e. near 1). Although apparently the estimated R^2 as large as 0.999 in panel E is calculated correctly, because the residuals are near 0, and it is well known that R^2 has been defined as a function of residuals in some texts (Kvalseth, 1985, Alexander et al., 2015), but this situation needs more explanations.

Panel E in Figure 3 is an extreme example of what first was described by Hamilton (1987), a suppression situation with $R^2 > r_{y1}^2 + r_{y2}^2$ in which r_{y1} and r_{y2} are both close to 0 but R^2 and $|r_{12}|$ are both near 1. Hamilton (1987) shows that under the condition of $R^2 > r_{y1}^2 + r_{y2}^2$ whenever $R^2 = 1$ and $r_{y2} = 0$ the following equality can be derived from formula (3) above:

$$r_{12}^2 = 1 - r_{y1}^2 \tag{5}$$

Note that by moving the $-r_{y1}^2$ to the left side of the equality (5) the following equality can be obtained:

$$R^2 = r_{12}^2 + r_{y1}^2 = 1 \tag{6}$$

Readers see that under a set of conditions defined by Hamilton (1987) including $R^2 > r_{y1}^2 + r_{y2}^2$, $R^2 = 1$, and $r_{y2} = 0$, if r_{y1} is also approximately close to 0, as it is the case in panel E of Figure 3, formula (3) tends to approximately substitute the value of r_{12}^2 for the value of R^2. It is possible to generate countless cases of Hamilton's extreme examples in which r_{12}^2 constitutes the major part of R^2 (for another instance see panel C of Figure 4). However, it is an obvious mistake to consider r_{12}^2 as the largest part of R^2 since it is only a proportion of inter-correlation between x_1 and x_2 themselves. One might argue that Hamilton's extreme examples never occur in real empirical studies, and therefore such a mistake would never occur in the real world. However, the authors

show in the next sections that substituting a proportion of r_{12} for the value of R^2 is not limited to Hamilton's extreme examples, but this phenomenon occurs in all different suppression situations.

When Hamilton's extreme example is the case, the slop of the regression surface also cannot be considered as a correct slop, because it causes an incorrect replacement of R^2 with a proportion of r_{12} by allocating inflated regression coefficients (IRC) to both x_1 and x_2 in the equation. IRC can be seen when one compares a regression model affected by high multicollinearity with an equivalent model with the same values of r_{y1} and r_{y2} but r_{12}= 0.

Readers know that in a two-predictor model in which r_{12}= 0, then $r_{y1} = \hat{\beta}_1$ and $r_{y2} = \hat{\beta}_2$, while in cases where $r_{12} \neq 0$ both $|\hat{\beta}_1|$ and $|\hat{\beta}_2|$ deviate from the respective $|r_{y1}|$ and $|r_{y2}|$ values. Also it is well known that both β coefficients and zero-order correlations (r_{yx}) are standardized measures. By using these principles, the authors suggest quantifying the severity of IRC by a novel index that hereafter is referred to as absolute beta-to-correlation ratio (or $|BC|$). The $|BC|$ is defined as follows:

$$|BC| = \left| \frac{\textit{the standardized regression coefficient}}{\textit{the respective zero-order correlation with "y"}} \right| \tag{7}$$

In Figure 3, panel E, the $|BC|$ for $\hat{\beta}_1$ equals 315.61 and it means that $|\hat{\beta}_1|$ is more than 315 times greater than $|\hat{\beta}_1|$ in an equivalent model with r_{12} = 0. And the $|BC|$ for $\hat{\beta}_2$ in panel E equals 49019.45 and it means that $|\hat{\beta}_2|$ is more than *49000 times greater than* $|\hat{\beta}_2|$ in an equivalent model with r_{12} = 0. In contrast, scatterplots from redundancy regions (panels B, D, and F in Figure 3) show no sign of IRC, because all $|BC|$ ratios $\leq$ 1 . For example, in panel F of Figure 3, relatively large values of r_{y1} and r_{y2}, but not necessarily a large value of r_{12}, are needed to obtain a R^2 value as large as 0.998. In fact, the $|BC|$ ratios for those RTM's drawn from redundancy regions are always equal to or smaller than 1 indicating the absence of IRC as it is evident from panels B, D, and F in Figure 3.

The scatterplots in Figure 4 help further explain the issue of IRC in enhancement regions compared to region III (suppression). Note that panels A and B as well as panels C and D are matched for R^2 values in Figure 4. Panels A and C also are matched for zero-order correlations with *y*. The possible interval of r_{12} in both panels A and C of Figure 4 is between -0.995 and 0.9992. A comparison between the two enhancement situations in panels A and C reveals that to obtain a R^2 value of 0.128, a $|r_{12}| = 0.956$ is needed (see panel A of Figure 4). And then in panel C only a 0.038 increase in $|r_{12}|$ is needed to obtain a R^2 value of 0.997. Again, *y* is almost independent from both x_1 and x_2 in both panels A and C. But in panel A, the value of $|r_{12}| = 0.956$ is not strong enough to produce an orthogonal regression surface through generating a large IRC to obtain a R^2 value near 1. Indeed, panel A needs only a 0.038 increase in $|r_{12}|$ value to

perform as well as panel C of Figure 4 in enhancing the R^2 up to 0.997. The $|BC|$ ratios are 17.36 and 39.8 respectively for $\hat{\beta}_1$ and $\hat{\beta}_2$ in panel A of Figure 4 compared to 135.43 and 315.34 respectively for $\hat{\beta}_1$ and $\hat{\beta}_2$ in panel C of Figure 4.

Similarly, IRC is always present in RTM's drawn from region III (suppression) (see panels B and D in Figure 4). For instance, the $|BC|$ ratios for panel B of Figure 4 are 1.135 and 0.784 respectively for $\hat{\beta}_1$ and $\hat{\beta}_2$, while they are more sever for panel D of Figure 4 as they are 3.41 and 2.76 respectively for $\hat{\beta}_1$ and $\hat{\beta}_2$.

So far the readers have seen that IRC may not occur in two-predictor models falling within redundancy regions while it is always present in models falling within region III (suppression), region I or region IV (enhancement). These conclusions have already been verified by the definitions presented by Friedman and Wall (2005) for each of the four regions on their graphs.

By referring to the important issue of statistical control in two-predictor linear regression, the next section presents the results of further case studies on RTM's, which call on researchers to be more cautious about the issue of IRC in suppression situations.

4. New Mathematical Reasoning: The Statistical Control Function

In this section the authors show that comparing the mechanisms of statistical control between regression models affected by suppression effects with those not affected can provide important new insights into *the effects of multicollinearity on the results of two-predictor regression models*. When a second predictor x_2 is entered into the regression equation, multicollinearity between x_1 and x_2 raises the issue of statistical control. To better understand the effects of multicollinearity the authors suggest equality (8) that can be derived from formula (3) by moving the terms $1 - r_{12}^2$ from the denominator to the left side of the equation, multiplying them by R^2 and then moving the term $-R^2.r_{12}^2$ to the right side:

$$R^2 = r_{y1}^2 + r_{y2}^2 - \left(2\, r_{y1}r_{y2}r_{12}\right) + R^2 r_{12}^2 \tag{8}$$

Of course, equality (8) is not an optimum way for calculating R^2, but it is still important because it helps figure out the role of multicollinearity by partitioning R^2 into two parts: a) the sum of the first two terms (i.e. $r_{y1}^2 + r_{y2}^2$) which we call the collinearity-independent part (*CIP*), and b) the sum of the second two terms (i.e. $-2\, r_{y1}r_{y2}r_{12} + R^2 r_{12}^2$), which we call the collinearity-dependent part (*CDP*). It should be noted that when calculating R^2, the terms $-2\, r_{y1}r_{y2}r_{12} + R^2 r_{12}^2$ or *CDP* are added to the terms $r_{y1}^2 + r_{y2}^2$ or *CIP* in order *to control for* the common variance explained jointly by x_1 and x_2 in cases where multicollinearity is present. However, if $r_{12} = 0$, then the sum of the terms $-2\, r_{y1}r_{y2}r_{12} + R^2 r_{12}^2$ is equal to 0, but r_{12} is usually non-zero and accordingly the sum of the terms $-2\, r_{y1}r_{y2}r_{12} + R^2 r_{12}^2$ is usually non-zero. The

terms $-2\, r_{y1}r_{y2}r_{12} + R^2r_{12}^2$ here should be regarded as a proportion of r_{12} because r_{y1} and r_{y2} are held constant to study the effects of variations in r_{12}. Indeed, equality (8) shows that when redundancy is the case, the R^2 formula tends to subtract *some proportion of* r_{12} from $r_{y1}^2 + r_{y2}^2$ to prevent the estimated value of R^2 from containing any part of the common variance explained jointly by x_1 and x_2. Therefore, the terms $-2\, r_{y1}r_{y2}r_{12} + R^2r_{12}^2$ hereafter are called the statistical control part (*SCP*) that usually subtracts some proportion of r_{12} from $r_{y1}^2 + r_{y2}^2$. However, there is evidence that under the enhancement conditions, especially those described by Hamilton (1987), the *SCP* can become positive (see Table 3 below).

By obtaining equality (8) from formula (3) for the first time, Hamilton (1987) argues that in cases where $R^2 > r_{y1}^2 + r_{y2}^2$, r_{y2}= 0, and $R^2 = 1$, then the equality (5) can be derived from formula (3). In fact, by suggesting equality (5), Hamilton (1987) has been first to show that in extreme cases under the condition of $R^2 > r_{y1}^2 + r_{y2}^2$, whenever $R^2 = 1$, r_{y2}= 0, and r_{y1} is also approximately near 0, then formula (3) tends to approximately substitute the value of r_{12}^2 for the value of R^2. Generally, when enhancement is the case, the *SCP* is always positive (see Table 3 below) adding some proportion of r_{12} to the value of $r_{y1}^2 + r_{y2}^2$, which in turn leads to the condition of $R^2 > r_{y1}^2 + r_{y2}^2$.

So far it is evident that there is a statistical control function inherent in formula (3), which if carefully quantified can help explain why suppression situations occur. Readers know that if r_{12} = 0, then $R^2 = r_{y1}^2 + r_{y2}^2$, while in cases where $r_{12} \neq 0$, then the value of R^2 deviates from the value of $r_{y1}^2 + r_{y2}^2$ (see Table 3 below). This explains why many texts (e.g. Cohen et al., 2003, Darlington and Hayes, 2017) suggest the following formulas:

$$R_{y.12}^2 = r_{y1}^2 + sr_2^2 \tag{9}$$

$$sr_2 = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{1 - r_{12}^2}} \tag{10}$$

where sr_2 is the semipartial correlation of x_2 with y, and sr_2^2 is its squared value representing a proportion of the total variance in y explained by x_2 over and above the variance explained by x_1. In fact, when calculating R^2, sr_2^2 is used instead of r_{y2}^2 *to prevent* R^2 *from including* the common variance explained jointly by x_1 and x_2 in cases of multicollinearity (i.e. when $r_{12} \neq 0$). Here, again, if r_{12} = 0, then $sr_2^2 = r_{y2}^2$, while if $r_{12} \neq 0$ then sr_2^2 deviates from r_{y2}^2. Indeed, sr_2^2 in formula (9) can be divided into two parts:

$$sr_2^2 = r_{y2}^2 + SCP \tag{11}$$

And formula (9) can be rewritten as follows:

$$R_{y.12}^2 = r_{y1}^2 + r_{y2}^2 + SCP \tag{12}$$

Therefore, equality (11) gives another simple method for quantifying the *SCP*:

$$SCP = sr_2^2 - r_{y2}^2 \tag{13}$$

As a result when r_{y1}, r_{y2} and r_{12} are known, the statistical control part (*SCP*) also can be defined as a function of the combination of three correlations:

$$SCP = f(r_{y1}, r_{y2}, r_{12}) = \left(\frac{r_{y2} - r_{y1} r_{12}}{\sqrt{1 - r_{12}^2}} \right)^2 - r_{y2}^2 \tag{14}$$

Readers see that the first term in function (14) is equal to sr_2^2, and therefore function (14) is identical to equality (13).

As the readers may guess, there is also a collinearity-dependent part (CDP_B) in both $\hat{\beta}_1$ and $\hat{\beta}_2$ formulas, which help explain the reason why regression coefficients become inflated in suppression situations. The following equalities can be derived from formulas of $\hat{\beta}_1$ and $\hat{\beta}_2$ (see formula (2) above):

$$\hat{\beta}_1 = r_{y1} - r_{y2} r_{12} + \hat{\beta}_1 r_{12}^2 \tag{15}$$

$$\hat{\beta}_2 = r_{y2} - r_{y1} r_{12} + \hat{\beta}_2 r_{12}^2 \tag{16}$$

Similarly, equalities (15) and (16) each partition the respective standardized regression coefficients into two parts: a) the first term, which is the zero-order correlation with *y* (r_{y1} or r_{y2}), is called the collinearity-independent part (CIP_B) and b) the sum of the next two terms (i.e. $-r_{y2} r_{12} + \hat{\beta}_1 r_{12}^2$ in equality (15) and $-r_{y1} r_{12} + \hat{\beta}_2 r_{12}^2$ in equality (16)) is called the collinearity-dependent part (CDP_B). The authors suggest using CDP_{B1} as the collinearity-dependent part in $\hat{\beta}_1$ and CDP_{B2} as the collinearity-dependent part in $\hat{\beta}_2$. Here, again, the aim of adding CDP_B terms to each zero-order correlations is to penalize the regression coefficients for multicollinearity. However, the term "penalty" can be used strictly for CDP_{B1} and CDP_{B2} values as long as no kind of two-predictor suppression exists in the model, because only and only over the redundancy regions the signs of CDP_{B1} and CDP_{B2} are constantly opposite to the signs of r_{y1} and r_{y2}, making them to produce $|\hat{\beta}_1|$ and $|\hat{\beta}_2|$ values smaller than or equal to $|r_{y1}|$ and $|r_{y2}|$ (see Table 3 below). In contrast, in region III (suppression) as well as both region I and region IV (enhancement), the sign of CDP_{B1} is always similar to the sign of r_{y1} adding progressively greater proportions of r_{12} to r_{y1} to produce more and more inflated $\hat{\beta}_1$ values as $|r_{12}|$ increases to its maximum value (see Table 3 below). Interestingly, over both the region III (suppression) and the region IV (enhancement), always $|CDP_{B2}| > |r_{y2}|$ and the signs of CDP_{B2}'s are always opposite to the signs of the respective r_{y2}'s making them to produce inflated $\hat{\beta}_2$ values of the opposite signs compared to r_{y2}. Therefore, over the region III (suppression) and the region IV (enhancement) situations, CDP_{B2} subtracts progressively larger proportions of r_{12} from

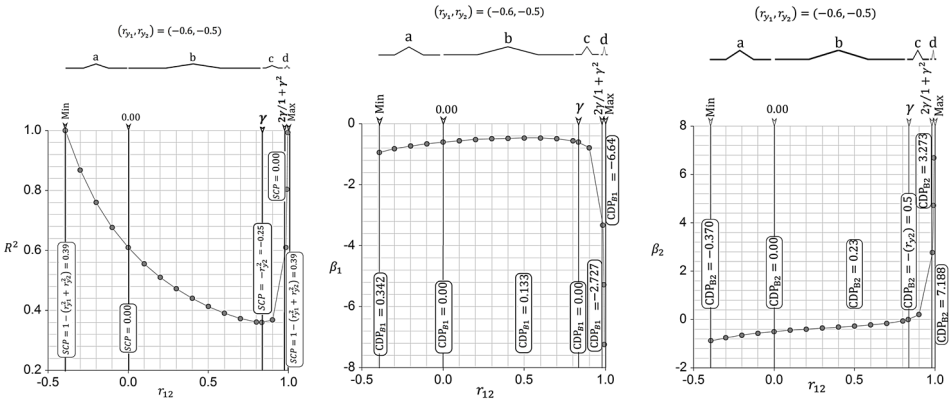
r_{y2} as $|r_{12}|$ increases to its maximum value (see Table 3 below). Finally, in region I (enhancement) the sign of CDP_{B2} values is always similar to the sign of r_{y2} adding progressively larger proportions of r_{12} to r_{y2} to produce inflated $\hat{\beta}_2$ values as $|r_{12}|$ increases to its maximum value (see Table 3 below).

To verify these observations, consider, for example, an arbitrary, fixed pair of r_{y1} and r_{y2} letting r_{12} vary over its possible limit. This arbitrary pair can be (-0.6, -0.5). Variations in the regression parameters in response to the variations in r_{12} for the pair (-0.6, -0.5) are shown in Table 3. To further discuss the mechanisms of statistical control, also for the pair (-0.6, -0.5), all the values of R^2, $\hat{\beta}_1$, and $\hat{\beta}_2$ are plotted against different values of r_{12} in panels A through C of Figure 5.

Table 3. Variations in the regression parameters according to the variation in r_{12} for the pair $r_{y1} = -0.6, r_{y2} = -0.5, n = 25$

γ	= 0.833333333				**Lower limit of r_{12}**	= -0.39282		
$2\gamma/1+\gamma^2$	= 0.983606557				**Upper limit of r_{12}**	= 0.9928203		
Range* of r_{12}	R^2	$\hat{\beta}_1$	$\hat{\beta}_2$	sr_2^2	*SCP*	CDP_{B1}	CDP_{B2}	**SE$\hat{\beta}$'s**
***Max*=0.992820323**	*1.000*	*-7.240*	*6.688*	*0.640*	*0.390*	*-6.640*	*7.188*	*0.000*
0.99	0.80402	-5.28	4.72	0.444	0.194	-4.68	5.22	0.669
***ratio*=0.983606557**	*0.610*	*-3.327*	*2.773*	*0.250*	*0.000*	*-2.727*	*3.273*	*0.738*
0.90	0.36842	-0.79	0.21	0.008	-0.241	-.189	0.710	0.389
γ = 0.833333333	*0.360*	*-0.600*	*0.000*	*0.000*	*-0.250*	*0.000*	*0.500*	*0.309*
0.80	0.36111	-0.56	-0.06	0.001	-0.249	0.044	0.44	0.284
0.70	0.37255	-0.49	-0.16	0.013	-0.237	0.11	0.34	0.236
0.60	0.39063	-0.47	-0.22	0.031	-0.219	0.131	0.28	0.208
0.50	0.41333	-0.47	-0.27	0.053	-0.197	0.133	0.23	0.189
0.40	0.44048	-0.48	-0.31	0.080	-0.17	0.123	0.19	0.174
0.30	0.47253	-0.49	-0.35	0.113	-0.137	0.105	0.148	0.162
0.20	0.51042	-0.52	-0.40	0.150	-0.099	0.079	0.10	0.152
0.10	0.55556	-0.56	-0.44	0.196	-0.054	0.044	0.055	0.143
0.00	*0.61000*	*-0.60*	*-0.50*	*0.250*	*0.000*	*0.000*	*0.000*	*0.133*
-0.10	0.67677	-0.66	-0.57	0.317	0.067	-0.056	-0.065	0.122
-0.20	0.76042	-0.73	-0.65	0.400	0.15	-0.129	-0.146	0.106
-0.30	0.86813	-0.82	-0.75	0.508	0.258	-0.224	-0.247	0.081
***Min*=-0.392820323**	*1.000*	*-0.942*	*-0.870*	*0.640*	*0.390*	*-0.342*	*-0.370*	*0.000*
-0.40	1.01190	-0.95	-0.88	0.652	0.4	-0.352	-0.381	-
-0.50	1.21333	-1.13	-1.07	0.853	0.6	-0.533	-0.567	-
-0.60	1.51563	-1.41	-1.34	1.156	0.9	-0.806	-0.844	-
-0.70	2.01961	-1.86	-1.80	1.660	1.4	-1.26	-1.304	-
-0.80	3.02778	-2.78	-2.72	2.668	2.41	-2.17	-2.22	-
-0.90	6.05263	-5.53	-5.47	5.693	5.44	-4.92	-4.97	-
-0.99	60.50251	-55.03	-54.97	60.143	59.9	-54.89	-54.48	-

Note: *SCP* = statistical control part; CDP_{B1} = collinearity-dependent part of $\hat{\beta}_1$; CDP_{B2} = collinearity-dependent part of $\hat{\beta}_1$; $SE\hat{\beta}$'s = standard errors of $\hat{\beta}'s$; Min = minimum allowed value of r_{12}; Max = maximum allowed value of r_{12}; ratio = $2\gamma/1+\gamma^2$; *: The possibility interval of r_{12} is highlighted in gray in r_{12} column. Note that only the highlighted area on the table falls within the allowed range of r_{12}.



A: Changes in R^2 According to Changes in Both r_{12} and SCP

a: Region I: Enhancement: When calculating the R^2 value, *SCP* adds progressively greater proportions of r_{12} to $(r_{y1} + r_{y2})$ as r_{12} approaches its minimum value.

b: **Region II: Redundancy:** SCP penalizes R^2 for multicollinearity by subtracting progressively greater proportions of r_{12} from $(r_{y1} + r_{y2})$ as r_{12} approaches γ.

c: **Region III: Suppression:** SCP subtracts progressively smaller proportions of r_{12} from $(r_{y1} + r_{y2})$ as r_{12} approaches $2\gamma/1 + \gamma^2$ until the penalty level against multicollinearity reaches 0 by $r_{12} = 2\gamma/1 + \gamma^2$.

d: Region IV: Enhancement: When calculating the R^2 value, SCP adds progressively greater proportions of r_{12} to $(r_{y1} + r_{y2})$ as r_{12} approaches its maximum value.

B: Changes in $\hat{\beta}_1$ According to Changes in Both r_{12} and CDP_{B1}

a: **Region I: Enhancement:** When calculating $\hat{\beta}_1$, CDP_{B1} adds progressively greater proportions of r_{12} to r_{y1} to create inflated $\hat{\beta}_1$ values as r_{12} approaches its minimum value. The signs of CDP_{B1} and r_{y1} are always similar in this region.

b: **Region II: Redundancy:** the CDP_{B1} penalizes $\hat{\beta}_1$ for multicollinearity by subtracting different proportions of r_{12} from r_{y1} when calculating $\hat{\beta}_1$. When r_{12} = 0.00 or $r_{12} = \gamma$ the penalty level against multicollinearity always is 0 and this explains why $\hat{\beta}_1 = r_{y1}$. The CDP_{B1} and the r_{y1} are always of the opposite signs in this region.

c: **Region III: Suppression:** CDP_{B1} adds progressively greater proportions of r_{12} to r_{y1} to create inflated $\hat{\beta}_1$ values as r_{12} approaches $2\gamma/1 + \gamma^2$. The signs of CDP_{B1} and r_{y1} are always similar in this region.

d: Region IV: Enhancement: CDP_{B1} adds progressively greater proportions of r_{12} to r_{y1} to create inflated $\hat{\beta}_1$ values as r_{12} approaches its maximum value. The sign of CDP_{B1} and r_{y1} are always similar in this region.

C: Changes in $\hat{\beta}_2$ According to Changes in Both r_{12} and CDP_{B2}

a: Region I: Enhancement: When calculating $\hat{\beta}_2$, CDP_{B2} adds progressively greater proportions of r_{12} to r_{y2} to create inflated $\hat{\beta}_2$ values as r_{12} approaches its minimum value. The signs of CDP_{B2} and r_{y2} are always similar in this region.

b: Region II: Redundancy: CDP_{B2} penalizes $\hat{\beta}_2$ for multicollinearity by subtracting progressively greater proportions of r_{12} from r_{y2} as r_{12} approaches γ. CDP_{B2} and r_{y2} are always of opposite signs in this region.

c: Region III: Suppression: Always $|CDP_{B2}| > |r_{y2}|$, CDP_{B2} and r_{y2} are always of opposite signs in this region, and CDP_{B2} subtracts progressively greater proportions of r_{12} from r_{y2} as r_{12} approaches $2\gamma/1 + \gamma^2$. Therefore, CDP_{B2} creates inflated $\hat{\beta}_2$ values of the opposite sign with respect to r_{y2}.

d: Region IV: Enhancement: Always $|CDP_{B2}| > |r_{y2}|$, CDP_{B2} and r_{y2} are always of opposite signs in this region, and CDP_{B2} subtracts progressively greater proportions of r_{12} from r_{y2} as r_{12} approaches its maximum value. CDP_{B2} creates inflated $\hat{\beta}_2$ values of the opposite sign with respect to r_{y2}.

Figure 5. Comparing the Statistical Control Mechanisms Among Suppression and Non-Suppression Situations

The possibility interval of r_{12} for the pair (-0.6, -0.5) is -0.39282 ≤ r_{12} ≤ 0.9928203. Table 3 and panels A through C in Figure 5 show that when the minimum allowed value of r_{12} is used (i.e. r_{12} = -0.39282) then the calculations indicate that $R^2 = r_{y1}^2 + sr_2^2 = (-0.6)^2 + 0.64 = 1$, $\hat{\beta}_1 = -0.942$, $\hat{\beta}_2 = -0.87$, sr_2= -0.8, $sr_2^2 = 0.64$, $SCP = sr_2^2 - r_{y2}^2$= 0.64 – 0.25 = 0.39, CDP_{B1} = -0.342, CDP_{B2} = -0.37. Because this is a region I situation (enhancement) (see the definitions in Table 1), therefore, the sign of the *SCP* is positive and the signs of CDP_{B1} and CDP_{B2} are both similar to the signs of r_{y1} and r_{y2}, respectively. Such conditions in region I (enhancement) cause *SCP* playing a role opposite to statistical control mechanism, because CDP_{B1} and CDP_{B2} in region I (enhancement) *add some proportions of* r_{12} to both r_{y1} and r_{y2}, instead of penalizing them for multicollinearity, a mechanism that causes inflation in $|\hat{\beta}_1|$ and $|\hat{\beta}_2|$ values. Therefore, it can be seen that $|\hat{\beta}_1|$ and $|\hat{\beta}_2|$ values for the pair (-0.6, -0.5), are respectively 1.57 and 1.74 times greater than $|\hat{\beta}_1|$ and $|\hat{\beta}_2|$ in an equivalent model with r_{12} = 0 (see Table 3). Panel A in Figure 5 also shows that, in examples where the minimum allowed r_{12} is used, $SCP = 1 - (r_{y1}^2 + r_{y2}^2) = 0.39$.

In contrast, Table 3 shows that if r_{12} = 0 then $R^2 = r_{y1}^2 + sr_2^2 = r_{y1}^2 + r_{y2}^2 = (-0.6)^2 + (-0.5)^2 = 0.61$, $\hat{\beta}_1 = r_{y1} = -0.6$, $\hat{\beta}_2 = r_{y2} = -0.5$, $sr_2 = r_{y2}$ = -0.5, $sr_2^2 = r_{y2}^2 = 0.25$, $SCP = sr_2^2 - r_{y2}^2$= 0.25 – 0.25 = 0, CDP_{B1} = 0, CDP_{B2} = 0 (see also panels A through C in Figure 5). Obviously, when r_{12} = 0 the R^2 value cannot exceed the value of $r_{y1}^2 + r_{y2}^2$.

Now, consider a condition in which $r_{12} = \gamma = \frac{-0.5}{-0.6} = 0.833333333$. In such a condition Table 3 shows that $R^2 = r_{y1}^2 + sr_2^2 = (-0.6)^2 + 0 = 0.36$, $\hat{\beta}_1 = -0.6$, $\hat{\beta}_2 = 0$, sr_2= 0, $sr_2^2 = 0$, SCP = $sr_2^2 - r_{y2}^2$= 0 – 0.25 = – 0.25, $CDP_{B1} = 0$, $CDP_{B2} = -(r_{y2}) = 0.5$. These results show that $r_{12} = \gamma$ is the end-point of the redundancy region, in which the statistical control function removes the entire part of x_2 by estimating $\hat{\beta}_2 = 0$ and $SCP = -r_{y2}^2$. Accordingly, panel A in Figure 5 shows that under such a condition $SCP = -0.25$. In fact, when $r_{12} = \gamma$, linear regression model assumes that any explained variance in y related to x_2 is in common with x_1 and therefore x_2 has no specific contribution to add to the explained variance in y.

And for the pair (-0.6, -0.5), if $r_{12} = \frac{2\gamma}{1+\gamma^2}$ = 0.983606557 then $R^2 = r_{y1}^2 + sr_2^2 = r_{y1}^2 + r_{y2}^2 = (-0.6)^2 + (-0.5)^2 = 0.61$, $\hat{\beta}_1 = -3.327$, $\hat{\beta}_2 = 2.773$, $|sr_2| = |r_{y2}| = 0.5$, $sr_2^2 = r_{y2}^2 = 0.25$, SCP = $sr_2^2 - r_{y2}^2$= 0.25 – 0.25 = 0, CDP_{B1} = -2.727, CDP_{B2} = 3.2726 (also see panels A through C in Figure 5). Although in the latter case, SCP is 0 and again $R^2 = r_{y1}^2 + r_{y2}^2$, contrary to situations where r_{12} =0, CDP_{B1} and CDP_{B2} here are quite large creating inflated $\hat{\beta}_1$ and $\hat{\beta}_2$ with $|\hat{\beta}_1|$ being 5.545 times greater than $|\hat{\beta}_1|$ in an equivalent model with r_{12} = 0 and $|\hat{\beta}_2|$ being 5.546 times greater than $|\hat{\beta}_2|$ in an equivalent model with r_{12} = 0. Another important insight here is that as $|r_{12}|$ increases beyond the value of $|\gamma|$ the statistical control mechanism is weakened

gradually so that by $|r_{12}| = \left|\frac{2\gamma}{1+\gamma^2}\right|$ the penalty level against multicollinearity reaches 0 (i.e. SCP = 0; see panels A through C in Figure 5).

Finally, if the maximum allowed value of r_{12} is used (i.e. r_{12} = 0.992820323) then $R^2 = r_{y1}^2 + sr_2^2 = (-0.6)^2 + 0.64 = 1$, $\hat{\beta}_1 = -7.24$, $\hat{\beta}_2 = 6.6881, sr_2 = 0.79999861$, $sr_2^2 = 0.64$, SCP = $sr_2^2 - r_{y2}^2 = 0.64 - 0.25 = 0.39$, $CDP_{B1} = -6.64$, $CDP_{B2} = 7.1881$. Again, here, $SCP = 1 - \left(r_{y1}^2 + r_{y2}^2\right) = 0.39$, but both CDP_{B1} and CDP_{B2} show that the IRC is much more sever compared to the case where the minimum allowed value of r_{12} is used. In this case, $\left|\hat{\beta}_1\right|$ and $\left|\hat{\beta}_2\right|$ are respectively 12.07 and 13.376 times greater than $\left|\hat{\beta}_1\right|$ and $\left|\hat{\beta}_2\right|$ in an equivalent model with $r_{12} = 0$.

5. Discussion

The concept of two-predictor suppression effects has been the subject of debate over terminology (Friedman and Wall, 2005), definition, and interpretation (Mendershausen, 1939, Horst, 1941, Meehl, 1945, Conger and Jackson, 1972, Conger, 1974, Tzelgov and Henik, 1991, Velicer, 1978, Cohen and Cohen, 1975, Lynn, 2003, Sharpe and Roberts, 1997, Shieh, 2001) for decades. However, one point of agreement has been the approach chosen by some researchers who agree that a suppressor variable showing "no or low" correlation with the criterion variable *y* but is correlated with another significant predictor x_1, can be included in the regression equation to increase the predictive validity of x_1 and it explains why they consider suppressor variables useful and even desirable for situations where the purpose of the study is prediction (Conger and Jackson, 1972, Horst, 1941, Pedhazur, 1997, Tzelgov and Henik, 1991, Watson et al., 2013, Friedman and Wall, 2005, Darlington and Hayes, 2017, Cohen et al., 2003). On the other hand, some texts have warned researchers against multicollinearity and suggest some "rules of thumb" to limit the magnitude of multicollinearity between predictor variables, especially when the purpose of the study is "theoretical explanation" (e.g. Cohen et al., 2003). They argue that highly correlated predictor variables, when simultaneously included in the regression equation, cause "instabilities" in different meanings: first, increased standard errors, as a function of high multicollinearity, may cause "instability" in estimating the regression coefficients (Cohen et al., 2003, Fox, 1997, Neter et al., 1996); second, computational inaccuracies are more likely to occur in calculating the inverses of matrices with highly correlated variables (Cohen and Cohen, 1983); and third, high levels of r_{12} can lead to rapid increase in $\hat{\beta}_2$, a condition in which "the interpretation of regression coefficients may become problematic" (Cohen et al., 2003). Friedman and Wall (2005) argue against the latter texts by presenting evidence that show the standard errors (*SE's*) of regression coefficients do not increase steadily with increasing multicollinearity and there are cases in which low standard errors are coincident with high multicollinearity and that *SE's* of regression

coefficients always become 0 when the multicollinearity for each given pair of r_{y1} and r_{y2} reaches its absolute maximum values (see Table 3). They also argue that the issue of computational accuracy is no longer problematic for the latest generations of regression algorithms (Friedman and Wall, 2005). And finally, Friedman and Wall (2005) conclude that when regressing y on two predictors there are no limits on multicollinearity except those warranting a nonnegative definite matrix. Although Friedman and Wall's observation concerning *SE's* of regression coefficients is quite correct, their final conclusion, which assumes no limits should be imposed on multicollinearity except nonnegative, definiteness limitation is incorrect. Similarly, as Cohen et al. (2003) observed, it is true that there is a rapid increase in $\hat{\beta}_2$ at high levels of r_{12}, but their agreement to use the suppressor variables in order to increase R^2 in cases where the main purpose of the study is increasing the predictive validity is misleading. As noted earlier in the introduction section, two important aspects of two-predictor suppression effects have been overlooked in the previous studies that have led researchers to misleading conclusions: first, failure to compare 3D scatterplots of suppression and non-suppression situations; and second, insufficient attention to the important issue of statistical control mechanisms in non-suppression compared to suppression situations. Taking into consideration these two important aspects, this study achieved significant findings as follows.

First, a closer look at the integral terms in R^2, $\hat{\beta}_1$, and $\hat{\beta}_2$ formulas indicates that these formulas consist of two separate parts (see Equalities 8, 15 and 16 above): the collinearity-independent part (*CIP*) and the collinearity-dependent part (*CDP*). The *CDP* terms in R^2, $\hat{\beta}_1$, and $\hat{\beta}_2$ formulas are associated with statistical control mechanisms, and therefore should be quantified and examined separately.

Second, the *CDP* terms in R^2 formula act differently in redundancy and suppression regions in terms of statistical control mechanisms (see Figure 5 panel A). While the *SCP* is always negative in redundancy regions penalizing R^2 for multicollinearity, the penalty level of *SCP* decreases progressively in region III (suppression), which in turn causes *SCP* to subtract progressively smaller proportions of r_{12} from r_{y2}^2 as r_{12} approaches $2\gamma/1+\gamma^2$. At $2\gamma/1+\gamma^2$ point, the penalty level of *SCP* against multicollinearity reaches 0. Beyond the $2\gamma/1+\gamma^2$ ratio, in region IV (enhancement), *SCP* becomes positive and adds progressively greater proportions of r_{12} to r_{y2}^2 as r_{12} approaches its absolute maximum value. As mentioned earlier, according to the definitions in Table 1 and Table 2, when r_{y1} and r_{y2} have similar signs, the region covering all r_{12}'s < 0 create the "region I" (enhancement) (or reciprocal suppression), but when r_{y1} and r_{y2} are of opposite signs, the region covering all r_{12}'s > 0 produces the "region I" (enhancement) (another type of reciprocal suppression). It should be noted that *SCP* is positive in both types of "region I" situations, adding progressively greater proportions of r_{12} to r_{y2}^2 as r_{12} approaches its absolute maximum

values. For example, for the pair (-0.6, -0.5), panel A in Figure 5 shows that *SCP* is positive and equal to $1 - (r_{y1}^2 + r_{y2}^2)$ both at the upper limit and at the lower limit of r_{12}, whereas in cases where $r_{12} = 0$, also *SCP* = 0; if $r_{12} = \gamma$, $SCP = -(r_{y2}^2)$; and if $r_{12} = 2\gamma/1 + \gamma^2$, *SCP* = 0.

According to these findings, the authors suggest renaming the regions suggested by Friedman and Wall (2005) in terms of their statistical control functioning. Therefore, the following labels are suggested: "region I: statistical anti-control", "region II: statistical control", "region III: statistical de-control", and "region IV: statistical anti-control", respectively for "region I: enhancement", "region II: redundancy", "region III: suppression", and "region IV: enhancement". In fact, the aim of these "relabelling" is to show that all different two-predictor suppression effects are different kinds of "dysregulations in statistical control" and that the "correct statistical control" can occur only and only in "region II: redundancy". The authors emphasize that no proportions of r_{12} can replace the R^2 value, and therefore the results produced by two-predictor suppression effects are completely erroneous and misleading.

Third, the *CDP* terms in formulas of both $\hat{\beta}_1$ and $\hat{\beta}_2$ also function differently in redundancy and suppression regions (see Figure 5, panels B and C). The signs of both CDP_{B1} and CDP_{B2} values in redundancy regions are always opposite to the signs of r_{y1} and r_{y2} and they always subtract different proportions of r_{12} from r_{y1} and r_{y2} to penalize the resulting $\hat{\beta}_1$ and $\hat{\beta}_2$ values for multicollinearity and to produce $\hat{\beta}_1$ and $\hat{\beta}_2$ values, which are always smaller than or equal to r_{y1} and r_{y2}, respectively. In contrast, in region III (suppression) the signs of CDP_{B1} values are always similar to the sign of r_{y1}, adding progressively greater proportions of r_{12} to r_{y1} to produce inflated $\hat{\beta}_1$ values as r_{12} approaches $2\gamma/1 + \gamma^2$, whereas the signs of CDP_{B2} values are always opposite to the sign of r_{y2} in region III (suppression), but always $|CDP_{B2}| > |r_{y2}|$ in this region, a condition in which CDP_{B2} produces inflated $\hat{\beta}_2$ values of the opposite sign compared to r_{y2}. Similarly, in region IV (enhancement) the signs of CDP_{B1} values are always similar to the sign of r_{y1}, creating inflated $\hat{\beta}_1$ values as r_{12} approaches its absolute maximum value, whereas the signs of CDP_{B2} values again are always opposite to the sign of r_{y2}, but always $|CDP_{B2}| > |r_{y2}|$ in this region, a condition that cause CDP_{B2} to produce inflated $\hat{\beta}_2$ values of the opposite sign compared to r_{y2}. In contrast, in region I (enhancement), the signs of both CDP_{B1} and CDP_{B2} values are always similar to the signs of r_{y1} and r_{y2} adding gradually greater proportions of r_{12} to the zero-order correlations to create progressively more inflated $\hat{\beta}_1$ and $\hat{\beta}_2$ values as r_{12} approaches its absolute maximum value. These findings show that the statistical control mechanisms can correctly adjust the slope of the regression surface only and only in redundancy regions, while the slope of the regression surface unjustifiably increases in all the three suppression regions in such a way that geometrically speaking the regression surface sharply cuts the plane spanned by both x_1 and x_2; a condition that can be called "slope

dysregulation" (see Figure 3 and Figure 4). Again, the authors emphasize that no proportions of r_{12} can be added to the values of regression coefficients, and therefore the slope regulations affected by two-predictor suppression effects are completely erroneous and misleading.

6. Conclusion

This study depicts a clear picture of the performance of the statistical control function in different suppression and non-suppression situations, and provides a mathematical proof indicating that the statistical control function does not work correctly in suppression situations. These findings provide evidence that the regression parameters affected by suppression effects should be regarded as incorrect estimations. This study also introduces an algorithm that can generate numerous simulated datasets showing all different kinds of suppression and non-suppression situations known so far, and therefore they help resolve the theoretical complexities related to two-predictor suppression situations by expanding the pervious knowledge in this field. Based on these results, researchers are strongly recommended to examine their linear regression models to make sure that their results are not affected by suppression effects. These findings also provide important implications for the issue of "effect size" in linear regression and can change the educational contents and materials of the topic of two-predictor suppression effects in linear regression.

Like any other research, this study also involves important limitations. First, the case studies and examples include only models with two predictors. Second, only continuous quantitative variables are included, and further investigation on regression with categorical variables or a combination of continuous and categorical variables remains to be carried out. The implications of these findings for the issue of "effect size" in linear regression also need to be investigated in the future. Future research should focus on providing researchers with other applied algorithms or packages to help them detect suppression effects in their actual datasets for regression models with two or more predictors. Finally, an important question is how these findings and tools can be best incorporated into educational contents and materials.

References

Alexander, d. L. J., Tropsha, A. & Winkler, D. A., (2015). Beware of R(2): Simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of chemical information and modeling,* 55**,** pp. 1316–1322.

Brown, N., (2005). *Graphical illustration of two-predictor suppression effects, v0.05* [Online]. Available: https://steamtraen.shinyapps.io/suppressiongraphics/ [Accessed August 30 2020].

Cohen, J., Cohen, P., (1975). *Applied multiple regression/correlation analysis for the behavioral sciences,* Oxford, England, Lawrence Erlbaum.

Cohen, J., Cohen, P., (1983). *Applied multiple regression/correlation analysis for the behavioral sciences,* Hillsdale, N.J, Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. & Aiken, L., (2003). *Applied multiple regression/correlation analysis for the behavioral sciences,* Mahwah, NJ, Lawrence Erlbaum Associates.

Conger, A. J., (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement,* 34**,** pp. 35–46.

Conger, A. J., Jackson, D. N., (1972). Suppressor variables, prediction, and the interpretation of psychological relationships. *Educational and Psychological Measurement,* 32**,** pp. 579–599.

Currie, I., Korabinski, A., (1984). Some comments on bivariate regression. *Journal of the Royal Statistical Society: Series D (The Statistician),* 33**,** pp. 283–293.

Darlington, R. B., (1968). Multiple regression in psychological research and practice. *Psychological Bulletin,* 69**,** 161–182.

Darlington, R. B., Hayes, A. F., (2017). *Regression analysis and linear models: concepts, applications, and implementation,* New York, The Guilford Press.

Fox, J., (1997). *Applied regression analysis, linear models, and related methods,* Thousand Oaks, CA, US, Sage Publications, Inc.

Friedman, L., Wall, M., (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician,* 59**,** pp. 127–136.

Hamilton, D., (1987). Sometimes R 2 > r 2 yx 1 + r 2 yx 2 : Correlated variables are not always redundant. *The American Statistician,* 41**,** pp. 129–132.

Holling, H., (1983). Suppressor structures in the general linear model. *Educational and Psychological Measurement,* 43**,** pp. 1–9.

Horst, P., (1941). *The prediction of personal adjustment: A survey of logical problems and research techniques, with illustrative application to problems of vocational*

selection, school success, marriage, and crime, New York, NY, US, Social Science Research Council.

Kvalseth, T. O., (1985). Cautionary note about R². *The American Statistician,* 39**,** pp. 279–285.

Ludlow, L., Klein, K., (2014). Suppressor variables: The difference between 'is' versus 'acting as'. *Journal of Statistics Education,* 22**,** null-null.

Lynn, H. S., (2003). Suppression and confounding in action. *The American Statistician,* 57**,** pp. 58–61.

Mcfatter, R. M., (1979). The use of structural equation models in interpreting regression equations including suppressor and enhancer variables. *Applied Psychological Measurement,* 3**,** 123–135.

Meehl, P. E., (1945). A simple algebraic development of Horst's suppressor variables. *The American Journal of Psychology,* 58**,** pp. 550–554.

Mendershausen, H., (1939). Clearing variates in confluence analysis. *Journal of the American Statistical Association,* 34**,** pp. 93–105.

Nazifi, M., Fadishei, H., (2021a). *Supsim: A Python package simulating two-predictor suppression and non-suppression situations* [Online]. Available: https://github.com/fadishei/supsim [Accessed 11 May 2021].

Nazifi, M., Fadishei, H., (2021b). *Supsim: A Novel Computerized Algorithm Simulating Two-Predictor Suppression and Non-Suppression Situations* [Online]. Available: https://youtu.be/6K82yDp-fNM [Accessed 10 November 2021].

Nazifi, M., Fadishei, H., (2021c). *Supsim Project* [Online]. Available: https://supsim.netlify.app/supsim [Accessed 11 May 2021].

Neill, J. J., (1973). *Tests of the equality of two dependent correlations.* Doctoral dissertation, University of California, Ann Arbor, Ann Arbor, MI: University Microfilms No.: 74–7671.

Neter, J., Kutner, M., Nuchtsheim, C. & Wasserman, W., (1996). *Applied linear statistical models (4th ed),* Chicago, Irwin.

Pedhazur, E. (1997). *Multiple regression in behavioral research: Explanation and prediction,* Wadsworth, Thomson Learning.

Sharpe, N. R., Roberts, R. A., (1997). The relationship among sums of squares, correlation coefficients, and suppression. *The American Statistician,* 51**,** pp. 46–48.

Shieh, G., (2001). The inequality between the coefficient of determination and the sum of squared simple correlation coefficients. *The American Statistician,* 55**,** pp. 121–124.

Tzelgov, J., Henik, A., (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin,* 109**,** pp. 524–536.

Velicer, W. F., (1978). Suppressor variables and the semipartial correlation coefficient. *Educational and Psychological Measurement,* 38**,** pp. 953–958.

Watson, D., Clark, L. A., Chmielewski, M. & Kotov, R., (2013). The value of suppressor effects in explicating the construct validity of symptom measures. *Psychological Assessment,* 25**,** pp. 929–941.

Whuber, (2017). Generate a random variable with a defined correlation to an existing variable(s). Stack Exchange Inc.

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 203–215, DOI 10.2478/stattrans-2022-0050
Received – 10.10.2020; accepted – 19.05.2022

k-th record estimator of the scale parameter of the α-stable distribution

Michał Stachura[1], **Barbara Wodecka**[2]

ABSTRACT

Various techniques of scale parameter estimation have been proposed in the case of alpha stable distributions. In the paper, the authors present an estimation technique that involves the *k*-th record theory. Although this theory is over 40 years old, its implementation in the classical extreme value theory – being the other cornerstone of the presented approach – is quite new, and tempting. Several theoretical properties of the introduced scale parameter estimators are presented. With the use of Monte Carlo methods, a comparative analysis is performed between the approach based on *k*-th records and approaches based on Hill's and Pickands' estimators. Additionally, the paper uses a real-life data set to illustrate how to effectively apply the *k*-th record estimator of the scale parameter. The research indicates several advantages of the *k*-th record approach over its other counterparts, especially when dealing with incomplete information about the underlying sample.

Key words: stable distribution, scale parameter estimator, *k*-th record values.

1. Introduction

Specificity of many financial data sets (regarded as proper time-series) imposes that the so-called heavy-tailed distributions constitute an attractive alternative way of modelling such data. Amongst these distributions, the class of α-stable ones gained one of prominent places.

There are several methods of estimation of stability index α. However, for a complete recognition of a theoretical α-stable distribution that approximates empirical data, it is necessary to estimate the other parameters of the distribution, including the scale parameter σ as well. For instance, this holistic look is the most appropriate approach when calculating risks measures such as VaR or CVaR (see, e.g. Stoyanov et al. 2006, Khindanova et al. 2001).

[1] Department of Economics and Finance, Faculty of Law and Social Sciences, The Jan Kochanowski University, Kielce, Poland. E-mail: michal.stachura@ujk.edu.pl. ORCID: https://orcid.org/0000-0002-0115-3522.

[2] Department of Economics and Finance, Faculty of Law and Social Sciences, The Jan Kochanowski University, Kielce, Poland. E-mail: barbara.wodecka@ujk.edu.pl. ORCID: https://orcid.org/0000-0002-2427-1572.

Therefore, the present paper: 1) describes construction of k-th record estimators of parameter σ, in the case of stability index $\alpha < 2$, and 2) reveals some theoretical properties of the estimators introduced in the paper. The main goal of this article is to compare the quality of k-th record estimators of parameter σ with the two estimators of this parameter that are based on Hill's and Pickand's estimators. Such a comparative analysis is conducted by simulation research concerning some arbitrarily chosen range of α-stable distribution parameters ($1.8 \leq \alpha \leq 1.99$, $\beta = \mu = 0$, $0.01 \leq \sigma \leq 100$). Additionally, the paper is supplemented by an empirical example concerning energy prices quoted at the Nord Pool Spot.

The procedure for estimating the sigma parameter of the stable distribution described in this paper is part of a broader research trend that explores methods implementing the possibility of using k-th records in estimation. In the literature on the subject, one can find proposals for estimating the parameters of other distributions, such as: Gumbel's, Burr's, power, Weibull's, Rayleigh's, logistic or Pareto's ones (for instance see: Ahsanullah 1990, Malinowska et al. 2005). Moreover, k-th records, apart from the more classical approach, appear as a tool in Bayesian estimation (see: Malinowska and Szynal 2004).

2. Theoretical background

From now on, let $X_1, X_2, X_3, \ldots$ be *independent and identically distributed* (i.i.d.) random variables with a common *cumulative distribution function* (cdf) F. For any fixed $n \in \mathbb{N}_+$, the order statistics of a sample $X_1, X_2, \ldots, X_n$ are denoted by $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$.

The main theorem of the *extreme value theory* (EVT) states that if there exist constants $a_n > 0$, b_n for $n \in \mathbb{N}_+$, and some non-degenerate distribution function G such that for all $x \in \mathbb{R}$ holds $\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n:n}-b_n}{a_n} \leq x\right) = G(x)$, then there exists a constant $\gamma \in \mathbb{R}$ such that the limit distribution G has the form:

$$G(x) = G_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right) & 1+\gamma x > 0 \quad \gamma \neq 0 \\ \exp(-e^{-x}) & x \in \mathbb{R} \quad \gamma = 0 \end{cases}.$$

The parameter γ is called the *extreme value index* (EVI), and it impacts the right tail asymptotics of the common cdf F (e.g. see de Haan and Ferreira 2006).

Classical estimators of EVI are based on upper order statistics. Among wide variety of such estimators, the most popular are Pickands' and Hill's ones (see Gomes et al. 2008), given respectively by formulas:

$$\hat{\gamma}_{\mathrm{P}}^k = \log_2 \frac{X_{n-k:n}-X_{n-2k:n}}{X_{n-2k:n}-X_{n-4k:n}}, \quad \hat{\gamma}_{\mathrm{H}}^k = \tfrac{1}{k}\sum_{i=0}^{k-1} \ln X_{n-i:n} - \ln X_{n-k:n}. \qquad \text{(1A, B)}$$

for any fixed $k \in \{1,2, \ldots [n/4]\}$ (case 1A), or $k \in \{1,2, \ldots, n-1\}$ (case 1B).

An alternative, proposed by Berred (1995), is based on the notion of k-th records, which were defined by Dziubdziela and Kopociński (1976). So for a fixed $k \in \mathbb{N}_+$, the k-th record times $\{T_n^{(k)}\}$, and the k-th record values $\{R_n^{(k)}\}$ are defined by recurrence relations:

$$T_1^{(k)} = k,\ T_n^{(k)} = \min\{\, j \in \mathbb{N} : j > T_{n-1}^{(k)}, X_j > X_{T_{n-1}^{(k)}-k+1:\, T_{n-1}^{(k)}} \}, \text{ for } n \geq 2,$$

$$R_n^{(k)} = X_{T_n^{(k)}-k+1:\, T_n^{(k)}}.$$

In other words, a sequence of k-th record values $R_1^{(k)} < R_2^{(k)} < R_3^{(k)} < \ldots$ is constructed by eliminating repetitions in the non-decreasing sequence of k-th order statistics $X_{1:k} \leq X_{2:k+1} \leq X_{3:k+2} \leq \ldots$, while $T_1^{(k)} < T_2^{(k)} < T_3^{(k)} < \ldots$ are the appearance numbers (the so-called *record times*) of the succeeding record values.

The original Berred's estimator based on the k-th record values is of the form:

$$\hat{\gamma}_{\mathrm{B}}^{k} = \ln \frac{R_{N(k,n)}^{(k)} - R_{N(k,n)-k}^{(k)}}{R_{N(k,n)-k}^{(k)} - R_{N(k,n)-2k}^{(k)}}, \tag{2}$$

where $N(k,n)$ is a random number of k-th records values in a sample of size n.

Pickands' and Berred's estimators are convenient for any real γ (these estimators are additionally invariant under any linear transformation – with a positive slope – of data, which is fully concordant with the linear transformation appearing in the main EVT theorem), while Hill's one is proper for $\gamma > 0$ only. Moreover, Berred's estimator value depends on sample order, which allows resampling, since i.i.d. property is assumed. (The mentioned resampling makes sense only if data do not represent any time series.)

We recall one of equivalent definitions of α-stable distribution in order to assume the parametrization we use. Thus, a random variable X has α-stable distribution (noted as: $X \sim S(\alpha, \beta, \mu, \sigma)$) if the logarithm of its characteristic function ϕ is given by the following formula:

$$\ln \phi(t) = \begin{cases} i\mu t - \sigma^{\alpha}|t|^{\alpha}\left(1 - i\beta \operatorname{sign}(t) \tan\frac{\pi\alpha}{2}\right), & \alpha \neq 1 \\ i\mu t - \sigma t\left(1 + i\beta\frac{2}{\pi}\operatorname{sign}(t)\ln|t|\right), & \alpha = 1 \end{cases},$$

where $\alpha \in (0, 2\rangle$ is the stability index, $\beta \in \langle -1, 1\rangle$ is the skewness parameter, $\sigma \in (0, \infty)$ is the scale parameter, $\mu \in \mathbb{R}$ is the location parameter. It should be also mentioned that in α-stable case the following relation holds: $\gamma = 1/\alpha$ for $\alpha \in (0, 2)$, and $\gamma = 0$ for $\alpha = 2$, which reveals discontinuous functional dependence of α-stable tails asymptotics on the stability parameter value. Thus, the tails of the stable distributions have a power decay (are the so-called "heavy tails") if they are distinct from normal distribution (see Nolan 2011, Weron 2001).

Let $Z = |X|$ for $X \sim S(\alpha, \beta, \mu, \sigma)$, and let G and Q be the cdf, and quantile function of variable Z, respectively. Moreover, let $k = k_n$ be an increasing sequence of natural numbers such that the following condition is fulfilled:

$$k_n \to \infty \quad \text{and} \quad k_n/n \to 0, \quad \text{as } n \to \infty.$$

Basic properties of α-stable distribution tail yield that:

$$1 - G(x) \sim C_\alpha \sigma^\alpha x^{-\alpha}, \quad \text{as } x \to \infty, \tag{3}$$

for the constant $C_\alpha = \frac{2}{\pi}\Gamma(\alpha)\sin\frac{\pi\alpha}{2}$, and the gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ (see Samorodnitsky and Taqqu 1994, Nolan 2011). As a consequence (see Meraghni and Necir 2007) we obtain:

$$Q\left(1 - \frac{k}{n}\right)\left(\frac{k}{n}\,\frac{\pi}{2\Gamma(\alpha)\sin\frac{\pi\alpha}{2}}\right)^{1/\alpha} \to \sigma, \quad \text{as } n \to \infty. \tag{4}$$

The last convergence enables straightforward construction of estimators in the following manner. For a given sample Z_1, Z_2, …, Z_n of independent copies of Z, an unknown quantile $Q\left(1 - \frac{k}{n}\right)$ may be substituted by the appropriate order statistic $Z_{n-k:n}$ taken out of that sample. Additionally, if parameter α is also unknown, it may be substituted by any of its estimators, let us say $\hat{\alpha}_{(\mathrm{E})}^{n,k}$. It may be Hill's, Pickand's, Dekkers-Einmahl-de Haan's one (e.g. see de Haan and Ferreira 2006), to mention but a few. In the α-stable case, owing to the formula (3), these estimators may be applied to sample $Z_1, Z_2, \ldots, Z_n$ instead of sample $X_1, X_2, \ldots, X_n$.

Therefore, the estimator of the scale parameter takes the form:

$$\hat{\sigma}_{(\mathrm{E})}^{n,k} = Z_{n-k:n}\left(\frac{k\pi}{2n\Gamma\left(\hat{\alpha}_{(\mathrm{E})}^{n,k}\right)\sin\frac{\pi\hat{\alpha}_{(\mathrm{E})}^{n,k}}{2}}\right)^{1/\hat{\alpha}_{(\mathrm{E})}^{n,k}}, \tag{5}$$

which is quite general, but limited for 'order statistics' case.

It occurs that k-th records may be applied in the convergence (4), which leads to the following estimator:

$$\hat{\sigma}_{(\mathrm{R})}^{n,k} = R_{N(n,k)}^{(k)}\left(\frac{k\pi}{2n\Gamma\left(\hat{\alpha}_{(\mathrm{R})}^{n,k}\right)\sin\frac{\pi\hat{\alpha}_{(\mathrm{R})}^{n,k}}{2}}\right)^{1/\hat{\alpha}_{(\mathrm{R})}^{n,k}}, \tag{6}$$

as the ideas from original proofs of Meraghni and Necir (2007), concerning properties of the estimator (5), may be straightforwardly adapted to the 'k-th records' case.

To do this, it suffices to notice that $R_{N(n,k)}^{(k)} =^d Z_{n-k+1:n}$ (as $n \to \infty$) for any continuous probability distribution, where '$=^d$' designates equality in distribution

(see Wodecka 2016, Lemma 2.18). Moreover, the key is a direct consequence of the formula (3) that $Q\left(1-\frac{k-1}{n}\right)/Q\left(1-\frac{k}{n}\right)\to 1$ (as $n\to\infty$) for variable $Z=|X|$ defined above herein. So, the replacement of a proper order statistic $Z_{n-k:n}$ in (5) by one of its nearest neighbour $R^{(k)}_{N(n,k)}$ creates the formula (6).

As a result, the estimator $\hat{\sigma}^{n,k}_{(\mathrm{R})}$ is consistent if $k=k_n{\sim}dn^{\theta}$, as $n\to\infty$, for some constants $d>0$ and $\theta\in(0,1)$ (see Wodecka 2016, Theorem 2.19). Additionally, $\frac{\sqrt{k}}{\log\frac{k}{n}}\left(\log\hat{\sigma}^{n,k}_{(\mathrm{R})}-\log\sigma\right)\to^{\mathcal{D}}\mathcal{N}\left(0,\frac{e^{2/\alpha}+1}{\left(e^{1/\alpha}-1\right)^2\alpha^2}\right)$, as $n\to\infty$, where $\to^{\mathcal{D}}$ stands for convergence in distribution, which means that the estimator $\hat{\sigma}^{n,k}_{(\mathrm{R})}$ has asymptotically log-normal property (see Wodecka 2016, Theorem 2.20).

Moreover, in contrast to the order statistics, the formula (6) allows to estimate σ even in case of unknown sample size. For this purpose, one may use, for instance, the following estimators of the sample size:

$$\hat{n}_{\psi}=\psi^{-1}\left(\frac{N(k,n)}{k}+\psi(k)\right)-1,\ \hat{n}_{l}=k\,exp\left(\frac{N(k,n)}{k}\right)-1. \qquad (7\mathrm{A, B})$$

The above holds, since: a) $\mathbb{E}\big(N(k,n)\big)=k\sum_{i=k}^{n}\frac{1}{i}=k\big(\psi(n+1)-\psi(k)\big)$, where ψ is the digamma function $\psi(x)=\Gamma'(x)/\Gamma(x)$, and it has logarithmic asymptotics in infinity, and b) $Var\big(N(k,n)\big)=k\sum_{i=k}^{n}\frac{1}{i}-k^2\sum_{i=k}^{n}\frac{1}{i^2}$ is relatively small, as: $0<Var\big(N(k,n)\big)<\mathbb{E}\big(N(k,n)\big)$.

3. Study of the quality of estimators

3.1. Comparing estimators

For a while, let us consider quite general perspective, and let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of an unknown parameter θ. We assume that we wish to assess which of these estimators is "better" than the other. One of the criteria for solving this question is the *Pitman nearness measure* (see Pitman 1937) given as:

$$\mathbf{P}\big(\hat{\theta}_1,\hat{\theta}_2\big|\theta\big)=\mathbb{P}\big(\,\big|\hat{\theta}_1-\theta\big|<\big|\hat{\theta}_2-\theta\big|\,\big),$$

which indicates that $\hat{\theta}_1$ is *Pitman-closer estimator* than $\hat{\theta}_2$, if $\mathbb{P}\big(\hat{\theta}_1=\hat{\theta}_2\big)=0$, and $\mathbf{P}\big(\hat{\theta}_1,\hat{\theta}_2\big|\theta\big)>\frac{1}{2}$. The measure is very natural and intuitive, and additionally – as it preserves bivariate relation of both estimators, regarded as joint vector $(\hat{\theta}_1,\hat{\theta}_2)$ – it is very advisable, in contrast to such measures that rely only on univariate (marginal) distributions of both compared estimators.

Therefore, in the sequel we select Pitman nearness criterion as the main one, and we use it in every case that provides large enough bivariate sample size, in a sense of pairwise completeness. Otherwise, we decide to employ an analogue of the commonly

known mean square error. The chosen measure is given by $\eta_i = \left|\mathrm{Me}(\hat{\theta}_i) - \theta\right| + \frac{\mathrm{IQR}(\hat{\theta}_i)}{2}$ for $i \in \{1,2\}$, and we say that $\hat{\theta}_1$ is better than $\hat{\theta}_2$, providing that $\eta_1 < \eta_2$. We prefer the positional measure to classical ones since this approach is unlimited by incongruity to any theoretical assumptions including existence of the high order moments of distribution, as long as we consider the α-stable case (see: Stachura 2017).

3.2. Simulation study

In order to compare estimates of parameter σ based on record value theory with several selected estimates based on classical order statistic approach, simulation research is executed as follows, in the case of stability index $\alpha < 2$. (The simulation research, and additionally all the calculations and plots presented hereunder, are accomplished in R environment (R Core Team 2018).)

Firstly, for a fixed pair of parameters α and σ – taken from arbitrarily chosen ranges $\alpha \in \{1.8, 1.82, 1.84, 1.86, 1.88, 1.9, 1.92, 1.94, 1.96, 1.98, 1.99\}$, $\sigma \in \{0.01, 0.1, 1, 10, 100\}$ – and $\beta = 0, \mu = 0$, and for a fixed n – out of $\{50, 80, 110\}$ – pseudo-random i.i.d. sample of size n is generated (with the use of the R's package stabledist by Wuertz et al. 2016). The choice of α's range is motivated by the reason that the research by Stachura and Wodecka (2016), and Wodecka (2016) – including α's from 0.1 to 1.9 by 0.1 step – showed that the values of estimates were alarmingly discrepant near $\alpha = 2$, so the authors decided to examine the case of $\alpha \geq 1.8$ far more accurately, taking a tiny step 0.02. Besides, this new range integrates with α's detected in empirical research in financial data (just about 1.6 – 1.9 see e.g. Weron 2004). Next,

a. with respect to formulas (1A), (1A), (2), and the relation $\alpha = 1/\gamma$, estimators $\hat{\alpha}_{\mathrm{H}}^k$, $\hat{\alpha}_{\mathrm{P}}^k$, $\hat{\alpha}_{\mathrm{B}}^k$ are calculated on the basis of absolute values of a sample, for each possible k, which means $k \in K_n = \{1, 2, \ldots, [n/4] - 1\}$ (k-th records – necessary for $\hat{\alpha}_{\mathrm{B}}^k$ – are calculated in the R's package Records by Chrapek 2012)
b. each estimate $\hat{\alpha}_{\mathrm{H}}^k$, $\hat{\alpha}_{\mathrm{P}}^k$, $\hat{\alpha}_{\mathrm{B}}^k$ that is beyond the interval $(0, 2)$, is rejected and, as a consequence, omitted in the sequel (this is the reason why we deal with the already mentioned meaningful pairwise incompleteness of bivariate samples of estimates),
c. for all the other cases – based on formulas (5), (6) – estimates $\hat{\sigma}_{\mathrm{H}}^k$, $\hat{\sigma}_{\mathrm{P}}^k$, $\hat{\sigma}_{\mathrm{B}}^k$ are computed based on known sample size n,
d. concurrently with $\hat{\sigma}_{\mathrm{B}}^k$, considering formulas (7A, B), two additional estimates $\hat{\sigma}_{\psi}^k$, $\hat{\sigma}_{l}^k$ are calculated as if a sample size n was unknown.

Secondly, the previous step is replicated $J = 10000$ times independently, so that for any set of given α, σ, n, k we get five sequences $\hat{\sigma}_{\mathrm{H}}^k$, $\hat{\sigma}_{\mathrm{P}}^k$, $\hat{\sigma}_{\mathrm{B}}^k$, $\hat{\sigma}_{\psi}^k$, $\hat{\sigma}_{l}^k$ of sizes at most J (because of marginal incompleteness).

Thirdly, we perform "internal" comparative analysis of estimators $\hat{\sigma}_{\mathrm{H}}^k$, $\hat{\sigma}_{\mathrm{P}}^k$, $\hat{\sigma}_{\mathrm{B}}^k$, $\hat{\sigma}_{\psi}^k$, $\hat{\sigma}_{l}^k$ (within these 5 types of estimators separately) in order to indicate the best k for any

given n. The Pitman nearness measure is evaluated for any pair of distinct $k_1, k_2 \in K_n$, given a set of α, σ, and n. A demonstrative Table 1 presents values of Pitman nearness measure for estimates based on Berred's approach with known sample size ($\hat{\sigma}_{\mathrm{B}}^{k}$), with fixed $n = 50$, $\alpha = 1.9$, $\sigma = 1$. Tables of Pitman nearness measure for other estimators and other values of α, σ, n provide quite similar tables of matrices, whose dimensions vary depending on sample sizes.

Next, all the indications of which k provides better (in the sense of being Pitman-closer estimate) within the same sample size, against other k's, are counted up. This procedure leads to optimal choices of k's for a given estimator type and sample size n, which is presented in Table 2.

Table 1. Pitman-closer measures, comparing all possible k's (order k_1 is assigned to the first of compared estimators) – selected case of for $\hat{\sigma}_{\mathrm{B}}^{k}$, $n = 50$ $\alpha = 1.9$, $\sigma = 1$.

k_1 \ k_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	-	0.250	0.197	0.170	0.175	0.199	0.247	0.245	0.226	0.406	0.250
2	0.750	-	0.300	0.295	0.239	0.255	0.290	0.242	0.262	0.346	0.400
3	0.803	0.700	-	0.323	0.329	0.268	0.444	0.315	0.333	0.438	0.556
4	0.830	0.705	0.677	-	0.362	0.345	0.399	0.300	0.244	0.474	0.429
5	0.825	0.761	0.671	0.638	-	0.369	0.414	0.400	0.388	0.280	0.600
6	0.801	0.745	0.732	0.655	0.631	-	0.517	0.458	0.412	0.478	0.875
7	0.753	0.710	0.556	0.601	0.586	0.483	-	0.463	0.476	0.382	0.357
8	0.755	0.758	0.685	0.700	0.600	0.542	0.537	-	0.478	0.444	0.438
9	0.774	0.738	0.667	0.756	0.612	0.588	0.524	0.522	-	0.500	0.522
10	0.594	0.654	0.563	0.526	0.720	0.522	0.618	0.556	0.500	-	0.588
11	0.750	0.600	0.444	0.571	0.400	0.125	0.643	0.563	0.478	0.412	-

Source: own study.

Table 2. Optimal choices of k's for 5 types of estimators.

n	**50**	**80**	**110**
$\hat{\sigma}_{\mathrm{H}}$	5	6	7
$\hat{\sigma}_{\mathrm{P}}$	10	13	15
$\hat{\sigma}_{\mathrm{B}}$	8	10	12
$\hat{\sigma}_{\psi}$	10	13	14
$\hat{\sigma}_{l}$	10	13	14

Source: own study.

Fourthly, we perform "external" comparative analysis of the best estimators of each type. In contrast to the "internal" case, we are forced to rely on the previously introduced measure η. For a given sample size, and values of both parameters α and σ, measures η of each estimator are calculated, and then ranked. Demonstrative values of these measures, for fixed $n = 50$, $\alpha = 1.9$ and all five types of estimators, are gathered

in Table 3, while Table 4 includes their corresponding ranks (from 0 – the best to 4 – the worst).

Table 3. Values of measure η – case of $n = 50$ $\alpha = 1.9$, all σ's.

σ	0.01	0.1	1	10	100
$\hat{\sigma}_{\mathrm{H}}$	0.00897	0.08565	1.0553	8.634	85.68
$\hat{\sigma}_{\mathrm{P}}$	0.00666	0.05361	0.6323	6.975	70.52
$\hat{\sigma}_{\mathrm{B}}$	0.00467	0.06508	0.5912	5.935	68.1
$\hat{\sigma}_{\psi}$	0.00558	0.05985	0.5223	4.615	59.62
$\hat{\sigma}_{l}$	0.00541	0.05804	0.5032	4.393	57.66

Source: own study.

Next, within the type of estimator, single ranks are summed up in the whole range of σ's and the sums are ranked ("combined ranks") as the preconceived approach to estimation assumes naturally that the value of parameter σ is unknown – see the two last columns of demonstrative Table 4.

Finally, within the type of estimator, "combined ranks" are summed up simultaneously in the range of all sample sizes and all values of parameter α (partial illustration of this procedure is contained in Table 5). Again, the sums obtained in this way are ranked ("total ranks").

Table 4. Ranks of η values – case of $n = 50$ $\alpha = 1.9$, all σ's.

σ	0.01	0.1	1	10	100	sums of ranks	combined ranks
$\hat{\sigma}_{\mathrm{H}}$	4	4	4	4	4	20	4
$\hat{\sigma}_{\mathrm{P}}$	3	0	3	3	3	12	3
$\hat{\sigma}_{\mathrm{B}}$	0	3	2	2	2	9	2
$\hat{\sigma}_{\psi}$	2	2	1	1	1	7	1
$\hat{\sigma}_{l}$	1	1	0	0	0	2	0

Source: own study.

Table 5. Total ranks of η values.

σ	$n = 50$ $\alpha = 1.8$	$n = 50$ $\alpha = 1.82$	⋮	$n = 50$ $\alpha = 1.9$	⋮	$n = 80$ $\alpha = 1.9$	⋮	$n = 110$ $\alpha = 1.98$	$n = 110$ $\alpha = 1.99$	sums of combined ranks	total ranks
$\hat{\sigma}_{\mathrm{H}}$	4	4	...	4	...	4	...	4	4	130	4
$\hat{\sigma}_{\mathrm{P}}$	3	2	...	3	...	3	...	3	3	82	3
$\hat{\sigma}_{\mathrm{B}}$	2	3	...	2	...	1.5	...	2	2	69.5	2
$\hat{\sigma}_{\psi}$	1	1	...	1	...	0	...	0	0	15	0
$\hat{\sigma}_{l}$	0	0	...	0	...	1.5	...	1	1	33.5	1

Source: own study.

The reported procedure of making rankings shows that the 'k-th-record' estimators of the scale parameter are appraised to be the best ones. Interestingly, regardless of quite small discrepancies in the values of measure η, especially good performance characterises estimators assuming unknown sample size.

Moreover, quite similar indications may be noticed globally, with the use of scaled measure η. To do so, every value of η is divided by the adequate σ, which allows to carry out a comparative analysis of estimates that are obtained for different σ's. (It is imposed by a simple fact that $\sigma X \sim S(\alpha, 0,0,\sigma)$ for $X \sim S(\alpha, 0,0,1)$, and any $\sigma > 0$). Within almost all values of α, the best estimates of the scale parameter are those based on k-th-records (see Table 6). Additionally, the total sums of the scaled measure η confirm previous insights, and accentuate approximate quality level of estimation based on Pickands' and Berred's approaches.

Table 6. Summed values of measure η scaled by σ – within groups of α's.

α	$\widehat{\sigma}_{\mathrm{H}}$	$\widehat{\sigma}_{\mathrm{P}}$	$\widehat{\sigma}_{\mathrm{B}}$	$\widehat{\sigma}_{\psi}$	$\widehat{\sigma}_{l}$
1.8	11.41176	10.69419	10.63437	9.02193	9.29989
1.82	11.34495	10.10213	10.58824	9.28663	9.57081
1.84	12.10306	10.89211	10.28613	8.56202	8.77059
1.86	12.48158	10.64509	10.39769	8.55502	8.68862
1.88	13.09939	10.42322	9.78006	7.89665	8.18492
1.9	13.92617	10.54751	10.10855	9.11400	9.33091
1.92	13.89563	10.12217	9.98363	9.08145	9.29733
1.94	13.91820	10.44230	9.94078	8.01853	8.22405
1.96	14.83072	10.07188	9.39138	8.93096	9.15942
1.98	15.78266	10.75358	9.93393	8.53742	8.62024
1.99	19.33139	10.30731	10.04061	8.06486	8.21329
Total	152.12551	115.00149	111.08538	95.06947	97.36007

Source: own study.

4. Empirical example

To illustrate how the introduced estimation works in practice we consider electric energy prices in Finland quoted in euro at the Nord Pool Spot (www.nordpoolgroup.com). The chosen time series represents weekly prices from the 10th week of 2018 to the 9th week of 2020, which makes the sample size to be $n = 104$ (time span of two years). Figure 1 illustrates the mentioned data, and suggests SARMA-GARCH approach as an appropriate way to model the series. Such types of models are effectively applied for electricity market data (see for instance Aiube et al. 2013, Stachura and Wodecka 2016).

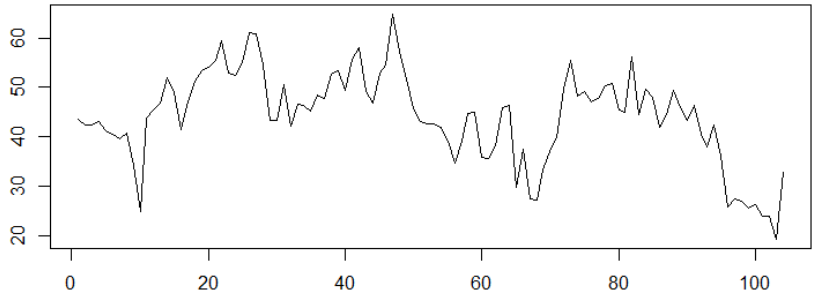


Figure 1. Week electricity prices in Finland.

Source: own study.

Amongst several initially estimated models (the R's package fGarch by Wuertz et al. 2020), the $\mathrm{SARMA}(1,1)_{13} \times \mathrm{GARCH}(1,1)$ occurred to be the best. Its residuals may be recognized as a random sample (Wald-Wolfowitz runs test p-value $= 0.9147$ – calculated with use of the R's package randtests by Caeiro and Mateus 2014) taken from normal distribution (Jarque-Bera test p-value $= 0.9201$, Shapiro-Wilk test p-value $= 0.9945$ – both calculated with the use of the R's package fGarch by Wuertz et al. 2020). The detected normality may as well indicate that residuals' distribution is α-stable with the stability parameter close to 2.

We decide to approximate the distribution of the residuals with a stable distribution $S(\alpha, 0, 0, \sigma)$. To do so, we fix record order $k = 3$. As formula (3) holds, we use formula (2) for absolute values of the residuals, obtaining $\hat{\alpha}_{\mathrm{B}}^{k} = 1.812339$. Then, formula (6) yields $\hat{\sigma}_{\mathrm{B}}^{k} = 0.895889$. It occurs that such gained approximation is accepted in view of two goodness of fit tests (Anderson-Darling test p-value $= 0.1051$ – calculated with use of the R's package goftest by Faraway et al. 2019, Kolmogorov-Smirnov test p-value $= 0.3532$).

5. Conclusions

The presentation of simulation research results gives some straightforward conclusions, which are as follows:

- 'k-th record' approach to estimation of the scale parameter σ is at least as good as the other classical methods presented herein (also, or even especially, assuming unknown sample size).
- 'k-th record' approach gives globally quite comparable results to Pickands' approach, which should not be surprising, as Berred's estimator is an analogue of Pickands' one.

- Estimation of σ based on Hill's estimator is distinctly characterized by the lowest stability in the sense that scale parameter estimates become more and more biased as stability index α tends to 2.
- 'k-th record' approach seems to be "unbeatable" in the region of stability index α very close to 2.

Concluding in general, it must be also remarked that the insights, hereinbefore specified, should be perceived essentially as the advantages of the 'k-th record' approach over the others presented, since the Berred's estimator, and the scale parameter estimator based on it, may be employed in cases of incomplete information about an underlying sample.

On the one hand, this incompleteness may be very useful if an analysed database must stay undisclosed, even for a researcher/statistician working on it, or more, the data are only partially recorded (i.e. record values of a proper order or several orders). On the other hand, if in contrary an analysed database is absolutely fulfilled and disclosed, the 'k-th record' approach opens up opportunities to make use of permutation methods in order to make repeated estimation that leads to much more precise results. Obviously, the key to success in the latter case is that the data correspond to i.i.d. random sample.

However, it should be pointed out that the 'k-th record' approach still requires a complete recognition of theoretical properties of the 'k-th record' estimator of the scale parameter, at least in a range of enhancing the results of Wodecka (2016) in the context of how fast is the 'k-th record' estimators' convergence.

References

Aiube, F. L., Baidya T. K. N., Blank, F. F., Mattos, A. B., Saboia W. and Siddiqui, A. S., (2013). Modeling Hourly European Electricity Spot Prices via a SARMA-GARCH Approach. *Working Paper of Stockholm University*, (260041), pp. 1–46.

Ahsanullah, M., (1990). Estimation of the parameters of the Gumbel distribution based on the m record values. *Comput. Statist. Quart*, 6, pp. 231–239.

Berred, M., (1995). K-record values and the extreme-value index. *J. Stat. Plan. Inference*, 45, pp. 49–63.

Caeiro, F., Mateus, A., (2014). Randtests: *Testing randomness in R*. R package version 1.0, https://CRAN.R-project.org/package=randtests

Chrapek, M., (2012). *Records: Record Values and Record Times*. R package version 1.0, https://CRAN.R-project.org/package=Records.

Dziubdziela, W., Kopocinski, B., (1976), Limiting properties of the k-th record values. *Zastosowania Matematyki*, 15, pp. 187–190.

Faraway, J., Marsaglia, G., Marsaglia, J. and Baddeley, A., (2019). Goftest: *Classical Goodness-of-Fit Tests for Univariate Distributions*. R package version 1.2-2, https://CRAN.R-project.org/package=goftest

Gomes, M. I., E Castro, L. C., Fraga, Alves, M. I. and Pestana, D., (2008). Statistics of extremes for IID data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions. *Extremes*, 11, pp. 3–34.

De Haan, L., Ferreira, A., (2006). Extreme Value Theory. An Introduction. *Springer*, New York.

Khindanova, I., Rachev, S. and Schwartz, E., (2001). Stable Modeling of Value at Risk. *Mathematical and Computer Modelling*, 34, pp. 1223–1259.

Meraghni, D., Necir, A., (2007). Estimating the Scale Parameter of a Lévy-stable Distribution via the Extreme Value Approach. *Methodol Comput Appl Probab*, 9(4), pp. 557–572.

Malinowska, I., Pawlas, P. and Szynal, D., (2005). Estimation of the parameters of Gumbel and Burr distributions in terms of kth record values. *Applicationes Mathematicae*, 32, pp. 375–393.

Malinowska, I., Szynal, D., (2004). On a family of Bayesian estimators and predictors for a Gumbel model based on the kth lower records. *Applicationes Mathematicae*, 31, pp. 107–115.

Nolan, J. P., (2011). Stable Distributions – Models for Heavy Tailed Data. *Birkhäuser*, Boston.

Pitman, E., (1937). The "closest" estimates of statistical parameters. *Mathematical Proceedings of the Cambridge Philosophical Society*, 33(2), pp. 212–222.

R Core Team, (2018). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria, URL https://www.R-project.org/.

Samorodnitsky, G., Taqqu, M. S., (1994). Stable Non-Gaussian Random Processes: *Stochastic Models with Infinite Variance*, Chapman & Hall, New York.

Stachura, M., (2017). On improved estimation of the extreme value index with the use of a shifted Hill's estimator. *Research Papers of the Wroclaw University of Economics*, 482, pp. 252–260.

Stachura, M., Wodecka, B., (2016). Wybrane aspekty i zastosowania modeli zdarzeń ekstremalnych, *Research Papers of the Wroclaw University of Economics,* 427, pp. 205–214.

Stoyanov, S., Samorodnitsky, G. and Rachev, S. T., (2006). Computing the portfolio Conditional Value-at-Risk in the α-stable case. *Probability and Mathematical Statistics,* 26(1), pp. 1–22.

Weron, R., (2001). Levy-stable distributions revisited: tail index > 2 does not exclude the Levy-stable regime. *International Journal of Modern Physics* C, 12 (2), pp. 209–223.

Weron, R., (2004). *Computationally intensive Value at Risk calculations*, Papers, Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE), No. 2004, 32.

Wodecka, B., (2016). *Wybrane aspekty i zastosowania modeli zdarzeń ekstremalnych*. Estymacja modeli na podstawie teorii wartości rekordowych, PhD Thesis, available at: http://hdl.handle.net/11089/20449.

Wuertz, D., Maechler, M. and Rmetrics Core Team Members, (2016). *Stabledist: Stable Distribution Functions. R package version 0.7-1*, https://CRAN.R-project.org/package=stabledist.

Wuertz, D., Setz, T., Chalabi, Y., Boudt, C., Chausse, P. and Miklovac, M., (2020). fGarch: Rmetrics – *Autoregressive Conditional Heteroskedastic Modelling*. R package version 3042.83.2, https://CRAN.R-project.org/package=fGarch

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 217–222

About the Authors

Asogwa Oluchukwu Chukwuemeka is a Lecturer I at the Department of Mathematics and Statistics, Faculty of Physical Sciences, Alex Ekwueme Federal University Ndufu Alike, Nigeria. His research interests are design and analysis of experiments, response surface designs, construction of designs, multivariate statistical analysis, statistical inference and data analysis in particular. Oluchukwu has published over 18 research papers in international/national journals and conferences. He is an active member of many national scientific professional bodies.

Elshaarawy Rasha S. received the PhD degree in Statistics from Cairo University, Egypt in 2022. She has been the Head of Statistical Department at the Ministry of Youth and Sports since 2012 until now. Her research interests include: ranked set sampling, stress strength models, multicomponent stress-strength models, nonparametric hypothesis testing and its applications, queueing theory, probability distributions, linear and nonlinear models.

Fadishei Hamid is an Assistant Professor at the Engineering Department, University of Bojnord, Iran. He received both his MSc and PhD degrees in Computer Engineering from Amirkabir University of Technology and Ferdowsi University of Mashhad, respectively. His research interests include data mining, big data systems and analytics, machine learning, and internet of things.

Hanagal David D. is an Emeritus Professor at the Department of Statistics, Savitribai Phule Pune University, India. An elected fellow of the Royal Statistical Society, UK, he serves on editorial boards of several respected international journals. He has authored four books and three book chapters, and published over 150 research publications in leading journals. He has delivered over 100 invited talks on many national and international platforms of repute worldwide. He also has worked as a visiting Professor at several universities in the USA, Germany, and Mexico. His research interests include statistical inference, selection problems, reliability, survival analysis, frailty models, Bayesian inference, stress-strength models, bootstrapping, censoring schemes, distribution theory, multivariate models, characterizations, repair, and replacement models, software reliability, quality loss index.

Hassan Amal Soliman is a Full Professor of Statistics at the Department of Mathematical Statistics, Faculty of Graduate Studies for Statistical Research at Cairo University, Egypt. She earned her PhD in Statistics from Cairo University's Faculty of

Graduate Studies for Statistical Research. Her main research interests are: probability theory, record values, ranked set sampling, stress-strength models, accelerated life tests and goodness of fit tests. Professor Hassan has published over 150 research papers in international/national journals and conferences. Currently, she is a member of two editorial boards. She has been a reviewer for numerous prestigious journals.

Hindls Richard, Professor of Statistics. In 2001–2006 he was a Dean of the Faculty of Informatics and Statistics of the University of Economics in Prague, and in 2006–2014 he was the President of the University of Economics in Prague. He focuses on the economic aggregates, the analysis of time series and the application of statistical methods in the auditing. He has published 40 books and about 250 articles. He is active in many institutions and scientific councils (Czech Statistical Council, Association of National Accounting Paris, European Advisory Committee for Economic and Social Statistics and others). In 2012, he received the Medal for the development of cooperation between Poland and the Czech Republic in the area of statistics (awarded by the Polish Statistical Society on the occasion of its 100th anniversary).

Hronová Stanislava, Professor at the Department of Economic Statistics in University of Economics, Prague. From 2001 to 2006 she was a Vice-Dean for Research of the Faculty of Informatics and Statistics, in 2006–2014 the Vice-President for Research of the University of Economics, Prague. From 2010 to 2014 she was a member of the Government Research and Development Council; from 2014–2022 she was a Vice-president of the Czech Science Foundation. She is interested in national accounts and economics statistics. She has co-operated with the Czech Statistical Office in the area of statistical methodology. She was awarded the French Ordre des Palmes Académiques. She is an active member of many scientific bodies. She is Editor-in-Chief of Statistika journal and a member of the Editorial Board of Politická ekonomie journal and Silesian Statistical Review.

Jiratampradab Arisa is a graduate student at the Department of Statistics, Faculty of Science, Kasetsart University. Her research interests are probability theory, statistical inference, and regression analysis.

Liu Yang is a PhD candidate in Statistics at Worcester Polytechnic Institute, Massachusetts, USA. Her research interests include Bayesian small area estimation, nonparametric Bayesian and data integration. Prior to joining WPI, she was a research statistician in the Neuroscience Statistics Lab at Massachusetts Institute of Technology, Massachusetts, USA.

Lubos Marek, Professor of Statistics. He works as the Head of Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, the Czech Republic. He studied the Mathematical Statistics at Charles University, Prague. His main research field concentrates on data analysis, probability, stochastic

processes and time series analysis. He worked as a member of some scientific projects. He worked 7 years as the Head of research project „Methods of knowledge acquisition from data and their use in economical decision-making". He is an author of several textbooks and monographs, many conference papers and journal articles. He is a member of Czech and international statistical societies, and member of several scientific boards.

Molefe Wilford is statistician by profession and a trained in survey research methodology. He holds a PhD in Survey Research Methodology (sample design for small area estimation) from the University of Wollongong, New South Wales, Australia. He is currently a senior lecturer and the Head of the Department of Statistics, an active researcher, whose primary research interest lies in sample survey methodology, and a consultant on a wide range of issues and for a varied client base. He also holds an MSc in Statistics from the University of Sheffield, United Kingdom, a Post-graduate diploma in Science from the University of Sheffield and BA in Statistics degree from the University of Botswana.

Nagy Heba F. is a Lecturer of in the Department of Mathematical Statistics, at the Faculty of Graduate Studies for Statistical Research, Cairo University, Giza, Egypt. Her main research interests include: ranked set sampling, stress strength models, reliability estimation, estimation of unknown parameters, entropy estimation, goodness of fit tests, probability distributions and record values.

Nandram Balgobin has been a Full Professor of Statistics at Worcester Polytechnic Institute since 2003. He does research in Bayesian statistics and small area estimation interfacing. He is an Adjunct Professor at three other major universities. He has written numerous research articles in statistical journals such as the Journal of the American Statistical Association, where he was a two-term Associate Editor. He has supervised many MS and PhD dissertations both nationally and internationally. He is a fellow of the ASA and an elected member of the ISI and Sigma Xi. In 2003–2005, he was Sinclair Professor of Mathematical Sciences at WPI and in 2006 he received the prestigious SPAIG award for WPI from the ASA. Since 2015, he has been a Senior Research Scientist at the National Agricultural Statistics Service, USDA.

Nazifi Morteza is an Assistant Professor at the Department of Psychology, University of Bojnord, Iran. He is Editor-in-Chief of the journal of "Mental Health: Research and Practice (MHRP)", which has been recently published by the University of Bojnord. His main area of interest is child and adolescent psychology and education. He is also interested in multivariate statistical analysis as well as advanced research methodology in psychology.

Obasi John is a Graduate/Researcher of the Department of Statistics, University of Nigeria. He is presently working as the Head of Business Unit of Ajatextechno and Data

Solutions. His research interests include design and analysis of experiments, Bayesian methods and data science for business. He is a professional member of Nigerian Statistical Association and Nigerian Mathematical Society.

Oladugba Abimibola Victoria is a Lecturer/Researcher at the Department of Statistics, Faculty of Physical Sciences, University of Nigeria, Nsukka. Her areas of research interest include: Response Surface Methodology, Design and Analysis of Experiments, Categorical Data Analysis and, Missing and Outlier Data Analysis. Abimibola has published over 30 articles in national/international journals and conferences. She is a Fellow, Royal Statistical Society (*FRSS*), and a member of the Nigerian Statistical Association, the International Statistical Institute and Caucus of Women in Statistics.

Pandey Arvind, MSc (Statistics), JRF-CSIR (Mathematical Sciences-2004), NET-CSIR (Mathematical Sciences – 2002), GATE (Engineering Sciences – 2001), PhD (Modeling Frailty for Bivariate Survival Data), is an Associate Professor at the Department of Statistics and the Head of Department of Data Science at the Central University of Rajasthan, Ajmer, India. His research interest areas are survival analysis, statistical inference, distribution theory, and frailty models. He has written 5 book chapters and over 75 research papers that have been published in prestigious journals. He has completed two minor research projects funded by UGC.

Shukla Diwakar is presently working as a Professor in Statistics, Department of Mathematics and Statistics, H.S. Gour University, Sagar, MP, India. He has obtained MSc (Statistics), PhD (Statistics), MTech (Computer Science) and has over 30 years of teaching experience. During his PhD, he was recipient of Research Fellowship of CSIR. Until now he has published over 150 research papers and supervised 20 PhD theses. He was a recipient of MPCOST Young Scientist Award, ISAS Young Scientist Medal and UGC Career Award. He authored 8 books and is a member of 11 learned bodies. His area of research interest are sampling theory, graph theory, stochastic modelling, big data, data mining, computer network, scheduling and operating system.

Singh Thakur Narendra is an Assistant Professor in Statistics with over 10 years of teaching and research experience. He has published over 30 research papers as author and co-author in various journals. He is a life member of Indian Science Congress Association (ISCA), Journal of Reliability and Statistical Studies (JRSS), Pantnagar (UP) and Madhya Bharti, A Journal of Dr. Harisingh Gour Central University, Sagar (M.P). He has published two books in survey sampling for missing observations estimation as an author and co-author. He has reviewed several research papers in different national and international journals. His area of research interest is Sampling Theory.

Stachura Michał is an Assistant Professor at the Department of Economics and Finance of the Faculty of Law and Social Sciences of the Jan Kochanowski University in Kielce, Poland. The area of his research interests includes the theory and applications of

statistics and econometrics in finance, health economics, and broadly understood social sciences. He has been involved in several interdisciplinary research projects in cooperation with various universities and national research institutes, where he applied his expertise in geological, legal, and systemic risk studies.

Sulewski Piotr graduated in Mathematics in 1996. Since then he has been working at the Institute of Mathematics at Pomeranian University in Słupsk. He received PhD in reliability theory in 2001 from the Systems Research Institute of Polish Academy of Sciences in Warsaw. His research interests concern mathematical statistics, computational methods in statistics and reliability theory. Piotr Sulewski has published about 50 research papers in international/national journals. He has also published three books/monographs.

Suntornchost Jiraphan is an Assistant Professor at the Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University. Her research interests are probability theory, mathematical statistics, statistical modelling, stochastic processes, actuarial mathematics, and mathematical finance.

Supapakorn Thidaporn is an assistant professor at the Department of Statistics, Faculty of Science, Kasetsart University. Her research interests are probability theory, mathematical statistics, statistical inference, statistical modelling, and forecasting techniques.

Szymkowiak Magdalena received her PhD degree in Mathematics from Adam Mickiewicz University in Poznan and she was an Associate Professor at the Institute of Mathematics at 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 discrete mathematics, data mining, reliability theory, characteristics of distributions and stochastic orders.

Tharshan Ramajeyam is a Lecturer at the Department of Mathematics and Statistics, University of Jaffna, Sri Lanka. He received his undergraduate degree from University of Jaffna, Sri Lanka and MSc degree in Statistics from Wright State University, USA. His research interests are in the areas of finite mixture models, models for over-dispersed count data, statistical inference, and data analysis. Currently, he is a PhD student in Statistics at the Postgraduate Institute of Science, University of Peradeniya, Sri Lanka.

Tyagi Shikhar, MSc (Statistics), MPhil (Statistics), PhD (Some Contributions to Frailty Models for Survival Data) has been an Assistant Professor at the Department of Data Science, Christ Deemed to be University, Bangalore, India since September 2022. He also worked as an Assistant Professor at the Department of Mathesmatics, GITAM Deemed to be University, Bangalore, India. He worked as a referee for several journals.

He is conducting research in several fields such as probability theory, survival analysis, parametric Bayesian inference, copula model, and distribution theory, and he specializes in frailty models. He has 13 research paper publications and has authored 5 book chapters in different reputed national and international journals.

Wijekoon Pushpakanthie received her PhD degree from the University of Dortmund, Germany. She is currently a Senior Professor at the Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka. The scope of her research interests include biased estimation, Stein-Rule estimation, and preliminary test estimation in the linear regression model, misspecified linear models, binomial and Poisson mixture models, improved methods in estimation in the exponential family of distributions and multivariate statistical methods. She has published over 50 research papers in journals and conferences throughout her professional career. She has also served as a referee for various reputed national and international journals. Professor Wijekoon is also an active member of many scientific professional bodies.

Wodecka Barbara is an Assistant Professor at the Department of Economics and Finance of the Faculty of Law and Social Sciences of the Jan Kochanowski University in Kielce, Poland. Her area of research includes mainly modelling of extreme events using the record value theory in broadly understood social sciences.

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 223–228

Acknowledgements to Reviewers

The Editor and Editorial Board of the 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 23, Numbers 1–5, the authors` work has benefited from their feedback.

Abid Salah Hamza, Department of Mathematics, Mustansiriyah University, Iraq

Adeleke Maradesa, Department of Statistics, Federal University of Technology Akure, Nigeria

Al-Qati Ahmed Baqer Jaafar, Department of Mathematics, University of Thi-Qar, Iraq

Amin Muhammad, Department of Statistics, Nuclear Institute for Food and Agriculture (NIFA), Pakistan

Andrzejczak Karol, Department of Mathematics, Poznan University of Technology, Poland

Bartl David, Department of Informatics and Mathematics, Silesian University in Opava, Czech Republic

Baszczyńska Aleksandra, Department of Statistical Methods, University of Lodz, Poland

Bąk Andrzej, Department of Econometrics and Informatics, Wroclaw University of Economics and Business, Poland

Benaissa Samir, Department of Mathematics, University of Sidi-Bel-Abbes, Algeria

Betti Gianni, Department of Political Economy and Statistics, University of Siena, Italy

Bhatti Ishaq, Department of Economics, Finance & Marketing, La Trobe University, Australia

Bieniek Milena, Institute of Management and Quality Sciences, Maria Curie-Sklodowska University, Poland

Bieszk-Stolorz Beata, Department of Econometrics and Statistics, University of Szczecin, Poland

Brombin Chiarra, Faculty of Applied Statistics, Vita-Salute San Raffaele University, Italy

Bwanakare Second, Department of Economics and Finance, Cardinal Stefan Wyszyński University in Warsaw, Poland

Chen Lu, Department of Statistics, USDA, USA

Chessa Antonio, Statistics Netherlands, The Netherlands

Chesneau Christophe, Department of Mathematics, University of Caen, France

Chipman Jonathan, Department of Biostatistics, University of Utah School of Medicine, USA

Chouaf Abdelhak, Department of Probability-Statistics, Djillali Liabes University, Algeria

Cui Yifan, Center for Data Science, Zhejiang University, China

Czapiewski Konrad, Department of Rural Geography and Local Development, Polish Academy of Science, Poland

Dietrich Alexander, Department of Economics, University of Tübingen, Germany

Dihidar Kajal, Sampling and Official Statistics Unit, Indian Statistical Institute, India

Domański Czesław, University of Lodz, Poland

Dryver Arthur, Department of Statistics, School of Applied Statistics, Thailand

Dziechciarz Józef, Department of Econometrics and Informatics, Wroclaw University of Economics and Business, Poland

Ezzebsa Abdali, Department of Mathematics, Guelma University, Algeria

Farooq Dar Qaiser, Division of International Trade, Incheon National University, South Korea

Gatnar Eugeniusz, Department of Economic and Financial Analysis, University of Economics in Katowice, Poland

Goroncy Agnieszka, Department of Mathematics and Computer Science, Nicolaus Copernicus University, Poland

Górka Joanna, Department of Econometrics and Statistics, Nicolaus Copernicus University in Toruń, Poland

Graczyk Małgorzata, Department of Mathematical and Statistical Methods Poznan University of Life Sciences, Poland

Grzenda Wioletta, Department of Statistics and Demography, SGH Warsaw School of Economics, Poland

Guo Huizhen, Department of Mathematics, School of Natural Sciences& Mathematics, USA,

Młodak Andrzej, Inter-faculty Department of Mathematics and Statistics, The president Stanislaw Wojciechowski Calisia University, Poland

Nwogu-Ikojo Leonard Eluma, Department of Mathematics, University of Nigeria, Nigeria

Okrasa Włodzimierz, Cardinal Stefan Wyszyński University in Warsaw & Statistics Poland, Poland

Olubusoye Olusanya Elisa, Department of Statistics, University of Ibadan, Nigeria

Omay Tolga, Department of Economics, Atilim University, Turkey

Oral Evrim, Department of Statistics, LSU Health Sciences Center New Orleans, USA

Osaulenko Oleksandr, National Academy of Statistics, Accounting and Audit, Ukraine

Quintana Fernando A., Department of Statistics, University Catholics (UC) in Chile, Chile

Panek Tomasz, Department of Statistics and Demography, SGH Warsaw School of Economics, Poland

Pascual Francis, Department of Mathematics and Statistics, Washington State University, USA

Peer Bilal Ahmad, Department of Mathematical Sciences, Islamic University of Science and Technology (IUST), India

Prasad Shakti, Department of Basic & Applied Science, National Institute of Technology, India

Raczkiewicz Dorota, Unit of Demographic, SGH Warsaw School of Economics, Poland

Rejchel Wojciech, Department of Mathematical Statistics and Data Analysis, Nicolaus Copernicus University in Torun, Poland

Rogala Tomasz, Department of Mathematics, Cardinal Stefan Wyszyński University in Warsaw, Poland

Rozkrut Dominik, President of Statistics Poland, Poland

Rybicki Wojciech, Department of Management, General Tadeusz Kościuszko Military University of Land Forces, Poland

Saegusa Takumi, Division of Statistics, University of Maryland, USA

Safari Hadi, School of Mathematical and Statistical Sciences, Southern Illinois University, USA

Sangnawakij Patarawan, Department of Mathematics and Statistics, Thammasat University, Thailand

Wilak Kamil, Department of Statistics, Poznan University of Economics and Business, Poland

Wycinka Ewa, Department of Statistics, University of Gdansk, Poland

Wywiał Janusz, Department of Statistics, Econometrics and Mathematics, University of Economics in Katowice, Poland

Yin Jiani, Department of Biostatistics, Takeda Pharmaceuticals, USA

Zimková Emília, Department of Finance and Accounting, Matej Bel University in Banská Bystrica, Slovakia

Żądło Tomasz, Department of Statistics, Econometrics and Mathematics, University of Economics in Katowice, Poland

STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 229–234

Index of Authors, Volume 23, 2022

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).