Volume 23, Number 3, September 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

Research articles

Other articles

XXXVIII Multivariate Statistical Analysis 2019, Lodz. Conference Papers

Research Communicates and Letters

Conference acknowledgement

Volume 23, Number 3, September 2022

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, 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, September 2022
Vol. 23, No. 23, 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, September 2022
Vol. 23, No. 3, pp. VII–XII

From the Editor

It is with great pleasure that we present our readers with the September issue consisting of 12 articles arranged, as usual, in three sections: *Original research articles*, *Other articles*, and *Research communicates and letters.* A wide spectrum of topics is discussed in these papers by authors from a large group of countries: USA, India, Iran, Algeria, Saudi Arabia, Sri Lanka, Nigeria, Thailand, and Poland.

Research articles

The issue starts with the paper by **Jacek Białek**, **Tomasz Panek**, and **Jan Zwierzchowski** entitled ***Assessing the effect of new data sources on the Consumer Price Index: a deterministic approach to uncertainty and sensitivity.*** The authors discuss the use of alternative sources of data about prices (scanned and scraped data) in the analysis of price dynamics with selecting the appropriate formula of the price index at the elementary group (5-digit) level as the one of the greatest challenges of the official statistics. The empirical study was based on data for February and March 2021, while scanner and scraped data about selected categories of food products were obtained from one retail chain operating hundreds of points of sale in Poland and selling products online. It was found that the choice of a data source has the most significant impact on the final value of the index at the elementary group level, while the choice of the aggregation formula used to consolidate different data sources is of secondary importance. The results indicate that consumer price indices calculated for the elementary groups of interest are characterised by a relatively low robustness to changes of a data source about consumer prices and by a relatively high robustness to changes of index formulas used for calculating price indices at the level of subgroups and elementary groups. For all elementary groups of interest, the first assumption, i.e. the choice of a given data source, has the biggest impact on the final value of the price index. Which index formula is used for aggregating indices for subgroups into elementary group indices has much less influence on the final results. The effect of choosing a particular index for aggregating indices derived from different sources is negligible.

Yeil Kwon in the article entitled ***A comparison of the method of moments estimator and maximum likelihood estimator for the success probability in the Fibonacci-type probability distribution*** shows that a Fibonacci-type probability distribution provides the probabilistic models for establishing stopping rules associated with the number of consecutive successes. It can be interpreted as a generalized version of a geometric distribution. After revisiting the Fibonacci-type probability distribution to explore its definition, moments and properties, the authors proposed numerical methods to obtain two estimators of the success probability: the method of moments estimator (MME) and maximum likelihood estimator (MLE). The ways both of them performed were compared in terms of the mean squared error. A numerical study demonstrated that MLE tends to outperform MME for most of the parameter space with various sample sizes. A Fibonacci-type probability distribution can be employed to determine the probabilistic behaviour of a random variable N defined by the number of Bernoulli trials with a success probability p until there are k-consecutive successes. To compare MME with MLE, the authors used the computational methods to obtain MLE by approximating the maximum likelihood function using the pmf of N defined recursively. The result of the simulation discloses that, for both MLE and MME, the biases are considerably smaller than the variances under all of the values of p and the sample sizes, indicating that the variance explains the majority of MSE.

Mriganka Mouli Choudhury, **Rahul Bhattacharya**, and **Sudhansu S. Maiti** discuss ***Estimation of $P(X \leq Y)$ for discrete distributions with non-identical support***. The Uniformly Minimum Variance Unbiased (UMVU) and the Maximum Likelihood (ML) estimations of $R = P(X \leq Y)$ and the associated variance are considered for independent discrete random variables X and Y. Assuming a discrete uniform distribution for X and the distribution of Y as a member of the discrete one parameter exponential family of distributions, theoretical expressions of such quantities are derived. Similar expressions are obtained when X and Y interchange their roles and both variables are from the discrete uniform distribution. A simulation study is carried out to compare the estimators numerically. A real application based on demand-supply system data is provided. The UMVU and ML estimations of $P(X \leq Y)$ considering a discrete uniform distribution to represent stress and/or strength were discussed. However, an assumption of equal (but unknown) probability for stress and/or strength is less practical. Consequently, the authors intend further development with a general class of distributions to model stress and/or strength, allowing non-identical and parameter dependent supports.

In the next paper, ***Interval shrinkage estimation of parameter of exponential distribution in presence of outliers under loss functions***, **Parviz Nasiri** examines estimators based on an interval shrinkage with equal weights point shrinkage estimators for all individual target points $\theta^{-} \in (\theta 0,\theta 1)$ for exponentially distributed observations in the presence of outliers drawn from a uniform distribution. The estimators obtained from both shrinkage and interval shrinkage were compared, showing that the estimators obtained via the interval shrinkage method perform better. The symmetric and asymmetric loss functions were also used to calculate the estimators. Finally, a numerical study and illustrative examples were provided to describe the results. It is shown that the interval shrinkage estimator is better than the shrinkage estimator. Using different loss functions can also improve the performance of the estimator. The proposed method can be extended for Bayesian interval shrinkage estimation and other positive data distributions as well as for the presence of outliers from other distributions.

Adam Szulc focuses on ***Polish inequality statistics reconsidered: are the poor really that poor?*** The author critically analyses the data problem typically present in studies of income inequality. According to most empirical studies based on household surveys and considering the European standards, the recent income inequality in Poland is moderate and decreased significantly after reaching its peaks during the first decade of the 21st century. These findings were challenged by Brzeziński et al. (2022), who placed Polish income inequality among the highest in Europe. Such a conclusion was possible when the household survey data and information on personal income tax are analysed jointly. In this study the above-mentioned findings are explored using 2014 and 2015 data employing additional corrections to the household survey incomes. Incomes of the poorest people are replaced by their predictions made on a large set of well-being correlates, using the hierarchical correlation reconstruction. Applying this method together with the corrections based on Brzeziński's et al. results reduces the 2014 and 2015 revised Gini indices, still keeping them above the values obtained with the use of the survey data only. It seems that the hierarchical correlation reconstruction offers more accurate proxies to the actual low incomes, while matching tax data provides better proxies to the top incomes. The results of the present study only partly confirm findings by Brzeziński et al. (2022) on the serious underestimation of the Polish inequality indices. Corrections of the 2014 and 2015 survey income data applied to both tails of the distribution also results in inequality growth, however not so high and not for all types of inequality measures.

In their paper ***New polynomial exponential distribution: properties and applications,*** **Abdelfateh Beghriche**, **Halim Zeghdoudi**, **Vinton Raman**, and **Sarra Chouia** discuss the general concept of the XLindley distribution. Forms of density and hazard rate functions are investigated in the manuscript. Moreover, precise formulations for several numerical properties of distributions are derived. Extreme order statistics are established using stochastic ordering, the moment method, the maximum likelihood estimation, entropies and the limiting distribution. The authors demonstrate the new family's adaptability by applying it to a variety of real-world datasets, and a suggested family of distributions with only one parameter. Moments, distribution function, characteristic function, failure rate, stochastic order, maximum likelihood approach, and method of moments were among the properties studied. The Lindley and Zeghdoudi distributions lack the flexibility needed to examine and model many forms of data related to lifetime data and survival analysis. The NPED distribution, on the other hand, is adaptable, straightforward, and simple to use. The novel distribution was used to evaluate two real data sets and was compared to existing distributions (Lindley, exponential, Zeghdoudi, Exponential Exponential and Xgamma). The comparison's findings support the NPED distribution's quality adjustment. The authors anticipate that the new distribution family will entice many additional life data, reliability analysis, and actuarial science applications, and it can employ a more general distribution with two parameters in future experiments.

Varathan Nagarajah's paper ***An improved ridge type of estimator for logistic regression*** demonstrates how to overcome the effect of multicollinearity in logistic regression. The proposed estimator is called a modified almost unbiased ridge logistic estimator (MAURLE). It is obtained by combining the ridge estimator and the almost unbiased ridge estimator. In order to assess the superiority of the proposed estimator over the existing estimators, theoretical comparisons based on the mean square error and the scalar mean square error criterion are presented. A Monte Carlo simulation study is carried out to compare the performance of the proposed estimator with the existing ones. Finally, a real data example is provided to support the findings. The superiority conditions for the proposed estimator with the existing MLE, LRE, and AURLE estimators are derived with respect to the MSE and SMSE criterions. Further, from the real data application and the Monte Carlo simulation study the authors notice that the proposed estimator performs well compared to MLE, LRE, and AURLE when the multicollinearity among the explanatory variables is high.

Sergiusz Herman examines ***Impact of restrictions on the COVID-19 pandemic situation in Poland*** focusing on the question of how the lockdown introduced in Poland affected the spread of the pandemic in the country. The study used the synthetic control method to this end. The analysis was carried on the basis of data from the Local Data Bank and a government website on the state of the epidemic in Poland. Results show that lockdown is an efficient tool that curbs the spread of the COVID-19 pandemic in Poland. Thanks to it, the number of new cases in the analysed region (in Poland) has diminished significantly (a drop of 9500 cases in three weeks was observed). Such a conclusion can be drawn from the performance of the placebo-in-space and placebo-in-time analyses. The research included the construction of the synthetic region that well illustrated the tendency of the pandemic development (in the Warmińsko-Mazurskie region) before the lockdown. After that the virus spread trajectories started to differ considerably.

In the next paper, **Wachirapond Permpoonsinsup** and **Rapin Sunthornwat** present ***Modified exponential time series model with prediction of total Covid-19 cases in Belgium, Czech Republic, Poland, and Switzerland***. The outbreaks in Belgium, the Czech Republic, Poland and Switzerland entered the second wave and was exponentially increasing between July and November, 2020. Consequently, the authors estimated the compound growth rate, to develop a modified exponential time-series model compared with the hyperbolic time-series model, and the optimal parameters for the models based on the exponential least-squares, three selected points, partial-sums methods, and the hyperbolic least-squares for the daily COVID-19 cases in Belgium, the Czech Republic, Poland and Switzerland. The speed and spreading power of COVID-19 infections were obtained by using derivative and root-mean-squared methods, respectively. The optimal forecasting model with the estimated parameters was selected based on having the lowest RMSPE and the highest 2R. The results show that the exponential least-squares method was the most suitable for the parameter estimation. The compound growth rate of COVID-19 infection was the highest in Switzerland, and the speed and spreading power of COVID-19 infection were the highest in Poland between July and November, 2020. In conclusion, the authors maintain that the exponential least-squares method was relatively the most appropriate method for parameter estimation for the modified exponential time-series model for the daily COVID-19 cases, in all four countries.

Ahmad Aijaz, S. Qurat ul Ain's, Ahmad Afaq, and **Rajnee Tripathi's** article ***Poisson area-biased Ailamujia Distribution with applications in environmental and medical science***. The authors used compounding to develop a new distribution. A new Poisson area-biased Ailamujia distribution has been formulated to analyse count data. It was created by combining two distributions: the Poisson and area-biased Ailamujia distributions, using the compounding technique, to analyse count data. Several distributional properties of the formulated distribution have been studied. The distribution's ageing characteristics were determined and expressed explicitly. A variety of diagrams were used to demonstrate the characteristics of the probability mass function (pmf) and the cumulative distribution function (cdf). The parameter of the developed model was estimated using the well-known maximum likelihood estimation approach. Finally, two data sets were employed to demonstrate the effectiveness of the investigated distribution. It was shown that the Poisson area-biased Ailamujia distribution provides an appropriate fit for the two count data sets.

Other articles

The 38th Multivariate Statistical Analysis 2019, Lodz. Conference Papers. In the paper ***Triads or tetrads? Comparison of two methods for measuring the similarity in preferences under incomplete block design***, **Artur Zaborski** compares two incomplete methods for measuring the similarity of preferences, i.e. the triad method and the tetrad method. These methods can be used whenever similarities are measured on an ordinal scale. They have been compared in terms of their labour intensity and ability to map the known structure of objects, even when all pairs of objects in subgroups are not equal. The article indicates the possibility of reducing the number of sets presented to respondents in such a way that each pair of objects appears just as often, but less than their potential maximum number. In the example for 9 objects it was shown that scaling based on 8 tetrads gave a good solution. It was also demonstrated that the choice of the incomplete sets has no significant effect on the results of nonmetric multidimensional preference scaling, even when all pairs of objects cannot be presented equally frequently. This conclusion is particularly relevant for the creation of reduced sets when the number of objects does not allow to fulfil the condition of an equal number of pairs. The analysis indicated that the tetrad method can be used if each pair of objects appears in sets at least once, while for the method of triads each pair should appear at least twice.

Research Communicates and Letters

The section presents the paper by **Michael C. Ugwu** and **Mbanefo S. Madukaife** entitled ***Two-stage cluster sampling with unequal probability sampling in the first stage and ranked set sampling in the second stage***. The authors introduce a new sampling design, namely a two-stage cluster sampling, where probability proportional to size with replacement (PPSWR) is used in the first stage unit and ranked set sampling in the second in order to address the issue of marked variability in the sizes of population units concerned with first stage sampling. An unbiased estimator of the population mean and total has been obtained, as well as the variance of the mean estimator. The authors calculated the relative efficiency of the new sampling design to the two-stage cluster sampling with simple random sampling in the first stage and ranked set sampling in the second stage. The results demonstrated that the new sampling design is more efficient as it produces a better estimator for estimating the population mean than a similar design built with simple random sampling in the first stage and ranked set sampling in the second stage units under the condition of significant variation in the sizes of the first stage units.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 1–25, DOI 10.2478/stattrans-2022-0027
Received – 23.02.2022; accepted – 09.06.2022

Assessing the effect of new data sources on the Consumer Price Index: a deterministic approach to uncertainty and sensitivity

Jacek Białek[1], Tomasz Panek[2], Jan Zwierzchowski[3]

ABSTRACT

One of the greatest challenges facing official statistics in the 21st century is the use of alternative sources of data about prices (scanned and scraped data) in the analysis of price dynamics, which also involves selecting the appropriate formula of the price index at the elementary group (5-digit) level. When consumer price indices of goods and services are constructed, a number of subjective decisions are made at different stages, e.g. regarding the choice of data sources and types of indices used for the purpose of estimation. All of these decisions can affect the bias of consumer price indices, i.e. the extent to which they contribute to the overall uncertainty about the resulting index values. By measuring how robust consumer price indices are, one can assess the impact that the decisions made at the different stages of index construction have on the index values. This assessment involves analysing uncertainty and sensitivity. The purpose of the study described in the article was to determine how much and in which direction the consumer price index changes when including scanner and scraped data in the analysis, in addition to the data on prices collected by enumerators. The impact of these new data sources was assessed by analysing uncertainty and sensitivity under the deterministic approach. To the best of the authors' knowledge, it is a novel application of robustness analysis to measure inflation using new data sources. The empirical study was based on data for February and March 2021, while scanner and scraped data about selected categories of food products were obtained from one retail chain operating hundreds of points of sale in Poland and selling products online. It was found that the choice of a data source has the most significant impact on the final value of the index at the elementary group level, while the choice of the aggregation formula used to consolidate different data sources is of secondary importance.

Key words: price indices, scraped data, scanner data, robustness analysis, inflation.

JEL: C43, E31

[1] University of Lodz, Department of Statistical Methods, E-mail: jacek.bialek@uni.lodz.pl; Statistics Poland, Department of Trade and Services, Poland. E-mail: J.Bialek@stat.gov.pl. ORCID: https://orcid.org/0000-0002-0952-5327.

[2] Warsaw School of Economics, Institute of Statistics and Demography, Poland. E-mail:tompa@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-1034-7222.

[3] Warsaw School of Economics, Institute of Statistics and Demography, Poland. E-mail:jzwier@sgh.waw.pl.

1. Introduction

Traditionally, data used to measure inflation are provided by enumerators who collect information about prices and characteristics of selected products in randomly selected shops located in the so-called price survey regions (there are 207 price survey regions in Poland, where enumerators visit about 35,000 shops). Pandemic-related difficulties affecting the traditional method of price data collection, such as the requirement of social distancing and shop closures, and the growing volume of online shopping have provided stronger motivation to intensify work on the use of alternative data sources, such as scanner data and scraped data.

For the purpose of this article, **scanner data** are defined as detailed data about consumer products obtained by scanning bar codes at electronic points of sale (CPI Manual, 2004). The list of barcodes most commonly used by retail chains includes GTIN (Global Trade Item Number) or its European version – EAN (European Article Number), PLU (Price Look-Up) and SKU (Stock Keeping Unit). Product codes, including the code assigned by a given chain store, together with the product label, are used for classifying products using 5-digit ECOICOP codes and below this level of aggregation (Chessa, 2015; 2016, Białek and Beręsewicz, 2021). One of the benefits of using scanner data is that they contain information about the level of consumption even at the lowest level of aggregation.

Scraped data are collected automatically from websites by a special computer programme called a scraper. The programme collects "raw" data, which are then cleaned and formatted to enable further analysis. Scraped data can be collected with greater frequency than data from other sources (usually they are collected daily), which is useful for understanding consumer behaviour patterns and data variability. One should bear in mind, however, that scanner data represent actual prices, while scraped data represent 'merely' offered prices, without providing any information about the level of consumption.

One of the biggest challenges facing official statistics in the 21st century is the use of alternative sources of data about prices (scanner and scraped data) in the analysis of consumer price dynamics, which also involves choosing the appropriate formula of the price index at the elementary group (5-digit) level (Chessa, 2015; 2016, de Haan et. al., 2021). Decisions about whether or not to include a given data source in the measurement of inflation, as well as the choice of a price index formula, can have a measurable impact on the bias of consumer price indices (Saisana, Saltelli and Tarantola, 2005; Sharpe and Salzman, 2004; Nardo et al., 2005 and Nardo et al., 2011), i.e. they contribute to the overall uncertainty about the resulting index values.

When applied to consumer price indices, **robustness analysis** can help to assess the impact of decisions made at different stages of index construction on the value of the index; the process involves uncertainty analysis and sensitivity analysis.

The underlying idea of **uncertainty analysis** is to construct a model linking input variables representing sources of uncertainty (assumptions made at different stages of constructing a consumer price index), determine distribution functions of input variables, and, based on this information, determine the distribution of the consumer price index or output variables used for measuring the impact of changes in the underlying assumptions of the index on its value. Uncertainty analysis itself consists in analysing parameters of the distribution of the consumer price index.

The goal of **sensitivity analysis** is to assess what share of the variance of the consumer price index is due to each of the identified sources of uncertainty (each initial assumption). This is achieved by decomposing the variance into variances explained by particular input variables representing types of assumptions made at different stages of index construction. Sensitivity analysis is therefore closely connected with uncertainty analysis. By combining these two types of analysis, one can measure how robust consumer price indices are when one changes assumptions regarding their construction, in other words, one can analyse the impact of these assumptions on the value of the consumer price index, which in turn affects estimates of consumer price fluctuations. In practice, steps taken during uncertainty and sensitivity analysis of consumer price indices depend on what sources of uncertainty (stages of index construction) and what assumptions concerning these sources (variants of solutions adopted at these stages) are made during a particular analysis.

The purpose of the study described in the article was to determine how much and in which direction the consumer price index changes as a result of including scanner and scraped data, in addition to price data collected by enumerators. The impact of these new data sources was assessed by analysing uncertainty and sensitivity under the so-called **deterministic approach**. To the best of the authors' knowledge, it is a novel application of robustness analysis to measure inflation using new data sources. The empirical study was based on data for February and March 2021, while scanner and scraped data about selected categories of food products were obtained from one retail chain operating hundreds of points of sale in Poland and selling products online. It was found that the choice of a data source has the biggest impact on the final value of the index at the elementary group level, while the choice of the aggregation formula used to consolidate different data sources is of secondary importance.

The article has the following structure: section 2 provides a description of the study sample and the main stages during which data scanned and scraped by Statistics Poland are prepared for the analysis of consumer price dynamics in cooperation with one retail chain. Section 3 presents a description of the research method, in particular uncertainty and sensitivity analysis, and includes a list of index formulas used in the study. Section 4 is devoted to the analysis of the results, while section 5 contains a summary of the study and conclusions.

2. Study sample

For the last three years, a consortium consisting of Statistics Poland, the Institute of Computer Sciences of the Polish Academy of Sciences and Warsaw School of Economics has been running a project called *InstatCeny*, with the goal of exploiting alternative sources of data to calculate Consumer Price Index (CPI). This article presents preliminary results of a study based on data about food products that have been obtained for the project. It may be treated as some continuation of the previously started research on these group of products (Białek et al., 2021). Given the reference period of available data (Statistics Poland has been scraping price data about food products only since the start of 2021), price indices were calculated using three sources of data about prices observed in February and March 2021. Seven elementary groups of food products were selected: rice, raw and whole milk, fresh low fat milk, yoghurt, beverages and other milk products, sugar and coffee. Data collected by enumerators in 207 price survey regions included information about each representative of the selected elementary groups, in each case, the actual price was recalculated to a fixed unit of measurement (e.g. price per litre for milk, price per kilogram for rice, etc.). In the case of scanner and scraped prices, the list of product representatives used to create subcategories of elementary groups were extended to include 3 new items: yoghurt flavoured with chocolate and fruit, powdered sugar, ground coffee. As these subcategories were sufficiently well represented and homogenous, they were included in the estimation of price changes in the corresponding ECOICOP elementary groups, despite certain discrepancies with respect to the list of representatives in the classification.

2.1. Acquisition of scanner data

The manner in which scanner data are acquired differs depending on the retail chain that supplies them. Statistics Poland uses secure (encrypted) transfers and, in the case of one retail chain – direct downloads by means of an application programming interface (API). What types of variables are provided also depends on the supplier; as regards scanner data used in the study, in addition to codes enabling product classification, the transferred *csv* data frame contained information about product ID, unit of sale, transaction date, selling price, total, a flag relating to discounts, sales and promotions and the amount of VAT. The seven elementary groups of products provided by the retail chain and used in the study amounted to 32MB of scanner data per month.

2.2. Acquisition of scraped data

Data for the *InstatCeny* project are scraped using Python scripts that rely on the Selenium package (Białek et al., 2021). Scrapers developed by Statistics Poland have been running since the start of 2021 and scraped data are saved and archived in the form of JSON files. The range of variables included in scraped data is very similar to that found in files transferred directly by the retail chain. It turned out that products shown on the retail chain's website represent between 40 and 90% of products of the same category that can be found in the chain's stores. For example, at the start of 2021, 33 products were classified as rice in scanner data, compared to 27 in scraped data. In the case of coffee, the ratio was 275 to 152. The reason why not all products found on shelves are included in the online offering is that websites probably only feature the most popular products. The elementary groups of products provided by the retail chain and used in the study amounted to 4MB of scraped data per month.

2.3. Preparation of data from alternative sources for processing

After scanner and scraped data had been cleaned (i.e. standardising names, removing incorrect data and unusual prices), all products were classified into appropriate elementary groups and the 6-digit codes. Products found in both alternative sources were classified into categories on the basis of product labels and previously created dictionaries of key words and phrases. The process was performed using the *data_selecting* and *data_matching* functions from the *PriceIndices* R package (Białek, 2021). Text labels were compared using the Jaro-Winkler distance (Jaro, 1989; Winkler, 1990), with the threshold distance (above which two labels were regarded as different) set to 0.02. Next, the sample of scanned products was filtered in order to remove products with extreme price fluctuations (3% of cases) as well as those with relatively low levels of sales (up to 25% of products, depending on the category). In the case of scraped data, only extreme price changes were filtered out, with cut-off values equal to 0.25 and 3 for the ratio of March to February prices, which had a negligible effect on the sample size (only two products from the yoghurt category were removed). It should be noted that the "monthly price" is determined differently for each data source. In the case of scanner data, it is defined as the ratio of total sales of a product in value to its total sales in quantity, which is known as unit sales. In the case of data scraped each month (regardless of whether or not the product was sold), the monthly price is calculated as the average of all observations scraped in a given month.

Figure A.1 in the appendix shows an illustrative comparison of monthly prices for March 2021 from the three data sources regarding the four most numerously represented product categories (coffee beans, natural yoghurt, long-grain rice, wheat flour). In general, scraped prices are characterised by the highest level of variability, while prices collected by enumerators are the least variable. However, this pattern does not hold for all categories of products: e.g. prices of long-grain rice from scanner data show the biggest fluctuations. The relatively smallest number of price outliers were observed in scanner data, which is not surprising given that these data were subjected to three kinds of filters, as described above. The biggest amount of noise was found in scraped data, despite the application of the price outlier filter at the level of GTIN code. As regards price outliers in data collected by enumerators, three such cases were recorded with respect to long-grain rice, which is a kind of exception. There is no doubt, however, that differences in average prices obtained from the three sources can be considerable, with scraped prices, on average, tending to be higher than those from the other two sources (again with the exception of long-grain rice).

3. Method description

The analysis of consumer price indices involved three sources of uncertainty, representing three kinds of decisions made during index construction: what data source with consumer prices is used, what formula is used for aggregating indices of 6-digit subgroups into 5-digit elementary groups within each data source, what formula is used for aggregating indices of price changes within elementary groups based on different data sources into total indices for each of the elementary groups. The purpose of the analysis was to assess the robustness of consumer price indices calculated for 7 elementary groups of products within the food division of the COICOP classification.

Consumer price indices for the selected elementary groups were estimated using six different sources of information about consumer prices of products classified into these elementary groups in February and March 2021:

- consumer prices surveyed by enumerators,
- data from the IT system of the retail chain (scanner data),
- online prices of the retail chain (scraped data),
- consumer prices surveyed by enumerators combined with retailer's online prices,
- consumer prices surveyed by enumerators combined with retailer's scanner data,
- all three data sources combined.

Consumer price indices for the selected elementary groups are calculated by aggregating price indices of 6-digit subgroups within each elementary groups. Data collected by enumerators and those from online price listings do not contain information about quantities of products purchased within 6-digit categories of each subgroup (or about the share of each subgroup in the total value of products sold within each elementary group). As a result, consumer price indices for 6-digit subgroups were aggregated into elementary groups' indices (5-digit level) using unweighted geometric mean, known as the Jevons index (CPI Manual, 2004):

$$_{J}I_{G} = \sqrt[n]{\prod_{i=1}^{n} I_{G,g_i}}, \qquad G=1,2,\dots,7, \tag{1}$$

where:

I_{G,g_i} , - price indices for the *i*-th subgroup of the subgroup *G*-th elementary group,

$_{J}I_{G}$ - the Jevons index for *G*-th elementary group.

In contrast, scanner data contain information not only about prices but also about amounts of products purchased within particular subgroups of each elementary group. These amounts were also used as weights in the process of aggregating subgroups of elementary groups for the other data sources. In this way, in addition to using the unweighted Jevons formula, price indices for subgroups could also be aggregated by employing weighted indices, namely the Laspeyres, Paasche, Fisher and Törnqvist indices (CPI Manual, 2004):

$$_{L}I_{G} = \sum_{i=1}^{n} w_{G,g_i}^{0} I_{G,g_i}, \qquad G=1,2,\dots,7, \tag{2}$$

$$_{P}I_{G} = \frac{1}{\sum_{i=1}^{n} w_{G,g_i}^{1} \frac{1}{I_{G,g_i}}}, \qquad G=1,2,\dots,7, \tag{3}$$

$$_{F}I_{G} = \sqrt{_{L}I_{G} \cdot {_{P}I_{G}}}, \qquad G=1,2,\dots,7, \tag{4}$$

$$_{T}I_{G} = \prod_{i=1}^{n} (I_{G,g_i})^{\frac{w_{G,g_i}^{0}+w_{G,g_i}^{1}}{2}}, \qquad G=1,2,\dots,7, \tag{5}$$

where:

$_{L}I_{G}$, $_{P}I_{G}$, $_{F}I_{G}$, $_{T}I_{G}$ - price indices proposed by Laspeyres (1871), Paasche (1874), Fisher (1922) and Törnqvist (1936), respectively, for the *G*-th elementary group,

$w_{G,g_i}^{0}, w_{G,g_i}^{1}$ – weight of the *i*-th subgroup of the *G*-th elementary group in the base period (February 2021) and in the reference period (March 2021).

The weight for a given subgroup in the base period is calculated by dividing the total value of products in this subgroup sold in the base period by the total value of sold products in all subgroups of a given elementary group. The same method is used to calculate weights for subgroups in the reference period (weight values are presented in Table A.1 in the Appendix).

When price indices are calculated using more than one source of data, the price index for each elementary group is calculated by aggregating price indices for these elementary groups that were calculated separately on the basis of each data source. This aggregation was performed using the Young index or the geometric Young index (Białek, 2017):

$$ {}_{Y}I_G \;= (I_G^N)^{w_N^\tau} + (I_G^W)^{w_W^\tau} + (I_G^S)^{w_S^\tau}, \qquad G=1,2,\ldots,7, \tag{6}$$

$$ {}_{YG}I_G \;= (I_G^N)^{w_N^\tau}(I_G^W)^{w_W^\tau}(I_G^S)^{w_S^\tau}, \qquad G=1,2,\ldots,7, \tag{7}$$

where:

w_N^τ , w_W^τ , w_S^τ – index weight for the *G*-th elementary group, established on the basis of a period more distant than the base period (in our case, it was 2020), calculated using prices surveyed by enumerators, scraped prices and scanner data, respectively.

${}_{Y}I_G$, ${}_{YG}I_G$ - the Young index or the geometric Young index for the *G*-th elementary group.

In the case of price indices calculated on the basis of more than one data source, shares of purchases within particular elementary group from a given source in total sales from all sources combined were used as weights. Shares of products purchased online were estimated on the basis of information obtained from the Household Budget Survey conducted by Statistics Poland. Shares of purchases in chain stores were obtained from databases maintained by Passport GMID, Euromonitor International as well as domestic market surveys conducted by Statistics Poland (Table A.2 in the Appendix).

Price indices for the analysed subgroups are shown in Table A.3 in the Appendix. In the analysis, the consumer price index for a given elementary group is a composite index, based on consumer price indices for its subgroups using the different data sources. When price data come from more than one source, the composite index is calculated in two steps. The first step consists in calculating price indices for each elementary group using data from each source. In the second step, price indices for a given elementary group calculated in the first step are aggregated into a single price index for this particular elementary group, which takes into account information from all data sources.

From a formal perspective, the dependence of the consumer price index for a given elementary group on input variables representing the assumptions made during index construction can be expressed using the following model:

$$I_G = f_{rs}\big(\boldsymbol{I}_{G,g}^Z, \boldsymbol{w}_{G,g}^S, \boldsymbol{w}_G^Z\big), \qquad G=1,2,\ldots,7, \tag{8}$$

where:

$\boldsymbol{I}_G$ – value of the consumer price index for the *G*-th elementary group,

$\boldsymbol{I}_{G,g}^Z$ – a vector of price indices for subgroups of the *G*-th elementary group calculated on the basis of different sources of price data (*Z*=*N*,*W*,*S*),

$\boldsymbol{w}_{G,g}^{S}$ – a vector of weights of price indices for subgroups of the G-th elementary group calculated on the basis of scanner data (Laspeyres, Paasche, Fisher and Törnqvist indices),

$\boldsymbol{w}_{G}^{Z}$ - a vector of weights of price indices for subgroups of the G-th elementary group calculated on the basis of different sources of price data (Z=N,W,S),

f_{rs} – a function transforming consumer price indices for subgroups calculated on the basis of sources of price data and by applying weights assigned to price indices for subgroups and elementary groups obtained from the s-th source of price information (s-th combination of sources of price data when data come from different sources) and the r-th index formula (the r-th combination of index formulas when data come from more than one source).

The purpose of uncertainty and sensitivity analysis was to examine changes in the values of the composite index of consumer prices identified at the level of elementary groups, which result from changes of assumptions made during the estimation of the index.

3.1. The main idea of the analysis of uncertainty

Uncertainty analysis aims at quantifying the variability of the output that is due to the variability of the input. One can distinguish two main groups of methods for analysing uncertainty, namely probabilistic and deterministic methods. Probabilistic methods involve simulations based on different assumptions regarding the construction of a composite index (in our study, the consumer price index for a given elementary group), which are treated as inputs of uncertainty. The simulation model should reflect the probabilistic nature of the phenomenon of interest (Saisana, Saltelli and Tarantola, 2005; OECD, 2008, Panek 2016).

The most commonly used probabilistic method of uncertainty analysis is the Monte Carlo method (Saisana, Saltelli and Tarantola, 2005; OECD, 2008; Panek, 2016). The Monte Carlo approach involves a multivariate assessment of the proposed model with quasi-randomly selected parameters (input variables describing assumptions made during the construction of the composite index). The procedure consists of a few steps. In the first step, we determine the probability distribution function of each input variable (each assumption) X_k, k=1,2,…,z. In our study, the first input variable (X_1) represents the choice of a source of data about consumer prices, the second input variable (X_2) represents index formulas used for aggregating subgroup indices within each data source into elementary group indices, and the third input variable (X_3) refers to index formulas used for aggregating elementary group indices calculated on the basis of different data sources into composite indices for each elementary group. All these assumptions are random variables with discrete distributions. We then randomly generate N combinations of independent input variables $\boldsymbol{X}^l$, l=1,2,...,N. The set $X^l = X_1^l, X_2^l, \dots, X_k^l$ of combinations of input variables constitutes a sample. Such a sample can be generated using different sampling methods, such as simple random sampling, stratified sampling, or quasi-random sampling (Saltelli, Chan and Scott, 2000). For each sample (combination of assumptions), the model is assessed by calculating values of input variables (in our study, values of the consumer price index). The sequence $\boldsymbol{Y}^l$, can be used to estimate empirical probability distribution functions of particular input variables and their In deterministic methods we do not use the simulation model. Instead of the simulation model, the value of a composite index (in our study, the consumer price index for a given elementary group) for all possible N combinations of independent assumptions $\boldsymbol{X}^l$. for a given elementary group, is calculated (see Table A.4 in the Appendix). Values of this index determine its distribution, which is used to assess how robust the consumer price index for a given elementary group is to changes of its underlying assumptions regarding the choice of data sources and index formulas used in the process of aggregation.

3.2. The main idea of the analysis of sensitivity

The purpose of the analysis of sensitivity is to determine how particular assumptions (input variables) underlying a given price index values.

When a few sources of uncertainty are simultaneously taken into account when modelling a composite index, a nonlinear model can be used. In the analysis of sensitivity, good results have been achieved by applying methods based on the analysis of variance (Chan et al., 2000; Saltelli et al., 2008; Panek, 2016). As in the case of uncertainty analysis, depending on data characteristics, sensitivity analysis can be performed by applying probabilistic or deterministic methods. In the probabilistic approach, sensitivity indices are usually estimated by means of Sobol's method (1993), modified by A. Saltelli (2002). Sobol's method, as already mentioned, uses quasi-random sampling to determine distributions of input variables. Sensitivity of a composite index (the consumer price index for a given elementary group) to different parameters (input variables, i.e. assumptions made during its construction) is assessed on the basis of sensitivity indices, which are calculated after decomposing the total output variance $D^2(Y)$ of the output variable Y (in our study, the consumer price index):

$$D^2(Y) = \sum_{k=1}^{z} V_k + \sum_{k=1}^{z} \sum_{\substack{k\prime=1 \\ k<k\prime}}^{z} V_{k,k\prime} + \dots + V_{1\dots z}, \tag{9}$$

where:

$$V_k = D_{X_k}^2\left[E_{X_{-k}}(Y|X_k)\right], \tag{10}$$

$$V_{kk\prime} = D_{X_k X_{k\prime}}^2\left[E_{X_{-kk\prime}}(Y|X_k, X_{k\prime})\right] - D_{X_k}^2\left[E_{X_{-k}}(Y|X_k)\right] - D_{X_{k\prime}}^2\left[E_{X_{-k\prime}}(Y|X_{k\prime})\right] \tag{11}$$

The first term of equation (11) provides assessment of the direct impact of the input variable *X_k*, (in our example the input variable represents the choice of a particular method at a given stage of constructing a composite index) on the total output variance of the *Y* output variable. The second term of equation (11) represents the impact of the interaction between the k-th and k'-th input variable on the total variance of the *Y* output variable (the indirect impact of variable X_k on the total output variance of the *Y* output variable).

The direct impact of input variables on the values of a composite index (assuming there are no interactions between input variables) is measured by first-order sensitivity indices:

$$S_k = \frac{V_k}{V} = \frac{D^2_{X_k}\left[E_{X_{-k}}(Y|X_k)\right]}{D^2(Y)}, \qquad k=1,2,\ldots,z. \tag{12}$$

The model without interactions between input variables is known as an additive model. In this case, the sum of all first-order sensitivity indices is equal to 1 ($\sum_{k=1}^{z} S_k = 1$).

In the case of a non-additive model, one needs to estimate higher-order sensitivity indices, which measure the interaction effects among the set of input variables. However, in practice, they are rarely calculated, since given a model with k input variables, the number of sensitivity indices that would have to be estimated equals 2^k-1. For this reason, the impact of interactions between input variables on the variance of the output variable is calculated indirectly. In the first step, we calculate total sensitivity indices which measure the total impact of the input variable X_k on the on the total output variance of the *Y* output variable, i.e. the direct impact as well as that resulting from interactions between all possible combinations of other input variables:

$$S_k^T = \frac{D^2(Y)-D^2_{X_{-k}}\left[E_{X_k}(Y|X_{-k})\right]\left[E_{X_k}(Y|X_{-k})\right]}{D^2(Y)} = \frac{E_{X_{-k}}\left(D^2_{X_k}(Y|X_{-k})\right)}{D^2(Y)}. \tag{13}$$

The analysis of sensitivity takes into account three types of assumptions, which enable us to calculate three total sensitivity indices:

$$S_1^T = \frac{D^2(Y)-D^2_{X_{-k}}\left[E_{X_k}(Y|X_{-k})\right]\left[E_{X_k}(Y|X_{-k})\right]}{D^2(Y)} = S_1 + S_{12} + S_{13} + S_{123}\ , \tag{14}$$

$$S_2^T = S_2 + S_{23} + S_{23} + S_{123}\ , \tag{15}$$

$$S_3^T = S_3 + S_{13} + S_{23} + S_{123}\ , \tag{16}$$

Ultimately, the impact of interactions between input variables (their indirect impact) on the variance of the output variable is calculated as a difference between their indirect and indirect impacts ($S_k^T - S_k$). Considerable differences between S_k^T and S_k indicate the significant role of interactions between the k-th input variable in shaping the value of the Y output variable, which in turn indicates a high degree of correlation between input variables. The analysis of interactions between input variables helps to understand the model structure. Variance-based methods of sensitivity analysis, both for independent and dependent input variables were described by Saltelli et al. (2008).

Under the deterministic approach we do not use quasi-random sampling to determine distributions of input variables. Instead of applying quasi-random sampling, sensitivity indices are calculated using values of price indices for particular elementary groups, estimated for each combination of sources of price data and index formulas.

4. Results

4.1. Uncertainty analysis

We opted for the deterministic approach because the number of assumptions used in the study was relatively small and it was possible to analyse all possible combinations.

Values of price indices for particular elementary groups, which were calculated for each combination of each combination of sources of price data and index formulas are presented in Table A.4 in the Appendix. For each elementary group, Figure 1 shows:

- the value of the index estimated by means of the method currently used by Statistics Poland (price data are surveyed by enumerators, consumer price indices for subgroups are aggregated using the Jevons index),
- the value of the index estimated on the basis of combined data sources and preferred aggregation formulas (consumer price indices for subgroups were aggregated into the price index for their respective elementary groups using the Törnqvist index, consumer price indices for a given elementary group, calculated on the basis of different data sources were aggregated into a single index using the geometric Young index),
- the minimum and maximum values of each index.

It should be noted that, owing to the limited availability of data, the robustness analysis of consumer price indices includes changes that were observed over one month (between February and March 2021). However, in view of the short period covered by the analysis, obtained results cannot constitute a basis for drawing conclusion of a general nature. For that purpose it is necessary to base the analysis on a longer period.

Given high levels of inflation in Poland both in 2000 and in 2021, differences in annual estimates of inflation based on price indices calculated from different sources of data are likely to be much higher than those calculated on the basis of monthly data (assuming that the trend of changes observed between February and March 2021 was to continue throughout the year, annual price changes would be 12 times bigger).

The biggest range of consumer price indices (difference between the maximum and minimum estimates of price indices), resulting from various combinations of different data sources and different formulas was obtained for the sugar elementary group – 5.1 percentage points.

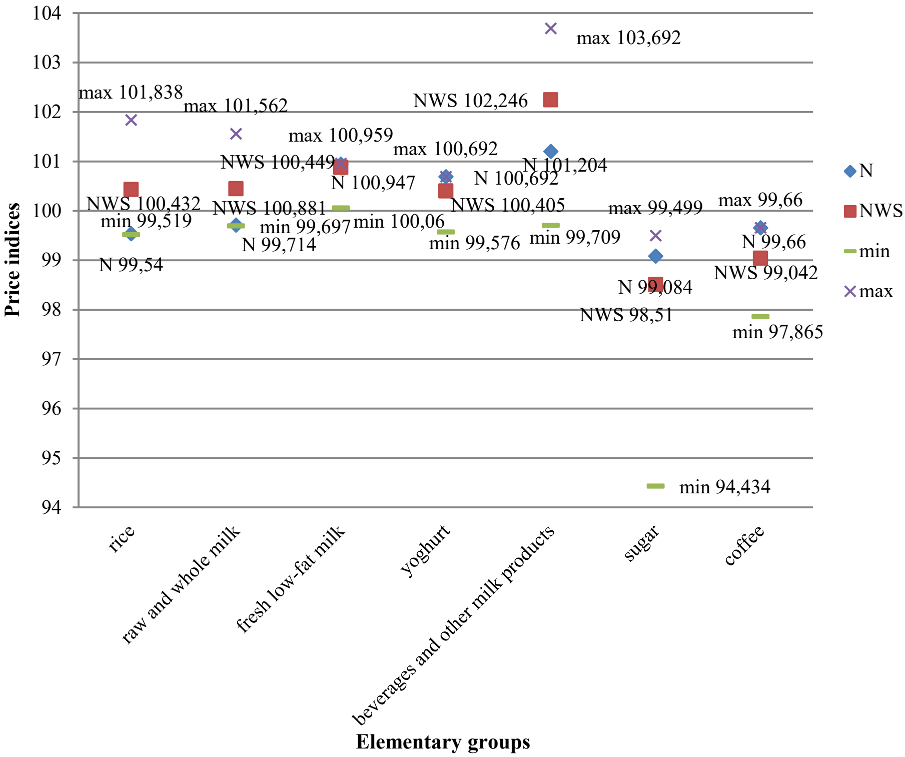


Figure 1. Results of uncertainty analysis for consumer price indices

Source: authors' work based on data in Table A.4.

The smallest range of consumer price indices can be observed for the elementary group of fresh low fat milk – 0.9 percentage point.

4.2. Sensitivity analysis

In the study described above, the analysis of sensitivity was conducted using the deterministic approach in which sensitivity indices were calculated on the basis of values of consumer price indices for particular elementary groups, estimated for each combination of data sources and index formulas used. The analysis involved seven variants of data sources, five index formulas for aggregating subgroups into indices for elementary groups and two formulas for aggregating indices calculated from different data sources. This resulted in a total of 70 combinations of assumptions.

In the deterministic model, total model variance V is estimated as variance from all calculated index values for all combinations of assumptions. In order to determine the variance component directly associated with a given assumption V_k (formula 10), it is necessary to calculate the mean index value for each possible value of this assumption. The number of mean values depends on the number of assumption variants. The variance of these means constitutes the estimator of V_k.

In order to calculate $E_{X_{-k}}\left(D^2_{X_k}(Y|X_{-k})\right)$ (formula 13) for the k-th assumption, one needs to consider all possible combinations of the remaining assumptions (denoted as k). For each such combination, one calculates the variance of values obtained for the final index. The construction of indices obtained from a given combination will differ only with respect to assumption k. The mean value calculated from all these variances constitutes the estimator of $E_{X_{-k}}\left(D^2_{X_k}(Y|X_{-k})\right)$.

Results of these calculations are presented in Table 1. The choice of a particular data source accounted directly for as much as 85% or more of total variability for all data sources, except for the yoghurt elementary group. The impact of the index formula used for aggregating indices for subgroups into elementary group indices accounted for only 0.09% of variability for fresh whole milk, 3% for beverages and other milk products, 4% for fresh low fat milk, 5.4% for coffee, 6.9% for rice, 9.3% for sugar and 16.5% for yoghurt. For all elementary groups, the impact of the index formula used for aggregating indices derived from various data sources into one composite index was negligible – less than 0.0001% of total variance.

For all elementary groups and all assumptions considered in the study, total impacts of interactions (13) are higher than their direct impacts (10). The bigger the differences between these two values, the more interactions between the assumptions contribute to the total variability of the final indices. The yoghurt elementary group is particularly noteworthy in this respect since the total impacts for the first two assumptions exceed 90%, while their direct impacts are equal to 21.9% and 16.5%, respectively, which implies that in the case of this elementary group these two assumptions strongly interact with each other. In other words, when one of them is replaced, values of the final index are not greatly affected. However, certain combinations of these assumptions can result in relatively disparate values of the index.

Table 1. Sensitivity indices of consumer price indices

Product elementary groups	**Values of sensitivity indices in %**								
	price data source			**index formulas – subgroups**			**index formulas – elementary groups**		
	S_1	S_1^T	S_1^T - S_1	S_2	S_2^T	S_2^T - S_2	S_3	S_3^T	S_3^T - S_3
Rice	88.22	107.09	18.87	6.88	14.51	7.64	0.00004	0.00035	0.00031
Raw and whole milk	99.82	114.89	15.07	0.09	0.22	0.13	0.00001	0.00008	0.00007
Fresh low fat milk	93.44	110.34	16.89	4.06	8.08	4.02	0.00000	0.00000	0.00000
Yoghurt	21.89	96.01	74.12	16.52	96.24	79.73	0.00000	0.00007	0.00007
Beverages and other milk products	96.69	111.51	14.82	3.04	4.08	1.04	0.00006	0.00036	0.00031
Sugar	85.45	104.28	18.83	9.32	17.93	8.61	0.00000	0.00001	0.00000
Coffee	84.42	108.75	24.33	5.44	19.20	13.77	0.00001	0.00006	0.00005

Source: authors' calculations based on data obtained from Statistics Poland and from a retail chain.

5. Conclusions

In the case of four elementary groups (fresh low fat milk, yoghurt, sugar and coffee) inflation indicators obtained by applying the method currently used by Statistics Poland are higher than measures produced by applying the preferred method and lower in the case of the remaining three elementary groups (rice, fresh and whole milk and beverages and other milk products). In comparison with the preferred method, the official method underestimates the rate of inflation to the most (1 p.p.) in the elementary group of beverages and other milk products. The maximum overestimation of the rate of inflation (0.6 p.p.) can be observed in the coffee elementary group.

The results indicate that consumer price indices calculated for the elementary groups of interest are characterised by a relatively low robustness to changes of a data source about consumer prices and by a relatively high robustness to changes of index formulas used for calculating price indices at the level of subgroups and elementary groups.

For all elementary groups of interest, the first assumption, i.e. the choice of a given data source, has the biggest impact on the final value of the price index. Which index formula is used for aggregating indices for subgroups into elementary group indices has much less influence on the final results. The effect of choosing a particular index for aggregating indices derived from different sources is negligible.

The authors are fully aware that the results are only a preliminary assessment of the impact of new data sources on the measurement of changes in consumer prices. These conclusions cannot be generalised because, for one thing, they are based on data from only two months and, secondly, they only refer to selected elementary groups of food products. Nonetheless, the study is an important starting point for further, more comprehensive studies into the robustness of price indices in the measurement of inflation.

The relatively low stability of price indices as a consequence of adopting different assumptions underlying their construction means that one needs to be very cautious when attempting to include new data sources in the measurement process. The danger associated with such attempts is that changes manifested in price indices, rather than reflect actual price changes, may well be merely the result of including new data sources.

Acknowledgement

The article is the outcome of the project entitled “The construction of an integrated system of price statistics” (INSTATCENY), funded by the National Centre for Research and Development (NCBiR) (1st edition of the GOSPOSTRATEG competition, No. 1/382525/14 / NCBR / 2018). Authors would like to thank the National Centre for Research and Development for financing this publication.

References

Białek J., (2017). Approximation of the Fisher price index by using Lowe, Young and AG Mean indices. *Communications in Statistics – Simulation and Computation*, 46(8), pp. 6454–6467.

Białek, J., Dominiczak-Astin, A. and Turek, D., (2021). Porównanie cen i wskaźników cen konsumpcyjnych: tradycyjna metoda uzyskiwania danych a źródła alternatywne. *Wiadomości Statystyczne. The Polish Statistician*, 66(9), pp. 32–69.

Białek, J., (2021). PriceIndices – a New R Package for Bilateral and Multilateral Price Index Calculations. *Statistika – Statistics and Economy Journal*, Vol. 2/2021, pp. 122–141, Czech Statistical Office, Praga.

Białek, J., Beręsewicz, M., (2021). Scanner data in inflation measurement: From raw data to price indices. *Statistical Journal of the IAOS*, Vol. 37, pp. 1315–1336.

Chan, K., Tarantola, S., Saltelli, A. and Sobol, I. M., (2000). Variance based methods, in: A. Saltelli, K. Chan i M. Scott (eds.). *Sensitivity analysis*, Wiley, New York, pp. 167–197.

Chessa, A., (2015). Towards a generic price index method for scanner data in the Dutch CPI. *In 14th meeting of the Ottawa Group*, Tokyo, pp. 20–22.

Chessa, A., (2016). A new methodology for processing scanner data in the Dutch CPI. *Eurostat review of National Accounts and Macroeconomic Indicators*, Vol. 1, pp. 49–69.

International Labour Office, (2004). *Consumer price index manual: Theory and practice*, Geneva.

De Haan, J., Hendriks, R. and Scholz, M., (2021). *Price measurement using scanner data: Time-product dummy versus time dummy hedonic indexes*. Review of Income and Wealth 67(2), pp. 394–417.

Fisher, I., (1922). *The making of index numbers: a study of their varieties, tests, and reliability*, Number 1, Houghton Mifflin.

Jaro, M., (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, Vol. 84, pp. 414–420.

Laspeyres, K., (1871). IX. Die berechnung einer mittleren waarenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik,* 16(1), pp. 296–318.

Nardo, M., Saisana, M., Saltelli, A. and Tarantola, S., (2011). Tools for composite indicators building. *Paperback – European Commission*, Dictus Publishing.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A. and Giovannini, E., (2005). Handbook on constructing composite indicators: Methodology and user guide. OECD, *Statistics Working Paper*.

OECD, (2008). Handbook on constructing composite indicators. Methodology and user guide. *OECD Publications*, Paris.

Panek, T., (2016). *Quality of Life – from conception to measurement*. Warsaw School of Economic Press, Warsaw.

Paasche, H., (1874). Über die preisentwicklung der letzten jahre nach den hamburger börsennotirungen. *Jahrbücher für Nationalökonomie und Statistik*, pp. 168–178

Saisana, M., Saltelli A. and Tarantola S., (2005). Uncertainty and sensitivity techniques as tools for the analysis and validation of composite indicators. *Journal of the Royal Statistical Society*, A., Vol. 168(2), pp. 307–323.

Saltelli, A., (2002). Making best use of model valuations to compute sensitivity indices. *Computer Physics Communications*, Vol. 145, pp. 280–297.

Saltelli A., Chan K. and Scott, E. M. (red.), (2000). *Sensitivity analysis*. John Wiley & Sons, New York.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S., (2008). *Global sensitivity analysis*. The primer. John Wiley & Sons, Chichester.

Sharpe, A., Salzman, J. (2004). *Methodological choices encountered in the construction of composite indices of economic and social well-being. Center for the Study of Living Standards*, Ottawa, CAN.

Sobol, I. M., (1993). Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling and Computational Experiment*, Vol. 1, pp. 407–414.

Törnqvist, L., (1936). The Bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin*, pp. 1–8.

Winkler, W., (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 354–359.

Appendix

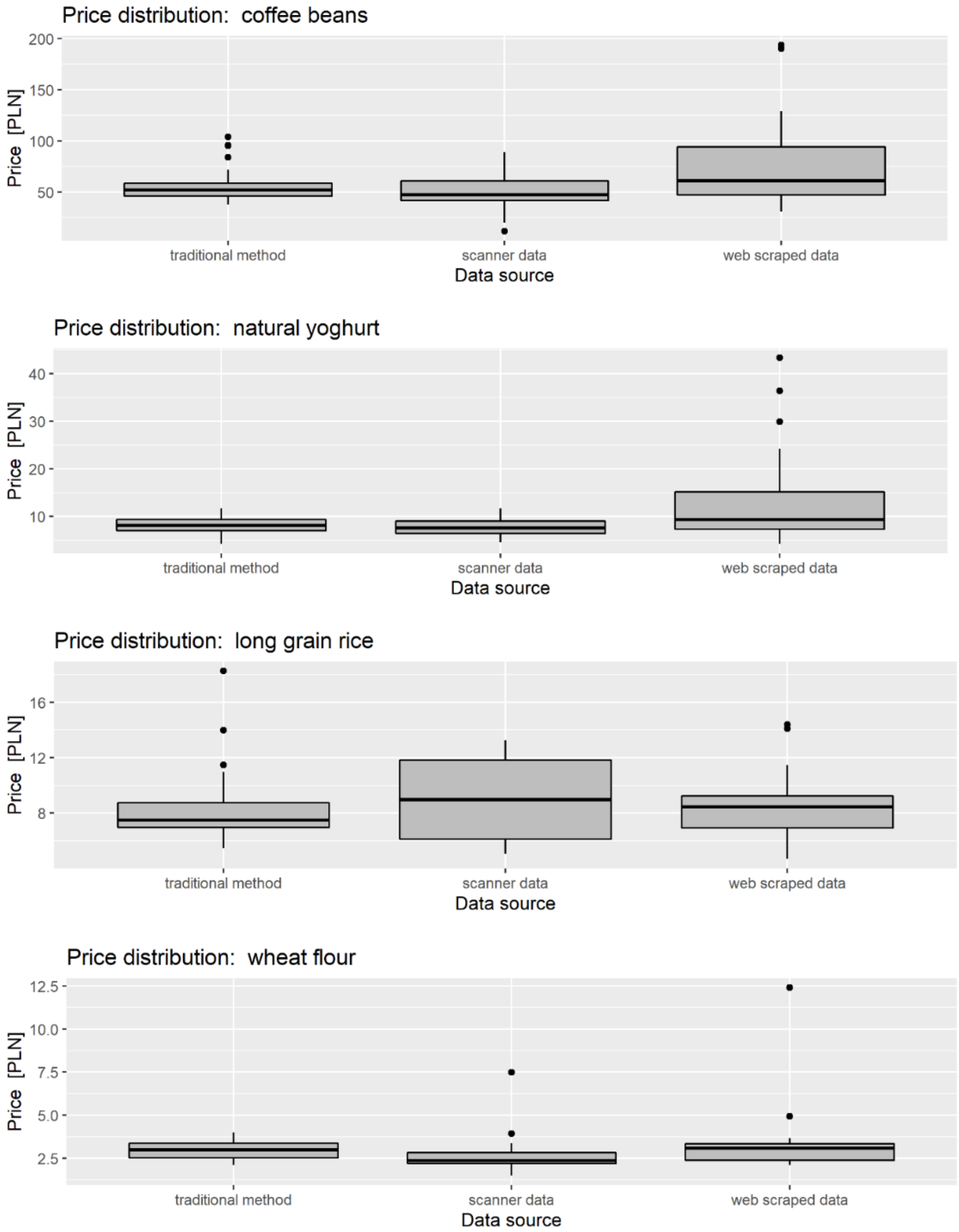


Figure A.1. Box plots for prices of selected categories of food products (based on 3 data sources for March 2021).

Source: authors' calculations based on data obtained from Statistics Poland and from a retail chain.

Table A.1. Weights of subgroups (within their respective elementary groups)

Elementary groups and subgroups	Weights	
	February 2021	March 2021
RICE		
long grain rice	0.654	0.678
white rice	0.346	0.322
RAW AND WHOLE MILK		
UHT whole milk	0.474	0.499
pasteurised whole milk	0.526	0.501
FRESH LOW FAT MILK		
UHT low fat milk	0.420	0.421
goat milk	0.034	0.036
pasteurised low fat milk	0.546	0.544
YOGHURT		
Actimel	0.100	0.097
fruit flavoured yoghurt	0.317	0.307
chocolate and nuts flavoured yoghurt	0.006	0.006
drinkable yogurt	0.282	0.294
natural yoghurt	0.294	0.297
BEVERAGES AND OTHER MILK PRODUCTS		
kefir	0.211	0.224
buttermilk	0.096	0.101
monte	0.229	0.201
homogenised cheese	0.464	0.473
SUGAR		
cane sugar	0.136	0.126
white sugar	0.790	0.778
powdered sugar	0.075	0.096
COFFEE		
instant coffee	0.083	0.073
coffee beans	0.546	0.512
ground coffee	0.371	0.415

Source: authors' calculations based on data obtained from a retail chain.

Table A.2. Elementary group weights depending on the place / mode of purchase

Elementary groups	**Weights**		
	price survey data (shops excluding chain stores)	**scanner data (chain stores)**	**web scraped data**
Rice	34.65	64.36	0.99
Fresh whole milk	34.86	64.73	0.41
Fresh low fat milk	34.85	64.72	0.43
Yoghurt	34.85	64.72	0.43
Beverages and other milk products	34.84	64.71	0.45
Sugar	39.82	59.72	0.46
Coffee	41.28	57.00	1.72

Source: authors' calculations based on data obtained from databases maintained by Passport GMID, *Euromonitor International and domestic market surveys conducted by Statistics Poland.*

Table A.3. Price indices of consumer products by price data source and index formula, February 2021-March 2021

Data source and combinations of index formulas	**Indices (February 2021=100)**						
	rice	**fresh whole milk**	**fresh low fat milk**	**yoghurt**	**beverages and other milk products**	**sugar**	**coffee**
All data sources							
Jevons x Young	101.023	100.425	100.692	99.966	102.809	99.323	99.473
Jevons x Geom. Young	101.017	100.424	100.692	99.964	102.802	99.323	99.472
Laspeyres x Young	100.498	100.484	100.882	100.430	102.391	98.511	99.048
Laspeyres x Geom. Young	100.495	100.483	100.882	100.430	102.383	98.511	99.046
Paasche x Young	100.372	100.418	100.880	100.381	102.121	98.510	99.039
Paasche x Geom. Young	100.370	100.416	100.880	100.381	102.116	98.509	99.038
Fisher x Young	100.435	100.451	100.881	100.406	102.256	98.511	99.043
Fisher x Geom. Young	100.433	100.449	100.881	100.405	102.250	98.510	99.042
Tornqvist x Young	100.434	100.451	100.881	100.405	102.252	98.510	99.043
Tornqvist x Geom. Young	100.432	100.449	100.881	100.405	102.246	98.510	99.042

Table A.3. Price indices of consumer products by price data source and index formula, February 2021-March 2021 (cont.)

Data source and combinations of index formulas	**Indices (February 2021=100)**						
	rice	**fresh whole milk**	**fresh low fat milk**	**yoghurt**	**beverages and other milk products**	**sugar**	**coffee**
Price survey data and scanner data							
Jevons x Young	101.034	100.421	100.695	99.966	102.821	99.333	99.501
Jevons x Geom. Young	101.028	100.419	100.695	99.965	102.814	99.333	99.501
Laspeyres x Young	100.504	100.480	100.885	100.432	102.403	98.530	99.061
Laspeyres x Geom. Young	100.502	100.478	100.885	100.432	102.396	98.530	99.060
Paasche x Young	100.378	100.413	100.883	100.383	102.132	98.528	99.052
Paasche x Geom. Young	100.376	100.412	100.883	100.383	102.127	98.528	99.051
Fisher x Young	100.441	100.446	100.884	100.407	102.267	98.529	99.056
Fisher x Geom. Young	100.439	100.445	100.884	100.407	102.261	98.529	99.055
Tornqvist x Young	100.441	100.446	100.884	100.407	102.264	98.529	99.056
Tornqvist x Geom. Young	100.438	100.445	100.884	100.407	102.258	98.529	99.055
Price survey data and web scraped data							
Jevons x Young	99.550	99.735	100.936	100.681	101.190	99.064	99.588
Jevons x geom. Young	99.550	99.735	100.936	100.681	101.190	99.064	99.588
Laspeyres x Young	99.530	99.718	100.948	100.200	100.754	98.715	99.528
Laspeyres x geom. Young	99.530	99.718	100.948	100.200	100.754	98.714	99.527
Paasche x Young	99.527	99.734	100.948	100.171	100.750	98.706	99.528
Paasche x geom. Young	99.527	99.734	100.948	100.171	100.750	98.705	99.528
Fisher x Young	99.529	99.726	100.948	100.185	100.752	98.711	99.528
Fisher x geom. Young	99.529	99.726	100.948	100.185	100.752	98.709	99.527
Tornqvist x Young	99.529	99.726	100.948	100.185	100.752	98.711	99.528
Tornqvist x geom. Young	99.529	99.726	100.948	100.185	100.752	98.709	99.527

Table A.3. Price indices of consumer products by price data source and index formula, February 2021-March 2021 (cont.)

Data source and combinations of index formulas	Indices (February 2021=100)						
	rice	fresh whole milk	fresh low fat milk	yoghurt	beverages and other milk products	sugar	coffee
Scanner and web scraped data							
Jevons x Young	101.809	100.806	100.556	99.577	103.667	99.482	99.341
Jevons x geom. Young	101.808	100.806	100.556	99.577	103.666	99.482	99.340
Laspeyres x Young	101.015	100.906	100.841	100.552	103.258	98.345	98.674
Laspeyres x geom. Young	101.015	100.906	100.841	100.552	103.258	98.344	98.674
Paasche x Young	100.825	100.795	100.838	100.492	102.847	98.347	98.661
Paasche x geom. Young	100.825	100.795	100.838	100.492	102.847	98.346	98.661
Fisher x Young	100.920	100.850	100.839	100.522	103.052	98.346	98.667
Fisher x geom. Young	100.920	100.850	100.839	100.522	103.052	98.345	98.667
Tornqvist x Young	100.919	100.850	100.839	100.522	103.048	98.346	98.667
Tornqvist x geom. Young	100.919	100.850	100.839	100.522	103.047	98.345	98.667
Price survey data							
Jevons x Young	99.540	99.714	100.947	100.692	101.204	99.084	99.660
Jevons x geom. Young	99.540	99.714	100.947	100.692	101.204	99.084	99.660
Laspeyres x Young	99.522	99.697	100.959	100.202	100.768	98.763	99.579
Laspeyres x geom. Young	99.522	99.697	100.959	100.202	100.768	98.763	99.579
Paasche x Young	99.519	99.713	100.958	100.173	100.763	98.756	99.577
Paasche x geom. Young	99.519	99.713	100.958	100.173	100.763	98.756	99.577
Fisher x Young	99.520	99.705	100.958	100.188	100.765	98.760	99.578
Fisher x geom. Young	99.520	99.705	100.958	100.188	100.765	98.760	99.578
Tornqvist x Young	99.520	99.705	100.958	100.188	100.765	98.760	99.578
Tornqvist x geom. Young	99.520	99.705	100.958	100.188	100.765	98.760	99.578

Table A.3. Price indices of consumer products by price data source and index formula, February 2021-March 2021 (cont.)

Data source and combinations of index formulas	Indices (February 2021=100)						
	rice	fresh whole milk	fresh low fat milk	yoghurt	beverages and other milk products	sugar	coffee
Scanner data							
Jevons x Young	101.838	100.801	100.559	99.576	103.692	99.499	99.385
Jevons x geom. Young	101.838	100.801	100.559	99.576	103.692	99.499	99.385
Laspeyres x Young	101.033	100.901	100.846	100.556	103.283	98.374	98.685
Laspeyres x geom. Young	101.033	100.901	100.846	100.556	103.283	98.374	98.685
Paasche x Young	100.840	100.790	100.842	100.496	102.869	98.377	98.671
Paasche x geom. Young	100.840	100.790	100.842	100.496	102.869	98.377	98.671
Fisher x Young	100.937	100.846	100.844	100.526	103.076	98.376	98.678
Fisher x geom. Young	100.937	100.846	100.844	100.526	103.076	98.376	98.678
Tornqvist x Young	100.936	100.846	100.844	100.525	103.071	98.376	98.678
Tornqvist x geom. Young	100.936	100.846	100.844	100.525	103.071	98.376	98.678
Web scraped data							
Jevons x Young	99.885	101.522	100.060	99.757	100.104	97.319	97.865
Jevons x geom. Young	99.885	101.522	100.060	99.757	100.104	97.319	97.865
Laspeyres x Young	99.831	101.562	100.098	99.987	99.715	94.493	98.292
Laspeyres x geom. Young	99.831	101.562	100.098	99.987	99.715	94.493	98.292
Paasche x Young	99.823	101.521	100.098	99.978	99.709	94.434	98.339
Paasche x geom. Young	99.823	101.521	100.098	99.978	99.709	94.434	98.339
Fisher x Young	99.827	101.542	100.098	99.983	99.712	94.464	98.316
Fisher x geom. Young	99.827	101.542	100.098	99.983	99.712	94.464	98.316
Tornqvist x Young	99.827	101.542	100.098	99.983	99.712	94.462	98.316
Tornqvist x geom. Young	99.827	101.542	100.098	99.983	99.712	94.462	98.316

Source: authors' calculations based on data obtained from Statistics Poland and from a retail chain.

Table A.4. Price indices of consumer products for subgroups, February 2021 – March 2021

Elementary groups and subgroups	**Indices (February 2021=100)**		
	price survey data	**scanner data**	**web scraped data**
RICE			
long grain rice	99.48	99.14	99.71
white rice	99.60	104.61	100.06
RAW AND MILK			
UHT whole milk	100.07	99.14	100.8
pasteurised whole milk	99.36	102.49	102.25
FRESH LOW FAT MILK			
UHT low fat milk	100.84	100.57	100
goat milk	100.95	100	100
pasteurised low fat milk	101.05	101.11	100.18
YOGHURT			
Actimel	103.25	106.16	100.15
fruit flavoured yoghurt	99.61	100.49	100.59
chocolate and nuts flavoured yoghurt	-	92.31	98.75
drinkable yogurt	99.04	99.62	100.34
natural yoghurt	100.92	99.79	98.97
BEVERAGES AND OTHER MILK PRODUCTS			
kefir	101.90	101.45	100.01
buttermilk	102.01	102.14	101.31
monte	101.08	110.91	100
homogenised cheese	99.84	100.59	99.11
SUGAR	**99.08**	**99.50**	**97.32**
cane sugar	99.54	100.83	101.17
white sugar	98.63	97.81	93.02
powdered sugar	-	99.88	97.94
COFFEE			
instant coffee	99.77	101.17	96.56
coffee beans	99.55	98.27	98.13
ground coffee	-	98.74	98.92

Source: authors' calculations based on data obtained from Statistics Poland and from a retail chain.

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 27–41, DOI 10.2478/stattrans-2022-0028
Received – 28.01.2021; accepted – 25.03.2022

A comparison of the method of moments estimator and maximum likelihood estimator for the success probability in the Fibonacci-type probability distribution

Yeil Kwon[1]

ABSTRACT

A Fibonacci-type probability distribution provides the probabilistic models for establishing stopping rules associated with the number of consecutive successes. It can be interpreted as a generalized version of a geometric distribution. In this article, after revisiting the Fibonacci-type probability distribution to explore its definition, moments and properties, we proposed numerical methods to obtain two estimators of the success probability: the method of moments estimator (MME) and maximum likelihood estimator (MLE). The ways both of them performed were compared in terms of the mean squared error. A numerical study demonsrated that the MLE tends to outperform the MME for most of the parameter space with various sample sizes.

Key words: Fibonacci probability distribution, generalized polynacci distribution, factorial moment generating function, method of moments, maximum likelihood estimator.

1. Introduction

A geometric random variable is defined by the number of independent Bernoulli trials until the first success with a success probability p. As a generalized version of the geometric random variable, a negative binomial random variable is defined by the number of independent Bernoulli trials until r successes. The negative binomial random variable does not require the r successes to be consecutive. It seems natural to be interested in the case in which we stop the Bernoulli trials after reaching r *consecutive* successes. For example, what is the probability that we need 10 independent Bernoulli trials to have three consecutive successes (Moivre, 1756)?

A Fibonacci-type probability distribution describes the behavior of a random variable N defined by the number of independent Bernoulli trials until the k-th consecutive success with a success probability p. Shane (1973) derived a probability mass function and distribution function of N using polynacci polynomials, and Turner (1979) approached the same problem with the Pascal-T triangle. Philippou et al. (1982, 1983) developed a new formula for the probability function for N in terms of the multinomial coefficient and Fibonacci polynomials of order k. Philippou also made a significant contribution to deriving the convolutions of Fibonacci-type polynomials (Philippou et al., 1985; Philippou & Makri, 1985; Philippou & Georghiou, 1989) and the distribution of the multivariate Fibonacci-type polynomials of order k (Philippou & Antzoulakos, 1990, 1991).

[1] University of Central Arkansas, USA. E-mail: ykwon1@uca.edu. ORCID: https://orcid.org/0000-0002-1663-5401.

A Fibonacci-type probability distribution potentially can be applied to numerous areas such as quality control, engineering, and transportation. For example, we can find the direct applications of negative binomial distribution in quality control (Das, 2003; Ma & Zhang,1996). Using a Fibonacci-type probability distribution can be an alternative way to improve the quality control process. Suppose a production line supervisor wants to make a stopping rule to control a defective rate. The supervisor can set a rule to stop the production line for inspection when three consecutive defectives are observed. In general, "consecutive defectives" indicate another type of hidden risk of the production line that cannot be captured by a stopping rule based on the negative binomial distribution. Thus, the stopping rule based on the Fibonacci-type probability distribution is an attractive method for multi-dimensional quality control. We can find similar applications: the number of flight operations until three successive accidents, the number of digital signals in specific data transmission devices until five consecutive missing signals. Needless to say, when we have the observations from the Fibonacci-type probability distribution, the precise estimation of the success probability p is one of the most critical procedures in data analysis.

This paper aims to find the estimators for the success probability p and examine their performances when we have observations from the Fibonacci-type probability distribution. We revisit the important results on the Fibonacci-type probability distribution in Section 2 and find the estimators for the success probability p using the moments of N and the likelihood function in Section 3. Although the moments of N are represented in an explicit function of p, the method of moment estimate for p is obtained using a numerical method because it is the solution to the k-th degree polynomial in p. Furthermore, since the Fibonacci-type probability distribution is defined as a recursive form, it is difficult to find the maximum likelihood function in an explicit form. We propose a numerical method to approximate the likelihood function to find the maximum likelihood estimate for p. In Section 4, we provide the numerical results, illustrating the performances of the two estimators in terms of the mean squared error (MSE). The simulation study demonstrates that the maximum likelihood estimator (MLE) has a smaller MSE than the method of moment estimator (MME) for $p > 1/2$ under various sample sizes.

2. Fibonacci-type Probability Distribution

2.1. Fibonacci Probability Distribution where $k = 2$ and $p = 1/2$

Let N be the number of coin flips until we have the first consecutive heads with $p = \Pr(H) = 1/2$. Examining a few cases, $\Pr(N=2) = \Pr(HH) = 1/4$, $\Pr(N=3) = \Pr(THH) = 1/8$, $\Pr(N=4) = \Pr(HTHH) + \Pr(TTHH) = 2/16$, $\Pr(N=5) = \Pr(HTTHH)$ $+ \Pr(THTHH) + \Pr(TTTHH) = 3/32$. In general, we can represent $\Pr(N=n)$, for a positive integer $n \geq 4$, with the following structure:

$$\Pr(N=n) = \Pr(n-3 \text{ flips with no consecutive heads})\Pr(THH) = \frac{C_{n-3}}{2^n},$$

where C_{n-3} stands for the number of arrangements of $n-3$ coin flip results with no consecutive heads. It is evident that $C_{n-3} = C_{n-4} + C_{n-5}$. Let E be the event where $n-3$ flips occur with no consecutive heads. The event E can be split into two cases:

① $n-3$ flips with no consecutive heads with the last flip being tail (T), or

② $n-3$ flips with no consecutive heads with the second last flip being tail (T)

Then, the number of arrangements for case ① is C_{n-4}, and the number of arrangements for case ② is C_{n-5}, which implies that C_n forms the Fibonacci sequence. Therefore, the probability mass function of the random variable N can be provided by the following:

$$\Pr(N=n)=f_{n-1}\left(\frac{1}{2}\right)^n, \quad n=2,3,4,\ldots, \tag{1}$$

where f_n is the n-th Fibonacci number with $f_0=0, f_1=1, f_2=1, f_3=2, f_4=3, f_5=5, f_6=8,\ldots$. Shane (1973) named the probability mass function (1) as Fibonacci probability distribution because (1) contains Fibonacci numbers. Let $s=\sum_{n=2}^{\infty}\frac{f_{n-1}}{2^n}$, which is the sum of all the probabilities of (1). Then, $s=1$ due to the following:

$$s=\frac{f_1}{2^2}+\sum_{n=3}^{\infty}\left(\frac{f_{n-2}}{2^n}+\frac{f_{n-3}}{2^n}\right)=\frac{1}{4}+\frac{1}{2}\sum_{n=2}^{\infty}\frac{f_{n-1}}{2^n}+\frac{1}{4}\left(\frac{f_0}{2}+\sum_{n=2}^{\infty}\frac{f_{n-1}}{2^n}\right)=\frac{1}{4}+\frac{1}{2}s+\frac{1}{4}s,$$

implying that $s=1$. The next problem we are interested in is $E(N)$, the expected value of N. Proposition (1) plays an important role in computing $E(N)$.

Proposition 1 *Let $m(y)$ be an infinite series of y defined by $m(y)=\sum_{n=2}^{\infty}f_{n-1}y^n$. Then, for $|y|<1/\varphi$,*

$$m(y)=\frac{y^2}{1-y-y^2},$$

where $\varphi=\lim_{n\to\infty}f_n/f_{n-1}$.

Proof:

$$\begin{aligned}
m(y)&=y^2+\sum_{n=3}^{\infty}f_{n-1}y^n=y^2+\sum_{n=2}^{\infty}f_n y^{n+1}\\
&=y^2+y\left(\sum_{n=2}^{\infty}(f_{n-2}+f_{n-1})\,y^n\right)=y^2+y\left(\sum_{n=1}^{\infty}f_{n-1}\,y^{n+1}+\sum_{n=2}^{\infty}f_{n-1}y^n\right)\\
&=y^2+y\left(y\sum_{n=2}^{\infty}f_{n-1}\,y^n+\sum_{n=2}^{\infty}f_{n-1}y^n\right)=y^2+y^2m(y)+ym(y).
\end{aligned}$$

The radius of convergence of $m(y)$ is given by $|y|<1/\varphi$ since $m(y)$ converges when

$$\lim_{n\to\infty}\left|\frac{f_n y^{n+1}}{f_{n-1}y^n}\right|=\lim_{n\to\infty}\left|\frac{f_n}{f_{n-1}}\right||y|=\varphi\,|y|<1.$$

□

For a probability mass function of N in (1), we can also show $\sum_{n=2}^{\infty}\Pr(N=n)=1$ by using Proposition (1) because $\sum_{n=2}^{\infty}\Pr(N=n)=\sum_{n=2}^{\infty}\frac{f_{n-1}}{2^n},=m(1/2)=1$. Next, we find a way to calculate the mean and variance of N. From Proposition 1, we have

$$m'(y)=\sum_{n=2}^{\infty} n f_{n-1} y^{n-1}=\frac{y(2-y)}{(1-y-y^2)^2},$$

and

$$m''(y)=\sum_{n=2}^{\infty} n(n-1) f_{n-1} y^{n-2}=\frac{2(1+3y^2-y^3)}{(1-y-y^2)^3}.$$

As $ym'(y)=\sum_{n=2}^{\infty} n f_{n-1} y^n,$ the expected value of N can be obtained by

$$E(N)=\sum_{n=2}^{\infty} n\Pr(N=n)=\sum_{n=2}^{\infty} n f_{n-1}\left(\frac{1}{2}\right)^n=ym'(y)\Big|_{y=1/2}=6.$$

Furthermore, by using simple algebra, we can find

$$y^2 m''(y)=\sum_{n=2}^{\infty} n^2 f_{n-1} y^n-\sum_{n=2}^{\infty} n f_{n-1} y^n,$$

resulting in

$$\sum_{n=2}^{\infty} n^2 f_{n-1} y^n=y^2 m''(y)+ym'(y). \tag{2}$$

Therefore, with $y=1/2$ in (2),

$$E(N^2)=\sum_{n=2}^{\infty} n^2 f_{n-1}\left(\frac{1}{2}\right)^n=\left(\frac{1}{2}\right)^2 m''\left(\frac{1}{2}\right)+\left(\frac{1}{2}\right)m'\left(\frac{1}{2}\right)=52+6=58,$$

and

$$V(N)=E(N^2)-E^2(N)=58-6^2=22.$$

The factorial moment generating function of a random variable X, with a probability mass (or density) function $f(x)$, is defined by

$$g(t)=E\left(t^X\right),$$

if this expectation exists for all values of $t\in(1-h,1+h)$. One of the well-known properties of the factorial moment generating function is that it satisfies

$$g^{(r)}(t)\Big|_{t=1}=E\left[X(X-1)\cdots(X-r+1)\right],$$

giving us factorial moments as we can infer from its name. For example, $g'(1)=E(X)$, $g''(1)=E\left[X(X-1)\right]$, and $g^{(3)}(1)=E\left[X(X-1)(X-2)\right]$. In particular, $V(X)=E(X^2)-E^2(X)=g''(1)+g'(1)-[g'(1)]^2$.

From Proposition (1), the factorial moment generating function of random variable N can be calculated using $m(y)$ because

$$g(t) = E\left(t^N\right) = \sum_{n=2}^{\infty} t^n \Pr(N=n) = \sum_{n=2}^{\infty} f_{n-1}\left(\frac{t}{2}\right)^n = m\left(\frac{t}{2}\right) = \frac{t^2}{4-2t-t^2}.$$

Moreover, since

$$g'(t) = \frac{2t(4-t)}{(4-2t-t^2)^2} \qquad \text{and} \qquad g''(t) = \frac{4(8+6t-t^3)}{(4-2t-t^2)^3},$$

we obtain $E(N) = g'(1) = 6$, $E(N(N-1)) = g''(1) = 52$, and $V(N) = E(N(N-1)) + E(N) - E^2(N) = 22$, the same value of variance for N obtained by using (2).

2.2. Fibonacci Distribution Function where $k = 2$ and $p \neq 1/2$

Next, we consider the case in which $p = p(H) \neq 1/2$. Thus, in this subsection, N is defined as the number of coin flips until we have the first consecutive heads with $p = \Pr(H) \neq 1/2$. Let $q = \Pr(T) = 1-p$, then $\Pr(N=2) = \Pr(HH) = p^2$ and $\Pr(N=3) = \Pr(THH) = qp^2$. For a positive integer $n \geq 4$, $\Pr(N=n)$ can be described as follows:

$$\begin{aligned}\Pr(N=n) &= \Pr(n-3 \text{ flips with no consecutive heads})\Pr(THH)\\ &= \Pr(n-3 \text{ flips with no consecutive heads})\,qp^2.\end{aligned}$$

Let $G_n(p,q)$ denote $\Pr(N=n)$ and examine the first several cases of $\Pr(N=n)$. Then

$$\begin{aligned}
G_2(p,q) &= \Pr(N=2) = p^2,\\
G_3(p,q) &= \Pr(N=3) = p^2q(1),\\
G_4(p,q) &= \Pr(N=4) = p^2q(q+p),\\
G_5(p,q) &= \Pr(N=5) = p^2q(q^2+2pq),\\
G_6(p,q) &= \Pr(N=6) = p^2q(q^3+3pq^2+p^2q),\\
G_7(p,q) &= \Pr(N=7) = p^2q(q^4+4pq^3+3p^2q^2).
\end{aligned}$$

Unlike (1), each $\Pr(N=n)$ does not include a Fibonacci number. Shane (1973) used the polynacci numbers and the polynacci polynomials to find the probability function of N for $p \neq 1/2$. However, we propose an alternative representation of $G_n(p,q)$ by using Fibonacci-type polynomials with a similar idea applied to (1). The probability mass function of N with the success probability p, is obtained as follows:

$$f(n) = \Pr(N=n) = G_n(p,q), \qquad n = 2,3,4,\ldots. \tag{3}$$

where $G_n(p,q)$ is a Fibonacci-type polynomial defined as

$$G_n(p,q) = \begin{cases} p/q, & \text{if } n = 0, \\ 0, & \text{if } n = 1, \\ qG_{n-1}(p,q) + pqG_{n-2}(p,q), & \text{if } n \geq 2. \end{cases} \tag{4}$$

$G_n(p,q)$ in (3) is a valid probability mass function since $\sum_{n=2}^{\infty} \Pr(N = n) = 1$:

$$\begin{aligned}
\sum_{n=2}^{\infty} G_n(p,q) &= q\sum_{n=2}^{\infty} G_{n-1}(p,q) + pq\sum_{n=2}^{\infty} G_{n-2}(p,q) \\
&= q\sum_{k=1}^{\infty} G_k(p,q) + pq\sum_{l=0}^{\infty} G_l(p,q) \\
&= q\sum_{k=2}^{\infty} G_k(p,q) + qG_1(p,q) + pq\sum_{l=2}^{\infty} G_l(p,q) + pqG_0(p,q) + pqG_1(p,q) \\
&= q\sum_{k=2}^{\infty} G_k(p,q) + pq\sum_{l=2}^{\infty} G_l(p,q) + p^2.
\end{aligned}$$

Therefore,

$$\sum_{n=2}^{\infty} G_n(p,q) = \frac{p^2}{1-q-pq} = 1.$$

Theorem 1 *Let N denote the number of coin flips until we have the first consecutive heads with $p = \Pr(H) \neq 1/2$ and $q = \Pr(T) = 1 - p$. The factorial moment generating function of N, $g(t)$, is given by*

$$g(t) = E\left(t^N\right) = \frac{p^2t^2}{1-qt-pqt^2},$$

where $G_n(p,q)$ is a Fibonacci-type polynomials defined in (4).

Proof:

$$\begin{aligned}
g(t) &= \sum_{n=2}^{\infty} \Big(qG_{n-1}(p,q) + pqG_{n-2}(p,q)\Big)t^n \\
&= \sum_{n=2}^{\infty} qG_{n-1}(p,q)t^n + \sum_{n=2}^{\infty} (pq)G_{n-2}(p,q)t^n \\
&= qt\sum_{n=2}^{\infty} G_{n-1}(p,q)t^{n-1} + pqt^2\sum_{n=2}^{\infty} G_{n-2}(p,q)t^{n-2} = qt\sum_{k=1}^{\infty} G_k(p,q)t^k + pqt^2\sum_{l=0}^{\infty} G_l(p,q)t^l \\
&= qt\left(\sum_{k=2}^{\infty} G_k(p,q)t^k + G_1(p,q)t\right) + pqt^2\left(\sum_{l=2}^{\infty} G_l(p,q)t^l + G_0(p,q)t^0 + G_1(p,q)t\right) \\
&= qtg(t) + pqt^2g(t) + pqt^2G_0(p,q) \qquad \Big(\because G_1(p,q) = 0\Big) \\
&= qtg(t) + pqt^2g(t) + p^2t^2.
\end{aligned}$$

Therefore, by solving the equation for $g(t)$, we have $g(t) = p^2t^2/(1-qt-pqt^2)$. □

In particular, when $p = q = 1/2$, we have $g(t) = t^2/(4-2t-t^2)$, $g'(t) = (8-2t^2)/(4-2t-t^2)^2$, and $g''(t) = 4\left(8+6t^2-t^3\right)/(4-2t-t^2)^3$. Thus, $E(N) = g'(1) = 6$, and $V(N) = E\left(N(N-1)\right) + E\left(N\right) - E^2\left(N\right) = g''(1) + g'(1) - \left(g'(1)\right)^2 = 22$, which are the same results as we obtained in Section 2.1. From Theorem 1, we have the expected value and the variance of N as

$$E(N) = g'(1) = \left.\frac{p^2t(2-qt)}{(1-qt-pqt^2)^2}\right|_{t=1} = \frac{1+p}{p^2}.$$

and

$$V\left(N\right) = g''(1) + g'(1) - \left(g'(1)\right)^2 = \frac{(1-p)(1+3p+p^2)}{p^4},$$

respectively, because $E\left(N(N-1)\right)$ can be obtained by

$$E\left(N(N-1)\right) = g''(1) = \left.\frac{2p^2(1+3pqt^2-pq^2t^3)}{(1-qt-pqt^2)^3}\right|_{t=1} = \frac{2(1+2p-p^2-p^3)}{p^4}.$$

2.3. Fibonacci-Type Probability Distribution with Order k

In this section, we discuss the most generalized version of the Fibonacci-type probability distribution. Let N denote the number of Bernoulli trials until we have the first k consecutive successes with the success probability p. It can be structured as follows:

$$\begin{aligned}\Pr(N=n) &= \Big(1 - \Pr\big((n-k-1) \text{ flips with } k \text{ consecutive successes}\big)\Big)\Pr(F\ \underbrace{S\cdots S}_{k \text{ successes}}\ ) \\ &= \Big(1 - \Pr\big((n-k-1) \text{ flips with } k \text{ consecutive successes}\big)\Big)qp^k.\end{aligned}$$

Cleary, $\Pr(N=k) = p^k$. If $k+1 \le n \le 2k$,
$\Pr\big((n-k-1) \text{ flips with } k \text{ consecutive successes}\big) = 0$, since $n-k-1 < k$.
Hence, $\Pr(N=n) = p^kq$, for $k+1 \le n \le 2k$. We examine several cases for $n \ge 2k+1$,

$$\begin{aligned}&H_{2k+1}(p,q) = \Pr(N=2k+1) = p^kq(1-p^k),\\ &H_{2k+2}(p,q) = \Pr(N=2k+2) = p^kq(1-2p^kq-p^{k+1}) = p^kq(1-p^k(1+q)),\\ &H_{2k+3}(p,q) = \Pr(N=2k+3) = p^kq(1-3p^kq^2-4p^{k+1}q-p^{k+2}) = p^kq(1-p^k(1+2q)),\\ &H_{2k+4}(p,q) = \Pr(N=2k+4) = p^kq(1-4p^kq^3-9p^{k+1}q^2-6p^{k+2}q-p^{k+3}) = \\ &\qquad p^kq(1-p^k(1+3q)).\end{aligned}$$

Philippou et al. (1983) found that the probability mass function for N is given by

$$f(n) = \Pr(N=n) = p^nF_{n+1-k}\left(\frac{q}{p}\right), \quad n = k, k+1, k+2, \ldots, \tag{5}$$

where

$$F_n(y) = \begin{cases} n, & \text{if } n = 0, 1, \\ y\sum_{i=1}^{n} F_{n-i}(y), & \text{if } 2 \le n \le k, \\ y\sum_{i=1}^{k} F_{n-i}(y), & \text{if } n \ge k+1, \end{cases} \tag{6}$$

and (5) and (6) can be re-represented in a simpler form (Philippou & Makri, 1985) as

$$\Pr(N=n)=H_n(p,q)=\begin{cases} p^k, & \text{if } n=k,\\ p^kq, & \text{if } k+1\le n\le 2k,\\ H_{n-1}-p^kqH_{n-1-k}, & \text{if } n\ge 2k+1. \end{cases} \tag{7}$$

It is easy to show $\sum_{n=k}^{\infty}\Pr(N=n)=1$ based on (7) because

$$\begin{aligned}
\sum_{n=k}^{\infty}H_n(p,q) &= p^k+\sum_{n=k+1}^{2k}p^kq+\sum_{n=2k+1}^{\infty}\Big(H_{n-1}(p,q)-p^kqH_{n-1-k}(p,q)\Big)\\
&= p^k+kp^kq+\sum_{m=2k}^{\infty}H_m(p,q)-p^kq\sum_{l=k}^{\infty}H_l(p,q)\\
&= p^k+kp^kq+\sum_{m=2k}^{\infty}H_m(p,q)+(k-1)p^kq+p^k\\
&-(k-1)p^kq-p^k-p^kq\sum_{l=k}^{\infty}H_l(p,q)\\
&= p^kq+\sum_{m=k}^{\infty}H_m(p,q)-p^kq\sum_{l=k}^{\infty}H_l(p,q),
\end{aligned}$$

which implies $\sum_{n=k}^{\infty}H_n(p,q)=1$.

Theorem 2 *Let N denote the number of Bernoulli trials until we have the first k consecutive successes with* $0<p=\Pr(success)<1$*, and* $q=\Pr(fail)=1-p$*. Then, the factorial moment generating function of N*, $h(t)$*, is given by*

$$h(t)=E\left(t^N\right)=\frac{p^kt^k(1-pt)}{1-t+p^kqt^{k+1}}, \tag{8}$$

where $H_n(p,q)$ *is a Fibonacci-type polynomials defined in (7).*

Proof:

$$\begin{aligned}
h(t) &= t^kp^k+\sum_{n=k+1}^{2k}t^np^kq+\sum_{n=2k+1}^{\infty}t^n\Big(H_{n-1}(p,q)-p^kqH_{n-1-k}(p,q)\Big)\\
&= t^kp^k+p^kq\left(\frac{t^{k+1}(1-t^k)}{1-t}\right)+\sum_{m=2k}^{\infty}t^{m+1}H_m(p,q)-p^kq\sum_{l=k}^{\infty}t^{l+k+1}H_l(p,q)\\
&= t^kp^k+p^kq\left(\frac{t^{k+1}(1-t^k)}{1-t}\right)+t\left(\sum_{m=2k}^{\infty}t^mH_m(p,q)+t^kp^k+p^kq\left(\frac{t^{k+1}(1-t^{k-1})}{1-t}\right)\right)\\
&\quad -t\left(t^kp^k+p^kq\left(\frac{t^{k+1}(1-t^{k-1})}{1-t}\right)\right)-p^kqt^{k+1}\sum_{l=k}^{\infty}t^lH_l(p,q)\\
&= t^kp^k(1-t)+p^kqt^{k+1}+h(t)t-h(t)p^kqt^{k+1},
\end{aligned}$$

indicating

$$h(t) = \frac{p^k t^k (1 - t + qt)}{1 - t + p^k q t^{k+1}} = \frac{p^k t^k (1 - pt)}{1 - t + p^k q t^{k+1}}.$$

The expected value of N can be computed by using $h(t)$ in (8). Moreover, since

$$h'(t) = \frac{p^k t^k \left(k(t-1)(p - 1/t) + q - p^k q t^k \right)}{(1 - t + p^k q t^{k+1})^2},$$

the expected value of N is given by

$$E(N) = h'(t)\Big|_{t=1} = \frac{1 - p^k}{p^k q}. \tag{9}$$

□

Corollary 1 *Let $N^{(k)}$ be the number of Bernoulli trials until we have the first k consecutive successes with $0 < p = \Pr(success) < 1$, and $q = \Pr(fail) = 1 - p$. Then,*

$$E\left(N^{(k+1)}\right) = \frac{1}{p} E\left(N^{(k)} + 1\right).$$

Proof: From (9), we have

$$E\left(N^{(k+1)}\right) = \frac{1 - p^{k+1}}{p^{k+1} q} = \frac{1}{p}\left(\frac{1 - p^k}{p^k q} + \frac{p^k - p^{k+1}}{p^k q} \right) = \frac{1}{p} E\left(N^{(k)} + 1\right),$$

since $p^k - p^{k+1} = p^k(1 - p) = p^k q$.

□

Corollary 1 indicates that the expected value of $N^{(k)}$ increases exponentially when k increases with a growth factor of $1/p$. Hence, it escalates dramatically as p is close to 0, for example, when $p = 1/2$, $E\left(N^{(2)}\right) = 6$, $E\left(N^{(3)}\right) = 14$ and $E\left(N^{(4)}\right) = 30$. For $p = 1/10$, $E\left(N^{(2)}\right) = 110$, $E\left(N^{(3)}\right) = 1110$ and $E\left(N^{(4)}\right) = 11110$.

3. Estimation Methods for p

3.1. Method of Moments Estimator (MME) for p

Assume a random sample with size m is given as

$$N_1, N_2, \ldots, N_m \overset{iid}{\sim} f(n), \tag{10}$$

where $f(n)$ is a pmf defined in (7). The first sample moment (the sample mean) is given by $\bar{N} = \sum_{i=1}^{m} N_i/m$. By replacing $E(N)$ with $\bar{N}$ in (9), we have the following equation:

$$\bar{N}p^{k+1} - (\bar{N}+1)p^k + 1 = (p-1)(\bar{N}p^k - p^{k-1} - \cdots - p - 1) = 0.$$

Thus, a positive root in $(0,1]$ of the equation $\bar{N}p^k - p^{k-1} - \cdots - p - 1 = 0$ becomes the MME for p. Proposition 2 indicates that a unique $\hat{p}_{mme}$ exists, and, in particular, if $\bar{N} = k$, the MME for p becomes 1. We have $\bar{N} = k$ only when all N_i's $(i = 1, \ldots, m)$ in (10) are equal to k, and this event occurs with the probability p^{km}. For example, if $p = 1/2$, $k = 2$ and $m = 10$, $\Pr(\bar{N} = k) = 1/2048$.

Proposition 2 *Let $\bar{N} = \sum_{i=1}^{m} n_i/m$ be a sample mean obtained from (10). Then, for $\bar{N} > k$,*

$$r(p) = \bar{N}p^k - p^{k-1} - \cdots - p - 1 \tag{11}$$

has only one zero in $(0,1)$. When $\bar{N} = k$, the solution of the equation $r(p) = 0$ is given by $p = 1$.

Proof:

First, suppose $\bar{N} > k$. We know $r(0) = -1 < 0$ and $r(1) = \bar{N} - k > 0$. Because $r(p)$ is a continuous function on $[0,1]$, by the intermediate value theorem, it has at least one zero in $(0,1)$. Furthermore, from the Descartes' rule (Albert, 1943), as there is one sign change in the coefficients of $r(p)$, the equation $r(p) = 0$ has exactly one positive root. Therefore, $r(p)$ has only one zero in $(0,1)$. In particular, when $\bar{N} = k$, $r(p)$ can be factored as

$$r(p) = (p-1)(kp^{k-1} + (k-1)p^{k-2} + (k-2)p^{k-3} + \cdots + 3p^2 + 2p + 1).$$

In addition, it turns out $kp^{k-1} + (k-1)p^{k-2} + (k-2)p^{k-3} + \cdots + 3p^2 + 2p + 1 = 0$ has no positive root by using the Descartes' rule again. Hence, the only root of $r(p) = 0$ on $[0,1]$ is $p = 1$.

□

For instance, when $k = 2$, the MME for p is the solution to the quadratic equation $\bar{N}p^2 - p - 1 = 0$, and it turns out to be

$$\hat{p}_{mme} = \frac{1 + \sqrt{1 + 4\bar{N}}}{2\bar{N}}.$$

When $p \geq 3$, $\hat{p}_{mme}$ can be obtained by finding the root of (11) with the numerical methods such as Newton's method and Halley's method.

3.2. Maximum Likelihood Estimator (MLE) for p

Under the same assumption of (10), the log-likelihood function $l(p)$ is

$$l(p) = \sum_{i=1}^{m} \ln \left[H_{n_i}(p,q) \right],$$

and, the maximum likelihood estimator of p is given by

$$\hat{p}_{mle} = \underset{0<p<1}{\arg\max} \sum_{i=1}^{m} \ln \left[H_{n_i}(p,q) \right].$$

However, because $H_{n_i}(p,q)$ is not provided in a closed form but in a recursive form, analytical derivation for the MLE of p is extremely challenging. Hence, a numerical method is proposed as follows:

- (Step 1) Discretize the values of $p \in (0,1)$, for example, $p_d = 0.01, 0.02, \ldots, 0.98, 0.99$.
- (Step 2) Given $n_1, \ldots, n_m$, and k, calculate $f(n_i) = H_{n_i}(p_d, q_d)$ based on (7) for all $i = 1, \ldots, m$ and all the discretized values of p_d.
- (Step 3) Approximate the log-likelihood function $l(p)$ by computing
$$l(p_d) = \sum_{i=1}^{m} \ln \left[H_{n_i}(p_d, q_d) \right].$$
- (Step 4) Find the optimal value of p_d^*, which maximizes $l(p_d)$.

4. Numerical Study

In this section, we compare the performance of the MME ($\hat{p}_{mme}$) and the computationally driven MLE ($\hat{p}_{mle}$) for the success probability p in terms of the MSE. In general, the MSE of an estimator $\hat{\theta}$ for a parameter θ is defined by $E(\hat{\theta} - \theta)^2$, and it can be decomposed as the sum of the variance of $\hat{\theta}$ and the squared bias of $\hat{\theta}$. In this simulation study, the MSE is estimated and decomposed by

$$\widehat{MSE}\left(\hat{\theta},\, \theta\right) = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}_r - \theta \right)^2 = \widehat{Bias}^2(\hat{\theta}) + \widehat{Var}(\hat{\theta}),$$

where

$$\widehat{Bias}^2(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} \left(\bar{\hat{\theta}} - \theta \right)^2, \quad \widehat{Var}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}_r - \bar{\hat{\theta}} \right)^2, \quad \text{and} \quad \bar{\hat{\theta}} = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}_r,$$

with R the total number of simulations, and $\hat{\theta}_r$ the estimate for θ in the r-th repetition. Here, $\hat{\theta}$ represents both $\hat{p}_{mme}$ and $\hat{p}_{mle}$. We set the true success probability $p = 0.1, 0.3, 0.5, 0.7, 0.9$, the sample size $m = 5, 10, 20, 30, 50$, and the number of simulation $R = 500$.

Table 1 displays the results of the simulation when $k = 2$. Table 2 is in the same format as Table 1 but presents the results when $k = 4$. In other words, Table 1 illustrates the results of the numerical study with a random variable N defined by the number of Bernoulli trials until we have two consecutive successes. The estimated squared bias, variance, and MSE for each estimator are reported with units in 10^{-3}. Ratio columns display the ratios of the estimated squared bias, variance, and MSE for $\hat{p}_{mme}$ and $\hat{p}_{mle}$. Hence, the values of the ratio that are greater than 1 imply that $\hat{p}_{mle}$ outperforms $\hat{p}_{mme}$. For example, in Table 1, for $m = 10$ and $p = 0.7$, the value in the ratio column of $\widehat{Bias}^2$ is computed as

Table 1. Squared bias, variance, and mean squared error of $\hat{p}_{mme}$ and $\hat{p}_{mle}$ for $k = 2$ (unit: 10^{-3}). The ratio columns represent the values of the squared bias (or variance) of $\hat{p}_{mme}$ divided by the squared bias (or variance) of $\hat{p}_{mle}$.

$k = 2$		$\widehat{Bias^2}$			$\widehat{Var}$			$\widehat{MSE}$		
Sample Size	p	$\hat{p}_{mme}$	$\hat{p}_{mle}$	Ratio	$\hat{p}_{mme}$	$\hat{p}_{mle}$	Ratio	$\hat{p}_{mme}$	$\hat{p}_{mle}$	Ratio
$m = 5$	0.1	0.0817	0.0819	0.997	0.800	0.804	0.994	0.884	0.886	0.997
	0.3	0.4885	0.4767	1.025	6.494	6.396	1.015	6.982	6.873	1.016
	0.5	1.9970	1.8523	1.078	14.11	13.84	1.019	16.11	15.70	1.026
	0.7	0.7481	0.5482	1.365	13.79	13.21	1.043	14.54	13.76	1.056
	0.9	0.0844	0.0514	1.642	6.963	6.676	1.043	7.047	6.728	1.047
$m = 10$	0.1	0.0455	0.0457	0.997	0.366	0.366	1.000	0.412	0.412	1.000
	0.3	0.1191	0.1133	1.051	2.752	2.733	1.007	2.871	2.846	1.009
	0.5	0.2651	0.2239	1.184	6.578	6.464	1.018	6.843	6.688	1.023
	0.7	0.0467	0.0285	1.637	7.321	7.092	1.032	7.368	7.120	1.035
	0.9	0.0015	0.0001	11.28	3.900	3.783	1.031	3.902	3.783	1.031
$m = 20$	0.1	0.0079	0.0079	1.012	0.141	0.141	0.999	0.149	0.149	1.000
	0.3	0.0211	0.0206	1.025	1.267	1.264	1.002	1.288	1.285	1.003
	0.5	0.0603	0.0561	1.075	2.703	2.684	1.007	2.763	2.741	1.008
	0.7	0.0678	0.0464	1.461	3.600	3.429	1.050	3.668	3.475	1.055
	0.9	0.0112	0.0086	1.294	2.134	2.036	1.048	2.145	2.045	1.049
$m = 30$	0.1	0.0034	0.0034	1.001	0.099	0.098	1.003	0.102	0.102	1.003
	0.3	0.0034	0.0031	1.092	0.751	0.752	0.999	0.754	0.755	1.000
	0.5	0.0430	0.0419	1.027	1.832	1.826	1.003	1.875	1.868	1.004
	0.7	0.0294	0.0229	1.285	2.343	2.261	1.037	2.373	2.284	1.039
	0.9	0.0103	0.0046	2.222	1.321	1.230	1.074	1.332	1.235	1.079
$m = 50$	0.1	0.0006	0.0006	1.003	0.060	0.059	1.002	0.060	0.060	1.002
	0.3	0.0000	0.0000	2.376	0.530	0.531	0.998	0.531	0.531	1.000
	0.5	0.0072	0.0074	0.972	1.167	1.165	1.002	1.175	1.172	1.002
	0.7	0.0055	0.0029	1.926	1.456	1.411	1.032	1.461	1.414	1.034
	0.9	0.0001	0.0013	0.109	0.803	0.752	1.068	0.803	0.753	1.066

$\left(0.0467 \times 10^{-3}\right) / \left(0.0285 \times 10^{-3}\right) = 1.637$. This indicates the squared bias of $\hat{p}_{mme}$ is 63.7% greater than that of $\hat{p}_{mle}$ on average. We can interpret the numbers in the ratio columns of $\widehat{Var}$ and $\widehat{MSE}$ in the same manner. As for the decomposition of the MSE, Tables 1 and 2 show the variance (compared with the squared bias) explains a major portion of the MSE for both of the estimators. Except for the case with a small sample size m, and a small success probability p, more than 95% of the MSE is explained by the variance approximately. For the magnitude of the bias, the squared bias of $\hat{p}_{mle}$ is smaller than that of $\hat{p}_{mme}$ for most cases. In the variance comparison, although the values in the ratio column are not as large as the ratio values of the squared bias, the variance of $\hat{p}_{mle}$ is smaller than that of $\hat{p}_{mme}$ for most of the values of p and sample sizes. The MSE ratio of $\hat{p}_{mme}$ and $\hat{p}_{mle}$ exhibits a pattern similar to the variance ratio due to the substantial contribution of the variance to the MSE. When p is small, the MSE difference between $\hat{p}_{mme}$ and $\hat{p}_{mle}$ is not significantly large. However, for moderate and large values of p, the MSE of $\hat{p}_{mle}$ is smaller than that of $\hat{p}_{mme}$ for all sample sizes. The improvement caused by $\hat{p}_{mle}$ tends to be larger when p is closer to 1.

Table 2. Squared bias, variance, and mean squared error of $\hat{p}_{mme}$ and $\hat{p}_{mle}$ for $k = 4$ (unit: 10^{-3}). The ratio columns represent the values of the squared bias (or variance) of $\hat{p}_{mme}$ divided by the squared bias (or variance) of $\hat{p}_{mle}$.

$k=4$		$\widehat{Bias^2}$			$\widehat{Var}$			$\widehat{MSE}$		
Sample Size	p	$\hat{p}_{mme}$	$\hat{p}_{mle}$	Ratio	$\hat{p}_{mme}$	$\hat{p}_{mle}$	Ratio	$\hat{p}_{mme}$	$\hat{p}_{mle}$	Ratio
$m=5$	0.1	0.0166	0.0166	1.002	0.166	0.166	0.997	0.182	0.183	0.998
	0.3	0.1365	0.1347	1.013	1.717	1.709	1.005	1.854	1.844	1.006
	0.5	0.3298	0.3148	1.048	5.189	5.085	1.020	5.519	5.400	1.022
	0.7	0.3761	0.3158	1.191	6.316	6.066	1.041	6.692	6.382	1.049
	0.9	0.2024	0.1163	1.741	4.015	3.836	1.047	4.218	3.953	1.067
$m=10$	0.1	0.0010	0.0010	0.993	0.075	0.075	1.005	0.076	0.076	1.005
	0.3	0.0455	0.0454	1.004	0.698	0.697	1.001	0.743	0.743	1.001
	0.5	0.0818	0.0791	1.034	2.133	2.119	1.007	2.215	2.198	1.008
	0.7	0.0340	0.0194	1.754	2.852	2.745	1.039	2.886	2.765	1.044
	0.9	0.0269	0.0117	2.311	2.022	1.908	1.060	2.049	1.920	1.067
$m=20$	0.1	0.0011	0.0011	0.967	0.036	0.036	0.996	0.037	0.037	0.995
	0.3	0.0091	0.0092	0.995	0.356	0.356	1.000	0.365	0.365	1.000
	0.5	0.0136	0.0132	1.027	1.036	1.037	0.999	1.049	1.049	1.000
	0.7	0.0067	0.0051	1.313	1.609	1.582	1.017	1.616	1.587	1.018
	0.9	0.0014	0.0008	1.733	1.068	0.963	1.108	1.069	0.964	1.109
$m=30$	0.1	0.0004	0.0004	0.994	0.022	0.023	0.991	0.023	0.023	0.991
	0.3	0.0014	0.0014	0.983	0.249	0.250	0.999	0.251	0.250	1.000
	0.5	0.0017	0.0017	0.996	0.622	0.622	1.001	0.624	0.623	1.001
	0.7	0.0145	0.0123	1.180	0.926	0.896	1.034	0.941	0.908	1.036
	0.9	0.0033	0.0016	2.033	0.638	0.577	1.107	0.642	0.578	1.110
$m=50$	0.1	0.0001	0.0002	0.868	0.015	0.015	0.994	0.015	0.015	0.993
	0.3	0.0002	0.0002	1.013	0.140	0.140	1.000	0.140	0.140	1.000
	0.5	0.0013	0.0012	1.042	0.367	0.363	1.010	0.368	0.364	1.010
	0.7	0.0061	0.0059	1.035	0.594	0.592	1.004	0.600	0.598	1.003
	0.9	0.0013	0.0012	1.119	0.414	0.386	1.072	0.416	0.388	1.072

5. Conclusion

A Fibonacci-type probability distribution can be employed to determine the probabilistic behavior of a random variable N defined by the number of Bernoulli trials with a success probability p until we have k-consecutive successes. When $p = 1/2$, it can be expressed as an implicit form with the Fibonacci numbers. When $p \neq 1/2$, the Fibonacci-type probability distribution is represented in terms of Fibonacci-type polynomials recursively. We calculated the first and second moments of N by using the factorial moment generating function. In particular, the expected value of N increases exponentially with a growth factor of $1/p$ when the number of consecutive successes k increases by 1, while the expected value of a negative binomial random variable increases linearly for the unit increase of the number of successes. To compare MME with MLE, we used the computational methods to obtain the MLE by approximating the maximum likelihood function using the pmf of N defined recursively. The result of the simulation discloses that, for both MLE and MME, the biases are considerably smaller than the variances under all of the values of p and the sample sizes, indicating that the variance explains the majority of the MSE. Furthermore, we can see, in terms of the MSE, the MLE performs better than MME for a wide range of p, especially when p is greater than 1/2 for the various sample sizes.

References

Albert, G. E., (1943). An Inductive Proof of Descartes' Rule of Signs, *The American Mathematical Monthly*, 50(3), pp. 178–180.

Das, N., (2003). Study on Implementing Control Charts Assuming Negative Binomial Distribution with Varying Sample Size in a Software Industry. *Software Quality Professional*, 6, pp. 38–39.

Ma, Y., Zhang, Y., (1996). Q control charts for negative binomial distribution. *Computers & Industrial Engineering*, 31(3), pp. 813–816.

De Moivre, A. (1756). *The Doctrine of Chances*, 3rd ed., Chelsea Publishing Co., New York (reprint, 1967).

Philippou, A. N., Antzoulakos, D., (1990). Multivariate Fibonacci polynomials of order k and the multiparameter negative binomial distribution of the same order, G. E. Bergum (ed.) A. N. Philippou (ed.) A. F. Horadam (ed.), *Applications of Fibonacci Numbers* 3, Kluwer Academic Publishers, pp. 273–279.

Philippou, A. N., Antzoulakos, D., (1991).Generalized Multivariate Fibonacci Polynomials of Order k and the Multivariate Negative Binomial Distributions of the Same Order. *The Fibonacci Quarterly*, 29, 322–328.

Philippou, A. N., Georghiou, C., (1989). Convolutions of Fibonacci-type polynomials of order k and the negative binomial distributions of the same order. *The Fibonacci Quarterly*, 27(3), pp. 209–216.

Philippou, A. N., Georghiou, C., Philippou, G. N., (1983). Fibonacci polynomials of order k, multinomial expansions and probability. *International Journal of Mathematics and Mathematical Sciences*, 6(3), pp. 545–550.

Philippou, A. N., Georghiou, C., Philippou, G. N., (1985). Fibonacci-type polynomials of order k with probability applications. *The Fibonacci Quarterly*, 23(2), pp. 100–105.

Philippou, A. N., Makri, D. S., (1985). Longest Success Runs and Fibonacci-type Polynomials of Order k. *The Fibonacci Quarterly*, 23(4), pp. 338–346.

Philippou, A. N., Muwafi, A. A., (1982). Waiting for the *k*-th consecutive success and the Fibonacci sequence of order *k*. *The Fibonacci Quarterly*, 20(1), pp. 28–32.

Shane, H. D., (1973). A Fibonacci Probability Function. *The Fibonacci Quarterly*, 11(6), pp. 517–522.

Turner, S. J., (1979). Probability via the *N*-th Order Fibonacci-*T* Sequence. *The Fibonacci Quarterly*, 17(1), pp. 23–28.

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 43–64, DOI 10.2478/stattrans-2022-0029
Received – 19.04.2020; accepted –20.07.2021

Estimation of $P(X \leq Y)$ for discrete distributions with non-identical support

Mriganka Mouli Choudhury[1], **Rahul Bhattacharya**[2], **Sudhansu S. Maiti**[3]

ABSTRACT

The Uniformly Minimum Variance Unbiased (UMVU) and the Maximum Likelihood (ML) estimations of $R = P(X \leq Y)$ and the associated variance are considered for independent discrete random variables X and Y. Assuming a discrete uniform distribution for X and the distribution of Y as a member of the discrete one parameter exponential family of distributions, theoretical expressions of such quantities are derived. Similar expressions are obtained when X and Y interchange their roles and both variables are from the discrete uniform distribution. A simulation study is carried out to compare the estimators numerically. A real application based on demand-supply system data is provided.

Key words: stress-strength model, uniformly minimum variance unbiased, maximum likelihood.

1. Introduction

In the stress-strength reliability literature, the quantity $R = P(X \leq Y)$, where X defines stress and Y defines strength, is the well-known reliability function, although a lot of development was carried out in the last few decades to explore inferential aspects concerning R, under the assumption of continuous X and Y. However, X and Y may be both discrete random variables and inference on R is required in many situations. For example, in demand analysis, if the number of demanded items is considered as the stress random variable X and the corresponding number of items supplied is regarded as strength random variable Y, then X and Y are both discrete and R represents the sensitivity of demand-supply system. Another example considered is the working of regular life gadgets, like scanners and Xerox machines, where measurable resistible voltage shocks are applied to the bulbs of the machines in a time interval. Then, the number of applied shocks may define stress (X) and the number of shocks the machine can withstand may define strength (Y) and consequently R measures the reliability of the system, where stress and strength random variables are both discrete. A comprehensive account of the details of stress-strength reliability can be found in the book-length coverage of Kotz et al. (2003).

[1] Department of Statistics, Visva-Bharati University, Santiniketan-731 235, West Bengal, India. E-mail: mmc.uttarpara@gmail.com. ORCID: https://orcid.org/0000-0003-1686-6389.

[2] Department of Statistics, University of Calcutta, 35, Ballygunge Circular Road, Kolkata -700019, India. E-mail: rahul_bhattya@yahoo.com. ORCID: https://orcid.org/0000-0002-3748-717X.

[3] Department of Statistics, Visva-Bharati University, Santiniketan-731 235, West Bengal, India. E-mail: dssm1@rediffmail.com.ORCID: https://orcid.org/0000-0001-8906-6513.

However, most of the authors considered identical distributions to represent stress and strength random variables. For example, Maiti (1995) took Geometric distribution to calculate UMVU and ML estimators of R, Ivshin and Lumelskii (1995) and Sathe and Dixit (2001) considered Negative Binomial distribution to represent both stress and strength random variables. Further, Balyaev and Lumelskii (1995) and Barbiero (2003) assumed Poisson distributions for both stress and strength random variables. On the contrary, Obradovic et al. (2015) regarded two different distributions for stress and strength random variables in a recent work. Specifically, Geometric distribution is used to model stress and Poisson distribution is assumed for the strength random variable.

In all these works, supports of the stress and strength random variables are assumed to be identical and the distributions are members of One Parameter Exponential Family (OPEF) of distributions. Strength and/ or stress random variable may be uniform, that is, supports may depend on the unknown parameters.

Assuming continuous uniform distributions, Ivshin (1996) and Ali et al. (2005) explored different inferential properties of R. But as far as our knowledge goes, no discrete counterpart of such work is developed. A motivating example may be the following. Suppose someone forgets his computer password. Now, if X denotes that he inputs the right password in second draw, then X follows Discrete Uniform distribution. Again, if Y represents the number of attempts the computer allows, then Y follows Poisson random variable. So R is the probability that he can open the computer successfully and hence Discrete Uniform-Poisson model is more appropriate. Consequently, in this work, we derive theoretical expressions of UMVU estimator of R and UMVU estimator of the associated variance assuming different discrete distributions for stress and strength random variables. In particular, assuming Y to be a member of the discrete OPEF and X as discrete uniform, we derive the UMVU estimator and then do the same when both X and Y are discrete uniform with different supports. UMVU estimation of R and the derivation of the UMVU estimators of the variances of the UMVU estimators are provided in Section 2. Section 3 gives the derivation of ML estimators of R for earlier mentioned combinations. In Section 4, we provide simplified expressions of R, associated three dimensional plots and also UMVU and ML estimators of R for specific members of OPEF. We compare efficiency of UMVU and ML estimators of R numerically for various combinations in Section 5. A real application based on demand-supply system data is discussed in Section 6. Finally, Section 7 concludes with a discussion of the related issues.

2. UMVU estimation of R

Derivation of the UMVU estimator of R depends on the nature of the distributions of stress (X) and strength (Y) random variables. Consequently, we start with the derivation for regular family of discrete distributions and then extend the methodology to cover distributions with parameter dependent supports. But, if either of $\mathscr{S}_{\mathscr{X}}$ (i.e. support of X) and $\mathscr{S}_{\mathscr{Y}}$ (i.e. support of Y) involves an unknown parameter, we need to develop afresh. Although a number of discrete distributions are available in the literature, we consider the Discrete Uniform distribution to model stress and discrete OPEF to represent strength and derive UMVU estimator of R. Similar expressions are also obtained for the combination (OPEF, Discrete Uniform). Further, considering (Discrete Uniform, Discrete Uniform) combination for (X,Y), we develop UMVU estimation of R.

2.1. UMVU estimation of R for regular family of discrete distributions

Suppose stress and strength are independent random variables having distributions in the regular family of discrete distributions with the supports (identical and/or non-identical) $\mathscr{S}_{\mathscr{X}}$ and $\mathscr{S}_{\mathscr{Y}}$, respectively. Naturally, $\mathscr{S}_{\mathscr{X}}$ and $\mathscr{S}_{\mathscr{Y}}$ are independent of the parameters and $\mathscr{S}_{\mathscr{X}} \bigcap \mathscr{S}_{\mathscr{Y}}$ is non-empty. Further, assume that single but different parameters are involved in the distributions of X and Y and complete sufficient statistics T_X and T_Y exist for the family of distributions of X and Y, respectively. Since we can write

$$R = P(X \leq Y) = \sum_{j \in \mathscr{S}_{\mathscr{Y}}} P(X \leq j) P(Y = j),$$

Blackwellisation (Rao, 1973) ensures that $\phi_j(T_Y) = P(Y_1 = j/T_Y)$ is the UMVU estimator of $P(Y = j)$ and $\phi_j(T_X) = P(X_1 \leq j/T_X)$ is that of $P(X \leq j)$ for every fixed $j \in \mathscr{S}_{\mathscr{Y}}$. Then, due to independence of the distributions of X and Y and the assumption of parameter independent of supports $\mathscr{S}_{\mathscr{X}}$ and $\mathscr{S}_{\mathscr{Y}}$ give the UMVU estimator of R as

$$\widehat{R}_{UMVUE} = \sum_{j \in \mathscr{S}_{\mathscr{Y}}} \phi_j(T_Y) \phi_j(T_X).$$

The available UMVU estimators of R (Kotz et al., 2003) can all be derived from the above expression.

2.2. UMVU estimation of R for (Discrete Uniform, OPEF) combination

Suppose X has a Discrete Uniform distribution over $\{1, 2, ..., N\}$ with probability mass function (PMF)

$$\begin{aligned} P(X = x) &= \frac{1}{N} \qquad \text{if } x = 1, ..., N \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

and Y has OPEF with PMF

$$\begin{aligned} P(Y = y) &= c(\theta) h(y) exp(q(\theta) t(y)) \qquad \text{if } y = 0, 1, 2, ... \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

where, $c(\theta) = \left[\sum_{y=0}^{\infty} h(y) exp(q(\theta) t(y))\right]^{-1}$. Then, under the independence of X and Y, we have

$$R = P(X \leq Y) \quad = \quad E\left[\frac{Y}{N}\right].$$

However, if the roles of X and Y are interchanged, we get the following expression:

$$R = P(X \leq Y) \quad = \quad 1 - E\left[\frac{X-1}{N}\right].$$

In order to facilitate UMVU estimation of R, we assume that $(X_1, X_2, ..., X_{n_1})$ and $(Y_1, Y_2, ..., Y_{n_2})$ are independent samples from the distributions of X and Y, respectively. Then, it is well known that complete sufficient statistics for N and θ exist (Lehmann and Casella, 1998) and are respectively, $T_X = X_{(n_1)} = \max(X_1, X_2, ..., X_{n_1})$ and $T_Y = \sum_{i=1}^{n_2} t(Y_i)$ with respective PMFs

$$\begin{aligned} P(X_{(n_1)} = t_x) &= \frac{t_x^{n_1} - (t_x - 1)^{n_1}}{N^{n_1}} \quad \text{if } t_x = 1, 2, ..., N \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

and

$$\begin{aligned} P(T_Y = t_y) &= [c(\theta)]^{n_2} h_0(t_y) exp(q(\theta) t_y) \quad \text{if } t_y = 0, 1, 2, ... \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

where $h_0(t_y)$ is the sum of $\prod_{j=1}^{n_2} h(y_j)$ over all $(y_1, y_2, ..., y_{n_2})$ for which $\sum_{j=1}^{n_2} t(y_j) = t_y$ (Ferguson, 1967).

Since the indicator function $I[X_1 \leq Y_1]$ is unbiased for R, the Rao-Blackwell theorem coupled with Lehman-Scheffe theorem (Lehmann and Casella, 1998) expresses the UMVU estimator of R as

$$\begin{aligned} \widehat{R}_{UMVUE} &= E(I[X_1 \leq Y_1] | X_{(n_1)} = t_x, T_Y = t_y) \\ &= P(X_1 \leq Y_1 | X_{(n_1)} = t_x, T_Y = t_y) \\ &= \frac{P(X_1 \leq Y_1, X_{(n_1)} = t_x, T_Y = t_y)}{P(X_{(n_1)} = t_x, T_Y = t_y)} \\ &= \frac{\sum_{y=1}^{\min(t_x, t_y)} P(X_1 \leq y, X_{(n_1)} = t_x, T_Y = t_y, Y_1 = y)}{P(X_{(n_1)} = t_x) P(T_Y = t_y)} \\ &= \frac{\sum_{y=1}^{\min(t_x, t_y)} P(X_1 \leq y, X_{(n_1)} = t_x) P(Y_1 = y, \sum_{i=1}^{n_2} t(Y_i) = t_y)}{P(X_{(n_1)} = t_x) P(T_Y = t_y)} \\ &= \frac{\sum_{y=1}^{\min(t_x, t_y)} h(y) h_0(t_y - t(y)) \sum_{x=1}^{y} P(X_1 = x | X_{(n_1)} = t_x)}{h_0(t_y)}. \end{aligned}$$

Further, using the fact that

$$\begin{aligned} P(X_1 = x | X_{(n_1)} = t_x) &= \frac{t_x^{n_1 - 1} - (t_x - 1)^{n_1 - 1}}{t_x^{n_1} - (t_x - 1)^{n_1}} \qquad \text{if } x = 1, 2, ..., (t_x - 1) \\ &= \frac{t_x^{n_1 - 1}}{t_x^{n_1} - (t_x - 1)^{n_1}} \qquad \text{if } x = t_x, \end{aligned}$$

$\widehat{R}_{UMVUE}$ can be simplified as

$$\begin{aligned}\widehat{R}_{UMVUE} &= \frac{1}{[T_X^{n_1}-(T_X-1)^{n_1}]h_0(T_Y)}\times\\ &\sum_{y=1}^{\min(T_X,T_Y)} h(y)h_0(T_Y-t(y))\times\\ &\sum_{x=1}^{y}\left\{T_X^{n_1-1}-(T_X-1)^{n_1-1}I[x\neq T_X]+T_X^{n_1-1}I[x=T_X]\right\}.\end{aligned}$$

However, for (OPEF, Discrete Uniform) combination), in a similar way, we derive the UMVU estimator of R as

$$\begin{aligned}\widehat{R}_{UMVUE} &= \frac{1}{[T_Y^{n_2}-(T_Y-1)^{n_2}]h_0(T_X)}\times\\ &\sum_{y=1}^{\min(T_X,T_Y)}\left\{(T_Y^{n_2-1}-(T_Y-1)^{n_2-1})I[y\neq T_Y]+T_Y^{n_2-1}I[y=T_Y]\right\}\times\\ &\sum_{x=0}^{y} h(x)h_0(T_X-t(x)).\end{aligned}$$

2.3. UMVU estimation of R for (Discrete Uniform, Discrete Uniform) combination

Now, assume that the distributions of both X and Y are Discrete Uniform with respective parameters N_1 and N_2. Then, the expression of R takes the form:

$$\begin{aligned}R=P(X\leq Y) &= \frac{2N_2-N_1+1}{2N_2} \qquad \text{if } N_1<N_2\\ &= \frac{N_2+1}{2N_1} \qquad \text{if } N_1\geq N_2.\end{aligned}$$

It is well known that for such distributions, complete sufficient statistics exist and are given by $T_X=X_{(n_1)}=\max(X_1,X_2,......,X_{n_1})$ and $T_Y=Y_{(n_2)}=\max(Y_1,Y_2,......,Y_{n_2})$, respectively. Thus the UMVU estimator of R takes the form

$$\begin{aligned}
\widehat{R}_{UMVUE} &= E(I[X_1 \leq Y_1]|X_{(n_1)} = t_x, Y_{(n_2)} = t_y) \\
&= P(X_1 \leq Y_1|X_{(n_1)} = t_x, Y_{(n_2)} = t_y) \\
&= \frac{P(X_1 \leq Y_1, X_{(n_1)} = t_x, Y_{(n_2)} = t_y)}{P(X_{(n_1)} = t_x, Y_{(n_2)} = t_y)} \\
&= \frac{\sum_{y=1}^{\min(t_x,t_y)} P(X_1 \leq y, X_{(n_1)} = t_x, Y_{(n_2)} = t_y, Y_1 = y)}{P(X_{(n_1)} = t_x)P(Y_{(n_2)} = t_y)} \\
&= \frac{\sum_{y=1}^{\min(t_x,t_y)} P(X_1 \leq y, X_{(n_1)} = t_x)P(Y_1 = y, Y_{(n_2)} = t_y)}{P(X_{(n_1)} = t_x)P(Y_{(n_2)} = t_y)} \\
&= \sum_{y=1}^{\min(t_x,t_y)} \sum_{x=1}^{y} P(X_1 = x|X_{(n_1)} = t_x)P(Y_1 = y|Y_{(n_2)} = t_y).
\end{aligned}$$

Now, using the expressions of the conditional PMF's $P(X_1 = x|X_{(n_1)} = t_x)$ and $P(Y_1 = y|Y_{(n_2)} = t_y)$, we derive the simplified expression

$$\begin{aligned}
\widehat{R}_{UMVUE} = {} & \frac{1}{\left\{T_X^{n_1} - (T_X - 1)^{n_1}\right\}\left\{T_Y^{n_2} - (T_Y - 1)^{n_2}\right\}} \times \\
& \sum_{y=1}^{\min(T_X,T_Y)} \left\{(T_Y^{n_2-1} - (T_Y - 1)^{n_2-1})I[y \neq T_Y] + T_Y^{n_2-1}I[y = T_Y]\right\} \times \\
& \sum_{x=1}^{y} \left\{(T_X^{n_1-1} - (T_X - 1)^{n_1-1})I[x \neq T_X] + T_X^{n_1-1}I[x = T_X]\right\}.
\end{aligned}$$

2.4. UMVU estimation of $Var(\widehat{R}_{UMVUE})$

For the UMVU estimation of $Var(\widehat{R}_{UMVUE})$, we consider the representation

$$\begin{aligned}
Var(\widehat{R}_{UMVUE}) &= E([\widehat{R}_{UMVUE}]^2) - E^2(\widehat{R}_{UMVUE}) \\
&= E([\widehat{R}_{UMVUE}]^2) - R^2
\end{aligned}$$

Therefore, if we can derive the UMVU estimator $\widehat{Q}_{UMVUE}$ of $Q = R^2$, then we can write

$$Var(\widehat{R}_{UMVUE}) = E([\widehat{R}_{UMVUE}]^2) - E(\widehat{Q}_{UMVUE}),$$

and hence obtain the UMVU estimator of $Var(\widehat{R}_{UMVUE})$ as $\widehat{Var}(\widehat{R}_{UMVUE}) = [\widehat{R}_{UMVUE}]^2 - \widehat{Q}_{UMVUE}$. Consequently, we move our attention to deriving $\widehat{Q}_{UMVUE}$.

For the relevant derivation, first of all, we note that the events $(X_1 \leq Y_1)$ and $(X_2 \leq Y_2)$ are independent and so are the corresponding indicator functions $I[X_1 \leq Y_1]$ and $I[X_2 \leq Y_2]$.

Then, naturally $I[X_1 \leq Y_1, X_2 \leq Y_2] = I[X_1 \leq Y_1]I[X_2 \leq Y_2]$ is unbiased for $Q = R^2$ and corresponding to the (Discrete Uniform, One Parameter Exponential family) combination, we derive

$$\begin{aligned}
\widehat{Q}_{UMVUE} &= E(I[X_1 \leq Y_1, X_2 \leq Y_2]|T_X = t_x, T_Y = t_y) \\
&= P(X_1 \leq Y_1, X_2 \leq Y_2|T_X = t_x, T_Y = t_y) \\
&= \frac{P(X_1 \leq Y_1, X_2 \leq Y_2, T_X = t_x, T_Y = t_y)}{P(T_X = t_x, T_Y = t_y)} \\
&= \frac{\sum_{y_1=1}^{\min(t_x,t_y)} \sum_{y_2=1}^{y_1} P(X_1 \leq y_1, X_2 \leq y_2, T_X = t_x, T_Y = t_y, Y_1 = y_1, Y_2 = y_2)}{P(T_X = t_x)P(T_Y = t_y)} \\
&= \frac{\sum_{y_1=1}^{\min(t_x,t_y)} \sum_{y_2=1}^{y_1} P(X_1 \leq y_1, X_2 \leq y_2, T_X = t_x)P(T_Y = t_y, Y_1 = y_1, Y_2 = y_2)}{P(T_X = t_x)P(T_Y = t_y)}.
\end{aligned}$$

Since, for $x_2 \leq x_1$,

$$\begin{aligned}
P(X_1 = x_1, X_2 = x_2 / X_{(n_1)} = t_x) &= \frac{t_x^{n_1-2} - (t_x-1)^{n_1-2}}{t_x^{n_1} - (t_x-1)^{n_1}} \qquad \text{if } x_1 = 1,2,...,(t_x-1) \\
&= \frac{t_x^{n_1-2}}{t_x^{n_1} - (t_x-1)^{n_1}} \qquad \text{if } x_1 = t_x,
\end{aligned}$$

we get the simplified expression of $\widehat{Q}_{UMVUE}$ for the (Discrete Uniform, OPEF) combination as

$$\begin{aligned}
\widehat{Q}_{UMVUE} = & \frac{1}{h_0(T_Y)[T_X^{n_1} - (T_X-1)^{n_1}]} \times \\
& \sum_{y_1=1}^{\min(T_X,T_Y)} \sum_{y_2=1}^{y_1} h(y_1)h(y_2)h_0(T_Y - t(y_1) - t(y_2)) \times \\
& \sum_{x_1=1}^{y_1} \sum_{x_2=1}^{y_2} \left\{T_X^{n_1-2} - (T_X-1)^{n_1-2} I[x_1 \neq T_X] + T_X^{n_1-2} I[x_1 = T_X]\right\} I[x_2 \leq x_1].
\end{aligned}$$

In a similar way, we derive the UMVU estimator of Q for (OPEF, Discrete Uniform) combination as

$$\begin{aligned}
\widehat{Q}_{UMVUE} = & \frac{1}{h_0(T_X)[T_Y^{n_2} - (T_Y-1)^{n_2}]} \times \\
& \sum_{y_1=1}^{\min(T_X,T_Y)} \sum_{y_2=1}^{y_1} \left\{(T_Y^{n_2-2} - (T_Y-1)^{n_2-2}) I[y_1 \neq T_Y] + T_Y^{n_2-2} I[y_1 = T_Y]\right\} I[y_2 \leq y_1] \\
& \times \sum_{x_1=0}^{y_1} \sum_{x_2=0}^{y_2} h(x_1)h(x_2)h_0(T_X - t(x_1) - t(x_2)).
\end{aligned}$$

Finally, replacing T_Y by $Y_{(n_2)}$ and proceeding in a similar manner, we get the UMVU estimator of Q for (Discrete Uniform, Discrete Uniform) combination as

$$\begin{aligned}\widehat{Q}_{UMVUE} &= \frac{1}{[T_X^{n_1}-(T_X-1)^{n_1}][T_Y^{n_2}-(T_Y-1)^{n_2}]}\times \\ &\sum_{y_1=1}^{\min(T_X,T_Y)}\sum_{y_2=1}^{y_1}\left\{T_Y^{n_1-2}-(T_Y-1)^{n_1-2}I[y_1\neq T_Y]+T_Y^{n_1-2}I[y_1=T_Y]\right\}I[y_2\leq y_1] \\ &\times\sum_{x_1=1}^{y_1}\sum_{x_2=1}^{y_2}\left\{T_X^{n_1-2}-(T_X-1)^{n_1-2}I[x_1\neq T_X]+T_X^{n_1-2}I[x_1=T_X]\right\}I[x_2\leq x_1].\end{aligned}$$

3. Maximum likelihood (ML) estimation of R

Suppose $(X_1,X_2,.....,X_{n_1})$ and $(Y_1,Y_2,.....,Y_{n_2})$ are independent random samples from the distribution of random variables X and Y respectively. It is well known that ML estimator $\widehat{N}_{MLE}$ of N is $X_{(n_1)}$ and ML estimator $\widehat{\theta}_{MLE}$ of θ is obtained by solving

$$\frac{n_2c^{(1)}(\theta)}{c(\theta)}+q^{(1)}(\theta)\sum_{i=1}^{n_2}t(y_i)=0,$$

where superscript indicate the first order derivative.

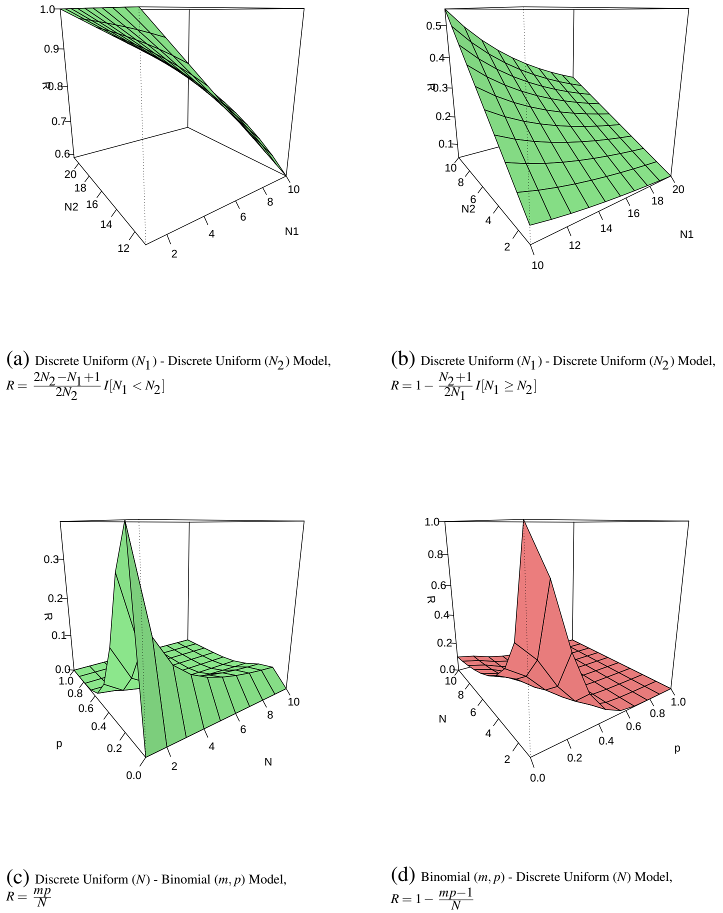
Then, by virtue of the invariance property of ML estimator we can obtain $\widehat{R}_{MLE}$ by substituting values of θ and N by $\widehat{\theta}_{MLE}$ and $\widehat{N}_{MLE}$ in the corresponding expression of R. Similarly, ML estimator of R can be obtained for OPEF - Discrete Uniform combination. $\widehat{R}_{MLE}$ for Discrete Uniform - Discrete Uniform combination can be written as

$$\begin{aligned}\widehat{R}_{MLE} &= \frac{2Y_{(n_2)}-X_{(n_1)}+1}{2Y_{(n_2)}} \qquad \text{if } X_{(n_1)}<Y_{(n_2)} \\ &= \frac{Y_{(n_2)}+1}{2X_{(n_1)}} \qquad \text{if } X_{(n_1)}\geq Y_{(n_2)}.\end{aligned}$$

4. Expressions of R and $\widehat{R}_{UMVUE}$

Theoretical expressions of R and its UMVU estimators are derived for discrete OPEF and Discrete Uniform (DU) distributions in the previous section. Now, we provide such expressions in the simplified form for different members of OPEF. In particular, apart from the well-known Binomial, Poisson, Negative Binomial (Neg Bin), Log Series distributions, we consider One Parameter Discrete Lindley (OPDL) distribution of Hussain et al. (2016). OPDL is also a member of the discrete OPEF having the PMF, $P(X=x)=(1-\phi)^2(1+x)\phi^x$, where $x=0,1,2,....$, $0<\phi<1$ and complete sufficient statistic $T_X=\sum_{i=1}^n X_i$.

Further, for brevity, we provide three dimensional plots (Figure 1-4) of R for different members of OPEF along with the concerned expressions of R. We also provide simplified expressions of UMVU and ML estimators of R for different combinations of stress and strength distributions in Tables 1-2.



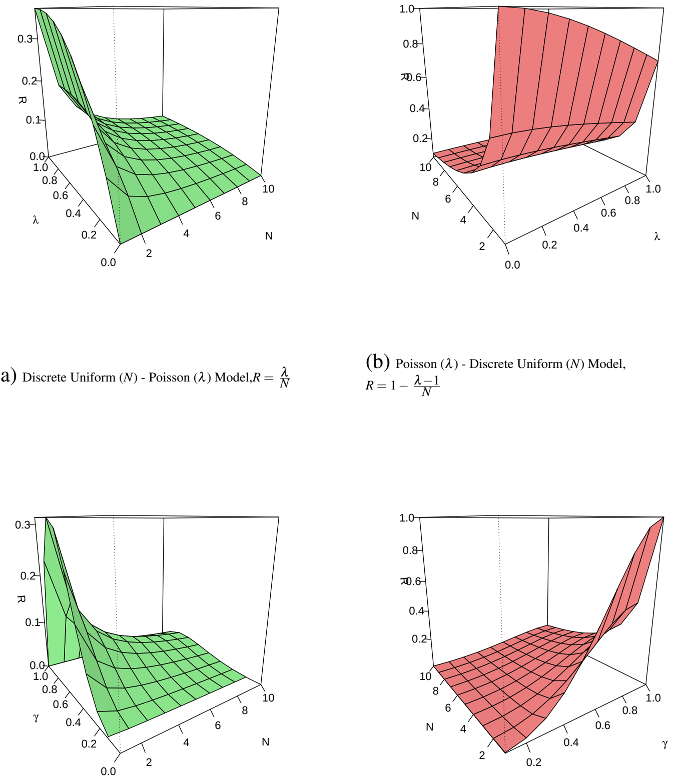
(a) Discrete Uniform (N_1) - Discrete Uniform (N_2) Model, $R = \frac{2N_2-N_1+1}{2N_2} I[N_1 < N_2]$

(b) Discrete Uniform (N_1) - Discrete Uniform (N_2) Model, $R = 1 - \frac{N_2+1}{2N_1} I[N_1 \geq N_2]$

(c) Discrete Uniform (N) - Binomial (m,p) Model, $R = \frac{mp}{N}$

(d) Binomial (m,p) - Discrete Uniform (N) Model, $R = 1 - \frac{mp-1}{N}$

Figure 1: Plot of R for Discrete Uniform - Discrete Uniform and Discrete Uniform - Binomial, Binomial - Discrete Uniform models



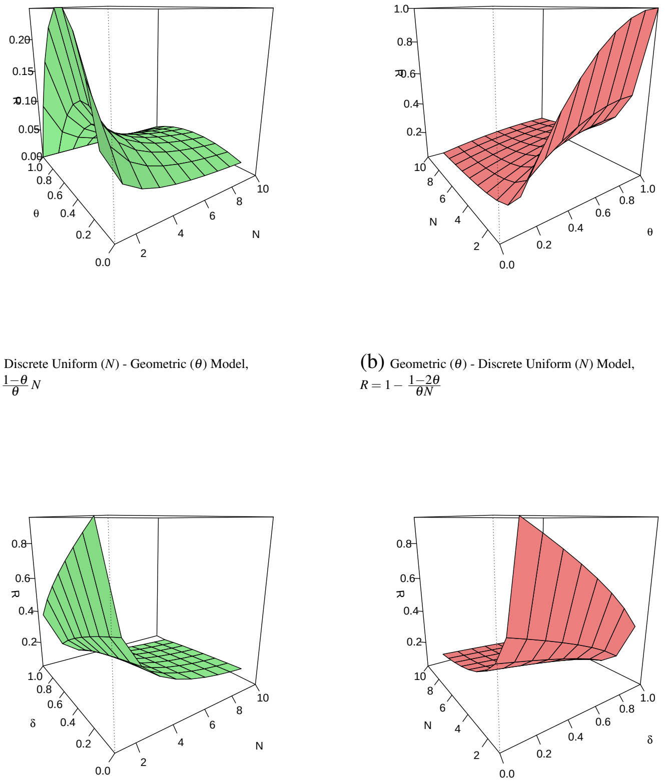
(a) Discrete Uniform (N) - Poisson (λ) Model,$R = \frac{\lambda}{N}$

(b) Poisson (λ) - Discrete Uniform (N) Model, $R = 1 - \frac{\lambda - 1}{N}$

(c) Discrete Uniform (N) - Negative Binomial (r, γ) Model, $R = \frac{r(1-\gamma)}{\gamma N}$

(d) Negative Binomial (r, γ) - Discrete Uniform (N) Model, $R = 1 - \frac{r-(r+1)\gamma}{\gamma N}$

Figure 2: Plot of R for and Discrete Uniform - Poisson, Poisson - Discrete Uniform and Discrete Uniform - Negative Binomial, Negative Binomial - Discrete Uniform models



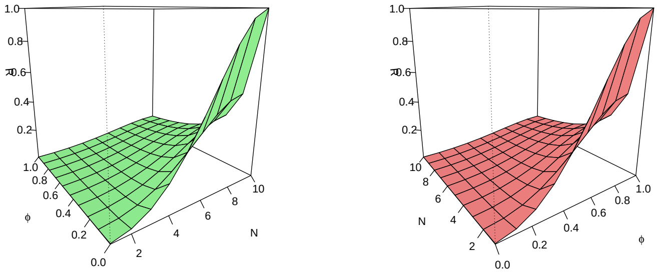
(a) Discrete Uniform (N) - Geometric (θ) Model, $R = \frac{1-\theta}{\theta} N$

(b) Geometric (θ) - Discrete Uniform (N) Model, $R = 1 - \frac{1-2\theta}{\theta N}$

(c) Discrete Uniform (N) - Log series (δ) Model, $R = -\frac{\delta}{N(1-\delta)\log(1-\delta)}$

(d) Log series (δ) - Discrete Uniform (N) Model, $R = 1 + \frac{\lambda+(1-\lambda)\log(1-\delta)}{N(1-\delta)\log(1-\delta)}$

Figure 3: Plot of R for Discrete Uniform - Geometric, Geometric - Discrete Uniform and Discrete Uniform - Log series, Log series - Discrete Uniform models



(a) Discrete Uniform (N) - OPDL (ϕ) Model, $R = \frac{2\phi}{(1-\phi)N}$

(b) OPDL (ϕ) - Discrete Uniform (N) Model, $R = 1 - \frac{3\phi-1}{N(1-\phi)}$

Figure 4: Plot of R for Discrete Uniform - OPDL and OPDL - Discrete Uniform model

5. Simulation study

In this section, it is of our interest to compare the efficiency of the estimates $\widehat{R}_{UMVUE}$ and $\widehat{R}_{MLE}$. Although estimators of $Var(\widehat{R}_{UMVUE})$ have a closed form, neither $MSE(\widehat{R}_{MLE})$ nor its estimator is analytically tractable. Therefore for the purpose of comparison, we run a simulation study with specific choices of (n_1, n_2) and different choices of stress and strength distributions.

For each such choice, we estimate, $\widehat{R}_{UMVUE}$ and $\widehat{R}_{MLE}$ together with their MSE. Finally, we report the empirical relative efficiency (ERE), defined by

$$ERE = \frac{MSE(\widehat{R}_{MLE})}{Var(\widehat{R}_{UMVUE})}$$

for different choices of parameters and distributions. Naturally $\widehat{R}_{UMVUE}$ is better or worse than $\widehat{R}_{MLE}$ as efficiency exceeds or does not exceed unity. Figure of Tables 3-8 reveal the superiority of $\widehat{R}_{MLE}$ over $\widehat{R}_{UMVUE}$ for most of the assumed configuration.

Table 1: UMVU and ML estimators of R for Discrete Uniform - Binomial, Discrete Uniform - Poisson, Discrete Uniform - Negative Binomial, Discrete Uniform - Geometric, Discrete Uniform - Log series, Discrete Uniform - OPDL models

$Model$	$\widehat{R}_{UMVUE}$	$\widehat{R}_{MLE}$
$DiscreteUniform(N)$ $Binomial(m,p)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)} \binom{m}{y}\binom{m(n_2-1)}{T_Y-y} \sum_{x=1}^{y}\left\{(T_X^{n_1-1}-(T_X-1)^{n_1-1})I[x\neq T_X]+T_X^{n_1-1}I[x=T_X]\right\}}{[T_X^{n_1}-(T_X-1)^{n_1}]\binom{mn_2}{T_Y}}$	$\frac{T_Y}{n_2 T_X}$
$DiscreteUniform(N)$ $Poisson(\lambda)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)} \frac{(n_2-1)^{T_Y-y}}{y!(T_Y-y)!} \sum_{x=1}^{y}\left\{(T_X^{n_1-1}-(T_X-1)^{n_1-1})I[x\neq T_X]+T_X^{n_1-1}I[x=T_X]\right\}}{[T_X^{n_1}-(T_X-1)^{n_1}]\frac{n_2^{T_Y}}{T_Y!}}$	$\frac{T_Y}{n_2 T_X}$
$DiscreteUniform(N)$ $NegativeBinomial(r,\gamma)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)} \binom{r+y-1}{y}\binom{(n_2-1)r+T_Y-y-1}{T_Y-y} \sum_{x=1}^{y}\left\{(T_X^{n_1-1}-(T_X-1)^{n_1-1})I[x\neq T_X]+T_X^{n_1-1}I[x=T_X]\right\}}{[T_X^{n_1}-(T_X-1)^{n_1}]\binom{n_2r+T_Y-1}{T_Y}}$	$\frac{n_2 r^2}{T_Y T_X}$
$DiscreteUniform(N)$ $Geometric(\theta)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)} \binom{n_2+T_Y-y-2}{T_Y-y} \sum_{x=1}^{y}\left\{(T_X^{n_1-1}-(T_X-1)^{n_1-1})I[x\neq T_X]+T_X^{n_1-1}I[x=T_X]\right\}}{[T_X^{n_1}-(T_X-1)^{n_1}]\binom{n_2+T_Y-1}{T_Y}}$	$\frac{n_2}{T_Y T_X}$
$DiscreteUniform(N)$ $Logseries(\delta)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)} \frac{(n_2-1)!\lvert S(T_Y-y,n_2-1)\rvert}{y(T_Y-y)!} \sum_{x=1}^{y}\left\{(T_X^{n_1-1}-(T_X-1)^{n_1-1})I[x\neq T_X]\right\}}{[T_X^{n_1}-(T_X-1)^{n_1}]\frac{n_2!\lvert S(T_Y,n_2)\rvert}{T_Y!}}$	$-\frac{\widehat{\delta}_{MLE}}{T_X(1-\widehat{\delta}_{MLE})\log(1-\widehat{\delta}_{MLE})}$
$DiscreteUniform(N)$ $OPDL(\phi)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)} (y+1)\binom{2n_2+T_Y-y-3}{T_Y-y} \sum_{x=1}^{y}\left\{(T_X^{n_1-1}-(T_X-1)^{n_1-1})I[x\neq T_X]+T_X^{n_1-1}I[x=T_X]\right\}}{[T_X^{n_1}-(T_X-1)^{n_1}]\binom{2n_2+T_Y-1}{T_Y}}$	$\frac{2log(\frac{2n_2+T_Y}{T_Y})}{(1-T_Y log(\frac{2n_2+T_Y}{T_Y}))}$

Table 2: UMVU and ML estimators of R for Binomial - Discrete Uniform, Poisson - Discrete Uniform, Negative Binomial - Discrete Uniform, Geometric - Discrete Uniform, Log series - Discrete Uniform, OPDL - Discrete Uniform models

$Model$	$\widehat{R}_{UMVUE}$	$\widehat{R}_{MLE}$
$Binomial(m,p)$ $DiscreteUniform(N)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)}\left\{(T_Y^{n_2-1}-(T_Y-1)^{n_2-1})I[y\neq T_Y]+T_Y^{n_2-1}I[y=T_Y]\right\}\sum_{x=0}^{y}\binom{m}{x}\binom{m(n_1-1)}{T_X-x}}{[T_Y^{n_2}-(T_Y-1)^{n_2}]\binom{mn_1}{T_X}}$	$1-\frac{T_X-n_1}{T_Y n_1}$
$Poisson(\lambda)$ $DiscreteUniform(N)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)}\left\{(T_Y^{n_2-1}-(T_Y-1)^{n_2-1})I[y\neq T_Y]+T_Y^{n_2-1}I[y=T_Y]\right\}\sum_{x=0}^{y}\frac{(n_1-1)^{T_X-x}}{x!(T_X-x)!}}{[T_Y^{n_2}-(T_Y-1)^{n_2}]\frac{n_1^{T_X}}{T_X!}}$	$1-\frac{T_X-n_1}{T_Y n_1}$
$NegativeBinomial(r,\gamma)$ $DiscreteUniform(N)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)}\left\{(T_Y^{n_2-1}-(T_Y-1)^{n_2-1})I[y\neq T_Y]+T_Y^{n_2-1}I[y=T_Y]\right\}\sum_{x=0}^{y}\binom{r+x-1}{x}\binom{(n_1-1)r+T_X-x-1}{T_X-x}}{[T_Y^{n_2}-(T_Y-1)^{n_2}]\binom{n_1 r+T_X-1}{T_X}}$	$1-\frac{n_1 r^2-T_X}{T_Y T_X}$
$Geometric(\theta)$ $DiscreteUniform(N)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)}\left\{(T_Y^{n_2-1}-(T_Y-1)^{n_2-1})I[y\neq T_Y]+T_Y^{n_2-1}I[y=T_Y]\right\}\sum_{x=0}^{y}\binom{n_1+T_X-x-2}{T_X-x}}{[T_Y^{n_2}-(T_Y-1)^{n_2}]\binom{n_1+T_X-1}{T_X}}$	$1-\frac{n_1-T_X}{T_Y T_X}$
$Logseries(\delta)$ $DiscreteUniform(N)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)}\left\{(T_Y^{n_2-1}-(T_Y-1)^{n_2-1})I[y\neq T_Y]+T_Y^{n_2-1}I[y=T_Y]\right\}\sum_{x=1}^{y}\frac{(n_1-1)!\lvert S(T_X-x,n_1-1)\rvert}{x(T_X-x)!}}{[T_Y^{n_2}-(T_Y-1)^{n_2}]\frac{n_1!\lvert S(T_X,n_1)\rvert}{T_X!}}$	$1+\frac{\log(1-\widehat{\delta}_{MLE})}{N(1-\widehat{\delta}_{MLE})\log(1-\widehat{\delta}_{MLE})}$
$OPDL(\phi)$ $DiscreteUniform(N)$	$\frac{\sum_{y=1}^{\min(T_X,T_Y)}\left\{(T_Y^{n_2-1}-(T_Y-1)^{n_2-1})I[y\neq T_Y]+T_Y^{n_2-1}I[y=T_Y]\right\}\sum_{x=0}^{y}(x+1)\binom{2n_1+T_X-x-3}{T_X-x}}{[T_Y^{n_2}-(T_Y-1)^{n_2}]\binom{2n_1+T_X-1}{T_X}}$	$1-\frac{3log(\frac{2n_1+T_X}{T_X})-1}{T_Y(1-log(\frac{2n_1+T_X}{T_X}))}$

$S(u,k)$ is the Stirling number of the first kind. $\widehat{\delta}_{MLE}$ is a root of equation $(1-\delta)^{\frac{1-\delta}{\delta}}=e^{-\frac{1}{\bar{x}}}$. The first (second) distribution in each cell of the first column represents Stress (Strength) distribution.

Table 3: Poisson(λ) and DU(N)

		Poisson – DU		*DU – Poisson*	
(n_1, n_2)	(λ, N)	R	ERE	R	ERE
(15 , 30)	(0.3 , 15)	0.064	0.430	0.015	0.006
(15 , 30)	(0.5 , 15)	0.061	1.577	0.020	1.728
(15 , 30)	(0.8 , 15)	0.054	0.523	0.024	0.359
(15 , 30)	(0.3 , 30)	0.032	1.947	0.007	2.125
(15 , 30)	(0.5 , 30)	0.030	0.040	0.010	0.403
(15 , 30)	(0.8 , 30)	0.027	0.939	0.012	0.435
(30 , 30)	(0.3 , 15)	0.064	0.041	0.015	1.922
(30 , 30)	(0.5 , 15)	0.061	0.032	0.020	0.783
(30 , 30)	(0.8 , 15)	0.054	0.956	0.024	0.005
(30 , 30)	(0.3 , 30)	0.032	0.110	0.007	0.533
(30 , 30)	(0.5 , 30)	0.030	4.819	0.010	0.829
(30 , 30)	(0.8 , 30)	0.027	3.461	0.012	0.034
(45 , 30)	(0.3 , 15)	0.064	0.246	0.015	0.432
(45 , 30)	(0.5 , 15)	0.061	0.398	0.020	0.602
(45 , 30)	(0.8 , 15)	0.054	0.102	0.024	0.205
(45 , 30)	(0.3 , 30)	0.032	0.202	0.007	1.148
(45 , 30)	(0.5 , 30)	0.030	1.701	0.010	0.015
(45 , 30)	(0.8 , 30)	0.027	0.601	0.012	0.060

Table 4: Binomial($m = 8$, p) and DU(N)

		Binomial − DU		*DU − Binomial*	
$(n_1\ ,n_2)$	$(p\ ,N)$	R	ERE	R	ERE
(15 , 30)	(0.3 , 15)	0.017	0.077	0.013	0.013
(15 , 30)	(0.5 , 15)	0.002	0.138	0.002	0.001
(15 , 30)	(0.8 , 15)	1.84×10^{-6}	0.415	5.46×10^{-6}	1.096
(15 , 30)	(0.3 , 30)	0.008	0.102	0.007	0.252
(15 , 30)	(0.5 , 30)	0.001	0.001	0.001	0.726
(15 , 30)	(0.8 , 30)	2.82×10^{-6}	0.387	2.73×10^{-6}	1.396
(30 , 30)	(0.3 , 15)	0.017	1.604	0.013	1.923
(30 , 30)	(0.5 , 15)	0.002	0.553	0.002	0.207
(30 , 30)	(0.8 , 15)	5.63×10^{-6}	0.859	5.46×10^{-6}	0.487
(30 , 30)	(0.3 , 30)	0.008	0.241	0.007	2.175
(30 , 30)	(0.5 , 30)	0.001	0.598	0.001	0.330
(30 , 30)	(0.8 , 30)	2.82×10^{-6}	0.214	2.73×10^{-6}	0.498
(45 , 30)	(0.3 , 15)	0.017	0.002	0.013	2.436
(45 , 30)	(0.5 , 15)	0.002	1.181	0.002	0.438
(45 , 30)	(0.8 , 15)	5.63×10^{-6}	0.642	5.46×10^{-6}	1.456
(45 , 30)	(0.3 , 30)	0.008	0.005	0.007	0.405
(45 , 30)	(0.5 , 30)	0.001	0.072	0.001	0.162
(45 , 30)	(0.8 , 30)	2.82×10^{-6}	0.266	2.73×10^{-6}	0.633

Table 5: Geometric(θ) and DU(N)

		$Geometric - DU$		$DU - Geometric$	
(n_1 , n_2)	(θ , N)	R	ERE	R	ERE
(15 , 30)	(0.3 , 15)	0.034	0.041	0.014	0.664
(15 , 30)	(0.5 , 15)	0.050	0.791	0.017	0.867
(15 , 30)	(0.8 , 15)	0.064	0.290	0.011	0.731
(15 , 30)	(0.3 , 30)	0.017	0.808	0.007	0.505
(15 , 30)	(0.5 , 30)	0.025	0.001	0.008	2.873
(15 , 30)	(0.8 , 30)	0.032	0.030	0.005	0.058
(30 , 30)	(0.3 , 15)	0.034	2.547	0.014	2.256
(30 , 30)	(0.5 , 15)	0.050	0.045	0.017	1.628
(30 , 30)	(0.8 , 15)	0.064	0.102	0.011	0.567
(30 , 30)	(0.3 , 30)	0.017	0.929	0.007	1.621
(30 , 30)	(0.5 , 30)	0.025	0.469	0.008	0.879
(30 , 30)	(0.8 , 30)	0.032	0.791	0.005	0.761
(45 , 30)	(0.3 , 15)	0.034	2.036	0.014	0.120
(45 , 30)	(0.5 , 15)	0.050	0.935	0.017	0.183
(45 , 30)	(0.8 , 15)	0.064	0.519	0.011	1.833
(45 , 30)	(0.3 , 30)	0.017	0.845	0.007	0.663
(45 , 30)	(0.5 , 30)	0.025	0.122	0.008	0.984
(45 , 30)	(0.8 , 30)	0.032	0.364	0.005	0.020

Table 6: Neg Bin($r = 3$, γ) and DU(N)

		$NegBin - DU$		$DU - NegBin$	
$(n_1\ , n_2)$	$(\gamma\ , N)$	R	ERE	R	ERE
(15 , 30)	(0.3 , 15)	0.006	0.014	0.004	0.085
(15 , 30)	(0.5 , 15)	0.021	0.154	0.012	4.031
(15 , 30)	(0.8 , 15)	0.055	4.173	0.020	0.118
(15 , 30)	(0.3 , 30)	0.003	0.015	0.002	1.790
(15 , 30)	(0.5 , 30)	0.010	2.942	0.006	2.391
(15 , 30)	(0.8 , 30)	0.027	0.447	0.010	4.302
(30 , 30)	(0.3 , 15)	0.006	0.630	0.004	1.851
(30 , 30)	(0.5 , 15)	0.021	1.315	0.012	0.397
(30 , 30)	(0.8 , 15)	0.055	0.828	0.020	0.209
(30 , 30)	(0.3 , 30)	0.003	0.014	0.002	1.082
(30 , 30)	(0.5 , 30)	0.010	0.003	0.006	0.031
(30 , 30)	(0.8 , 30)	0.027	1.847	0.010	0.307
(45 , 30)	(0.3 , 15)	0.006	0.543	0.004	0.443
(45 , 30)	(0.5 , 15)	0.021	1.050	0.012	3.684
(45 , 30)	(0.8 , 15)	0.055	0.313	0.020	0.076
(45 , 30)	(0.3 , 30)	0.003	1.123	0.002	1.337
(45 , 30)	(0.5 , 30)	0.010	0.006	0.006	2.508
(45 , 30)	(0.8 , 30)	0.027	2.156	0.010	0.698

Table 7: OPDL(ϕ) and DU(N)

		$OPDL-DU$		$DU-OPDL$	
$(n_1\ ,n_2)$	$(\phi\ ,N)$	R	ERE	R	ERE
(15 , 30)	(0.3 , 15)	0.013	0.004	0.007	0.003
(15 , 30)	(0.5 , 15)	0.026	0.001	0.013	0.020
(15 , 30)	(0.8 , 15)	0.041	0.089	0.017	0.031
(15 , 30)	(0.3 , 30)	0.006	0.021	0.004	0.003
(15 , 30)	(0.5 , 30)	0.013	0.033	0.006	0.041
(15 , 30)	(0.8 , 30)	0.021	0.019	0.008	0.005
(30 , 30)	(0.3 , 15)	0.013	0.026	0.007	0.017
(30 , 30)	(0.5 , 15)	0.026	0.059	0.013	0.005
(30 , 30)	(0.8 , 15)	0.041	0.832	0.017	0.045
(30 , 30)	(0.3 , 30)	0.006	0.004	0.004	0.025
(30 , 30)	(0.5 , 30)	0.013	0.009	0.006	0.065
(30 , 30)	(0.8 , 30)	0.021	0.063	0.008	0.013
(45 , 30)	(0.3 , 15)	0.013	0.001	0.007	0.003
(45 , 30)	(0.5 , 15)	0.026	0.003	0.013	0.202
(45 , 30)	(0.8 , 15)	0.041	0.886	0.017	0.090
(45 , 30)	(0.3 , 30)	0.006	0.003	0.004	0.098
(45 , 30)	(0.5 , 30)	0.013	0.012	0.006	0.103
(45 , 30)	(0.8 , 30)	0.021	0.003	0.008	0.006

Table 8: Log Series(δ) and DU (N)

		$LogSeries - DU$		$DU - LogSeries$	
(n_1 , n_2)	(δ , N)	R	ERE	R	ERE
(15 , 30)	(0.3 , 15)	0.080	0.115	0.056	0.023
(15 , 30)	(0.5 , 15)	0.096	0.014	0.048	0.535
(15 , 30)	(0.8 , 15)	0.165	0.192	0.033	0.143
(15 , 30)	(0.3 , 30)	0.040	0.910	0.028	0.683
(15 , 30)	(0.5 , 30)	0.048	0.004	0.024	3.074
(15 , 30)	(0.8 , 30)	0.082	0.238	0.016	0.014
(30 , 30)	(0.3 , 15)	0.080	0.666	0.056	1.627
(30 , 30)	(0.5 , 15)	0.096	0.060	0.048	0.226
(30 , 30)	(0.8 , 15)	0.165	0.073	0.033	1.028
(30 , 30)	(0.3 , 30)	0.040	0.009	0.028	0.150
(30 , 30)	(0.5 , 30)	0.048	0.021	0.024	3.418
(30 , 30)	(0.8 , 30)	0.082	1.262	0.016	1.915
(45 , 30)	(0.3 , 15)	0.080	0.004	0.056	1.156
(45 , 30)	(0.5 , 15)	0.096	0.068	0.048	0.050
(45 , 30)	(0.8 , 15)	0.165	0.323	0.033	0.235
(45 , 30)	(0.3 , 30)	0.040	0.031	0.028	1.785
(45 , 30)	(0.5 , 30)	0.048	0.041	0.024	3.425
(45 , 30)	(0.8 , 30)	0.082	0.386	0.016	0.050

6. A real application

A uniform distribution may be used to model demand-supply system data [Hadley and Whitin (1963), Wanke (2008)]. Here, we use a demand-supply system data (Naikan et al. 2014) of spare parts from an auto ancillary unit in India, reported in Table 9. We fit Discrete Uniform with $\hat{N}_1 = 56$ for demand (X) and Discrete Uniform with $\hat{N}_2 = 45$ for supply (Y). Now, by using expression of UMVU and ML estimators of R, we obtain $\widehat{R}_{MLE} = 0.409$ and $\widehat{R}_{UMVUE} = 0.411$. Therefore, the expressions derived theoretically are well applicable in real problems to estimate reliability (R) of demand-supply system data.

Table 9: Demand and supply system data for spare parts

Week	Demand	Supply	Week	Demand	Supply	Week	Demand	Supply
1	25	30	12	50	43	23	27	31
2	38	37	13	27	31	24	36	36
3	2	16	14	19	21	25	25	30
4	28	32	15	27	31	26	30	33
5	23	29	16	18	27	27	27	31
6	7	21	17	18	27	28	17	26
8	23	29	18	34	35	29	22	29
9	56	45	19	34	35	30	12	24
10	48	41	20	34	35			
11	6	21	21	26	31			

7. Concluding Remarks

We have discussed so far the UMVU and ML estimation of $P(X \leq Y)$ considering a discrete uniform distribution to represent stress and/or strength. However, an assumption of equal (but unknown) probability for stress and/or strength is less practical. Consequently, we intend further development with a general class of distributions to model stress and/or strength, allowing non-identical and parameter dependent supports.

Acknowledgements

The authors are grateful to the Editor, the Associate Editor and the anonymous Reviewer for their valuable and insightful suggestions to improve the quality of the submitted manuscript.

References

Ali, M.M, Pal, M, Woo, J, (2005). Inference On $P(Y < X)$ in Generalized Uniform Distributions. *Calcutta Statistical Association Bulletin*, 57, pp. 35–48.

Belyaev, Y, Lumelskii, Y, (1988). Multidimensional Poisson Walks. *Journal of Mathematical Sciences*, 40, pp. 162–165.

Barbiero, A, (2013). Inference on Reliability of Stress-Strength Models for Poisson Data. *Journal of Quality and Reliability Engineering*, 2013, 8 pages.

Ferguson, S. T, (1967). Mathematical Statistics: A Decision Theoretic Approach. *Academic Press*.

Hussain, T, Aslam, M, Ahmad, M, (2016). A Two Parameter Discrete Lindley Distribution. *Revista Colombiana de Estadistica*, 39(1), pp. 45–61.

Ivshin, V, V, Lumelskii, Ya, P, (1995). Statistical estimation problems in "Stress-Strength" models.*Perm University Press, Perm, Russia*.

Ivshin, V, V, (1996). Unbiased estimation of $P(X < Y)$ and their variances in the case of Uniform and Two-Parameter Exponential distributions. *Journal of Mathematical Sciences*, 81, pp. 2790–2793.

Kotz, S, Lumelskii, Y, Pensky, M, (2003). The stress-strength model and its generalizations. *Singapore: World Scientific*.

Lehmann, E. L, Casella, G, (1998). Theory of Point Estimation. *New York: Springer*.

Maiti, S.S, (1995). Estimation of $P(X \leq Y)$ in geometric case. *Journal of Indian Statistical Association*, 33, pp. 87–91.

Obradovic, M, Jovanovic, M, Milosevic, B, Jevremovic, V, (2015). Estimation of $P(X \leq Y)$ for Geometric-Poisson model. *Hacettepe Journal of Mathematics and Statistics*, 44(4), pp. 949–964.

Rao, C. R, (1973). Linear Statistical Inference and Its Application. *John Wiley & Sons, Inc.*.

Sathe, Y.S, Dixit, U.J, (2001). Estimation of $P(X \leq Y)$ in the negative binomial distribution. *Journal of Statistical Planning and Inference*, 93, pp. 83–92.

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 65–78, DOI 10.2478/stattrans-2022-0030
Received – 21.12.2021; accepted – 26.05.2022

Interval shrinkage estimation of the parameter of exponential distribution in the presence of outliers under loss functions

Parviz Nasiri[1]

ABSTRACT

In this paper, we studied estimators based on an interval shrinkage with equal weights point shrinkage estimators for all individual target points $\bar{\theta} \in (\theta_0, \theta_1)$ for exponentially distributed observations in the presence of outliers drawn from a uniform distribution. Estimators obtained from both shrinkage and interval shrinkage were compared, showing that the estimators obtained via the interval shrinkage method perform better. Symmetric and asymmetric loss functions were also used to calculate the estimators. Finally, a numerical study and illustrative examples were provided to describe the results.

Key words: interval information, mean square error, shrinkage estimator, exponential distribution, uniform distribution, outliers, Linex loss function.

1. Introduction

We are interested in working on an exponential distribution due to its various applications in life testing in case we encounter some outliers. Suppose $(X_1, X_2, ..., X_n)$is a random sample of size n whose k out of n observations seem to be outliers and taken from a uniform distribution. Studying previous works shows that Epstein and Sobel (1954) obtained the minimum variance unbiased estimator (MVUE) for scale parameter and location parameter of exponential distribution. Bhattacharia and Srivastava (1974) work on a shrinkage estimator for scale parameter. Stein (1956) proposes non-sample information in shrinkage estimation. The shrinkage estimation contents are an innovative combination of classical estimators of parameter and a guess value for it, which is called a shrinkage target. Based on Hawkins (1980) an outlier is an observation that deviates so much from other observations and it might have been generated by a different procedure. Dixit and Nasiri (2001) estimate parameters of exponential distribution in the presence of outliers generated from uniform distribution. Nasiri and Jabbari (2009) discuss estimation of parameters of the generalized exponential distribution in the existence of outliers. Finally, Golosnoy and Liesenfeld (2011) obtain an interval for shrinkage estimators. Based of this review, we show that the interval shrinkage estimator does better than another estimator. This paper is in some way related to the investigation by Nasiri and Ebrahimi (2019), whenever now we consider the outliers generated from uniform distribution.

[1] Department of Statistics, University of Payam Noor, 19395-4697, Tehran, Iran. E-mail: pnasiri@pnu.ac.ir. ORCID:https://orcid.org/0000-0002-0827-4853.

The LINEX loss function was introduced by Varian (1975), and several others including Zellner (1986), Basu and Ebrahimi Rojo (1987) and Soliman (2000), who have used this loss function in different estimation and prediction problems. The LINEX loss function is given by:

$$L(\Delta) = e^{a\Delta} - a\Delta - 1, \quad a \neq 0,$$

With $\Delta = \frac{\hat{\theta}}{\theta}$, where $\hat{\theta}$ is an estimate of θ and a represents the shape parameter of the loss function. The behaviour of the LINEX loss function changes with the choice of a. Particularly, if a is close to zero (see Pandey (1997)), this loss function is almost equivalent to the Squared Error Loss Function(SELF) and therefore almost symmetric.

In shrinkage estimation when θ_g, a guess value of θ is available, the shrinkage estimator and its properties following Thompson (1968) is defined as

$$\hat{\theta}_{sh} = \theta_g + \omega(\hat{\theta} - \theta_g), \quad 0 \leq w < 1 \tag{1}$$

To find ω we have to consider MSE of estimator as:

$$MSE(\hat{\theta}_{sh}) = E[\hat{\theta}_{sh} - \theta]^2$$

In equation (1), to obtain MSE($\hat{\theta}_{sh}$), we consider $\hat{\theta}_{sh}$ as the shrinkage estimator, that is $\hat{\theta}_{sh} = \theta_g + \omega(\hat{\theta} - \theta)$, where $0 \leq \omega < 1$ and θ_g is our guess from parameter space (see Thompson 1968). Hence, MSE($\hat{\theta}_{sh}$) $= E(\hat{\theta}_{sh} - \theta)^2 = E(\theta_g + \omega(\hat{\theta} - \theta_g) - \theta)^2$, so

$$\begin{aligned} MSE(\hat{\theta}) &= E[\theta_g + \omega(\hat{\theta} - \theta_g) - \theta]^2 & (2)\\ &= E[\omega(\hat{\theta} - \theta) + (\omega - 1)(\theta - \theta_g)]^2 \\ &= \omega^2 MSE(\hat{\theta}, \theta) + (\omega - 1)^2 * (\theta - \theta_g)^2 + 2\omega(\omega - 1)(\theta - \theta_g)E(\hat{\theta} - \theta) \\ &= \frac{\omega^2\theta^2}{n} + (\omega - 1)^2(\theta - \theta_g)^2, & (3) \end{aligned}$$

Now, we have to minimize the MSE,

$$\frac{dMSE(\hat{\theta}_{sh})}{d\omega} = \frac{2\omega\theta^2}{n} + 2(\theta_g - \theta)^2(\omega - 1) = 0, \tag{4}$$

$$\omega^* = \frac{(\theta_g - \theta)^2}{\frac{\theta^2}{n} + (\theta_g - \theta)^2}, \tag{5}$$

So the shrinkage estimator is given by

$$\hat{\theta}_{sh} = \theta_g + [\frac{(\theta_g - \theta)^2}{\frac{\theta^2}{n} + (\theta_g - \theta)^2}](\hat{\theta} - \theta_g) \tag{6}$$

and

$$\begin{aligned} MSE(\hat{\theta}_{sh}) &= [\hat{\theta}_{sh} - \theta]^2 \\ &= E[\theta_g + B_1(\hat{\theta} - \theta_g) - \theta]^2 \\ &= {B_1}^2 MSE(\hat{\theta}) + (1 - B_1)^2(\theta_g - \theta)^2, \end{aligned}$$

where

$$B_1 = \frac{(\theta_g - \theta)^2}{\frac{\theta^2}{n} + (\theta_g - \theta)^2}. \tag{7}$$

In Section 2, we have obtained the joint distribution of $(X_1, X_2, ..., X_n)$ in the presence of k outliers. In Section 3, 4 and 5 we deal with the shrinkage estimator with the presence of outliers, a feasible interval shrinkage estimator and an interval shrinkage estimator under LINEX loss function. In Section 6, we compare the MSE and LINEX risk of the interval shrinkage estimators.

2. Joint distribution of $(X_1, X_2, ..., X_n)$ with presence of outliers

Let $X_1, X_2, ..., X_n$ be n non-negative continuous random variables such that for a given combination $(i_1, i_2, ..., i_{n-k})$ of the integers $(1, 2, ..., n)$, the following conditions hold:

a) The random variables $X_{i_1}, X_{i_2}, ..., X_{i_{n-k}}$ are independent, each having the probability density function $f(x)$.

b) The remaining random variables are also independent, each having the probability density function $g(x)$.

c) The two sets of the random variables are also independent.

d) Further, it is assumed that the combinations $(i_1, i_2, ..., i_{n-k})$ of the integers $(1, 2, 3, ..., n)$ are chosen at random with equal probability $[c(n,k)]^{-1}$ for each combination, where

$$c(n,k) = \frac{n!}{k!(n-k)!}$$

The joint density of $X_1, X_2, ..., X_n$ is given as (See Dixit and Nasiri (2001))

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i) \sum_{(i_1, i_2, ..., i_{n-k})} \prod_{j=1}^{k} [c(n,k)]^{-1} \frac{g(x_{i_j})}{f(x_{i_j})}$$

Dixit and Nasiri (2001) consider estimation of parameters of an exponential distribution in the presence of outliers generated from a uniform distribution. So, if we have random variables $(X_1, X_2, ..., X_n)$ such that k of them are a distribution with pdf $f_1(x;\theta)$

$$f_1(x;\theta) = \frac{1}{\theta}, 0 < x < \theta, \tag{8}$$

and the remaining $(n-k)$ random variables are distributed with pdf $f_2(x;\theta)$ function

$$f_2(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, x > 0, \theta > 0, \tag{9}$$

then the joint distribution of $(X_1, X_2, ..., X_n)$ is

$$f(x_1, x_2, ..., x_n; \theta) = \left[\frac{k!(n-k)!}{n!}\right]^{-1} \prod_{i=1}^{n} f_2(x_i, \theta) \sum^{*} \prod_{j=1}^{k} \frac{f_1(x_{A_j}; \theta)}{f_2(x_{A_j}; \theta)}, \tag{10}$$

where

$$\sum^{*} = \sum_{A_1=1}^{n-k+1} \sum_{A_2=A_1+1}^{n-k+2} \cdots \sum_{A_k=A_{k-1}+1}^{n}$$

.

For $f_1(x;\theta)$ and $f_2(x;\theta)$, $f(x_1, x_2, ..., x_n; \theta)$ is

$$\begin{aligned} f(x_1, x_2, ..., x_n; \theta) &= \frac{k!(n-k)!}{n!} \; \frac{e^{\frac{-\Sigma x_i}{\theta}}}{\theta^{n-k}} \sum^{*} \prod_{j=1}^{k} \frac{\frac{1}{\theta} I_{(0,\theta)}(x_{A_j})}{e^{\frac{-x_{A_j}}{\theta}}} \\ &= \frac{k!(n-k)!}{n! \; \theta^n} \; e^{\frac{-\Sigma x_i}{\theta}} \sum^{*} \prod_{j=1}^{k} \frac{I_{(0,\theta)}(x_{A_j})}{e^{\frac{-x_{A_j}}{\theta}}} \\ &= \frac{k!(n-k)!}{n! \; \theta^n} \; e^{\frac{-\Sigma x_i}{\theta}} \sum^{*} \prod_{j=1}^{k} e^{\frac{x_{A_j}}{\theta}} I_{(0,\theta)}(x_{A_j}), \end{aligned}$$

For $k = 1$; $f(x_1, x_2, ..., x_n; \theta) = \frac{1}{n\theta^n} e^{\frac{-\Sigma x_i}{\theta}} \sum_{A_1=1}^{n} e^{\frac{x_{A_1}}{\theta}} I(\theta - x_{A_1})$.
For $k = 2$; $f(x_1, x_2, ..., x_n; \theta) = \frac{2}{n(n-1)\theta^n} e^{\frac{-\Sigma x_i}{\theta}} \sum_{A_1=1}^{n-1} \sum_{A_2=A_1+1}^{n} e^{\frac{x_{A_1}+x_{A_2}}{\theta}} \; I(x_{A_1} - \theta) \; I(x_{A_2} - \theta)$.

Dixit (1987), based on the joint distribution

$$f(x_1, x_2, ..., x_n) = \prod_{j=1}^{n} \frac{g(x_{ij})}{f(x_{ij})} [C(n,k)]^{-1}$$

show that the marginal distribution of X_i is given by

$$h(x_i) = \frac{k}{n} g(x_i) + \frac{n-k}{n} f(x_i)$$

Hence,

$$\begin{aligned} f(x;\theta) &= \frac{k}{n} f_1(x;\theta) + \frac{n-k}{n} f_2(x;\theta) \\ &= \frac{k}{n}\frac{1}{\theta} I_{(0,\theta)}(x) + \frac{n-k}{n\theta}\ e^{\frac{-x}{\theta}} I_{(0,\infty)}(x), \end{aligned} \tag{11}$$

So we have

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) = E(X) = \frac{k}{n}\int_0^{\theta} \frac{1}{\theta} x\, dx + \frac{n-k}{n}\int_0^{\infty} \frac{1}{\theta} x e^{-\frac{x}{\theta}} dx = \frac{(2n-k)\theta}{2n} \\ V(\bar{X}) &= \Big(1 - \frac{2k}{3n} + \frac{k^2}{4n^2}\Big)\frac{\theta^2}{n}, \end{aligned} \tag{12}$$

It is easy to show that

$$\hat{\theta} = \frac{2n}{2n-k}\bar{X}. \tag{13}$$

which is unbiased with expectation and variance as:

$$E(\hat{\theta}) = \theta \qquad \textit{and} \qquad V(\hat{\theta}) = A^2 C \frac{\theta^2}{n}, \tag{14}$$

where $A = \frac{2n}{2n-k}$ and $C = \Big(1 - \frac{2k}{3n} + \frac{k^2}{4n^2}\Big)$.

Note: The sample size n and the number of outliers k are given parameters. But in the actual application, k is unknown and should be estimated. One of the methods is that k can be selected by evaluating the likelihood for different values of k choosing the one that maximizes the likelihood.

3. Feasible interval shrinkage estimator

In 2011, Golosnoy and Liesenfeld (2011) show the shrinkage estimator towards the interval $\theta \in [\theta_0, \theta_1] \subset R$ for unbiased conventional sample estimator of $\hat{\theta}$ with $E(\hat{\theta}) = \theta$ is given by

$$\begin{aligned} \widetilde{\theta}_{sh} =& \hat{\theta} + \sqrt{V(\hat{\theta})}\frac{\theta - \hat{\theta}}{\theta_1 - \theta_0}\Big[arctan\Big(\frac{\theta_1 - \theta}{\sqrt{V(\hat{\theta})}}\Big) - arctan\Big(\frac{\theta_0 - \theta}{\sqrt{V(\hat{\theta})}}\Big)\Big] \\ &+ \frac{V(\hat{\theta})}{2(\theta_1 - \theta_0)} \ln\Big(\frac{V(\hat{\theta}) + (\theta_1 - \theta)^2}{V(\hat{\theta}) + (\theta_0 - \theta)^2}\Big), \end{aligned} \tag{15}$$

and

$$E(\widetilde{\theta}_{sh}) = \hat{\theta} + \frac{V(\hat{\theta})}{2(\theta_1 - \theta_0)} \ln \Big[\frac{V(\hat{\theta}) + (\theta_1 - \hat{\theta})^2}{V(\hat{\theta}) + (\theta_0 - \hat{\theta})^2}\Big], \tag{16}$$

for $E(\hat{\theta}) = \theta$, we have

$$\widetilde{\theta}_{sh} = \hat{\theta} + \frac{V(\hat{\theta})}{2(\theta_1 - \theta_0)} ln \Big[\frac{V(\hat{\theta}) + (\theta_1 - \hat{\theta})^2}{V(\hat{\theta}) + (\theta_0 - \hat{\theta})^2}\Big]. \tag{17}$$

For different values of lower and upper bound of the interval, when θ_1 is far from θ_0 or $V(\hat{\theta})$ approaches zero, the $MSE(\hat{\theta})$ decreases. Furthermore, if $\hat{\theta}$ is considered as the median of the interval, $\theta_m = (\theta_0 + \theta_1)/2$, then $(\theta_1 - \hat{\theta}) = \frac{\theta_1 - \theta_0}{2}$ and $(\theta_0 - \hat{\theta}) = \frac{\theta_0 - \theta_1}{2}$. In this case, the equation(16) can be written as:

$$\begin{aligned}\widetilde{\theta}_{sh} &= \hat{\theta} + \frac{V(\hat{\theta})}{2(\theta_1 - \theta_0)} ln \Big[\frac{V(\hat{\theta}) + \frac{(\theta_1 - \theta_0)^2}{4}}{V(\hat{\theta}) + \frac{(\theta_0 - \theta_1)^2}{4}}\Big] \\ &= \hat{\theta} + \frac{V(\hat{\theta})}{2(\theta_1 - \theta_0)} ln(1) = \hat{\theta},\end{aligned}$$

$\widetilde{\theta}_{sh}$ approaches $\hat{\theta}$.

Note that the expectation and variance of $\widetilde{\theta}_{sh}$ is not easy since $\widetilde{\theta}_{sh}$ is not linear $\hat{\theta}$. Golosnoy and Liesenfeld (2011) suggest to find $\widetilde{\theta}_{sh}$ by using the first order Taylor expansion around the median point θ_m. We also define $\theta_d = (\theta_1 - \theta_0)/2$, so the equation would be as follows:

$$\widetilde{\theta}_{sh} = \theta_m + (\hat{\theta} - \theta_m) \frac{\partial \hat{\theta}(\theta_m)}{\partial \hat{\theta}} + \frac{(\hat{\theta} - \theta_m)^2}{2} \frac{\partial^2 \hat{\theta}(\theta_m)}{\partial \hat{\theta}^2} = 0$$

where

$$\frac{\partial \hat{\theta}(\theta_m)}{\partial \hat{\theta}} = 1 + \frac{V(\hat{\theta})}{\theta_1 - \theta_0} \left(\frac{\theta_0 - \hat{\theta}}{V(\hat{\theta}) + (\theta_0 - \hat{\theta})^2} + \frac{\theta_1 - \hat{\theta}}{V(\hat{\theta}) + (\theta_0 - \hat{\theta})^2} \right),$$

and

$$\frac{\partial^2 \hat{\theta}(\theta_m)}{\partial \hat{\theta}^2} = 0.$$

The resulting estimator is

$$\widetilde{\theta}_{sh} = \hat{\theta}\left(1 - \frac{V(\hat{\theta})}{V(\hat{\theta}) + \left(\left(\frac{\theta_1-\theta_0}{2}\right)^2\right)}\right) + \theta_m \frac{V(\hat{\theta})}{V(\hat{\theta}) + \left(\left(\frac{\theta_1-\theta_0}{2}\right)^2\right)}$$

We also define $\theta_d = \frac{\theta_1-\theta_0}{2}$, so the equation would be as follows:

$$\widetilde{\widetilde{\theta}}_{sh} = \hat{\theta}\left[1 - \frac{V(\hat{\theta})}{V(\hat{\theta}) + \theta_d^2}\right] + \theta_m \frac{V(\hat{\theta})}{V(\hat{\theta}) + \theta_d^2}. \tag{18}$$

For $\frac{V(\hat{\theta})}{V(\hat{\theta})+\theta_d^2}$ is constant, its variance is equal zero. So, we can easily show that

$$E(\widetilde{\widetilde{\theta}}_{sh}) = \theta - (\theta - \theta_m)\frac{V(\hat{\theta})}{V(\hat{\theta}) + \theta_d^2},$$

and

$$V(\widetilde{\widetilde{\theta}}_{sh}) = V(\hat{\theta})\left(1 - \frac{V(\hat{\theta})}{V(\hat{\theta}) + \theta_d^2}\right)^2.$$

Let $1 - \frac{V(A\bar{X})}{V(A\bar{X})+\theta_d^2} = B_2$ then

$$\widetilde{\widetilde{\theta}}_{sh} = AB_2\bar{X} + (1-B_2)\theta_m,$$

so

$$\begin{aligned}
MSE(\widetilde{\widetilde{\theta}}_{sh}) &= E\left(\widetilde{\widetilde{\theta}}_{sh} - \theta\right)^2 \\
&= E\left(B_2A\bar{X} + (1-B_2)\theta_m - \theta\right)^2 \\
&= E\left(B_2(A\bar{X} - \theta) + B_2\theta + (1-B_2)\theta_m - \theta\right)^2 \\
&= E\left(B_2(A\bar{X} - \theta) + (1-B_2)\theta + (1-B_2)\theta_m\right)^2 \\
&= E\left(B_2(A\bar{X} - \theta) + (1-B_2)(\theta - \theta_m)\right)^2 \\
&= B_2^2 MSE(\hat{\theta}_{Sh-outlier}) + (1-B_2)^2(\theta - \theta_m)^2,
\end{aligned}$$

resulting

$$MSE(\widetilde{\widetilde{\theta}}_{sh}) \leq MSE(\hat{\theta}_{Sh-outlier}).$$

4. Interval shrinkage estimation under LINEX loss function

In decision theory and quality assurance filed, loss functions are used to reflect the monetary loss or economic loss caused by deterioration of the product characteristics from the target quality. However, Berger (1985) even emphasized that the loss function should be bounded and concave, because the loss function also mimics the negative of the utility, whereas the squared-error loss, Taguchi quadratic loss in quality control, or absolute error loss is unbounded and even disturb the convexity. In some decision problems, some types of asymmetric losses are proposed. One of the most eminent examples is LINEX, which was proposed by Varian (1975) and populated by Zellner (1986).

Consider LINEX loss function for $\tilde{\tilde{\theta}}$

$$L(\Delta)=e^{a\Delta}-a\Delta-1,\ \ \Delta=\frac{\tilde{\tilde{\theta}}}{\theta}.$$

which

$$\Delta=\frac{\tilde{\tilde{\theta}}}{\theta}=\frac{AB_2}{\theta}\bar{X}+(1-B_2)\frac{\theta_m}{\theta}.$$

where $A=\frac{2n}{2n-k}$, $B_2=1-\frac{V(A\bar{X})}{V(A\bar{X})+\theta_d^2}$. In this case the risk under LINEX loss function is obtained by

$$R=E(L(\Delta))=E(e^{a\Delta}-a\Delta-1)=E(e^{a\Delta})-aE(\Delta)-1$$

where

$$\begin{aligned}aE(\Delta)=aE\left(\frac{\tilde{\tilde{\theta}}}{\theta}\right)&=\frac{a}{\theta}E(\tilde{\tilde{\theta}})=\frac{a}{\theta}E(AB_2\bar{X}+(1-B_2)\theta_m)\\&=\frac{aAB_2}{\theta}E(\bar{X})+(1-B_2)\frac{\theta_m}{\theta}\\&=\frac{aAB_2}{\theta}\left(\frac{2n-k}{2n}\theta\right)+(1-B_2)\frac{\theta_m}{\theta}\\&=\frac{aAB_2(2n-k)}{2n}+(1-B_2)\frac{\theta_m}{\theta}\end{aligned}$$

$$\begin{aligned}E(e^{a\Delta})&=E\left(e^{\frac{a\tilde{\theta}}{\theta}}\right)=E\left(e^{\frac{a}{\theta}(AB_2\bar{X}+(1-B_2)\theta_m)}\right)\\&=e^{\frac{a(1-B_2)\theta_m}{\theta}}E\left(e^{\frac{aAB_2\bar{X}}{\theta}}\right)=e^{\frac{a(1-B_2)\theta_m}{\theta}}E\left(e^{\frac{aAB_2}{n\theta}(X_1+X_2+...+X_n)}\right)\\&=e^{\frac{a(1-B_2)\theta_m}{\theta}}E\left(e^{\frac{aAB_2}{\theta}X_1}e^{\frac{aAB_2}{\theta}X_2}...e^{\frac{aAB_2}{\theta}X_n}\right)=e^{\frac{a(1-B_2)\theta_m}{\theta}}\left[E\left(e^{\frac{aAB_2}{n\theta}X}\right)\right]^n\end{aligned}$$

such that

$$
\begin{aligned}
E\left(e^{\frac{aAB_2}{n\theta}X}\right) &= \frac{k}{n}\int_0^{\theta}\frac{1}{\theta}e^{\frac{aAB_2}{n\theta}x}dx+\frac{n-k}{n}\int_0^{\infty}\frac{1}{\theta}e^{\frac{aAB_2}{n\theta}x}e^{\frac{-x}{\theta}}dx \\
&= \frac{k}{n\theta}\left[\frac{n\theta}{aAB_2}e^{\frac{aAB_2}{n\theta}x|_0^{\theta}}\right]+\frac{n-k}{n\theta}\int_0^{\infty}e^{-\left(1-\frac{aAB_2}{n}\right)\frac{x}{\theta}}dx \\
&= \frac{k}{n\theta}\left[\frac{n\theta}{aAB_2}e^{\frac{aAB_2}{n\theta}}-\frac{n\theta}{aAB_2}\right]=\frac{n-k}{n\theta}\left(\frac{n\theta}{n-aAB_2}\right) \\
&= \frac{k}{aAB_2}e^{\frac{aAB_2}{n}}-\frac{k}{aAB_2}+\frac{n-k}{n-aAB_2}
\end{aligned}
$$

Hence,

$$
R=e^{\frac{a(1-B_2)\theta m}{\theta}}\left[\frac{k}{aAB_2}e^{\frac{aAB_2}{2}}-\frac{K}{aAB_2}+\frac{n-k}{n-aAB_2}\right]^n-\frac{aAB_2(2n-k)}{2n}+(1-B_2)\frac{\theta_m}{\theta}-1
$$

5. Numerical Study

To compare the performance of mean square error (MSE) and LINEX risk of the interval shrinkage estimator $\tilde{\tilde{\theta}}_{sh}$, we carry out simulation study using R software and the results are shown in Tables 1 to 4. The shape parameter takes different values. Samples were generated with sizes $n = 10(10)(50)$ using of R software. The MSE and LINEX loss function of the interval shrinkage estimator decrease when the sample size increases. Meantime, for $k = 1, n = 9$ means that one sample is generated from the uniform distribution and 9 samples are generated from the exponential distribution. Here, the number of replicated cases is $N = 1000$. In cases of $a = -0.01$ and $a = 0.01$, they are very close to each other. It is also worth mentioning, based on the results of Tables 1 and 2, when the value of "a" tends to zero, the results are the same.

Table 1. $k = 1, \quad \theta = 4, \quad \theta_g = 3.2, \quad \theta_0 = 3.7, \quad \theta_1 = 4.2$

	a=-1		a=-0.25		a=-0.01	
n	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$
10	0.0082	0.0054	0.0082	0.0008	0.0077	0.0012
20	0.0042	0.0052	0.0055	0.0007	0.0060	0.0011
30	0.0037	0.0048	0.0043	0.0007	0.0060	0.0011
40	0.0027	0.0039	0.0030	0.0006	0.0033	0.0013
50	0.0027	0.0038	0.0025	0.0005	0.0023	0.0013

Table 2. $k = 2, \quad \theta = 4, \quad \theta_g = 3.2, \quad \theta_0 = 3.7, \quad \theta_1 = 4.2$

	a=-1		a=-0.25		a=-0.01	
n	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$
10	0.0012	0.0024	0.0021	0.0004	0.0013	0.0019
20	0.0016	0.0021	0.0017	0.0003	0.0016	0.0014
30	0.0015	0.0029	0.0017	0.0003	0.0017	0.0013
40	0.0019	0.0029	0.0017	0.0004	0.0017	0.0010
50	0.0016	0.0049	0.0016	0.0004	0.0017	0.0017

Table 3. $k = 1, \quad \theta = 4, \quad \theta_g = 3.2, \quad \theta_0 = 3.7, \quad \theta_1 = 4.2$

	a=0.01		a=0.25		a=1	
n	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$
10	0.0083	0.0013	0.0082	0.0024	0.0082	0.0006
20	0.0058	0.0013	0.0046	0.0020	0.0052	0.0005
30	0.0044	0.0010	0.0041	0.0024	0.0043	0.0004
40	0.0029	0.0013	0.0032	0.0024	0.0028	0.0003
50	0.0026	0.0015	0.0026	0.0016	0.0026	0.0002

Table 4. $k = 2, \quad \theta = 4, \quad \theta_g = 3.2, \quad \theta_0 = 3.7, \quad \theta_1 = 4.2$

	a=0.01		a=0.25		a=1	
n	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$	$MES(\tilde{\tilde{\theta}}_{sh})$	$R(\tilde{\tilde{\theta}}_{sh})$
10	0.0025	0.0005	0.0017	0.0005	0.0018	0.0002
20	0.0018	0.0010	0.0017	0.0010	0.0016	0.0002
30	0.0017	0.0017	0.0017	0.0014	0.0015	0.0002
40	0.0016	0.0016	0.0016	0.0013	0.0015	0.0004
50	0.0018	0.0018	0.0015	0.0012	0.0017	0.0001

6. Practical Example

In order to illustrate the methodology proposed in this paaper, we consider, Nelson (1982) concerning the data on time to break-down of an insulating fluid between electrodes at a voltage of 34 KV (Kilo-Volts). Data are as follows:
0.19 0.78 0.96 1.31 2.78 3.16 4.15 4.67 4.85 6.50 7.35 8.01 8.27 12.06 31.75 32.52 33.91 36.71 72.89

In the initial evaluation, one-sample Kolmogorov-Smirnov test results show that the data follow an exponential distribution. Figures 1 and 2 have been reported to be checked for the presence of outlier's data. Figure 1 shows the presence of one outlier. Investigation of this

result is based on theoretical and interval shrinkage estimation. Here, to find the number of outliers or k, we consider $\theta \in (14,15)$ and based on the sample information
$n = 19$, $\sum_{i=1}^{19} x_i = 272.82$; $\bar{x} = 14.35895$. Note that in Table 5, to determine the value of k we have

$$\hat{V}(\tilde{\tilde{\theta}}) = \hat{V}(\hat{\theta})\left(1 - \frac{\hat{V}(\hat{\theta})}{\hat{V}(\hat{\theta}) + \theta_d^2}\right)$$

such that

$$\hat{V}(\hat{\theta}) = \left[\frac{2n}{2n-k}\right]^2 \left(1 - \frac{2k}{3n} - \frac{k^2}{4n^2}\right)\frac{\hat{\theta}^2}{n}$$

According to the estimator of $\hat{V}(\tilde{\tilde{\theta}})$, it can be said that the increase in the value of k is greater than the increase in the estimator. But by rotating the value of the maximum likelihood, the value of k is determined.

Table 5.

k	$\tilde{\tilde{\theta}}_{sh}$	$V(\tilde{\tilde{\theta}}_{sh})$	Σ^*	$L(\tilde{\tilde{\theta}}\|x)$
0	14.49682	0.0055031	1	5.790283×10^{-31}
1	14.50518	0.0051386	19.29345	$5.879082\times10^{-31*}$
2	14.51281	0.0047826	172.2583	$5.831662\times10^{-31*}$
3	14.51969	0.00443596	943.3984	5.635563×10^{-31}
4	14.52582	0.00409897	3540.8260	5.287464×10^{-31}

According to the results of Table 5, the likelihood function with respect to k is maximized when k is equal to 1. So, the number of outliers is 1 and $\tilde{\tilde{\theta}}_{sh} = 14.51281$.

7. Conclusion

In an experimental situation, many a time an experimenter comes across some of the observations which are far removed from the main body of the data and hence are outliers. In this paper, shrinkage and interval shrinkage estimators are discussed for the first time with the presence of outliers generated from a uniform distribution and it is shown that the interval shrinkage estimator is better than the shrinkage estimator. Using different loss functions can also improve the performance of the estimator. It may be mentioned that the proposed method can be extended for Bayesian interval shrinkage estimation and other positive data distribution as well as for the presence of outliers from other distributions.

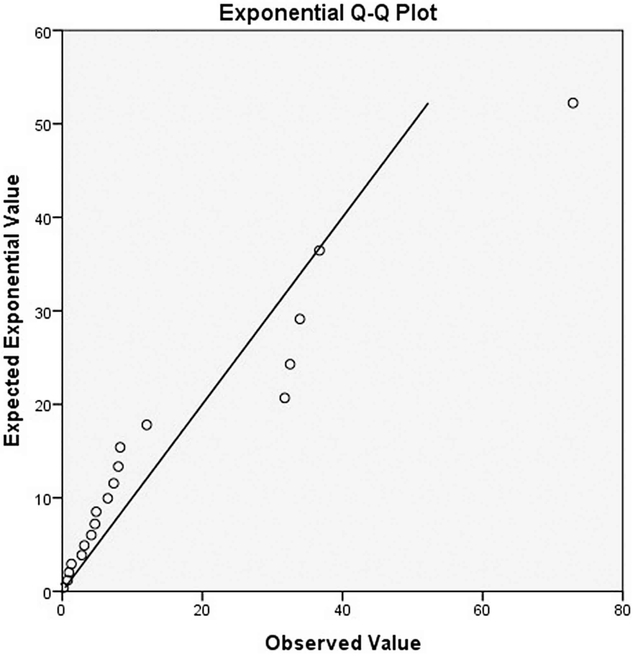


Figure 1. Exponential Q-Q Plot

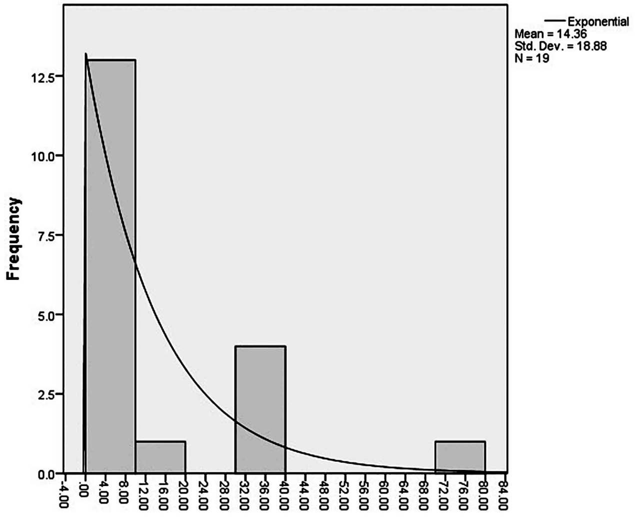


Figure 2. Frequency Distribution

8. Acknowledgment

The author would like to thank the editor and the two reviewers for their valuable comments which improved the quality of the paper.

References

Basu, A. P., & Ebrahimi, N., (1991). Bayesian approach to life testing and reliability estimation using asymmetric loss function. *Journal of statistical planning and inference*, 29(1–2), pp. 21–31.

Berger, J., (1985). *Statistical decision theory and Bayesian analysis*. Springer, second edition.

Bhattacharya, S. K., & Srivastava, V. K., (1974). A preliminary test procedure in life testing. *Journal of the American Statistical Association*, 69(347), pp. 726–729.

Dixit, U. J. & Nasiri, P., (2001). Estimation of parameters of the exponential distribution in the presence of outlier generated from uniform distribution, Metron 49, (3-4), pp. 187–198.

Epstein, B., & Sobel, M., (1954). Some theorems relevant to life testing from an exponential distribution. *The Annals of Mathematical Statistics*, pp. 373–381.

Golosnoy, V., & Liesenfeld, R., (2011). Interval shrinkage estimators. *Journal of Applied Statistics*, 38(3), pp. 465–477.

Hawkins, D. M., (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

Nasiri, P. & Ebrahimi, F., (2019), Interval Shrinkage Estimators of Scale Parameter of Exponential Distribution in the Presence of Outliers, Malaysian Journal of Mathematical Sciences, 13(1), pp. 75–85.

Nasiri, P., & Jabbari Nooghabi, M., (2009). Estimation of $P[Y < X]$ for generalized exponential distribution in presence of outlier. *Iranian Journal of Numerical Analysis and Optimization*, 2(1), pp. 69–80.

Nelson, W. B., (1982). Applied life Data Analysis. Wiley, New York.

Pandey, B. N., (1997). Testimator of the scale parameter of the exponential distribution using LINEX loss function. *Communications in statistics-theory and methods*, 26(9), pp. 2191–2202.

Roio, J., (1987). On the admissibility of $c[Xbar]+d$ with respect to the LINEX loss function. *Communications in Statistics-Theory and Methods*, 16(12), pp. 3745–3748.

Soliman, A. A., (2000). Comparison of LINEX and quadratic Bayes estimators for the Rayleigh distribution. *Communications in Statistics-theory and Methods*, 29(1), pp. 95–107.

Stein, C., (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *In Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, Vol. 1, No. 1, pp. 197–206.

Thompson, J. R., (1968). Accuracy borrowing in the estimation of the mean by shrinkage to an interval. *Journal of the American Statistical Association*, 63(323), pp. 953–963.

Varian, H. R., (1975). A Bayesian approach to real estate assessment. *Studies in Bayesian econometric and statistics in Honor of Leonard J. Savage*, pp. 195–208.

Zellner, A., (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394), pp. 446–451.

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 79–94, DOI 10.2478/stattrans-2022-0031
Received – 07.02.2022; accepted – 13.04.2022

Polish inequality statistics reconsidered: are the poor really that poor?

Adam Szulc[1]

ABSTRACT

In the present study income inequality in Poland is evaluated using corrected income data to provide more reliable estimates. According to most empirical studies based on household surveys and considering the European standards, the recent income inequality in Poland is moderate and decreased significantly after reaching its peaks during the first decade of the 21st century. These findings were challenged by Brzeziński et al. (2022), who placed Polish income inequality among the highest in Europe. Such a conclusion was possible when combining the household survey data with information on personal income tax. In the present study the above-mentioned findings are further explored using 2014 and 2015 data and employing additional corrections to the household survey incomes. Incomes of the poorest people are replaced by their predictions made on a large set of well-being correlates, using the hierarchical correlation reconstruction. Applying this method together with the corrections based on Brzeziński's et al. results reduces the 2014 and 2015 revised Gini indices, still keeping them above the values obtained with the use of the survey data only. It seems that the hierarchical correlation reconstruction offers more accurate proxies to the actual low incomes, while matching tax data provides better proxies to the top incomes.

Key words: inequality indices, household income imputation, income correlates.

1. Introduction

According to most of the empirical studies based on household surveys, recent income inequality in Poland is moderate, considering the European standards, and decreased significantly after peaks reached during the first decade of the 21st century (time series for the official Gini indices covering 1995–2015 period may be found in Brzeziński et al., 2022). However, a prevailing part of those studies ignore the problem of the data quality and representativity, although there are reasons to assume that nominally low declared incomes are frequently underestimated, especially in tails

[1] Warsaw School of Economics, Institute of Statistics and Demography, Poland. E-mail: aszulc@sgh.waw.pl. ORCID: https://orcid.org/0000-0003-2646-2468.

of the distributions. This affects also official inequality measures in Poland, which may be substantially underestimated, as demonstrated by Brzeziński et al. (2022) on the basis of combined survey and tax return data for 1995–2015 period. Further consequences of prospective underestimation of the official inequality are of political nature, as concluded by those authors: underrating by the previous governments importance of the (real) inequality and degree of the redistribution might be one of the reasons for reaching the parliament majority by Law and Justice (*Prawo i Sprawiedliwość*) party in 2015 election. Although this hypothesis is hardly testable empirically, it seems to be obvious that the social rhetoric represented by this party was widely accepted by the voters. On the other hand, according to Bussolo et al. (2021) the demand for redistribution in Poland between 1992 and 2009 was at a moderate level, as compared to several European countries included into that study. Moreover, other results presented by Bussolo et al. do not claim correlation between the demand for redistribution and the (in)equality perception. Nevertheless, calculation of more accurate inequality indices definitely may shed more light on the abovementioned issues in Poland, especially on discrepancy between the official indicators and the inequality perception. In this study some estimates obtained by Brzeziński et al. (2022) are utilised to correct incomes in the upper tails of the distributions for 2014 and 2015 years. Corrections of the household survey incomes are also performed at the bottom tails, which is an added value of the present research. Incomes of the poorest people are replaced by their predictions estimated on a large set of well-being correlates, using the so-called hierarchical correlation reconstruction (Duda, 2018, Szulc and Duda, 2018). This should yield more accurate inequality measures, as compared to the official ones and to those based solely on the top incomes corrections.

Several sources of non-random errors leading to underestimation of household survey incomes may be pronounced: i/ allocating too large portion of the revenues to production when completing the questionnaires (this applies to self-employed incomes, including farmers), ii/ incorrect tax adjustment, iii/ intentional misreporting, and iv/ seasonality of the revenues. For a comprehensive discussion of household survey measurement errors see Moore et al. (2000) and Kasprzyk (2005), while non-response issues are discussed in Lepkowski (2005). A discussion of the Polish household survey data quality may be found in Kośny (2019). Generally, two approaches to handling the data errors in research on inequality and poverty may be observed in the literature. In the first one additional datasets, usually tax registers, are utilised. Household survey data are combined with administrative records to provide more reliable income statistics at the upper tails of the distribution (Jenkins, 2017, Bartels and Metzing, 2018, Blanchet, 2018, Medeiros et al., 2018, Davern et al., 2019, Brzeziński et al., 2022). The literature on "decontamination" of low declared incomes is rather narrow. Nicoletti et al. (2011) proposed the so-called partial identification approach, taking into account the whole range of the distribution. This allows calculation of bounds for the poverty rates instead of the point estimates. Pudney and Francavilla (2006) employed a data "decontamination" procedure based on observing discrepancies between income and other well-being indicators (like consumption or household durables) ranks for Albania. This procedure, utilising non-parametric regression, is supposed to produce more reliable poverty rates. In this research the so-called hierarchical correlation reconstruction method (hereafter: HCREC) proposed by Duda (2018) is utilised. This methods yields estimates of the income distribution function, conditional on the household attributes correlated with well-being. As they are mainly nonmonetary and/or relatively stable in time, it may be assumed that they are more reliable and therefore can provide more accurate proxies to the household well-being and then to the actual incomes. Moreover, better reliability of the declared incomes in the middle range of the distribution than in the extreme ones is assumed. In this approach no additional information but survey data is required (this applies also to the methods proposed by Nicoletti et al., 2011, and Pudney and Francavilla, 2006).

The remaining part of the paper is organised as follows. In Section 2 the database is described. In Section 3 the main principles of two methods of data imputation are presented. Section 4 reports results of the empirical inequality comparisons. Section 5 concludes.

2. The data issues

The individual data employed in this research come from 2014 and 2015 household budget surveys being carried by Statistics Poland (Główny Urząd Statystyczny). It encompasses, inter alia, information on the households' disposable income and its components, expenditures, assets, durables, dwelling conditions, demographic and socio-economic attributes, and answers to subjective income questions. The samples covered more than 37,000 households and 101,000 persons per year. The reference period of observation is one month. More methodological details on Polish HBS may be found in Główny Urząd Statystyczny – Statistics Poland (2015). For a brief description of the tax data in Poland applicable to this study see Kośny (2019).

Except the disposable income numerous household variables are used in the present research in order to provide estimates of the corrected declared incomes. They may be of financial type and then continuous (remaining equivalent cash at the end of the month, shares of expenditures on the luxury goods and on the food) but most of them are nonmonetary and discrete (demographic attributes, dwelling and neighbourhood characteristics, possession of durables, main income source, subjective evaluations). For the full list of the variables employed in the imputations based on HCREC method see Duda and Szulc (2020). All calculations are performed for equivalent units, using the total household incomes and assuming equal distribution between the household members.

As mentioned in the Introduction, misestimation of the incomes affects mainly tails of the distribution. Since the disposable income is calculated as a difference between the household net revenues and spending on production, allocating too large portion of the revenues to a latter component affects mainly producers' (including farmers) households. Overestimation of the cost of production is quite frequent and leads to underestimation of disposable incomes, making them, in some cases, negative. Although negative disposable incomes constitute only about 0.9% of the whole 2015 sample, there is no reason to believe that the positive ones are free of such a bias. A meta study of the problem may be found in Hlasny et al. (2022). Errors caused by seasonality and by intentional misreporting of incomes may affect most of types of the households. It seems to be rational to suppose that the majority of well-being correlates, like household conditions or possession of durables, are much more stable in time and less likely to be intentionally misreported than the disposable incomes. Assuming moreover that the income data in the middle of the distribution are relatively reliable and the relations between income and welfare correlates are stable for the whole range of the distribution, it is possible to reduce impact of the abovementioned data errors applying imputations based on the HCREC method. However, this technique seems to be rather unproductive at the high ranges of the distribution, due to very low share of the extreme incomes which would result in a serious downward bias of the estimates. As mentioned above, it is possible to handle underestimation of highest incomes by matching survey data with tax registers. In the present research this method is embedded by replacing top 1% or top 5% incomes by estimates of the Pareto distribution obtained by Brzeziński et al. (2022) after matching the tax and the household survey data.

Underestimation of low incomes in the Polish household surveys becomes evident when they are confronted with a simple multidimensional household well-being indicator. The one employed in the present study covers equivalent income, dwelling conditions (esp. dwelling size and quality, presence of various appliances, neighbourhood), household equipment with durables and subjective evaluations of own material position. Each of those components was transformed into [0, 1] interval. Hence, at each dimension of well-being the households or people may be compared directly. For the non-binary, continuous or ordinal variables the multidimensional poverty indicator for each household is calculated as weighted mean of the following components:

$$f_i = f(y) = \frac{Y_{max} - y_i}{Y_{max} - Y_{min}}$$

where y_i stands for *i-th* well-being individual component, e.g. equivalent income, dwelling size per capita or subjective income evaluation. This concept represents a more general method referred to as fuzzy set approach to poverty measurement (for more details see Panek, 2006). In order to relax impact of the outliers, for the continuous, non-limited variables the minimum and maximum values in the above equation were replaced by percentiles of rank 0.05 and 0.95, respectively, with due censoring of y_i values beyond these limits. The highest weight, 0.4, is attached (arbitrarily) to the monetary dimension (equivalent income), the remaining three equal 0.2.

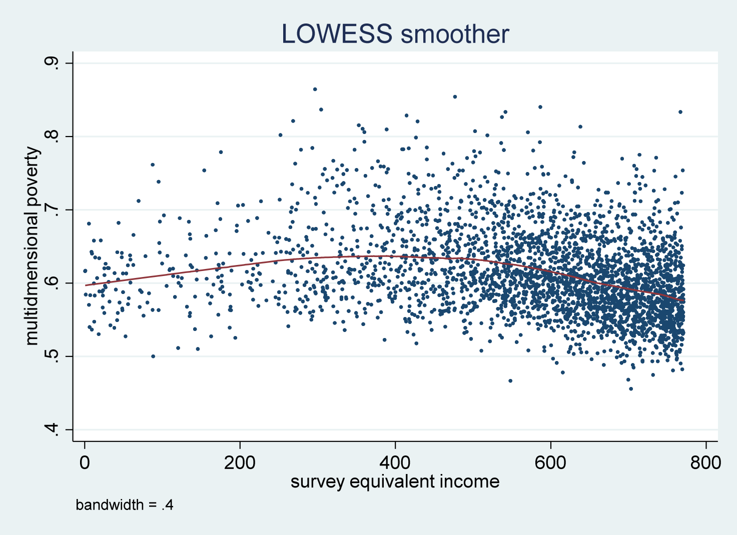


Figure 1a. Nonparametric estimation of multidimensional poverty on survey incomes (PLN per month), below the first decile.

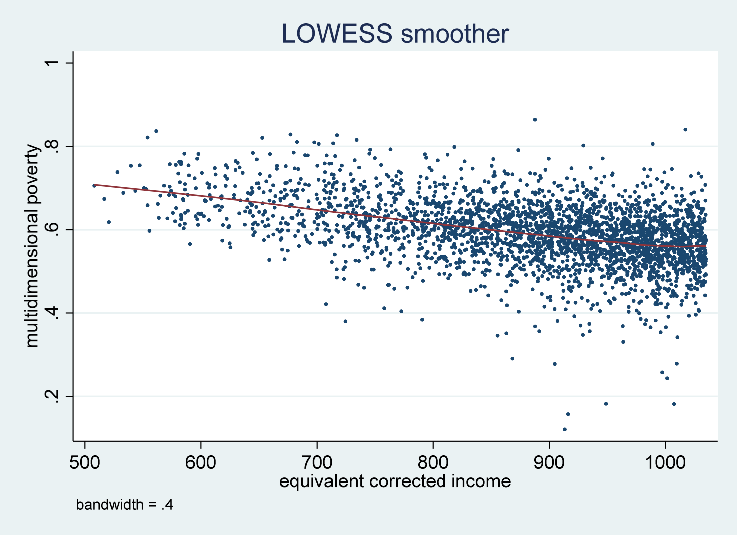


Figure 1b. Nonparametric estimation of multidimensional poverty on imputed incomes (PLN per month), below the first decile.

Figures 1a and 1b display the results of nonparametric LOWESS estimation (for the econometric details see Cleveland, 1979) of the abovementioned multidimensional poverty indicator on the equivalent income declared by the households and on income corrected by means of HCREC, respectively. In the first case, a nonsensical, positive correlation between both variables may be observed for the lowest incomes. The results obtained for the corrected incomes appear to be much more acceptable (see Figure 1b). A positive correlation between the declared income and poverty resulted also in reporting a counterproductive effect of the social transfers on the multidimensional poverty (although the respective indicator included also equivalent incomes). This nonsensical result did not appear when low incomes were replaced by their predictions estimated by means of HCREC (see Duda and Szulc, 2020 for details). Naturally, the components of the multidimensional poverty index and the set of income correlated yielding HCREC estimates do not intersect, as it would result in upper bias in the correlation measures.

3. Income imputations

3.1. Low incomes

The method of imputation applied in the present study, referred to as HCREC, allows to predict conditional probability distribution of an exogenous variable (here: household equivalent income) based on values of endogenous variables (here: income correlates). First, the marginal distribution of the predicted variable is normalized to uniform distribution on [0, 1] using empirical distribution function ($x = EDF(y) \in [0, 1\,]$). Then a density of its conditional distribution is predicted as a linear combination of orthonormal polynomials using coefficients modelled as linear combinations of the remaining variables. Once the conditional density function is estimated, selected declared incomes (here: those below first quintile) may be replaced by the respective theoretical estimates. HCREC offers two advantages, as compared to a standard regression. First, due to employing high order polynomials instead of assuming a priori a functional form, it can fit virtually all types of distribution. Second, except conditional expected values, it is possible to estimate the entire probability distribution as well as a large set of the moments. For more technical details see Duda (2018), and Duda and Szulc (2018). For an empirical application in the measurement of social transfers impact on poverty see Duda and Szulc (2020).

3.2. High incomes

This type of correction utilises findings by Brzeziński et al. (2022), who estimated Pareto I distribution function using the tax data and then replaced top 1% or 5% of equivalent incomes by the predictions. Pareto I cumulative distribution function for income x is defined as follows:

$$x' = x_m \sqrt[\alpha]{1 - F'(x)}$$

where F'(x) represents a cumulative distribution function estimated using a whole sample of the survey data.

3.3. Comparing the survey and the corrected income distributions

In the present study correction of the low incomes utilising HCREC is principally applied to bottom 20%. The impact of corrections applied to alternative low ranges of the distribution (bottom 5%, 10%, 15% and 25%) is also investigated further and the results are reported in Table 4. Figure 2 displays differences in the density functions using the declared (survey) and the bottom corrected incomes. To make the plot more readable, the highest incomes are not included in the subsample. Similar comparisons, using cumulative distribution functions, made for top 1% and 5% corrected incomes are presented in Figures 3 and 4 for the survey equivalent incomes exceeding, approximately, 95th and 90th centiles.

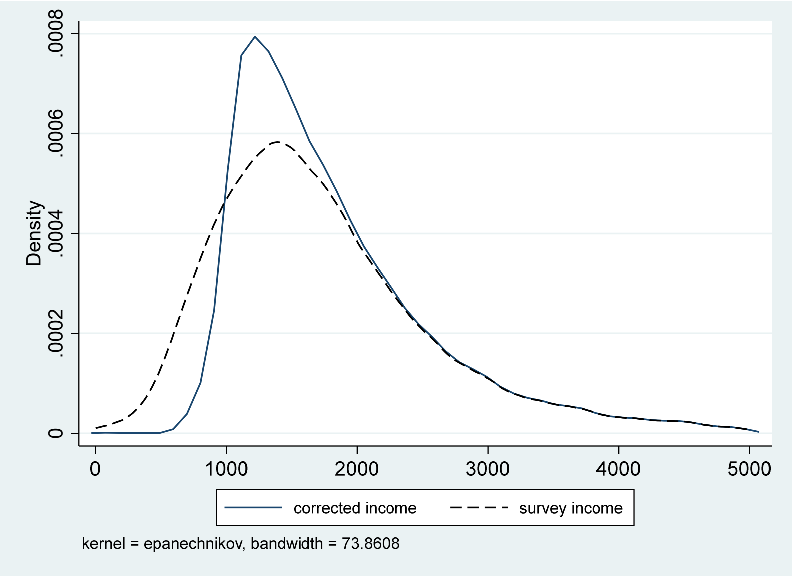


Figure 2. Kernel density functions for the survey and the corrected monthly equivalent incomes below 5000 PLN, 2015 data.

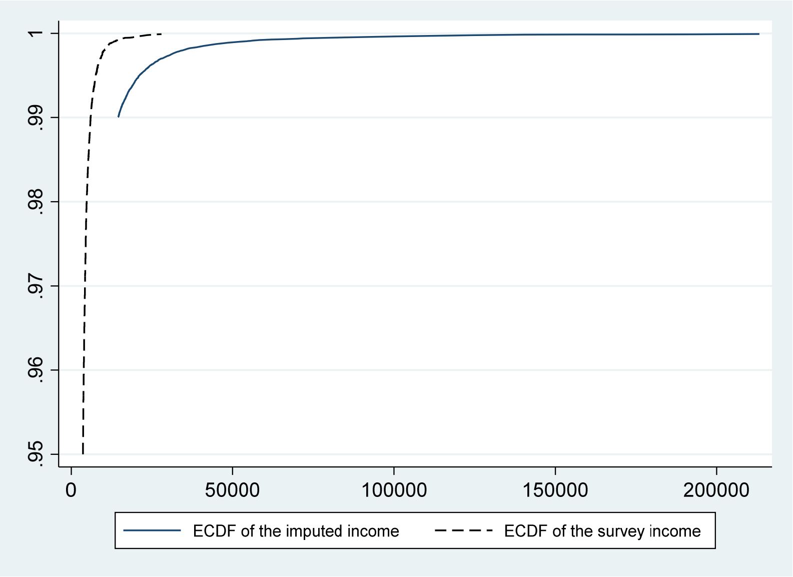


Figure 3. Empirical cumulative distribution function for the survey and the imputed (top 1%) monthly equivalent income, 2015 data.

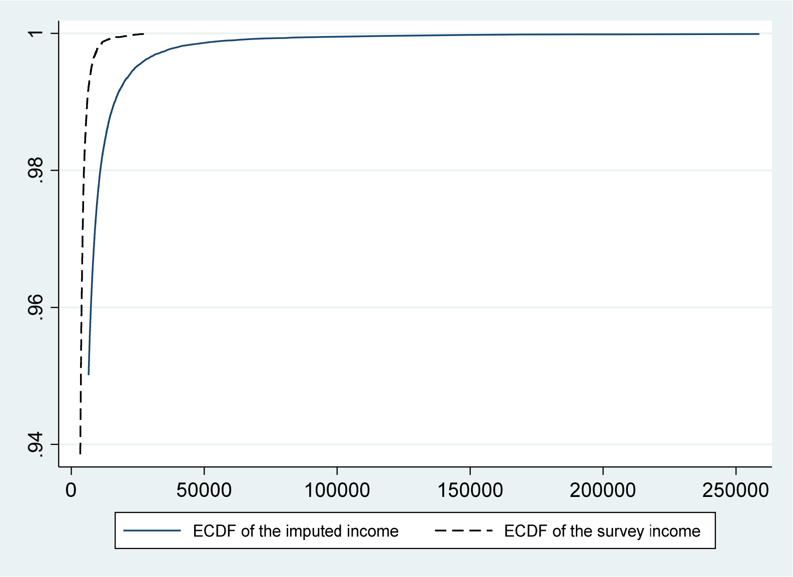


Figure 4. Empirical cumulative distribution function for the survey and the imputed (top 5%) monthly equivalent income, 2015 data.

Table 1 displays mean values of the survey and the corrected incomes in low and top ranges of the distribution for 2014 and 2015, in 2015 prices. Applied corrections raised enormously top 1% and 5% incomes, as compared to the declared ones: in 2014 by 233%/79% and by 248%/85% in 2015. Increases among the poorest 20% were less massive: by 65% in 2014 and by 63% in 2015. Nevertheless, since much larger share of the nominally poor people this growth mitigated significantly the income inequality growth caused by rising the highest incomes. The final results of changes in the inequality are reported in the succeeding section.

Table 1. Changes in mean equivalent incomes due to bottom 20%, top 1% and top 5% corrections, 2015 prices, PLN per month.

Range	Type of income			
	Raw survey	Top 1% corrected	Top 5% corrected	Bottom 20% corrected
		2014		
Bottom 20%	696	-	-	1148 *(+65%)*
Top 1%	9054	30139 *(+233%)*	36895 *(+307%)*	-
Top 5%	5299	9510 *(+79%)*	15002 *(+183%)*	-
		2015		
Bottom 20%	728	-	-	1189 *(+63%)*
Top 1%	9261	32198 *(+248%)*	37519 *(+305%)*	-
Top 5%	5379	9964 *(+85%)*	15067 *(180%)*	-

Legend: in parentheses growth rates, as compared to the survey incomes

Source: own calculation based on the household budget survey.

4. Inequality in Poland after income imputations

The final results on changes in the inequality indices are summarised in Table 2 (top 1%) and Table 3 (top 5%). Similarly to comparisons made in the previous section, inequality indices (Gini, Theil, and 90/10 and 75/25 percentile ratios) are calculated using the raw survey incomes and those corrected at low and high ranges, separately and altogether. It should be noted that corrections of the top incomes are not exactly the same as those proposed by Brzeziński et al. (2022). This is due to the various weighting systems employed in both studies. The one applied in the present study uses the survey weights (the only available), while that of Brzeziński et al. utilises the tax information for that purpose.

Table 2. Inequality indices for the survey and the corrected incomes: top 1% and bottom 20%.

	Raw survey income	Area of income correction		
		Top 1%	Bottom 20%	Both
2014	Gini			
	0.308	0.382	0.263	0.339
	90/10			
	3.89	3.89	2.85	2.85
	75/25			
	1.98	1.98	1.77	1.77
	Theil			
	0.175	0.455	0.135	0.403
2015	Gini			
	0.303	0.382	0.258	0.338
	90/10			
	3.79	3.79	2.78	2.78
	75/25			
	1.95	1.95	1.75	1.75
	Theil			
	0.171	0.470	0.132	0.416

Source: own calculation based on the household budget survey.

As might be expected, applying income correction to the bottom 20% (see Figure 2 for detailed changes in the income distribution) reduces inequality, as compared to that calculated using the survey data only. Gini indices drop by, approximately, 15%. On the other hand, modifying top incomes increases Gini indices, up to 0.38 or by 25% (when top 1% incomes are corrected) and up to 0.45 or by 47-48% (top 5% incomes corrected). Applying both corrections simultaneously places Gini indices between both extremes but still well above, by 10-34%, those calculated with the use of the survey data only. Although the bottom corrections were applied to a larger portion of the sample, the top corrections resulted in much higher increases in the extreme incomes (see Figures 2-4). In a similar way changes in the incomes revised also the Theil index, however with a much higher magnitude. Final estimates raised as much as by 220%, which confirms empirically high sensitivity of this formula to extremes values (proved theoretically by Cowell and Flachaire, 2007). It is worth mentioning that the modifications of top 1% and even top 5% incomes left inequality measures based on the percentile ratios (75/25 and 90/10) unchanged. This is because 75th and 90th centiles derived from the survey data are still below the values derived from the modified incomes. Correcting bottom incomes reduces inequality measures of that type and the amount of this reduction is greater for the 90/10 ratios. In Table 4, the impact of the area of the bottom incomes corrections on inequality is examined by comparing Gini index for the following ranges of the distribution: 10%, 15%, 20%, and 25%, calculated together with top 1% and top 5% corrections. As might be expected, the wider range of the bottom correction, the stronger the reducing effect, however the differences in the size of the changes are rather moderate: from 0.020 to 0.022.

Table 3. Inequality indices for the survey and the imputed incomes: corrections to top 5% and bottom 20%.

	Raw survey income	Area of income correction		
		Top 5%	Bottom 20%	Both
2014	Gini			
	0.308	0.454	0.263	0.413
	90/10			
	3.89	3.89	2.85	2.85
	75/25			
	1.98	1.98	1.77	1.78
	Theil			
	0.175	0.619	0.135	0.561
2015	Gini			
	0.303	0.447	0.258	0.405
	90/10			
	3.79	3.79	2.78	2.78
	75/25			
	1.95	1.95	1.75	1.75
	Theil			
	0.171	0.604	0.132	0.546

Source: own calculation based on the household budget survey.

Additionally, Theil indices are decomposed into between-group and within-group inequality. The following subgroups, created on the basis of the main source of the household income, are observed:

- blue collar employees,
- white collar employees,
- farmers,
- self-employed and those living on a property income,
- retirement pensioners,
- invalid pensioners,
- living on social transfers.

Table 4. Gini index sensitivity to the area of income correction.

Survey income	Area of income correction							
	Top 1% and bottom:				Top 5% and bottom:			
	10%	15%	20%	25%	10%	15%	20%	25%
	2014							
0.308	0.355	0.346	0.339	0.333	0.428	0.420	0.413	0.407
	2015							
0.303	0.354	0.345	0.338	0.332	0.420	0.412	0.405	0.400

Source: own calculation based on the household budget survey.

A question about within- and between-group components of the overall inequality may be, less formally, translated into the question: which gap is, on average, larger – between a rich and a poor employee (say) or between an employee and a pensioner (say)? Probably in all similar calculations based on household incomes performed for a large variety of the countries, the within-group inequality is substantially higher than the between-group one. The present results[2] definitely confirm this rule: depending on the type of income the within-group component ranges from 85% to 89% of the overall inequality. Lower values are obtained when the survey incomes and the incomes with corrections for bottom 20% only are used. When top 1% or top 5% corrections are applied, the within-group component rises to around 89%, irrespective of whether bottom 20% correction is applied or not. Rising highest incomes results in huge rises in the within-group inequality for all groups, however at certainly different paces. The largest increases may be observed for the farmers and the self-employed, which hardly can surprise. Not much smaller inequality increases in within-group inequality were experienced by the white collars households. Changing a type of the income left inequality rankings between groups nearly unaffected.

5. Conclusions and further studies

The results of the present study only partly confirm findings by Brzeziński et al. (2022) on the serious underestimation of the Polish inequality indices. Corrections of the 2014 and 2015 survey income data applied to both tails of the distribution also results in inequality growth, however not so high and not for all types of inequality measures. Possible overestimation of the income inequality by Brzeziński et al. stems from restricting the survey income corrections to the highest ranges of the distribution

[2] Since similarity of the outcomes for 2014 and 2015 only results for the latter year are reported. Detailed results are available upon request.

(top 1% and top 5%). Applying corrections also to the bottom tail of the distribution, which may be informally called making the "fake poor" non poor, leads to lower and probably more reliable estimates of the income inequality in Poland. Nevertheless, the final indices are still well above those calculated solely by means of the survey data (Gini index is higher by at least 10%) but also well below those calculated after correcting highest incomes only (Gini index is lower by at least 9%).

Assuming better reliability of the corrected incomes than of the raw survey data, there is bad and good news. Potentially bad news is a rise in the inequality in Poland, as compared to that based on the survey data. Good news is that the rise in the inequality measures is due to an upward correction of the high incomes ("making the rich more rich"), not due to a downward correction of the low incomes ("making the poor more poor"). One more good news is a reduction of the poverty incidence and depth estimates, due to income corrections applied to the bottom ranges of the distribution (see Duda and Szulc, 2020). Less sizable growth in income disparities put partly in question Brzeziński's et al. (2022) hypothesis on impact of inequality perception by the voters on the results of 2015 election in Poland. Moreover, applying income corrections to both tails of the distribution even decreased inequality measures defined as extreme percentile ratios (75/25 and 90/10). The question which measures of inequality, the latter ones or Gini indices, are better proxies to inequality perception is another issue worth further research. One more argument against the hypothesis under consideration may be pronounced referring to 2015 election campaign. The winner's (Law and Justice) rhetoric was rather pro-poor than anti-rich (with two exceptions: they announced increased taxation for banks and foreign hypermarkets, see Prawo i Sprawiedliwość, 2015). Informally speaking, the future government promised to be Santa Claus rather than Robin Hood.

Another point worth consideration is the source of income growths estimated for some rich people. As pointed out by Brzeziński et al. (2022), it followed a substantial reduction of the personal income tax progressivity. Without deciding whether this growth itself was advantageous or not, the recent changes in the tax system in Poland, referred to as Polish Deal ("Polski Ład"), started in 1st January 2022 make room for further studies in this field. The declared features of those changes are, inter alia: a minor increase in a tax burden for the richest people and an enlargement of tax exemptions for (at least) the less privileged groups. Another relevant reform introduced by the government after 2015 is a reconstruction of the system of social cash transfers. In April 2016 the family support program, referred to as "Family 500+", was launched. Its principal details may be found in Brzeziński and Najsztub (2017) and in Michoń (2021). It ensures monthly unconditional support of tax-free 500 PLN (26% of mean equivalent income in 2016) per each child in families with two or more children and means tested support of same amount for families with one child. The transfers to families with children resulted, inter alia, in reduction of Gini index for equivalent incomes between 2015 and 2017 by 9%, using survey data only. It seems, however, that Family 500+ cash transfers had no impact on underreporting incomes in the household surveys[3].

Acknowledgement

The author is grateful to Jarosław Duda for providing the estimates of imputed incomes obtained by means of the hierarchical correlation reconstruction and for a fruitful discussion. All remaining errors are solely the responsibility of the author.

References

Altman, E. I., (1968). Financial Ratios, Discriminant Analysis and the Prediction of the Corporate Bankruptcy. *The Journal of Finance*, Vol. 23, pp. 589–609.

Bartels, C., Metzing, M., (2019). An integrated approach for a top-corrected income distribution. *Journal of Economic Inequality*, Vol. 17, pp. 125–143.

Blanchet, T., Flores, I. and Morgan, M., (2018). The weight of the rich: Improving surveys using tax data. *WID.world Working Paper Series No. 2018/12*, World Inequality Lab, retrieved from https://pdfs.semanticscholar.org/71d4/c87af7224d185bdd9adee4ea22fcd1edc879.pdf

Brzeziński, M., Myck, M. and Najsztub, M., (2022). Sharing the gains of transition: Evaluating changes in income inequality and redistribution in Poland using combined survey and tax return data, forthcoming in *European Journal of Political Economy*.

Brzeziński, M., Najsztub, M., (2017). The impact of "Family 500+ programme on household incomes, poverty and inequality. *Polityka Społeczna*, Vol. 44, pp. 16–25.

Bussolo, M., Ferrer-I-Carbonell, Giolbas, A. and Torre, I., (2021). *I perceive therefore I demand: the formation of inequality perceptions and demand for redistribution*, Review of Income and Wealth, Vol. 67, pp. 835–871.

Cleveland, W. S., (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association,* Vol. 74, pp. 829–836.

Cowell, F. A., Flachaire, E., (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics*, Vol. 141, pp. 1044–1072.

[3] Preliminary, not published findings available upon request.

Davern, M. E., Meyer, B. D. and Mittag, N. K., (2019). Creating improved survey data products using linked administrative-survey data. *Journal of Survey Statistics and Methodology*, Vol. 7, pp. 440–463.

Duda, J., (2018). *Hierarchical correlation reconstruction with missing data*, arXiv preprint, arXiv:1804.06218, retrieved from: https://arxiv.org/abs/1804.06218

Duda, J., Szulc, A., (2018). *Credibility evaluation of income data with hierarchical correlation reconstruction*, arXiv: 1812.08040, retrieved from: https://arxiv.org/abs/1812.08040

Duda, J., Szulc, A., (2020). Social benefits versus monetary and multidimensional poverty in Poland: imputed income exercise. In: Tsounis, N., Vlachvei, A. (eds.) Advances in cross-section data methods in applied economic research. ICOAE 2019. *Springer Proceedings in Business and Economics*. Springer.

Główny Urząd Statystyczny – Statistics Poland, (2015). Budżety gospodarstw domowych – *Household budget survey*, Warsaw.

Hlasny, V., Ceriani, L. and Verme, P., (2022)., Bottom incomes and the measurement of poverty and inequality, forthcoming in Review of Income and Wealth.

Jenkins, S. P., (2017). Pareto models, top incomes and recent trends in UK income inequality. *Economica*, Vo. 84, pp. 261–289.

Kasprzyk, D., (2005). *Measurement error in household surveys: sources and measurement*, in: Household Sample Surveys in Developing and Transition Countries, United Nations, New York.

Kośny, M., (2019). Upper tail of the income distribution in tax records and survey data. Evidence from Poland. *Argumenta Oeconomica*, Vol. 42, pp. 55–80.

Lepkowski, J., (2005). *Non-observation error in household surveys in developing countries*, in: Household Sample Surveys in Developing and Transition Countries, United Nations, New York.

Medeiros, M., De Castro Galvão, J. and De Azevedo Nazareno, L., (2018). Correcting the underestimation of top incomes: Combining data from income tax reports and the Brazilian 2010 census. Social Indicators Research, Vol. 135, pp. 233–244.

Michoń, P., (2021). *Deservingness for "Family 500+" Benefit in Poland: Qualitative Study of Internet Debates*, Social Indicators Research, Vol. 157, pp. 203–223.

Moore, J. C., Stinson, L. L. and Welniak, E. J., (2000). Income measurement error in surveys: A review. *Journal of Official Statistics*, Vol. 16, pp. 331–361.

Nicoletti, C., Peracchi, F. and Foliano, F., (2011). Estimating income poverty in the presence of missing data and measurement error. *Journal of Business and Economic Statistics*, Vol. 29, pp. 61–72.

Prawo i Sprawiedliwość – Law and Justice, (2015). *Myśląc Polska 2015*, proceedings of the conference, retrieved from: http://pis.org.pl/dokumenty?page=2

Panek, T., (2006). *Multidimensional fuzzy relative poverty dynamic measures in Poland*. In: Lemmi, A. and G. Betti (eds.), Fuzzy Set Approach to Multidimensional Poverty Measurement, Springer.

Pudney, S., Francavilla, F., (2006). *Income mis-measurement and the estimation of poverty rates*. An analysis of income poverty in Albania. ISER Working Paper 2006–35. Colchester: University of Essex, retrieved from: https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2006-35.pdf

STATISTICS IN TRANSITION new series, September 2022
Vol. 23 No. 3, pp. 95–112, DOI 10.2478/stattrans-2022-0032
Received – 31.01.2020; accepted – 31.01.2022

New polynomial exponential distribution: properties and applications

Abdelfateh Beghriche[1], Halim Zeghdoudi[2], Vinoth Raman[3], Sarra Chouia[4]

ABSTRACT

The study describes the general concept of the XLindley distribution. Forms of density and hazard rate functions are investigated. Moreover, precise formulations for several numerical properties of distributions are derived. Extreme order statistics are established using stochastic ordering, the moment method, the maximum likelihood estimation, entropies and the limiting distribution. We demonstrate the new family's adaptability by applying it to a variety of real-world datasets.

Key words: exponential distribution, Xgamma distribution, Lindley distribution, quantile function stochastic ordering, maximum-likelihood estimation, XLindley distribution.

1. Introduction

Statistical models can be used to describe and predict real-world events. In recent years, a variety of distributions have been employed for data modelling in a variety of domains. Recent advances have centred on establishing new families that extend well-known distributions while still allowing for a great deal of flexibility in data modelling in practice. Several distributions have been proposed in the statistical literature to modify lifetime data, including the Lindley, exponential, gamma, Weibull, Zeghdoudi, and Xgamma distributions.

[1] Department of Mathematics, Faculty of Exact Sciences, University the Brothers Mentouri Constantine 1, 25019, Algeria. E-mail: b.abdelfateh1@yahoo.fr.

[2] Laps laboratory, Badji-Mokhtar University, Box 12, Annaba, 23000, Algeria. E-mail: zeghdoudihalim@yahoo.fr. ORCID: https://orcid.org/0000-0002-4759-5529.

[3] Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Kingdom of Saudi Arabia. E-mail: vrrangan@iau.edu.sa. ORCID: https://orcid.org/ 0000-0002-3815-2312.

[4] Laps laboratory, Badji-Mokhtar University, Box 12, Annaba, 23000, Algeria. E-mail: chouiasarra3@gmail.com.

In this paper, we investigate a new polynomial exponential family that includes the distributions of XLindley and Xgamma, as well as Zeghdoudi as special instances, to introduce a new family of single-parameter continuous distributions. The existing literature on modelling survival data, biological sciences, and actuarial sciences will benefit from this new family of distributions.

Assume X is a random variable with values in the range [0, ∞], and the distribution of X depends on an indeterminate parameter θ with values in the range [0, ∞]. The distribution of X can be absolutely continuous or discrete. The distribution of X is a new one-parameter polynomial exponential family and the probability density function is expressed as

$$f_{NPED}(\mathrm{x},\theta) = \frac{P(x,\theta)e^{-\theta x}}{\sum_{k=0}^{n} a_{k,\theta}\frac{k!}{\theta^{k+1}}};\ x,\theta > 0 \tag{1}$$

where $P(x,\theta) = \sum_{k=0}^{n} a_{k,\theta}x^{k}$, and $a_{k,\theta}$ depend on k and θ.

The following is the format of this research paper:

Section 2 covers the survival and hazard rate functions, moments stochastic orders, mean deviations, extreme domain of attraction, constraint force estimate parameter, the Lorenz curve, and entropies of the new polynomial exponential distribution (NPED). Sections 3 and 4 look at estimating maximum likelihood distribution parameters and inferring a random sample from the XLindley and Xgamma distributions. Finally, various real-world applications demonstrate the superior performance of the XLindley and Xgamma distributions, two special examples of the (NPED) family, as compared to the exponential, Lindley, Zeghdoudi, and exponential distributions.

2. Statistical and reliability measures of some properties of NPED distribution

We present some key statistical and reliability measures, as well as various NPED features, in this section.

2.1. Density and distribution functions

The first derivative of f_{NPED}:

$$\frac{d}{dx}f_{NPED}(\mathrm{x},\theta) = \frac{[(a_{1,\theta}-\theta a_{0,\theta})+\cdots+(na_{n,\theta}-\theta a_{n-1,\theta})x^{n-1}+a_{n,\theta}x^{n}]e^{-\theta x}}{\sum_{k=0}^{n} a_{k,\theta}\frac{k!}{\theta^{k+1}}} = 0 \tag{2}$$

gives $x_1, x_2, \ldots., x_n$ solutions.

The NPED cumulative distribution function (CDF) is derived in (3).

$$F_{NPED}(\mathrm{x}) = 1 - \frac{\sum_{k=0}^{n}\frac{a_{k,\theta}\Gamma(k+1,x\theta)}{\theta^{k+1}}}{\sum_{k=0}^{n} a_{k,\theta}\frac{k!}{\theta^{k+1}}};\ x,\theta > 0 \tag{3}$$

2.2. Survival and hazard rate functions

$$S_{NPED}(\mathrm{x}) = 1 - F_{NPED}(\mathrm{x}) = \frac{\sum_{k=0}^{n} \frac{a_{k,\theta}\Gamma(k+1,x\theta)}{\theta^{k+1}}}{\sum_{k=0}^{n} a_{k,\theta}\frac{k!}{\theta^{k+1}}} \;;\; x, \theta > 0 \tag{4}$$

$$h_{NPED}(\mathrm{x}) = \frac{f_{NPED}(x)}{1-F_{NPED}(x)} = \frac{\sum_{k=0}^{n} a_{k,\theta}x^{k}e^{-\theta x}}{\sum_{k=0}^{n}\frac{a_{k,\theta}\Gamma(k+1,x\theta)}{\theta^{k+1}}} \;;\; x, \theta > 0 \tag{5}$$

Let equation (4) and (5) be the survival and hazard rate function, respectively.

<u>Proposition 1</u>. Let $h_\theta(\mathrm{x})$be the hazard rate function of X . Then, $h_\theta(\mathrm{x})$ is increasing for:

$$\sum_{k=0}^{n}(k+1)(m-2k)a_{m-k,\theta}a_{k+1,\theta} \geq 0, m = 0, \dots\dots, 2n-1.$$

Proof. According to Glaser (1980) and from the density function (2) we have:

$$\rho(x) = -\frac{f'_{NPED}(x;\theta)}{f_{NPED}(x;\theta)} = -\frac{\sum_{k=1}^{n} k a_{k,\theta}x^{k-1}}{\sum_{k=0}^{n} a_{k,\theta}x^{k}} + \theta. \tag{6}$$

After simple computations, we obtain:

$$\rho'(x) = \frac{\sum_{m=0}^{2n}\sum_{k=0}^{m}(k+1)(m-2k)a_{m-k,\theta}a_{k+1,\theta}x^{m-1}}{(\sum_{k=0}^{n} a_{k,\theta}x^{k})^{2}} + \theta \tag{7}$$

Which implies that $h_\theta(\mathrm{x})$is increasing for:

$$\sum_{k=0}^{n}(k+1)(m-2k)a_{m-k,\theta}a_{k+1,\theta} \geq 0, m = 0, \dots\dots, 2n-1$$

2.3. Moments and related measures

The k^{th} moment about the origin of $NPED$ is:

$$E(X^{i}) = \frac{\sum_{k=0}^{n}\frac{a_{k,\theta}(k+i+1)!}{\theta^{k+i+1}}}{\sum_{k=0}^{n} a_{k,\theta}\frac{k!}{\theta^{k+1}}}; i = 1,2, \dots \tag{8}$$

Corollary 1. Let $X \sim NPED(\theta)$, the mean of X is:

$$E(X) = \frac{\sum_{k=0}^{n}\frac{a_{k,\theta}(k+1)!}{\theta^{k+2}}}{\sum_{k=0}^{n} a_{k,\theta}\frac{k!}{\theta^{k+1}}} \;. \tag{9}$$

Theorem 1. Let $X \sim NPED(\theta)$, $me = median(X)$and $\mu = E(X)$. Then, $me < \mu$.

Proof. According to the increasing of $F(X)$ for all x and θ.

$$F_{NPED}(me) = \frac{1}{2}$$

and

$$F_{NPED}(\mu) = 1 - h(\theta) \textstyle\sum_{k=0}^{n} \frac{a_{k,\theta}\Gamma\left(k+1,\theta h(\theta) \sum_{k=0}^{n} a_{k,\theta}\frac{(k+1)!}{\theta^{k+2}}\right)}{\theta^{k+1}}.$$

Note that $\frac{1}{2} < F(\mu) < 1$. It is easy to check that $F(me) < F(\mu)$. At the other end we have $me < \mu$.

2.4. Stochastic orders

Definition 1. Consider two random variables X and Y. X is said to be smaller than Y in the:

a) Stochastic order $X \prec_S Y$if $F_X(t) \geq F_Y(t), \forall t$.

b) Convex order $X \prec_{CX} Y$ Nif for all convex functions Φ and provided expectation exist, $E[\Phi(X)] \leq E[\Phi(Y)]$.

c) Hazard rate order $X \prec_{hr} Y$, if $h_X(t) \geq h_Y(t), \forall t$.

d) Likelihood ratio order $X \prec_{lr} Y$, if $\frac{f_X(t)}{f_Y(t)}$ is decreasing in *t*.

Remark 1. Likelihood ratio order⇒Hazard rate order⇒Stochastic order.

If E (X) = E (Y), then convex order⇔stochastic order.

Theorem 2. Let $X_i \sim NPED(\theta_i) i = 1,2$ be two random variables. If $\theta_1 \geq \theta_2$, then $X_1 \prec_{lr} X_2$,$X_1 \prec_{hr} X_2$,$X_1 \prec_S X_2$.

Proof. We have:

$$\frac{f_{X_1}(t)}{f_{X_2}(t)} = \frac{\sum_{k=0}^{n} a_{k,\theta}\frac{(k+1)!}{{\theta_2}^{k+2}}}{\sum_{k=0}^{n} a_{k,\theta}\frac{(k+1)!}{{\theta_1}^{k+2}}} e^{-(\theta_1-\theta_2)}. \tag{11}$$

For simplification, we use $\ln(\frac{f_{X_1}(t)}{f_{X_2}(t)})$. Now, we can find

$$\frac{d}{dt}\ln\left(\frac{f_{X_1}(t)}{f_{X_2}(t)}\right) = -(\theta_1 - \theta_2).$$

To this end, if $\theta_1 \geq \theta_2$,, we have $\frac{d}{dt}\ln\left(\frac{f_{X_1}(t)}{f_{X_2}(t)}\right)$ ≤0. This means that $X_1 \prec_{lr} X_2$. Also, according to Remark 1 the theorem is proved.

2.5. Mean deviations

These are two mean deviations: about Mean and Median, defined as: $MD_1 = \int_0^\infty |x-\mu| f(x)dx$ and $MD_2 = \int_0^\infty |x - me| f(x)dx$ respectively, where $\mu = E(X)$ and $me = Median(X)$.

The measures MD_1 and MD_2 can be computed using the following simplified formulas:

$$MD_1 = 2\mu F(\mu) - 2\int_0^{\mu} xf(x)\,dx$$

$$MD_2 = \mu - 2\int_0^{me} xf(x)\,dx$$

2.6. Extreme domain of attraction

As to the extreme value stability, the $F_{NPED}$is in the Gumbel extreme value domain of attraction, that is, there exist two sequences $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$of real numbers such that for any $x\epsilon R$, we have

$$\lim_{x\to+\infty} P\left(\frac{M_n-b_{n,\theta}}{a_{n,\theta}} \leq x\right) = \lim_{x\to+\infty} F_{NPED}\left(a_{n,\theta}x + b_{n,\theta}\right)^n = e^{(-e^{-x})} \tag{12}$$

This follows from Formula 1.2.4 in theorem 1.2.1 (Laurens de Haan, Ana Ferreira (2006)) since we have

$$\lim_{t\to+\infty} \frac{1 - F_{NPED}(t + xf(t))}{1 - F_{NPED}(t)} = \lim_{t\to+\infty} \frac{f_{NPED}(t + xf(t))}{f_{NPED}(t)}$$
$$= \lim_{t\to+\infty} \frac{\sum_{k=0}^{n} a_{k,\theta}(t+xf(t))^{k+1}e^{-\theta(t+xf(t))}}{\sum_{k=0}^{n} a_{k,\theta}t^{k+1}e^{(-\theta t)}} = e^{-x} \tag{13}$$

(Such formula is called Γ -variation). Then, F_{NPED} lies in the Gumbelextreme domain of attraction. In his case, $f(t) = \frac{1}{\theta}$.

So, for (as in the invoked theorem) $a_{n,\theta} = f\left({F^{-1}}_{NPED}\left(1 - \frac{1}{n}\right)\right) = \frac{1}{\theta}$ and $b_{n,\theta} = {F^{-1}}_{NPED}\left(1 - \frac{1}{n}\right)$, we have:

$$\lim_{x\to+\infty} F_{NPED}\left(a_{n,\theta}x + b_{n,\theta}\right)^n = e^{(-e^{-x})}$$

2.7. Estimation of the Stress-Strength Parameter and Lorenz curve

Because it evaluates the system performance, the stress-strength parameter (R) is crucial in the reliability analysis. Furthermore, R indicates the likelihood of a system failure; the system breaks when the applied stress exceeds its strength, i.e.

$R = P(X > Y)$. Here, $X \sim NPED(\theta_1)$, denotes the strength of a system subject to stress Y, and $Y \sim NPED(\theta_2)$,, X and Y are independent of each other. In our case, the stress-strength parameter R is given by:

$$R = P(X > Y) = \int_0^\infty S_X(y)f_Y(y)dy$$

$$= \frac{\int_0^\infty \sum_{k=0}^{n} \frac{a_{k,\theta}\Gamma(k+2,y\theta_1)}{{\theta_1}^{k+2}} \sum_{k=0}^{n} a_{k,\theta}y^{k+1}e^{(-\theta_2 y)}dy}{\left(\sum_{k=0}^{n} a_{k,\theta}\frac{(k+1)!}{{\theta_1}^{k+2}}\right)\left(\sum_{k=0}^{n} a_{k,\theta}\frac{(k+1)!}{{\theta_2}^{k+2}}\right)}$$

The Lorenz curve is a well-known way of describing income and wealth distributions. The graph of the ratio is the Lorenz curve for a positive random variable X. Against $F(x)$with the properties $L(p) \leq p, L(0) = 0$and $L(1) = 1$. If X represents annual income, $L(p)$ is the proportion of total income that accrues to individuals with the $100\%p$ lowest incomes.

If all individuals earn the same income then $L(p) = p$ for all p. The area between the line $L(p) = p$ and the Lorenz curve can be used to calculate income inequality or, more broadly, the variability of X. The Lorenz curve is well known for the exponential distribution and is given by:

$$L(p) = p\{p + (1 - p)\log(1 - p)\}$$

For the $NPED$ distribution in (3),

$$E(X/X \leq x)F_{NPED}(x) \sum_{k=0}^{n} a_{k,\theta} \frac{(k+2)!}{\theta^{k+3}} (\frac{1-\sum_{k=0}^{n} \frac{a_{k,\theta}\Gamma(k+2,x\theta)}{\theta^{k+2}})}{(\sum_{k=0}^{n} a_{k,\theta}\frac{(k+1)!}{\theta^{k+2}})^{2}} \tag{14}$$

2.8. Entropies

It is commonly understood that entropy and information can be used to calculate the degree of uncertainty in a probability distribution. However, many correlations have been created based on the features of entropy.

The entropy of a random variable X is a measure of the uncertainty's variation. The entropy of Rényi is defined as:

$$J(\gamma) = \frac{1}{1-\gamma}\log\{\int_{0}^{\infty} f^{\gamma}(x)\,dx\}$$

where $\gamma > 0$ and $\gamma \neq 1$. For the $NPED$ distribution in (2), note that for γ integer we have:

$$\int f_{NPED}{}^{\gamma}(x)dx = \frac{\int (\sum_{k=0}^{n} a_{k,\theta}x^{k})^{\gamma}\, e^{(-\theta\gamma x)}dx}{(\sum_{k=0}^{n} a_{k,\theta}\, \frac{k!}{\theta^{k+1}})^{\gamma}}$$

$$= \frac{\sum_{k=0}^{n} b_{k,\theta}(\gamma) \int x^{k\gamma}\, e^{(-\theta\gamma x)}dx}{(\sum_{k=0}^{n} a_{k,\theta}\, \frac{k!}{\theta^{k+1}})^{\gamma}}$$

where: $\int x^{k\gamma}\, e^{(-\theta\gamma x)}dx = -\frac{1}{(\theta\gamma)^{k\gamma+1}}\Gamma(k\gamma + 1, x\gamma\theta)$ and $b_{k,\theta}(\gamma)$ in function $a_{k,\theta}$ and γ. Now, the Rényi entropy is given by:

$$J(\gamma) = \frac{1}{1-\gamma}\log\left(\frac{\sum_{k=0}^{n} b_{k,\theta}(\gamma)\frac{(k\gamma)!\Gamma(k\gamma+1)}{(\theta\gamma)^{k\gamma+1}}}{\left(\sum_{k=0}^{n} a_{k,\theta}\frac{k!}{\theta^{k+1}}\right)}\right) \tag{15}$$

2.9. Estimation and inference

Let $X_1, \dots\dots X_n$ be a random sample of $NPED$. The ln-likelihood function $lnl(x_i;\theta)$ is given by:

$$lnl(x_i;\theta) = nlnh(\theta) + \sum_{i=1}^{n} \ln(\sum_{k=0}^{m} a_{k,\theta} {x_i}^k) - \theta \sum_{i=1}^{n} x_i \tag{16}$$

The derivative of $lnl(x_i;\theta)$ with respect to θ is:

$$\frac{lnl(x_i;\theta)}{d\theta} = \frac{n\dot{h}(\theta)}{h(\theta)} + \sum_{i=1}^{n} \frac{\dot{p}(x_i,\theta)}{p(x_i,\theta)} - \sum_{i=1}^{n} x_i$$

The Method of Moments (MoM) and Maximum Likelihood (ML) estimators of the parameter are the same after using $\boldsymbol{NPED}$ (16), and they may be found by solving the following non-linear equation:

$$\frac{\dot{h}(\theta)}{h(\theta)} + \frac{1}{n} \sum_{i=1}^{n} \frac{\dot{p}(x_i,\theta)}{p(x_i,\theta)} - \bar{x} = 0$$

where:

$$\dot{h}(\theta) = \frac{dh(\theta)}{d\theta} \ and \ \dot{p}(\theta) = \frac{dp(\theta)}{d\theta}$$

$$h(\theta)[\sum_{k=0}^{m} \frac{k!}{\theta^{k+2}} \left(a_{k,\theta}(k+1) - \dot{a_{k,\theta}}\theta\right] + \frac{1}{n} \sum_{i=1}^{n} \frac{\dot{p}(x_i,\theta)}{p(x_i,\theta)} - \bar{x} = 0 \tag{17}$$

Although this equation is difficult to answer, we can consider a specific scenario in which, $p(x_i,\theta) = (2 + \theta + x_i) \ and \ h(\theta) = \frac{\theta^2}{(1+\theta)^2}$. This case will be studied in Section 3.

3. XLindley distribution and some properties

In this section, we present the XLindley (XL) distribution, which belongs to the new polynomial exponential family of distributions.

A random variable *X* is said to possess an XL distribution if it has the following form:

$$f_{XL}(x;\theta) = \frac{\theta^2(2+\theta+x)}{(1+\theta)^2} e^{-\theta x} \qquad x, \theta > 0 \tag{18}$$

Note that the XL distribution is a member of the new polynomial exponential family where $n = 1, a_{0,\theta} = 2 + \theta, a_{1,\theta} = 1$using formula (1). Therefore, the mode of XL is given by

$$mode(X) = -\frac{\theta^2+2\theta-1}{\theta} \ for \ x, 0 < \theta < \sqrt{2} - 1 \tag{19}$$

We can find easily the CDF of the XL distribution

$$F_{XL}(x;\theta) = 1 - \left(1 + \frac{\theta x}{(1+\theta)^2}\right) e^{-\theta x} \qquad x, \theta > 0 \tag{20}$$

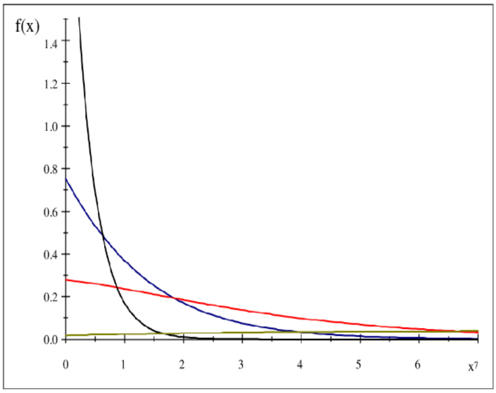


Figure.1. Plots of the density function for some parameter values of θ

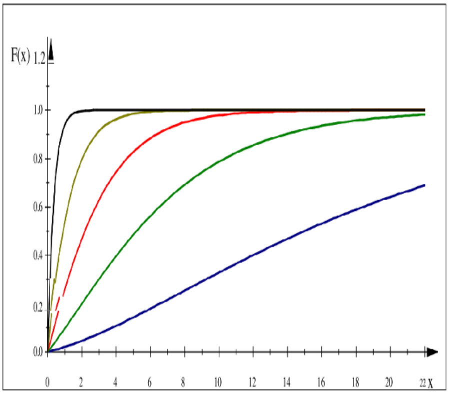


Figure.2. Plots of the cumulative function for some parameters values of θ

3.1. Survival and hazard rate function

For a continuous distribution, the survival function and the failure rate (hazard rate) functions are defined as:

$$S_{XL}(x;\theta) = 1 - F_{IXL}(x;\theta) = (1 + \frac{\theta x}{(1+\theta)^2})e^{-\theta x} \qquad x, \theta > 0 \tag{21}$$

$$h_{XL}(x;\theta) = \frac{f'_{XL}(x;\theta)}{1 - F_{XL}(x;\theta)} = \frac{\theta^2(x+\theta+2)}{(1+\theta)^2 + \theta x} \qquad x, \theta > 0 \tag{22}$$

Let equation (21) and (22) be the survival and hazard rate function, respectively.

<u>**Proposition 2.**</u> Let h_{XL} be the hazard rate function of *X*. Then, h_{XL} is increasing.

Proof. According to Glaser (1980) and from the density function (18):

$$\rho(x) = -\frac{f'_{XL}(x)}{f_{XL}(x;\theta)} = \frac{x\theta + \theta^2 - 2\theta - 1}{x + \theta + 2}$$

It follows that:

$$\rho'(x) = \frac{1}{(x + \theta + 2)^2}$$

Imply that h_{XL} is increasing.

3.2. Moments and related measures

The r^{th} moment about the origin of the XLindley distribution can be obtained as:

$$\mu_r' = E(X^r) = \int_0^\infty x^r \, f_{XL}(x)dx$$

$$= \int_0^\infty x^r \frac{\theta^2(2+\theta+x)}{(1+\theta)^2} e^{-\theta x} dx$$

$$\frac{\theta^2}{(1+\theta)^2} \int_0^\infty x^r \,(2+\theta+x)e^{-\theta x} dx$$

Finally, using gamma integral and little algebraic simplification, we get a general expression for the r^{th} factorial moment of XL distribution as:

$$\mu_r' = \frac{(\theta^2+2\theta+r+1)r!}{(1+\theta)^2\theta^r} \tag{23}$$

The first four moments can be derived by substituting $r = 1; 2; 3$ and 4 in (23), and then using the relationship between moments about origin and moments about mean, the first four moments about origin of the XL distribution may be obtained as follows:

$$\mu_1' = \frac{(\theta^2+2\theta+2)}{(1+\theta)^2\theta} = \frac{(1+\theta)^2+1}{(1+\theta)^2\theta} = \frac{1}{\theta} + \frac{1}{(1+\theta)^2\theta}$$

$$\mu_2' = \frac{2(\theta^2+2\theta+3)}{(1+\theta)^2\theta^2}$$

$$\mu_3' = \frac{6(\theta^2+2\theta+4)}{(1+\theta)^2\theta^3}$$

$$\mu_4' = \frac{24(\theta^2+2\theta+5)}{(1+\theta)^2\theta^4}$$

Let $X{\sim}XL(\theta)$, the mean, variance for X be:

$$\mu_1' = E(X) = \frac{(1+\theta)^2+1}{(1+\theta)^2\theta} \tag{24}$$

$$E(X^2) = \frac{2(\theta^2+2\theta+3)}{(1+\theta)^2\theta^2}$$

$$\mu_2 = Var(X) = \frac{(1+\theta)^4+4\theta^2+6\theta+1}{(1+\theta)^4\theta^2}$$

3.3. Estimation of parameter

3.3.1. Maximum Likelihood Estimation (MLE)

Let $X_i \sim XL(\theta)$, $i = 1, \dots\dots, n$ be n random variables, the ln-likelihood function, $lnl(x_i;\theta)$ is:

$$L(\theta) = (\tfrac{\theta^2}{(1+\theta)^2})^n \prod_{i=1}^{n}(2+\theta+x_i)\, e^{-\theta\sum_{i=1}^{n} x_i} \tag{25}$$

The logarithm of the likelihood function is:

$$lnl(x_i;\theta) = 2nlog\theta - 2nlog(\theta+1) + \sum_{i=1}^{n} log(2+\theta+x_i) - \theta\sum_{i=1}^{n} x_i$$

$$lnl(x_i;\theta) = 2n[log\theta - log(\theta+1)] + \sum_{i=1}^{n} log(2+\theta+x_i) - \theta\sum_{i=1}^{n} x_i \tag{26}$$

The derivatives of $lnl(x_i;\theta)$ with respect to θ are:

$$\frac{lnl(x_i;\theta)}{\delta\theta} = 0$$

$$\frac{\delta lnl(x_i;\theta)}{\delta\theta} = \frac{2n}{\theta} - \frac{2n}{1+\theta} + \sum_{i=1}^{n}\frac{1}{2+\theta+x_i} - \sum_{i=1}^{n} x_i$$

$$\frac{\delta lnl(x_i;\theta)}{\delta\theta} = \frac{2}{\theta} - \frac{2}{1+\theta} + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2+\theta+x_i} - \bar{X}$$

$$\frac{\delta lnl(x_i;\theta)}{\delta\theta} = \frac{2}{\theta(1+\theta)} + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2+\theta+x_i} - \bar{X} \tag{27}$$

To obtain the MLE of θ:$\hat{\theta}_{MLE}$ we can maximize equation (27) directly with respect toθ, or we can solve the non-linear equation$\frac{\delta lnl(x_i;\theta)}{\delta\theta} = 0$. Note that $\widehat{\theta}_{MLE}$ cannot be solved analytically; numerical iteration techniques, such as the Newton-Raphson algorithm, are thus adopted to solve the logarithm of the likelihood equation for which (27) is maximized.

3.3.2. Method of Moments Estimation (MME)

Let $\bar{X}$ be the sample mean, equating sample mean and population mean $E(X)$,

$$E(X) = \sum_{i=1}^{n}\frac{x_i}{n} \tag{28}$$

When we plug in the expression of $E(X)$ from equation (24) and solve the equation for θ, we get

$$\bar{X} = \frac{(1+\theta)^2+1}{(1+\theta)^2\theta} = \frac{\theta^2+2\theta+2}{\theta^3+2\theta^2+\theta}$$

We obtain equation of 3rd degree $\bar{X}\theta^3 + \theta^2(2\bar{X} - 1) + \theta(\bar{X} - 2) - 2 = 0$. We take the real part for the solution

$$\hat{\theta}_{MLE} = -\frac{1}{3\bar{X}}(2\bar{X} - 1) + \frac{\frac{2}{9\bar{X}}+\frac{1}{9\bar{X}^2}+\frac{1}{9}}{\sqrt{\sqrt{\frac{1}{27\bar{X}}+\frac{13}{36\bar{X}^2}+\frac{1}{9\bar{X}^3}+\frac{1}{27\bar{X}^4}+\frac{11}{18\bar{X}}+\frac{1}{9\bar{X}^2}+\frac{1}{27\bar{X}^3}+\frac{1}{27}}}} + \sqrt[3]{\sqrt{\frac{1}{27\bar{X}} + \frac{13}{36\bar{X}^2} + \frac{1}{9\bar{X}^3} + \frac{1}{27\bar{X}^4} + \frac{11}{18\bar{X}} + \frac{1}{9\bar{X}^2} + \frac{1}{27\bar{X}^3} + \frac{1}{27}}} \quad (29)$$

3.4. Simulation

The behaviour of the estimators for a finite sample size (n) is investigated in this subsection. A simulation study consisting of the following steps is being carried out N=10000 times for selected values of (θ, n), where $\theta = 0.05; 0.25; 1; 2; 5$ and $n = 20; 50; 100$.

– Generate U_i Uniform (0; 1), $i = 1, \dots \dots, n$.

– Generate Y_i Exponential$(\theta), i = 1, \dots \dots, n$.

– Generate Z_i Lindley$(\theta), i = 1, \dots \dots, n$.

– If $U_i \leq p(\theta)$, then set $X_i = Y_i$ otherwise, set $X_i = Z_i$, $i = 1, \dots \dots, n$

$$verage\ bias(\theta) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta).$$

And the average square error:

$$MSE(\theta) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta)^2$$

Table 1. Average bias of the estimator $\widehat{\theta}$

Bias	$\theta = 0.05$	$\theta = 0.25$	$\theta = 1$	$\theta = 2$	$\theta = 5$
$n = 20$	0.00131	0.01002	0.0456	0.2451	0.7512
$n = 50$	0.00095	0.0124	0.0106	0.1162	0.1421
$n = 100$	0.00011	0.00251	0.0122	0.0423	0.0506

Table 2. The average square error of the estimator $\widehat{\theta}$

MSE	$\theta = 0.05$	$\theta = 0.25$	$\theta = 1$	$\theta = 2$	$\theta = 5$
$n = 20$	$1,03.10^{-6}$	0.000113	0.00236	0.0654	0.6177
$n = 50$	$2, 55.10^{-7}$	0.000214	0.000162	0.01233	0.03135
$n = 100$	$1,04.10^{-8}$	$1.34.10^{-5}$	0.000216	0.00184	0.00301

Table 1 and 2 show the outcomes of the simulation. The simulation analysis yielded the following conclusions:

- for some given value of θ, the average of bias of θ and the mean square error of θ decrease as the sample size n increases,
- the mean square error (MSE) gets higher and following a similar way for larger value of θ as we mentioned before.

3.5. Application and goodness of fit

Data set 1: Survival times (in months) of 94 Sierra Leone individuals infected with Ebola virus. It is available at *https://apps.who.int/gho/data/node.ebola-sitrep*. In table 3, we compare the Lindley (LD), Zeghdoudi, exponential, XGamma, and XL distributions using data set 1.

Table 3. Comparison between LD, XG, ZD, Exp and XL distributions.

Survival time m=3.17 , s=2.095	**Obsfreq**	**LD** $\hat{\theta}=0.522$	**Xgamma** $\hat{\theta}=0.689$	**ZD** $\hat{\theta}=0.852$	**Exp** $\hat{\theta}=0.315$	**XL** $\hat{\theta}=0.467$
[0,2]	45	38. 262	37. 652	30. 339	43. 937	41. 028
[2,4]	22	28. 164	27. 197	37.27	23. 4	25. 855
[4,6]	17	15. 075	16. 342	17.743	12. 463	13. 984
[6,8]	7	7. 1187	7. 7769	6.1658	6. 6375	6. 9986
[8,10]	3	3. 1423	3. 2015	1.828	3. 5351	3. 3409
Total	94	94	94	94	94	94
χ^2	-	2. 7899	3. 2040	14.236	1. 8619	**1. 6446**

4. Exponential-gamma $(3, \boldsymbol{\theta})$ (X gamma) distribution and its applications

In this section, we give an overview on Exponential-gamma Eg $(\boldsymbol{\theta})$ (X gamma) distribution (see Subhradev (2016)), which is a member of the NPED. A random variable X is said to possess Eg$(\boldsymbol{\theta})$ distribution if it has the following form:

$$f_{EG}(x;\theta)=\frac{\theta^2}{(1+\theta)}\left(1+\frac{\theta}{2}x^2\right)e^{-\theta x} \qquad x,\theta>0 \tag{30}$$

Note that the *Eg* distribution is a member of the NPED family where:
$n=2, a_{0,\theta}=1, a_{1,\theta}=0, a_{2,\theta}=\frac{\theta}{2}$, using formula (1).

Therefore, the mode of *Eg* $(\boldsymbol{\theta})$ distribution is given by:

$$mode(X)=\frac{1+\sqrt{1-2\theta}}{\theta} \; for \;\; 0<\theta<\frac{1}{2} \tag{31}$$

We can find easily the CDF of the *Eg* $(\boldsymbol{\theta})$ distribution:

$$F_{Eg}(x;\theta)=1-\frac{(1+\theta+\theta x+\frac{\theta^2x^2}{2})}{(1+\theta)}e^{-\theta x} \qquad x,\theta>0 \tag{32}$$

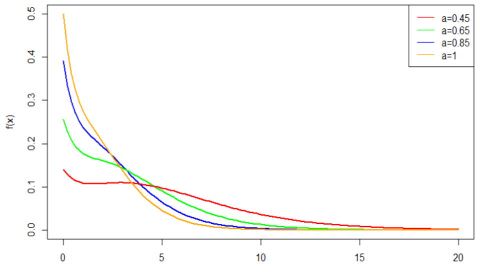


Figure 3. Plots of the density function for some parameters values of θ

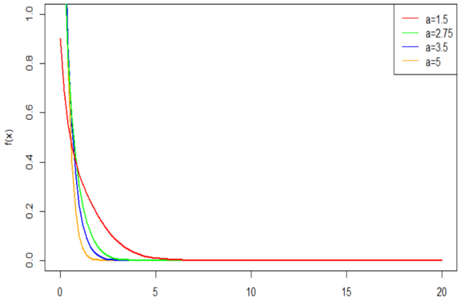


Figure4. Plots of the cumulative function for some parameters values of θ

4.1. Survival and hazard rate function

For a continuous distribution, the survival function and failure rate (hazard rate) functions are defined as:

$$S_{Eg}(x;\theta) = 1 - F_{Eg}(x;\theta) = \frac{(1+\theta+\theta x+\frac{\theta^2 x^2}{2})}{(1+\theta)} e^{-\theta x} \qquad x, \theta > 0 \qquad (33)$$

4.2. Moments and related measures

The r^{th} moment about the origin of the *Eg* ($\boldsymbol{\theta}$) distribution can be obtained as:

$$\mu_r' = E(X^r) = \frac{r!(\theta+r+a_r)}{\theta^r(1+\theta)} \qquad (34)$$

where $a_r = a_{r-1} + r$ for $r = 1,2,3, \dots ..$with $a_0 = 0$ and $a_1 = 2$. In particular,

$$\mu_1' = \frac{(\theta+3)}{\theta(\theta+1)} = Mean(X) = \mu$$

$$\mu_2' = \frac{2(\theta+6)}{\theta^2(\theta+1)}, \mu_3' = \frac{6(\theta+10)}{\theta^3(\theta+1)}, \mu_4' = \frac{24(\theta+15)}{\theta^4(\theta+1)}$$

It is to be noted that, for the exponential distribution with parameter θ, the r^{th} order moment about origin is

$$\mu_r' = \frac{r!}{\theta^r}$$

The j^{th} order central moment of the *Eg* ($\boldsymbol{\theta}$) is

$$\mu_j = E\left[(X-\mu)^j\right] = \sum_{r=0}^{j} \binom{j}{r} \mu_r' (-\mu)^{j-r}.$$ In particular,

$$\mu_2 = \frac{(\theta^2+8\theta+3)}{\theta^2(1+\theta)^2} = \text{var(X)} = \sigma^2$$

$$\mu_3 = \frac{2(\theta^3+15\theta^2+9\theta+3)}{\theta^3(1+\theta)^3}$$

$$\mu_4 = \frac{3(5\theta^4+88\theta^3+310\theta^2+288\theta+177)}{\theta^4(1+\theta)^4}$$

4.3. Estimation of parameter

Let $X_i \sim Eg(\theta)$distribution, $i = 1, \ldots\ldots, n$ be n random variables. The ln-likelihood function, $lnl(x_i; \theta)$ is:

$$L(\theta) = \prod_{i=1}^{n} f(x_i{:}\,\theta) = \prod_{i=1}^{n} \frac{\theta^2(1+\frac{\theta}{2}x_i^2)e^{-\theta x_i}}{1+\theta}$$

The logarithm of the likelihood function is:

$$\log L\,(x_i; \theta) = 2nlog\theta - nlog(1+\theta) + \textstyle\sum_{i=1}^{n}[\log\left(1+\frac{\theta}{2}x_i^2\right) - \theta x_i] \quad (35)$$

The derivatives of $\log L\,(x_i; \theta)$with respect to θ are:

$$\frac{\delta L}{\delta \theta} = \frac{2n}{\theta} - \frac{n}{1+\theta} + \sum_{i=1}^{n}\left(\frac{x_i^2}{2(1+\frac{\theta}{2}x_i^2)} - x_i\right)$$

We get the likelihood equation as a system of nonlinear equations in θ by setting the left side of the above equation to zero. The MLE of θ in this system is obtained by solving it in θ. It is simple to calculate numerically using a statistical software tool such as the nlmpackage in R programming with arbitrary initial values.

The Fisher information about θ,$I(\theta)$, is

$$I(\theta) = E\left\{-\frac{\partial^2}{\partial^2\theta^2} lnf(X,\theta)\right\} = E\left\{\frac{2}{\theta^2} - \frac{1}{(1+\theta)^2} + \frac{x^4}{4}\frac{1}{\left(1+\frac{\theta}{2}x^2\right)^2}\right\}$$

$$= \frac{2}{\theta^2} - \frac{1}{(1+\theta)^2} + E\left\{\frac{x^4}{4}\frac{1}{(1+\frac{\theta}{2}x^2)^2}\right\} \quad (36)$$

Then the asymptotic $100(1-\alpha)\%$ confidence interval for θ is given by $\hat{\theta} \pm z\alpha/_{2} \frac{I^{-1/_{2}}}{\sqrt{n}}$.

4.4. Simulation

Table 4. Average bias and MSE of the estimator $\hat{\theta}$

θ	n	**Bias**	**MSE**
1	50	-0.00086	$3.65\,e^{-05}$
	100	0.00040	$1.56\,e^{-05}$
	500	$1.32\,e^{-05}$	$8.56\,e^{-08}$
1.5	50	-0.000061	$2.64\,e^{-05}$
	100	-0.00063	$3.34\,e^{-05}$
	500	$-3.92\,e^{-06}$	$7.63\,e^{-09}$
1.85	50	0.00174	0.000153
	100	0.00090	$8.61\,e^{-05}$
	500	0.000168	$1.4097\,e^{-05}$

4.5. Data analysis and applications

Application of the Eg distribution is illustrated in two examples.

Data set 2: The data set is taken from Klein and Berger. It shows the survival data on the death times of 26 psychiatric inpatients admitted to the University of Iowa hospitals during the years 1935-1948.

Table. 5. The survival data on the death times of psychiatric inpatients.

1	1	2	22	30	28	32	11	14	36	31	33	33
37	35	25	31	22	26	24	35	34	30	35	40	39

To evaluate the data, we used three different distributions: ED, EED, and Eg distributions. Table 6 shows the estimated unknown parameters, as well as the accompanying Kolmogorov-Sminrov (*K-S*) test statistic and *LogL* values for three alternative models.

Table 6. The estimates, K-S test statistic and $log - likelihood$ for the data set 2

Model	**Estimates**	***K-S***	***LogL***
ED	$\hat{\theta} = 0.0378$	0.377	-112.321
EED	$\hat{a} = 1.797, \hat{b} = 0.052$	0.318	-109.998
Eg	$\hat{\theta} = 0.0105$	**0.3146**	**-104.611**

We present the *p-value*, corresponding Akaikes Information Criterion (AIC) (see Akaike, H. (1974) and Bayesian Information Criterion (*BIC*) in the following table 7.

Table 7. The *p-value, AIC* and *BIC* of the models on the base data set 1

Model	***p-value***	***AIC***	***BIC***
ED	0.001	224.264	225.518
EED	0.011	221.974	224.490
Eg	**0.057**	**211.171**	**212.429**

Table 6 provides the fitted distributions' parameter MLEs and log likelihood values, while table 7 shows the AIC, BIC, and p-value values. Tables 6 and 7 show that the Eg (θ) distribution is a strong rival to the other distributions chosen to suit the dataset here.

Data set 3: Chen (Gupta R. D. and Kundu D. (1999)) gave type-II censoring data of samples with complete unit failures: 0.29, 1.44, 8.38, 8.66, 10.20, 11.04, 13.44, 14.37, 17.05, 17.13, and 18.35. Table 8 shows the estimated unknown parameters, as well as the accompanying Kolmogorov-Smirnov (*K-S*) test statistic and *Log L* values for three alternative models.

Table 8. The estimates, *K-S* test statistic and log-likelihood for the data set 2

Model	**Estimates**	***K-S***	***LogL***
ED	$\hat{\theta} = 0.091$	0.3622	-40.432
EED	$\hat{a} = 1.355, \hat{b} = 0.109$	0.3183	-38.523
Eg	$\hat{\theta} = 0.237$	**0.251**	**-35.642**

We present the *p-value*, corresponding *AIC* and *BIC* for the data set in 2 in Table 9.

Table 9. The $p-values$, (AIC) and (BIC) of the models based on the data set 3

Model	***p-value***	***AIC***	***BIC***
ED	0.098	76.635	77.033
EED	0.172	78.093	78.889
Eg	**0.462**	**72.504**	**72.902**

The parameter MLEs and log-likelihood values of the fitted distributions are shown in table 8, and the values of *AIC*, *BIC*, and *p-values* are shown in Table 9. Tables 8 and 9 show that the *Eg (θ)* is a strong rival to the other distributions employed to suit the dataset here.

5. Conclusions

We have suggested a family of distributions with only one parameter in this paper. Moments, distribution function, characteristic function, failure rate, stochastic order, maximum likelihood approach, and method of moments were among the properties studied.

The Lindley and Zeghdoudi distributions lack the flexibility needed to examine and model many forms of data related to lifetime data and survival analysis. The $NPED$ distribution, on the other hand, is adaptable, straightforward, and simple to use. The novel distribution was used to evaluate two real data sets and was compared to existing distributions (Lindley, exponential, Zeghdoudi, Exponential Exponential and Xgamma). The comparison's findings support the $NPED$ distribution's quality adjustment. We anticipate that our new distribution family will entice many additional life data, reliability analysis, and actuarial science applications.

We can employ a more general distribution with two parameters in future experiments, and

$$f_{New}(x,\theta) = h(\theta)p(x,\theta)cos\theta\exp(-\theta x)$$

where $h(\theta)$ is real-valued functions on $[0,\infty]$, and where $p(x,\theta) = b(\theta) + x^k$.

Acknowledgement

The authors acknowledge the Editor-in-chief, *Prof. Dr. Patryk Barszcz* and the referees of this journal for constant encouragement to finalize the paper.

References

Akaike, H., (1974). A new look at the statistical model identification, IEEE Transactions on Automatic Control, AC-19, pp. 716–723.

Asgharzadeh, A., Hassan, S. and Bakouch, L. E., (2013). Pareto Poisson-Lindley distribution and its application. *Journal of Applied Statistics*, Vol. 40, No. 8, pp. 1–18.

Bashir, S., Rasul, M., (2015). Some properties of the weighted Lindley distribution. *International Journal of Economic and Business Review*, 3 (8), pp. 11–17.

Beghriche, A. F., Zeghdoudi, H., (2019). A Size Biased Gamma Lindley Distribution. *Thailand Statistician*, Vol.17, No 2, pp. 179–189.

Fisher, R. A., (1934). The effects of methods of ascertainment upon the estimation of frequencies. *Ann. Eugenics*, , 6, pp. 13–25.

Ghitany, M. E, Atieh, B and Nadarajah, S., (2008b). Lindley distribution and its applications. Mathematics and Computers in Simulation, 78, pp. 493–506.

Glaser, R. E., (1980). Bathtub and related failure rate characterizations. *J. Amer.Statist. Assoc.*, 75, pp. 667–672.

Gupta R. D., Kundu D., (1999). Generalized Exponential distribution. *Australian and New Zealand Journal of Statistics*, Vol. 41, No. 2, pp.173–188.

Laurens de Haan, Ferreira A., (2006). Extreme value theory: An introduction. *Springer*.

Lawless, J. F., (2003). Statistical models and methods for lifetime data. Wiley, New Y [16] Sen, Subhradev; Maiti, Sudhansu S.; and Chandra, N., (2016). The Xgamma Distribution: Statistical Properties and Application. *Journal of Modern Applied Statistical Methods*, Vol. 15, Iss. 1, Article 38.

Lindley, D. V., (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Society*, series B, 20, pp. 102–107.

Patil, G. P., Rao, C. R., (1977). Weighted distribution survey of their applications, In P. R. Krishnaiah, (E ds.), Applications of statistics, pp. 383–405, Amsterdam, North Holland.

Patil, G. P., Rao, C. R., (1978). Weighted distributions and size biased sampling with applications to wild life populations and human families. *Biometrics*, 34, pp. 179–189.

Sen Subhradev, Maiti Sudhansu S. and Chandra, N., (2016). The xgamma Distribution: Statistical Properties and Application. *Journal of Modern Applied Statistical Methods*, Vol. 15, Iss. 1, Article 38. DOI: 10.22237/jmasm/1462077420.

Zeghdoudi, H., Bouchahed, L., (2018). A new and unified approach in generalizing the Lindley's distribution with applications. *Statistics in Transition new series,* Vol. 19, No. 1, pp. 61–74,

Zeghdoudi, H., Messadia, M., (2018). Zeghdoudi Distribution and its Applications. *International Journal of Computing Science and Mathematics***,** Vol. 9, No.1, pp. 58–65.

Zeghdoudi, H.,. Nedjar, S., (2016c). A pseudo Lindley distribution and its application. *Afr. Stat.*, Vol. 11, No. 1, pp. 923–932.

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 113–126, DOI 10.2478/stattrans-2022-0033
Received – 07.12.2020; accepted – 14.06.2022

An improved ridge type estimator for logistic regression

Nagarajah Varathan[1]

ABSTRACT

In this paper, an improved ridge type estimator is introduced to overcome the effect of multicollinearity in logistic regression. The proposed estimator is called a modified almost unbiased ridge logistic estimator. It is obtained by combining the ridge estimator and the almost unbiased ridge estimator. In order to asses the superiority of the proposed estimator over the existing estimators, theoretical comparisons based on the mean square error and the scalar mean square error criterion are presented. A Monte Carlo simulation study is carried out to compare the performance of the proposed estimator with the existing ones. Finally, a real data example is provided to support the findings.

Key words: Logistic Regression, Multicollinearity, ridge estimator, Modified almost unbiased ridge logistic estimator, Mean square error.

1. Introduction

The general form of logistic regression model is

$$y_i = \pi_i + \varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where ε_i are independent with mean zero and variance $\pi_i(1-\pi_i)$ and π_i is the expected value of the response y_i when the ith value of dependent variable follows the Bernoulli distribution with parameter π_i as

$$\pi_i = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}, \tag{2}$$

where x_i is the ith row of X, which is an $n \times p$ data matrix with p explanatory variables and β is a $p \times 1$ vector of coefficients. The Maximum likelihood method is the most common estimation technique to estimate the parameter vector β, and the maximum likelihood estimator (MLE) of β based on the sample model (1) can be obtained as follows:

$$\hat{\beta}_{MLE} = (X'\hat{W}X)^{-1}X'\hat{W}z, \tag{3}$$

where z is the column vector with ith element equals $logit(\hat{\pi}_i) + \frac{y_i-\hat{\pi}_i}{\hat{\pi}_i(1-\hat{\pi}_i)}$ and $\hat{W} = diag[\hat{\pi}_i(1-\hat{\pi}_i)]$, which is asymptotically unbiased estimate of β. The asymptotic covariance matrix of $\hat{\beta}_{MLE}$ is

$$Cov(\hat{\beta}_{MLE}) = (X'\hat{W}X)^{-1}. \tag{4}$$

[1] Department of Mathematics and Statistics, University of Jaffna, Sri Lanka. E-mail: varathan@univ.jfn.ac.lk. ORCID: https://orcid.org/0000-0003-2014-8144.

The asymptotic MSE and SMSE of $\hat{\beta}_{MLE}$ are

$$\begin{aligned} MSE[\hat{\beta}_{MLE}] &= Cov[\hat{\beta}_{MLE}] + B[\hat{\beta}_{MLE}]B'[\hat{\beta}_{MLE}] \\ &= \{X'\hat{W}X\}^{-1} \\ &= C^{-1} \end{aligned} \tag{5}$$

and

$$SMSE[\hat{\beta}_{MLE}] = tr[C^{-1}] \tag{6}$$

Since C is a positive definite matrix there exists an orthogonal matrix P such that $P'CP = \Delta = diag(\lambda_1, \lambda_2, ..., \lambda_p)$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$ are the ordered eigen values of C. Then,

$$SMSE[\hat{\beta}_{MLE}] = \sum_{j=1}^{p} \frac{1}{\lambda_j}.$$

The logistic regression model becomes unstable when there exists strong dependence among explanatory variables. This situation is referred to as multicollinearity. When the multicollinearity presents among explanatory variables, the estimation of the model parameters becomes inaccurate because of the need to invert near-singular information matrix $X'\hat{W}X$. As a result, the estimates have large variances and large confidence intervals, which produces inefficient estimates.

To overcome the problem of multicollinearity in the logistic regression, many estimators have been proposed in the literature alternative to the MLE. The most popular estimator to deal with this problem is called the Logistic Ridge Estimator (LRE), and was first proposed by Schaefer et al. (1984). Later, Aguilera et al. (2006) introduced the Principal Component Logistic Estimator (PCLE), Nja et al. (2013) proposed the Modified Logistic Ridge Regression Estimator (MLRE), Mansson et al. (2012) introduced the Liu-Estimator in logistic regression, Inan and Erdogan (2013) proposed Liu-type estimator, Xinfeng (2015) proposed the Almost Unbiased Liu Logistic Estimator (AULLE), Wu and Asar (2016) proposed the Almost Unbiased Ridge Logistic Estimator (AURLE), Varathan and Wijekoon (2019) proposed the Modified Almost Unbiased Liu Logistic Estimator (MAULLE), Jadhav (2020) proposed the Linearized ridge logistic estimator (LRLE), and the Modified ridge type logistic estimator was proposed by Lukman et al. (2020).

In this research a new estimator is proposed by combining AURLE and LRE. Further, we compare the performance of the proposed MAURLE estimator with the existing MLE, LRE and AURLE estimators in the mean square error sense.

The organization of the paper is as follows. The construction of the proposed estimator is given in Section 2. In Section 3, the asymptotic properties of the estimators are given. In Section 4, the conditions for superiority of the proposed MAURLE estimator over the existing MLE, LRE, and AURLE estimators are derived with respect to mean square error (MSE) criterion. In Section 5, the conditions for superiority of the proposed MAURLE estimator over the existing MLE, LRE, and AURLE estimators are derived with respect to scalar mean square error (SMSE) criterion. The detail Monte Carlo simulation study is given to investigate the performance of the proposed estimator with some existing estimators in Section 6. A real data application is discussed in Section 7. Finally, some conclusive remarks are given in Section 8.

2. Construction of the proposed estimator

The new estimator is constructed by considering the Logistic ridge estimator (LRE) (Schaefer et al., 1984) and the Almost Unbiased Ridge Logistic Estimator (AURLE) (Wu and Asar, 2016). Note that LRE and AURLE are defined as

$$\begin{aligned}\hat{\beta}_{LRE} &= (X'\hat{W}X+kI)^{-1}X'\hat{W}X\hat{\beta}_{MLE} \\ &= (C+kI)^{-1}C\hat{\beta}_{MLE} \\ &= Z_k\hat{\beta}_{MLE}\end{aligned} \tag{7}$$

where $Z_k = (C+kI)^{-1}C$ and k is the ridge parameter, $k \geq 0$.

$$\hat{\beta}_{AURLE} = W_k\hat{\beta}_{MLE} \tag{8}$$

where $W_k = I - k^2(C+kI)^{-2}$, $k \geq 0$.

By substituting $\hat{\beta}_{LRE}$ in place of $\hat{\beta}_{MLE}$ in the estimator AURLE in (2.2), we propose a new estimator which is named the Modified almost unbiased ridge logistic estimator (MAURLE) and defined as

$$\begin{aligned}\hat{\beta}_{MAURLE} &= W_k\hat{\beta}_{LRE} \\ &= W_kZ_k\hat{\beta}_{MLE} \\ &= F_k\hat{\beta}_{MLE}\end{aligned} \tag{9}$$

where

$$\begin{aligned}F_k &= W_kZ_k \\ &= [I-k^2(C+kI)^{-2}][(C+kI)^{-1}C]\end{aligned} \tag{10}$$

$$\begin{aligned}SMSE(\hat{\beta}_{MAURLE}) &= \sum_{j=1}^{p}\frac{\lambda_j^3(\lambda_j+2k)^2}{(\lambda_j+k)^6} \\ &\quad +\sum_{j=1}^{p}\frac{(k^3+3k^2\lambda j+k\lambda j^2)^2}{(\lambda_j+k)^6}\alpha_j^2\end{aligned} \tag{11}$$

3. Asymptotic properties of the proposed estimator

The mean vector, dispersion matrix and the bias vector of $\hat{\beta}_{MAURLE}$ are

$$\begin{aligned} E[\hat{\beta}_{MAURLE}] &= E[F_k\hat{\beta}_{MLE}] \\ &= F_k\beta, \end{aligned} \tag{12}$$

$$\begin{aligned} D(\hat{\beta}_{MAURLE}) &= Cov(\hat{\beta}_{MAURLE}) \\ &= Cov(F_k\hat{\beta}_{MLE}) \\ &= F_kC^{-1}F_k', \end{aligned} \tag{13}$$

and

$$\begin{aligned} Bias(\hat{\beta}_{MAURLE}) &= E[\hat{\beta}_{MAURLE}]-\beta \\ &= [F_k-I]\beta \\ &= \delta_{MAURLE} \end{aligned} \tag{14}$$

Consequently, the mean square error and scalar mean square error can be obtained as,

$$\begin{aligned} MSE(\hat{\beta}_{MAURLE}) &= D(\hat{\beta}_{MAURLE})+Bias(\hat{\beta}_{MAURLE})Bias(\hat{\beta}_{MAURLE})' \\ &= F_kC^{-1}F_k'+(F_k-I)\beta\beta'(F_k-I)' \end{aligned} \tag{15}$$

where

$$\begin{aligned} F_k &= W_kZ_k \\ &= [I-k^2(C+kI)^{-2}][(C+kI)^{-1}C] \\ &= (C+kI)^{-2}C(C+2kI)(C+kI)^{-1}C \\ &> 0 \text{ is a positive definite matrix.} \end{aligned} \tag{16}$$

and

$$\begin{aligned} SMSE(\hat{\beta}_{MAURLE}) &= \sum_{j=1}^{p}\frac{\lambda_j^3(\lambda_j+2k)^2}{(\lambda_j+k)^6} \\ &\quad +\sum_{j=1}^{p}\frac{(k^3+3k^2\lambda j+k\lambda j^2)^2}{(\lambda_j+k)^6}\alpha_j^2 \end{aligned} \tag{17}$$

where α_j is the jth element of $P'\beta$, P is an orthogonal matrix such that $P'CP=\Delta=diag(\lambda_1,\lambda_2,...,\lambda_p)$, where $\lambda_1\geq\lambda_2\geq...\geq\lambda_p>0$ are the ordered eigen values of C.

4. Mean square error comparison of estimators

To check the performance of the proposed MAURLE estimator with the existing MLE, LRE, and AURLE estimators, we compare the corresponding mean square errors of the estimators.

Note that, following Schaefer et al. (1984) and Wu and Asar (2016), the mean square errors of LRE and AURLE are respectively given by
$MSE[\hat{\beta}_{LRE}] = Z_k C^{-1} Z_k' + \delta_{LRE}\delta_{LRE}'$; where $\delta_{LRE} = (Z_k - I)\beta$
$MSE[\hat{\beta}_{AURLE}] = W_k C^{-1} W_k' + \delta_{AURLE}\delta_{AURLE}'$; where $\delta_{AURLE} = (W_k - I)\beta$

(I). **MAURLE Versus MLE**

$$\begin{aligned} MSE(\hat{\beta}_{MLE}) - MSE(\hat{\beta}_{MAURLE}) &= \{D(\hat{\beta}_{MLE}) + B(\hat{\beta}_{MLE})B'(\hat{\beta}_{MLE})\} \\ &\quad -\{D(\hat{\beta}_{MAURLE}) + B(\hat{\beta}_{MAURLE})B'(\hat{\beta}_{MAURLE})\} \\ &= C^{-1} - \{F_k C^{-1} F_k' + \delta_{MAURLE}\delta_{MAURLE}'\} \\ &= U_1 - V_1 \qquad (18) \end{aligned}$$

where $U_1 = C^{-1}$ and $V_1 = F_k C^{-1} F_k' + \delta_{MAURLE}\delta_{MAURLE}'$. One can obviously say that $F_k C^{-1} F_k'$ and U_1 are positive definite matrices and $\delta_{MAURLE}\delta_{MAURLE}'$ is non-negative definite matrix. Further by Lemma 1 (see Appendix), it is clear that V_1 is a positive definite matrix. By Lemma 2 (see Appendix), if $\lambda_{\max}(V_1 U_1^{-1}) < 1$, then $U_1 - V_1$ is a positive definite matrix, where $\lambda_{\max}(V_1 U_1^{-1})$ is the largest eigen value of $V_1 U_1^{-1}$. Based on the above arguments, the following theorem can stated.
Theorem 1: The MAURLE estimator is superior to MLE if and only if $\lambda_{\max}(V_1 U_1^{-1}) < 1$.

(II). **MAURLE Versus LRE**

$$\begin{aligned} MSE(\hat{\beta}_{LRE}) - MSE(\hat{\beta}_{MAURLE}) &= \{D(\hat{\beta}_{LRE}) + B(\hat{\beta}_{LRE})B'(\hat{\beta}_{LRE})\} \\ &\quad -\{D(\hat{\beta}_{MAURLE}) + B(\hat{\beta}_{MAURLE})B'(\hat{\beta}_{MAURLE})\} \\ &= \{Z_k C^{-1} Z_k' + \delta_{LRE}\delta_{LRE}'\} \\ &\quad -\{F_k C^{-1} F_k' + \delta_{MAURLE}\delta_{MAURLE}'\} \\ &= U_2 - V_2 \qquad (19) \end{aligned}$$

where $U_2 = Z_k C^{-1} Z_k' + \delta_{LRE}\delta_{LRE}'$ and $V_2 = F_k C^{-1} F_k' + \delta_{MAURLE}\delta_{MAURLE}'$. One can easily say that $F_k C^{-1} F_k'$ and $Z_k C^{-1} Z_k'$ are positive definite matrices and $\delta_{LRE}\delta_{LRE}'$ and $\delta_{MAURLE}\delta_{MAURLE}'$ are non-negative definite matrices. Further by Lemma 1, it is clear that U_2 and V_2 are positive definite matrices. By Lemma 2, if $\lambda_{\max}(V_2 U_2^{-1}) < 1$, then $U_2 - V_2$ is a positive definite matrix, where $\lambda_{\max}(V_2 U_2^{-1})$ is the largest eigen value of $V_2 U_2^{-1}$. Based on the above results, the following theorem can be stated.
Theorem 2: The MAURLE estimator is superior to LRE if and only if $\lambda_{\max}(V_2 U_2^{-1}) < 1$.

(III). **MAURLE Versus AURLE**

$$\begin{aligned} MSE(\hat{\beta}_{AURLE}) - MSE(\hat{\beta}_{MAURLE}) &= \{D(\hat{\beta}_{AURLE}) + B(\hat{\beta}_{AURLE})B'(\hat{\beta}_{AURLE})\} \\ &\quad -\{D(\hat{\beta}_{MAURLE}) + B(\hat{\beta}_{MAURLE})B'(\hat{\beta}_{MAURLE})\} \\ &= \{W_kC^{-1}W_k' + \delta_{AURLE}\delta_{AURLE}'\} \\ &\quad -\{F_kC^{-1}F_k' + \delta_{MAURLE}\delta_{MAURLE}'\} \\ &= U_3 - V_3 \qquad (20) \end{aligned}$$

where $U_3 = W_kC^{-1}W_k' + \delta_{AURLE}\delta_{AURLE}'$ and $V_3 = F_kC^{-1}F_k' + \delta_{MAURLE}\delta_{MAURLE}'$. One can easily say that $F_kC^{-1}F_k'$ and $W_kC^{-1}W_k'$ are positive definite matrices and $\delta_{AURLE}\delta_{AURLE}'$ and $\delta_{MAURLE}\delta_{MAURLE}'$ are non-negative definite matrices. Further by Lemma 1, it is clear that U_3 and V_3 are positive definite matrices. By Lemma 2, if $\lambda_{\max}(V_3U_3^{-1}) < 1$, then $U_3 - V_3$ is a positive definite matrix, where $\lambda_{\max}(V_3U_3^{-1})$ is the largest eigen value of $V_3U_3^{-1}$. Based on the above results, the foolowing theorem can be stated.

Theorem 3: The MAURLE estimator is superior to AURLE if and only if $\lambda_{\max}(V_3U_3^{-1}) < 1$.

5. Scalar mean square error comparison

In this section, we compare the scalar mean square error of the proposed MAURLE estimator with the existing MLE, LRE, and AURLE estimators. According to Schaefer et al. (1984) and Wu and Asar (2016), the mean square errors of LRE and AURLE are respectively given by:

$SMSE[\hat{\beta}_{LRE}] = \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j+k)^2} + \sum_{j=1}^{p} \frac{k^2\alpha_j^2}{(\lambda_j+k)^2}$

$SMSE[\hat{\beta}_{AURLE}] = \sum_{j=1}^{p} \frac{\lambda_j(\lambda_j+2k)^2}{(\lambda_j+k)^4} + \sum_{j=1}^{p} \frac{k^4\alpha_j^2}{(\lambda_j+k)^4}$

(I). **MAURLE Versus MLE**

$$\begin{aligned} SMSE(\hat{\beta}_{MLE}) - SMSE(\hat{\beta}_{MAURLE}) & \\ &= \sum_{j=1}^{p}\frac{1}{\lambda_j} - [\sum_{j=1}^{p}\frac{\lambda_j^3(\lambda_j+2k)^2}{(\lambda_j+k)^6} \\ &\quad + \sum_{j=1}^{p}\frac{(k^3+3k^2\lambda j+k\lambda j^2)^2}{(\lambda_j+k)^6}\alpha_j^2] \\ &= \sum_{j=1}^{p}\frac{(\lambda_j+k)^6-(\lambda j^4\lambda_j+2k)^2}{\lambda_j(\lambda_j+k)^6} \\ &\quad - \sum_{j=1}^{p}\frac{(k^3+3k^2\lambda j+k\lambda j^2)^2\alpha_j^2}{(\lambda_j+k)^6} \\ &= \Delta_1^* \qquad (21) \end{aligned}$$

Based on the above comparison, it can be noted that MAURLE is superior to MLE in the SMSE sense if and only if $\Delta_1^* > 0$.

(II). **MAURLE Versus LRE**

$$\begin{aligned}
SMSE(\hat{\beta}_{LRE}) - SMSE(\hat{\beta}_{MAURLE}) &= \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j+k)^2} + \sum_{j=1}^{p} \frac{k^2\alpha_j^2}{(\lambda_j+k)^2} - \sum_{j=1}^{p} \frac{\lambda_j^3(\lambda_j+2k)^2}{(\lambda_j+k)^6} \\
&\quad - \sum_{j=1}^{p} \frac{(k^3+3k^2\lambda j+k\lambda j^2)^2}{(\lambda_j+k)^6}\alpha_j^2 \\
&= \sum_{j=1}^{p} \frac{\lambda_j(\lambda_j+k)^4 - \lambda_j^3(\lambda_j+2k)^2}{(\lambda_j+k)^6} \\
&\quad + \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j+k)^6} k^2(\lambda_j+k)^4 \\
&\quad - \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j+k)^6} (k^2+3k^2\lambda j+k\lambda j^2)^2 \\
&= \Delta_2^* \qquad (22)
\end{aligned}$$

From the above comparison, it can be concluded that MAURLE is superior to LRE in the SMSE sense if and only if $\Delta_2^* > 0$.

(III). **MAURLE Versus AURLE**

$$\begin{aligned}
SMSE(\hat{\beta}_{AURLE}) - SMSE(\hat{\beta}_{MAURLE}) &= \sum_{j=1}^{p} \frac{\lambda_j(\lambda_j+2k)^2}{(\lambda_j+k)^4} + \sum_{j=1}^{p} \frac{k^4\alpha_j^2}{(\lambda_j+k)^4} \\
&\quad - \sum_{j=1}^{p} \frac{\lambda_j^3(\lambda_j+2k)^2}{(\lambda_j+k)^6} \\
&\quad - \sum_{j=1}^{p} \frac{(k^3+3k^2\lambda j+k\lambda j^2)^2}{(\lambda_j+k)^6}\alpha_j^2 \\
&= \sum_{j=1}^{p} \frac{\lambda_j(\lambda_j+2k)^2(\lambda_j+k)^2 - \lambda_j^3(\lambda_j+2k)^2}{(\lambda_j+k)^6} \\
&\quad + \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j+k)^6} k^2(\lambda_j+k)^4 \\
&\quad - \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j+k)^6} (k^2+3k^2\lambda j+k\lambda j^2)^2 \\
&= \Delta_3^* \qquad (23)
\end{aligned}$$

Based on the above comparison, it can be said that MAURLE is superior to AURLE in the SMSE sense if and only if $\Delta_3^* > 0$.

6. Monte Carlo Simulation study

In this simulation study we compare the performance of proposed MAURLE estimator with the existing MLE, LRE, and AURLE estimators in the scalar mean square error criteria. The sample sizes *n*= 20, 50, and 100 are considered. Following McDonald and Galarneau (1975) and Kibria (2003), we generate the explanatory variables as follows:

$$x_{ij} = (1-\rho^2)^{1/2} z_{ij} + \rho z_{i,p+1}, i = 1,2,...,n, \quad j = 1,2,...,p$$

where z_{ij} are pseudo- random numbers from standardized normal distribution and ρ^2 represents the correlation between any two explanatory variables. For the multicollinearity, different levels of ρ, such as ρ= 0.9, 0.95, 0.99 and 0.999 are used. Further, for the biasing parameter k, we consider some selected values in the range $0 < k < 1$. The simulation is repeated 1000 times by generating new pseudo- random numbers, and the simulated SMSE values of the estimators are obtained using the following equation.

$$SM\hat{S}E(\hat{\beta}) \quad = \quad \frac{1}{1000} \sum_{r=1}^{1000} (\hat{\beta}_r - \beta)'(\hat{\beta}_r - \beta)$$

where $\hat{\beta}_r$ is any estimator considered in the r^{th} simulation. The simulated scalar mean square errors of estimators are reported for different values of d, ρ, and n in Tables 1 - 3.

Table 1: The estimated MSE values for different k when $n = 20$

		$\rho = 0.9$	$\rho = 0.95$	$\rho = 0.99$	$\rho = 0.999$
$k = 0.1$	MLE	66.8246	135.8973	698.0638	7023.9774
	LRE	33.6541	45.7819	64.8760	37.0857
	AURLE	48.3103	74.0211	137.5215	108.2149
	MAURLE	**30.1161**	**37.8386**	**40.0568**	**10.4364**
$k = 0.2$	MLE	66.8246	135.8973	698.0638	7023.9774
	LRE	25.0801	30.8425	33.9479	15.0765
	AURLE	39.5509	54.7585	77.9639	40.7710
	MAURLE	**21.1814**	**23.5606**	**18.3130**	**5.8744**
$k = 0.3$	MLE	66.8246	135.8973	698.0638	7023.9774
	LRE	20.4920	23.7132	22.6380	9.7370
	AURLE	33.8980	44.0296	53.3815	23.4383
	MAURLE	**16.7268**	**17.3486**	**11.6077**	**5.0841**
$k = 0.4$	MLE	66.8246	135.8973	698.0638	7023.9774
	LRE	17.6061	19.5145	16.9926	7.6181
	AURLE	29.8427	37.0186	40.0643	16.3491
	MAURLE	**14.1271**	**13.9827**	**8.7388**	**4.8320**
$k = 0.5$	MLE	66.8246	135.8973	698.06386	7023.9774
	LRE	15.6389	16.7747	13.7330	6.5873
	AURLE	26.7593	32.0282	31.8030	12.7009
	MAURLE	**12.4977**	**11.9727**	**7.3495**	**4.7941**
$k = 0.6$	MLE	66.8246	135.8973	698.0638	7023.9774
	LRE	14.2318	14.8758	11.6889	6.0421
	AURLE	24.3250	28.2792	26.2346	10.5413
	MAURLE	**11.4422**	**10.7165**	**6.6648**	**4.8842**
$k = 0.7$	MLE	66.8246	135.8973	698.0638	7023.9774
	LRE	13.1936	13.5077	10.3412	5.7533
	AURLE	22.3515	25.3560	22.2638	9.1379
	MAURLE	**10.7527**	**9.9205**	**6.3619**	**5.0612**
$k = 0.8$	MLE	66.8246	135.8973	698.06386	7023.9774
	LRE	12.4114	12.4961	9.4255	5.6153
	AURLE	20.7202	23.0144	19.3148	8.1649
	MAURLE	**10.3085**	**9.4233**	**6.2837**	**5.2989**
$k = 0.9$	MLE	66.8246	135.8973	698.0638	7023.9774
	LRE	11.8137	11.7350	8.7934	**5.5725**
	AURLE	19.3512	21.0998	17.0568	7.4588
	MAURLE	**10.0342**	**9.1285**	**6.3450**	5.5784

Table 2: The estimated MSE values for different k when $n = 50$

		$\rho = 0.9$	$\rho = 0.95$	$\rho = 0.99$	$\rho = 0.999$
$k = 0.1$	MLE	12.1526	22.3349	102.87942	1008.1701
	LRE	11.1597	18.8554	51.6995	59.7663
	AURLE	12.0889	21.8856	81.6874	165.7703
	MAURLE	**11.1069**	**18.5386**	**44.1989**	**19.2800**
$k = 0.2$	MLE	12.1526	22.3349	102.8794	1008.1701
	LRE	10.3536	16.4178	34.0311	22.8387
	AURLE	11.9366	21.0209	63.2961	68.7981
	MAURLE	**10.2011**	**15.6829**	**25.0449**	**5.3163**
$k = 0.3$	MLE	12.1526	22.3349	102.8794	1008.1701
	LRE	9.6877	14.5909	24.9083	12.7929
	AURLE	11.7342	20.0544	50.7573	38.4925
	MAURLE	**9.4312**	**13.5255**	**16.2785**	**3.1866**
$k = 0.4$	MLE	12.1526	22.3349	102.8794	1008.1701
	LRE	9.1311	13.1668	19.4025	8.6338
	AURLE	11.5041	19.0916	41.8789	25.0999
	MAURLE	**8.7818**	**11.8640**	**11.5592**	**2.6042**
$k = 0.5$	MLE	12.1526	22.3349	102.8794	1008.1701
	LRE	8.6620	12.0274	15.7713	6.5190
	AURLE	11.2596	18.1716	35.3344	17.9971
	MAURLE	**8.2368**	**10.5637**	**8.7654**	**2.3981**
$k = 0.6$	MLE	12.1526	22.3348	102.8794	1008.1701
	LRE	8.2642	11.0987	13.2338	5.3059
	AURLE	11.0095	17.3081	30.3489	13.7716
	MAURLE	**7.7814**	**9.5338**	**7.0081**	**2.3256**
$k = 0.7$	MLE	12.1526	22.3348	102.8794	1008.1701
	LRE	7.9257	10.3315	11.3873	4.5556
	AURLE	10.7591	16.5045	26.4491	11.0495
	MAURLE	**7.4028**	**8.7115**	**5.8587**	**2.3167**
$k = 0.8$	MLE	12.1526	22.3348	102.8794	1008.1701
	LRE	7.6369	9.6912	10.0031	4.0692
	AURLE	10.5121	15.7596	23.3318	9.1899
	MAURLE	**7.0903**	**8.0521**	**5.0896**	**2.3459**
$k = 0.9$	MLE	12.1526	22.3348	102.8794	1008.1701
	LRE	7.3902	9.1526	8.9421	3.7456
	AURLE	10.2707	15.0700	20.7951	7.8607
	MAURLE	**6.8348**	**7.5224**	**4.5706**	**2.4017**

Table 3: The estimated MSE values for different k when $n = 100$

		$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.99$
$k = 0.1$	MLE	5.0169	8.9865	40.3501	392.0983
	LRE	4.8749	8.4967	30.7126	72.6411
	AURLE	5.0144	8.9688	38.7927	170.6357
	MAURLE	**4.8725**	**8.4807**	**29.6393**	**37.2474**
$k = 0.2$	MLE	5.0169	8.9865	40.3501	392.0983
	LRE	4.7454	8.0613	24.5283	31.8073
	AURLE	5.0072	8.9218	35.9068	88.1286
	MAURLE	**4.7370**	**8.0075**	**22.1917**	**10.0957**
$k = 0.3$	MLE	5.0169	8.9865	40.3501	392.0983
	LRE	4.6275	7.6724	20.2308	18.2460
	AURLE	4.9959	8.8527	32.8470	53.9915
	MAURLE	**4.6107**	**7.5696**	**17.0628**	**4.4179**
$k = 0.4$	MLE	5.0169	8.9865	40.3501	392.0983
	LRE	4.5201	7.3234	17.0880	12.0657
	AURLE	4.9811	8.7669	29.9663	36.6658
	MAURLE	**4.4936**	**7.1672**	**13.4521**	**2.6290**
$k = 0.5$	MLE	5.0169	8.9865	40.3501	392.0983
	LRE	4.4226	7.0095	14.7042	8.7289
	AURLE	4.9633	8.6686	27.3633	26.6724
	MAURLE	**4.3859**	**6.7990**	**10.8469**	**1.9207**
$k = 0.6$	MLE	5.0169	8.9865	40.3501	392.0983
	LRE	4.3342	6.7263	12.8461	6.7242
	AURLE	4.9427	8.5610	25.0478	20.3845
	MAURLE	**4.2874**	**6.4634**	**8.9250**	**1.5979**
$k = 0.7$	MLE	5.0169	8.9865	40.3501	392.0983
	LRE	4.2543	6.4704	11.3661	5.4283
	AURLE	4.9200	8.4465	22.9995	16.1707
	MAURLE	**4.1979**	**6.1586**	**7.4803**	**1.4386**
$k = 0.8$	MLE	5.0169	8.9865	40.3501	392.0983
	LRE	4.1821	6.2388	10.1670	4.5452
	AURLE	4.8952	8.3273	21.1888	13.2082
	MAURLE	**4.1172**	**5.8824**	**6.3771**	**1.3585**
$k = 0.9$	MLE	5.0169	8.9865	40.3501	392.0983
	LRE	4.1173	6.0288	9.1817	3.9194
	AURLE	4.8688	8.2047	19.5857	11.0458
	MAURLE	**4.0451**	**5.6329**	**5.5238**	**1.3212**

From the results of Tables 1 - 3 it can be observed that, the proposed MAURLE estimator outperforms the MLE, LRE, and AURLE estimators in the scalar mean square error sense for almost all the values of biasing parameter k in the range $0 < k < 1$ and for all sample sizes $n = 20$, 50, and 100, except the case of k=0.9, ρ=0.999, and n=20. The LRE gives the second best performance compared to MLE, and AURLE for all the values of k, ρ, and n considered in this study. It is further noted that, comparatively MLE gives the worst performance by giving large values of SMSE.

7. A real data application

To illustrate the performance of the proposed MAURLE estimator with the existing MLE, LRE and AURLE estimators, in this research, we consider a real data application, which is obtained from the Statistics Sweden website (http://www.scb.se/). This example was used in Mansson et al. (2012), Asar and Genç (2016), and Wu and Asar (2016) to illustrate results of their papers. The data describes the information of 100 municipalities of Sweden. The following variables are considered in this study.
$x1$: Population,
$x2$: Number unemployed people,
$x3$: Number of newly constructed buildings,
$x4$: Number of bankrupt firms,
y: Net population change and is defined as

$$y = \begin{cases} 1 & ; \quad \text{if there is an increase in the population;} \\ 0 & ; \quad o/w. \end{cases}$$

Note that the Variance Inflation Factor (VIF) values for the above data are 488.17, 344.26, 44.99, and 50.71. VIF measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. According to the literature, multicollinearity is high if VIF > 10. Hence, a clear high multicollinearity exists in this data set. Further, the condition number, which is used as a measure of the degree of multicollinearity, is obtained as 188. This indicates the sign of severe multicollinearity in this data set.
The SMSE values of MLE, LRE, AURLE, and MAURLE for some selected values of biasing parameter k in the range $0 < k < 1$ are given in the Table 4. Results reveal that the proposed MAURLE estimator outperforms the MLE, LRE, and AURLE estimators in the SMSE sense, with respect to all values of k in the range $0 < k < 1$, and LRE performs well compared to MLE and AURLE.

Table 4: The SMSE values (in 10^{-4}) of estimators for the real world application data

	MLE	LRE	AURLE	MAURLE
$k = 0.1$	7.596226	7.595180	7.759226	**7.595180**
$k = 0.2$	7.596226	7.594137	7.759225	**7.541370**
$k = 0.3$	7.596226	7.593097	7.759225	**7.593096**
$k = 0.4$	7.596226	7.592059	7.759225	**7.592058**
$k = 0.5$	7.596226	7.591024	7.759224	**7.591022**
$k = 0.6$	7.596226	7.589991	7.759223	**7.589988**
$k = 0.7$	7.596226	7.588961	7.759222	**7.588957**
$k = 0.8$	7.596226	7.587934	7.759221	**7.587929**
$k = 0.9$	7.596226	7.586909	7.759220	**7.586903**

8. Concluding Remarks

In this paper, an improved estimator called Modified almost unbiased ridge logistic estimator (MAURLE) is proposed for logistic regression model when the multicollinearity problem exists. The superiority conditions for the proposed estimator with the existing MLE, LRE, and AURLE estimators are derived with respect to MSE and SMSE criterions. Further, from the real data application and the Monte Carlo simulation study we notice that the proposed estimator performs well compared to MLE, LRE, and AURLE when the multicollinearity among the explanatory variables is high.

References

Aguilera, A. M., Escabias, M., Valderrama, M. J., (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis,* 50, pp. 1905–1924.

Asar, Y., Genç, A., (2016). New Shrinkage Parameters for the Liu-type Logistic Estimators. *Communications in Statistics- Simulation and Computation,* 45(3), pp. 1094–1103.

Farebrother, R. W., (1976). Further results on the mean square error of ridge regression. *J. R. Stat. Soc. Ser B.*, 38, pp. 248–250.

Inan, D., Erdogan, B. E., (2013). Liu-Type logistic estimator. *Communications in Statistics-Simulation and Computation*, 42, pp. 1578–1586.

Jadhav, N. H., (2020). On linearized ridge logistic estimator in the presence of multicollinearity. *Comput Stat.*, 35, pp. 667–687.

Kibria, B. M. G., (2003). Performance of some new ridge regression estimators.*Commun. Statist. Theor. Meth.*, 32, pp. 419–435.

Lukman, A.F., Emmanuel, A., Clement, O.A. et al., (2020). A Modified Ridge-Type Logistic Estimator. *Iran J Sci Technol Trans Sci.*, 44, pp. 437–443.

Mansson, G., Kibria, B. M. G., Shukur, G., (2012). On Liu estimators for the logit regression model. *The Royal Institute of Techonology, Centre of Excellence for Science and Innovation Studies (CESIS)*, Sweden, Paper No. 259.

McDonald, G. C., and Galarneau, D. I., (1975). A Monte Carlo evaluation of some ridge type estimators. *Journal of the American Statistical Association*, 70, pp. 407–416.

Nja, M. E., Ogoke, U. P., Nduka, E. C., (2013). The logistic regression model with a modified weight function. *Journal of Statistical and Econometric Method,* Vol. 2, No. 4, pp. 161–171.

Rao, C. R.,and Toutenburg, H.,(1995). *Linear Models:Least Squares and Alternatives, Second Edition* .Springer-Verlag New York, Inc.

Rao, C. R., Toutenburg, H., Shalabh and Heumann, C., (2008). *Linear Models and Generalizations*. Springer. Berlin.

Schaefer, R. L., Roi, L. D., Wolfe, R. A., (1984). A ridge logistic estimator. *Commun. Statist. Theor. Meth*, 13, pp. 99–113.

Trenkler, G. and Toutenburg, H., (1990). Mean Square Error Matrix Comparisons between Biased Estimators-An Overview of Recent Results. *Statistical Papers*, 31, pp. 165–179, http://dx.doi.org/10.1007/BF02924687.

Varathan, N., Wijekoon, P., (2019). Modified almost unbiased Liu estimator in logistic regression. *Communications in Statistics - Simulation and Computation,* DOI: 10.1080/03610918.2019.1626888.

Wu, J., Asar, Y., (2016). On almost unbiased ridge logistic estimator for the logistic regression model. *Hacettepe Journal of Mathematics and Statistics*, 45(3), pp. 989–998, DOI: 10.15672/HJMS.20156911030.

Xinfeng, C., (2015). On the almost unbiased ridge and Liu estimator in the logistic regression model. *International Conference on Social Science, Education Management and Sports Education*. Atlantis Press, pp. 1663–1665.

Appendix

Lemma 1: (Rao and Toutenburg, 1995) Let $A : n \times n$ and $B : n \times n$ such that A is positive definite and B is non-negative definite. Then $(A+B)$ is positive definite.

Lemma 2: (Rao et al., 2008) Let the two $n \times n$ matrices M be positive definite, N be non-negative definite, then $M-N$ is positive definite if and only if $\lambda_{\max}(NM^{-1}) < 1$.

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 127–146, DOI 10.2478/stattrans-2022-0034
Received – 12.08.2021; accepted – 01.06.2022

Impact of restrictions on the COVID-19 pandemic situation in Poland

Sergiusz Herman[1]

ABSTRACT

The COVID-19 pandemic has had a substantial impact on public health all over the world. In order to prevent the spread of the virus, the majority of countries introduced restrictions which entailed considerable economic and social costs. The main goal of the article is to study how the lockdown introduced in Poland affected the spread of the pandemic in the country. The study used synthetic control method to this end. The analysis was carried on the basis of data from the Local Data Bank and a government website on the state of the epidemic in Poland.

The results indicated that the lockdown significantly curbed the spread of the COVID-19 pandemic in Poland. Restrictions led to the substantial drop in infections – by 9500 cases – in three weeks. The results seem to stay the same despite the change of assumptions in the study. Such conclusion can be drawn from the performance of the placebo-in-space and placebo-in-time analyses.

Key words: COVID-19, coronavirus, lockdown, synthetic control method, treatment effect.

1. Introduction

The Coronavirus disease 2019 (COVID-19) is an infectious respiratory disease caused by a SARS-CoV-2 virus. The first known case was identified in Wuhan, Central China, in November 2019. The virus has rapidly spread around the world. As a result, the World Health Organization declared the outbreak of a pandemic on 11 March 2020. Until 31 June 2021, more than 182 million cases were identified around the world and 3.9 million people died from the coronavirus.

Due to the rapid spread of the virus, countries around the world introduced restrictions to stop the spread of the coronavirus. China was the first country to do so. At the beginning of March 2020, Italy was the first European country to impose

[1] Department of Econometrics, Poznań University of Economics and Business, Poland. E-mail: sergiusz.herman@ue.poznan.pl. ORCID: https://orcid.org/0000-0002-2753-1982.

a national lockdown. By the end of the month, similar actions were taken by the majority of European countries, which introduced various restrictions for their citizens.

In the literature, there are studies on COVID-19 restrictions. Some authors focus on negative impacts of a lockdown and evaluate its social and economic costs (Bonaccorsi et al., 2020; Palomino, Rodríguez and Sebastian, 2020; Coccia, 2021; Ke and Hsiao, 2021, Wu et al., 2021, Zhang et al., 2022). More applicable for this article are studies that concern a positive impact of introduced restrictions on the pace of spread of the COVID-19 pandemic in the world. For instance, the studies that demonstrated that making masks mandatory in public spaces considerably reduces the spread of the virus (Mitze et al., 2020; Zhang et al., 2020, Bo et al., 2021, Chernozhukov et al., 2021). Other authors try to study the impact of lockdown on the outbreak of the pandemic. Studies on the topic mostly concern China (Lai et al., 2020; Ruan et al., 2020; Tian, Luo, et al., 2021; Tian, Tan, et al., 2021) and US (Bayat et al., 2020, Courtemanche et al., 2020; Siedner et al., 2020; Abouk and Heydari, 2021; Li et al., 2021). There are still not many similar studies concerning Europe. Among others, there were studies on the impact of school openings in Italy (Alfano, Ercolano and Cicatiello, 2021), lockdown in Madrid or London (Goodman-Bacon and Marcus, 2020) or a lack of restrictions on the health of citizens (Cho, 2020; Born, Dietrich and Müller, 2021). Also, there are international studies (Alfano and Ercolano, 2020; Fountoulakis et al., 2020; Piovani et al., 2021). Mendez-Brito et al. (2021) presented more research connected to the discussed subject.

In their research, the authors mentioned above used diversified methods, such as: synthetic control method (Bayat et al., 2020; Cho, 2020; Mitze et al., 2020; Alfano, Ercolano and Cicatiello, 2021; Born, Dietrich and Müller, 2021; Tian, Luo, et al., 2021; Tian, Tan, et al., 2021), linear regression (Fountoulakis et al., 2020; Siedner et al., 2020; Zhang et al., 2020), event studies (Courtemanche et al., 2020; Abouk and Heydari, 2021; Li et al., 2021), structural equation model (Chernozhukov et al., 2021), SEIR model (Lai et al., 2020), difference-in-differences analyses (Goodman-Bacon and Marcus, 2020; Abouk and Heydari, 2021), panel analysis (Alfano and Ercolano, 2020; Piovani et al., 2021), generalized linear mixed model (Bo et al., 2021).

The main goal of the research is to study the impact of the lockdown in Poland on the spread of the coronavirus pandemic. The spread of this pandemic shown completely different behaviour in different regions (voivodships). In the research, therefore, a data-driven, non-parametric way to look at things is required. For this reason the synthetic control method was used in the study. It allowed the author to determine how the pandemic would have had spread in the region similar to Warmińsko-Mazurskie voivodship if it had not been for the lockdown. There is not much research on the impact of restrictions on the spread of the pandemic in European countries. It is crucial to conduct such analyses due to the risk of another wave of COVID-19. The results might be imperative for the government to determine effective actions against the spread of the pandemic. To the best knowledge of the author, there is no other analysis in the Polish literature that uses the synthetic control method.

The article has the following structure. In the second part, the author presented statistics on the spread of the coronavirus pandemic in Poland and restrictions implemented subsequently. The third part includes a description of the methodology of the study – the synthetic control method. Next parts present a research sample and results of the empirical analysis. The article ends with the summary.

2. The spread of COVID-19 in Poland and subsequent restrictions

First cases of the coronavirus in Europe were identified at the end of January 2020. First, they were recorded in France, Germany and Italy. In Poland, a month later, on 4 March 2020, patient zero was documented in Lubuskie voivodship. Next cases on the subsequent days as well as a rapid spread of the disease in Western Europe led to first restrictions in Poland. On 10 March 2020, mass gatherings were cancelled. Two days later, schools and cultural institutions were closed. On 24 March, the movement of the population and gatherings (up to 2 people) were restricted. A week later, next restrictions were imposed; among others: hotels, restaurants and hair salons were closed, and religious gatherings were limited. On 16 April, similarly to other European countries, it became mandatory to cover nose and mouth in public spaces. Owing to the rapid introduction of restrictions as well as self-discipline of the society, a daily number of cases was stagnant in the studied period (Figure 1).

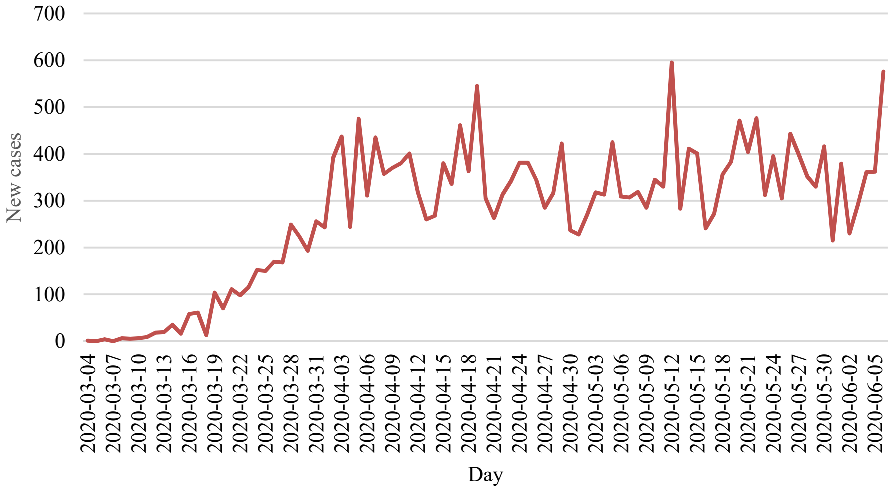


Figure 1. Daily new coronavirus cases in Poland for the period 4.03.2020 – 6.06.2020.
Source: author's work on the basis of Ritchie et al. (2021).

As a result, as of 20 April, restrictions were gradually lifted. It was divided in the 7-14-day intervals and lasted until 6 June 2020. During summertime, daily numbers of new cases were growing (reaching the maximum value of 903) but remained stable. Restrictions were implemented locally (in poviats).

The situation got considerably worse in autumn. This is well-illustrated in Figure 2. This is when the second wave of pandemic hit Poland. Due to a dramatic growth in new cases, all restrictions lifted during summertime were reintroduced. The worst situation was in November. In the subsequent months, the situation was gradually improving. Therefore, first restrictions were lifted on 18 January – schools were opened for younger kids, shopping malls, museums and art galleries were opened on 1 February, and hotels, swimming pools, cinemas and theatres were opened on 12 February.

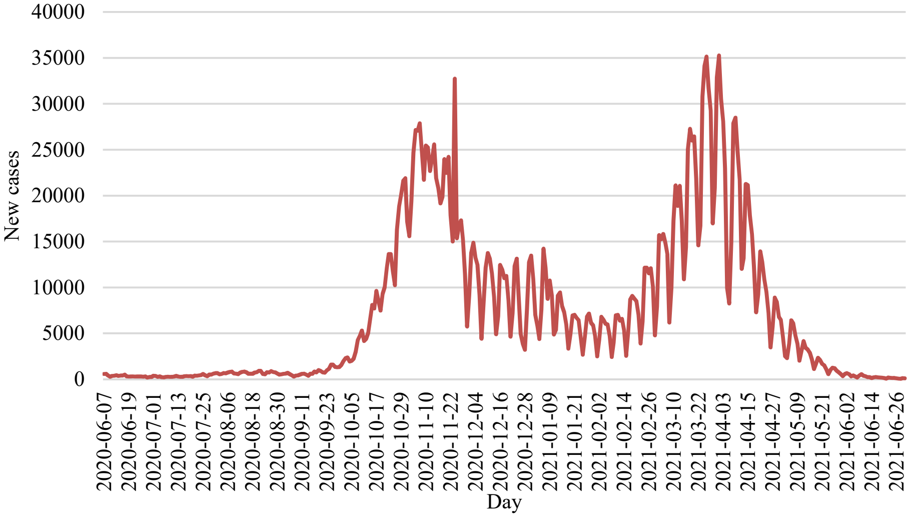


Figure 2. Daily new coronavirus cases in Poland between 7.06.2020 – 30.06.2021

Source: author's work on the basis of Ritchie et al. (2021).

It was not long before another lockdown was introduced. The third, most serious, wave of pandemic hit Poland. From the beginning of March, the daily number of new cases was rapidly growing reaching the maximum of 35000 cases a day. The hospitals were overwhelmed with COVID-19 patients. To reduce the strain on healthcare system, temporary hospitals were opened around Poland. The pandemic was spreading at different pace around the country. Therefore, first, the government introduced lockdown only in some regions. First, it was introduced in Warmińsko-Mazurskie (27.02), then Pomorskie (13.03), and Mazowieckie and Lubuskie (15.03). Figure 2 illustrates that the daily number of new coronavirus cases was growing in the studied period. Therefore, the lockdown was introduced in the entire country on 20 March. After the peak of new cases at the turn of March and April, the situation began to improve. From 20 April, the introduced restrictions were being lifted.

3. Research methodology

The synthetic control method, proposed by Abadie and Gardeazabal (Abadie and Gardeazabal, 2003) and further developed by Abadie, Diamond and Hainueller (Abadie, Diamond and Hainmueller, 2010, 2011, 2015; Abadie, 2021) was used in the study. This method is used in comparative case study research. In this kind of research, authors compare the outcome of one or multiple units affected by the treatment or intervention with the outcome of one or multiple units not affected by them. In the synthetic control method, the assumption is that only one unit was affected by the intervention. The goal of the research is to identify the impact of the treatment on the outcome of the research subject.

The assumption is that gathered data included $J+1$ units ($j = 1,2, \dots, J + 1$), where the first unit ($j = 1$) is a treated unit, whereas remaining units $j = 2, \dots. J + 1$ belong to the donor pool and constitute a set of potential comparative units not affected by the treatment. The data were gathered from T periods, where first T_0 are periods prior to the treatment (periods $1,2, \dots, T_0$) . The Y_{jt} outcome of interest can be observed for each j unit and t period. For each j unit, there is k set of $X_{1j,} \dots, X_{kj}$ predictors of the outcome that may include outcomes before the Y_{jt} treatment. Vectors with $(kx1)\boldsymbol{X_1}, \dots. \boldsymbol{X_{J+1}}$ dimensions include components for $j = 1, \dots. J + 1$ units. The $\boldsymbol{X}_0 = \left[X_2 \dots X_{J+1}\right]$ matrix with (kxJ) dimensions includes predictors of J untreated units. The outcome of interest for $t > T_0$ and for the studied ($j = 1$) unit may be defined with the equation:

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N \tag{1}$$

where Y_{1t}^I and Y_{1t}^N are denoted as the outcome for the unit affected by the intervention or in the absence of the intervention respectively. The (1) equation allows for the fact that the impact of the political intervention may vary over time. The intervention might not have an immediate effect – it can be accumulated over time. The Y_{1t}^I values are known. The purpose of the synthetic control method is to estimate the outcome for the studied unit in the absence of the Y_{1t}^N intervention. The method is based on the assumption that the linear combination of units not affected by the intervention will better illustrate how the unit reacts to the intervention. In order to construe the synthetic control unit, the author defined the $(Jx1)$ weights vector where weights are denoted as $W = (w_2, \dots, w_{J+1})'$. When the W weights vector is known, the Y_{1t}^N and τ_{1t} estimators are respectively:

$$\hat{Y}_{1t}^N = \sum_{j=1}^{J+1} w_j Y_{jt} \tag{2}$$

$$\hat{\tau}_{1t} = Y_{1t}^I - \hat{Y}_{1t}^N \tag{3}$$

The weights meet the assumptions: $w_j \geq 0 \;\; j = 2,, \ldots J$ and $w_2 + \cdots + w_{J+1} = 1$. The main challenge is to estimate the $w_2, \ldots, w_{J+1}$ weights. Abadie and Gardeazabal (2003) as well as Abadie *et al.* (2010) suggest choosing weights in such a way that characteristics of the synthetic control unit were best illustrating characteristics of the unit affected by the intervention $(j = 1)$. It means that, taking into account non-negative $v_1, \ldots \, v_k$ values, they suggest choosing $W^* = \left(w_2^*, \ldots, w_{J+1}^*\right)$ synthetic control that minimizes the distance defined as:

$$\|\boldsymbol{X_1} - \boldsymbol{X_0 W}\| = \sqrt{(\boldsymbol{X_1} - \boldsymbol{X_0 W})' \boldsymbol{V} (\boldsymbol{X_1} - \boldsymbol{X_0 W})} = \left(\sum_{h=1}^{k} v_h (X_{h1} - w_2 X_{h2} - \cdots - w_{J+1} X_{hJ+1})^2\right)^{\frac{1}{2}} \tag{4}$$

under conditions:

$$0 \leq w_j \quad j = 2,, \ldots J \quad w_2 + \cdots + w_{J+1} = 1$$

Positive $v_1, \ldots \, v_k$ values illustrate validity of each $X_{11}, \ldots \, X_{k1}$ predictive variable. For a given set of $v_1, \ldots \, v_k$ values, the minimizing of the (4) equation constitutes the problem of the square optimization. The question is how to choose the $\boldsymbol{V}$ vector. Abadie and Gardeazabal (2003) as well as Abadie *et al.* (2010) suggest choosing the $\boldsymbol{V}$ vector that minimizes the mean squared prediction error (MSPE) for the outcome over some set of pre-intervention periods. In other words, the $Z_1(T_P x 1)$ is a vector of the outcome for the treated unit over some set of pre-intervention periods and $Z_0(T_P x J)$ is a matrix of corresponding values for units from the donor pool, where T_P $(1 \leq T_P \leq T_0)$ is a number of pre-intervention periods for which the mean squared prediction error (MSPE) is minimized. Then, the V^* is chosen to minimize:

$$arg\min_{V \in \gamma} (Z_1 - Z_0 W^*(V))'(Z_1 - Z_0 W^*(V)) \tag{5}$$

where γ is a set of all positive (KxK) diagonal matrices. In the end, the embedded optimization problem is solved, which minimized the above (5) equation for $W^*(V^*)$ defined by the (4) equation.

4. Research sample

The goal of the research is to study the impact of restrictions imposed in Poland on the spread of the COVID-19 pandemic. For this purpose, the author used Warmińsko-Mazurskie region (voivodship), where lockdown was introduced on 27 February 2021. As mentioned before, this is the first region in Poland where the lockdown was introduced during the third wave of the coronavirus pandemic. The research covers the period of 36 days, which is period of two weeks prior to restrictions and three weeks after they were introduced. It is the period between the lifting of restrictions in Poland (12 February 2021) and their reintroduction in the entire country (20 March 2021).

It was assumed that restrictions introduced on 13 and 15 March in three regions had not shown the desired effect – to curb the spread of the pandemic. The assumption was based on the fact that according to the literature, the average period between the infection and a positive test result is 11.7 days (Mitze et al., 2020).

Apart from Warmińsko-Mazurskie region, all 15 voivodships in Poland constituted the set of potential comparative units – the donor pool. The Y_{jt} outcome variable responsible for the spread of the pandemic in every voivodships was the accumulated number of new infections from 12 February 2021 until 20 march 2021 (36 observations for each region). The predictive variables were:

- accumulated number of cases a day and seven days prior to the restrictions (2 observations for each voivodship),
- average daily number of cases in the last 7 days prior to the restrictions (1 observation for each voivodship),
- young-age dependency ratio (ratio of people aged 14 and younger to the 15–64 age group) in 2020 (1 observation for each voivodship),
- share of people living in cities in 2020 (1 observation for each voivodship),
- doctors per 10 thousand citizens in 2020 (1 observation for each voivodship),
- pharmacies per 10 thousand citizens in 2020 (1 observation for each voivodship),
- number of people vaccinated per 10 thousand citizens on a day prior to restrictions in 2020 (1 observation for each voivodship),
- number of recoveries per 10 thousand citizens on a day prior to restrictions in 2020 (1 observation for each voivodship).

They were chosen based on the literature review and the availability of data. Data on the spread of the pandemic in Poland were taken from the government website on the pandemic in Poland (Ministry of Health 2021), whereas data on demographics and healthcare were taken from the Local Data Bank (GUS 2021). Calculations were made in the R statistical environment.

5. Results of the empirical study

Using the methodology and data described in previous parts, based on the two-week period prior to restrictions (13.02.2021-26.02.2021), the author construed the synthetic control unit – synthetic Warmińsko-Mazurskie region (voivodship). Table 1 includes weights created for this purpose. Based on that, it is safe to say that the synthetic region is the combination of three regions: Kujawsko-Pomorskie, Śląskie and Mazowieckie.

Table 1. Synthetic weights for Warmińsko-Mazurskie

Region	Synthetic control weight	Region	Synthetic control weight
Dolnośląskie	0.001	Podkarpackie	0.001
Kujawsko-Pomorskie	0.745	Podlaskie	0.000
Łódzkie	0.001	Pomorskie	0.000
Lubelskie	0.000	Śląskie	0.196
Lubuskie	0.000	Świętokrzyskie	0.000
Małopolskie	0.000	Wielkopolskie	0.000
Mazowieckie	0.055	Zachodniopomorskie	0.001
Opolskie	0.000		

Source: own calculation based on GUS (2021) and Ministry of Health (2021).

Table 2 presents values of predictive variables used in the study for two units, that is Warmińsko-Mazurskie and synthetic Warmińsko-Mazurskie regions as well as an average value for the 15 remaining regions.

Table 2. Averages for predictive variables of accumulated number of cases

Variable	Warmińsko-Mazurskie		Average of 15 control voivodships
	Real	Synthetic	
Accumulated number of cases a day prior to the restrictions	8944.0	8878.1	6808.4
Accumulated number of cases seven days prior to the restrictions	4378.0	4414.2	3306.9
Average daily number of cases in the last 7 days prior to the restrictions	704.3	715.8	562.1
Young-age dependency ratio (persons aged 14 years and below per 100 of population aged 15-64 years)	22.5	22.8	22.8
Doctors per 10 thousand citizens	42.4	58.5	55.7
Pharmacies per 10 thousand citizens	2.9	2.9	3.1
Share of people living in cities	59.0	62.4	58.3
Number of people vaccinated per 10 thousand citizens on a day prior to restrictions	799.1	785.9	848.3
Number of recoveries per 10 thousand citizens on a day prior to restrictions	575.6	534.8	432.0
	RMSPE	126.6	

Source: own calculation based on GUS (2021) and Ministry of Health (2021).

According to the results, for the vast majority of predictive variables, the synthetic Warmińsko-Mazurskie region is more similar to the actual region than to all other studied regions. In other words, the determined linear combination of regions better reflects characteristics of the studied unit than the average of the donor pool. The table also presents the root mean square prediction error (RMSPE). It measures the difference between outcome variables (accumulated number of cases) for Warmińsko-Mazurskie region and its synthetic equivalent for the period prior to restrictions. Figure 3 presents slight differences. Trajectories of the pandemic for both analysed units are aligned in the studied period. The synthetic region almost perfectly illustrates outcome variables for the actual Warmińsko-Mazurskie region.

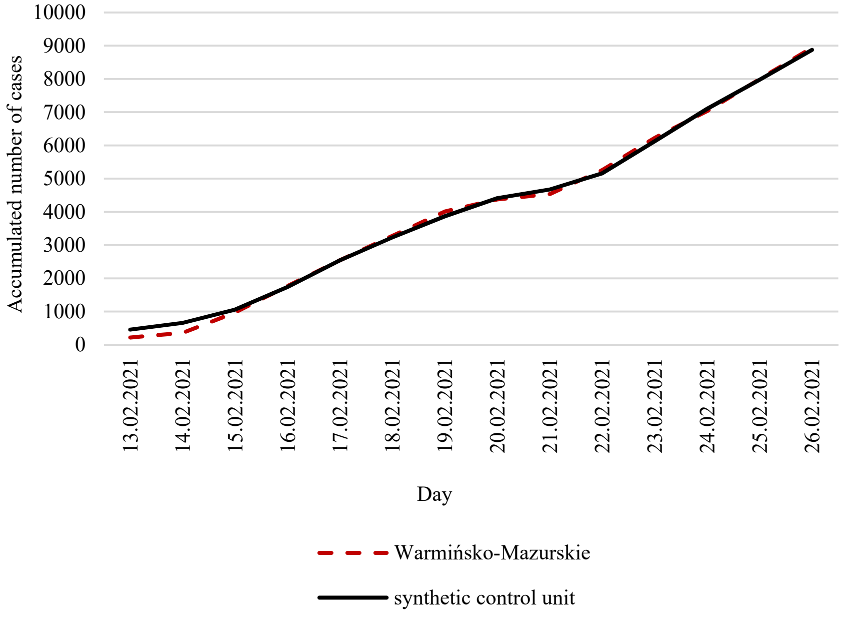


Figure 3. Accumulated number of cases for Warmińsko-Mazurskie region and synthetic Warmińsko-Mazurskie region in the period prior to restrictions

Source: own calculation based on GUS (2021) and Ministry of Health (2021).

For the goal of the research, it is crucial how the variable responsible for the accumulated number of new cases for the synthetic unit changes after restrictions were introduced. The figure presents data on the spread of the pandemic in the studied period.

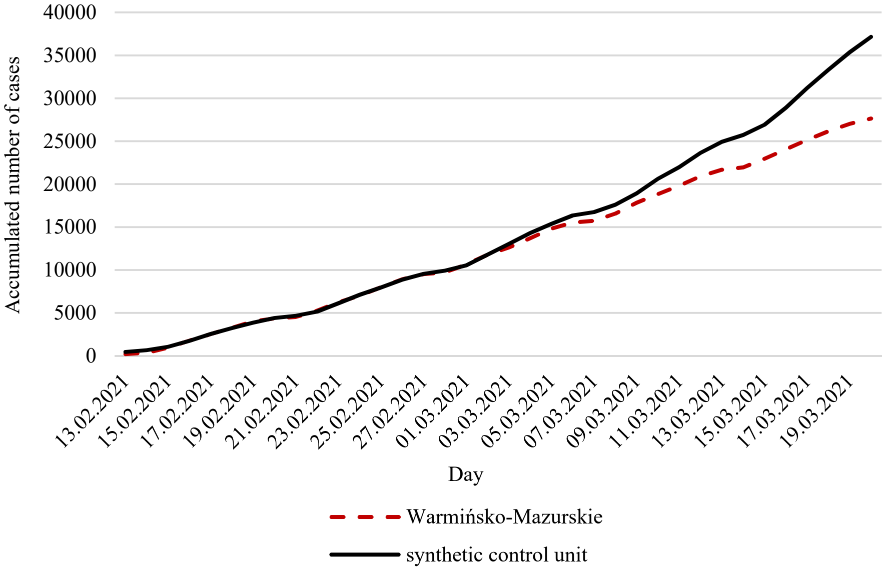


Figure 4. Accumulated number of cases in Warmińsko-Mazurskie region and for the synthetic control unit in the studied period

Source: own calculation based on GUS (2021) and Ministry of Health (2021).

The figure shows that after 27 February there is a growing discrepancy between presented trajectories that are responsible for the spread of the pandemic. Over time, values for the synthetic unit are more and more different compared to actual numbers in the studied region in Poland. It is worth reminding that the synthetic unit is responsible for the situation with an absence of the studied intervention; that is if the lockdown had not been introduced. According to the results, only 10 days after the restrictions were imposed, the accumulated number of cases for the Warmińsko-Mazurskie region would have been lower by 1100 compared to the synthetic unit. On the last studied day, the difference grew to more than 9500 cases. Thus, the conclusion is that if it had not been for the lockdown introduced in the Warmińsko-Mazurskie region, on 20 March 2020 the number of new cases would have been higher by 34% compared to the reality.

To assess the validity of the results, placebo studies were conducted. As a result, it was verified whether differences visible in Figure 4 stem from the introduced restrictions or lack of prognostic abilities of the adopted method. Placebo studies were conducted in two dimensions: time and space. In the first scenario, the entire analysis was repeated assuming that the restrictions were introduced earlier, e.g. on 22 January 2020. Figure 5 presents accumulated number of cases.

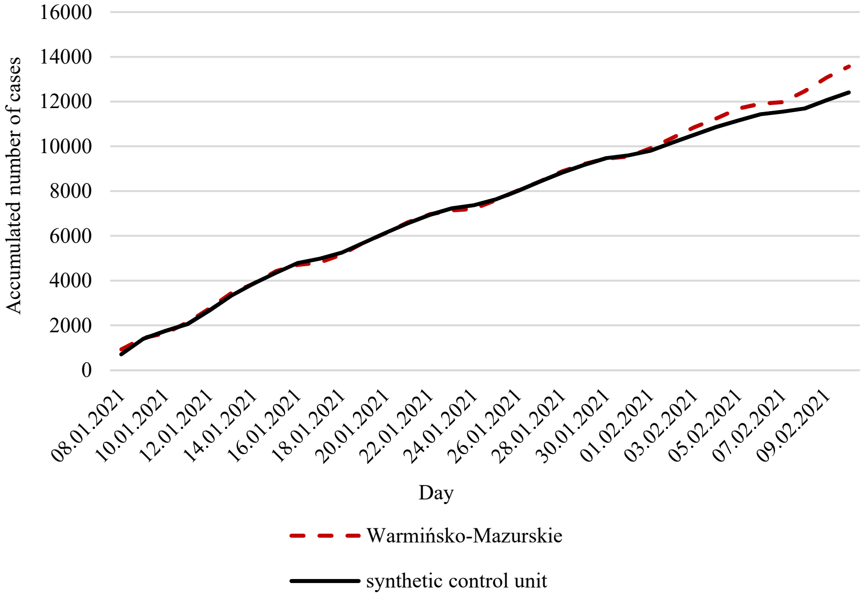


Figure 5. Placebo-in-time tests for (pseudo) treatment effects in the period 22 January to 12 February

Source: own calculation based on GUS (2021) and Ministry of Health (2021).

According to the figure, the dynamics of the pandemic evolution in Warmińsko-Mazurskie region and its synthetic equivalent is similar. Only the last 4 days show slightly higher discrepancies between the curves presented on the figure. Most importantly, unlike in the case of Figure 4, values for the synthetic unit are lower compared to the Warmińsko-Mazurskie region.

In the second placebo study, analyses are redone assuming that units affected by the intervention (lockdown) are all 15 regions from the donor pool. As a result, it was possible to compare the estimated outcome of restrictions for the Warmińsko-Mazurskie region with the distribution of placebo outcomes for other regions. The assumption is that the studied outcome for the Warmińsko-Mazurskie region is relevant if the gap (defined as the difference between the accumulated number of cases for the actual region and the synthetic one) for the Warmińsko-Mazurskie is high in comparison with the gaps from the donor pool. The results are shown in Figure 6. The figure presents the distribution of the gaps. It is clear that Warmińsko-Mazurskie voivodship is one of the regions for which the studied difference is negative. Most importantly, it is the highest (absolute value) on the last day of the studied period. Results of the placebo studies show that the outcome presented in Figure 4 is the actual outcome of restrictions introduced in Warmińsko-Mazurskie voivodship.

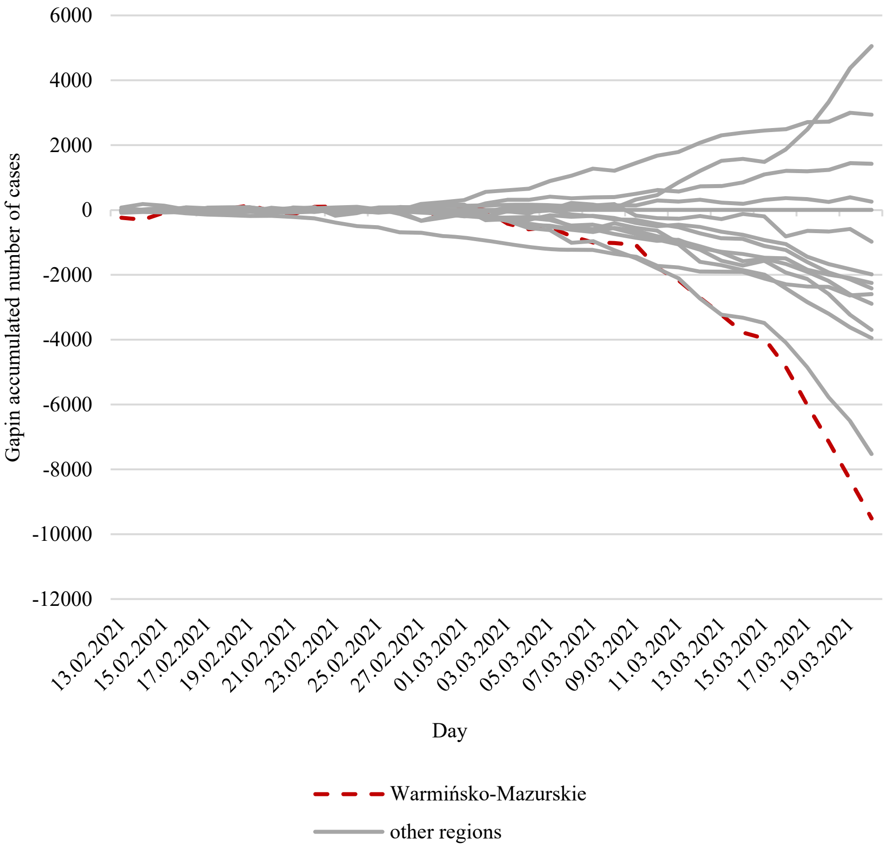


Figure 6. Placebo-in-space tests (all voivodships)

Source: own calculation based on GUS (2021) and Ministry of Health (2021).

The last part of the study includes robustness tests and is connected with the sensitivity analysis determining how the adopted method is affected by the change of the research sample. First, it was studied how the outcomes would be changing if different regions were taken to the donor pool. Table 1 shows that three regions (Kujawsko-Pomorskie, Śląskie and Mazowieckie) are crucial for the construction of the synthetic Warmińsko-Mazurskie region. Using variables presented earlier, the methodology was used three more times and one of the above-mentioned regions was left out from the donor pool in next iterations. The outcomes on the accumulated number of cases in Warmińsko-Mazurskie region and its synthetic equivalents are presented in Figure 7.

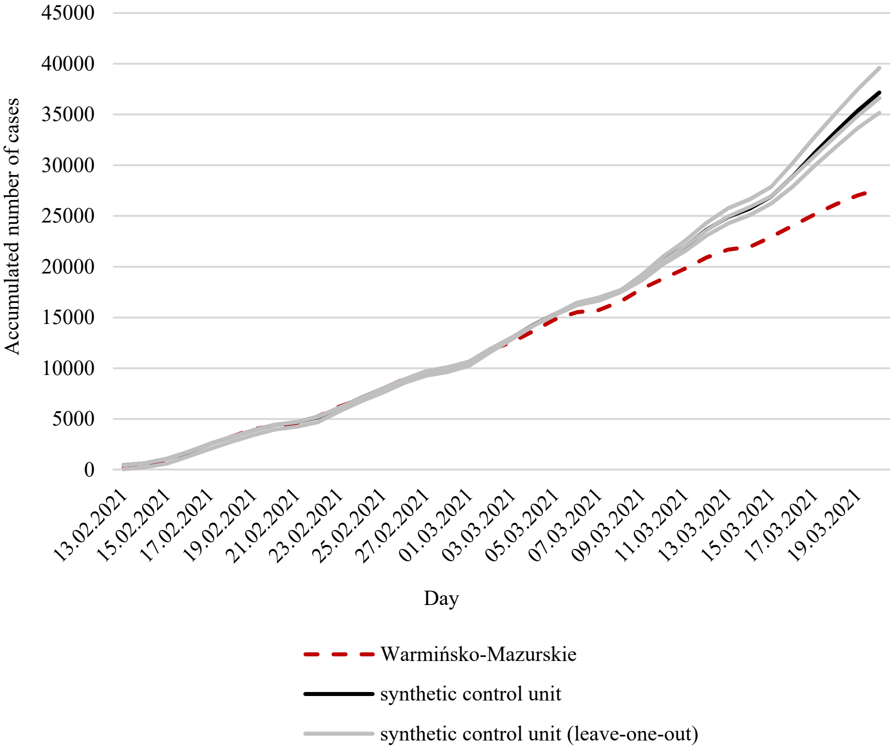


Figure 7. Leave-One-Out Distribution of the Synthetic Control for Warmińsko-Mazurskie

Source: own calculation based on GUS (2021) and Ministry of Health (2021).

The figure shows that outcomes obtained at the beginning of the research are quite resistant to leaving out the most important regions from the donor pool. The leave-one-out synthetic controls produce a very similar effect of restrictions introduced in Warmińsko-Mazurskie. In all three cases introducing lockdown causes decrease in the number of cases. Only in one case (after deleting the Mazowieckie region) the effect of lockdown would be higher than for whole quota of donors. The second robustness test included the change of predictive variables. To conduct the test, calculations were repeated twice. First, not taking into account demographics and healthcare data in Poland, then, not including variables on the dynamics of the pandemic in Poland. Table 3 shows weights for particular variables in presented calculations and root mean square prediction errors.

Table 3. Weights for analysed predictive variables

Variable	**Synthetic control unit**	**Synthetic control unit (without variables on demographics and healthcare)**	**Synthetic control unit (without variables on the dynamics of the pandemic)**
Accumulated number of cases a day prior to the restrictions	0.624	0.870	-
Accumulated number of cases seven days prior to the restrictions	0.195	0.096	-
Average daily number of cases in the last 7 days prior to the restrictions	0.177	0.030	-
Young-age dependency ratio (persons aged 14 years and below per 100 of population aged 15-64 years)	0.004	-	0.000
Doctors per 10 thousand citizens	0.000	-	0.840
Pharmacies per 10 thousand citizens	0.000	-	0.005
Share of people living in cities	0.000	-	0.107
Number of people vaccinated per 10 thousand citizens on a day prior to restrictions	0.000	0.004	0.000
Number of recoveries per 10 thousand citizens on a day prior to restrictions	0.001	0.000	0.048
RMSPE	126.600	126.352	158.691

Source: own calculation based on GUS (2021) and Ministry of Health (2021)

The results show that leaving out variables on demographics and healthcare does not have much impact on the alignment of the synthetic unit to the actual one. More significant impact (higher RMSPE value) can be observed if variables on the dynamics of the pandemic are excluded from the study. However, the value of the error is not very high. The same conclusion can be drawn from analysing the figure on the accumulated number of cases in the studied period (Figure 8).

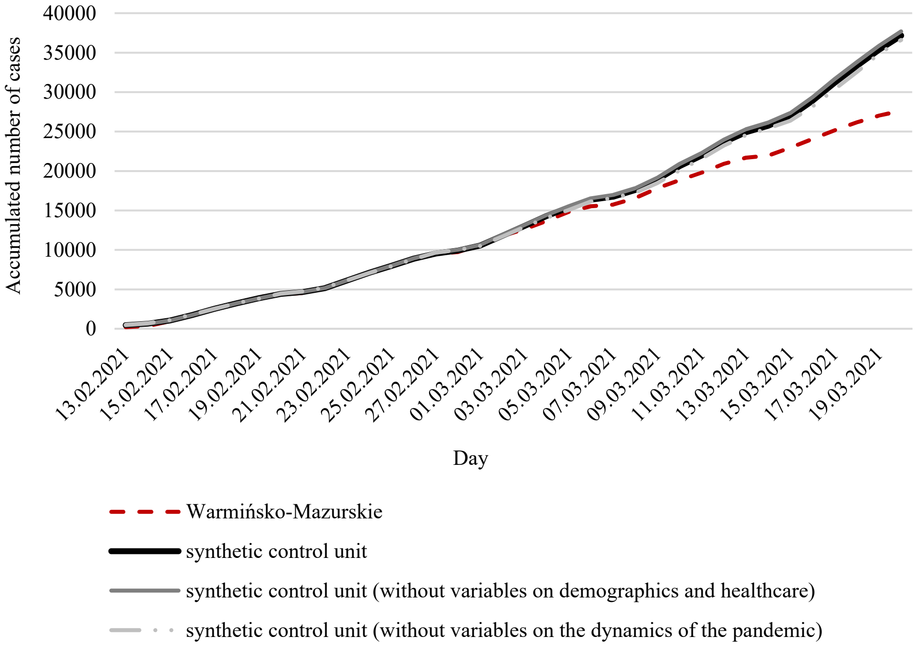


Figure 8. Accumulated number of cases for Warmińsko-Mazurskie and synthetic regions with a different set of predictive variables

Source: own calculation based on GUS (2021) and Ministry of Health (2021).

It is clear that a day before the restrictions were introduced, all curves have a very similar position. What is more, after the restrictions were implemented, their dynamics are not changing significantly. Therefore, the synthetic control method and its results are resistant to the change of assumptions made at the beginning of the study.

6. Conclusions

Due to the rapid spread of the pandemic, governments in many countries decided to introduce various restrictions. Their main goal was to reduce contact between people, which is how the virus transmits. Closing of shopping malls, cinemas, schools, hotels involves high economic and social costs. Since lockdowns are put in place repeatedly and for a longer period, the society has been rebelling against them more and more often. Therefore, before the next pandemic wave, it is essential to estimate the validity of restriction on life and health of citizens.

The main goal of the research was to study the impact of lockdown on the spread of COVID-19 in Poland. Results show that lockdown is an efficient tool that curbs the spread of the COVID-19 pandemic. Its introduction has significantly limited the number of new cases in the analysed region in Poland. The research included the construction of the synthetic region that well illustrated the tendency of the pandemic development in the Warmińsko-Mazurskie region before the lockdown. After that, with time, the virus spread trajectories started to differ considerably. Results indicate that imposed restrictions decreased the number of coronavirus cases by 9500 people in 21 days. Placebo-in-space and placebo-in-time studies proved that results are reliable and resistant to the change of research assumptions.

Acknowledgments

The project financed within the Regional Initiative for Excellence programme of the Minister of Science and Higher Education of Poland, years 2019-2022, grant no. 004/RID/2018/19, financing 3,000,000 PLN.

References

Abadie, A., (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, 59(2), pp. 391–425, doi: 10.1257/jel.20191450.

Abadie, A., Diamond, A. and Hainmueller, A. J., (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco control program. *Journal of the American Statistical Association*, 105(490), pp. 493–505, doi: 10.1198/jasa.2009.ap08746.

Abadie, A., Diamond, A. and Hainmueller, J., (2011). Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, 42(13), pp. 1–17, doi: 10.18637/jss.v042.i13.

Abadie, A., Diamond, A., and Hainmueller, J., (2015). Comparative Politics and the Synthetic Control Method. American Journal of Political Science, 59(2), pp. 495–510, doi: 10.1111/ajps.12116.

Abadie, A., and Gardeazabal, J., (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review*, 93(1), pp. 113–132, doi: 10.1257/000282803321455188.

Abouk, R. and Heydari, B. (2021). *The Immediate Effect of COVID-19 Policies on Social-Distancing Behavior in the United States*, Public Health Reports, 136(2), pp. 245–252, doi: 10.1177/0033354920976575.

Alfano, V., and Ercolano, S., (2020). The Efficacy of Lockdown Against COVID-19: A Cross-Country Panel Analysis. *Applied Health Economics and Health Policy*, 18(4), pp. 509–517, doi: 10.1007/s40258-020-00596-3.

Alfano, V., Ercolano, S. and Cicatiello, L., (2021). School openings and the COVID-19 outbreak in Italy. A provincial-level analysis using the synthetic control method, *Health Policy*, 125(9), pp. 1200–1207, doi: 10.1016/j.healthpol.2021.06.010.

Bayat, N., Morrin, C., Wang, Y., and Misra, V., (2020). *Synthetic control, synthetic interventions, and COVID-19 spread: Exploring the impact of lockdown measures and herd immunity*, arXiv preprint arXiv:2009.09987.

Bo, Y., Guo, C., Lin, C., Zeng, Y., Li, H. B., Zhang, Y., Hossain, M. S., Chan, J., Yeung, D. W., Kwok, K. O., Wong, S. Y. S, Lau, A. K. H., and Lao, X. Q., (2021). Effectiveness of non-pharmaceutical interventions on COVID-19 transmission in 190 countries from 23 January to 13 April 2020. *International Journal of Infectious Diseases*, 102, pp. 247–253, doi: 10.1016/j.ijid.2020.10.066.

Bonaccorsi, G., Pierri, F., Cinelli, M., Flori, A., Galeazzi, A., Porcelli, F., Schmidt, A. L., Valensise, C. M., Scala, A., Quattrociocchi, W., and Pammolli, F., (2020). Economic and social consequences of human mobility restrictions under COVID-19. *Proceedings of the National Academy of Sciences*, 117(27), 15530–15535, https://doi.org/10.1073/pnas.2007658117

Born, B., Dietrich, A. M., and Müller, G. J., (2021). The lockdown effect: A counterfactual for Sweden. *PLoS ONE*, 16(4 April 2021), pp. 1–13, doi: 10.1371/journal.pone.0249732.

Chernozhukov, V., Kasahara, H., and Schrimpf, P., (2021). Causal impact of masks, policies, behavior on early COVID-19 pandemic in the U.S. *Journal of Econometrics*, 220(1), pp. 23–62, https://doi.org/10.1016/j.jeconom.2020.09.003

Cho, S.W., (2020). Quantifying the impact of nonpharmaceutical interventions during the COVID-19 outbreak: The case of Sweden. *The Econometrics Journal*, 23(3), pp. 323–344, doi: 10.1093/ectj/utaa025.

Coccia, M., (2021). The relation between length of lockdown, numbers of infected people and deaths of Covid-19, and economic growth of countries: Lessons learned to cope with future pandemics similar to Covid-19 and to constrain the deterioration of economic system. *Science of The Total Environment*, 775, pp. 145801, doi: https://doi.org/10.1016/j.scitotenv.2021.145801.

Courtemanche, C., Garuccio, J., Le, A., Pinkston, J., and Yelowitz, A., (2020). Strong social distancing measures in the United States reduced the COVID-19 growth rate. *Health Affairs*, 39(7), pp. 1237–1246, https://doi.org/10.1377/hlthaff.2020.00608

Fountoulakis, K. N., Fountoulakis, N. K., Koupidis, S. A., and Prezerakos, P. E., (2020). Factors determining different death rates because of the COVID-19 outbreak among countries. *Journal of Public Health*, 42(4), pp. 681–687, https://doi.org/10.1093/pubmed/fdaa119

Goodman-Bacon, A., and Marcus, J., (2020). Using Difference-in-Differences to Identify Causal Effects of COVID-19 Policies. *Survey Research Methods*, 14(2 SE-Design proposals), pp. 153–158. doi: 10.18148/srm/2020.v14i2.7723.

GUS, (2021). *Bank Danych lokalnych*, URL https://bdl.stat.gov.pl/BDL/start [Accessed 31 July 2021]

Mendez-Brito, A., El Bcheraoui, C. and Pozo-Martin, F., (2021). Systematic review of empirical studies comparing the effectiveness of non-pharmaceutical interventions against COVID-19. *Journal of Infection*, 83(3), pp. 281–293, doi: 10.1016/j.jinf.2021.06.018.

Ke, X., and Hsiao, C., (2021). Economic impact of the most drastic lockdown during COVID-19 pandemic—The experience of Hubei, China. *Journal of Applied Econometrics*, March, 1–23, https://doi.org/10.1002/jae.2871

Lai, S., Ruktanonchai, N. W., Zhou, L., Prosper, O., Luo, W., Floyd, J. R., Wesolowski, A., Santillana, M., Zhang, C., Du, X., Yu, H., and Tatem, A. J., (2020). Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature*, 585(7825), pp. 410–413, https://doi.org/10.1038/s41586-020-2293-x

Li, Y., Li, M., Rice, M., Zhang, H., Sha, D., Li, M., Su, Y., and Yang, C., (2021). The impact of policy measures on human mobility, COVID-19 cases, and mortality in the US: A spatiotemporal perspective. *International Journal of Environmental Research and Public Health*, 18(3), pp. 1–25, https://doi.org/10.3390/ijerph18030996

Mendez-Brito, A., El Bcheraoui, C. and Pozo-Martin, F. (2021). Systematic review of empirical studies comparing the effectiveness of non-pharmaceutical interventions against COVID-19. *Journal of Infection*, 83(3), pp. 281–293, doi: 10.1016/j.jinf.2021.06.018.

Ministry of Health, (2021). *Raport zakażeń koronawirusem (SARS-CoV-2)*, URL https://www.gov.pl/web/koronawirus/wykaz-zarazen-koronawirusem-sars-cov-2 [Accessed 31 July 2021]

Mitze, T., Kosfeld, R., Rode, J. and Wälde, K., (2020). Face masks considerably reduce COVID-19 cases in Germany. *Proceedings of the National Academy of Sciences*, 117(51), pp. 32293–32301, https://doi.org/10.1073/pnas.2015954117

Palomino, J. C., Rodríguez, J. G., Sebastian, R., (2020). Wage inequality and poverty effects of lockdown and social distancing in Europe. *European Economic Review*, 129, pp. 103564, doi: https://doi.org/10.1016/j.euroecorev.2020.103564.

Piovani, D., Christodoulou, M. N., Hadjidemetriou, A., Pantavou, K., Zaza, P., Bagos, P. G., Bonovas, S., and Nikolopoulos, G. K., (2021). Effect of early application of social distancing interventions on COVID-19 mortality over the first pandemic wave: An analysis of longitudinal data from 37 countries. *Journal of Infection*, 82(1), pp. 133–142, https://doi.org/10.1016/j.jinf.2020.11.033

Ritchie, H., Ortiz-Ospina, E., Beltekian, D., Mathieu, E., Hasell, J., Macdonald, B., Giattino, C., Appel, C., Rodés-Guirao, L., and Roser, M., (2021). *Coronavirus pandemic (COVID-19).* Our World in Data, URL https://ourworldindata.org/coronavirus [Accessed 31 July 2021]

Ruan, L., Wen, M., Zeng, Q., Chen, C., Huang, S., Yang, S., Yang, J., Wang, J., Hu, Y., Ding, S., Zhang, Y., Zhang, H., Feng, Y., Jin, K., and Zhuge, Q., (2020). New Measures for the Coronavirus Disease 2019 Response: A Lesson From the Wenzhou Experience. *Clinical Infectious Diseases*, 71(15), pp. 866–869, https://doi.org/10.1093/cid/ciaa386

Siedner, M. J., Harling, G., Reynolds, Z., Gilbert, R. F., Haneuse, S., Venkataramani, A. S., and Tsai, A. C., (2020). Social distancing to slow the US COVID-19 epidemic: Longitudinal pretest–posttest comparison group study. *PLoS Medicine*, 17(8 August), pp. 1–12, https://doi.org/10.1371/JOURNAL.PMED.1003244

Tian, T., Tan, J., Luo, W., Jiang, Y., Chen, M., Yang, S., Wen, C., Pan, W., and Wang, X., (2021). The Effects of Stringent and Mild Interventions for Coronavirus Pandemic. *Journal of the American Statistical Association*, 116(534), pp. 481–491, https://doi.org/10.1080/01621459.2021.1897015

Tian, T., Luo, W., Tan, J., Jiang, Y., Chen, M., Pan, W., Yang, S., Zhao, J., Wang, X., and Zhang, H., (2021). The timing and effectiveness of implementing mild interventions of COVID-19 in large industrial regions via a synthetic control method. *Statistics and Its Interface*, 14(1), pp. 3–12, https://doi.org/10.4310/20-SII634

Wu, S., Yao, M., Deng, C., Marsiglia, F. F., and Duan, W., (2021). Social Isolation and Anxiety Disorder During the COVID-19 Pandemic and Lockdown in China. *Journal of Affective Disorders*, 294, pp. 10–16, https://doi.org/https://doi.org/10.1016/j.jad.2021.06.067

Zhang, R., Li, Y., Zhang, A. L., Wang, Y., and Molina, M. J., (2020). Identifying airborne transmission as the dominant route for the spread of COVID-19. *Proceedings of the National Academy of Sciences*, 117(26), 14857 LP–14863, https://doi.org/10.1073/pnas.2009637117\

Zhang, H., Li, P., Zhang, Z., Li, W., Chen, J., Song, X., Shibasaki, R., and Yan, J. (2022). Epidemic versus economic performances of the COVID-19 lockdown: A big data driven analysis. *Cities*, 120, 103502, https://doi.org/10.1016/j.cities.2021.103502

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 147–165, DOI 10.2478/stattrans-2022-0035
Received – 24.12.2020; accepted – 08.04.2022

Modified exponential time series model with prediction of total COVID-19 cases in Belgium, Czech Republic, Poland and Switzerland

Wachirapond Permpoonsinsup[1], Rapin Sunthornwat[2]

ABSTRACT

The coronavirus (COVID-19) pandemic affected every country worldwide. In particular, outbreaks in Belgium, the Czech Republic, Poland and Switzerland entered the second wave and was exponentially increasing between July and November, 2020. The aims of the study are: to estimate the compound growth rate, to develop a modified exponential time-series model compared with the hyperbolic time-series model, and to estimate the optimal parameters for the models based on the exponential least-squares, three selected points, partial-sums methods, and the hyperbolic least-squares for the daily COVID-19 cases in Belgium, the Czech Republic, Poland and Switzerland. The speed and spreading power of COVID-19 infections were obtained by using derivative and root-mean-squared methods, respectively. The results show that the exponential least-squares method was the most suitable for the parameter estimation. The compound growth rate of COVID-19 infection was the highest in Switzerland, and the speed and spreading power of COVID-19 infection were the highest in Poland between July and November, 2020.

Key words: COVID-19, modified exponential time-series model, method of parameter estimation, compound growth rate.

1. Introduction

Since the end of 2019, the coronavirus disease 2019 (COVID-19) outbreak caused by the SARS-Cov-2 virus, which started in Wuhan of Hubei Province, China, has spread throughout the world. The outbreak in Europe has entered the second wave with increasing numbers of COVID-19 cases in many countries. As of November 12, 2020,

[1] Industrial Technology and Innovation Management Program, Faculty of Science and Technology, Pathumwan Institute of Technology, Bangkok, Thailand. E-mail: wachirapond@pit.ac.th. ORCID: https://orcid.org/0000-0003-4033-493X.

[2] Industrial Technology and Innovation Management Program, Faculty of Science and Technology, Pathumwan Institute of Technology, Bangkok, Thailand. E-mail: rapin@pit.ac.th, Corresponding Author. ORCID: https://orcid.org/0000-0001-8981-5107.

there were 12,914,903 cases and 306,504 deaths in Europe (Worldometer, 2020). Especially, the second-wave COVID-19 outbreaks in Belgium, the Czech Republic, Poland, and Switzerland rapidly spread with what looks like an exponential function; at the same time point, there were 507,475 cases and 13,561 deaths in Belgium, 438,805 cases and 5,570 deaths in the Czech Republic, 641,496 cases and 9,080 deaths in Poland, and 243,472 cases and 3,113 deaths in Switzerland. Measures imposed by the governments of these countries, such as lockdown policies, face mask-wearing in public areas, encouraging hand washing, avoiding public areas, and prohibiting people from assembling were launched to control and protect the population from the spread of COVID-19. In addition, physical and social distancing have remained in practice in many areas, while learning from home for students and working from home have become necessary policies. Thus, the spread of COVID-19 has gone mainly unchallenged due to a lack of medical equipment and personnel to fight the pandemic.

Forecasting the number of COVID-19 cases (the number of people contracting the disease) is essential for planning the necessary provisions for medical treatment (hospital beds, ventilators, personal protective equipment, etc.). Many models for forecasting COVID-19 cases comprising time-series data have been studied. For example, linear regression analysis, machine learning, vector support regression machine, and autoregressive integrated moving average (ARIMA) models based on the linear relationship between the time variable and dependent variables have been popular for establishing models for forecasting COVID-19 cases. Forecasting new cases and new deaths from COVID-19 in Ethiopia was investigated by Argaru (2020). Linear regression analysis of COVID-19 data comprising new cases, deaths, the number of days, and recoveries in May and June to estimate the parameters for a forecasting model has been reported. The relationship between COVID-19 data and time has been analysed by using Pearson's correlation analysis. The results show that there is a correlation between new COVID-19 cases and deaths, while the number of days and new recoveries were significant to the new deaths. For the COVID-19 outbreak in Henan province, China, linear regression analysis was adopted to estimate parameters for constructing a forecasting model and to study the relationship between the number of people from Wuhan who had travelled and the number of cases in 18 cities in Henan province. The results show a statistically significant linear correlation between the number of people traveling from Wuhan and the number of cases (Cheng, 2020).

The COVID-19 outbreak in India has caused socioeconomic recession and mounting deaths. Both multiple and linear regression analyses have been adopted to predict the number of deaths and to study correlations in the COVID-19 data from India, with the ability of the developed predictive model being based on autoregression (Ghosal et al., 2020). The dependent variable (the number of active cases) in the forecasting model was correlated with the independent variables (the number of cases, deceased, and recovered). In a study of public health responses in various countries by Rath, Tripathy and Tripathy (2020), the performances of forecasting models were compared based on the coefficient of determination and correlation. Correlations between intervention scores, daily new cases, and doubling time were significant for identifying epidemiological changes in the spread of COVID-19. In research involving linear and polynomial regression analysis for predicting the COVID-19 fatality rate in Nigeria by Suleiman et al. (2020), the results reveal that the polynomial regression model is suitable for predicting the COVID-19 fatality rate in this particular country.

In other studies, linear regression analysis was employed by Melik-Huseynov et al. (2020) to estimate new cases of COVID-19. Simple regression analysis was applied by Losif et al. (2020) to study the incidence of correlation between the COVID-19 peak or plateau and air traffic volume. Forecasting models based on polynomial regression were investigated by Ekum and Ogunsanya (2020) to forecast new cases of COVID-19; their results show that the cubic polynomial regression model performed better than other polynomial regressions. Calculating the fatality rate based on linear regression analysis and comparing its efficacy among countries affected by the COVID-19 pandemic was conducted by Hoseinpour et al. (2020). Support vector regression as a predictive model for the duration of spread and analysis of growth and transmission rates was used to evaluate the correlation between COVID-19 outbreaks and weather conditions by Yadav, Perumal and Srinivas (2020). Machine learning was applied as a forecasting model based on linear regression, least absolute shrinkage and selection operator, support vector machine, and exponential smoothing for the number of COVID-19 patients, new infection cases, deaths, and recoveries by Rustam et al. (2020); their results proved that exponential smoothing offered the best performance. Linear regression and support vector machine analyses have been used for predicting the number of COVID-19 cases to aid decision-making by the government in India (Likhesh et al., 2020). Prediction models and comparison between the susceptible-exposed/infectious-recovered model and regression analysis were used to predict the number of COVID-19 cases in India by Pandey et al. (2020).

ARIMA models have often been applied to COVID-19 time-series data. An ARIMA model and regression analysis were used to estimate the mortality rate of COVID-19 by Chaurasia and Pal (2020). Forecasting COVID-19 time-series data based on an ARIMA model in the US, Brazil, India, Russia, and Spain was investigated by Sahai et al. (2020). An ARIMA model was created to forecast new cases and deaths from COVID-19 time-series data by Yang et al. (2020). An ARIMA model for short-term prediction was developed by Fang, Wang and Pan (2020) to predict COVID-19 cases, deaths, and recoveries in Russia. An ARIMA model was applied by Benvenuto et al. (2020) to a COVID-19 time-series dataset from the Johns Hopkins database for forecasting the trend and incidence of COVID-19 outbreak. Singh et al. (2020)

developed an ARIMA model for predicting confirmed cases, deaths, and recoveries with a spatial map showing the intensity of each criterion. Moreover, they used the Akaike information criterion to validate the ARIMA model.

Because linear regression analysis, linear machine learning, linear support vector regression, and ARIMA model are dependent on the linear combination of time as the independent variable to predict dependent variables, our aim was to develop a modified exponential time-series model compared with hyperbolic time-series model that is nonlinear and uses the exponential growth rate to forecast the number of COVID-19 cases increasing rapidly each day in Belgium, the Czech Republic, Poland, and Switzerland. Herein, the accuracy and validation of the developed model are reported, while its ability to predict the speed and spreading power of the daily COVID-19 cases are illustrated.

2. Methods

2.1. Derivation of the modified exponential time-series model

A time series is a sequence of observations taken sequentially in time (George et al., 2015). The number of COVID-19 cases per day is an example of a time series. It can be represented by modelling its curve as the solution of a differential equation with time as the independent variable. In this research, a modified exponential curve is adopted to analyse and forecast the daily COVID-19 cases as follows.

Let $y(t)$ be the number of total daily COVID-19 cases at time t. Differential equation which represents the speed of COVID-19 cases and is solved into the modified exponential curve $y(t)$ can be derived as follows:

$$\frac{dy}{dt} = b\ln(c)c^t;\ y(0) = a + b;\ a,b,c \in \Re \tag{1}$$

Taking integral both sides of Equation (1), the result becomes

$$\int \frac{dy}{b\ln(c)} = \int c^t dt$$

$y(t) = bc^t + Kb\ln(c)$ where K is an arbitrary constant.

With initial condition $y(0) = a + b$, the K value can be carried out as

$$K = \frac{a}{b\ln(c)}$$

Therefore, a modified exponential curve is

$$y(t) = a + bc^t \tag{2}$$

where a,b,c are the parameters.

2.2. Compound growth rate and parameter estimations with the algorithm

Let $\{y(t)\}_{t=1}^{n}$ be a time series of the number of COVID-19 cases. The dependent variable y , which is related to independent variable t has the modified exponential relationship to t as in Equation (2). In this research, the exponential least-squares method, three selected points, partial-sums method, and the hyperbolic least-squares method were employed for the estimation of parameters as follows.

2.2.1. The exponential least-squares method

The exponential least-squares method is to seek an approximating function that best fits the data points (Kharab and Guenther, 2012). It is based on the sum of squares error (SSE) defined by

$$SSE=(y-\hat{y})^2 \tag{3}$$

where y is an actual value of time series and $\hat{y}$ is a forecasted value of time series.

The partial derivative is taken into both sides of Equation (3). Then, it is determined to be zero for evaluating the parameters a,b,c based on the minimum of the sum of squares error.

$$\frac{\partial}{\partial a}SSE=\frac{\partial}{\partial a}(y-\hat{y})^2=0,$$
$$\frac{\partial}{\partial b}SSE=\frac{\partial}{\partial b}(y-\hat{y})^2=0,$$
$$\frac{\partial}{\partial c}SSE=\frac{\partial}{\partial c}(y-\hat{y})^2=0.$$

In addition, the compound growth rate of the time series with initial value y_0 is defined as

$$y(t)=y_0(1+r)^t \tag{4}$$

Rewriting Equation (2), the result becomes

$$Y(t)=A+Bt$$

where $A=\ln(a)+\ln(b)=\ln(ab),\ B=\ln(c),\ Y=\ln(y).$

An estimate of the compound growth rate r based on the least squares estimation for estimation of parameters A and B is given by

$$\hat{B}=\frac{\dfrac{\sum_{t=1}^{n}t^2y(t)}{n}-\dfrac{\sum_{t=1}^{n}ty(t)\sum_{t=1}^{n}t}{n}}{\sum_{t=1}^{n}t^2-\dfrac{\left(\sum_{t=1}^{n}t\right)^2}{n}} \text{ and } \hat{A}=\overline{Y}-\hat{B}\overline{t}$$

The estimator of parameter b is given by

$$\hat{c} = \exp(\hat{B}) = 1 + \hat{r}$$

where $\hat{r}$ is the estimator of the compound growth rate.

Therefore, an estimate of the compound growth rate is given by

$$\hat{r} = \hat{c} - 1 \tag{5}$$

Also, Student's T-test is a statistic for significant test of the compound growth rate given by

$$T = \frac{\hat{B}}{SE(\hat{B})};\ df = n - 2. \tag{6}$$

The decision of the significance of the compound growth rate is dependent on the comparison between the calculated value of $|T|$ with the critical value of T or on the consideration of p-value.

2.2.2. The three selected points method

The estimation of parameters of the modified exponential curve is represented by the three selected points method (Das and Chakrabarty, 2017). Three points of the time series coordinates $(t_1^*, y_1^*), (t_2^*, y_2^*), (t_3^*, y_3^*)$ along the time series $\{y(t)\}_{t=1}^{n}$ are selected to estimate parameters a, b, c.

$$y_1^* = a + bc^{t_1^*} \tag{7}$$

$$y_2^* = a + bc^{t_2^*} \tag{8}$$

$$y_3^* = a + bc^{t_3^*} \tag{9}$$

where $h = t_2^* - t_1^* = t_3^* - t_2^*$.

By the algebraic way, the system of Equations (7)-(9) is solved for the estimation of parameters $\hat{a}, \hat{b}, \hat{c}$ as

$$\hat{c} = \left[\frac{y_3^* - y_2^*}{y_2^* - y_1^*}\right]^{\frac{1}{h}}$$

$$\hat{b} = \frac{y_3^* - y_2^*}{\hat{c}^{t_2^*}(\hat{c}^h - 1)} = \frac{y_2^* - y_1^*}{\hat{c}^{t_1^*}(\hat{c}^h - 1)}$$

$$\hat{a} = y_3^* - \hat{b}\hat{c}^{t_3^*} = y_2^* - \hat{b}\hat{c}^{t_2^*} = y_1^* - \hat{b}\hat{c}^{t_1^*}$$

2.2.3. The partial sums method

The partial sums method (Ikaya et al., 2005) is based on the partition of the time series data into three categories with equal length n points. The dependent variable is $y = \{y_1, y_2, y_3, ..., y_n; y_{n+1}, y_{n+2}, y_{n+3}, ..., y_{2n};\ y_{2n+1}, y_{2n+2}, y_{2n+3}, ..., y_{3n}\}$ and the time independent variable is

$$t = \{1, 2, 3, ..., n; n+1, n+2, n+3, ..., 2n;\ 2n+1, 2n+2, 2n+3, ..., 3n\}.$$

Let S_1, S_2, S_3 be the partial sums of the partitions of the dependent variable y. Thus,

$$S_1 = \sum_{t=1}^{n} y_t = an + \frac{bc(c^n - 1)}{c-1} \tag{10}$$

$$S_2 = \sum_{t=n+1}^{2n} y_t = an + \frac{bc^{n+1}(c^n - 1)}{c-1} \tag{11}$$

$$S_3 = \sum_{t=2n+1}^{3n} y_t = an + \frac{bc^{2n+1}(c^n - 1)}{c-1} \tag{12}$$

The algebraic way is adopted to carry out Equations (10)-(12) for the estimation of parameters $\hat{a}, \hat{b}, \hat{c}$ as

$$\hat{c} = \left[\frac{S_3 - S_2}{S_2 - S_1}\right]^{\frac{1}{n}}$$

$$\hat{b} = \frac{(S_2 - S_1)(\hat{c}-1)}{\hat{c}}\left[\frac{S_3 - S_2}{S_2 - S_1} - 1\right]^{-2}$$

$$\hat{a} = \frac{S_1(\hat{c}-1) - \hat{b}\hat{c}(\hat{c}^n - 1)}{n(\hat{c}-1)} = \frac{S_2(\hat{c}-1) - \hat{b}\hat{c}^{n+1}(c^n - 1)}{n(\hat{c}-1)} = \frac{S_3(\hat{c}-1) - \hat{b}\hat{c}^{2n+1}(\hat{c}^n - 1)}{n(\hat{c}-1)}$$

2.2.4. The hyperbolic least-squares method

The hyperbolic least-squares method (Kharab and Guenther, 2012) is a nonlinear model estimation. It is the fitting given observations with hyperbolic time-series model, which is given as

$$y(t) = a + \frac{b}{t}$$

Setting $Y(t) = y(t)$, $\alpha = a$, $\beta = b$, and $T = \frac{1}{t}$, the hyperbolic time-series model can be transformed as

$$Y(T) = \alpha + \beta T$$

Then, the least-squares method is applied to the estimation of parameters a and b.

2.2.5. Statistics for the accuracy and validation of the time-series model and spreading power

In this section, the accuracy and validation of the time series model are measured. The spreading power is also measured. The measurement of validation of the time series model is evaluated by the Root Mean Squared Percentage Error ($RMSPE$) as

$$RMSPE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{y(t) - \hat{y}(t)}{y(t)}\right)^2} \tag{13}$$

where $y(t)$ is an actual value of y and $\hat{y}(t)$ is a forecasted value of y.

For accuracy of the time series model, the coefficient of determination (R^2) is defined as

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{t=1}^{n}\left[y(t) - \hat{y}(t)\right]^2}{\sum_{t=1}^{n}\left[y(t) - \bar{y}\right]^2} \tag{14}$$

where RSS is the Sum of Squares of Residuals,
TSS is the Total Sum of Squares.

The Root Mean Square (RMS) (Jones, 2019) is measured as the spreading power of the COVID-19 cases time series. The RMS can be defined as

$$RMS = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left[y(t)\right]^2} \tag{15}$$

2.2.6. The algorithm for evaluating the parameters, estimating the derivative, root-mean-square (RMS), RMS percentage error (RMSPE), and estimating the compound growth rate

In this section, the algorithm for this research is demonstrated.

Algorithm

```
Input: total COVID-19 cases
  $y \leftarrow$ total COVID-19 cases
  $n \leftarrow$ length(y)
  $t \leftarrow 1 : n$
  $Para_{ExpoLeast}, RMSPE_{ExpoLeast}, R^2_{ExpoLeast} \leftarrow$ Expoleastsquare(modifiedexpo, $t, y$ )
  $Para_{Three}, RMSPE_{Three}, R^2_{Three} \leftarrow$ Threepoints(modifiedexpo, $t, y$ )
  $Para_{Partial}, RMSPE_{Partial}, R^2_{Partial} \leftarrow$ Partialsums(modifiedexpo, $t, y$ )
  $Para_{HyperLeast}, RMSPE_{HyperLeast}, R^2_{HyperLeast} \leftarrow$ Hyperleastsquare(hypebolic, $t, y$ )

 The optimal estimate parameter is based on the minimum of $RMSPE$ and the maximum of $R^2$
  $del \leftarrow 1$
  For $t \leftarrow 2 : n-1$
  $dy(t) \leftarrow (y(t+1) - y(t-1)) / 2 / del$
End
  $SS, SS_1, SS_2, SS_3, SS_4 \leftarrow 0$
```

```
For $i \leftarrow 1:n$
  $SS \leftarrow SS + (y(i))^2$
  $SS_1 \leftarrow SS_1 + t^2 y(t)$
  $SS_2 \leftarrow SS_2 + ty(t)$
  $SS_3 \leftarrow SS_3 + t$
  $SS_4 \leftarrow SS_4 + t^2$
End
  $RMS \leftarrow sqrt(\frac{1}{n}SS)$
  $\hat{B} = \frac{\frac{SS_1}{n} - \frac{(SS_2)(SS_3)}{n}}{SS_4 - \frac{(SS_3)^2}{n}}$
  $\hat{c} \leftarrow \exp(\hat{B})$
  $\hat{r} \leftarrow \hat{c} - 1$
Output: estimate parameters, estimate derivative, $RMS$, $RMSPE$, $\hat{r}$
```

2.3. Data collection

The sampled countries, Belgium, the Czech Republic, Poland, and Switzerland, are selected for investigation because the spreading of COVID-19 in these countries is severe outbreak at the second wave in the manner of exponential outbreak. The total COVID-19 cases were only collected to model in the first stage of COVID-19 outbreak because the first stage is exponentially increasing. Then, the total COVID cases will pass the inflation point and will be converged to the carrying capacity (Areepong and Sunthornwat, 2021). For selected four countries, the outbreak in the first stage of the second wave started in the second wave between July and November, 2020. The data for this research are the number of daily total COVID-19 cases in Belgium, the Czech Republic, Poland, and Switzerland. The data is collected at the Worldometers website (Worldometer, 2020). This website reveals the real time data about world population, government and economics, society and media, environment, food, water, energy, health, as well as COVID-19 statistics. The duration of time for collecting data for making the forecasting model is dependent on the severity in each country. The time range for collection of data in Belgium is from July 15, 2020 ($t = 0$) to November 3, 2020 ($t = 111$). The time range for collection of da ta in Czech Republic is from August 23, 2020 ($t = 0$) to November 3, 2020 (t = 72). The time range for collection of data in Poland is from September 1, 2020 ($t = 0$) to November 3, 2020 ($t = 63$). The time range for collection of data in Switzerland is from September 2, 2020 ($t = 0$) to November 3, 2020 ($t = 62$). For the out of sample data, the time range is extended to 5 days from the last day of the data for making the forecasting model.

3. Results

The results of this research concern estimating the parameters and forecasting using the models, as well as the compound growth rate and spreading power of COVID-19 in Belgium, the Czech Republic, Poland, and Switzerland.

3.1. The parameters and forecasting models for the number of daily COVID-19 cases in Belgium

Here, we present the estimated parameters and forecasting models for the daily COVID-19 cases in Belgium. The estimated parameters evaluated by each method are as follows: $\hat{a} = 72207.129$, $\hat{b} = 316.821$, and $\hat{c} = 1.066$ by using the exponential least-squares method; $\hat{a} = 62757.544$, $\hat{b} = 114.456$, and $\hat{c} = 1.086$ by using the three selected points method; $\hat{a} = 67327.213$, $\hat{b} = 754.656$, and $\hat{c} = 1.058$ by using the partial-sums method; and $\hat{a} = 138260.53$ and $\hat{b} = -188328.83$ by using the hyperbolic least-squares method. The forecasting models based on the three methods and estimating the parameters of the daily COVID-19 cases in Belgium are shown in Figure 1.

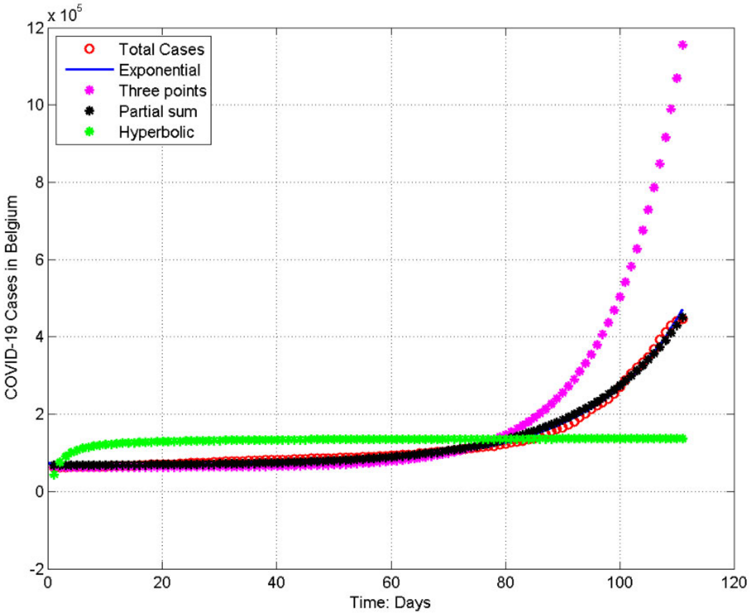


Figure 1. Estimation of the daily COVID-19 cases in Belgium

3.2. The parameters and forecasting models for the daily COVID-19 cases in the Czech Republic

Here, we present the estimated parameters and forecasting models for the daily COVID-19 cases in the Czech Republic. The estimated parameters evaluated by each method are as follows: $\hat{a} = 9077.568$, $\hat{b} = 9261.801$, and $\hat{c} = 1.053$ by using the least-squares method; $\hat{a} = 21701.501$, $\hat{b} = 221.499$, and $\hat{c} = 1.125$ by using the three selected points method; $\hat{a} = 3408.564$, $\hat{b} = 13198.966$, and $\hat{c} = 1.047$ by using the partial-sums method; and $\hat{a} = 125497.33$ and $\hat{b} = -235555.95$ by using the hyperbolic least-squares method. The forecasting models based on the three methods and estimates of the parameters for the daily COVID-19 cases in the Czech Republic are shown in Figure 2.

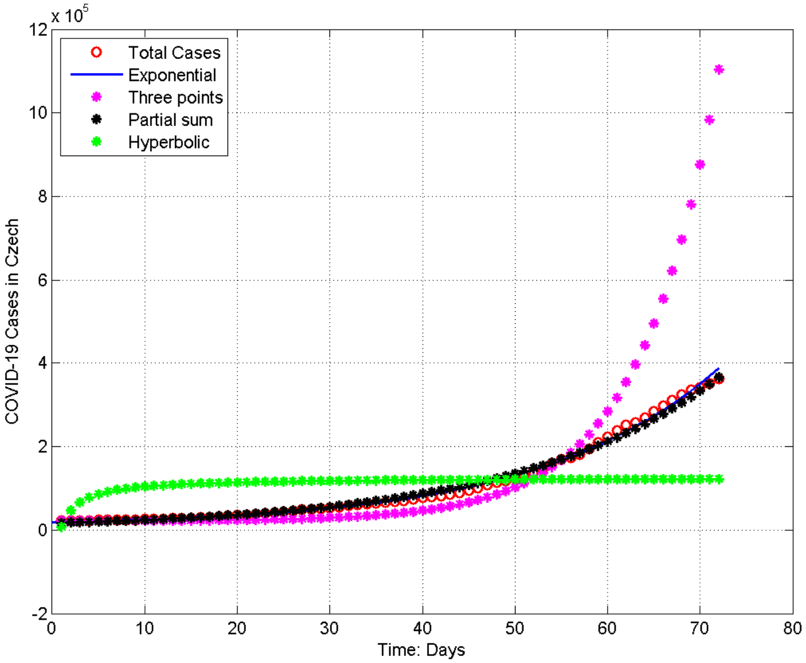


Figure 2. Estimation of the daily COVID-19 cases in the Czech Republic

3.3. The parameters and forecasting models for the daily COVID-19 cases in Poland

Here, we present the estimated parameters and forecasting models for the daily COVID-19 cases in Poland. The estimate parameters evaluated by the three methods are as follows: $\hat{a} = 64655.708$, $\hat{b} = 3192.084$, and $\hat{c} = 1.078$ by using the least-squares method; $\hat{a} = 47097.000$, $\hat{b} = 20825.000$, and $\hat{c} = 1.029$ by using the three selected points method; $\hat{a} = 67958.474$, $\hat{b} = 1649.921$, and $\hat{c} = 1.099$ by using the partial-sums method; and $\hat{a} = 157216.97$ and $\hat{b} = -198139.63$ by using the hyperbolic least-squares method. The forecasting models based on the three methods and the estimated parameters for the daily COVID-19 cases in Poland are shown in Figure 3.

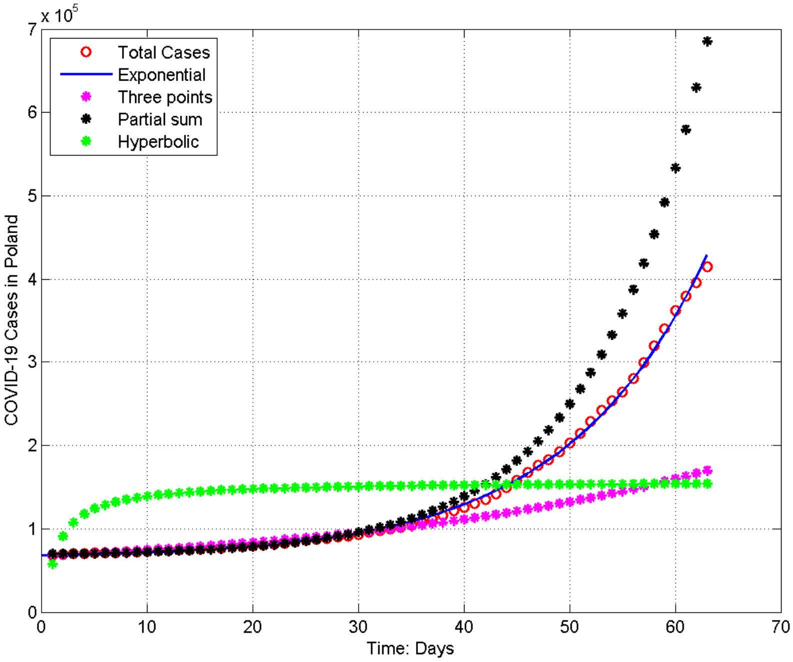


Figure 3. Estimation of the daily COVID-19 cases in Poland

3.4. Parameters and forecasting models for the daily COVID-19 cases in Switzerland

Here, we present the estimate parameters and forecasting models for the daily COVID-19 cases in Switzerland. The estimated parameters evaluated by each method are as follows: $\hat{a} = 44629.163$, $\hat{b} = 699.939$, and $\hat{c} = 1.089$ by using the least-squares method; $\hat{a} = 42258.5002$, $\hat{b} = 504.499$, and $\hat{c} = 1.119$ by using the three selected points method; $\hat{a} = 55456.458$, $\hat{b} = 901.478$, and $\hat{c} = 1.089$ by using the partial-sums method; and $\hat{a} = 77679.67$ and $\hat{b} = -75580.83$ by using the hyperbolic least-squares method.. The forecasting models based on the three methods and the estimated parameters for the daily COVID-19 cases in Switzerland are shown in Figure 4.

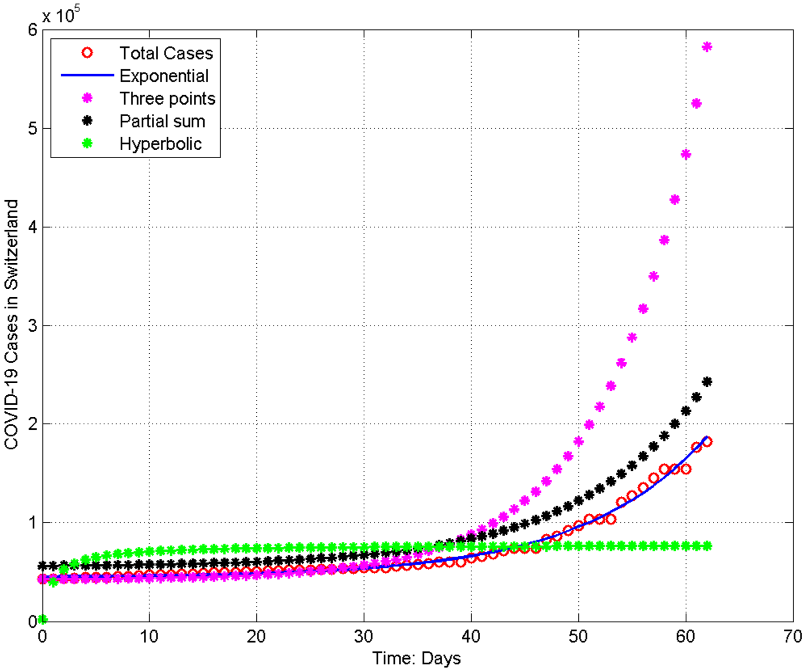


Figure 4. Estimation of the daily COVID-19 cases in Switzerland

3.5. Comparison of the spread of COVID-19 and appropriate parameters

The derivatives of the speed and spreading power of the daily COVID-19 cases in Belgium, the Czech Republic, Poland, and Switzerland are shown in Figures 5 and 6, respectively. In addition, the compound growth rate and forecasting model validation based on $RMSPE$ and R^2 values along with testing of the significance of the compound growth rate based on p-values are reported in Table 1. Moreover, forecasting of the daily COVID-19 cases for the out-of-sample data for November 4 – 8, 2020, and comparing the forecasted values with the actual values are summarized in Table 2. The results indicate that the increase in the speed and spreading power of the daily COVID-19 cases in Belgium was higher than for the other countries. Moreover, the compound growth rate for each country was statistically significant (p-value < 0.05). The exponential least-squares method provided the best fitting of the parameter estimations for the four countries, as indicated by the lowest $RMSPE$ and the highest R^2 values.

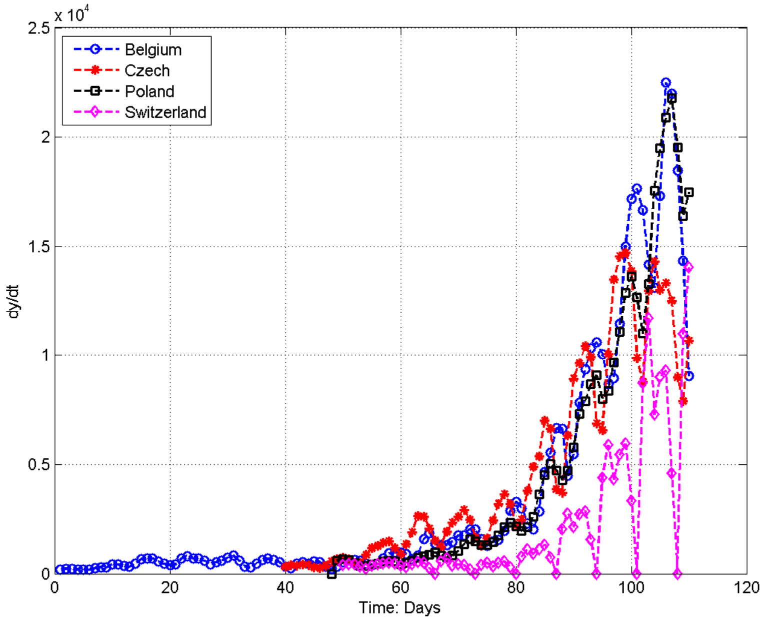


Figure 5. Estimated speed of the increase in daily COVID-19 cases

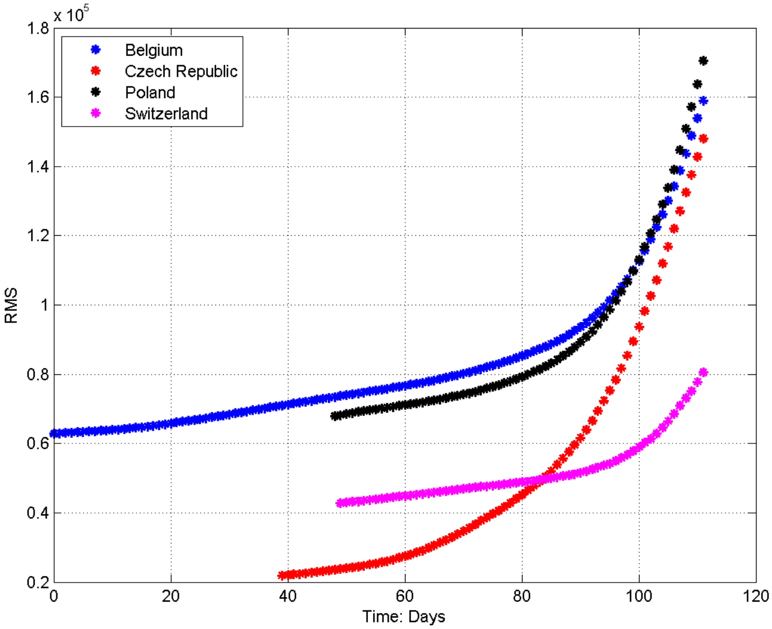


Figure 6. The spreading power for the daily COVID-19 cases

Table 1. The compound growth rate and forecasting model validation for the daily COVID-19 cases.

Country	Estimate B	Compound Growth Rate	Standard Error	t-test	p-value	*RMSPE*	R^2
Belgium	0.064	0.066	0.001	22.879	1.267e-43	0.456	0.982
						0.064	0.992
						0.063*	0.995*
						0.594	0.051
Czech Republic	0.052	0.053	3.613e-07	23.266	6.066e-35	0.543	0.867
						0.091	0.994
						0.058*	0.996*
						1.711	0.101
Poland	0.075	0.078	0.001	22.236	3.056e-31	0.248	0.904
						0.207	0.989
						0.020*	0.999*
						0.641	0.089
Switzer land	0.086	0.090	0.001	17.460	2.076e-25	0.735	0.978
						0.266	0.993
						0.034*	0.994*
						0.406	0.089

Note: * the best value. For each country, the top, middle, and bottom rows are for the models using the three points, partial-sums, exponential least-squares methods, and hyperbolic least-squares methods, respectively.

Table 2. Forecasting COVID-19 cases for the out-of-sample data using the least-squares method.

Country	Time					*RMSPE*	R^2
	Nov 4, 2020	Nov 5, 2020	Nov 6, 2020	Nov 7, 2020	Nov 8, 2020		
Belgium	453310	468213	479341	488044	494168	0.179	0.959
	497600.183	525854.870	555986.239	588118.939	622385.899		
Czech Republic	378717	391949	403497	411219	414827	0.134	0.943
	406855.462	427880.541	450016.929	473323.365	497861.694		
Poland	439536	466679	493765	521640	546425	0.061	0.997
	456977.627	487608.675	520631.283	556232.174	594612.651		
Switzerland	192376	202504	211913	211913	211913	0.137	0.735
	199713.233	213594.438	228718.115	245195.477	263147.688		

Note: For each country, the top line is the actual value and the bottom line is the forecasted value based on the exponential least-squares method.

4. Conclusions

In this research, we applied a modified exponential time series model to forecast daily COVID-19 cases. Belgium, the Czech Republic, Poland, and Switzerland were selected for this research because their curves for the daily COVID-19 cases in the second wave were exponentially increasing. Parameter estimation of the modified exponential time-series model was conducted by using the exponential least-squares method, the three selected points method, and the partial-sums methods. The hyperbolic least-squares time-series model, the other nonlinear model, which is a hyperbolic form, is applied to be compared with the previous models. The optimal forecasting model with the estimated parameters was selected based on having the lowest $RMSPE$ and the highest R^2 . Moreover, the compound growth rate, speed, and spreading power of the daily COVID-19 cases were evaluated and compared. The findings show that the exponential least-squares method was the most appropriate method for parameter estimation for the modified exponential time-series model for the daily COVID-19 cases in all four countries. The compound growth rates were statistically significant for each country, with that of Switzerland being slightly higher than in the other countries. Moreover, the speed and spreading power of the daily COVID-19 cases in Belgium were higher than the other countries. When applying the optimal least-squares model to predict the daily COVID-19 cases from the out-of-sample data, the forecasted daily COVID-19 cases were in good agreement with the actual values. Changing the parameters of the modified exponential time-series model made the forecasting model less accurate. Hence, the modified exponential time-series model is suitable for short-term forecasting, and parameter estimation should be evaluated again if the accuracy of the forecasting model is reduced. However, the limitation of this research is that the modified exponential time-series model is effectively used for the first stage of the outbreak because the total COVID-19 cases will exponentially increase in the first stage of the outbreak. Consequently, the total COVID-19 cases will pass the inflation point and converge to the carrying capacity. Future research will encompass other variables related to the COVID-19 situation, such as the number of active cases and the number of deaths, to enable the authorities effectively control a COVID-19 outbreak and protect the population from it.

Acknowledgement

The authors wish to thank Professor Dr. Yupaporn Areepong, Department of Applied Statistics, King Mongkut's University of Technology North Bangkok, Thailand, for support during the conduct of this research.

References

Areepong, Y., Sunthornwat, R., (2021). EWMA control chart based on its first hitting time and coronavirus alert levels for monitoring symmetric COVID-19 cases. *Asian Pacific Journal of Tropical Medicine*, 14(8), pp. 364–374.

Argaru, A. S., (2020). Linear regression model for predictions of COVID-19 New cases and new deaths based on May/June Data in Ethiopia, *Research Square*, doi:10.21203/rs.3.rs-61667/v1.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S. and Ciccozzi, M., (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset, *Data in Brief*, 29(105340), pp. 1–4.

Chaurasia, V., Pal, S., (2020). COVID-19 pandemic: ARIMA and regression model-based worldwide death cases predictions, *SN Computer Science*, 1(288), pp. 1–12.

Cheng, Y., (2020). Linear regression analysis of COVID-19 outbreak and control in Henan province caused by the output population from Wuhan. *medrxiv*, doi: 10.1101/2020.05.03.20089193.

Das, B., Chakrabarty, D., (2017). Representation of numerical data by exponential curve. *Journal of Mathematics and Systems Sciences*, 13(1-2), pp. 1–6.

Ekum, M. and Ogunsanya, A., (2020). Application of Hierarchical Polynomial Regression Models to Predict Transmission of COVID-19 at Global Level. *International Journal of Clinical Biostatistics and Biometrics*, 6(1), pp. 1–18.

Fang, L., Wang, D. and Pan, G., (2020). Analysis and Estimation of COVID-19 Spreading in Russia Based on ARIMA Model. *SN Comprehensive Clinical Medicine*, doi: 10.1007/s42399-020-00555-y.

George, E. P. B., Gwilym M. J., Gregory, C. R. and Greta, M. L., (2015). *Time Series Analysis: Forecasting and Control.* 5th ed. Hoboken, New Jersey: John Wiley & Sons, Inc.

Ghosal, S., Sengupta, S., Majumder, M. and Sinha, B., (2020). Linear regression analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020). *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 4(4), pp.311–315.

Hoseinpour, D. A., Alizadeh, M., Derakhshan, P., Babazadeh, P. and Jahandideh, A., (2020). Understanding epidemic data and statistics: A case study of COVID-19. *Journal of Medical Virology*, 92(7), pp. 868–882.

Ikaya, K. K. Z., Balcióglu, M. S., Yolcu, H. I. and Karabag, K., (2005). The application of exponential method in the analysis of growth curve for Japanese quail. *European Poultry Science*, 69(5), pp. 193–198.

Jones, A. R., (2019). *Probability, statistics and other frightening stuff*. 2nd ed. Abingdon, UK: Routledge.

Kharab, A., Guenther, B., (2012). *An introduction to numerical methods: A MATLAB approach*. 3rd dd. Boca Raton, FL, USA: CRC Press.

Likhesh, K., Yogesh, P., Puja, K. and Vishal, J., (2020). Prediction of coronavirus Covid-19 cases using linear regression and support vector machine. *International Journal of Advanced Science and Technology*, 29(5s), pp.1911–1924.

Melik-Huseynov, D. V., Karyakin, N. N., Blagonravova, A. S., Klimko, V. I., Bavrina, A. P., Drugova, O. V., Saperkin, N. V. and Kovalishena, O. V., (2020). Regression models predicting the number of deaths from the new coronavirus infection. *Modern Technologies in Medicine*, 12(2), pp.6–13.

Pandey, G., Chaudhary, P., Gupta, R. and Pal, S., (2020). SEIR and regression model based COVID-19 outbreak predictions in India. *medrxiv*, doi: 10.1101/2020.04.01.20049825.

Rath, S., Tripathy, A. and Tripathy, A., (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(5), pp. 1467–1474.

Rustam, F., Reshi, A., Mehmood, A., Ullah, S., On, B., Aslam, W. and Choi, G. S., (2020). COVID-19 future forecasting using supervised machine learning models. *IEEE Access*, 8(2020), pp. 101489–101499.

Sahai, A. K., Rath, N., Sood, V. and Singh, M. P., (2020). ARIMA modelling and forecasting of COVID-19 in top five affected countries. *Diabetes and Metabolic Syndrome*, 14(5), pp. 1419–1427.

Singh, R. K., Rani, M., Bhagavathula, A. S., Sah, R., Rodriguez-Morales, A. J., Kalita, H., Nanda, C., Sharma, S., Sharma, Y. D., Rabaan, A. A., Rahmani, J. and Kumar, P., (2020). Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced autoregressive integrated moving average (ARIMA) model. *JMIR Public Health and Surveillance*, 6(2), pp. 1–10.

Suleiman, A. A., Suleiman, A., Abdullahi, U. A. and Suleiman, S. A., (2020). Estimation of the case fatality rate of COVID-19 epidemiological data in Nigeria using statistical regression analysis. *Biosafety and Health*, 3(1) pp. 4–7, doi:10.1016/j.bsheal.2020.09.003.

WORLDOMETER, (2020). *COVID-19 coronavirus pandemic*, [online]. Available at: <https://www.worldometers.info/coronavirus/country> [Accessed 12 November 2020].

Yadav, M., Perumal, M. and Srinivas, M., (2020). Analysis on novel coronavirus (COVID-19) using machine learning methods. *Chaos, Solitons and Fractals*, 139 (110050), pp. 1–12.

Yang, Q., Wang, J., Ma, H. and Wang, X., (2020). Research on COVID-19 based on ARIMA model—taking Hubei, China as an example to see the epidemic in Italy. *Journal of Infection and Public Health*, 13(10), pp.1415–1418.

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 167–184, DOI 10.2478/stattrans-2022-036
Received – 11.02.2021; accepted – 14.12.2021

Poisson area-biased Ailamujia Distribution and its applications in environmental and medical sciences

Ahmad Aijaz[1], S. Qurat ul Ain[2], Ahmad Afaq[3], Rajnee Tripathi[4]

ABSTRACT

In this paper, a new Poisson area-biased Ailamujia distribution has been formulated to analyse count data. It was created by combining two distributions: the Poisson and area-biased Ailamujia distributions, using the compounding technique. Several distributional properties of the formulated distribution were studied. Its ageing characteristics were determined and expressed explicitly. A variety of diagrams were used to demonstrate the characteristics of the probability mass function (pmf) and the cumulative distribution function (cdf). The parameter of the developed model was estimated by employing the maximum likelihood estimation approach. Finally, two data sets were used to demonstrate the effectiveness of the investigated distribution.

Key words: compound technique, Poisson distribution, area-biased Ailamujia distribution, reliability analysis, order statistics, maximum likelihood estimator.

Mathematics subject classification: 60E05, 62E15.

1. Introduction

In probability distributions, discrete distributions are very essential. Researches are focused extensively in past years to build new discrete models for assessing count data. There are a variety of procedures for developing new distributions in the statistics literature. Extensions to classical distributions can be made by adding additional parameters to them. Transmutation, discretization of continuous distributions, Marshall-Olkin method, compounding, and other approaches were examples. Classical distributions frequently fail to offer an acceptable fit to observable data. This became imperative for researchers to investigate new probability models in order to overcome

[1] Corresponding author's. Department of Mathematics, Bhagwant University, Ajmer, Rajasthan, India. E-mail: ahmadaijaz4488@gmail.com.

[2] Department of Mathematics, Bhagwant University, Ajmer, Rajasthan, India.

[3] Department of Mathematical Sciences, Islamic University of Science & Technology, Awantipora, Kashmir.

[4] Department of Mathematics, Bhagwant University, Ajmer, Rajasthan, India.

the drawbacks of classical distributions. The compounding of distributions has attracted the attention of researchers over the last decade. The compounding approach is most commonly used when the parameter of one distribution is a random variable that follows another distribution, as in the case of count data. The compounding of distributions occurs when two separate distributions are combined. It makes no odds whether they are discrete or continuous in character. Based upon parent distribution, the resultant distribution from compounding may be continuous or discrete.

The concept of weighted models can be traced back from Fisher (1935). Later on weighted models were briefly discussed by C.R. Rao (1964), when sample observations have an unequal probability of choosing. Thus, in such situation we add weights to the distribution to model bias.

Suppose Y denotes random variable with pmf $p(y)$, then pmf of weighted variable Y_w is defined by

$$P(y;\theta)=\frac{w(y)f(y)}{E[w(y)]};y>0$$

where $w(y)=y^k$ is a non–negative weight function. For $k=2$ we get area-biased distributions.

In this study, we have used compounding approach to create a new distribution by combining Poisson and area-biased Ailamujia distribution. The newly established distribution is called "Poisson area-biased Ailamujia distribution". Compounding distributions have extensive applications in several sectors of research such as biomedicine, insurance, engineering, and communications, among others. Researchers in this field have worked extensively, and they have made significant contributions to compounding research that has been tracked back to 1920. The inception of compounding models has been traced from Greenwood and Yule (1920). Sankaran (1970), Gerstenkorn (1993,1996), Mahmodi et al. (2010), Zamani and Ismail (2010), Gupta and Ong (2004), Shanker (2017), Shi(2012), Subhradev sen (2018), Giovani Carrara Rodrigues et al. (2018), Shanker et al. (2019), This study proposes a novel probability model known as the Poisson area-biased Ailamujia distribution, which is derived via the compounding process, and discusses its many mathematical aspects.

2. Definition of Poisson Area-Biased Ailamujia Distribution

Consider a random variable Y follows Poisson distribution i:e $Y\sim P(\lambda)$ and assume that the parameter of $P(\lambda)$ follows area-biased Ailamujia distribution with parameter θ. The distribution obtained by compounding Poisson with area-biased Ailamujia distribution follows a discrete distribution whose probability mass function is denoted as PABAD (Y,θ). The probability function of the obtained model PABAD (θ) is given by the following theorem.

Theorem 2.1. The probability mass function of a discrete Poisson area-biased Ailamujia distribution PABAD(Y,θ) is given as

$$P(Y=y)=\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4\frac{(y+1)(y+2)(y+3)}{(2\theta+y)^y}\quad ;y=0,1,2..\theta>0$$

Proof: The probability mass function of the discrete Poisson area-biased Ailamujia distribution PABAD(Y,θ) may be obtained as

If $Y\sim P(\lambda)$, the probability mass function (pmf) of the Poisson distribution is given by

$$f(Y|\lambda)=\frac{e^{-\lambda}\lambda^y}{y!};y=0,1,2,...;\lambda>0$$

As the parameter λ follows area-biased Ailamujia distribution with probability density function (pdf)

$$g(\lambda;\theta)=\frac{(2\theta)^4}{6}\lambda^3e^{-2\theta\lambda};\lambda>0,\theta>0$$

We have

$$P(Y=y)=\int_0^\infty f(Y|\lambda).g(\lambda;\theta)d\lambda$$

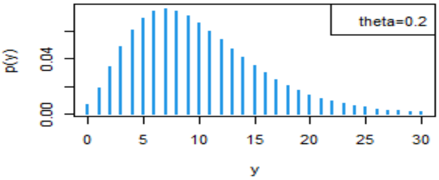
$$=\int_0^\infty\frac{e^{-\lambda}\lambda^y}{y!}\frac{(2\theta)^4}{6}\lambda^3e^{-2\theta\lambda}d\lambda$$

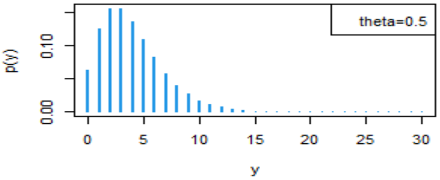
$$=\frac{(2\theta)^4}{6y!}\int_0^\infty\lambda^{y+3}e^{-(2\theta+1)\lambda}d\lambda$$

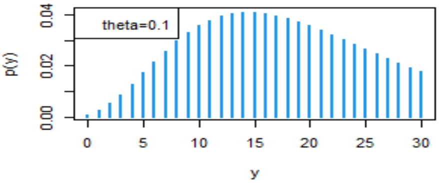
$$=\frac{(2\theta)^4}{6y!}\frac{(y+3)!}{(2\theta+1)^{y+4}}$$

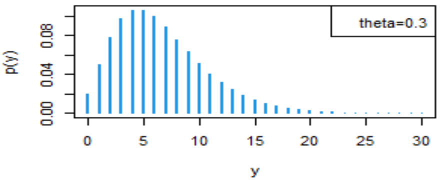
$$=\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4\frac{(y+1)(y+2)(y+3)}{(2\theta+1)^y}\quad ;y=0,1,2,...;\theta>0\qquad(2.1)$$

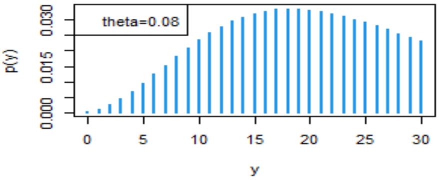
The following six graphs illustrate the behaviour of pmf of the Poisson area-biased Ailamujia distribution for different values of parameter

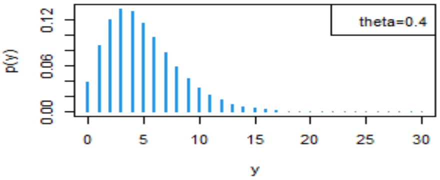












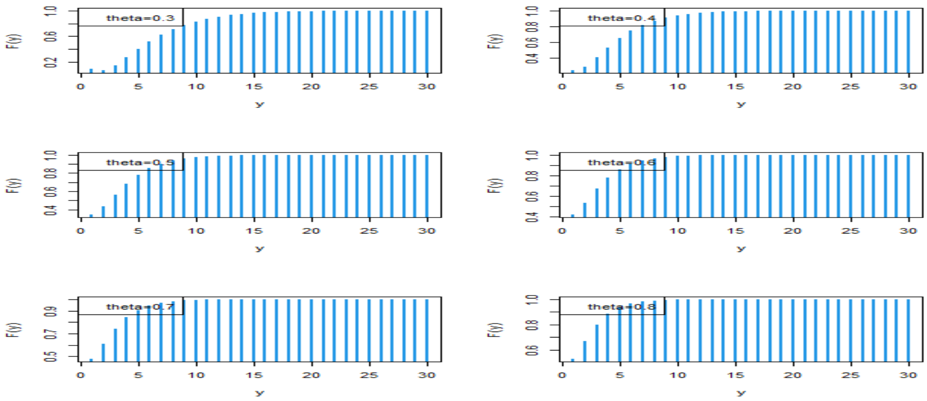
The corresponding cumulative distribution function (cdf) of the discrete Poisson area-biased Ailamujia distribution is given as

$$F(Y=y)=p_r(Y\leq y)=1-p_r(Y>y)$$

$$=1-\sum_{w=y+1}^{\infty}P(w)$$

$$=1-\frac{\left\{\begin{array}{l}8\theta^3y^3+4\theta^2(7\theta+3)y^2+\left(208\theta^3+60\theta^2+24\theta+12\right)y\\+\left(52\theta^2+16\theta+6\right)(2\theta+1)\end{array}\right\}}{6(2\theta+1)^{y+4}}\quad ;y=0,1,2,..;\theta>0 \qquad (2.2)$$

The following six graphs illustrate the behaviour of cdf of the Poisson area-biased Ailamujia distribution for different values of parameter



3. Statistical Measures of Poisson Area-Biased Ailamujia Distribution

In this section several statistical measures of the Poisson area-biased Ailamujia distribution has been studied. They include are moments, moment generating function (mgf) and probability generation function (pgf).

3.1. Moments of Poisson Area-Biased Ailamujia Distribution.

The r^{th} factorial moment of the Poisson area-biased Ailamujia distribution is denoted as $\mu_{(r)}'$ and can be obtained by

$$\mu_{(r)}' = E\left[E\left(Y^{(r)}\middle|\lambda\right)\right], \text{ where } Y^{(r)} = Y(Y-1)(Y-2)\ldots(Y-r+1)$$

$$= \frac{(2\theta)^4}{6}\int_0^\infty\left[\sum_{y=0}^{\infty} y^{(r)}\frac{e^{-\lambda}\lambda^y}{y!}\right]\lambda^3 e^{-2\theta\lambda}d\lambda$$

$$= \frac{(2\theta)^4}{6}\int_0^\infty\left[\lambda^r\sum_{y=r}^{\infty}\frac{e^{-\lambda}\lambda^{y-r}}{(y-r)!}\right]\lambda^3 e^{-2\theta\lambda}d\lambda$$

Taking $y+r$ in place of y within the bracket, we get

$$\mu_{(r)}' = \frac{(2\theta)^4}{6}\int_0^\infty\left[\lambda^r\left(\sum_{y=0}^{\infty}\frac{e^{-\lambda}\lambda^y}{y!}\right)\right]\lambda^3 e^{-2\theta\lambda}d\lambda$$

$$\mu_{(r)}{}' = \frac{(2\theta)^4}{6} \int_0^\infty \lambda^{3+r} e^{-2\theta\lambda} d\lambda$$

$$= \frac{(2\theta)^4}{6} \frac{\Gamma(r+4)}{(2\theta)^{r+4}} = \frac{(r+3)!}{6(2\theta)^r} \tag{3.1}$$

Substituting $r = 1,2,3,4$ in (3.1), the first four factorial moments can be obtained, and using the relationship between factorial moments and moments about origin, the first four moments about origin of the PABAD (2.1) are obtained as

$$\mu_1{}' = \frac{2}{\theta}, \qquad \mu_2{}' = \frac{5+2\theta}{\theta^2},$$

$$\mu_3{}' = \frac{2\theta^2 + 15\theta + 15}{\theta^3}, \qquad \mu_4{}' = \frac{4\theta^3 + 70\theta^2 + 180\theta + 105}{2\theta^4}.$$

The moments about mean of the Poisson area-biased Ailamujia distribution are obtained by using the relationship between moments about mean and moments about origin

$$\mu_2 = \frac{2\theta+1}{\theta^2}$$

$$\mu_3 = \frac{2\theta^3 + 3\theta + 1}{\theta^3}$$

$$\mu_4 = \frac{-(28\theta^3 - 70\theta^2 - 36\theta - 9)}{2\theta^4}$$

The coefficient of variation (C.V), coefficient of skewness $\left(\sqrt{\beta_1}\right)$, coefficient of kurtosis (β_2), index of dispersion (γ) of the Poisson area-biased Ailamujia distribution are determined as

$$C.V = \frac{\sigma}{\mu_1^{'}} = \frac{\sqrt{1+2\theta}}{2}$$

$$\sqrt{\beta_1} = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}} = \frac{(2\theta^3 + 3\theta + 1)}{(5+2\theta)^3}$$

$$\beta_2 = \frac{\mu_4}{(\mu_2)} = \frac{-(28\theta^3 - 70\theta^2 - 36\theta - 9)}{2(5+2\theta)^2}$$

$$\gamma = \frac{\sigma^2}{\mu_1^{'}} = \frac{2\theta+1}{2\theta}$$

Table1. The numerical values of the mean, variance, skewness, kurtosis, coefficient of variation and index of dispersion for some values of parameter θ

θ	μ	σ^2	$\sqrt{\beta_1}$	β_2	C.V	γ
0.5	4.00	8.000	0.012	0.569	0.707	2.000
0.6	3.333	6.111	0.013	0.647	0.741	1.833
0.7	2.857	4.897	0.014	0.718	0.774	1.714
0.8	2.500	4.062	0.015	0.783	0.806	1.625
0.9	2.222	3.456	0.016	0.840	0.689	1.555
1	2.00	3.000	0.017	0.887	0.836	1.500
2	1.00	1.250	0.031	0.845	0.866	1.250
3	0.666	0.777	0.048	-0.037	1.118	1.166
4	0.500	0.562	0.064	-1.535	1.322	1.125
5	0.400	0.440	0.078	-3.468	1.658	1.100

3.2. Generating Functions (pgf, mgf, ch.f) of Poisson Area-Biased Ailamujia Distribution

In this section we study pgf, mgf and characteristics function (ch.f) of the Poisson area-biased Ailamujia distribution.

Theorem.3.2.1. If $Y \sim PABAD(\theta)$then the probability generating function $P_Y(t)$ is

$$p_Y(t)=\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4\left\{\frac{6t^4}{(2\theta+1-t)^4}+\frac{12t^3}{(2\theta+1-t)^3}+\frac{11t^2}{(2\theta+1-t)^2}+\frac{18t}{(2\theta+1-t)}+6\right\}$$

Proof: The probability generating function (pgf) of the Poisson area-biased Ailamujia distribution is defined as

$$P_Y(t)=E(t)=\sum_{y=0}^{\infty}t^y P(y)$$

$$=\sum_{y=0}^{\infty}\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4\frac{\left(y^3+6y^2+11y+6\right)}{(2\theta+1)^y}t^y$$

$$=\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4\sum_{y=0}^{\infty}\left(y^3+6y^2+11y+6\right)\left(\frac{t}{2\theta+1}\right)^y$$

$$=\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4\left\{\sum_{y=0}^{\infty}y^3\left(\frac{t}{2\theta+1}\right)^y+6\sum_{y=0}^{\infty}y^2\left(\frac{t}{2\theta+1}\right)^y+11\sum_{y=0}^{\infty}y\left(\frac{t}{2\theta+1}\right)^y+6\sum_{y=0}^{\infty}\left(\frac{t}{2\theta+1}\right)^y\right\}$$

$$=\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4\left\{\frac{(2\theta+1)\left(t^3+4t^2+t\right)}{(2\theta+1-t)^4}+\frac{t^2(2\theta+1)+t(2\theta+1)^2}{(2\theta+1-t)^3}+\frac{(2\theta+1)t}{(2\theta+1-t)^2}+\frac{(2\theta+1)}{(2\theta+1-t)}\right\}$$

$$=\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4\left\{\frac{6t^4}{(2\theta+1-t)^4}+\frac{12t^3}{(2\theta+1-t)^3}+\frac{11t^2}{(2\theta+1-t)^2}+\frac{18t}{(2\theta+1-t)}+6\right\}$$

Theorem 3.2.2. If $Y \sim PABAD(\theta)$ then the moment generating function $M_Y(t)$ is

$$M_Y(t)=\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4\left\{\frac{6e^{4t}}{\left(2\theta+1-e^t\right)^4}+\frac{12e^{3t}}{\left(2\theta+1-e^t\right)^3}+\frac{11e^{2t}}{\left(2\theta+1-e^t\right)^2}+\frac{18e^t}{\left(2\theta+1-e^t\right)}+6\right\}$$

Proof: Since the moment generating function is a generalization of the probability generating function with the relationship given as

$$M_Y(t)=P_Y\left(e^t\right)$$

So that

$$M_Y(t)=\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4\left\{\frac{6e^{4t}}{\left(2\theta+1-e^t\right)^4}+\frac{12e^{3t}}{\left(2\theta+1-e^t\right)^3}+\frac{11e^{2t}}{\left(2\theta+1-e^t\right)^2}+\frac{18e^t}{\left(2\theta+1-e^t\right)}+6\right\}$$

Similarly, the relationship between mgf and ch.f is defined as

$$M_Y(t)=\phi_Y(it)$$

$$\phi_Y(it)=\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4\left\{\frac{6e^{4it}}{\left(2\theta+1-e^{it}\right)^4}+\frac{12e^{3it}}{\left(2\theta+1-e^{it}\right)^3}+\frac{11e^{2it}}{\left(2\theta+1-e^{it}\right)^2}+\frac{18e^{it}}{\left(2\theta+1-e^{it}\right)}+6\right\}$$

4. Reliability Measures of Poisson Area-Biased Ailamujia Distribution

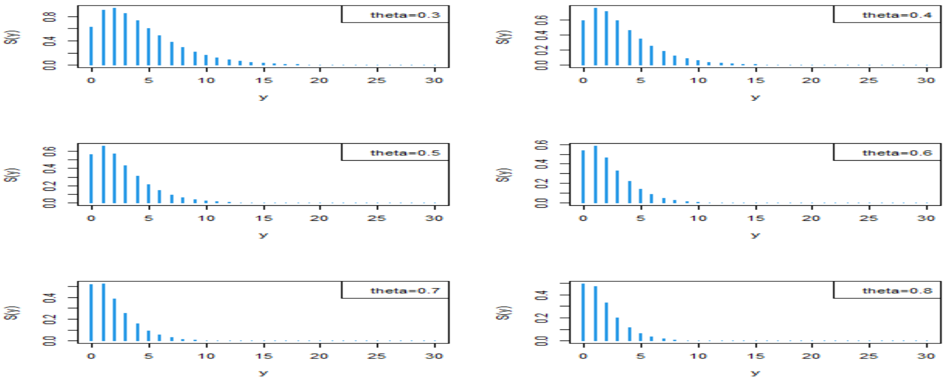
The reliability of the majority of the system decreases with time. So, the chance that a device that is operating until period "t" would fail after that period is referred to as the device's reliability. Suppose Y is a continuous random variable with cdf $F(y)$; $y>0$. Then its reliability function, which is also called survival function, is defined as

$$S(y)=p_r(Y>y)=\int_y^\infty f(y)dy=1-F(y)$$

The survival function of the discrete Poisson area-biased Ailamujia distribution is given as

$$R(y,\theta)=S(y,\theta)=\frac{\left\{\begin{matrix}8\theta^3y^3+4\theta^2(7\theta+3)y^2+\left(208\theta^3+60\theta^2+24\theta+12\right)y\\+\left(52\theta^2+16\theta+6\right)(2\theta+1)\end{matrix}\right\}}{6(2\theta+1)^{y+4}} \tag{4.1}$$

The following six graphs show the behaviour of the survival function of the Poisson area-biased Ailamujia distribution for different values of parameter.



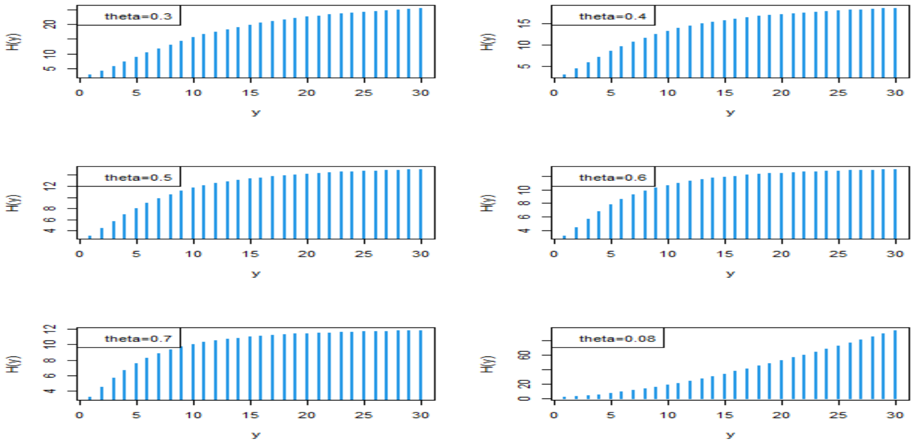
The hazard rate function is described as an indicator of the system's proclivity to collapse within a certain time interval. The hazard rate function of a random variable y is given as

$$H(y,\theta)=\frac{f(y,\theta)}{S(y,\theta)} \tag{4.2}$$

Substituting (2.2) and (4.1) into (4.2), we get

$$H(y,\theta)=\frac{(2\theta)^4\left(y^3+6y^2+11y+6\right)}{\left\{\begin{matrix}8\theta^3y^3+4\theta^2(7\theta+3)y^2+\left(208\theta^3+60\theta^2+24\theta+12\right)y\\+\left(52\theta^2+16\theta+6\right)(2\theta+1)\end{matrix}\right\}}$$

The following six graphs show the behaviour of the hazard function of the Poisson area-biased Ailamujia distribution for different values of parameter.

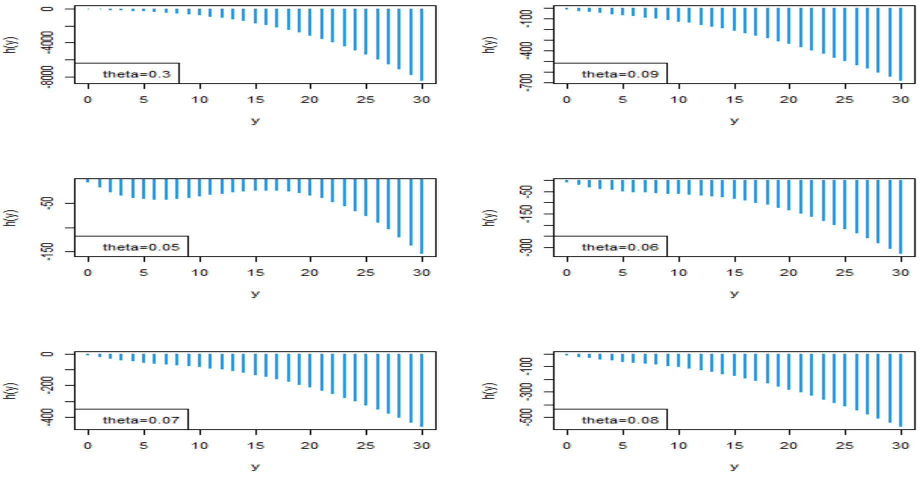


The reverse hazard rate function denoted as h_r is given by

$$h_r(y,\theta)=\frac{f(y,\theta)}{F(y,\theta)}$$

$$h_r(y,\theta)=\frac{(2\theta)^4\left(y^3+6y^2+11y+6\right)}{6(2\theta+1)^{y+4}-\left\{\begin{array}{l}8\theta^3y^3+4\theta^2(7\theta+3)y^2+\left(208\theta^3+60\theta^2+240\theta+12\right)y\\+\left(52\theta^2+160\theta+6\right)(2\theta+1)\end{array}\right\}}$$

The following six graphs shows the behaviour of reverse hazard function of the Poisson area-biased Ailamujia distribution for different values of parameter



5. Recurrence Relation of Poisson Area-Biased Ailamujia Distribution.

If $Y \sim PABAD(\theta)$ then probability mass function of Y is

$$P(Y=y)=\frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4 \frac{(y+1)(y+2)(y+3)}{(2\theta+1)^y} \quad ; y=0,1,2,\ldots,; \theta>0$$

The recurrence relation of the Poisson area-biased Ailamujia distribution is given by

$$\frac{P(y+1)}{P(y)}=\frac{(y+2)(y+3)(y+4)}{(y+1)(y+2)(y+3)}\left\{\frac{(2\theta+1)^y}{(2\theta+1)^{y+1}}\right\}$$

$$P(y+1)=\frac{(y+4)}{(y+1)}\frac{1}{(2\theta+1)}P(y)$$

This represents the recurrence relation.

6. Method of Estimation

6.1. Method of Moments (MOM)

Suppose $y_1, y_2, \ldots, y_n$ denotes a random sample of size n from the Poisson area-biased Ailamujia distribution. Now, to obtain sample moments, we replace population moments with sample moments.

$$\mu_1' = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{y}=\frac{2}{\theta} \Rightarrow \hat{\theta}=\frac{2}{\bar{y}}$$

Theorem 6.1.1. The MOM estimator $\hat{\theta}$ of θ is positively biased.

Proof: Let us suppose $\hat{\theta}=\varphi(\bar{y})$, where $\varphi(u)=\frac{2}{u}, u>0$ so that $\varphi''(u)=\frac{4}{u^3}>0$

Then, $\varphi(u)$ is strictly convex. Hence by Jensen's inequality, we have

$$E(\varphi(\bar{u}))>\varphi(E(\bar{u}))$$

Thus,

$$\varphi(E(\bar{u}))=\varphi(\mu)=\varphi\left(\frac{2}{\theta}\right)=\theta$$

We obtain

$$E(\hat{\theta})>\theta$$

Theorem 6.1.2. The MOM estimator $\hat{\theta}$ of θ is consistent and asymptotically normal

$$\sqrt{n}\left(\hat{\theta}-\theta\right)\xrightarrow{d}N\left(0,\upsilon^2(n)\right)$$

where $\upsilon^2(n)=\dfrac{\theta^2(2\theta+1)}{4}$

Proof: Consistency: since $\mu<\infty$, then $\bar{Y}\xrightarrow{p}\mu$. Also, since $\varphi(\mu)$ is a continuous function at $u=\mu$, then $\varphi(\bar{Y})\rightarrow\varphi(\mu)$, $\hat{\theta}\xrightarrow{p}\theta$.

Asymptotically normality:

As $\sigma^2<\infty$, then by applying central limit theorem we have

$$\sqrt{n}\left(\bar{Y}-\mu\right)\xrightarrow{d}N\left(0,\sigma^2\right)$$

since $\varphi(\mu)$ is differentiable function and $\varphi'(\mu)\neq 0$. Then, by applying the delta method we have

$$\sqrt{n}\left(\varphi(\bar{Y})-\varphi(\mu)\right)\xrightarrow{d}N\left(0,[\varphi'(\mu)]^2\sigma^2\right)$$

Finally, we have

$$\varphi(\bar{Y})=\hat{\theta}\ ,\ \varphi(\mu)=\theta$$

and

$$\varphi'(\mu)=\frac{-2}{\mu^2}=-\frac{\theta^2}{2}$$

The theorem follows, as a result of this asymptotic $100(1-\alpha)\%$ the confidence interval for θ is given as

$$\hat{\theta}\pm Z_{\frac{\alpha}{2}}\frac{v(\hat{\theta})}{\sqrt{n}}$$

where $Z_{\frac{\alpha}{2}}$ denotes the $\left(1-\dfrac{\alpha}{2}\right)$ percentile of the standard normal distribution.

6.2. Maximum Likelihood Estimation of Poisson Area-Biased Ailamujia Distribution.

Let $y_1, y_2, \ldots, y_n$ denote a random sample of size n from the Poisson area-biased Ailamujia distribution. Then, its likelihood function is given by

$$l = \prod_{i=1}^{n} p(y_i, \theta)$$

$$= \prod_{i=1}^{n} \frac{1}{6}\left(\frac{2\theta}{2\theta+1}\right)^4 \frac{\left(y_i^3 + 6y_i^2 + 11y_i + 6\right)}{(2\theta+1)^{y_i}}$$

$$= \left(\frac{1}{6}\right)^n \left(\frac{2\theta}{2\theta+1}\right)^n \prod_{i=1}^{n} \frac{\left(y_i^3 + 6y_i^2 + 11y_i + 6\right)}{(2\theta+1)^{y_i}}$$

The log likelihood function is obtained as

$$\log l = -n\log(6) + n\log\left(\frac{2\theta}{2\theta+1}\right) + \sum_{i=1}^{n} \log\left(y_i^3 + 6y_i^2 + 11y_i + 6\right) - \sum_{i=1}^{n} \log(2\theta+1)y_i$$

Differentiate w.r.t to θ, we get

$$\frac{\partial \log l}{\partial \theta} = \frac{4n}{\theta} - \frac{8n}{(2\theta+1)} - 2\sum_{i=1}^{n} \frac{y_i}{(2\theta+1)}$$

Substituting $\frac{\partial \log l}{\partial \theta} = 0$, we get the required $\hat{\theta}_{mle}$

$$\hat{\theta} = \frac{2}{\bar{y}}$$

7. Application to real data sets

In this section the goodness of fit of area-biased Poisson Ailamujia distribution (PABAD) has been proposed for two real count data sets. And we show that the established distribution perform better than size-biased Poisson Ailamujia distribution (PSBAD), Poisson Ailamujia distribution (PAD) and Poisson distribution (PD), Poisson Lindley distribution (PLD) and Poisson Shunkar distribution (PSD).

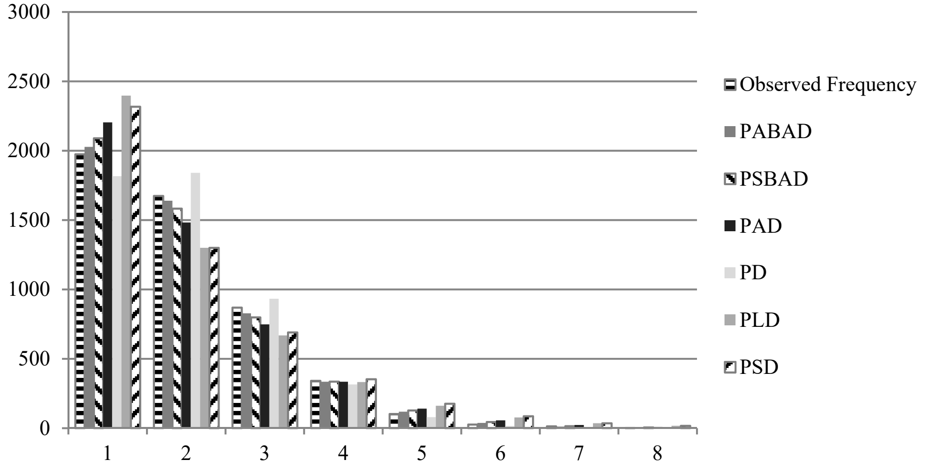
Data set 1: The first data set represents the number of micronuclei after exposure at dose 4 Gy of γ radiation, counted using the cytochalasim B method and available in reference (10).

In order to compare the above distribution models, we consider the criteria like AIC (Akaike Information criterion), AICC (corrected Akaike information criterion), BIC (Bayesian information criterion). Among the above distributions, the better distribution is considered to have lesser values of AIC, AICC.

Table 7.1. Number of micronuclei

Number of micronuclei	Observed frequency	Expected frequency					
		PABAD	**PSBAD**	**PAD**	**PD**	**PLD**	**PSD**
0	1974	2027.28	2089.44	2203.66	1816.07	2396.79	2316.91
1	1674	1638.94	1582.51	1481.90	1839.91	1300.33	1299.64
2	869	828.12	799.09	747.49	932.13	668.82	689.95
3	342	334.74	336.23	335.13	314.81	332.15	353.33
4	102	118.39	127.33	140.85	79.74	160.91	176.42
5	26	38.29	45.01	56.84	16.16	76.53	86.44
6	13	11.61	15.15	22.30	2.73	35.88	35.88
7	2	3.35	4.92	8.57	0.40	16.63	16.63
Total	5002	5000.72	4999.76	4999.84	5001.12	4988.04	4975.2
ML estimates (Standard Error)		1.9739 (0.031)	1.4804 (0.024)	0.9869 (0.0170)	1.0131 (0.014)	1.3873 (0.022)	1.3197 (0.018)
$-\log L$		6740.37	6752.8	6794.42	6767.91	6918.36	6931.20
AIC		13482.3	13507.2	13590.2	13537.2	13836.72	13864.1
AICC		13482.6	13507.8	13590.8	13537.82	13838.73	13864.41
BIC		13489.5	13514.3	13597.4	13544.4	13845.25	13870.2
χ^2		11.25	32.98	105.07	92.73	281.35	273.89
df		5	5	7	4	7	7
p-value		0.1280	$2.6*10^{-5}$	$9.6*10^{-20}$	$3.3*10^{-17}$	$3.4*10^{-66}$	$2.4*10^{-62}$

The following histogram represents the number of micronuclei for the proposed model when compared with other models.

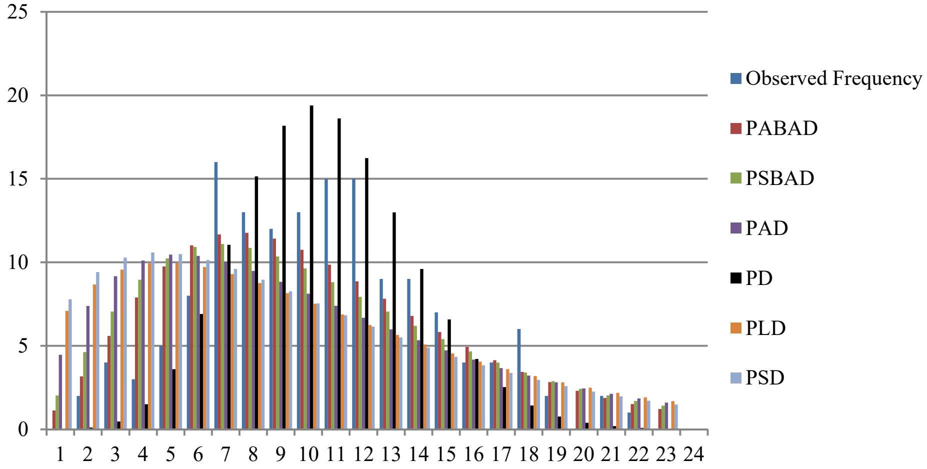


Data set 2: Data on the macroscopic fresh-water fauna in dredge samples from the bottom of water ber Lake is due to Juday (1942) and Thomas (1949).

Table 7.2. Microcalanus Nauplii

Microcalnus Nauplii	Observed frequency	Expected frequency					
		PABAD	**PSBAD**	**PAD**	**PD**	**PLD**	**PSD**
0	0	1.13	2.03	4.46	0.02	7.09	7.79
1	2	3.17	4.63	7.38	0.10	8.67	9.4
2	4	5.60	7.06	9.16	0.47	9.56	10.30
3	3	7.81	8.96	10.11	1.50	9.94	10.58
4	5	9.76	10.24	10.46	3.60	9.97	10.49
5	8	11.07	10.92	10.39	6.10	9.71	10.13
6	16	11.67	11.09	10.03	11.05	9.29	9.60
7	13	11.76	10.86	9.49	15.15	8.75	8.96
8	12	11.42	10.35	8.83	18.18	8.15	8.26
9	13	10.75	9.64	8.12	19.39	7.57	7.54
10	15	9.86	8.81	7.40	18.62	6.88	6.83
11	15	8.86	7.92	6.68	16.24	6.25	6.14
12	9	7.82	7.05	5.99	12.99	5.65	5.50
13	9	6.71	6.20	5.34	9.60	5.07	4.90
14	7	5.82	5.40	4.79	6.59	4.54	4.34
15	4	4.93	4.66	4.18	4.21	4.05	3.83
16	4	4.13	3.99	3.68	2.53	3.60	3.37
17	6	3.43	3.40	3.22	1.43	3.19	2.96
18	2	2.83	2.88	2.81	0.77	2.82	2.59
19	0	2.31	2.42	2.45	0.39	2.48	2.26
20	2	1.88	2.03	2.13	0.19	2.18	1.97
21	1	1.51	1.70	1.85	0.09	1.9	1.70
22	0	1.21	1.40	1.60	0.04	1.68	1.48
Total	150	145.52	143.63	140.41	149.98	138.93	140.89
ML estimates (Standard Error)		0.2083 (0.0101)	0.1562 (0.008)	0.1041 (0.006)	9.6000 (0.2529)	0.1907 (0.0120)	0.2024 (0.0125)
$-\log L$		435.85	443.80	459.20	441.62	467.25	461.16
AIC		873.71	889.60	920.41	885.24	936.51	924.33
AICC		873.74	889.63	920.44	885.27	936.53	924.36
BIC		876.72	892.61	923.42	888.25	939.52	927.34
χ^2		23.38	32.45	57.54	109.12	75.22	71.37
df		11	12	11	9	13	12
p-value		0.31875	0.03275	$3.0*10^{-5}$	$1.3*10^{-15}$	$9.5*10^{-8}$	$4.0*10^{-7}$

The following histogram represents the number of micronuclei for the proposed model when compared with other models.



From Table 1 and 2, it has been observed that the discrete Poisson area-biased Ailamujia distribution have the lesser AIC, AICC, $-2\log l$, BIC and χ^2 values along with higher p-values as compared to size-biased Poisson Ailamujia distribution (PSBAD), Poisson Ailamujia distribution (PAD), Poisson distribution (PD), Poisson Lindley distribution (PLD) and Poisson Shanker distribution (PSD). It is evident from the above arguments that the proposed distribution provides better fit than the compared ones.

8. Concluding Remarks

The aim of this study is to use compounding to develop a new distribution for count data termed the "Poisson area-biased Ailamujia distribution". Different distributional features of the newly formed distribution have been obtained and analysed. The parameter of the proposed distribution has been estimated by the known method of maximum likelihood estimation. Eventually, the model's efficiency was assessed using two count data sets, and it was revealed that the Poisson area-biased Ailamujia distribution provides an appropriate fit for the two count data sets.

References

Fisher, R., (1934). *The effect of methods of ascertainment*, Annals Eugenics, 6, pp. 13–25.

Rao, C. R., (1965). *On Discrete Distributions Arising Out of Methods of Ascertainment.* Sankhya; the Ind. J. Statist, series A, 27(2/4), pp. 311–324.

Gerstenkorn, T., (1993). A compound of the generalized gamma distribution with the exponential one. Recherches surles deformations,16(1), pp. 5–10.

Gerstenkorn, T., (1996). *A compound of the Polya distribution with beta distribution. Random oper.* And Stoch. Equ., 4(2), pp. 103-110.

Giovani, C. R., Francisco, L., and Pedro, L. R., (2016). Poisson-Exponential distribution different methods of estimation. *Journal of applied statistics*, 15(45), pp. 128–144.

Gupta, R. C., Ong, S. H., (2004). A new generalization of negative binomial distribution. *Journal of computational statistics and data analysis*, 45, pp. 287–300.

Greenwood, M., Yule, G. U., (1920). *An inquiry into the nature of frequency distribution representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents*, J. Roy. Stat. Soc., 83, pp. 225–279.

Judcy, C., (1942). *Data on the macroscopic fresh-water fauna in dredge samples from the bottom of Weber Lake.*

Mahmoudi, E., Zakerzadeh H., (2010). Generalized Poisson-Lindley distribution. *Communications in statistics – theory and methods*, 39(10), pp. 1785–1798.

Puig, P., Valero J., (2006). Count Data Distributions, Some Characterizations With Applications. *Journal of the American Statistical Association*, 101, pp. 332–340.

Shanker, R., (2017). The discrete Poisson-Akash distribution. *International journal of probability and statistics*, 6(1), pp. 1–10.

Sankaran, M., (2010). The discrete Poisson-Lindley distribution. *Biometrics*, 26, pp. 145–149.

Subhradev, S., (2018). Quasi-Xgamma distribution. *Istatistik: journal of the Turkish statistical association*, 11(3), pp. 65–76.

Shanker, R., Shukla, K.K., (2019). A generalization of Poisson-Sujatha distribution and its application to ecology. *International journal of biomathematics*, 12(2), pp. 66–83.

Thomas, M., (1949). A generalization of Poisson's binomial limit for use in ecology. Biometrika, 36(2), pp. 18–25.

Wanbo, L., Shi, D., (2012). A new compounding life distribution: the Weibull-Poisson distribution. *Journal of applied statistics*, 39(1), pp. 21–38.

Zamani, H., Ismail, N., (2010). Negative binomial-Lindley distribution and its application. *Journal of mathematics and statistics*, 1, pp. 4–9.

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 185–198, DOI 10.2478/stattrans-2022-0037
Received – 05.01.2021; accepted – 09.06.2022

Triads or tetrads? Comparison of two methods for measuring the similarity in preferences under incomplete block design

Artur Zaborski[1]

ABSTRACT

The measurement of preferences can be based on historical observations of consumer behaviour or on data describing consumer intentions. In the latter case, the measure-ment of preferences is performed using methods which express consumer attitudes at the time of research. However, most of these methods are very laborious, especially when a large number of objects is tested. In such cases incomplete analyses may prove useful. An incomplete analysis involves the division of objects into subgroups, so that each pair of objects appears at exactly the same frequency and all objects are in each subgroup.

The purpose of the work is to compare two incomplete methods for measuring the similarity of preferences, i.e. the triad method and the tetrad method. These methods can be used whenever similarities are measured on an ordinal scale. They have been com-pared in terms of their labour intensity and ability to map the known structure of ob-jects, even when all pairs of objects in subgroups cannot be presented equally frequent-ly.

Key words: measurement of preferences, triads, tetrads, multidimensional scaling.

1. Introduction

Preferences represent the basic concept in the theory of economics and, in particular, in the consumer choice theory. They reflect consumers' attitudes developed in the process of mutual interactions between consumers and their environment. They take the form of a binary relationship based on axiomatic properties of reflexivity, transitivity and consistency (e.g. Varian, 2005, pp. 63–64). Even though the relationship of preferences is very easy to determine experimentally (e.g. using a questionnaire survey), the measurement aimed at quantifying preferences is a problematic one. There are no precise and unambiguous definitions of many concepts, therefore it is difficult to measure both the intensity and the level of the conditions described by these concepts.

[1] Wroclaw University of Economics, Faculty of Economics and Finance, Poland.
E-mail: artur.zaborski@ue.wroc.pl, ORCID ID: https://orcid.org/0000-0003-1374-2268.

An important tool in the study of the similarities of preferences is nonmetric multidimensional scaling, which is a technique for the analysis of similarity (or dissimilarity) data on a set of n objects (see, e.g. Borg and Groenen, 2005). Multidimensional scaling produces a multidimensional geometrical representation of objects in a low dimensional space (this is usually a two or three-dimensional space), where relationships between the objects correspond to geometric relationships of points representing objects on the perceptual map. In the nonmetric multidimensional scaling, dissimilarities are measured on the ordinal scale. In this case, given the dissimilarities δ_{ij} and δ_{kl} of two object pairs (O_i, O_j) and (O_k, O_l) from the set of n objects $O = (O_1, O_2, \ldots, O_n)$, the researcher is only interested which of the two dissimilarity δ_{ij} and δ_{kl} is greater (or smaller).

There are two ways of obtaining input dissimilarities in multidimensional scaling. When they are directly obtained from empirical subjective measurements of objects performed by subjects, they are called direct dissimilarities. By contrast, when they are not obtained from subjects, but calculated from a data matrix associated with these objects, they are labeled as derived dissimilarities. This article focuses only on direct dissimilarities.

When the number of objects is high, the number of direct assessments made by respondents becomes too large, and makes the dissimilarities task more difficult. In this article, two incomplete methods are proposed to solve this problem in order to make the similarity task easier, while keeping satisfactory scaling solutions. These methods are the method of triads and the method of tetrads. The idea of the presented methods is based on the theory of balanced incomplete block designs (see, e.g. Burton and Nerlove, 1976; Rink, 1987; Morris, 2010, pp. 109–111). The method of tetrads is an original proposal, the idea of which is based on the method of triads. These methods will be compared due to their labor intensity and the ability to map the known structure of preferences.

2. The methods of collecting preferences similarity data

The most important decision to be taken at the initial stage of preference scaling is the selection method for measuring similarities. So far, many more or less popular and widely used methods of direct similarities measurement have been developed (see, e.g. Bijmolt, 1996, pp. 30-31; Zaborski, 2001, pp. 40–43). There are three main approaches to collecting input similarities. The first approach is based on rankings and similarity ratings of the pairs of objects, the second uses grouping and sorting tasks in order to calculate similarities, and the third approach consists of pairwise comparisons of similarities. Some of them, suggested in the literature, are presented in Table 1.

Table 1. The methods of collecting similarities data

Method	Description
Sorting	The subject has to sort the objects into a number of groups, with relatively similar objects in each group
Paired comparisons	For all pairs of objects the subject has to indicate the most preferred object
Ratings	The subject has to rate each pair of objects on an ordinal scale, where the extreme values of the scale represent the maximum dissimilarity and maximum similarity of preferences
Ranking	The subject has to arrange the objects from the most to the least preferred
Ranking of pairs	The subject is requested to arrange all possible pairs of objects in order of decreasing similarity of preferences
Pick *k* out of *n*	The subject is asked to pick a number of objects which s/he considers most similar to a particular reference object. This process has to be done several times while rotating the reference object
Conditional ranking	One object is presented to the subject as a reference object, and the remaining objects have to be ordered on the basis of their preference similarity with the reference object. Each of the objects is in turn presented as the reference
Dyads	For each pair of pairs of objects (dyad) the subject has to select a more similar pair of the two
Triads	The subject has to indicate which objects of combinations of tree objects form the most similar pair, and which form the least similar pair

Source: Zaborski (2017).

The differences in the application of various measurement methods may result from the number of objects simultaneously presented to the respondents (e.g. in the method consisting in ranking, sorting or conditional ordering of similarities the respondents simultaneously assess all objects, while in the course of pairwise comparison or triad method, only two or three objects are presented in a sequence), the difficulty in assessing similarities (e.g. ordering for the entire set, especially with a large number of objects, is more problematic than selecting the preferred object from two or three items) and the total number of required ratings (in the case of ranking, it is just one assessment, and, e.g. for the triad method the number of assessments is a cubic function of the number of objects).

Table 2. Effects of the similarity data collection methods

Effects	Preference collection methods				
	ST	RT	CR	RN	TR
Subjective feelings:					
Fatigue	++	+	–	++	–
Boredom	+	+	–	+	–
Ease of expressing preferences	+	+	+	+/–	+
Command clarity	+	+	+	+	+
Preference judgements:					
Completion time	++	+	+/–	++	–
Missing values	+	++	+/–	+/–	–
Preference scaling results:					
Goodness of fit to the data	+	+	+	+	+
Recovery of known distances	–	+	+	+/–	+

Explanation: ST – sorting, RT – ratings, CR – conditional ranking, RN – ranking, TR – triads, ++ = very good, + = good, +/– = medium, – = poor

Source: own work based on Bijmolt, 1996, pp. 41-48; Zaborski, 2003.

The selection of a method affects subjective feelings of the respondents, i.e. fatigue, weariness resulting from making numerous assessments, or difficulties in expressing similarity assessments. As a result, the collected data may be incomplete or assessments which do not always fully reflect the respondents' attitudes may occur. Table 2 presents the impact of different preference collecting methods on subjective feelings of respondents, preference judgements and preference scaling results. It shows that by using methods that are not labor intensive, i.e. sorting or ranking, we are not able to fully reproduce the known structure of preferences. With ranking procedures, the respondent may become frustrated if asked to rank many more objects, and he/she may skip the question or select the most and least preferred, ignoring the rest. On the other hand, paired comparison methods require a large number of observations. When the number of objects becomes large, deriving all pairs can become tedious and time-consuming. The respondent may become tired answering the large number of paired comparisons that are necessary to collect similarity data. In such cases, incomplete tests may be helpful. The triad and tetrad methods presented in this paper under the incomplete block design allow for a significant reduction of the above-mentioned limitations, resulting from the use of other methods included in Table 1.

3. Presentation of methods

In the method of triads (see Roskam, 1970; Burton and Nerlove, 1976) the subject is asked to consider all possible groups of three objects (O_i, O_j, O_k) ($i, j, k = 1, 2, \ldots, n$, where $i \neq j \neq k \neq i$) at a time, taken from the full set of n objects $O = (O_1, O_2, \ldots, O_n)$. The subject has to indicate which two objects of each combination form the most similar pair, and which two objects form the least similar pair. On this basis the triad is formed, where the most similar objects are placed as the first and the second, and the least similar as the first and the third one. For example, if (O_i, O_j) is the most similar pair and (O_j, O_k) is the least similar pair, the triad is (O_j, O_i, O_k).

In the method of tetrads the respondent also has the task to indicate the most similar pair and the least similar pair, but for all possible groups of four objects (O_i, O_j, O_k, O_l) $i, j, k, l = 1, 2, \ldots, n$, where $i \neq j \neq k \neq l \neq i \neq k$ and $j \neq l$. On this basis the tetrad is formed, where the most similar objects are placed as the first and the second, and the least similar as the first and the fourth one. For example, if (O_i, O_j) is the most similar pair and (O_j, O_l) is the least similar pair, the tetrad is (O_j, O_i, O_k, O_l). If the object from the most similar pair (O_i, O_j) is not present in a pair of the least similar objects then the most similar objects are placed as the second and the third. In this situation one should also ask the respondent to indicate the second most similar pair of objects. For example, if (O_i, O_j) is the most similar pair, (O_i, O_k) is the second similar pair and (O_k, O_l) is the least similar pair, the tetrad is $(O_k, O_i, O_j, O_l)^*$.

Although the advantage of the methods presented above is a relative simplicity of the judgments required of the subjects, so they can be useful techniques for preference data collection, the number of triads and tetrads increases very rapidly with the number of objects. The number of ratings which a respondent must make for n objects in the method of triads is equal to the number of three element combinations of n-element set and it amounts to:

$$C_n^3 = \frac{n(n-1)(n-2)}{6}. \tag{1}$$

For tetrads it is a four element combinations of n-element set:

$$C_n^4 = \frac{n(n-1)(n-2)(n-3)}{24}, \tag{2}$$

so beyond about n=7, the presentation of the full sets becomes totally unfeasible and very laborious for the subject.

When the number of triads or tetrads is considered too large to be practical, according to the theory of balanced incomplete block designs, it can be reduced in such a way that all pairs of objects are presented equally frequently, but less than their potential maximum number. If λ denotes the number of three or four-elements combinations (blocks) in which each pair of objects occurs, than the reduced number of blocks L_λ must satisfy both of these defining relations (see, e.g. Rink, 1987):

$$\begin{cases} nr = kL_\lambda \\ (n-1)\lambda = (k-1)r \end{cases}, \tag{3}$$

where:

k is the number of objects in one block (k=3 for triads and k=4 for tetrads),
r is the number of replication of each object in the reduced blocks,
λ=1,..., n – 2 for triads,
λ=1,..., (n – 1)(n – 2)/2 for tetrads.

According to the equations (3), the number of incomplete blocks in the method of triads is equal:

$$L_\lambda = C_n^3 \frac{\lambda}{n-2} = \frac{\lambda n(n-1)}{6}, \tag{4}$$

and in the method of tetrads:

$$L_\lambda = C_n^4 \frac{2\lambda}{(n-2)(n-3)} = \frac{\lambda n(n-1)}{12}. \tag{5}$$

The number of triads and tetrads for different values of λ and n is shown in Table 3. Because it is not possible to define a reduced number of blocks for all combinations of λ and n, not all the cells in Table 3 are filled.

Table 3. The number of triads and tetras for different values of λ and n

n	**Triads**						**Full set of triads**	**Tetrads**						**Full set of tetrads**
	λ							**λ**						
	1	**2**	**3**	**4**	**5**	**6**		**1**	**2**	**3**	**4**	**5**	**6**	
6	–	10	–	20	×	×	20	–	–	–	–	–	15	15
7	7	14	21	28	35	×	35	–	7	–	14	–	21	35
8	–	–	–	–	–	56	56	–	–	14	–	–	28	70
9	12	24	36	48	60	72	84	–	–	18	–	–	36	126
10	–	30	–	60	–	90	120	–	15	–	30	–	45	210
11	–	–	55	–	–	110	165	–	–	–	–	–	55	330
12	–	44	–	88	–	132	220	–	–	33	–	–	66	495
13	26	52	78	104	130	156	286	13	26	39	52	65	78	715
14	–	–	–	–	–	182	364	–	–	–	–	–	91	1001
15	35	70	105	140	175	210	455	–	–	–	–	–	105	1365
16	–	80	–	160	–	240	560	20	40	60	80	100	120	1820
17	–	–	136	–	–	272	680	–	–	68	–	–	136	2380

Source: own work.

For both methods it is possible to enter the judgement on paired comparisons into a similarity matrix. The creation of the triangular similarity matrix is possible by giving the pair of objects the number of points, which is equal to the number of pairs in a block, for which it can be assumed that the similarity is smaller than the similarity of a given pair. The number of points assigned to pairs from the set of hypothetical blocks (triads and tetrads) marked with the consecutive letters of the alphabet is presented in Table 4 and Table 5.

Table 4. Number of points for pairs in example triad

Objects	**Most similar pair**	**Least similar pair**	**Triad**	**Number of points for pairs in triads**		
ABC	AB	AC	ABC	AB=2	AC=0	BC=1

Source: own work.

Table 5. Number of points for pairs in example tetrads

Objects	**Most similar pair**	**Least similar pair**	**Tetrad**	**Number of points for pairs in tetrads**					
ABCD	AB	AD	ABCD	AB=5	AC=1	AD=0	BC=3	BD=1	CD=3
ABCD	AB	CD	CABD$^{*1)}$	AB=5	AC=4	AD=1	BC=1	BD=3	CD=0
			CBAD$^{*2)}$	AB=5	AC=1	AD=3	BC=4	BD=1	CD=0
			DABC$^{*3)}$	AB=5	AC=1	AD=4	BC=3	BD=1	CD=0
			DBAC$^{*4)}$	AB=5	AC=3	AD=1	BC=1	BD=4	CD=0

Explanation: the second most similar pair of objects is: 1) AC; 2) BC; 3) AD; 4) BD

Source: own work.

The value of an element p_{ij} in the i-th row and the j-th column of the similarity matrix is equal to the sum of points awarded to a pair consisting of the i-th and the j-th objects in all blocks.

To discover the similarity structure of preferences by using nonmetric multidimensional scaling, the similarity matrix should be transformed into a matrix of dissimilarities, especially if all pairs of objects in blocks cannot be presented equally frequently. The dissimilarities δ_{ij} are determined in accordance with the formula:

$$\delta_{ij} = 1 - \frac{p_{ij}}{\max r \cdot m_{ij}}, \quad (6)$$

where m_{ij} is the number of pairs (i, j) in blocks, max$r$ is the maximum number of points that can be obtained by a pair of objects in a block (for triads max$r = 2$ and for tetrads max$r = 5$). The denominator in the second component of the equation (6) indicates the maximum possible number of points for the pair (i, j), i.e. when in all blocks it was considered to be the pair of the most similar objects.

4. The comparison of methods

In order to make the study results independent on respondents' subjective effects (fatigue, boredom, task insight), the comparison of the presented above methods was made on the basis of the given distance matrix (see Table 6). The matrix shows the dissimilarities in the preferences of the University of the Third Age members in relation to the selected forms of activities (see Zaborski, 2014). As a result of multidimensional scaling based on the dissimilarity matrix, a configuration of points representing activities was obtained (Figure 1).

Table 6. The preferences dissimilarity matrix

Activities	1	2	3	4	5	6	7	8	9
1. English	0.000								
2. German	0.694	0.000							
3. Computer skills	1.372	1.128	0.000						
4. Gymnastics	0.908	1.111	0.766	0.000					
5. Yoga	0.596	1.007	1.062	0.370	0.000				
6. Swimming	1.117	1.276	0.712	0.209	0.568	0.000			
7. Weight training	1.395	1.413	0.530	0.522	0.892	0.342	0.000		
8. Nordic walking	0.754	1.291	1.333	0.578	0.318	0.723	1.065	0.000	
9. Painting and handcraft	1.196	0.663	0.637	1.071	1.190	1.138	1.104	1.507	0.000

Source: own work.

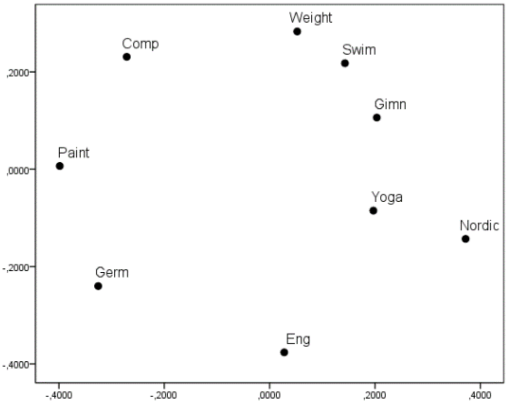


Figure 1. Preference map received based on the dissimilarity matrix

Source: own work.

In order to check how the incomplete study affects the preferences scaling results, five sets of triads (for λ=1, 2, …,5) and three sets for tetrads (for λ=3, λ=6 and λ=9) were generated. All sets are presented in Table 7.

Table 7. Sets of triads and tetrads

λ	Triads											
λ=1	1 2 3	6 4 5	8 7 9	7 4 1	9 2 5	3 6 8	1 6 9	8 4 2	3 5 7	8 5 1	7 6 2	3 9 4
λ=2	1 5 9	3 2 8	4 6 7	2 9 6	3 4 1	8 5 7	7 3 9	4 5 2	6 8 1	8 4 9	5 6 3	2 1 7
	1 2 3	6 4 5	8 7 9	7 4 1	9 2 5	3 6 8	1 6 9	8 4 2	3 7 5	8 5 1	7 6 2	3 9 4
λ=3	2 1 4	2 5 3	4 6 3	5 4 7	8 5 6	6 7 9	1 8 7	9 2 8	3 9 1	3 4 1	4 5 2	5 6 3
	4 6 7	8 5 7	8 6 9	7 9 1	2 1 8	3 9 2	2 1 6	7 3 2	8 4 3	5 4 9	6 5 1	7 6 2
	3 7 8	8 4 9	1 5 9	1 6 3	7 4 2	8 5 3	6 4 9	1 5 7	8 6 2	7 3 9	4 8 1	9 2 5
λ=4	the complement of triads set for λ=3											
λ=5	the complement of triads set for λ=2											
	Tetrads											
λ=3	1243	6512	2187	7351	6481	3961	8419	1597	9328	5429	7692	7432
	8539	6479	5463*	8562	3678*	8547						
λ=6	1243	6512	2187	7351	6481	3961	7681	8419	1597	8429	5429	7692
	7432	6562	8539	6479	5463*	8673*	8547	7491	8469*	4539	3748	6451
	2467*	4512	8423	2957	9218	2936	7391	8569	2587*	1263	5673*	8513
λ=9	7351	6512	2187	7439	7612	7382	7681	8419	1597	8429	5429	7692
	6592	6562	8539	3921	6432	1469*	8547	7491	8469*	4539	3748	6451
	1243	4512	8423	7351	6481	3961	7391	8569	2587*	1263	5673*	8513
	7432	2957	7645*	6479	5463*	8673*	7681	8479*	8563	8519	8542	3451
	2467*	5368	8412	2957	9218	2936						

Explanation: * – the most similar objects are placed second and third

Source: own work.

For each set similarity matrices were calculated, then they were transformed into dissimilarity matrix according to the formula (6) and the multidimensional scaling with the use of MINISSA program was performed. In the case of the method of triads the program TRISOSCAL, which uses MINISSA algorithm for multidimensional scaling, was used. MINISSA and TRISOSCAL are available in the multidimensional scaling package NewMDSX (Coxon and Davies, 1982). MINISSA performs the basic model of nonmetric MDS by taking data in the form of the full square symmetric matrix (or its lower triangle) of dissimilarities, whose elements are to be transformed to give the distances of the solution. This transformation will preserve the rank order of the input data.

The quality of matching the resulting points' configuration to the configuration determined based on the distance matrix (Table 6) was tested by the Procrustes statistic (see Cox and Cox, 2001; Borg and Groenen, 2005):

$$R^2 = \frac{\left\{tr\left(\mathbf{X}^{*T}\mathbf{Y}\mathbf{Y}^T\mathbf{X}^*\right)^{\frac{1}{2}}\right\}^2}{tr(\mathbf{X}^{*T}\mathbf{X}^*)tr(\mathbf{Y}^T\mathbf{Y})}, \tag{7}$$

where $\mathbf{X}^* = \mathbf{X}(\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X})^{\frac{1}{2}}(\mathbf{Y}^T\mathbf{X})^{-1}$ – optimally rotated configuration $\mathbf{X}$ ($\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n]^T$ – the configuration of points determined on the basis of the incomplete blocks), $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_n]^T$ – the configuration of points determined on the basis of the distance matrix. $R^2 \in (0; 1\rangle$, where 1 means a perfect matching. Because all configurations have the centroids at the origin and the average distance of points from the origin is equal to 1, the Procrustes analysis is limited only to the stage of optimal rotation. The quality of matching of the resulting configurations of points to the configuration on Figure 1 tested by the Procrustes statistic is presented in Table 8.

Table 8. Procrustes statistics for different sets of triads and tetrads

	Triads					Tetrads		
λ	λ=1	λ=2	λ=3	λ=4	λ=5	λ=3	λ=6	λ=9
L_λ	12	24	36	48	60	18	36	54
R^2	0.6071	0.9401	0.9438	0.9484	0.9749	0.9647	0.9483	0.9715

Source: own work.

With the exception of the set of twelve triads, the quality of matching the other sets does not differ significantly (they are in the range from 0.94 to 0.97) and should be considered as very good. Therefore, due to the practical application of both methods, the further study was limited to the triad sets for λ=2, λ=3 and λ=4, and the tetrad sets for λ=3 and λ=6. To verify how the choice of blocks affects the preference scaling results, nine sets of triads were generated (three for each value of λ), and six sets of tetrads (three for λ=3, and three for λ=6). As it was previously mentioned, it is not possible to determine reduced sets for all combinations of λ and *n*, and in consequence, all pairs of objects cannot be presented equally frequently. So each set was modified by subtracting randomly selected two, four and six triads/tetrads. Finally 36 sets of triads and 24 sets of tetrads were obtained. Based on the dissimilarity matrices for all sets, multidimensional scaling with the use of MINISSA program was performed. The quality of matching of the resulting configuration to the initial configuration (Figure 1) was tested by the Procrustes statistic. In addition, for λ=2 (in the case of triads) and for λ=3 (in the case of tetrads) each set was successively reduced by two triads/tetrads, until the value of the Procrustes statistics started to fall drastically. The reduction in the number of blocks was done in such a way that in each block (as much as possible) each pair was present at least once. The results of the study are presented in Table 9 and Table 10.

It can be seen that for all generated sets of tetrads results should be regarded as almost perfect. Even if the number of tetrads in sets was reduced by 10, the results indicate a very good matching in relation to the scaling carried out for the data set in Table 6. There is only a small difference in the obtained results between reduced (maximum to 8) sets of tetrads. The difference between the best and the worst solution for all sets in this group is less than 0.08 (excluding the results for Te_p^{-12} and Te_p^{-14}).

Table 9. Procrustes statistics for different sets of triads

	λ=2			λ=3			λ=4		
	Tr_1	Tr_2	Tr_3	Tr_4	Tr_5	Tr_6	Tr_7	Tr_8	Tr_9
Tr_p^{0}	0.9401	0.9009	0.9444	0.9438	0.9606	0.9402	0.9484	0.9572	0.9392
Tr_p^{-2}	0.9705	0.9009	0.9566	0.9535	0.9633	0.9398	0.9487	0.9535	0.9373
Tr_p^{-4}	0.8723	0.9181	0.9403	0.9434	0.9420	0.9355	0.9473	0.9508	0.9316
Tr_p^{-6}	0.8772	0.9010	0.8949	0.9313	0.9427	0.9364	0.9448	0.9581	0.9204
$\overline{Tr}_p$	0.9181			0.9444			0.9448		
CV(%)	3.74			1.04			1.16		
Tr_p^{-8}	0.8480	0.9033	0.9097						
Tr_p^{-10}	0.8610	0.8862	0.8126						

Explanation: Tr_p^{-k} – set Tr_p (p=1,2,…,9) reduced by k triads; CV – the coefficient of variation

Source: own work.

Table 10. Procrustes statistics for different sets of tetrads

	λ=3			λ=6		
	Te_1	Te_2	Te_3	Te_4	Te_5	Te_6
Te_p^{0}	0.9647	0.9727	0.9323	0.9483	0.9814	0.9518
Te_p^{-2}	0.9507	0.9674	0.9062	0.9412	0.9723	0.9556
Te_p^{-4}	0.9452	0.9226	0.9034	0.9431	0.9828	0.9622
Te_p^{-6}	0.9450	0.9356	0.9437	0.9541	0.9688	0.9447
$\overline{Te}_p$	0.9409			0.9589		
CV (%)	2.37			1.51		
Te_p^{-8}	0.9586	0.9658	0.9563			
Te_p^{-10}	0.9575	0.9425	0. 9184			
Te_p^{-12}	0.8213	0.9077	0.8125			
Te_p^{-14}	0.8235	0.4878	0.2434			

Explanation: Te_p^{-k} – set Te_p (p=1,2,…,6) reduced by k tetrads; CV – the coefficient of variation

Source: own work.

The coefficient of variation of the Procrustes statistic value for sets containing from 8 to 18 tetrads is about 0.024, and from 30 to 36 tetrads about 0.015. It attests the fact that the choice of a set of tetrads (as in the case of the method of triads) has no significant effect on the results of preference scaling, even when all pairs of objects cannot be presented equally frequently. The analysis showed that the results clearly deteriorated only when the number of tetrads in sets was less than 8, but in these cases, not all pairs appear in sets. In the case of the triads method, similar results were obtained when the number of triads in the set is over 20, which means that recovery of a known structure of preferences requires respondents to make about three times more assessments than in the method of tetrads.

5. Conclusions

The results of many studies (see, e.g. Humphreys, 1982; Bijmolt, 1996; Zaborski, 2003) indicate that preference scaling based on various direct methods of measuring dissimilarities gives similar solutions. However, the selection method affects subjective feelings of respondents, which may result in different quality of input data. Therefore, the choice of the method of measurement should be guided primarily by two criteria: the method should not be labour-intensive, and expressing opinions on similarities should not cause problems to respondents. The methods which are proposed in the article do not satisfy the first of the above conditions. In the case of the triads method the number of ratings which a respondent must make for *n* objects is equal to the number of three element combinations of an *n*-element set. In the method of tetrads it is the number of four element combinations of an *n*-element set. The article indicates the possibility of reducing the number of sets presented to respondents in such a way that each pair of objects appears equally frequently, but less than their potential maximum number. In the example for 9 objects it was shown that scaling based on 8 tetrads gave a good solution. Using the method of triads, where a respondent is asked to pick out the most similar and the least similar pair from the three element sets, obtaining comparable results requires over three times more assessments. It was also demonstrated that the choice of the incomplete sets has no significant effect on the results of nonmetric multidimensional preference scaling, even when all pairs of objects cannot be presented equally frequently. This conclusion is particularly relevant for the creation of reduced sets when the number of objects does not allow to fulfil the condition of an equal number of pairs. The analysis indicated that the tetrad method can be used if each pair of objects appears in sets at least once, while for the method of triads each pair should appear at least twice.

Acknowledgements

This paper was presented on the conference MSA 2019, which financed its publication. Organization of the international conference "Multivariate Statistical Analysis 2019" (MSA 2019) was supported from resources for popularization of scientific activities of the Minister of Science and Higher Education in the framework of agreement No 712/P-DUN/202019.

References

Bijmolt, T. H. A., (1996). *Multidimensional Scaling in Marketing: Towards Integrating Data Collection and Analysis*. Capelle a/d Ussel: Labyrint Publication.

Borg, I., Groenen, P. J. F, (2005). *Modern Multidimensional Scaling*. Theory and Applications. Second Edition. New York: Springer-Verlag.

Burton, M. L., Nerlove, S. B., (1976). Balanced design for triads tests: two examples from English. *Social Science Research*, 5, pp. 247–267.

Cox, T. F., Cox, M. A. A., (2001). *Multidimensional Scaling*. Second Edition London: Chapman and Hall.

Coxon, A. P. M, Davies, P. M. (1982). The user's guide to multidimensional scaling with special reference to the MDS(X) library of computer programs. London, *Heinemann Educational.*

Humphreys, M. A., (1982). Data collecting effects on nonmetric multidimensional scaling solutions. E*ducational and Psychological Measurement*, 42, pp. 1005–1022.

Morris, M., (2010). *Design of Experiments: An Introduction Based on Linear Models*, New York: Chapman and Hall/CRC.

Rink, D. R., (1987). An improved preference data collection method: Balanced incomplete block designs. *Journal of the Academy of Marketing Science*, 15(3), pp. 54–61.

Roskam, E. E. (1970), The methods of triads for multidimensional scaling, *Nederlands Tijdschrift Voor de Psychologie*, No. 25, pp. 404–417.

Varian, H. R., (2005), *Mikroekonomia*, PWN, Warszawa.

Zaborski, A., (2001). Skalowanie wielowymiarowe w badaniach marketingowych. Wrocław, *Wydawnictwo Akademii Ekonomicznej we Wrocławiu.*

Zaborski, A., (2003). *Wpływ alternatywnych metod pomiaru preferencji na wyniki skalowania wielowymiarowego*. Prace Naukowe AE w Katowicach, Analiza i prognozowanie zjawisk o charakterze niemetrycznym, pp. 59–69.

Zaborski, A., (2014). Analiza preferencji słuchaczy Uniwersytetu Trzeciego Wieku z wykorzystaniem wybranych metod niesymetrycznego skalowania wielowymiarowego. Studia Ekonomiczne. *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach*, 195(14), pp. 216–224.

Zaborski, A., (2017). The Influence of Triad Selection on the Preference Scaling Results. Acta Universitatis Lodziensis. *Folia Oeconomica*, 4(330), pp. 87–97.

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 199–214, DOI 10.2478/stattrans-2022-0038
Received – 24.08.2021; accepted – 26.05.2022

Two-stage cluster sampling with unequal probability sampling in the first stage and ranked set sampling in the second stage

Michael C. Ugwu[1], **Mbanefo S. Madukaife**[2]

ABSTRACT

In this research work we introduce a new sampling design, namely a two-stage cluster sampling, where probability proportional to size with replacement is used in the first stage unit and ranked set sampling in the second in order to address the issue of marked variability in the sizes of population units concerned with first stage sampling. We obtained an unbiased estimator of the population mean and total, as well as the variance of the mean estimator. We calculated the relative efficiency of the new sampling design to the two-stage cluster sampling with simple random sampling in the first stage and ranked set sampling in the second stage. The results demonstrated that the new sampling design is more efficient than the competing design when a significant variation is observed in the first stage units.

Key words: cluster sampling, population mean estimator, probability proportional to size sampling, ranked set sampling, relative efficiency.

1. Introduction

In scientific research, sample survey to a great extent plays a vital role, most especially in the presence of limited cost. This is because we need not possibly embark on complete enumeration which entails studying the entire population, in order to learn efficiently about the population characteristics of interest. Also in real life situations, there are occasions unlike in element sampling when a list of elements of the population is not available but it is easy (or possible) to obtain a list of segmented groups, known as clusters. Even when such list exists, it is sometimes uneconomical to obtain information from a sample of elements in the population due to the nature of the distribution of the population. In such cases, it becomes ideal to use cluster sampling technique to draw random sample from the population and when this

[1] Department of Statistics, University of Nigeria, Nsukka, Nigeria. ORCID: https://orcid.org/0000-0001-7356-9183

[2] Corresponding author. Department of Statistics, University of Nigeria, Nsukka. Nigeria. E-mail: mbanefo.madukaife@unn.edu.ng. ORCID: https://orcid.org/0000-0003-2823-4223.

technique is carried out in two phases, it becomes two-stage cluster sampling (Okafor, 2002).

In two-stage cluster sampling, the entire units of the population are at first grouped into say *N* clusters, each having M_i, $i = 1, 2, \ldots, N$ elements. Then a random sample of *n* clusters, say, is drawn from the *N* clusters, also known as the first stage units (FSU), as the first stage sample. From each of the *n* selected clusters, each of size M_i, $i = 1, 2, \ldots, n$ elements, a random sample of cluster elements of size m_i is also selected from M_i second stage units (SSU) as second stage sample. The common motivation of cluster sampling is to reduce cost by increasing sampling efficiency.

A good number of authors have applied two-stage cluster sampling in real life situations in order to enhance sampling efficiency. Some of them include Fears and Gail (2000), Stehman et al. (2009), Phillips et al. (2008), Horney et al. (2010) as well as Galway et al. (2012) and Dilip (2015). The efficiency of the design when applied to real life situations, however depends to great extent on the sampling techniques used in both stages of the design.

It could be recalled that in equal probability sampling, all the population units have equal chances of being selected in the sample regardless of the size of each unit. When units of clusters are of different sizes, it is appropriate to use probability proportional to size (PPS) sampling (Damon, 2018 & Ozturk, 2019). In this sampling plan, the probability of selection of a cluster element is in proportion to its size or measure of size of the element, so that larger clusters have greater chances of being selected than the smaller clusters, provided the sizes of units of clusters in the population are known and also have positive correlation with the variable under study. The choice of PPS scheme in the first-stage of two-stage sampling under variant cluster sizes has also been supported by Innocenti et al. (2019). Such a procedure of sample selection is also known as unequal probability sampling (Okafor, 2002). For a more detailed discussion on selection procedures and estimation in unequal probability sampling, see Shahbaz and Hanif (2010).

Optimum sampling methods that are cost friendly have been of great concern in the field of statistics, especially when the cost of measuring the population attribute under study is high. In situations where it is less costly to identify sampling units to be included in the sample and at the same time ranking them accordingly with respect to the attribute of interest than to directly measure the values, a ranked set sample (RSS) yields better efficiency than its simple random sample (SRS) counterpart under the same sample size (McIntyre, 1952 and Chen et al., 2003). Ranked set sampling was introduced by McIntyre (1952) and Halls and Dell (1966) while its theoretical basis was laid by Takahasi and Wakimoto (1968) and Dell and Clutter (1972). Also, it has been applied in real-life situations by a number of researchers including Chen et al. (2003).

In order to improve the efficiency of two-stage sampling, Nematollahi et al. (2008) introduced RSS in the second stage, with the first stage remaining as SRS scheme. They showed that the estimators obtained from the design have significant improvement in efficiency over the dominant case of SRS scheme on both stages. Regardless of the improvement observed in Nematollahi et al. (2008), the problem of sampling from variant cluster sizes in the first stage is not addressed. Innocenti et al. (2021) presented three options, namely: sampling clusters with probability proportional to cluster size, and then sampling the same number of individuals from each selected cluster in the second stage; sampling clusters with equal probability, and then sampling the same percentage of individuals from each sampled cluster in the second stage and sampling clusters with equal probability, and then sampling the same number of individuals per cluster in the second stage. These options, no doubt, addressed the underlying problem only in the first option. In what appears to be an overall improvement so far, in this direction, Ozturk (2019) obtained a frame work for a two-stage cluster sampling where probability proportional to size (PPS) sampling is applied in the first stage as well as RSS applied in the second stage of sampling.

It is well known that PPS can be carried out with or without replacement. However, PPS without replacement (PPSWOR) is more complex in application than PPS with replacement (PPSWR) and that is one of the major advantages of the later over the former. Additionally, when the study population is very large, sampling with replacement is always best suited. In this work therefore, we shall propose a cluster sampling design in two stages where PPSWR is applied in the first stage and RSS in the second stage. Section 2 gives the framework for PPSWR as well as RSS. In section 3, the estimators of population mean and total of the new sampling design as well as the variance of the estimators are derived. Section 4 gives the relative efficiency of the design over the earlier design proposed by Nematollahi et al. (2008) under significantly variant clusters in the first stage of sampling and the paper is concluded in section 5.

2. The new sampling design

In this paper, a two-stage cluster sampling where sampling is done among the first stage units by probability proportional to size sampling with replacement (PPSWR) and ranked set sampling (RSS) among the second stage units is proposed.

2.1. Probability proportional to size sampling with replacement

Suppose $U_1, U_2, U_3, \ldots, U_N$ have measure of sizes $X_1, X_2, X_3, \ldots, X_N$ respectively, where $X_i;\ i=1,2,3,\ldots,N$ is an integer value and $U_i; i=1,2,3,\ldots,N$ is the *i*th first stage unit. In a situation where the X_i's are not integers, they are all multiplied by an appropriate power of 10 to make them integers. Now, suppose a sample of size *n* units $\{U_i, i = 1,2,3...,n\}$ is to be selected from a population of *N* units, we first form a cumulative aggregate of sizes for each of the first stage units, U_i in the population. Then, the ranges to all the population units are obtained using the cumulative totals. Using a table of random numbers, one is required to select a number d between 1 and $X = \sum_{i=1}^{N} X_i$ inclusive. If the number d falls in the range of U_2, say, then it is selected in the sample. Another random number is drawn between 1 and *X* inclusive, and if the number drawn falls this time in the range of U_i, the unit U_i is selected. In other words, the unit chosen to be included in the sample is the unit whose range contains the drawn random number. The process of drawing a random number is repeated independently until *n* number of units is drawn into the sample. With this selection procedure, the *n* number of units are drawn with PPSWR, and the probability of drawing the *i*th unit from the population is $P_i = X_i / X$ where $\sum_{i=1}^{N} P_i = 1$.

From the foregoing technique according to Hansen and Hurwitz (1943), the unbiased estimator of the population mean is given by:

$$\bar{y}_{PPS} = \frac{1}{nN} \sum_{i=1}^{n} \frac{y_i}{p_i} = \frac{\hat{Y}_{PPS}}{N} \tag{1}$$

where $y_i, i = 1, 2, \ldots, n$ is the value of the variable of interest in the sample, $p_i = \frac{x_i}{X}$ is the probability of drawing the *i*th unit in the sample; x_i is the measure of size of the *i*th sample unit and $\hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$ is the unbiased estimator of the population total, *Y*.

Also, the variance of the sample mean is given by:

$$V(\bar{y}_{pps}) = \frac{1}{nN^2} \left(\sum_{i=1}^{N} \frac{Y_i^2}{P_i} - Y^2 \right) \tag{2}$$

where Y_i is the *i*th cluster total.

2.2. Ranked set sampling (RSS) procedure

The basic premise for RSS is that sampling units are drawn from infinite population or with replacement from a finite population under study and that the sampling units drawn from the population can be ranked by certain means, rather cheaply, devoid of actual measurement of the variable of interest which is either costly or time consuming, or both. It may be considered as a controlled random sampling design. Stokes (1980), Chen et al. (2003) and Al-Omari and Bouza (2014) describe ranked set sampling procedure as follows: (*i*) Randomly select from the study population sampling units of size m^2 (*ii*) Randomly allot the m^2 units selected into m independent sets where every set is of size m. (*iii*) The units in every set are ranked in line with the information about study variable by visual inspection, concomitant variable or through other methods that cost little or nothing. (*iv*) The samples are chosen for quantification by selecting from the first, second down to the *m*th set the lowest ranked unit, the second lowest ranked unit, up to the highest ranked unit from the *m*th set. The entire process from (*i*) to (*iv*) is called a cycle. (*v*) Repeat the cycle, say r times to get a ranked set sample of size rm out of the total of rm^2 units initially selected, see Table 1.

Each cycle of the selection process (i.e. from step *i* to *iv*) will result in measured observations $y_{11}, y_{22}, \ldots, y_{mm}$ into the sample assuming our variable of interest is Y and each of these observations is called judgment order statistic. If $m_1 = m_2 = \ldots = m_m$, that is the set sizes of the independent random samples are equal, the RSS is said to be balanced, else, it is unbalanced. The ranks which the units in the set receive may not necessarily correspond with the numerical layouts of the real values of Y. If they correspond with the numerical layouts, the ranking is said to be perfect, else, it is imperfect. The square brackets [.] are used to denote imperfect ranking in the subscripts of ranked observations while the round brackets (.) are used if the judgment order statistics are perfect.

The efficiency of RSS relies on the sampling allocation, either balanced or unbalanced. In balanced RSS, the rank order statistics has an equal allocation. Takahasi and Wakimoto (1968), Patil (2002) and Al-Omari and Bouza (2014) state that balanced RSS estimator has a variance not greater than its SRS estimator counterpart even in the presence of errors in ranking. This implies that no matter how bad RSS method is, it cannot be worse than SRS method if properly conducted. This no doubt, lies the goodness of the former over the later. Thus, from the measured ranked set sample, we can obtain unbiased estimators of population parameters, such as the population mean and variance.

Suppose y_i is the value of the variable of interest, Y_i for $i = 1, 2, \ldots, M$, where M is the population size. The set $\{Y_1, Y_2, \ldots, Y_m\}$ is a random sample from Y with pdf $f(y)$, finite mean μ and variance σ^2 and with a set of observed values $\{y_1, y_2, \ldots, y_m\}$. Let $Y_{j1}, Y_{j2}, \ldots, Y_{jm}$; $j = 1, 2, \ldots, m$ be a simple random sample drawn from the population with replacement. In some occasions, it is not an easy task ranking m units for large sample set of m, so we select a ranked set sample with small sample set of m and then replicate this sampling scheme up to r times. If that is well executed, it will turn out to produce r cycles, yielding the judgment order statistics value as it is displayed in Table 1. Let Y_{jl} represent the *j*th judgment ordered statistic value from the *j*th sample of size *m* coming from the *l*th cycle of size *r*, $j = 1, 2, \ldots, m; l = 1, 2, \ldots, r$ and with y_{jl} being the value of the observed variable.

According to Takahasi and Wakimoto (1968), based on RSS technique, the unbiased estimator of the mean and its variance are respectively obtained by:

$$\hat{\mu}_{rss} = \frac{1}{mr}\sum_{j=1}^{m}\sum_{l=1}^{r} Y_{jl} \tag{3}$$

and

$$Var(\hat{\mu}_{rss}) = \frac{1}{mr}\left[\sigma^2 - \frac{1}{mr}\sum_{j=1}^{m}\sum_{l=1}^{r}(\mu_{jl} - \mu)^2\right] = \frac{1}{mr}\left[\sigma^2 - \frac{1}{m}\sum_{j=1}^{m}(u_j - \mu)^2\right] \tag{4}$$

Table 1. Display of judgment order statistics (JOS) values from RSS when the cycle is replicated *r* times

Cycle	First JOS	Second JOS	…	mth JOS
Cycle 1	$Y_{[1]1}$	$Y_{[2]1}$	…	$Y_{[m]1}$
Cycle 2	$Y_{[1]2}$	$Y_{[2]2}$	…	$Y_{[m]2}$
⋮	⋮	⋮	⋮	⋮
Cycle *r*	$Y_{[1]r}$	$Y_{[2]r}$	…	$Y_{[m]rfy}$

2.3. The proposed two-stage cluster sampling design

Suppose there are *N* first stage units (FSU's) in the population where every *i*th FSU has M_i second stage units (SSU's) with expected value μ_i and variance σ_i^2. Let the sample size from FSU's be represented by n while m_i represents the sample size from SSU's in the *i*th selected FSU. First, a sample of n FSU's is selected from the population using probability proportional to size with replacement (PPSWR) in the first stage. Then from every *i*th selected FSU's, m_i second stage sampling units will be selected by ranked set sampling (RSS) scheme. Assuming RSS procedure where *r* = 1 is the case in the second stage, then out of every *i*th chosen FSU's, we draw m_i units using RSS procedure. The final sample can be displayed in the array of values given by:

$$\begin{array}{cccc} Y_{1[1]} & Y_{1[2]} & \cdots & Y_{1[m_1]} \\ Y_{2[1]} & Y_{2[2]} & \cdots & Y_{2[m_2]} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{n[1]} & Y_{n[2]} & \cdots & Y_{n[m_n]} \end{array} \tag{5}$$

This is as illustrated by Nematollahi et al. (2008), where $Y_{i[j]}$ denotes the variable of interest pertaining to the *j*th order of the *j*th random sample in *i*th selected FSU which are independent but not identically distributed. To maintain the attribute of independence of samples in RSS, the units selected from every *i*th FSU drawn for rank ordering in the *j*th sample set is carried by simple random sampling with replacement scheme.

If we consider RSS scheme with replication in the second stage sampling, then from every *i*th FSU selected, $m_i = r_i m_i'$ units will be drawn by RSS method in cycles r_i with fix sample size m'. Going by this, let $Y_{il[j]}$ represent the variable pertaining to the *j*th order of *j*th random sample in *l*th cycle from *i*th drawn FSU. Thus, the observations in (5) will form a random sample in every *i*th selected FSU while m_i and $Y_{i[j]}$ are replaced by m_i' and $Y_{il[j]}$ respectively.

3. Estimators of the population mean and total in the new sampling design

The mean estimator for two-stage cluster sampling with probability proportional to size sampling with replacement in the first stage units and RSS design in the second stage units is given by:

$$\overline{y}_{ppsrss} = \frac{1}{n\sum_{i=1}^{N} M_i} \sum_{i=1}^{n} \frac{M_i \overline{y}_i}{P_i} = \frac{1}{n\sum_{i=1}^{N} M_i} \sum_{i=1}^{n} \frac{\hat{Y}_i}{P_i} = \frac{\hat{Y}_{ppsrss}}{\sum_{i=1}^{N} M_i} \tag{6}$$

where $\overline{y}_i = \frac{1}{r_i m_i} \sum_{l=1}^{r_i} \sum_{j=1}^{m_i'} Y_{il[j]}$ is the sample mean of the variable pertaining to the *j*th ordered value from the *j*th random sample in *l*th cycle of sampling in *i*th selected FSU and

$$\hat{Y}_{ppsrss} = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{r_i} \sum_{j=1}^{m_i'} \frac{M_i}{r_i m_i' p_i} Y_{il[j]} \tag{7}$$

is the unbiased estimator of the population total *Y*. $p_i = \frac{x_i}{X}$ is the probability of selecting the *i*th unit in the first stage sample; x_i is the measure of size of the *i*th sample unit. In the case of the measure of size used in this work, $p_i = \frac{m_i}{\sum_{i=1}^{n} M_i}$

It is straight forward to show that the estimator in (6) is an unbiased estimator of the population mean. This is because we have:

$$E\left(\overline{y}_{ppsrss}\right)=E_1E_2\left(\frac{1}{n\sum\limits_{i=1}^{N}M_i}\sum_{i=1}^{n}\frac{M_i\overline{y}_i}{P_i}\right)=E_1\left(\frac{1}{\sum\limits_{i=1}^{N}M_i}\sum_{i=1}^{n}\frac{M_i}{nP_i}E_2\left(\overline{y}_i\right)\right)$$

$$=\frac{1}{n\sum\limits_{i=1}^{N}M_i}\sum_{i=1}^{n}\frac{\hat{Y}_i}{P_i}=\frac{\hat{Y}_{ppsrss}}{\sum\limits_{i=1}^{N}M_i}$$

$$E\left(\overline{y}_{ppsrss}\right)=E_1\left(\frac{1}{\sum\limits_{i=1}^{N}M_i}\sum_{i=1}^{n}\frac{\hat{Y}_i}{nP_i}\right)=\frac{Y_{ppsrss}}{\sum\limits_{i=1}^{N}M_i}=\overline{Y} \tag{8}$$

The variance of the unbiased estimator of the population mean is given by:

$$V(\overline{y}_{ppsrss})=\frac{1}{nM_0^{\;2}}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i}-Y\right)^2+\frac{1}{nM_0^{\;2}}\sum_{i=1}^{N}\frac{M_i^2}{P_i}\frac{\sigma_i^2}{m_i}-\frac{1}{n^2M_0^{\;2}}E_1\left[\sum_{i=1}^{n}\sum_{l=1}^{r_i}\sum_{j=1}^{m_i'}\frac{M_i^2}{P_i^2m_i^2}\left(\mu_{i[j]}-\mu_i\right)^2\right] \tag{9}$$

The result in (9) is derived as follows:

Without loss of generality, let the number of cycles in each FSU be one such that $r_1=r_2=\ldots=r_n=1$. Hence, $m_i=m_i'$. Then,

$$V(\overline{y}_{ppsrss})=V_1E_2(\overline{y}_{ppsrss})+E_1V_2(\overline{y}_{ppsrss}) \tag{10}$$

Considering $V_1E_2(\overline{y}_{ppsrss})$ gives the result:

$$V_1E_2(\overline{y}_{ppsrss})=V_1E_2\left(\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i\overline{y}_i}{P_i}\right),$$

where $\overline{y}_i=\frac{1}{m_i}\sum\limits_{j=1}^{m_i}Y_{i[j]}=\hat{\mu}_{i[j]}$

$$V_1E_2(\overline{y}_{ppsrss})=V_1\left[\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i}{P_i}E_2(\overline{y}_i)\right]=V_1\left[\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i}{P_i}\mu_i\right]$$

$$=\frac{1}{M_0^{\;2}}V_1\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{P_i}\right)$$

$$=\frac{1}{M_0^{\;2}}\left[\frac{1}{n}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i}-Y\right)^2\right] \tag{11}$$

Also considering $E_1V_2(\bar{y}_{ppsrss})$ in (10) gives the result:

$$E_1V_2(\bar{y}_{ppsrss}) = E_1V_2\left(\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{P_i}\right)$$

$$= E_1\left[V_2\left(\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{P_i}\right)\right] = E_1\left[\frac{1}{n^2M_0{}^2}\sum_{i=1}^{n}\frac{M_i^2}{P_i^2}V_2(\bar{y}_i)\right]$$

But Takahasi and Wakimoto (1968) have obtained that

$$V_2(\bar{y}_i) = \frac{1}{m_i}\left[\sigma_i^2 - \frac{1}{m_i}\sum_{j=1}^{m_i}\left(\mu_{i[j]} - \mu_i\right)^2\right]$$

where σ_i^2 is the variance of the variable of interest Y in the *i*th FSU and $\mu_{i[j]}$ is the expected value of $Y_{i[j]}$. Hence,

$$E_1V_2(\bar{y}_{ppsrss}) = E_1\left[\frac{1}{n^2M_0{}^2}\sum_{i=1}^{n}\frac{M_i^2}{P_i^2}\frac{1}{m_i}\left(\sigma_i^2 - \frac{1}{m_i}\sum_{j=1}^{m_i}\left(\mu_{i[j]} - \mu_i\right)^2\right)\right]$$

$$= E_1\left[\frac{1}{n^2M_0{}^2}\sum_{i=1}^{n}\frac{M_i^2}{P_i^2}\frac{1}{m_i}\sigma_i^2\right] - E_1\left[\frac{1}{n^2M_0{}^2}\sum_{i=1}^{n}\frac{M_i^2}{P_i^2m_i^2}\sum_{j=1}^{m_i}\left(\mu_{i[j]} - \mu_i\right)^2\right]$$

$$= \frac{1}{nM_0{}^2}\sum_{i=1}^{N}\frac{M_i^2}{P_im_i}\sigma_i^2 - \frac{1}{n^2M_0{}^2}E_1\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{M_i^2}{P_i^2m_i^2}\left(\mu_{i[j]} - \mu_i\right)^2\right] \tag{12}$$

Adding (11) and (12) gives the variance of $\bar{y}_{ppsrss}$ as:

$$\frac{1}{nM_0{}^2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i} - Y\right)^2 + \frac{1}{nM_0{}^2}\sum_{i=1}^{N}\frac{M_i^2}{P_i}\frac{\sigma_i^2}{m_i} - \frac{1}{n^2M_0{}^2}E_1\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{M_i^2}{P_i^2m_i^2}\left(\mu_{i[j]} - \mu_i\right)^2\right] \tag{13}$$

Now, if the number of cycles is r_i instead of one, (13) would have turned out to be (9).

4. Relative Efficiency

Relative efficiency of a sampling design ξ_1 over another ξ_2 based on an estimator $\hat{\theta}$ of a population parameter θ is a measure of relative overall quality of the designs evidenced in their estimators. Algebraically, the relative efficiency of ξ_1 over ξ_2, based on $\hat{\theta}$ is obtained by:

$$RE(\xi_1 \left| \xi_2 \right.) = \frac{\operatorname{var}\left(\hat{\theta}_{\xi_2}\right)}{\operatorname{var}\left(\hat{\theta}_{\xi_1}\right)} \tag{14}$$

where var(.) is a measure of variability of the estimators obtained from the two designs. Using (14), ξ_1 will be adjudged a more efficient design if $RE(\xi_1 \left| \xi_2 \right.)$ is greater than 1 and less efficient if otherwise.

The proposed sampling design and its associated estimators are applied to the greenhouses data obtained in the 2003 agricultural survey conducted in Iran as adopted from Nematollahi et al. (2008). The provinces or a set of provinces are considered as first stage units (FSU's) and greenhouses as second stage units (SSU's). For us to estimate the mean value of the greenhouses products and subsequently compare our proposed sample mean in (6) with the mean estimator ($\hat{\mu}^r_{TSCRSS}$) proposed by Nematollahi et al. (2008) for relative efficiency, a simulation study is carried out on this data. The sampling units are ranked based on the values of the greenhouses in the frame, and the ranking is assumed to be flawless. The study variable is also the same as the greenhouses values in our simulation survey, consequently, the sizes of the second stage units M_i are used as our measure of sizes.

4.1. Layout of the data selection

In this study, there are *N* = 25 first stage units (FSU's) or provinces in the frame. And every *i*th province contains a total of $M_i; i = 1, \ldots, N$ greenhouses that are regarded as second stage units as they appeared in Table 2. For the sake of demonstration of the methodology for the proposed estimator of the mean, a random sample of size *n* = 5 first stage units are selected from the population of *N* = 25 clusters, using unequal probability sampling (PPSWR). The FSU's selected in the first stage of sampling via PPSWR are marked asterisks (*) in Table 3. Out of every *i*th selected province, $m = rm'$ greenhouses (SSU's) were selected by RSS. This paper considers where $r = 4$ and $m' = 3$ to get a ranked set sample of size 12 units each in the second stage sampling.

Similarly, for the estimator according to Nematollahi et al. (2008), a random sample of size *n* = 5 is also selected from the population by simple random sampling without replacement. Out of every *i*th chosen FSU, a sample of SSU's, $m = rm'$ is selected by RSS. Also, $r = 4$ and $m' = 3$ are considered to get a ranked set sample of size 12 units each in the second stage sampling.

Table 2. The number of secondary sampling units in the first stage units

FSU's	M_i	FSU's	M_i	FSU's	M_i	FSU's	M_i	FSU's	M_i
1	42	6	61	11	27	16	750	21	30
2	169	7	680	12	26	17	32	22	26
3	538	8	936	13	14	18	275	23	18
4	38	9	167	14	40	19	14	24	14
5	33	10	20	15	93	20	20	25	84

Table 3. Cumulative table for selection of 5 provinces by PPSWR

FSU's	M_i	Cum. of M_i's	Prob (M_i)	FSU's	M_i	Cum. of M_i's	Prob (M_i)
1	42	42	0.010127803	14	40	2791	0.009645527
2	169	211	0.040752351	15	93	2884	0.022425850
3	538	749	0.129732337	16*	750	3634	0.180853629
4	38	787	0.009163251	17*	32	3666	0.007716422
5	33	820	0.007957560	18*	275	3941	0.066312997
6	61	881	0.014709429	19	14	3955	0.003375934
7*	680	1561	0.163973957	20	20	3975	0.004822763
8*	936	2497	0.225705329	21	30	4005	0.007234145
9	167	2664	0.040270075	22	26	4031	0.006269592
10	20	2684	0.004822763	23	18	4049	0.004340487
11	27	2711	0.006510731	24	14	4063	0.003375934
12	26	2737	0.006269592	25	84	4147	0.020255606
13	14	2751	0.003375934				

4.2. Computation of estimated means for the two competing designs

In order to obtain the estimated means and totals using the Nematollahi et al. (2008) estimators and the new proposed estimators, their mean estimators and computations are presented as follows:

$$\hat{\mu}^{r}_{TSCRSS} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} \sum_{l=1}^{r_i} \sum_{j=1}^{m_i'} \frac{M_i}{r_i m_i'} Y^{(j)}_{il[j]} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} M_i \hat{\mu}_i \qquad (15)$$

where $\bar{M} = \sum_{i=1}^{N} \frac{M_i}{N}$ and $\hat{\mu}_i = \frac{1}{r_i m_i'} \sum_{l=1}^{r_i} \sum_{j=1}^{m_i'} Y_{il[j]}^{(j)}$. The terms of the mean estimator in (15) are computed and presented in Table 4. Using the computed terms, the mean is estimated as $\hat{\mu}_{TSCRSS}^{r}$ = 28.25125.

Table 4. Calculation of the estimated population mean in Nematollahi et al. (2008)

FSU's	M_i	$m_i = r_i m_i'$	$\hat{\mu}_i$	$M_i \hat{\mu}_i$
13	38	12	14.0000	532.0000
8	938	12	12.0000	11256.0000
10	20	12	13.8333	276.6667
16	750	12	12.5833	9437.5000
2	169	12	11.4167	1929.4167

Also, the terms contained in the new proposed estimator of the mean in (6) are computed for the sample in Table 5 and the mean is estimated as $\bar{y}_{ppsrss}$ = 20.2204.

Table 5. Calculation of the estimated population mean using the new mean estimator

FSU's	M_i	p_i	$m = rm'$	y_i	$\frac{y_i}{p_i}$	$\bar{y}_i$	$\frac{M_i \bar{y}_i}{p_i}$
7	680	0.162213740	12	160	986.3529	13.33	55893.33
16	750	0.178912214	12	140	782.5067	11.67	48906.67
8	936	0.223282443	12	170	761.3675	14.17	59386.67
17	32	0.007633588	12	158	200698.0000	13.17	55194.67
18	275	0.065601145	12	154	2347.5200	12.83	53797.33

The entire process of sampling and computation is carried out using appropriate packages in the *R* statistical software.

It has been shown that the mean estimator proposed in this paper is unbiased. As a result, the appropriate measure of its variability to be used in this section is the variance. However, to ensure uniformity of computation with the estimator due to Nematollahi et al. (2008), the mean squared error (MSE) is used. Now, MSE of the two competing estimators are obtained empirically using 10000 replications of samples and computations. Precisely, the MSE of each estimator of the mean is obtained by:

$$MSE(\bar{y}_b) = \frac{1}{10000} \sum_{a=1}^{10000} (\bar{y}_{ab} - \mu)^2 \; ; \; b = 1, 2 \qquad (16)$$

where $\overline{y}_{a1}$ and $\overline{y}_{a2}$ denote $\overline{y}_{ppsrss}$ and $\hat{\mu}^{r}_{TSCRSS}$ respectively in ath replication of the sample, $a = 1, 2, \ldots, 10000$. The results for first stage sample sizes n = 4, 5, 6 and 8 are presented in Table 6 for the new estimator as MSE_1 and Nematollahi et al. (2008) estimator as MSE_2.

Table 6. Mean square errors corresponding to each mean estimator $\overline{y}_{ppsrss}$ and $\hat{\mu}^{r}_{TSCRSS}$

FSU Sample Size	r = 3, m' = 2		r = 3, m' = 3		r =3, m' = 4	
	MSE₁	MSE₂	MSE₁	MSE₂	MSE₁	MSE₂
4	227.4776	89.4379	193.0703	90.0924	204.7227	91.5629
5	55.4979	69.9861	51.0138	69.3528	55.0475	70.0885
6	25.8019	56.7316	26.5967	54.9603	26.0757	54.2575
8	9.2579	37.5100	9.0912	37.3246	9.2049	36.5891

Finally, the relative efficiency of the new estimator $\overline{y}_{ppsrss}$ to $\hat{\mu}^{r}_{TSCRSS}$ at different sample sizes in the two stages of the sampling designs are obtained by:

$$RE = \frac{MSE\left(\hat{\mu}^{r}_{TSCRSS}\right)}{MSE\left(\overline{y}_{ppsrss}\right)} \tag{17}$$

The computed relative efficiencies are presented in Table 7.

Table 7. Relative efficiencies of the new estimator $\overline{y}_{ppsrss}$ to $\hat{\mu}^{r}_{TSCRSS}$

Number of selected FSU's	r = 3 m' = 2	r = 3 m' = 3	r = 3 m' = 4
4	0.393	0.467	0.447
5	1.261	1.359	1.273
6	2.199	2.067	2.080
8	4.052	4.105	3.9749

From the results in Table 7, the relative efficiency of the new estimator compared to Nematollahi et al. (2008) estimator shows that the new estimator is more efficient for different sizes of first stage units sample except, in the case when n = 4. This suggests that as the sample size in the first stage of sampling increases, the relative efficiency of the new estimator keeps improving. For instance, when n = 5 in the first stage and m = 5 in the second stage, the relative efficiency improves from 0.39 to 1.26. Similarly, when n = 8 and m = 12, it changes from 2.08 when n = 6 and m = 12 to 3.97. However, it is important to note that the preferred performance of the new estimator to the Nematollahi et al. (2008) estimator may have been because the population in question has significantly varied sizes in the first stage units. If a situation where somewhat equality in sizes of the FSU's is encountered, this preference may not be guaranteed.

5. Conclusion

A new two-stage sampling design has been developed where probability proportional to size sampling with replacement (PPSWR) is used in the first stage and ranked set sampling is used in the second stage. The empirical comparative study carried out revealed that our new sampling design is more efficient as it produced better estimator for estimating the population mean than similar design built with simple random sampling in the first stage and ranked set sampling in the second stage units under the condition of significant variation in the sizes of the first stage units.

Acknowledgements

The authors wish to thank the two anonymous reviewers and the managing editor for their comments and criticisms which have greatly improved the quality of this paper.

References

Al-Omari A. I., Bouza C. N., (2014). Review of Ranked Set Sampling: Modifications and Applications. *Revista Investigacion Operacional*, 35(3), pp. 215–240.

Chen Z., Bai Z., Sinha B. K., (2003). *Ranked Set Sampling: Theory and Applications.* Springer, New York.

Damon, V., (2018). Advantages and disadvantages of multistage sampling. https://classroom.synonym.com/advantages-disadvantages-multistage-sampling-8544049.html

Dell, T. R., Clutter, J. I., (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28, pp. 545–553.

Dilip, N. C., (2015). Two-stage sampling design for estimation of total fertility rate: with an illustration for slum dweller married woman. *Electronic Journal of Applied statistical Analysis,* 8(1), pp. 112–121.

Fears, T. R., Gail, M. H., (2000). Analysis of a two-stage case-control study with cluster sampling of controls: Application to Nonmelanoma skin cancer. *Biometrics,* 56(1), pp. 190–198.

Galway, L. P., Bell, N., Shatari, S. AE. Al., Hagopian, A., Burnham, G., Flaxman, A., Weiss, W. M., Rajaratnam, J. and Takaro, T. K., (2012). A two-stage cluster sampling method using gridded population data, a GIS and Google Earth TM imagery in population-based mortality survey in Iraq. *International Journal of Health Geographics,* 11(12), pp. 1–9.

Halls, L. S., Dell, T. R., (1966). Trial of ranked set sampling for forage yields. *Forest Science,* 12(1), pp. 22–26.

Hansen, M. H., Hurwitz, W. N., (1943). On the theory of sampling from a finite population. *Annals of Mathematical Statistics,* 14, pp. 333–362.

Horney, J. J., Dickinson, M., Hsai, J., Williams, A. and Zotti, M., (2010). Two-stage cluster sampling with referral: Improving the efficiency of estimating unmet needs among pregnant and postpartum women after flooding in Northwest Georgia. *Remote Sensing of Environment,* 113(6), pp. 1236–1249.

Innocenti, F., Candel, M. J. J. M., Tan, F. E. S. and van Breukelen, G. J. P., (2019). Relative efficiencies of two-stage sampling schemes for mean estimation in multilevel populations when cluster size is informative. *Statistics in Medicine,* 38(10), pp. 1817–1834.

Innocenti, F., Candel, M. J. J. M., Tan, F. E. S. and van Breukelen, G. J. P., (2021). Optimal two-stage sampling for mean estimation in multilevel populations when cluster size is informative. *Statistical Methods in Medical Research,* 30(2), pp. 357–375.

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

Nematollahi, N., Salehi, M. M. and Saba, A. R., (2008). Two-stage cluster sampling with ranked set sampling in the secondary sampling frame. *Communications in Statistics–Theory and methods,* 37(15), pp. 2404–2415.

Okafor, F. C., (2002). *Sample Survey Theory with Application.* Afro-Orbis Publications Ltd. Nsukka.

Ozturk, O., (2019). Two-stage cluster samples with ranked set sampling designs. *Annals of the Institute of Statistical Mathematics,* 71, pp. 63–91.

Patil, G. P., (2002). Ranked set sampling. *Encyclopedia of Environmetrics,* 3, pp. 1684–1690.

Phillips, A. E., Boily, M. C., Lowndes, C. M., Garnett, G. P., Gurav, K., Ramesh, B. M., Anthony, J., Watts, R., Moses, S. and Alary, M., (2008). Sexual identity and its contribution to MSM risk behaviour in Bangaluru (Bangalore) India: The results of a two-stage cluster sampling survey. *Journal of LGBT Health Research,* 4, pp. 111–126.

Shahbaz, M. Q., Hanif, M., (2010). *Some developments in unequal probability sampling: selection procedures and estimators.* Lap Lambert Academic Publishing, GmbH & Co. KG, Deutschland.

Stehman, S. V., Wichham, J. D., Fattorini, L., Wade, T. D., Baffetta, F. and Smith, J. H., (2009). Estimating accuracy of land-cover composition from two-stage cluster sampling. *Remote Sensing of Environment,* 113(6), pp. 1236–1249.

Stokes, S. L., (1980). Estimation of variance using judgment ordered ranked set samples. *Biometrics*, 36, pp. 35–42.

Takahasi, K., Wakimoto, K., (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 21, pp. 249–255.

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 215–216,

On behalf of the Department of Statistical Methods of the University of Lodz we are pleased to announce the organization of the ***40th International Conference MSA'2022 joined with MASEP*** by the Department of Statistical Methods, Department of Economic and Social Statistics, Institute of Statistics and Demography of the University of Lodz, Polish Academy of Sciences and Polish Statistical Association.

The conference will take place on November 7–9, 2022 at the Training and Conference Center of the University of Lodz, Kopcińskiego 16/18 St., 90-232 Lodz.

Registration deadline is September 30, 2022.

The honorary patronage was provided by Rector of University of Lodz and the President of Statistics Poland.

Scientific Committee

Prof. Czesław Domański, University of Lodz – Committee Chair

Prof. dr hab. Sławomir Bukowski, Kazimierz Pułaski University of Technology and Humanities in Radom

Assoc. Prof. Francesca Greselin, University of Milano Biococca

Prof. Marie Huskova, Charles University in Prague

Prof. Krzysztof Jajuga, Wrocław University of Economics and Business

Prof. Mirosław Krzyśko, Adam Mickiewicz University in Poznan

Prof. Włodzimierz Okrasa, Cardinal Stefan Wyszyński University in Warsaw

Prof. Józef Pociecha, Cracow University of Economics

Prof. Mirosław Szreder, University of Gdansk

Prof. Janusz Wywiał, University of Economics in Katowice

Organizing Committee

Alina Jędrzejczak, prof., University of Lodz – Committee Chair

Marta Małecka, Ph.D., University of Lodz, e-mail: marta.malecka@uni.lodz.pl

Artur Mikulec, Ph.D., University of Lodz, e-mail: artur.mikulec@uni.lodz.pl

Elżbieta Zalewska, Ph.D., University of Lodz, e-mail: elzbieta.zalewska@uni.lodz.pl
Łukasz Ziarko, Ph.D., University of Lodz, e-mail: lukasz.ziarko@uni.lodz.pl

For more information please refer the conference website
https://sites.google.com/view/msa-lodz

STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 217–220,

About the Authors

Beghriche Abdelfateh is a faculty member at the University the Brothers Mentouri Constantine 1, Algeria. He received his PhD degree in Mathematics from the Department of Mathematics, Faculty of Exact Sciences, University the Brothers Mentouri Constantine 1. His research areas are in applied statistics and mathematical sciences.

Bhattacharya Rahul is a Full Professor of Statistics at the Department of Statistics, University of Calcutta, India. His research interests are statistical inference, biostatistics, reliability theory, data analysis and statistical computing, among others. Professor Bhattacharya has published over 50 research papers in international/national journals and conferences. Professor Bhattacharya is an active member of Calcutta Statistical Association.

Białek Jacek is a Professor at the University of Lodz in Poland, where he is a long-term employee of the Department of Statistical Methods. His research interests are: the theory and practice of price indices and the measurement of CPI and HICP. At the same time, he works at Statistics Poland, where he deals with the analysis of scanner and web-scraped data at the Department of Trade and Services. Jacek Białek is an author of over 90 scientific papers and the PriceIndices R-package, which is used for analysing scanner data and determining price indices. He is a member of the Editorial Board in International Journal of Statistics and Probability.

Choudhury Mriganka Mouli received his BSc degree in Mathematics from Calcutta University in 2008. He received MSc in Statistics and MPhil in Statistics & Computer application from West Bengal State University and Calcutta University in 2010 and 2012 respectively. He completed his PhD degree in Reliability Theory from Visva-Bharati University in 2021 and is currently working as an Assistant Teacher in Homra Palta High School, West Bengal. His research interest covers reliability theory, industrial statistics and statistical inference.

Chouia Sarra is a doctoral student of mathematics class at Badji Mokhtar University Annaba, Algeria. She received her MSc degree in Mathematics from Badji Mokhtar University. Her research areas are in applied statistics and actuarial science.

Herman Sergiusz is an Assistant Professor at the Department of Econometrics, Institute of Informatics and Quantitative Economics, Poznań University of Economics and

Business. His research interests are multivariate data analysis and classification methods.

Kwon Yeil is an Assistant Professor at the Department of Mathematics at the University of Central Arkansas in the US. His main areas of interest include Bayesian estimation, multiple testing, and Fibonacci-type probability distributions. One of his recent studies on the empirical Bayesian estimator for multiple variances was published in Biometrika. He is also interested in statistical education and works on various projects for statistical education with faculty members in the Mathematics Education Department.

Madukaife Mbanefo S. is a Senior Lecturer in the Department of Statistics, Faculty of Physical Sciences at the University of Nigeria, Nsukka. His main research interests include goodness-of-fit techniques of multivariate statistical distributions, characterisations of multivariate statistical distributions, multivariate statistical inference, multivariate computational statistics, classification and clustering techniques, as well as sample survey methodology. Dr. Madukaife is an active member of many scientific professional bodies.

Maiti Sudhansu S. is a Professor at the Department of Statistics, Visva-Bharati University. He received his undergraduate, postgraduate and PhD degree from Calcutta University. His research interest includes reliability/survival analysis, industrial statistics, distribution theory, information theory and Bayesian inference.

Nagarajah Varathan is currently working as a Senior Lecturer in Statistics at the Department of Mathematics and Statistics, Faculty of Science, University of Jaffna, Sri Lanka. He received his PhD degree in Statistics from the University of Peradeniya, Sri Lanka. Dr. Varathan received his MSc degree from Memorial University of Newfoundland, Canada, and his BSc degree from the University of Jaffna, Sri Lanka. His research interests are logistic regression, time series, and statistical inference in particular. Dr. Varathan has published over 30 research papers in international/national journals and conferences.

Nasiri Parviz is an Associate Professor at the Department of Statistics, Faculty of Science, University of Payame Noor. His research interests are distribution theory, multivariate statistical analysis, Bayesian inference, statistics modelling, statistical inference and data analysis in particular. Dr. Nasiri has published over 80 research papers in international/national journals and conferences. He has also published five books.

Panek Tomasz is a Full Professor at the Institute of Statistics and Demography, where he is Vice-Director. His research interests focus primarily on the issues of household living conditions (poverty, social inequalities, income inequalities and quality of life) as well as electronic media consumption and inflation measurement. Author, co-author and editor of over a hundred studies published both nationally and internationally. He has participated in dozens of national and international research projects and has led them many times. Full member, inter alia, of the International Statistical Institute, member of the International Association of Survey Statisticians and member of the Scientific Statistical Council of Statistics Poland. Associate Editor of the Statistical News and member of Editorial Board of the Polish Statistician. Member of the Program Council and the Scientific Council of the United Nations Global Compact in Poland.

Permpoonsinsup Wachirapond is an Assistant Professor at the Industrial Technology and Innovation Management Program, Faculty of Science and Technology, Pathumwan Institute of Technology. Her primary interests lie in mathematical modelling, applied statistics, artificial intelligent, optimization and computer science to address engineering and business problems and experimental mathematics and statistics using computers. Her research interests are not only to simply solve problems within the confines of a particular area, but also to look for connections with other areas to apply in the real-world problem.

Sunthornwat Rapin is an Assistant Professor of Statistics at Industrial Technology and Innovation Management Program, Faculty of Science and Technology, Pathumwan Institute of Technology, Bangkok, Thailand. His research interests are statistical quality control, time series analysis, mathematics and statistics for finance and accounting, abstract algebra for algebraic statistics, and differential and integral equations. He has focused in mathematics and statistics for COVID-19 problems. For academic activities, he is a reviewer in international journals such as Thailand Statistician and he is a member of Thai Statistical Association.

Szulc Adam is an Associate Professor at the Institute of Statistics and Demography of the Warsaw School of Economics. His fields of interests cover poverty and inequality measurement, consumer demand systems, equivalence scales, social policy evaluations, matching estimation. He is a member of the review boards of the Equilibrium – Quarterly Journal of Economics and Economic Policy and of the Argumenta Oeconomica.

Ugwu Michael C. is a postgraduate student in the Department of Statistics, Faculty of Physical Sciences at the University of Nigeria, Nsukka. His main research interests include sample survey methodology and inferential statistics.

Vinoth Raman is an Assistant Professor at the Deanship of Quality and Academic Accreditation Department at Imam Abdulrahman Bin Faisal University, Kingdom of Saudi Arabia. He received his PhD degree in Statistics from Annamalai University, Annamalainagar, Chidambaram, India. Doctor Vinoth has published over 60 research papers in international/national journals and conferences. His research areas are in statistical analysis, biostatistics and applied statistics.

Zaborski Artur is an Associate Professor at the Department of Econometrics and Computer Science, Faculty of Economics and Finance, Wroclaw University of Economics and Business. His main areas of interest include multivariate statistical analysis, in particular methods of multidimensional scaling, preference studies, measurement of similarity.

Zeghdoudi Halim is a faculty at the Department of Mathematics at the University of Badji-Mokhtar, Annaba-Algeria. He received his PhD degree in Mathematics and the highest academic degree (HDR) specializing Probability and Statistics from Badji-Mokhtar University, Annaba-Algeria. He also did his Post Doc at Waterford Institute of Technology – Cork Rd, Waterford, Ireland. Professr Zeghdoudi has published over 100 research papers in international/national journals and conferences. His research areas are in actuarial science, particles systems, dynamics systems and applied statistics. Currently, he is a member of two editorial boards: Frontiers in Applied Mathematics and Statistics – Asian Journal of Probability and Statistics.

Zwierzchowski Jan is an Associate Professor at the Institute of Statistics and Demography, Collegium of Economic Analysis, Warsaw School of Economics. His main areas of interest include: survey design, quality of life, economic inequality and poverty measurement. Currently, he is a member of the Editorial Board of Studia Demograficzne.

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).