**Volume 21, Number 5, December 2020** **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*
Secretary
Patryk Barszcz, *Statistics Poland, Warsaw, Poland, e-mail: p.barszcz@stat.gov.pl, phone number +48 22 — 608 33 66*
Technical Assistant
Rajmund Litkowiec, *Statistical Office in Rzeszów, Rzeszów, Poland, e-mail: r.litkowiec@stat.gov.pl*

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

## CONTENTS

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. III–VII*

# From the Editor

With this release of the *Statistics in Transition new series*, we conclude the 2020 edition of our quarterly issued journal, which this time has been expanded due to publishing recently an extraordinary, special issue devoted to statistical data integration. An international team of experts led by Partha Lahiri of the University Maryland – who served as a Guest Editor of the special issue – has succeeded with arranging for a set of original papers addressing frontiers in theoretical and application aspects of multiple data sources creation and use, delivered by leaders in the relevant topics. It begun with the invited paper by Malay Ghosh on small area estimation during the past decades, based on his 2019 Morris Hansen lecture:
https://sit.stat.gov.pl/SiT/SpecialIssue/August%202020/gus_sit_2020_04_special_issue.pdf.
The topics covered by the papers were grouped in four categories, as follows: (i) small area estimation, (ii) advances in probabilistic record linkage and analysis of linked data, (iii) statistical methods for longitudinal data, multiple-frame, and data fusion, and (iv) synthetic data for microsimulations, disclosure avoidance and multi-purpose inferences.

Traditionally, as the last in the annual cycle of publication, this issue provides us with the opportunity to express gratitude to all contributors to our success, i.e. to publishing articles of high quality guaranteed, among other things, by the participation of outstanding experts as reviewers in the double-blind review process. A list of the names of these people of merit for our journal is included in the Acknowledgements. On behalf of the Editorial Board, Associate Editors and the journal's readers I sincerely thank to all our partners and patrons.

The set of eight original scientific articles that make up this issue is opened by the article ***Estimating the population mean using a continuous sampling design dependent on an auxiliary variable*** by **Janusz Wywiał.** Its purpose is to estimate the mean of the variable under study using a sampling design which is dependent on the observation of a continuous auxiliary variable in the whole population. Auxiliary variable values observed in this population allow one to estimate the inclusion density function of the sampling design. The variance of the continuous version of the Horvitz-Thompson estimator under the proposed sampling design is compared with the variance of the mean of a simple random sample. The accuracy of the estimation strategies is analysed by means of simulation experiments.

In the paper entitled ***The Gamma Kumaraswamy-G family of distributions: theory, inference and applications*****, Rana Muhammad Imran Arshad, Muhammad Hussain Tahir, Christophe Chesneau,** and **Farrukh Jamal** introduce a new family of univariate continuous distributions called the Gamma Kumaraswamy-generated family of distributions. Most of its properties are studied in detail, including skewness, kurtosis, analytical comportments of the main functions, moments, stochastic ordering and order statistics, followed by a particular member of the family with four parameters, called the gamma Kumaraswamy exponential distribution. It has several advantages, including the corresponding probability density function which can have symmetrical, left-skewed, right-skewed and reversed-J shapes, while the corresponding hazard rate function can have (nearly) constant, increasing, decreasing, upside-down bathtub, and bathtub shapes. The inference on the gamma Kumaraswamy exponential model is performer using the method of maximum likelihood to estimate the model parameters. In order to demonstrate the importance of the new model, analyses on two practical data sets were carried out showing that the proposed model prevails over any of the other eight competitive models.

**Warisa Thangjai's** and **Suparat Niwitpong's** paper on ***Comparing particulate matter dispersion in Thailand using the Bayesian Confidence Intervals for ratio of coefficients of variation*** addresses the problem of measuring air pollution detected in Thailand. A high dispersion of PM is measured by a coefficient of variation of log-normal distribution applied to environmental data such as hazardous dust particle levels and daily rainfall data. The authors develop confidence interval estimation for the ratio of coefficients of variation of two log-normal distributions constructed using the Bayesian approach, and compare them with the existing approaches: the method of variance estimates recovery (MOVER), modified MOVER, and approximate fiducial approaches using their coverage probabilities and average lengths via Monte Carlo simulation. The simulation results show that the Bayesian confidence interval performed better than the others in terms of coverage probability and average length. The proposed approach and the existing approaches are illustrated using examples from data selected regions in the northern Thailand.

The next article, ***A new generalization of the Pareto distribution and its applications*** by **Ehab M. Almetwally** and **Hanan A. Haj Ahmad** takes up the problem of generalization of the Pareto distribution using the Marshall-Olkin generator and the method of alpha power transformation. The Authors demonstrate several desirable properties due to which the new model is appropriate for modelling right skewed data and how the hazard rate function and moments are obtained. Also, an estimation for the new model parameters is provided, through the application of the maximum likelihood and maximum product spacing methods, as well as the Bayesian estimation. Approximate confidence intervals are obtained by means of an asymptotic property of

the maximum likelihood and maximum product spacing methods, while the Bayes credible intervals are found by using the Monte Carlo Markov Chain under different loss functions. A simulation analysis is conducted to compare the estimation methods. Finally, the application of the proposed new distribution to three real-data examples is presented and its goodness-of-fit is demonstrated. Some comparisons to other models are made in order to prove the efficiency of the distribution under consideration including better data fit than some other sub models.

**Shakti Prasad's** paper ***Some linear regression type ratio exponential estimators for estimating the population mean based on quartile deviation and deciles*** deals with some linear regression type ratio exponential estimators for estimating the population mean using the known values of quartile deviation and deciles of an auxiliary variable in survey sampling. The expressions of the bias and the mean square error of the suggested estimators have been derived and comparison was made with the usual mean, usual ratio (Cochran (1977)), Kadilar and Cingi (2004, 2006) and Subzar et al. (2017) estimators. After the comparison, the condition which makes the suggested estimators more efficient than others is found. To verify the theoretical results, numerical results are performed on two natural population data sets.

In the next paper, ***Modelling bid-ask spread conditional distributions using hierarchical correlation reconstruction,*** **Jarosław Duda, Robert Syrek** and **Henryk Gurgul** discuss the problem of prediction of the exact values given that the information available is rarely sufficient; consequently, only conditional probability distributions are possible to be predicted. Hierarchical correlation reconstruction (HCR) methodology is used for such a prediction starting with normalized marginal distributions, nearly uniform. Next, joint densities are modelled as linear combinations of orthonormal polynomials, obtaining their decomposition into mixed moments. Each moment of the predicted variable is modelled separately as a linear combination of mixed moments of known variables using least squares linear regression. By combining these predicted moments, the predicted density is obtained as a polynomial, for which the expected value and other characteristics are calculated. An advantage of using this methodology is also its computational efficiency; estimating and evaluating a model with hundreds of parameters and thousands of data points by means of this methodology takes only a second on a computer, at relatively low-cost.

**Adetola Adedamola Adediran, Femi Barnabas Adebola, Olesegun Sunday Ewemooje** are discussing ***Unbiased estimator modelling in unrelated dichotomous randomized response*** constructed by incorporating an unrelated question into the alternative unbiased estimator in the dichotomous randomized response model (proposed by Ewemooje in 2019). An unbiased estimate and variance of the model are obtained, and the latter decreases as the proportion of the sensitive attribute π_A and the unrelated attribute π_U increases. The relative efficiency of the proposed model

over the earlier model (by Ewemooje) decreases as π_U increases and increases as π_U increases. Application of the proposed model also revealed its efficiency over the direct method in estimating the prevalence of examination malpractices among university students; for instance, the direct method gave an estimate of 19.0 percent, compared to the proposed method's estimate of 23.0 percent.

In the next paper, ***A Bayesian analysis of complete multiple breaks in a panel autoregressive (CMB-PAR(1)) time series model*** by **Varun Agiwal, Jitendra Kumar, Dahud Kehinde Shangodoyin** discussed is the problem of economic time series – such as GDP, real exchange rate and banking series – which are irregular by nature and often affected by a variety of discrepancies such as: political changes, policy reforms, import-export market instability, etc. The Authors propose to manage this problem using a generalised structural break time series model. The Bayesian approach is applied to estimate the model parameters and to obtain the highest posterior density interval. Strong evidence is observed to support the Bayes estimator and then it is compared with the maximum likelihood estimator. A simulation experiment is conducted and an empirical application on the SARRC association's GDP per capita time series is used to illustrate the performance of the proposed model.

In the section Other articles, there is one article based on conference presentation (Multivariate Statistical Analysis 2019, Łódź) by **Czesław Domański** and **Piotr Szczepocki** entitled ***Comparison of selected tests for univariate normality based on measures of moments.*** It deals with univariate normality, tests which are typically classified into tests based on empirical distribution, moments, regression and correlation, and other. The Authors present results of power comparisons of nine normality tests based on measures of moments via the Monte Carlo simulations. The effects on power of the sample size, significance level, and on the number of alternative distributions are investigated. None of the considered tests proved uniformly most powerful for all types of alternative distributions. However, the most powerful tests for different shape departures from normality (symmetric short-tailed, symmetric long-tailed or asymmetric) are indicated.

In the section containing articles classified as research communicates there are two papers. In the first one, ***Predicting Polish transport industry equilibrium characteristics as an inverse problem: An Entropy Econometrics Model*** by **Second Bwanakare** and **Marek Cierpiał-Wolan** the problem of decision-making process in the business environment is discussed, given that it is governed by a high degree of uncertainty and risk. Moreover, when detailed statistical information relating to the industry is missing, any decisions may become a matter of highly risky conjectures. The Authors propose a simultaneous equation model based on the entropy econometrics estimator for recovering some key industrial subsector long-term equilibrium characteristics under condition that only sparse, insufficient statistical

information is available (e.g. only aggregated data on the whole industry). The model is applied to the transportation equipment manufacturing industry in Poland, which is composed of eight sub-sectors, showing that all firms from different sub-sectors have to increase their steady-state concentration ratios, while the highest concentration corresponds to the lowest increase in profitability. The model outputs conform to the market tendency in this sector and should lead to further applications of the NCEE methodology in business activity on a world-wide scale.

The paper by **Piotr Zawada, Włodzimierz Okrasa** and **Jack Warchalowski** entitled ***Flow management system for maximising business revenue and profitability*** starts with an observation that most for-profit organisations must constantly improve their business strategies and approaches to remain competitive. Many of them choose to embark on Lean or Six Sigma journeys with the intention of maximising productivity and increasing sales. Despite a significant progress in the development of the Big 3 Improvement Methodologies (Lean, Six Sigma, Theory of Constraints (TOC)), many manufacturers still involve themselves in ineffective operations, resulting in longer-than-desired lead times, late deliveries, high inventories and considerable operational costs. All of these issues seriously challenge the company's competitiveness. The aim of the paper is to demonstrate the importance of effective analysis of maintaining certain level of inventory to gain a competitive advantage and using the company's key resources in the competitive struggle on the market while conducting continuous reporting of reasons for not achieving the assumed business goals.

**Włodzimierz Okrasa**

Editor

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. IX*

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

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. XI–XII*

# Editorial Policy

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

***

ABSTRACTING AND INDEXING DATABASES

*Statistics in Transition new series* is currently covered in:

**Databases indexing the journal:**

- 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
- Ulrichsweb & Ulrich's Periodicals Directory
- WanFang Data
- WorldCat (OCLC)
- Zenodo.

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 1–16, DOI 10.21307/stattrans-2020-052*
Received – 17.03.2020; accepted – 07.08.2020

# Estimating the population mean using a continuous sampling design dependent on an auxiliary variable

**Janusz L. Wywiał** [1]

## ABSTRACT

Continuous distribution of variables under study and auxiliary variables are considered. The purpose of the paper is to estimate the mean of the variable under study using a sampling design which is dependent on the observation of a continuous auxiliary variable in the whole population. Auxiliary variable values observed in this population allow to estimate the inclusion density function of the sampling design. The variance of the continuous version of the Horvitz-Thompson estimator under the proposed sampling design is compared with the variance of the mean of a simple random sample. The accuracy of the estimation strategies is analysed by means of simulation experiments.

**Key words:** continuous sampling design, Horvits-Thompson estimator, inclusion density, sampling scheme, bivariate gamma distribution, ratio estimator.

## 1. Introduction

Survey sampling theory is well developed for inference based on a finite and fixed population, where the variable under study as well as auxiliary variables are non-random (see, e.g. Särndal, Swenson, Wretman (1992) and Tillé (2006)). The estimation of population parameters is based on a sampling design defined as functions of auxiliary variable values observed in the whole population.

In this paper, the auxiliary variable is also treated as random. We assume that the continuous distribution function of the variable under study and the auxiliary variable (denoted by $X$ and $Y$ respectively) is known, or can be estimated. Values of $X$ and $Y$ are observed on the whole population of size $N$ and in the sample respectively. For instance, the joint distribution of these two variables can be suggested by economic theory. Tax registers are an example of auxiliary variable observation in the whole population.

Another example deals with application of statistics in auditing. Book values of accounting documents are inspected (audited) in order to assess the true values of the documents. Calculating the mean of the true values is one of the purposes of auditing. We can consider joint continuous distribution of the book values and the true values of the documents. The book values can be treated as observations of $X$ throughout the population of the documents, while values of $Y$ are observations of the variable under study. Our aim is to estimate the mean of $Y$ based on a sample selected according to a sampling design dependent on $X$. For example, Frost and Tamura (1986) and Wywiał (2018) considered gamma distribution for modelling book values in statistical auditing.

[1] University of Economics in Katowice, Poland. E-mail: janusz.wywial@ue.katowice.pl.
ORCID: https://orcid.org/0000-0002-3392-1688.

Benhenni and Cambanis (1992) and Thompson (1997) considered continuous sampling for Monte Carlo integration. Some continuous sampling designs were studied in Bąk (2014, 2018), Wilhelm, Tillé and Qualité (2017), and Wywiał (2016). The efficiency of estimation of parameters based on stratified and systematic samples was studied by, e.g. Cressie (1993) and Zubrzycki (1958). A sampling design dependent on the positively valued continuous auxiliary variable proposed by Cox and Snell (1979) was applied to financial auditing. The continuous sampling designs and inclusion density functions were defined by Cordy (1993), who also adapted the well-known Horvitz-Thompson (1952) estimator to estimate parameters. This paper draws on these two sources. In Section 2.1, the properties of the Horvitz-Thompson statistic for the continuous sampling design are presented. Next, in Section 2.2, these properties are generalized to the joint distribution of $Y$ and $X$. A continuous sampling design with inclusion function proportional to the density function of the auxiliary variable is considered in the third chapter. In the fourth chapter, the main results of the paper are used to construct the estimation strategies under the assumption that the sample was drawn from the continuous population defined by bivariate gamma distribution. The accuracy of these strategies is studied using simulation analysis. In the last chapter, the main conclusions are formulated.

## 2. Horvitz-Thompson statistic from sample selected according to continuous sampling design

### 2.1. Basic results

This section has been prepared according to Cordy (1993) results. Let the population $U \subset R^q$, $q = 1,2,....$ To simplify our analysis we assume that $q = 1$. The sample space, denoted by $\boldsymbol{S}_n = U^n$, is the set of ordered samples denoted by $\mathbf{y} = (y_1,...,y_n)$, $y_k \in U$, $k = 1,...,n$, where $y_i$ is the outcome of the variable observed in the first draw. Let $\mathbf{y}$ be a value of the $n$-dimensional random variable $\mathbf{Y} = (Y_1,...,Y_n)$ with density function $f(\mathbf{y}) = f(y_1,...,y_n)$. Let $f_i(y)$ and $f_{i,j}(y,y')$, $y \in U$, $y' \in U$, be marginal density functions of $Y_i$ and $(Y_i,Y_j)$ respectively, $j > i = 1,...,n$. The inclusion functions of the first order and the second order are defined respectively as follows:

$$\pi(y) = \sum_{i=1}^{n} f_i(y), \qquad \pi(y,y') = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} f_{i,j}(y,y'), \quad y \in U, y' \in U \tag{1}$$

and $\int_U \pi(y)dy = n$, $\int_U \int_U \pi(y,y')dydy' = n(n-1)$.

Let $f(y_i|y_{i-1},y_{i-2},...,y_1)$, $i = 1,...,n-1$ be the conditional density function of the randomly selected $y_i$ value in the $i$-th draw (provided that the values $(y_{i-1},y_{i-2},...,y_1)$ were drawn earlier). Therefore, the density function of the sampling design can be written as follows:

$$f(y_n,...,y_i,y_{i-1},...,y_1) = f(y_1)\prod_{i=2}^{n} f(y_i|y_{i-1},y_{i-2},...,y_1) \tag{2}$$

Let $g(y)$ be an integrable function $g: U \to R$. We estimate the following parameter:

$$\theta = \int_U g(y)dy. \tag{3}$$

The continuous version of the well-known Horvitz and Thompson (1952) estimator is:

$$T_{\mathbf{Y}} = \sum_{i=1}^{n} \frac{g(Y_i)}{\pi(Y_i)} \tag{4}$$

*Theorem* 2.1. [Cordy (1993)] The statistic $T_{\mathbf{Y}}$ is an unbiased estimator for $\theta$, if the function $g(y)$ is either bounded or non-negative, and $\pi(y) > 0$ for each $y \in U$.

*Theorem* 2.2 [Cordy (1993)] If the function $g(y)$ is bounded, $\pi(y) > 0$ for each $y \in U$, and $\int_U (1/\pi(y))dy < \infty$, then

$$V(T_{\mathbf{Y}}) = \int_U \frac{g^2(y)}{\pi(y)}dy + \int_U \int_U g(y)g(y')\frac{\pi(y,y') - \pi(y)\pi(y')}{\pi(y)\pi(y')}dydy' =$$
$$= \int_U \frac{g^2(y)}{\pi(y)}dy + \int_U \int_U g(y)g(y')\frac{\pi(y,y')}{\pi(y)\pi(y')}dydy' - \theta^2. \tag{5}$$

When, in addition, $\pi(y_i, y_j) > 0$ for all $y_i, y_j \in U$, $i \neq j = 1, ..., n$, an unbiased estimator of the variance in (5) is:

$$\hat{V}(T_{\mathbf{Y}}) = \sum_{i=1}^{n} \frac{g^2(Y_i)}{\pi^2(Y_i)} + \sum_{i=1}^{n} \sum_{j=1, i\neq j}^{n} g(Y_i)g(Y_j)\frac{\pi(Y_i,Y_j) - \pi(Y_i)\pi(Y_j)}{\pi(Y_i,Y_j)\pi(Y_i)\pi(Y_j)}$$

In particular, when $h(y)$ is a density function and $g(y) = \eta(y)h(y)$, then $\theta = E(\eta(Y))$. Of course if $\eta(y) = y$, then $\theta = E(Y)$.

When $Y_1, \ldots, Y_n$ is a random sample from a distribution with density $f(y)$, then the density function of the sampling design defined by (2) and its inclusion functions become as follows:

$$f(y_1, ..., y_n) = \prod_{i=1}^{n} f(y_i), \quad \pi(y) = nf(y), \quad \pi(y,y') = n(n-1)f(y)f(y'). \tag{6}$$

This allows us to transform expressions (4) and (5) as follows:

$$T_{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\eta(Y_i)h(Y_i)}{f(Y_i)}, \quad E(T_{\mathbf{Y}}) = \theta, \tag{7}$$

$$V(T_{\mathbf{Y}}) = \frac{1}{n}\left(\int_U \frac{\eta^2(y)h^2(y)}{f(y)}dy - \theta^2\right) =$$
$$= \frac{1}{n}\left(E\left(\frac{\eta^2(Y)h^2(Y)}{f^2(Y)}\right) - E^2\left(\frac{\eta(Y)h(Y)}{f(Y)}\right)\right) = \frac{1}{n}V\left(\frac{\eta(Y)h(Y)}{f(Y)}\right). \tag{8}$$

Sampling design $f(y_n,...,y_1)$, given by (6) provides what is known as the *importance sample* considered, e.g. by Bucklew (2004) and Ripley (1987). When the importance sample is drawn from density $h(y)$, then it becomes the well-known simple random sample defined as the sequence of independent and identically distributed random variable (see e.g. Wilks (1962)) and $\theta = E(Y) = \mu_y$ is estimated by means of the following statistic:

$$T_Y = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad V(T_Y) = V(\bar{Y}) = \frac{1}{n}V(Y) \tag{9}$$

where $V(Y) = \int_{-\infty}^{\infty}(y - E(Y))^2 f(y)dy$.

### 2.2. Estimation using auxiliary variable

Let $h(x,y)$, $(x,y) \in U \subseteq R^2$, be the density function. The marginal densities are: $h_1(x)$ and $h_2(y)$. $h(y|x) = h(x,y)/h_1(x)$ is the conditional density. Moreover, $\mu_y = E(Y) = \int_{-\infty}^{-\infty} yh_2(y)dy$, $\mu_x = E(X) = \int_{-\infty}^{-\infty} xh_1(x)dx$, $E(Y|x) = \int_{-\infty}^{-\infty} yh(y|x)dy$, $V(Y|x) = \int_{-\infty}^{-\infty}(y - E(Y|x))^2 h(y|x)dy$. Our purpose is estimation of parameter $\theta$, given by (3) where

$$g(x) = E(\eta(Y)|x)h_1(x) = h_1(x)\int_{-\infty}^{\infty}\eta(y)h(y|x)dy.$$

We set $\eta(y) = y$. Therefore:

$$g(x) = E(Y|x)h_1(x) = h_1(x)\int_{-\infty}^{\infty} yh(y|x)dy. \tag{10}$$

In this case:

$$\theta = \mu_y = \int_{-\infty}^{\infty} E(Y|x)h_1(x)dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yh(y|x)h_1(x)dxdy. \tag{11}$$

Parameter $\mu_y$ is estimated by means of the following statistic:

$$T_{\mathbf{X},\mathbf{Y}} = \sum_{i=1}^{n}\frac{Y_i h_1(X_i)}{\pi(X_i)} \tag{12}$$

where $\{X_i, i = 1,..,n\}$ is the sample drawn according to sampling design defined by expression (2) and $y_i$ should be replaced by $x_i$. Let us assume that:

$$h(y|x) = h(y_1,...,y_n|x_1,...,x_n) = \prod_{i=1}^{n} h(y_i|x_i) \tag{13}$$

*Theorem* 2.3 If $E(Y) < \infty$ and $\pi(x) > 0$ for all $(x,y) \in U$ and assumption (13) holds, then $E_{\mathbf{f(X)}}E_{\mathbf{h(Y/X)}}(T_{\mathbf{X},\mathbf{Y}}) = \mu_y$.

Proof: When in (4) we replace $g(Y_i)$ with $g(X_i)$, given by (10), then Theorem 2.1 let us

write

$$E_{\mathbf{f(X)}}\left(\sum_{i=1}^{n}\frac{g(X_i)}{\pi(X_i)}\right) \quad = \quad E_{\mathbf{f(X)}}\left(\sum_{i=1}^{n}\frac{E_{\mathbf{h(Y/X)}}(Y_i)h_1(X_i)}{\pi(X_i)}\right) \quad = \quad E_{\mathbf{f(X)}}E_{\mathbf{h(Y/X)}}\left(T_{\mathbf{X,Y}}\right).$$

This derivation shows that Theorem 2.3 is a special case of Theorem 2.1.

*Theorem* 2.4 If the function $E(Y)$ is bounded, $\pi(y) > 0$ for each $(x,y) \in U$, and $\int_U (1/\pi(y))dy < \infty$, then

$$V(T_{\mathbf{X,Y}}) = \int_U \frac{V(Y|x)h_1^2(x)}{\pi(x)}dx + \int_U \frac{E^2(Y|x)h_1^2(x)}{\pi(x)}dx + A \tag{14}$$

where

$$A = \int_U\int_U E(Y|x)h_1(x)E(Y|x')h_1(x')\frac{\pi(x,x') - \pi(x)\pi(x')}{\pi(x)\pi(x')}dxdx'$$

or

$$A = \int_U\int_U E(Y|x)h_1(x)E(Y|x')h_1(x')\frac{\pi(x,x')}{\pi(x)\pi(x')}dxdx' - E^2(Y).$$

Proof: Adding $E_{\mathbf{h(Y/X)}}(T_{\mathbf{X,Y}})$ to $E_{f(\mathbf{X})}E_{\mathbf{h(Y/X)}}(T_{\mathbf{X,Y}} - \mu_y)^2$ we have:

$$\begin{aligned} V(T_{\mathbf{X,Y}}) &= E_{f(\mathbf{X})}E_{\mathbf{h(Y/X)}}((T_{\mathbf{X,Y}} - E_{\mathbf{h(Y/X)}}(T_{\mathbf{X,Y}})) + (E_{\mathbf{h(Y/X)}}(T_{\mathbf{X,Y}}) - E(Y)))^2 = \\ &= E_{f(\mathbf{X})}E_{\mathbf{h(Y/X)}}\left(\left(\sum_{i=1}^{n}\frac{Y_i - E_{\mathbf{h(Y/X)}}(Y_i)h_1(X_i)}{\pi(X_i)}\right) + (E_{\mathbf{h(Y/X)}}(T_{\mathbf{X,Y}}) - \mu_y)\right)^2 = \\ &= E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{V_{\mathbf{h(Y/X)}}(Y_i)h_1^2(X_i)}{\pi^2(X_i)}\right) + E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{E_{\mathbf{h(Y/X)}}(Y_i)h_1(X_i)}{\pi(X_i)} - \mu_y\right)^2, \end{aligned}$$

because $E_{\mathbf{h(Y/X)}}(Y_i - E_{\mathbf{h(Y/X)}}(Y_i)) = 0$ and $E_{\mathbf{h(Y/X)}}(Y_i - E_{\mathbf{h(Y/X)}}(Y_i))^2 = V_{\mathbf{Y/X}}(Y_i)$. Continuing the derivation we have:

$$\begin{aligned} V(T_{\mathbf{X,Y}}) = \\ &= E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{V(Y_i|X_i)h_1^2(X_i)}{\pi^2(X_i)}\right) + E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{E(Y_i|X_i)h_1(X_i)}{\pi(X_i)} - \mu_y\right)^2. \end{aligned} \tag{15}$$

By setting $\frac{V(Y_i|X_i)h_1^2(X_i)}{\pi(X_i)} = g(X_i)$ Theorem 2.1 allows us to write the following:

$$E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{V(Y_i|X_i)h_1^2(X_i)}{\pi^2(X_i)}\right) = \int_U \frac{V(Y|x)h_1^2(x)}{\pi(x)}dx. \tag{16}$$

Similarly, by setting $E(Y_i|X_i)h_1(X_i) = g(X_i)$, the second term in (15) becomes:

$$E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{g(X_i)}{\pi(X_i)}-\mu_y\right)^2 = \\ = E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{g(X_i)}{\pi(X_i)}-E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{g(X_i)}{\pi(X_i)}\right)\right)^2 = V_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{g(X_i)}{\pi(X_i)}\right). \quad (17)$$

This, expression (16) and Theorem 2.2 lead straightforward to the conclusion of Theorem 2.4.

Similarly to expression (6) let us assume that

$$f(x_1,...,x_n) = \prod_{i=1}^{n} f(x_i), \quad \pi(x) = nf(x), \quad \pi(x,x') = n(n-1)f(x)f(x'). \quad (18)$$

This, expression (17) and Theorem 2.4 lead to the following:

$$V(T_{\mathbf{X},\mathbf{Y}}) = \frac{1}{n}\left(\int_U \frac{V(Y|x)h_1^2(x)}{f(x)}dx + \int_U \frac{E^2(Y|x)h_1^2(x)}{f(x)}dx - E^2(Y)\right) = \\ = \frac{1}{n}\left(E_{f(X)}\left(\frac{V(Y|X)h_1^2(X)}{f^2(X)}\right) + V_{f(X)}\left(\frac{E(Y|X)h_1(X)}{f(X)}\right)\right) \quad (19)$$

We estimate $\mu_y$ with the following sampling design:

$$f(x_1,...,x_n) = \prod_{i=1}^{n} h_1(x_i). \quad (20)$$

Under additional assumption that $E(Y|x) = ax$ where $a = \rho\sqrt{\frac{V(Y)}{V(X)}}$ and $\rho$ is the correlation coefficient between $X$ and $Y$ then expressions (12) and (19) lead to the following:

$$T_{\mathbf{X},\mathbf{Y}} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad E(\bar{Y}) = \mu_y, \qquad V(\bar{Y}) = \frac{V(Y)}{n}(1+\rho^2) \quad (21)$$

Hence, when $\rho \neq 0$, estimator $T_{\mathbf{X},\mathbf{Y}}$ of the mean based on sampling design, given by (20) is less accurate than the simple random sample mean.

# 3. Inclusion function of sampling design proportional to values of auxiliary variable

## 3.1. Density function of the auxiliary variable is known

After Cox and Snell (1979), let us consider the following sampling design:

$$f(x_1,...,x_n)=\prod_{i=1}^{n} f(x_i), \qquad f(x_i)=\frac{x_i h_1(x_i)}{\mu_x}. \tag{22}$$

where $\mu_x = E(X) = E(X_i)$ for all $i = 1,...,n$. In this case, according to (18) the inclusion function is proportional to the value of the auxiliary variable because $\pi(x) = \frac{nxh_1(x)}{\mu_x}$. Expression (12), (19), Theorems 2.3 and Theorem 2.4 lead to the following:

$$T_{\mathbf{X},\mathbf{Y}} = \hat{Y}_R = \frac{\mu_x}{n}\sum_{i=1}^{n}\frac{Y_i}{X_i}, \qquad E(\hat{Y}_R)=\mu_y, \tag{23}$$

$$V(T_{\mathbf{X},\mathbf{Y}}) = \frac{1}{n}\left(\mu_x\int_U \frac{V(Y|x)h_1(x)}{x}dx+\mu_x\int_U \frac{E^2(Y|x)h_1(x)}{x}dx-\mu_y^2\right) = \\ = \frac{\mu_x}{n}\int_U \frac{V(Y|x)h_1(x)}{x}dx+\frac{\mu_x}{n}V\left(\frac{E(Y|x)}{x}\right). \tag{24}$$

Statistic $\hat{Y}_R$ is an unbiased ratio-type estimator of $\mu_y$.

When parameter $\mu_x$ and other parameters of the auxiliary variable density function are known, the sample can be select. The following sections address selection when these parameters are estimated.

## 3.2. Estimated parameters of the auxiliary variable density function

The values $x_1,...,x_N$ of the auxiliary variable observed in whole population are regarded as a random sample from a distribution with density $h_1(x,\theta_1,...,\theta_r)$. Let $\hat{\theta}_1...\hat{\theta}_r$ and $\hat{\mu}_x$ be consistent estimators of parameters $\theta_1,...,\theta_r$ and $\mu_x$ respectively. According to expression (22) we have the following density function of sampling design:

$$\hat{f}(x_1,...,x_n)=f(x_1,...,x_n,\hat{\theta}_1...\hat{\theta}_r)=\prod_{i=1}^{n}\hat{f}(x_i), \quad \hat{f}(x)=\frac{xh_1(\hat{\theta}_1...\hat{\theta}_r)}{\hat{\mu}_x}. \tag{25}$$

Estimation of parameters could be based on data observed, e.g. in the previous round of a regularly conducted survey.

By replacing $\mu_x$ in eq. (23) with $\hat{\mu}_x$ we obtain the following estimator:

$$T_{\mathbf{X},\mathbf{Y}} = \tilde{Y}_R = \frac{\hat{\mu}_x}{n}\sum_{i=1}^{n}\frac{Y_i}{X_i}. \tag{26}$$

where $X_1,...,X_n$ is the sample drawn according to sampling design based on density $f(x_1,...,x_n, \hat{\theta}_1...\hat{\theta}_r)$. The variance of $\tilde{Y}_R$ could be estimated by means of the well-known parametric or non-parametric method of bootstrap.

### 3.3. Kernel estimator of the auxiliary variable density function

Density function $h_1(x)$ can be estimated by means of the following well-known kernel-type estimator on the basis of all observations of auxiliary variable in the population:

$$\tilde{h}_1(x) = \frac{1}{N}\sum_{i=1}^{N} k(x,x_i,\Delta), \quad \int_{-\infty}^{\infty} k(x,x_i,\Delta)dx = 1 \tag{27}$$

where $\Delta > 0$ is the bandwidth parameter. This leads to the following estimator of $f(x)$:

$$\tilde{f}(x) = \frac{x\tilde{h}_1(x)}{\tilde{\mu}_x} = \frac{\sum_{i=1}^{N} xk(x,x_i,\Delta)}{N\tilde{\mu}_x} \tag{28}$$

where:

$$\tilde{\mu}_x = \int_{-\infty}^{\infty} x\tilde{h}_1(x)dx \tag{29}$$

is the estimator of $\mu_x$.

Let us consider the following simple kernel function based on the uniform distribution:

$$k(x,x_i,\Delta) = \begin{cases} \frac{1}{2\Delta}, & x \in [x_i-\Delta; x_i+\Delta], \\ 0, & x \notin [x_i-\Delta; x_i+\Delta]. \end{cases} \tag{30}$$

For this kernel function we have:

$$\int_{-\infty}^{\infty} xk(x,x_i,\Delta)dx = x_i \quad \textbf{for} \quad i = 1,...N, \quad \textbf{and} \quad \tilde{\mu}_x = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{31}$$

Expression (28) leads to the following:

$$\tilde{f}(x) = \frac{1}{N\bar{x}}\sum_{i=1}^{N} xk(x,x_i,\Delta) = \frac{1}{N\bar{x}}\sum_{i=1}^{N} x_i\tilde{f}_i(x,x_i,\Delta) = \sum_{i=1}^{N} w_i\tilde{f}_i(x,x_i,\Delta) \tag{32}$$

where: $w_i = \frac{x_i}{N\bar{x}}$, for $i = 1,...,N$ and

$$\tilde{f}_i(x,x_i,\Delta) = \begin{cases} \frac{x}{2x_i\Delta}, & x \in [x_i-\Delta; x_i+\Delta], \\ 0, & x \notin [x_i-\Delta; x_i+\Delta] \end{cases} \tag{33}$$

where $\tilde{f}_i(x,x_i,\Delta)$ is the trapezoid density function of the probability distribution on interval

$[x_i-\Delta; x_i+\Delta]$. After simplifications we have:

$$\tilde{f}(x) = \frac{1}{2\Delta N\bar{x}} \sum_{i=1}^{N} xI(x, x_i, \Delta) \tag{34}$$

where:

$$I(x, x_i, \Delta) = \begin{cases} 1, & x \in [x_i-\Delta; x_i+\Delta], \\ 0, & x \notin [x_i-\Delta; x_i+\Delta]. \end{cases} \tag{35}$$

Expressions (32) and (33) allows us to derive the following distribution function estimator:

$$\tilde{F}(x) = \int_{-\infty}^{x} \tilde{f}(t)dt = \sum_{i=1}^{N} w_i \tilde{F}_i(x, x_i, \Delta) \tag{36}$$

where: $w_i = \frac{x_i}{N\bar{x}}$, for $i = 1, ..., N$ and

$$\tilde{F}_i(x, x_i, \Delta) = \begin{cases} 0, & x \in (-\infty; x_i-\Delta], \\ \frac{x^2-(x_i-\Delta)^2}{4x_i\Delta}, & x \in (x_i-\Delta; x_i+\Delta], \\ 1, & x \in [x_i+\Delta; \infty). \end{cases} \tag{37}$$

The inverse function to $\tilde{F}_i(x)$ (the quantile function), $i = 1, .., N$, is as follows:

$$x = \tilde{F}_i^{-1}(u) = \sqrt{4x_i\Delta u + (x_i-\Delta)^2}, \qquad z \in [0; 1] \tag{38}$$

where $u$ has uniform distribution on interval $[0; 1]$. This allows us to easily generate the pseudovalues of the trapezoid distribution on interval $[x_i-\Delta; x_i+\Delta]$.

### 3.4. Sampling schemes

Let us assume that observations of $\mathbf{x} = [x_1, ..., x_k, ..., x_N]$ are known book values or they are gathered from a census or surveys made on a previous occasion. Function $h_1(x)$ is also known. Our purpose is to select sample $\mathbf{x}_s = [x_1, ..., x_k, ..., x_n]$ as the sub-vector of $\mathbf{x}$ according to the sampling design defined by expression (22). In order to do this, values of vector $\mathbf{x}'_s = [x'_1, ..., x'_n]$ are generated by means of the quantile functions $x' = F^{-1}(u)$, where $u$ is the value of the uniformly distributed variable on interval $[0; 1]$, $F(x) = \int_{-\infty}^{x} f(t)dt$ and $f(t)$ are given by (22). Elements of $\mathbf{x}_s$ are selected from $\mathbf{x}$ according to

$$x_k = arg \min_{j=1,...,N} |x_j - x'_k|. \tag{39}$$

This algorithm could lead to a repetition of the elements in $\mathbf{x}_s$. If the algorithm yields a sample with duplicate elements, the sample is rejected and the algorithm repeated until a sample with no duplicates is obtained.

The next algorithm, which leads to drawing $\mathbf{x}_s$ without repetition, is explained by expression:

$$x_s = arg \min_{\mathbf{x}_s \in \mathbf{X}_s} (\mathbf{x}_s - \mathbf{x}'_s)(\mathbf{x}_s - \mathbf{x}'_s)^T \tag{40}$$

where $\mathbf{X}_s$ consists of all $n$-element combinations selected without replacement from $\mathbf{x}$. The complete data $\mathbf{d} = [(x_1, y_1), ..., (x_n, y_n)]$ are evaluated after observation values $y_j$, $j = 1, ..., n$ (observations of the variable under study) are attached to the appropriate elements of vector $\mathbf{x}_s$. This algorithm becomes simpler when elements of $\mathbf{x}'_s$ and $\mathbf{x}$ are ordered from the lowest to highest.

The next variant of the sampling design is as follows. Let us note that the kernel density function $\tilde{f}(x)$, defined by expression (32), could be treated as a mixture of density functions $\tilde{f}_i(x)$, $i = 1, ..., N$ given by (33). Therefore, the $k$-th element of vector $\mathbf{x}'_s$ could be generated as follows. Firstly, the value of index $i$ is randomly (with probability $w_i$) selected from the sequence $1, ..., N$. Next, the values $x'_k$ (k=1,...,n) are generated by means of the quantile function, given by (36)-(38). Finally, the elements of vector $\mathbf{x}_s$ could be selected according to expression (39) or (40).

The complete data $\mathbf{d} = [(x_1, y_1)...(x_n, y_n)]$ are evaluated after observation values $y_j$, $j = 1, ..., n$ are attached to appropriate elements of vector $\mathbf{x}$.

## 4. Estimation in the case of McKay's bivariate gamma distribution

Suppose the random variables $U_i$ have distributions with gamma densities

$$l_i(u_i) = l_i(u_i, \theta_i, c) = \frac{c^{\theta_i}}{\Gamma(\theta_i)} u_i^{\theta_i - 1} e^{-cu_i} \tag{41}$$

where: $u_i > 0$, $c > 0$, $\theta_i > 0$, $E(U_i) = \frac{\theta_i}{c}$, $V(U_i) = \frac{\theta_i}{c^2}$, $i = 0, 1, 01$, $\theta_{01} = \theta_0 + \theta_1$ and $U_{01} = U_0 + U_1$ provided $U_0$ and $U_1$ are independent. $\theta$ and $c$ are called the shape parameter and the scale parameter respectively.

The McKay's (1934) density function of joint probability distribution of $X = U_{01}$ and $Y = U_0$ takes the following form (see also Ghirtis (1967) and Kotz et al. (2000)):

$$l(x, y) = \frac{c^{\theta_{01}}}{\Gamma(\theta_0)\Gamma(\theta_1)} y^{\theta_0 - 1} (x - y)^{\theta_1 - 1} e^{-cx}, \quad x > y > 0. \tag{42}$$

This could be useful with valuation of damage supported by declared observed data as values of $X$. In this case $\mu_y$ is mean of the true valuation of damage.

According to expression (22), the sampling design density function is defined as follows:

$$f(x) = \frac{x}{\mu_x} l_{01}(x) \tag{43}$$

where $f(x)$ is also density function of gamma distribution with shape and scale parameters equal to $\theta_{01} + 1$ and $c$ respectively.

The conditional density function is:

$$l(y|x) = \frac{\Gamma(\theta_{01})}{\Gamma(\theta_0)\Gamma(\theta_1)} x^{-\theta_0} y^{\theta_0 - 1} \left(1 - \frac{y}{x}\right)^{\theta_1 - 1}, \qquad x > y.$$

Its first two moments are:

$$\begin{cases} E(Y|x) = xE(U) = \frac{\theta_0 x}{\theta_{01}}, \\ V(Y|x) = E(Y^2|x) - E^2(Y|x) = x^2 V(U) = \frac{\theta_0 \theta_1 x^2}{(\theta_{01})^2(\theta_{01}+1)} \end{cases} \tag{44}$$

where $U$ has the beta probability distribution with parameters $\theta_0$ and $\theta_1$.

Expressions (24) and (44) lead to the following:

$$V(\hat{Y}_R) = \frac{\theta_0}{nc}\left(\left(\frac{\theta_1}{\theta_{01}(\theta_{01}+1)} + \frac{\theta_0}{\theta_{01}}\right)E(X) - \frac{\theta_0}{c}\right).$$

By substituting the expression $\frac{\theta_{01}}{c}$ for $E(X)$ we obtain:

$$V(\hat{Y}_R) = \frac{\theta_0\theta_1}{nc^2(\theta_{01}+1)} = \frac{1}{n}\mu_y(\mu_x - \mu_y)\frac{\gamma_x^2}{1+\gamma_x^2}, \quad \gamma_x = \frac{\sigma_x}{\mu_x}.$$

Finally, we have:

$$V(\hat{Y}_R) = \frac{\theta_1}{\theta_{01}+1}V(\bar{Y}) < V(\bar{Y}) = \frac{\theta_0}{nc^2}. \tag{45}$$

The variation coefficient of the estimator is as follows:

$$\gamma(\hat{Y}_R) = 100\%\frac{\sqrt{V(\hat{Y}_R)}}{\mu_y}. \tag{46}$$

The relative efficiency coefficient takes the following form:

$$deff(\hat{Y}_R) = 100\%\frac{V(\hat{Y}_R)}{V(\bar{Y})} = \frac{100\%\theta_1}{\theta_{01}+1} < 100\%. \tag{47}$$

Hence, the estimator $\hat{Y}_R$ is more precise than $\bar{Y}$.

Parameters $\theta_0$ and $c$ of the auxiliary variable can be estimated based on the observed data $\mathbf{x} = [x_1, ..., x_N]$. The method of moments yields the following estimates of the parameters:

$$\hat{\theta}_{01} = \frac{\bar{x}^2}{\hat{v}_x} = \hat{\gamma}_x^{-2}, \quad \hat{\theta}_0 = \tilde{Y}_R\frac{\bar{x}}{\hat{v}_x}, \quad \hat{\theta}_1 = \frac{(\bar{x}-\tilde{Y}_R)\bar{x}}{\hat{v}_x}, \quad \hat{c} = \frac{\bar{x}}{\hat{v}_x} \tag{48}$$

where

$$\hat{v}_x = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{x})^2, \quad \bar{x} = \frac{1}{N}\sum_{k=1}^{N}x_k, \quad \hat{\gamma}_x = \frac{\hat{v}_x}{\bar{x}^2}.$$

We estimate the density $f(x)$ by

$$\hat{f}(x) = \frac{x}{\bar{x}}\hat{l}_{01}(x, \hat{\theta}_{01}, \hat{c}) \tag{49}$$

which is the gamma density with parameters $\hat{\theta}_{01}+1 = \bar{x}\hat{c}+1$ and $\hat{c}$. The expectation $\mu_y$ can be estimated using the statistic $\tilde{Y}_R$, given by (26).

Owing to (45), the variance $V(\hat{Y}_R)$ can be estimated by means of the following statistic:

$$\tilde{V}(\tilde{Y}_R,\hat{f}(x)) = \frac{1}{n}\tilde{Y}_R(\bar{x}-\tilde{Y}_R)\frac{\hat{\gamma}_x^2}{1+\hat{\gamma}_x^2}. \tag{50}$$

The variance could be estimated by means of the following non-parametric bootstrap method. Firstly, the value of the estimator $\hat{Y}_R$ is evaluated based on the data observed in the original sample $\mathbf{D} = [(Y_j,X_j), j = 1,...,n]$. Bootstrap samples will be denoted by $\mathbf{D}^{(k)} = \left[\left(Y_j^{(k)},X_j^{(k)}\right), j = 1,...,n\right]$, $k = 1,...,B$ which are independently drawn with replacement from sample $\mathbf{D}$. This leads to the following bootstrap-type estimators of variance:

$$\hat{V}\left(\tilde{Y}_R\right) = \frac{1}{B-1}\sum_{k=1}^{B}\left(\tilde{Y}_R^{(k)}-\tilde{Y}_R\right)^2, \quad \tilde{Y}_R^{(k)} = \frac{\bar{x}}{n}\sum_{k=1}^{n}\frac{Y_i^{(k)}}{X_i^{(k)}} \tag{51}$$

or

$$\hat{V}'\left(\tilde{Y}_R\right) = \frac{1}{B-1}\sum_{k=1}^{B}\left(\tilde{Y}_R^{(k)}-\bar{\tilde{Y}}_R\right)^2, \qquad \bar{\tilde{Y}}_R = \frac{1}{B}\sum_{k=1}^{B}\tilde{Y}_R^{(k)}. \tag{52}$$

We set that $B = 1000$.

**Example**

Let us suppose that the population data are generated according to bivariate gamma distribution defined by density $l(x,y)$, given by (42). We estimate $\mu_y$ by two methods denoted by $(\tilde{Y}_R,\tilde{f}(x))$ and $(\tilde{Y}_R,\hat{f}(x))$, explained by expressions (32) and (49) respectively. They are implemented in "R" language.

First, the program draws random samples $\mathbf{D}_i = \left[(Y_j,X_j)_i, j = 1,...,3000\right]$, $i = 1,...,T$ from McKay distribution. Next, the parameters of the inclusion density function are estimated. This allows us to draw the samples $\mathbf{D}_{1i} = \left[(Y_j,X_j)_i, j = 1,...,n\right]$ from $\mathbf{D}_i$ and evaluate the values of $\tilde{Y}_R^{(i)}$ of $\mu_y$, $i = 1,...,T$. This is replicated $T = 1000$-times. Results for some alternative sample sizes and the gamma density function parameters are in columns 1-6 of Table 1. Under the assumed parameters of gamma distribution, the true values of the variation coefficient and *deff* coefficient (given by expression (46) and (47) respectively) have been calculated. They are presented in columns 7 and 8 respectively. In columns 10 and 12 there are values of the relative bias coefficients of the variance estimation, given by the following expressions:

$$b_2 = 100\frac{\tilde{V}(\tilde{Y}_R,\hat{f}(x))}{\check{V}(\tilde{Y}_R,\hat{f}(x))}, \qquad b_2' = 100\frac{\bar{\tilde{V}}(\tilde{Y}_R,\tilde{f}(x))}{\check{V}(\tilde{Y}_R,\tilde{f}(x))}, \quad \bar{\tilde{V}}(\tilde{Y}_R,\tilde{f}(x)) = \frac{1}{T}\sum_{i=1}^{T}\hat{V}_i(\tilde{Y}_R,\tilde{f}(x)) \tag{53}$$

where $\hat{V}_i(\tilde{Y}_R,\tilde{f}(x))$ explains the right side of equation (51) for the bootstrap samples: $\mathbf{D}_{1i}^{(k)} = \left[\left(Y_j^{(k)},X_j^{(k)}\right)_i, j = 1,...,n\right]$, $k = 1,...,B$ drawn from $\mathbf{D}_{1i}$, $i = 1,...T$. In columns 9, 11 and

13, there are the following estimated relative efficiency coefficients:

$$d = 100\frac{\check{V}(\tilde{Y}_R, \hat{f}(x))}{\check{V}(\bar{Y})}, \qquad d' = 100\frac{\check{V}(\tilde{Y}_R, \tilde{f}(x))}{\check{V}(\bar{Y})}, \qquad e = 100\frac{\check{V}(\tilde{Y}_R, \tilde{f}(x))}{\check{V}(\tilde{Y}_R, \hat{f}(x))} \tag{54}$$

where $\tilde{y}_R$ is given by (26) and:

$$\check{V}(\tilde{Y}_R, .) = \frac{1}{T-1}\sum_{i=1}^{T}\left(\tilde{Y}_R^{(i)} - \bar{\tilde{Y}}_R\right)^2, \qquad \bar{\tilde{Y}}_R = \frac{1}{T-1}\sum_{i=1}^{T}\tilde{Y}_R^{(i)} \tag{55}$$

are evaluated based on samples $\mathbf{D}_{1i}$, $i = 1, ...T$.

Table 1. Relative efficiency and bias of the estimation methods.

| | | | | | | | | $(\tilde{Y}_R, \hat{f}(x))$ | | $(\tilde{Y}_R, \tilde{f}(x))$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | $\theta_1$ | $\theta_0$ | $c$ | $\mu_y$ | $\mu_x$ | $\gamma(\hat{Y}_R)$ | *deff* | $d$ | $b_2$ | $d'$ | $b_2'$ | $e$ |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 30 | 1 | 10 | 1 | 10 | 11 | 1.7 | 8.3 | 8.3 | 93.7 | 9.0 | 82.8 | 97.1 |
| 60 | 1 | 10 | 1 | 10 | 11 | 1.2 | 8.3 | 9.4 | 89.0 | 10.5 | 72.5 | 102.3 |
| 150 | 1 | 10 | 1 | 10 | 11 | 0.8 | 8.3 | 11.5 | 65.9 | 13.4 | 59.2 | 109.6 |
| 60 | 1 | 10 | 0.01 | 1000 | 1100 | 1.2 | 8.3 | 8.3 | 90.1 | 10.8 | 74.7 | 110.1 |
| 60 | 3 | 10 | 0.01 | 1000 | 1300 | 1.9 | 21.4 | 24.0 | 99.9 | 22.1 | 94.0 | 85.9 |
| 60 | 10 | 3 | 0.01 | 300 | 1300 | 6.3 | 71.4 | 65.2 | 105.6 | 75.1 | 91.7 | 99.9 |

Source: Own calculations.

Statistic $\check{V}(\bar{Y})$ is evaluated by replacing $\tilde{Y}_R$ with the sample mean in equation (55). The relative efficiency coefficients in columns 9 and 10 deal with the case when the sample is selected according to the inclusion density function defined by expression (49). The coefficients from columns 11-12 are calculated based on the data from the sample drawn according to the inclusion density function defined by expressions (32) and (33), where we assumed that the bandwidth parameter $\Delta = \sqrt{\hat{v}_x}$. Moreover, in this case variance of $\tilde{Y}_R$ is estimated by means of the bootstrap method based on expression (51). In column 13, there are values of the relative efficiency coefficient of the estimation methods $(\tilde{Y}_R, \hat{f}(x))$ and $(\tilde{Y}_R, \tilde{f}(x))$ denoted by $e$. This is evaluated based on expressions (54).

The simulation analysis allows us to calculate values of the relative bias coefficient of the mean estimation defined by $b_1 = 100\bar{\tilde{y}}_R/\mu_y$. Its values for both considered estimation methods oscillate between 98% and 101%. This confirms that both methods give unbiased estimates of the expected value of the variable under study. Therefore, the values of the coefficient $b_1$ are not presented in Table 1.

Column 7 shows that in the case when $\theta_1 > \theta_0$, a value of the variation coefficient of $\hat{Y}_R$ is larger then its value for $\theta_1 < \theta_0$. Column 8 allows us to conclude that the variance of the estimator under the continuous sampling design equal to the modified density function of the auxiliary variable has a lower value than the variance of the simple random sample mean. Column 9 gives the relative efficiency coefficient value evaluated under the assumption that the parameters of the inclusion density function are estimated. Values of this coefficient

differ from appropriate values of *deff* by no more than 4.2%. This is the effect of variability of the parameter estimators. Similarly (see column 11), the kernel-type estimator of the inclusion density function leads to the higher (but not by more than 4.3%) values of $d'$ than the appropriate values of *deff*.

The proposed estimators of the variances are quite significantly biased. Usually, they underestimate the variances (see columns 10 and 12). The bias depends on the parameter values of gamma distribution, and its level is not more than 11% of the true variance.

Efficiency of the two estimators is compared in the last column of Table 1. The relative efficiency coefficient, given in expression (54), oscillates between 85.9% and 110.1%. The estimators of the expected value have comparable accuracy. Both estimation methods are unbiased. Their variances differ from each other by not more than 14.1%. However, the method based on a kernel-type estimator of the inclusion density function is preferable because it does not entail the assumption of bivariate gamma distribution.

## 5. Conclusion

This paper contributes to research on estimating of the mean value of the variable under study using continuous sampling designs. The well-known properties of the conditional distribution of the variable under study under an assumed value of the auxiliary variable and results from Cordy (1993) allow us to construct the estimator of the mean of the variable under study. It has been shown that this estimator is unbiased. The theorems presented in this paper also deal with estimating parameters other than the mean. These results allow us to consider a particular (inspired by Cox and Snell (1979)) sampling design with inclusion function dependent on the auxiliary variable. This provides a ratio-type estimator of the mean value. Estimation of the inclusion density function by means of a kernel-type estimator is also proposed. It does not need additional assumptions about density functions. From the results of a simulation study, we conclude that the expected value can be estimated more efficiently than by the sample mean.

Perhaps, additional studies could show, if the considered estimation method can be useful in statistical applications like auditing, insurance problems, and analysis of joint distributions of income and expenditures. There are many possibilities for modifying the sampling designs represented by continuous inclusion functions and their estimators. For instance, other kernels can be applied. We could apply classical statistical inference procedures for large sample sizes. All the considered estimators could be shown as sums of independent identically distributed random variables. Therefore, the well-known asymptotic methods of statistical inference could be used to constructions of confidence intervals and statistical tests. Moreover, there are possibilities for applying well-known bootstrap techniques to test statistical hypotheses or confidence interval estimation.

## Acknowledgement

This paper is a result of a grant supported by the *National Science Centre, Poland*, no. 2016/21/B/HS4/00666.

## References


BĄK, T., (2014). Triangular method of spatial sampling. *Statistics in Transition*, Vol. 15, No. 1, pp. 9–22. http://stat.gov.pl/en/sit-en/issues-and-articles-sit.

BĄK, T., (2018). An extension of Horvitz-Thompson estimator used in adaptive cluster sampling to continuous universe. *Communications in Statistics – Theory and Methods*, vol. 46, Issue 19, pp. 9777–9786, DOI: 10.1080/03610926.2016.1218028.

BENHENNI, K., CAMBANIS, S., (1992). Sampling Designs for Estimating Integrals of Stochastic Processes. *The Annals of Statistics*, Vol. 20, No. 1, pp. 161–194.

BUCKLEW, J. A., (2004). *Introduction to Rare Event Simulation*. Springer, New York, Berlin, Heidelberg, Hong Kong, London, Milan, Paris, Tokyo.

CORDY, C. B., (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics and Probability Letters*, Vol. 18, pp. 353–362.

COX, D. R., SNELL, E. J., (1979). On sampling and the estimation of rare errors. *Biometrika*, Vol. 66, 1, pp. 125–32.

CHERIYAN, K. C., (1941). A bivariate correlated gamma-type distribution function. Journal of the Indian Mathematical Society, Vol. 5, pp. 133–144.

CRESSIE, N. A. C., (1993). *Statistics for Spatial Data*. Wiley, New York.

FROST, P. A., TAMURA, H., (1986). Accuracy of auxiliary information interval estimation in statistical auditing. *Journal of Accounting Research* 24, pp. 57–75.

GHIRTIS G. C., (1967). Some problems of statistical inference relating to double-gamma distribution. *Trabajos de Estatistica*, Vol. 18, pp. 67–87.

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

KOTZ, S., BALAKRISHNAN, JOHNSON, N. L., (2000). *Continuous Multivariate Distributions, Vol. 1: Models and Applications*. John Wiley & Sons, Inc., New York, Chichester, Wenheim, Brisbane, Sigapore, Toronto.

MCKAY, A. T., (1934). Sampling from batches. *Journal of the Royal Statistical Society* 2, pp. 207–216.

RIPLEY, B. D., (1987). *Stochastic Simulation*. Wiley, 1987, New York. Sarndal Särndal C. E., Swenson B., Wretman J. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York-Berlin-Heidelberg-London-Paris-Tokyo-Hong Kong-Barcelona-Budapest.

TILLÉ, Y., (2006). *Sampling Algorithms*. Springer.

THOMPSON, M. E., (1997). *Theory of Sample Survey.* Chapman & Hall, London, Weinheim, New York, Tokyo, Melbourne, Madras.

WILHELM M., TILLÉ, Y., QUALITÉ, M., (2017). Quasi-systematic sampling from a continuous population. *Computational Statistics & Data Analysis*, 105, pp. 11–23.

WILKS, S. S., (1962). *Mathematical Statistics*. John Wiley & Sons, New York, London.

WYWIAŁ, J. L., (2016). *Contributions to Testing Statistical Hypotheses in Auditing*. PWN, Warsaw.

WYWIAŁ, J. L., (2018). Application of two gamma distribution mixture to financial auditing. Sankhya B, Vol. 80, issue 1, pp. 1–18.

ZUBRZYCKI, S., (1958). Remarks on random, stratified and systematic sampling in a plane. *Colloquium Mathematicum*, Vol. 6, pp. 251–262. DOI: 10.4064/cm-6-1-251-264, http://matwbn.icm.edu.pl/ksiazki/cm/cm6/cm6135.pdf.

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 17–40, DOI 10.21307/stattrans-2020-053*
Received – 30.05.2020; accepted – 03.09.2020

# The Gamma Kumaraswamy-G family of distributions: theory, inference and applications

**Rana Muhammad Imran Arshad**[1],
**Muhammad Hussain Tahir**[2] , **Christophe Chesneau**[3],
**Farrukh Jamal**[4]

## ABSTRACT

In this paper, we introduce a new family of univariate continuous distributions called the Gamma Kumaraswamy-generated family of distributions. Most of its properties are studied in detail, including skewness, kurtosis, analytical comportments of the main functions, moments, stochastic ordering and order statistics. The next part of the paper focuses on a particular member of the family with four parameters, called the gamma Kumaraswamy exponential distribution. Among its advantages, the following should be mentioned: the corresponding probability density function can have symmetrical, left-skewed, right-skewed and reversed-J shapes, while the corresponding hazard rate function can have (nearly) constant, increasing, decreasing, upside-down bathtub, and bathtub shapes. Subsequently, the inference on the gamma Kumaraswamy exponential model is performed. The method of maximum likelihood is applied to estimate the model parameters. In order to demonstrate the importance of the new model, analyses on two practical data sets were carried out. The results proved more favourable for the studied model than for any of the other eight competitive models.

**Key words:** Kumaraswamy distribution, gamma distribution, generalised family, moments, stochastic ordering, maximum likelihood method, data analysis.

## 1. Introduction

In order to meet scientific requirements, modern experiments require high precision in data analysis. Unfortunately, in most situations this requirement cannot be achieved through the use of standard statistical models. For this reason, the creation of new flexible models, well adapted to the context, remains a passionate challenge for the statisticians. From a probabilistic point of view, attractive models can be derived from families of distributions enjoying desirable properties. Such families can be defined by the use of effective techniques introducing tuning parameters to well-established distributions. These families are often characterized by sophisticated but flexible functions, which can be handled thanks to

[1]Department of Statistics, The Islamia University of Bahawalpur, Punjab 63100, Pakistan.
E-mail: imranarshad.stat@gmail.com. ORCID: https://orcid.org/0000-0002-5687-4634.
[2]Department of Statistics, The Islamia University of Bahawalpur, Punjab 63100, Pakistan.
E-mail: mtahir.stat@gmail.com. ORCID: https://orcid.org/0000-0002-2157-3997.
[3]Department of Mathematics, Université de Caen, LMNO, Campus II, Science 3, 14032 Caen, France.
E-mail: christophe.chesneau@gmail.com. ORCID: https://orcid.org/0000-0002-1522-9292.
[4]Department of Statistics, The Islamia University of Bahawalpur, Punjab 63100, Pakistan.
E-mail: drfarrukh1982@gmail.com. ORCID: https://orcid.org/0000-0001-6192-9890.

the computational and analytical facilities available in modern programming software (as R, Maple, Mathematica...). In particular, the use of this software can easily tackle the problems involved in computing eventual special functions. Among the high impacted families of distributions, there are the beta-G family by Eugene *et al.* (2002) and Jones (2004), the Kumaraswamy-G (Kw-G) family by Cordeiro and de Castro (2011) and Ramos (2014), the Kumaraswamy Poisson-G (Kw-G) family by Ramos (2014), the McDonald-G (Mc-G) family by Alexander *et al.* (2012), the gamma-G type 1 family by Zografos and Balakrishnan (2009) and Amini *et al.* (2014), the gamma-G type 2 family by Ristic and Balakrishnan (2012) and Amini *et al.* (2014), the odd-gamma-G type 3 family by Torabi and Montazari (2012), the logistic-G family by Torabi and Montazari (2014), the odd exponentiated generated (odd exp-G) family by Cordeiro *et al.* (2013), the transformed-transformer (T-X) (Weibull-X and gamma-X) family by Alzaatreh *et al.* (2013a), the exponentiated T-X family by Alzaatreh *et al.* (2013b), the odd Weibull-G family by Bourguignon *et al.* (2014), the exponentiated half-logistic by Cordeiro *et al.* (2014), the T-X$\{$Y$\}$-quantile based approach family by Aljarrah *et al.* (2014), the T-R$\{$Y$\}$ family by Alzaatreh *et al.* (2014), the odd Burr-III-G family by Jamal *et al.* (2017), the Kumaraswamy odd Burr-G family by Nasir *et al.* (2018), the generalized odd gamma-G family by Hosseini *et al.* (2018), the truncated Cauchy power-G family by Aldahlan *et al.* (2019) and the type II general inverse exponential-G family by Jamal *et al.* (2020).

In this study, we introduce a new family of distributions derived to two important families: the Kumaraswamy-G and odd gamma-G families introduced by Cordeiro and de Castro (2011) and Torabi and Montazari (2012), respectively. Before going further in the motivation, let us briefly describe these two well-recognized families, beginning with the Kumaraswamy-G family of distributions. Let $a > 0$, $b > 0$, $G(x)$ be the cumulative distribution function (cdf) of an univariate continuous distribution and $g(x)$ be the corresponding probability distribution function (pdf). Then, the Kumaraswamy-G family of distributions is characterized by the cdf given by

$$H(x) = 1 - \{1 - G(x)^a\}^b, \qquad x \in \mathbb{R} \tag{1}$$

and the corresponding pdf can be expressed as

$$h(x) = abg(x)G(x)^{a-1}\{1 - G(x)^a\}^{b-1}, \qquad x \in \mathbb{R}. \tag{2}$$

Thus, the feature of the Kumaraswamy-G family is to add two shape parameters to the former distribution characterized by the cdf $G(x)$, increasing mechanically its flexible properties. This allows the construction of more flexible models to analyse a wide variety of data sets, as developed in Cordeiro and de Castro (2011) for the normal, Weibull, gamma, Gumbel and inverse Gaussian distributions. The Kumaraswamy-G family of distributions is also known to be a simple alternative to the beta-G family of distribution established by Eugene *et al.* (2002). The essentials of the standard Kumaraswamy distribution are detailed in Jones (2008). Current developments and extensions of the Kumaraswamy-G family of distributions can be found in, e.g. Paranaiba *et al.* (2012), de Pascoa *et al.* (2011), Ramos (2014), Gomes *et al.* (2014), Rodrigues and Silva (2015) and Jamal *et al.* (2019).

On the other side, Torabi and Montazari (2012) introduced the odd gamma-G family of distributions, briefly described below. Let $\alpha > 0$, $H(x)$ be the cdf of an univariate continuous distribution, $\bar{H}(x) = 1 - H(x)$ and $h(x)$ be the corresponding pdf. Let $\gamma_1(\alpha, z)$ be the regularized lower incomplete gamma function defined by $\gamma_1(\alpha, z) = \gamma(\alpha, z)/\Gamma(\alpha)$, where $\gamma(\alpha, z) = \int_0^z t^{\alpha-1} e^{-t} dt$ and $\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt$. Then, the odd gamma-G family of distributions "with $G = H$" is characterized by the cdf given as

$$F(x) = \gamma_1 \left( \alpha, \frac{H(x)}{\bar{H}(x)} \right), \qquad x \in \mathbb{R} \tag{3}$$

and the corresponding pdf is specified by

$$f(x) = \frac{1}{\Gamma(\alpha)} \frac{h(x) H(x)^{\alpha-1}}{\bar{H}(x)^{\alpha+1}} \exp\left( -\frac{H(x)}{\bar{H}(x)} \right), \qquad x \in \mathbb{R}. \tag{4}$$

The odd-gamma-G family of distributions gives an alternative to the useful gamma-G type 1 family of distributions introduced by Zografos and Balakrishnan (2009) in the following stochastic ordering sense: $F(x) \geq K(x)$, where $K(x) = \gamma_1 (\alpha, -\log[\bar{H}(x)])$ is the cdf corresponding to the gamma-G type 1 family of distributions. Also, the merits of the odd-gamma-G family have been highlighted in recent studies, including those of Torabi and Montazari (2012), Hosseini *et al.* (2018), Oluyede *et al.* (2018) and Nasir *et al.* (2020), via the exploration of various theoretical and practical aspects. In particular, it is shown that the parental distribution characterized by the cdf $H(x)$ can take the benefits of the considered polynomial-exponential transformation with $\alpha$ as the tuning parameter, allowing the construction of new flexible statistical models. In particular, for appropriated $H(x)$, the analyses of a wide broad range of real life data sets are favourable to the odd-gamma-G models in comparison to well-recognized competitors.

In the light of the previous arguments, a promising direction of work becomes the combination of the Kumaraswamy-G and odd gamma-G families via the composition technique of the respective cdfs. Thus, we aim to create a new generalized family of distributions benefiting of the respective qualities of these two families, aiming

- to skew any symmetrical distribution;
- to modulate the weight of the tails of any parental distribution;
- to increase the possible shapes of the (probabilistic or reliability) functions of the parental distribution;
- to construct new statistical models with better (fits) properties than other competitive models, or enlarging the horizon of fields of applications.

The proposed family is called the gamma Kumaraswamy-G (GKw-G) family of distributions. This study explores, in both theoretical and practical terms, the properties of the GKw-G family. A special member defined with the exponential distribution as the parent, called the GKw-E distribution, will serve as a statistical model. The complete analyses of

two practical data sets are proposed, showing that the GKw-E model presents better fit to eight notorious models in the field.

The rest of the article is organized as follows. In Section 2, we present the main functions and properties of the GKw-G family of distributions. In Section 3, the GKw-E distribution is introduced, as well as some of its structural properties. In Section 4, the GKw-E model parameters are estimated by the maximum likelihood method and a simulation study is performed to verify the convergence properties. Also, the usefulness of the GKw-E model is illustrated by means of two practical data sets. Finally, Section 5 offers some concluding remarks.

## 2. The gamma Kumaraswamy-G family of distributions

### 2.1. Presentation

We characterize the GKw-G family of distributions by the cdf of the odd gamma-H family of distributions given by (3), defined with the cdf $H(x)$ of the Kumaraswamy-G family of distributions given as (1). Hence, by noticing that $H(x)/\bar{H}(x) = \{1-G(x)^a\}^{-b} - 1$, the corresponding cdf is defined by

$$F(x) = \gamma_1\left(\alpha, \{1-G(x)^a\}^{-b} - 1\right), \qquad x \in \mathbb{R}. \tag{5}$$

One can remark that, if $b = 1$, this cdf becomes the one of the generalized odd gamma-G family introduced by Hosseini *et al.* (2018), that is $F(x) = \gamma_1(\alpha, G(x)^a/[1-G(x)^a]), x \in \mathbb{R}$. In this sense, the GKw-G family of distributions can be viewed as a generalization of this family. The parameter $b$ plays an important role, as we shall see later. The corresponding survival (sf) function is

$$S(x) = 1 - \gamma_1\left(\alpha, \{1-G(x)^a\}^{-b} - 1\right), \qquad x \in \mathbb{R}.$$

The pdf of the GKw-G family can be obtained by putting (1) and (2) into (4). More directly, upon almost everywhere differentiation of $F(x)$, it is obtained as

$$\begin{aligned} f(x) = \frac{ab}{\Gamma(\alpha)} g(x) G(x)^{a-1} \{1-G(x)^a\}^{-b-1} \left\{\{1-G(x)^a\}^{-b} - 1\right\}^{\alpha-1} \\ \times \exp\left[1 - \{1-G(x)^a\}^{-b}\right]. \qquad x \in \mathbb{R}. \end{aligned} \tag{6}$$

The corresponding hazard rate function (hrf) is obtained as $\pi(x) = f(x)/S(x)$, that is

$$\pi(x) = \frac{ab}{\Gamma(\alpha)} \frac{g(x) G(x)^{a-1} \{1-G(x)^a\}^{-b-1} \left\{\{1-G(x)^a\}^{-b} - 1\right\}^{\alpha-1} \exp\left[1 - \{1-G(x)^a\}^{-b}\right]}{1 - \gamma_1\left(\alpha, \{1-G(x)^a\}^{-b} - 1\right)}.$$

Some special members of the GKw-G family characterized by their cdfs are presented in Table 1.

Table 1: Some members of the GKw-G family of distributions characterized by their cdfs.

| cdf $G(x)$ | Support | GKw-G cdf $F(x)$ | Parameters |
|---|---|---|---|
| Uniform | $(0,\theta)$ | $\gamma_1\left(\alpha,\{1-(x/\theta)^a\}^{-b}-1\right)$ | $(\alpha,a,b,\theta)$ |
| Exponential | $(0,+\infty)$ | $\gamma_1\left(\alpha,\left\{1-[1-e^{-\lambda x}]^a\right\}^{-b}-1\right)$ | $(\alpha,a,b,\lambda)$ |
| Weibull | $(0,+\infty)$ | $\gamma_1\left(\alpha,\left\{1-[1-e^{-(\lambda x)^\beta}]^a\right\}^{-b}-1\right)$ | $(\alpha,a,b,\lambda)$ |
| Inverse Weibull | $(0,+\infty)$ | $\gamma_1\left(\alpha,\left\{1-e^{-a(\lambda/x)^\beta}\right\}^{-b}-1\right)$ | $(\alpha,a,b,\lambda,\beta)$ |
| Burr XII | $(0,+\infty)$ | $\gamma_1\left(\alpha,\{1-\{1-[1+(x/s)^c]^{-k}\}^a\}^{-b}-1\right)$ | $(\alpha,a,b,c,k,s)$ |
| Logistic | $\mathbb{R}$ | $\gamma_1\left(\alpha,\left\{1-[1+e^{-(x-\mu)/s}]^{-a}\right\}^{-b}-1\right)$ | $(\alpha,a,b,\mu,s)$ |
| Gumbel | $\mathbb{R}$ | $\gamma_1\left(\alpha,\left\{1-\exp(-ae^{-(x-\mu)/\sigma})\right\}^{-b}-1\right)$ | $(\alpha,a,b,\mu,\sigma)$ |
| Normal | $\mathbb{R}$ | $\gamma_1\left(\alpha,\{1-\Phi((x-\mu)/\sigma)^a\}^{-b}-1\right)$ | $(\alpha,a,b,\mu,\sigma)$ |
| Cauchy | $\mathbb{R}$ | $\gamma_1\left(\alpha,\{1-[(1/\pi)\arctan((x-x_0)/\theta)+1/2]^a\}^{-b}-1\right)$ | $(\alpha,a,b,x_0,\theta)$ |

Thanks to its simplicity in the definition, the special member of the GKw-G family based on the exponential distribution will be the object of all the attention in our applications.

Let $Q_G(x)$ be the quantile function corresponding to $G(x)$, that is, the function satisfying the following equation: $G(Q_G(p)) = Q_G(G(p)) = p$ for any $p \in (0,1)$. Then, the quantile function of the GKw-G family of distributions can be expressed as

$$Q(p) = Q_G\left(\left[1-\left\{1+\gamma_1^{-1}(\alpha,p)\right\}^{-1/b}\right]^{1/a}\right), \qquad p \in (0,1), \tag{7}$$

where $\gamma_1^{-1}(\alpha,p)$ denotes the inverse function of $\gamma_1(\alpha,p)$, i.e., satisfying $\gamma_1\left(\alpha,\gamma_1^{-1}(\alpha,p)\right) = \gamma_1^{-1}(\alpha,\gamma_1(\alpha,p)) = p$ for any $p \in (0,1)$. Further details on $\gamma_1^{-1}(\alpha,p)$ can be found in (Abramowitz and Stegun, 1965, Section 6.5). In particular, the median of the GKw-G family is specified by $M = Q(1/2)$. Also, the three quartiles are defined by $Q_1 = Q(1/4)$, $Q_2 = M$ and $Q_3 = Q(3/4)$, and the seven octiles by $O_1 = Q(1/8)$, $O_2 = Q(2/8) = Q_1$, $O_3 = Q(3/8)$, $O_4 = Q(4/8)$, $O_5 = Q(5/8)$, $O_6 = Q(6/8) = Q_3$ and $O_7 = Q(7/8)$.

The quantile function and its related values are useful to evaluate some properties of the GKw-G family, such as the skewness and kurtosis, as described below.

### 2.2. Skewness and kurtosis

A measure of the skewness of the GKw-G family is given by

$$S = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}. \tag{8}$$

In full generality, for given $G(x)$, $\alpha$, $a$ and $b$, when the corresponding GKw-G distribution is symmetric, we have $S = 0$, when it is right skewed, we have $S > 0$ and when it is left skewed, we have $S < 0$. See Kenney and Keeping (1962).

Also, a measure of the kurtosis of the GKw-G family of distributions is proposed by

$$K = \frac{O_3 - O_1 + O_7 - O_5}{O_6 - O_2}. \tag{9}$$

For given $G(x)$, $\alpha$, $a$ and $b$, as $K$ increases, the tail of the corresponding GKw-G distribution becomes heavier. We refer to Moors (1998).

The advantages of these measures are to be robust in presence of outliers and they always exist (even if the distribution does not admit moments).

### 2.3. Properties

Diverse and important properties of the new family are now described.

#### 2.3.1 Asymptotic properties

The two following propositions investigate the asymptotic properties of the cdf, sf, pdf and hrf of the GKw-G family of distributions.

**Proposition 2.1** *The asymptotic equivalences of the cdf, pdf and hrf of the GKw-G family when* $G(x) \to 0$ *are, respectively,*

$$F(x) \sim \frac{b^\alpha}{\alpha\Gamma(\alpha)} G(x)^{a\alpha}, \qquad f(x) \sim \frac{ab^\alpha}{\Gamma(\alpha)} g(x) G(x)^{a\alpha-1}, \qquad h(x) \sim \frac{ab^\alpha}{\Gamma(\alpha)} g(x) G(x)^{a\alpha-1}.$$

**Proof 2.1** *The proof follows from the following equivalences: when* $y \to 0$*, we have* $(1 - y^a)^{-b} \sim 1 + by^a$ *and* $\gamma_1(\alpha, y) \sim y^\alpha/(\alpha\Gamma(\alpha))$.

**Proposition 2.2** *The asymptotic equivalences of the sf, pdf and hrf of the GKw-G family when* $G(x) \to 1$ *are, respectively,*

$$S(x) \sim \frac{a^{-b(\alpha-1)}}{\Gamma(\alpha)} \{1 - G(x)\}^{-b(\alpha-1)} e^{1 - a^{-b}\{1-G(x)\}^{-b}},$$

$$f(x) \sim \frac{ba^{-\alpha b}}{\Gamma(\alpha)} g(x) \{1 - G(x)\}^{-\alpha b - 1} e^{1 - a^{-b}\{1-G(x)\}^{-b}}$$

*and*

$$h(x) \sim ba^{-b} g(x) \{1 - G(x)\}^{-b-1}.$$

**Proof 2.2** *The proof follows from the following equivalences: when* $y \to +\infty$*, we have* $\gamma_1(\alpha, y) \sim 1 - y^{\alpha-1}e^{-y}/\Gamma(\alpha)$ *and, when* $y \to 1$*, we have* $y^a \sim 1 - a(1 - y)$.

Propositions 2.1 and 2.2 are useful to understand the roles of $G(x)$, $g(x)$, $\alpha$, $a$ and $b$ on the asymptotic properties of the cdf, sf, pdf and hrf of the GKw-G family. In particular, we see that $b$ has a strong impact, mainly when $G(x) \to 1$.

### 2.3.2 Critical points

The analytical study of the pdf and hrf of the GKw-G family is crucial to understand their complexity. The critical points are essential in this regard. As usual, they can be determined by solving the following nonlinear equations $\partial \log[f(x)]/\partial x = 0$ and $\partial \log[h(x)]/\partial x = 0$, respectively, both obtained as

$$\frac{\partial g(x)/\partial x}{g(x)} + (a-1)\frac{g(x)}{G(x)} + a(b+1)\frac{g(x)G(x)^{a-1}}{1-G(x)^a} + ab(\alpha-1)\frac{g(x)G(x)^{a-1}\{1-G(x)^a\}^{-b-1}}{\{1-G(x)^a\}^{-b}-1} - abg(x)G(x)^{a-1}\{1-G(x)^a\}^{-b-1} = 0 \quad (10)$$

and

$$\frac{\partial g(x)/\partial x}{g(x)} + (a-1)\frac{g(x)}{G(x)} + a(b+1)\frac{g(x)G(x)^{a-1}}{1-G(x)^a} + ab(\alpha-1)\frac{g(x)G(x)^{a-1}\{1-G(x)^a\}^{-b-1}}{\{1-G(x)^a\}^{-b}-1} - abg(x)G(x)^{a-1}\{1-G(x)^a\}^{-b-1} + \frac{ab}{\Gamma(\alpha)}\frac{g(x)G(x)^{a-1}\{1-G(x)^a\}^{-b-1}\left\{\{1-G(x)^a\}^{-b}-1\right\}^{\alpha-1}\exp\left[1-\{1-G(x)^a\}^{-b}\right]}{1-\gamma_1\left(\alpha,\{1-G(x)^a\}^{-b}-1\right)} = 0. \quad (11)$$

The nature of the obtained critical points can be determined by investigating the signs of $\partial^2 \log[f(x)]/\partial x^2$ and $\partial^2 \log[h(x)]/\partial x^2$ taken at these points, respectively.

### 2.3.3 Some results in distribution

As usual, for any random variable $U$ following the uniform distribution over $(0,1)$, the random variable $X$ defined by $X = Q(U)$ has the cdf $F(x)$. For given $G(x)$, $\alpha$, $a$ and $b$, this characterization is useful to generate random values distributed according to the related GKw-G distribution through the inverse transform sampling.

Now, we say that a random variable follows the gamma distribution $\mathscr{G}_{am}(1,\alpha)$ if it has the cdf given by $K(x) = \gamma_1(\alpha,x)$, $x > 0$. If $X$ is a random variable having the cdf of the GKw-G family, then the random variable $Y$ defined by $Y = \{1-G(X)^a\}^{-b} - 1$ follows the gamma distribution $\mathscr{G}_{am}(1,\alpha)$.

Also, if $Y$ is a random variable following the gamma distribution $\mathscr{G}_{am}(1,\alpha)$, then the random variable $X$ defined by $X = Q_G\left(\left[1-\{1+Y\}^{-1/b}\right]^{1/a}\right)$ has the cdf of the GKw-G family.

### 2.3.4 Linear representations

This subsection is devoted to exploitable linear representations for the cdf and pdf of the GKw-G family.

**Proposition 2.3** *We have the following linear representations for the cdf and pdf of the GKw-G family of distributions:*

$$F(x)=\sum_{i=0}^{+\infty} w_i G(x)^{ai}, \quad f(x)=\sum_{i=1}^{+\infty} w_i \left[aig(x)G(x)^{ai-1}\right], \tag{12}$$

*where*

$$w_i=\sum_{j,k=0}^{+\infty}\frac{(-1)^{i+j+k}}{\Gamma(\alpha)k!(\alpha+k)}\binom{\alpha+k}{j}\binom{b(j-\alpha-k)}{i}$$

*and* $\binom{b}{a}$ *denotes the generalized binomial coefficient, i.e.* $\binom{b}{a}=b(b-1)\ldots(b-a+1)/a!$.

**Proof 2.3** *By using the regularized lower incomplete gamma function series expansion, i.e.*

$$\gamma_1(\alpha,y)=\sum_{k=0}^{+\infty}(-1)^k\frac{y^{\alpha+k}}{\Gamma(\alpha)k!(\alpha+k)}, \qquad y\geq 0,$$

*and after some simplifications, we can express* $F(x)$ *as*

$$\begin{aligned}F(x)&=\gamma_1\left(\alpha,\frac{1-\{1-G(x)^a\}^b}{\{1-G(x)^a\}^b}\right)\\&=\sum_{k=0}^{+\infty}\frac{(-1)^k}{\Gamma(\alpha)k!(\alpha+k)}\{1-G(x)^a\}^{-b(\alpha+k)}\underbrace{\left[1-\{1-G(x)^a\}^b\right]^{\alpha+k}}_{A}.\end{aligned}$$

*By virtue of the generalized binomial series expansion, the term A can expressed as*

$$A=\sum_{j=0}^{+\infty}(-1)^j\binom{\alpha+k}{j}\{1-G(x)^a\}^{bj}.$$

*By putting the previous equalities together, we get*

$$F(x)=\sum_{j,k=0}^{+\infty}\frac{(-1)^{j+k}}{\Gamma(\alpha)k!(\alpha+k)}\binom{\alpha+k}{j}\underbrace{\{1-G(x)^a\}^{b(j-\alpha-k)}}_{B}.$$

*By using again the generalized binomial series expansion, we get*

$$B=\sum_{i=0}^{+\infty}(-1)^i\binom{b(j-\alpha-k)}{i}G(x)^{ai}.$$

*The desired linear representation of* $F(x)$ *follows from the combination of all the equalities above. Upon differentiation, we derive the linear representation of* $f(x)$*. This completes the proof of Proposition 2.3.*

Since it depends on the well-known exp-G family of distributions (with parameter *ai* for any

integer $i$), the linear representations presented in Proposition 2.3 are useful to derive related analytical and numerical properties. Some of them are explored in the subsections below.

### 2.3.5 Moments and derivations

Here, we assume that all the presented integrals and sum exist (which is not necessarily the case, depending on the definition of $G(x)$, among others). Let $r$ be an integer. Then, the $r$-th ordinary moment of the GKw-G family is given as

$$\mu'_r = \int_{-\infty}^{+\infty} x^r f(x)dx = \int_{-\infty}^{+\infty} x^r \frac{ab}{\Gamma(\alpha)} g(x)G(x)^{a-1}\left\{1-G(x)^a\right\}^{-b-1}\left\{\left\{1-G(x)^a\right\}^{-b}-1\right\}^{\alpha-1} \times \exp\left[1-\left\{1-G(x)^a\right\}^{-b}\right]dx.$$

By using the quantile function in (7), with the change of variable $x = Q(p)$, we can express $\mu'_r$ as

$$\mu'_r = \int_0^1 Q(p)^r dp = \int_0^1 \left[Q_G\left(\left[1-\left\{1+\gamma_1^{-1}(\alpha,p)\right\}^{-1/b}\right]^{1/a}\right)\right]^r dp.$$

For given $G(x)$, $r$, $\alpha$, $a$ and $b$, this integral can be computed numerically via any mathematical software (R, Maple, Matlab, Mathematica…). Also, a linear representation of $\mu'_r$ can be deduced from Proposition 2.3. Indeed, owing to (12), we have

$$\mu'_r = \sum_{i=1}^{+\infty} w_i \int_{-\infty}^{+\infty} x^r \left[aig(x)G(x)^{ai-1}\right]dx = \sum_{i=1}^{+\infty} w_i ai \int_0^1 p^{ai-1} Q_G(p)^r dp.$$

Among others, one can deduce the mean defined by $\mu = \mu'_1$, the variance given by $\sigma^2 = \mu'_2 - (\mu'_1)^2$, the $r$-th central moment given as

$$\mu_r = \int_{-\infty}^{+\infty} (x-\mu'_1)^r f(x)dx = \sum_{k=0}^{r} \binom{r}{k} (-1)^k (\mu'_1)^k \mu'_{r-k}, \tag{13}$$

the coefficient of skewness given as $CS = \mu_3/\mu_2^{3/2}$, the coefficient of kurtosis obtained as $CK = \mu_4/\mu_2^2$ and the moment generating function given by

$$M(t) = \int_{-\infty}^{+\infty} e^{tx} f(x)dx = \sum_{r=0}^{+\infty} \frac{t^r}{r!}\mu'_r.$$

Alternatively, we can use (12) to have a linear representation for $M(t)$ without using moments. Indeed, we have

$$M(t) = \sum_{i=1}^{+\infty} w_i \int_{-\infty}^{+\infty} e^{tx}\left[aig(x)G(x)^{ai-1}\right]dx = \sum_{i=1}^{+\infty} w_i ai \int_0^1 p^{ai-1} e^{tQ_G(p)} dp.$$

Finally, let us mention that the incomplete moments can be expressed in a similar way, giving expressions for the Bonferroni and Lorenz curves, mean residual-life, mean waiting-time, mean deviation about the mean and mean deviation about the median. For similar developments, we refer to the methodology of Hosseini *et al.* (2018).

### 2.3.6 Stochastic ordering

We now prove a result on the stochastic ordering involving the GKw-G family of distributions with $a$ and $b$ as common parameters. Further details on stochastic ordering can be found in Shaked and Shanthikumar (1994).

**Proposition 2.4** *Let $X$ be a random variable having the pdf $f_1(x)$ given by* (6) *with parameters $\alpha_1$, $a$ and $b$ and $Y$ be a random variable having the pdf $f_2(x)$ given by* (6) *with parameters $\alpha_2$, $a$ and $b$. Then, if $\alpha_1 \leq \alpha_2$, we have $X \leq_{lr} Y$, i.e. $f_1(x)/f_2(x)$ is decreasing.*

**Proof 2.4** *We have*

$$\frac{f_1(x)}{f_2(x)} = \frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)} \left\{ \{1-G(x)^a\}^{-b} - 1 \right\}^{\alpha_1-\alpha_2}.$$

*By differentiating with respect to x, since $\alpha_1 \leq \alpha_2$, we have*

$$\frac{\partial}{\partial x}\frac{f_1(x)}{f_2(x)} =$$
$$\frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)}(\alpha_1-\alpha_2) \left\{ \{1-G(x)^a\}^{-b} - 1 \right\}^{\alpha_1-\alpha_2-1} abg(x)G(x)^{a-1} \{1-G(x)^a\}^{-b-1} \leq 0.$$

*Hence, we have $X \leq_{lr} Y$. This ends the proof of Proposition 2.4.*

## 2.4. Order statistics

The order statistics naturally arise in many applications involving data relating to survival testing studies. All the details can be found in the book of David and Nagaraja (2003). This subsection is devoted to the order statistics of the GKw-G family. Let $X_1, \ldots, X_n$ be the random sample from the GKw-G family and $X_{i:n}$ be the $i$-th order statistic. Then, the pdf of $X_{i:n}$ is given by

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!} f(x) F(x)^{i-1} [1-F(x)]^{n-i}, \qquad x \in \mathbb{R}. \tag{14}$$

Hence, by using (5) and (6), we have

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!} \frac{ab}{\Gamma(\alpha)} g(x) G(x)^{a-1} \{1-G(x)^a\}^{-b-1} \left\{ \{1-G(x)^a\}^{-b} - 1 \right\}^{\alpha-1}$$
$$\exp\left[1-\{1-G(x)^a\}^{-b}\right] \gamma_1\left(\alpha, \{1-G(x)^a\}^{-b} - 1\right)^{i-1} \left[1-\gamma_1\left(\alpha, \{1-G(x)^a\}^{-b} - 1\right)\right]^{n-i}.$$

In particular, the pdfs of $X_{1:n} = \inf(X_1, \ldots, X_n)$ and $X_{n:n} = \sup(X_1, \ldots, X_n)$ are given by $f_{1:n}(x)$ and $f_{n:n}(x)$, respectively.

The proposition below presents a result characterizing $f_{i:n}(x)$.

**Proposition 2.5** *The pdf of $X_{i:n}$ can be expressed as a linear combination of pdfs of the exp-G family of distributions.*

**Proof 2.5** *Let us consider the expression of $f_{i:n}(x)$ given by* (14)*. It follows from the binomial formula and* (12) *that*

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!}\sum_{j=0}^{n-i}\binom{n-i}{j}(-1)^j f(x)F(x)^{j+i-1}$$
$$= \frac{n!}{(i-1)!(n-i)!}\sum_{j=0}^{n-i}\binom{n-i}{j}(-1)^j\left\{\sum_{\ell=1}^{+\infty} w_\ell\left[a\ell g(x)G(x)^{a\ell-1}\right]\right\}\left[\sum_{k=0}^{+\infty} w_k G(x)^{ak}\right]^{j+i-1}.$$

*By virtue of a result established by (Gradshteyn and Ryzhik, 2000, Section 0.314), we have*

$$\left[\sum_{k=0}^{+\infty} w_k G(x)^{ak}\right]^{j+i-1} = \sum_{m=0}^{+\infty} d_{j+i-1,m}G(x)^{am},$$

*where $d_{j+i-1,0} = w_0^{j+i-1}$ and, for any integer $m \geq 1$,*

$$d_{j+i-1,m} = \frac{1}{mw_0}\sum_{k=1}^{m}(k(j+i)-m)w_k d_{j+i-1,m-k}.$$

*By putting the equalities above together, we obtain*

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!}\sum_{j=0}^{n-i}\sum_{\ell=1}^{+\infty}\sum_{m=0}^{+\infty}\binom{n-i}{j}(-1)^j w_\ell d_{j+i-1,m}\frac{\ell}{\ell+m}q_{\ell,m}(x), \qquad (15)$$

*where $q_{\ell,m}(x) = a(\ell+m)g(x)G(x)^{a(\ell+m)-1}$. Since $q_{\ell,m}(x)$ is a pdf of the exp-G family with parameter $a(\ell+m)$, the proof of Proposition 2.5 is complete.*

By using the existing results on the exp-G family, we can use Proposition 2.5 to derive mathematical properties of the distribution of the *i*-th order statistics, as moments and all the related quantities.

## 3. GKw-Exponential distribution

### 3.1. Definition

In this section, we focus our attention on the special member of the GKw-G family based on the exponential distribution. Hence, by substituting the cdf $G(x) = 1 - e^{-\lambda x}$, $x > 0$, into (5), the cdf of this special distribution is given by

$$F_{GKw-E}(x) = \gamma_1\left(\alpha, \left\{1 - \left(1 - e^{-\lambda x}\right)^a\right\}^{-b} - 1\right), \quad x > 0. \qquad (16)$$

The related distribution is called the GKw-Exponential (GKw-E) distribution. Naturally, the corresponding sf is

$$S_{GKw-E}(x) = 1 - \gamma_1\left(\alpha, \left\{1-\left(1-e^{-\lambda x}\right)^a\right\}^{-b} - 1\right), \quad x > 0.$$

The corresponding pdf is specified by

$$f_{GKw-E}(x) = \frac{ab\lambda}{\Gamma(\alpha)} e^{-\lambda x}\left(1-e^{-\lambda x}\right)^{a-1}\left\{1-\left(1-e^{-\lambda x}\right)^a\right\}^{-b-1}\left\{\left\{1-\left(1-e^{-\lambda x}\right)^a\right\}^{-b}-1\right\}^{\alpha-1} \times \exp\left[1-\left\{1-\left(1-e^{-\lambda x}\right)^a\right\}^{-b}\right]. \quad x > 0, \tag{17}$$

and the corresponding hrf is given as

$$\pi_{GKw-E}(x) = \frac{ab\lambda}{\Gamma(\alpha)} \frac{e^{-\lambda x}\left(1-e^{-\lambda x}\right)^{a-1}\left\{1-\left(1-e^{-\lambda x}\right)^a\right\}^{-b-1}\left\{\left\{1-\left(1-e^{-\lambda x}\right)^a\right\}^{-b}-1\right\}^{\alpha-1}}{1-\gamma_1\left(\alpha, \left\{1-\left(1-e^{-\lambda x}\right)^a\right\}^{-b}-1\right)} \times \exp\left[1-\left\{1-\left(1-e^{-\lambda x}\right)^a\right\}^{-b}\right], \qquad x > 0. \tag{18}$$

Let us now investigate some asymptotic properties of $F_{GKw-E}(x)$, $S_{GKw-E}(x)$, $f_{GKw-E}(x)$ and $h_{GKw-E}(x)$. When $x \to 0$, we have

$$F_{GKw-E}(x) \sim \frac{b^\alpha \lambda^{a\alpha}}{\alpha\Gamma(\alpha)} x^{a\alpha}, \quad f_{GKw-E}(x) \sim \frac{ab^\alpha \lambda^{a\alpha}}{\Gamma(\alpha)} x^{a\alpha-1}, \quad h_{GKw-E}(x) \sim \frac{ab^\alpha \lambda^{a\alpha}}{\Gamma(\alpha)} x^{a\alpha-1}.$$

The following limits follow. If $a\alpha < 1$, we have $f_{GKw-E}(x) \to +\infty$, if $a\alpha = 1$, we have $f_{GKw-E}(x) \to ab^{1/a}\lambda/\Gamma(\alpha)$, and if $a\alpha > 1$, we have $f_{GKw-E}(x) \to 0$. Similarly, if $a\alpha < 1$, we have $h_{GKw-E}(x) \to +\infty$, if $a\alpha = 1$, we have $h_{GKw-E}(x) \to ab^{1/a}\lambda/\Gamma(\alpha)$, and if $a\alpha > 1$, we have $h_{GKw-E}(x) \to 0$. When $x \to +\infty$, we have

$$S_{GKw-E}(x) \sim \frac{a^{-b(\alpha-1)}}{\Gamma(\alpha)} e^{\lambda b(\alpha-1)x} e^{1-a^{-b}e^{\lambda bx}}, \qquad f_{GKw-E}(x) \sim \frac{\lambda b a^{-\alpha b}}{\Gamma(\alpha)} e^{\lambda b\alpha x} e^{1-a^{-b}e^{\lambda bx}}$$

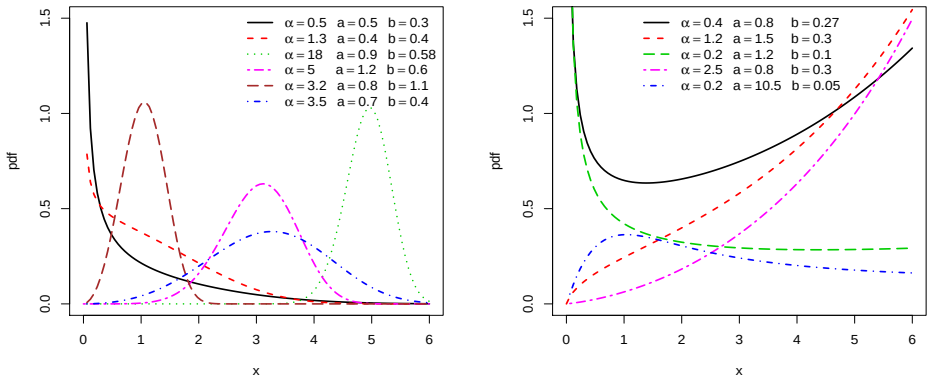
and

$$h_{GKw-E}(x) \sim \lambda b a^{-b} e^{\lambda bx}.$$

Hence, we have $f_{GKw-E}(x) \to 0$ and $h_{GKw-E}(x) \to +\infty$.

In order to give more concrete illustrations on their shapes, Figure 1 displays some plots of the GKw-E pdf and hrf for specified parameters values. It indicates that the GKw-E distribution can be right-skewed, left-skewed and reversed-J shaped, whereas the GKw-E hrf can produce various shapes such as increasing, decreasing, bathtub and upside-down

bathtub shapes.




(a) (b)

Figure 1: Plots of (a) GKw-E pdfs and (b) GKw-E hrfs for some parametric values with fixed $\lambda = 1$.

Since $Q_G(p) = -(1/\lambda)\log(1-p)$, based on (7), the GKw-E quantile function is given by

$$Q_{GKw-E}(p) = -\frac{1}{\lambda}\log\left[1-\left[1-\left\{1+\gamma_1^{-1}(\alpha,p)\right\}^{-1/b}\right]^{1/a}\right], \qquad p \in (0,1).$$

From this definition, the quartiles and octiles can be determined, as well as skewness and kurtosis, and some results on distributions, as the useful one: for a random variable $U$ following the uniform distribution on $(0,1)$, $Q_{GKw-E}(U)$ follows the GKw-E distribution.

### 3.2. Linear representation with applications

A result on linear representations of $F_{GKw-E}(x)$ and $f_{GKw-E}(x)$ in terms of exponential functions is presented below.

**Proposition 3.1** *We have the following linear representations for the cdf and pdf of the GKw-E distribution:*

$$F_{GKw-E}(x) = \sum_{m=0}^{+\infty} w_m^* e^{-\lambda m x}, \quad f_{GKw-E}(x) = \sum_{m=1}^{+\infty} w_m^{**} e^{-\lambda m x}, \quad x > 0,$$

*where*

$$w_m^* = \sum_{i,j,k=0}^{+\infty} \frac{(-1)^{i+j+k+m}}{\Gamma(\alpha)k!(\alpha+k)} \binom{\alpha+k}{j} \binom{b(j-\alpha-k)}{i} \binom{\alpha i}{m}, \quad w_m^{**} = -\lambda m w_m^*.$$

**Proof 3.1** *Let $G(x) = 1 - e^{-\lambda x}$ and $g(x) = \lambda e^{-\lambda x}$. Then, owing to Proposition 2.3, we have*

$$F_{GKw-E}(x) = \sum_{i=0}^{+\infty} w_i G(x)^{ai}, \quad f_{GKw-E}(x) = \sum_{i=1}^{+\infty} w_i \left[ aig(x)G(x)^{ai-1} \right],$$

*where*

$$w_i = \sum_{j,k=0}^{+\infty} \frac{(-1)^{i+j+k}}{\Gamma(\alpha)k!(\alpha+k)} \binom{\alpha+k}{j} \binom{b(j-\alpha-k)}{i}.$$

*Now, for any positive integer i, by virtue of the generalized binomial formula, we have*

$$G(x)^{\alpha i} = (1 - e^{-\lambda x})^{\alpha i} = \sum_{m=0}^{+\infty} \binom{\alpha i}{m} (-1)^m e^{-\lambda m x}.$$

*Therefore*

$$F_{GKw-E}(x) = \sum_{i=0}^{+\infty} w_i G(x)^{ai} = \sum_{m=0}^{+\infty} w_m^* e^{-\lambda m x},$$

*where $w_m^* = \sum_{i=0}^{+\infty} \binom{\alpha i}{m} (-1)^m w_i$. The desired expansion for the pdf is obtained by differentiating $F_{GKw-E}(x)$. This ends the proof of Proposition 3.1.*

Thanks to Proposition 3.1, several structural properties of the GKw-E distribution can be derived. Some of them are described below.

The $r$-th ordinary moment of the GKw-E distribution is defined by

$$\mu'_r = \sum_{m=1}^{+\infty} w_m^{**} \int_0^{+\infty} x^r e^{-\lambda m x} dx = \frac{1}{\lambda^{r+1}} \Gamma(r+1) \sum_{m=1}^{+\infty} w_m^{**} \frac{1}{m^{r+1}}.$$

Then, we can easily deduce the mean, the variance, the $r$-th central moment, the coefficient of skewness and the coefficient of kurtosis. The numerical values of these measures for some chosen parameters are collected in Table 2.

Table 2: First four moments, variance skewness and kurtosis of the GKw-E distribution for some parameter values.

| $(\alpha,a,b,\lambda)$ | $\mu'_1$ | $\mu'_2$ | $\mu'_3$ | $\mu'_4$ | $\sigma^2$ | $CS$ | $CK$ |
|---|---|---|---|---|---|---|---|
| $(0.5,0.5,0.5,0.5)$ | 0.7273 | 1.6188 | 5.0408 | 18.9862 | 1.0898 | 2.0021 | 10.2919 |
| $(2,0.5,0.5,0.5)$ | 2.8256 | 10.5036 | 45.5683 | 219.8988 | 2.5192 | 0.4129 | 11.6882 |
| $(4,0.5,0.5,0.5)$ | 4.8084 | 25.4660 | 144.9078 | 872.9899 | 2.3447 | -0.0260 | 77.1860 |
| $(0.5,2,0.5,0.5)$ | 2.1594 | 7.7311 | 35.2336 | 186.3619 | 3.0678 | 0.9841 | 5.0394 |
| $(0.5,3,0.5,0.5)$ | 2.7489 | 11.1864 | 56.3355 | 325.4536 | 3.6295 | 0.8140 | 4.6925 |
| $(0.5,4,0.5,0.5)$ | 3.2045 | 14.2449 | 76.8333 | 471.4195 | 3.9758 | 0.7194 | 4.8590 |
| $(2,3,0.5,0.5)$ | 6.0439 | 40.1030 | 285.8935 | 2158.8840 | 3.5734 | 0.0468 | 55.1839 |
| $(4,3,0.5,0.5)$ | 8.2805 | 71.2060 | 632.6335 | 5784.2410 | 2.6379 | -0.1592 | 355.3071 |
| $(2,2,1,0.5)$ | 3.1275 | 10.8143 | 40.2792 | 159.0419 | 1.0327 | -0.0030 | 136.0614 |
| $(2,2,1.5,0.5)$ | 2.3533 | 6.0758 | 16.8069 | 49.0889 | 0.5375 | -0.0555 | 237.7113 |
| $(2,2,1.5,0.1)$ | 11.7668 | 151.8967 | 2100.8630 | 30680.5800 | 13.4387 | -0.0583 | 0.1911 |

It is clear from Table 2 that the GKw-E distribution is numerically versatile in mean and variance. Also, the values of CS reveal that it can be right-skewed, almost symmetrical, and slightly left-skewed. The values of CK indicate that the GKw-E distribution can be mesokurtic, leptokurtic (thin bell shape) and platykurtic (flat bell shape). All these characteristics illustrate a certain flexibility of the GKw-E distribution, which remains attractive for modelling purposes.

In addition, the $r$-th incomplete moment is obtained as, for $t \geq 0$,

$$I_r(t) = \int_{-\infty}^{t} x^r f_{GKw-E}(x)dx = \sum_{m=1}^{+\infty} w_m^{**} \int_0^t x^r e^{-\lambda m x} dx = \frac{1}{\lambda^{r+1}} \sum_{m=1}^{+\infty} w_m^{**} \frac{1}{m^{r+1}} \gamma(r+1, \lambda m t).$$

The incomplete moments are useful to determine other important mathematical quantities such as the Bonferroni and Lorenz curves, mean residual-life, mean waiting-time, mean deviation about the mean and mean deviation about the median.

## 4. Estimation and application

In this section, we adopt the GKw-E distribution as a model and consider the estimation of the unknown parameters by the maximum likelihood method. In addition, the convergence of the obtained estimates is investigated through a simulation study and applications are given to two practical data sets.

### 4.1. Method of estimation

The usefulness of the maximum likelihood estimates (MLEs) in statistical inference is due to their theoretical and practical merits. The log-likelihood function for the vector of parameters $\Omega = \left(a, b, \alpha, \lambda\right)^{\top}$ is given by

$$\ell(\Omega) = n\log(a) + n\log(b) - n\log\left[\Gamma(\alpha)\right] + n\log(\lambda) - \lambda\sum_{i=1}^{n} x_i + (a-1)\sum_{i=1}^{n}\log\left[1-e^{-\lambda x_i}\right]$$
$$-(b+1)\sum_{i=1}^{n}\log\left[1-\left(1-e^{-\lambda x_i}\right)^{a}\right] + (\alpha-1)\sum_{i=1}^{n}\log\left[\left\{1-\left(1-e^{-\lambda x_i}\right)^{a}\right\}^{-b} - 1\right] + n$$
$$-\sum_{i=1}^{n}\left\{1-\left(1-e^{-\lambda x_i}\right)^{a}\right\}^{-b}.$$

The MLEs of the parameters are defined by $\widehat{\Omega} = \left(\widehat{a}, \widehat{b}, \widehat{\alpha}, \widehat{\lambda}\right)^{\top}$ making maximum the log-likelihood function $\ell(\Omega)$ with respect to $\Omega$. Since they have no closed forms, one can use standard statistical software to approximate them. Also, let us mention that the observed Fisher information for the MLEs can be computed, allowing the construction of confidence intervals for the parameters based on the limiting normal distribution. In particular, this is useful to examine the probability coverage of these intervals through simulation.

### 4.2. A numerical study

Now, we assess the performance of the maximum likelihood method for estimating the GKw-E parameters by using Monte Carlo simulations. The simulation study is repeated 5000 times each with sample sizes $n = 50$, 100, 200 and the following parameter scenarios are followed: I: $a = 0.5$, $b = 0.5$, $\alpha = 0.5$, and $\lambda = 1$, II: $a = 0.3$, $b = 1.5$, $\alpha = 0.7$, and $\lambda = 2.5$ and III: $a = 1.7$, $b = 0.7$, $\alpha = 0.2$, and $\lambda = 0.3$, IV: $a = 0.1$, $b = 2.5$, $\alpha = 1.1$, and $\lambda = 1.5$,V: $a = 2.5$, $b = 1.7$, $\alpha = 2.5$, and $\lambda = 1$, VI: $a = 1.8$, $b = 1.7$, $\alpha = 2.1$, and $\lambda = 0.1$. Under this setting, Table 3 gives the average biases (Bias) of the MLEs, mean square errors (MSEs) and model-based coverage probabilities (CPs) for the parameters $a$, $b$, $\alpha$ and $\lambda$. Based on these results, we conclude that the MLEs perform quite well in estimating the parameters. In addition, the CPs of the confidence intervals are quite close to the 95% nominal level. Therefore, the MLEs and their asymptotic results can be adopted to estimate and construct efficiently confidence intervals for the model parameters.

Table 3: Monte Carlo simulation results for the GKw-E distribution: Biases, MSEs and CPs.

| | | I | | | II | | | III | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | Bias | MSE | CP | Bias | MSE | CP | Bias | MSE | CP |
| $a$ | 50 | −0.015 | 0.051 | 0.98 | −0.008 | 0.044 | 0.94 | 0.810 | 14.386 | 0.85 |
| | 100 | 0.007 | 0.047 | 0.97 | 0.023 | 0.049 | 0.95 | 0.616 | 4.488 | 0.90 |
| | 200 | 0.039 | 0.045 | 0.96 | 0.004 | 0.037 | 0.95 | 0.576 | 2.908 | 0.95 |
| $b$ | 50 | −0.140 | 0.162 | 0.97 | −0.404 | 1.318 | 0.90 | 0.244 | 3.047 | 0.97 |
| | 100 | −0.125 | 0.127 | 0.97 | −0.217 | 0.918 | 0.96 | 0.307 | 2.484 | 0.98 |
| | 200 | −0.113 | 0.104 | 0.95 | −0.072 | 0.477 | 0.99 | 0.287 | 0.977 | 0.99 |
| $\alpha$ | 50 | 0.153 | 0.257 | 0.91 | 0.465 | 1.300 | 0.92 | 0.452 | 1.404 | 0.83 |
| | 100 | 0.084 | 0.116 | 0.91 | 0.307 | 0.710 | 0.93 | 0.225 | 0.989 | 0.89 |
| | 200 | 0.046 | 0.082 | 0.89 | 0.306 | 0.628 | 0.96 | 0.139 | 0.958 | 0.96 |
| $\lambda$ | 50 | 1.807 | 6.527 | 0.95 | 2.601 | 2.726 | 0.92 | 0.752 | 1.324 | 1.00 |
| | 100 | 1.461 | 4.742 | 0.94 | 1.136 | 1.129 | 0.93 | 0.555 | 1.002 | 1.00 |
| | 200 | 1.180 | 3.136 | 0.95 | 0.202 | 0.847 | 0.97 | 0.364 | 0.743 | 0.97 |
| | | IV | | | V | | | VI | | |
| | $n$ | Bias | MSE | CP | Bias | MSE | CP | Bias | MSE | CP |
| $a$ | 50 | −0.904 | 1.154 | 0.65 | 0.146 | 0.535 | 0.94 | 0.441 | 1.253 | 0.95 |
| | 100 | −0.665 | 0.461 | 0.92 | 0.164 | 0.309 | 0.95 | 0.194 | 0.579 | 0.96 |
| | 200 | −0.002 | 0.019 | 0.97 | 0.195 | 0.228 | 0.97 | 0.015 | 0.263 | 0.99 |
| $b$ | 50 | −0.032 | 0.349 | 0.98 | 0.172 | 0.241 | 1.00 | 0.018 | 0.893 | 0.95 |
| | 100 | 0.014 | 0.333 | 0.98 | 0.053 | 0.065 | 0.96 | 0.072 | 0.633 | 0.96 |
| | 200 | −0.051 | 0.052 | 0.96 | 0.001 | 0.031 | 0.97 | 0.136 | 0.438 | 0.98 |
| $\alpha$ | 50 | 0.477 | 0.480 | 0.89 | 0.311 | 0.163 | 0.99 | −0.158 | 0.112 | 0.97 |
| | 100 | 0.270 | 0.163 | 0.96 | 0.271 | 0.132 | 0.95 | −0.145 | 0.106 | 0.96 |
| | 200 | −0.051 | 0.052 | 0.98 | 0.222 | 0.100 | 0.96 | −0.148 | 0.110 | 0.97 |
| $\lambda$ | 50 | 0.337 | 0.601 | 0.99 | −0.062 | 0.022 | 0.95 | 0.179 | 0.298 | 0.95 |
| | 100 | 0.214 | 0.284 | 0.96 | −0.059 | 0.017 | 0.96 | 0.204 | 0.323 | 0.96 |
| | 200 | 0.243 | 0.814 | 0.98 | −0.051 | 0.011 | 0.98 | 0.253 | 0.392 | 0.97 |

### 4.3. Application

Here, we compare the proposed GKw-E model with well-known models in the fitting of two real data sets.

**Application 1.** The first data set is reported in Ristic and Balakrishnan (2012). The data represent the annual maximum precipitation (inches) for one rain gauge in Fort Collins, Colorado from 1900 through 1999. The data are as follows: 239, 232, 434, 85, 302, 174, 170, 121, 193, 168, 148, 116, 132, 132, 144, 183, 223, 96, 298, 97, 116, 146, 84, 230, 138, 170, 117, 115, 132, 125, 156, 124, 189, 193, 71, 176, 105, 93, 354, 60, 151, 160, 219, 142, 117, 87, 223, 215, 108, 354, 213, 306, 169, 184, 71, 98, 96, 218, 176, 121, 161, 321, 102, 269, 98, 271, 95, 212, 151, 136, 240, 162, 71, 110, 285, 215, 103, 443, 185, 199, 115, 134, 297, 187, 203, 146, 94, 129, 162, 112, 348, 95, 249, 103, 181, 152, 135, 463, 183, 241.

In the statistical literature, several models are appropriate to the analysis of such kinds of data. The most commonly used are the lognormal, generalized logistic (GL), Gumbel, gamma, Weibull and generalized binomial exponential 2 (GBE2) models. Several extensions have also been introduced by this purpose. Here, in order to highlight the potentiality of the GKw-E model, the comparison is made between the GKw-E model and eights noto-

rious models: the Kumaraswamy Weibull (Kw-W) model studied by Cordeiro *et al.* (2010), the Beta Weibull (BW) model elaborated by Lee *et al.* (2007), the exponentiated generalized Weibull (EGW) model by Oguntunde *et al.* (2015), the generalized binomial exponential 2 (GBE2) model introduced by Asgharzadeh *et al.* (2016), the generalized logistic (GL) model, and some classic models, which are the Gumbel, gamma and Weibull models. We estimate the unknown model parameters by the maximum likelihood method (as described in Section 4 for the GKw-E model). For the comparison of the models, we consider three well-known statistics: Akaike information criterion (AIC), Anderson-Darling ($A^*$), Cramér–von Mises ($W^*$) and Kolmogorov-Smirnov (K-S) measures, where lower values of these statistics indicate good fits.

Table 4 lists the MLEs and standard errors for the considered models. Table 5 lists the AIC, $A^*$, $W^*$ and K-S for the considered models. The values of the statistics in Table 5 indicate that the GKw-E model shows small values of the statistics and thus provides the best fit compared to the other models. Figure 2 shows the graphs of the estimated pdf of the GKw-E model over the histogram of the data, along with the graphs of the pdfs of the top four competitors.

Table 4: MLEs and their standard errors (in parentheses) for Precipitation data.

| | $\alpha$ | $\beta$ | $a$ | $b$ | $\mu$ | $\sigma$ | $\theta$ | $\lambda$ |
|---|---|---|---|---|---|---|---|---|
| GKw-E | 0.2975 | - | 67.1975 | 0.1802 | - | - | - | 0.0261 |
| | (0.1594) | - | (24.7418) | (0.0599) | - | - | - | (0.0072) |
| Kw-W | 0.0228 | 1.3122 | 13.4486 | 0.2461 | - | - | - | - |
| | (0.0053) | (0.2462) | (7.6120) | (0.1229) | - | - | - | - |
| BW | 0.0243 | 1.4375 | 12.6298 | 0.1734 | - | - | - | - |
| | (0.0033) | (0.0193) | (5.5638) | (0.0446) | - | - | - | - |
| EGW | 0.3105 | 0.7061 | 0.2357 | 27.1942 | - | - | - | - |
| | (0.0148) | (0.0117) | (0.0276) | (7.6257) | - | - | - | - |
| GBE2 | 9.0774 | - | - | - | - | - | 0.0222 | 0.0165 |
| | (1.9764) | - | - | - | - | - | (0.3265) | (0.0029) |
| GL | 13.5845 | 0.0174 | - | - | -8.5348 | - | - | - |
| | (6.8592) | (0.0015) | - | - | (35.1221) | - | - | - |
| Gumbel | - | - | - | - | 139.8754 | 57.8420 | - | - |
| | - | - | - | - | (6.0596) | (4.7356) | - | - |
| Gamma | 33.2955 | 5.2761 | - | - | - | - | - | - |
| | (4.7925) | (0.7239) | - | - | - | - | - | - |
| Weibull | 0.0051 | 2.2608 | - | - | - | - | - | - |
| | (0.0002) | (0.1628) | - | - | - | - | - | - |

Table 5: The statistics AIC, $A^*$, $W^*$ and K-S for Precipitation data.

| Distribution | AIC | $A^*$ | $W^*$ | K-S |
|---|---|---|---|---|
| GKw-E | 1137.2320 | 0.1664 | 0.0187 | 0.0421 |
| Kw-W | 1138.0280 | 0.1831 | 0.0212 | 0.0430 |
| BW | 1137.7220 | 0.1844 | 0.0210 | 0.0429 |
| EGW | 1138.7100 | 0.2045 | 0.0259 | 0.0481 |
| GBE2 | 1138.9210 | 0.3655 | 0.0482 | 0.0573 |
| GL | 1143.1390 | 0.6335 | 0.0872 | 0.0565 |
| Gumbel | 1139.2900 | 0.4990 | 0.0675 | 0.0640 |
| Gamma | 1141.9400 | 0.7732 | 0.1088 | 0.0600 |
| Weibull | 1156.2860 | 1.8272 | 0.2927 | 0.0950 |

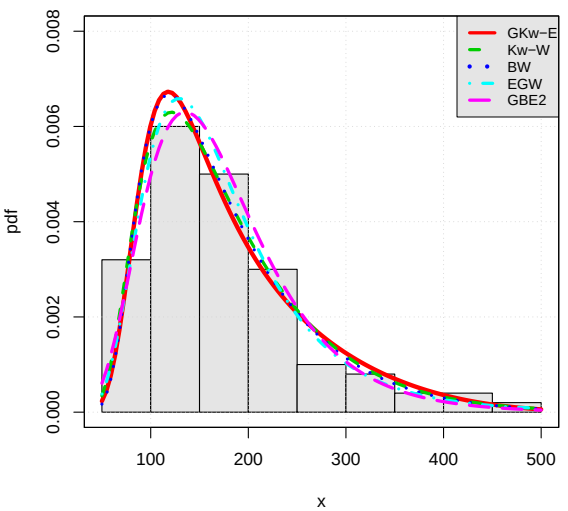



Figure 2: Estimated pdfs of the top models for Precipitation data.

**Application 2.** The second data set was reported by professor Jim Irish and can be obtained at http://www.statsci.org/data/oz/kiama.html. It is about the Kiama Blowhole eruptions. The data are as follows: 83, 51, 87, 60, 28, 95, 8, 27, 15, 10, 18, 16, 29, 54, 91, 8, 17, 55, 10, 35, 47, 77, 36, 17, 21, 36, 18, 40,10, 7, 34, 27, 28, 56, 8, 25, 68, 146, 89, 18, 73, 69, 9, 37, 10, 82, 29, 8, 60, 61, 61, 18, 169, 25, 8, 26, 11, 83, 11, 42, 17, 14, 9, 12.

Table 6 lists the MLEs and standard errors for the considered models. Table 7 lists the AIC, $A^*$, $W^*$ and K-S for the considered models. It is clear that the GKw-E model provides a better fit than the other tested models, because it has the smallest value among AIC, $A^*$, $W^*$ and K-S. Figure 3 shows the graphs of the estimated pdf of the GKw-E model over the histogram of the data, along with the graphs of the pdfs of the four main competitors.

Table 6: MLEs and their standard errors (in parentheses) for the Kiama Blowhole eruptions data.

| | $\alpha$ | $\beta$ | $a$ | $b$ | $\mu$ | $\sigma$ | $\theta$ | $\lambda$ |
|---|---|---|---|---|---|---|---|---|
| GKw-E | 0.4154 | - | 17.7076 | 0.0481 | - | - | - | 0.2063 |
| | (0.0545) | - | (0.2513) | (0.0072) | - | - | - | (0.0046) |
| Kw-W | 0.3410 | 0.8685 | 10.4397 | 0.1396 | - | - | - | - |
| | (0.0026) | (0.0022) | (0.0083) | (0.0168) | - | - | - | - |
| BW | 0.5484 | 0.7937 | 13.5819 | 0.1336 | - | - | - | - |
| | (0.0025) | (0.0025) | (4.8229) | (0.0177) | - | - | - | - |
| EGW | 2.5406 | 0.3714 | 0.7506 | 26.1285 | - | - | - | - |
| | (9.4366) | (0.2260) | (3.0932) | (0.8858) | - | - | - | - |
| GBE2 | 1.7325 | - | - | - | - | - | 0.0048 | 0.0350 |
| | (0.3190) | - | - | - | - | - | (0.5680) | (0.0111) |
| GL | 21.5045 | 0.0473 | - | - | -38.5692 | - | - | - |
| | (6.5526) | (0.0048) | - | - | (7.8114) | - | - | - |
| Gumbel | - | - | - | - | 25.6833 | 21.8407 | - | - |
| | - | - | - | - | (2.8506) | (2.3260) | - | - |
| Gamma | 24.5722 | 1.6207 | - | - | - | - | - | - |
| | (4.6509) | (0.2623) | - | - | - | - | - | - |
| Weibull | 0.0230 | 1.2701 | - | - | - | - | - | - |
| | (0.0023) | (0.1199) | - | - | - | - | - | - |

Table 7: The statistics AIC, $A^*$, $W^*$ and K-S for the Kiama Blowhole eruptions data.

| Distribution | AIC | $A^*$ | $W^*$ | K-S |
|---|---|---|---|---|
| GKw-E | 589.2545 | 0.4614 | 0.0530 | 0.0708 |
| Kw-W | 591.0460 | 0.6231 | 0.0819 | 0.0954 |
| BW | 591.6412 | 0.6366 | 0.0840 | 0.1023 |
| EGW | 595.9134 | 0.8324 | 0.1134 | 0.0946 |
| GBE2 | 597.3321 | 0.9009 | 0.1287 | 0.1227 |
| GL | 612.7799 | 1.5554 | 0.2440 | 0.1517 |
| Gumbel | 609.6039 | 1.5124 | 0.2361 | 0.1493 |
| Gamma | 595.7988 | 0.9220 | 0.1324 | 0.1215 |
| Weibull | 597.8029 | 1.0058 | 0.1467 | 0.1111 |

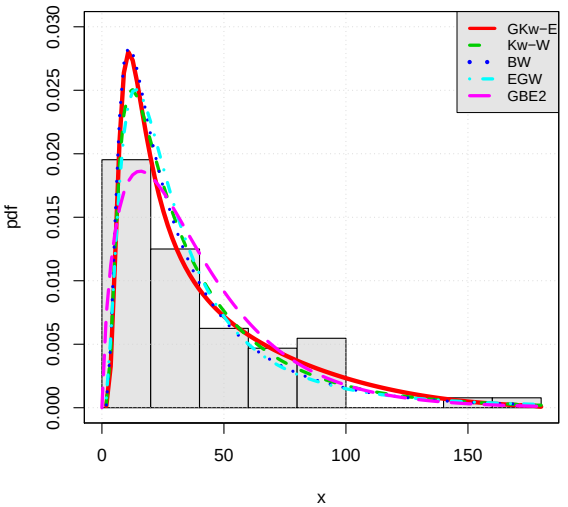



Figure 3: Estimated pdfs of the top models for Kiama Blowhole eruptions data.

## 5. Concluding remarks

In this paper, we introduce the GKw-G family of distributions, with a focus on a special model, the GKw-E model, defined with the exponential distribution as the parent. A complete theoretical treatment is developed, with a focus on the skewness, kurtosis, analytical comportments of the main functions, moments, stochastic ordering and order statistics. Then, the proposed family is considered from the statistical point of view. The maximum likelihood method is employed for estimating the model parameters. We analyse two practical data sets to demonstrate the usefulness of the new family, with fair comparison to other models. The results are strictly favourable to the GKw-E model. We hope that the proposed family and its generated models will attract wider applications in various areas such as engineering, survival and lifetime data, hydrology and economics.

### Acknowledgments

The authors are very grateful to two reviewers for constructive comments, which have helped to improve the final version of the paper.

### Conflict of interest

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

## References

ABRAMOWITZ, M., STEGUN, I. A., (1965). *Handbook of Mathematical Functions*, National Bureau of Standards, Applied Math. Series 55, Dover Publications.

ALDAHLAN, M. A., JAMAL, F., CHESNEAU, C., ELGARHY, M. and ELBATAL, I., (2019). The truncated Cauchy power family of distributions with inference and applications, *Entropy*, 22, p. 346.

ALEXANDER, C., CORDEIRO, G. M., ORTEGA, E. M. M. and SARABIA, J. M., (2012). Generalized beta-generated distributions. *Computational Statistics & Data Analysis*, 56, pp. 1880–1897.

ALJARRAH, M. A., LEE, C. and FAMOYE, F., (2014). On generating T-X family of distributions using quantile functions. *Journal of Statistical Distributions and Applications*, 1, Article No. 2.

ALZAATREH, A., FAMOYE, F. and LEE, C., (2014). T-normal family of distributions: A new approach to generalize the normal distribution. *Journal of Statistical Distributions and Applications*, 1, Article No. 16.

ALZAATREH, A., LEE, C. and FAMOYE, F., (2013a). A new method for generating families of distributions. *Metron*, 71, pp. 63–79.

ALZAGHAL, A., LEE, C. and FAMOYE, F., (2013b). Exponentiated T-X family of distributions with some applications. *International Journal of Probability and Statistics*, 2, pp. 31–49.

AMINI, M., MIRMOSTAFAEE, S. M. T. K. and AHMADI, J., (2014). Log-gamma-generated families of distributions. *Statistics*, 48, pp. 913–932.

ASGHARZADEH, A., BAKOUCH, H. S. and HABIBI, M., (2016). A generalized binomial exponential 2 distribution: modeling and applications to hydrologic events. *Journal of Applied Statistics*, 44, pp. 2368–2387.

BOURGUIGNON, M., SILVA, R. B. and CORDEIRO, G. M., (2014). The Weibull-G family of probability distributions. *Journal of Data Science*, 12, pp. 53–68.

CORDEIRO, G. M., ALIZADEH, M. and ORTEGA, E. M. M., (2014). The exponentiated half-logistic family of distributions: Properties and applications. *Journal of Probability and Statistics* Article ID 864396, 21 pages.

CORDEIRO, G. M., DE CASTRO, M., (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, 81, pp. 883–893.

CORDEIRO, G. M., ORTEGA, E. M. M. and DA CUNHA, D. C. C., (2013). The exponentiated generalized class of distributions. *Journal of Data Science*, 11, pp. 1–27.

CORDEIRO, G. M., ORTEGA, E. M. M. and NADARAJAH, S., (2010). The Kumaraswamy Weibull distribution with application to failure data. *Journal of the Franklin Institute*, 347, pp. 1399–1429.

DAVID, H. A., NAGARAJA, H. N., (2003). *Order Statistics*. John Wiley and Sons, New Jersey.

DE PASCOA, M. A. R., ORTEGA, E. M. M. and CORDEIRO, G. M., (2011). The Kumaraswamy Weibull distribution with application to failure data. *Journal of Franklin Institute*, 347, pp. 1399–1429.

EUGENE, N., LEE, C. and FAMOYE, F., (2002). Beta-normal distribution and its applications. *Communications in Statistics - Theory and Methods*, 31, pp. 497–512.

GOMES, A. E., DA SILVA, C. Q., CORDEIRO, G. M. and ORTEGA, E. M. M., (2014). A new lifetime model: The Kumaraswamy generalized Rayleigh distribution. *Journal of Statistical Computation and Simulation*, 84, pp. 290–309.

GRADSHTEYN, I. S., RYZHIK, I. M., (2000). *Table of Integrals, Series and Products*. Academic Press, New York.

HOSSEINI, B., AFSHARI, M. and ALIZADEH, M., (2018). The Generalized Odd Gamma-G Family of Distributions: Properties and Applications. *Austrian Journal of Statistics*, 47, pp. 69–89.

JAMAL, F., CHESNEAU, C. and ELGARHY, M., (2020). Type II general inverse exponential family of distributions, *Journal of Statistics and Management Systems* 23, 3, pp. 617–641.

JAMAL, F., NASIR, M. A., OZEL, G., ELGARHY, M. and KHAN, N. M., (2019). Generalized inverted Kumaraswamy generated family of distributions: theory and applications. *Journal of Applied Statistics*, 46, pp. 2927–2944.

JAMAL, F., NASIR, M. A., TAHIR, M. H. and MONTAZERI, N. H., (2017). The odd Burr-III family of distributions. *Journal of Statistics Applications and Probability*, 6, pp. 105–122.

JONES, M. C., (2004). Families of distributions arising from the distributions of order statistics. *Test*, 13, pp. 1–43.

JONES, M. C., (2008). Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, 6, pp. 70–81.

KENNEY, J., KEEPING, E., (1962). *Mathematics of Statistics*. Vol. 1, 3rd edition, Princeton:NJ, Van Nostrand.

LEE, C., FAMOYE, F. and OLUMOLADE, O., (2007). Beta-Weibull Distribution: Some Properties and Applications to Censored Data. *Journal of Modern Applied Statistical Methods*, 6, pp. 173–186.

MOORS, J. J. A., (1998). A quantile alternative for kurtosis. *Statistician*, 37, pp. 25–32.

NASIR, A., BAKOUCH, H. S. and JAMAL, F., (2018). Kumaraswamy Odd Burr G Family of Distributions with Applications to Reliability Data. *Studia Scientiarum Mathematicarum Hungarica*, 55, pp. 1–21.

NASIR, M. A., TAHIR, M. H., CHESNEAU, C., JAMAL, F. and SHAH, M. A. A., (2020). The odds generalized gamma-G family of distributions: Properties, regressions and applications. *Statistica*, 80, 1, pp. 3–38.

OGUNTUNDE, P. E., ODETUNMIBI, O. A. and ADEJUMO, A. O., (2015). On the Exponentiated Generalized Weibull Distribution: A Generalization of the Weibull Distribution. *Indian Journal of Science and Technology*, 8, pp. 1–7.

OLUYEDE, B. O., PU, S., MAKUBATE, B. and QIU, Y., (2018). The Gamma-Weibull-G Family of Distributions with Applications. *Austrian Journal of Statistics*, 47, pp. 45–76.

PARANAIBA, P. F., ORTEGA, E. M. M., CORDEIRO, G. M. and de Pascoa, M. A. D., (2012). The Kumaraswamy Burr XII distribution: Theory and practice. *Journal of Statistical Computation and Simulation*, 82, pp. 1–27.

RAMOS, M. W. A., (2014). Some new extended distributions: theory and applications, 88 f. Tese (Doutorado em Matemática Computacional). Universidade Federal de Pernambuco. Recife.

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

RODRIGUES, J. A., SILVA, A. P. C., (2015). The exponentiated Kumaraswamy-exponential distribution. *British Journal of Applied Science and Technology*, 10, pp. 1–12.

SHAKED, M., SHANTHIKUMAR, J. G., (1994). *Stochastic orders and their applications*. Academic Press, New York.

TORABI, H., MONTAZARI, N. H., (2012). The gamma-uniform distribution and its application. *Kybernetika*, 48, pp. 16–30.

TORABI, H., MONTAZARI, N. H., (2014). The logistic-uniform distribution and its application, *Communications in Statistics - Simulation and Computation*, 43, pp. 2551–2569.

ZOGRAFOS, K., BALAKRISHNAN, N., (2009). On families of beta- and generalized gamma-generated distributions and associated inference. *Statistical Methodology*, 6, pp. 344–362.

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 41–60, DOI 10.21307/stattrans-2020-054*
Received – 12.10.2019; accepted – 04.05.2020

# Comparing particulate matter dispersion in Thailand using the Bayesian Confidence Intervals for ratio of coefficients of variation

**Warisa Thangjai**[1], **Suparat Niwitpong**[2]

## ABSTRACT

Recently, harmful levels of air pollution have been detected in many provinces of Thailand. Particulate matter (PM) contains microscopic solids or liquid droplets that are so small that they can be inhaled and cause serious health problems. A high dispersion of PM is measured by a coefficient of variation of log-normal distribution. Since the log-normal distribution is often used to analyse environmental data such as hazardous dust particle levels and daily rainfall data. These data focus the statistical inference on the coefficient of variation. In this paper, we develop confidence interval estimation for the ratio of coefficients of variation of two log-normal distributions constructed using the Bayesian approach. These confidence intervals were then compared with the existing approaches: method of variance estimates recovery (MOVER), modified MOVER, and approximate fiducial approaches using their coverage probabilities and average lengths via Monte Carlo simulation. The simulation results show that the Bayesian confidence interval performed better than the others in terms of coverage probability and average length. The proposed approach and the existing approaches are illustrated using examples from data set PM10 level and PM2.5 level in the northern Thailand.

**Key words:** Bayesian approach, coefficient of variation, confidence interval, log-normal distribution, ratio.

## 1. Introduction

Nowadays, the problem of air pollution has received widespread attention in toxicology and epidemiology studies because it is associated with increased incidences of human disease and mortality rate (Xing et al., 2016). The effects on human health include the cardiovascular system, resulting in heart attacks and heart failure, and the respiratory tract, resulting in asthma and bronchitis. Smoke, dust, and smog create air pollution, which includes gaseous pollutants and particulate matter (PM): the gases include carbon monoxide, sulphur dioxide, ozone, and nitrogen dioxide, while PM is defined by size, e.g. PM2.5 ($\leq 2.5\ \mu m$) and PM10 ($\leq 10\ \mu m$), and so on. People are at high risk when they live in high PM levels. For PM2.5, both short-term and long-term exposure has been associated with increased hospital admission and absenteeism from school, work, etc. Exposure to

[1]Department of Statistics, Faculty of Science, Ramkhamhaeng University, Bangkok, 10240, Thailand. E-mail: wthangjai@yahoo.com. ORCID: https://orcid.org/0000-0002-9306-3742.

[2]Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand. E-mail: suparat.n@sci.kmutnb.ac.th. ORCID: https://orcid.org/0000-0003-3059-1131.

PM2.5 can also result in emergency room visits for asthma symptoms whereas exposure to the PM10 can result in hospitalization of chronic lung disease and/or premature death. Moreover, PM2.5 and PM10 can damage stone and culturally important objects such as monuments and statues. Thailand is a country located in Southeast Asia. It covers a total land area of approximately 513,000 $km^2$ and is divided into six regions used in geographic studies: north, northeast, central, east, west, and south. These are based on natural features and human cultural patterns. Recently, Thailand has faced the PM problem resulting in the deterioration of air quality. Harmful levels have been detected in the north region of Thailand, in Chiang Mai, Chiang Rai, Lampang, Mae Hong Son, Nan, Phrae, and Phayao provinces. The coefficient of variation can be used as a statistic to describe air quality and thus can be used to measure and manage air pollution risk.

Meanwhile, several authors have discussed which parameter should be used in statistical inference for a log-normal distribution (Lacey et al., 1997; Royston, 2001; Krishnamoorthy and Mathew, 2003; Hannig et al., 2006; Tian and Wu, 2007; Sharma and Singh, 2010; Harvey and van der Merwe, 2012; Lin and Wang, 2013; Rao and D'Cunha, 2016; Thangjai et al., 2016; Nam and Kwon, 2017; Hasan and Krishnamoorthy, 2017; Thangjai and Niwitpong, 2019). Furthermore, the coefficient of variation has been used in various applications (Tsim et al., 1991; Faupel-Badger et al., 2010). In addition, the confidence intervals for the coefficient of variation have received some attention recently (Niwitpong, 2013; Ng, 2014; Thangjai et al., 2016; Nam and Kwon, 2017; Hasan and Krishnamoorthy, 2017). The inference with the log-normal coefficient of variation is interesting. Nam and Kwon (2017) proposed the method of variance estimate recovery (MOVER) approach for constructing the confidence intervals for the ratio of coefficients of variation of log-normal distributions. Meanwhile, Hasan and Krishnamoorthy (2017) improved the confidence intervals for the ratio of coefficients of variation of log-normal distributions based on an alternative MOVER approach and the fiducial approach.

Both these approaches have produced classical statistics, and while some problems are best solved using these, others are best solved using the Bayesian approach. Therefore, in this paper, we extend the research idea from Hasan and Krishnamoorthy (2017) to develop the Bayesian approach for confidence interval estimation of the ratio of coefficients of variation of log-normal distributions. The Bayesian approach is a statistical method based on Bayes' theorem, which is used to update the probability. The method derives the posterior probability that is the result of a prior probability and a likelihood function. This is advantageous in the interpretation and construction of the Bayesian confidence interval, which makes it more straightforward than the classical confidence interval approaches. However, a disadvantage is that the Bayesian confidence interval requires more input than the classical approach (Casella and Berger, 2002). The Bayesian approach for parameter estimation has been addressed in several research papers (Harvey and van der Merwe, 2012; Rao and D'Cunha, 2016; Ma and Chen, 2018).

## 2. Methods

Suppose that random samples $X_1$ and $X_2$ follow two independent normal distributions with means $\mu_1$ and $\mu_2$ and variances $\sigma_1^2$ and $\sigma_2^2$, respectively. Also, suppose that $Y_1$ and $Y_2$

are random samples of sizes $n_1$ and $n_2$ from two independent log-normal distributions with parameters $\mu_1$, $\sigma_1^2$, $\mu_2$, and $\sigma_1^2$, respectively. The mean and variance of $Y_1$ are

$$E(Y_1) = \exp(\mu_1 + \sigma_1^2/2) \text{ and } Var(Y_1) = (\exp(\sigma_1^2) - 1)(\exp(2\mu_1 + \sigma_1^2)). \quad (1)$$

The coefficient of variation of $Y_1$ is

$$\tau_1 = E(Y_1)/\sqrt{Var(Y_1)} = \sqrt{\exp(\sigma_1^2) - 1}. \quad (2)$$

Similarly, the mean and variance of $Y_2$ are

$$E(Y_2) = \exp(\mu_2 + \sigma_2^2/2) \text{ and } Var(Y_2) = (\exp(\sigma_2^2) - 1)(\exp(2\mu_2 + \sigma_2^2)). \quad (3)$$

The coefficient of variation of $Y_2$ is

$$\tau_2 = E(Y_2)/\sqrt{Var(Y_2)} = \sqrt{\exp(\sigma_2^2) - 1}. \quad (4)$$

The ratio of two coefficients of variation is given by

$$\theta = \frac{\tau_1}{\tau_2} = \sqrt{\frac{\exp(\sigma_1^2) - 1}{\exp(\sigma_2^2) - 1}}. \quad (5)$$

The estimator of $\theta$ is

$$\hat{\theta} = \frac{\hat{\tau}_1}{\hat{\tau}_2} = \sqrt{\frac{\exp(S_1^2) - 1}{\exp(S_2^2) - 1}}, \quad (6)$$

where $S_1^2$ and $S_2^2$ are the variances of the log-transformed sample from a log-normal distributions.

This section describes the three existing confidence intervals. One is the MOVER confidence interval introduced by Nam and Kwon (2017). The modified MOVER and approximate fiducial confidence intervals are proposed by Hasan and Krishnamoorthy (2017). Furthermore, the Bayesian confidence interval, which is a novel approach, is presented.

### 2.1. Classical confidence intervals for ratio of coefficients of variation

Three confidence intervals for the ratio of coefficients of variation of log-normal distributions are presented.

#### 2.1.1 MOVER confidence interval for ratio of coefficients of variation

Donner and Zou (2002) and Zou and Donner (2008) describe a theorem of MOVER. The lower limit $L$ and the upper limit $U$ are used to derive the variance estimates for $\theta$, which is ranging from $L$ to $\hat{\theta}$ and from $\hat{\theta}$ to $U$. The variance estimate recovered from the lower tail of $\theta$ is $(\hat{\theta} - L)^2/z^2$, where $z$ denotes the $100(\alpha/2)$-th percentile of the standard normal distribution. Similarly, the variance estimate recovered from the upper tail of $\theta$ is

$(U-\hat{\theta})^2/z^2$. These variance estimates are used to construct the lower and upper limits of the confidence interval for $\theta$.

Nam and Kwon (2017) introduced the MOVER approach for constructing the confidence interval for the ratio of coefficients of variation of two log-normal distributions. The MOVER confidence interval can be obtained from the one for $\ln(\theta) = \ln(\tau_1) - \ln(\tau_2)$. The variances of $\ln(\hat{\tau}_1)$ and $\ln(\hat{\tau}_2)$ are given by

$$\hat{Var}(\ln(\hat{\tau}_1)) = \frac{\hat{\sigma}_1^2(1+\hat{\tau}_1^2)^2}{2n_1\hat{\tau}_1^4} \tag{7}$$

and

$$\hat{Var}(\ln(\hat{\tau}_2)) = \frac{\hat{\sigma}_2^2(1+\hat{\tau}_2^2)^2}{2n_2\hat{\tau}_2^4}, \tag{8}$$

where $\hat{\sigma}_1^2 = (n_1-1)S_1^2/n_1$ and $\hat{\sigma}_2^2 = (n_2-1)S_2^2/n_2$ are the maximum likelihood estimates of $\sigma_1^2$ and $\sigma_2^2$, respectively.

The confidence intervals of $\ln(\tau_1)$ and $\ln(\tau_2)$ are given by

$$[l_1', u_1'] = [\ln(\hat{\tau}_1) - z_{1-\alpha/2}\sqrt{\hat{Var}(\ln(\hat{\tau}_1))}, \ln(\hat{\tau}_1) + z_{1-\alpha/2}\sqrt{\hat{Var}(\ln(\hat{\tau}_1))}] \tag{9}$$

and

$$[l_2', u_2'] = [\ln(\hat{\tau}_2) - z_{1-\alpha/2}\sqrt{\hat{Var}(\ln(\hat{\tau}_2))}, \ln(\hat{\tau}_2) + z_{1-\alpha/2}\sqrt{\hat{Var}(\ln(\hat{\tau}_2))}], \tag{10}$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$-th percentile of the standard normal distribution and $\hat{Var}(\ln(\hat{\tau}_1))$ and $\hat{Var}(\ln(\hat{\tau}_2))$ are defined in Equation (7) and Equation (8).

The lower and upper limits of the confidence interval for $\ln(\theta) = \ln(\tau_1) - \ln(\tau_2)$ based on the MOVER approach are given by

$$L_{\theta.MOVER} = \ln(\hat{\tau}_1) - \ln(\hat{\tau}_2) - \sqrt{(\ln(\hat{\tau}_1) - l_1')^2 + (\ln(\hat{\tau}_2) - u_2')^2} \tag{11}$$

and

$$U_{\theta.MOVER} = \ln(\hat{\tau}_1) - \ln(\hat{\tau}_2) + \sqrt{(\ln(\hat{\tau}_1) - u_1')^2 + (\ln(\hat{\tau}_2) - l_2')^2}. \tag{12}$$

Therefore, the $100(1-\alpha)\%$ MOVER confidence interval for ratio of coefficients of variation $\theta$ is defined as

$$CI_{\theta.MOVER} = [L_{\theta.MOVER}, U_{\theta.MOVER}] = [\exp(L_{\theta.MOVER}), \exp(U_{\theta.MOVER})]. \tag{13}$$

### 2.1.2 Modified MOVER confidence interval for ratio of coefficients of variation

Hasan and Krishnamoorthy (2017) extended the research paper from Nam and Kwon (2017) to propose the new confidence interval for the ratio of coefficients of variation based on the MOVER approach. The new confidence interval is called modified MOVER confidence interval. Hasan and Krishnamoorthy (2017) used the exact confidence intervals for

$\tau_1^2$ and $\tau_2^2$ given by

$$[l_1'', u_1''] = [\exp\left(\frac{(n_1-1)S_1^2}{\chi^2_{n_1-1,\alpha/2}}\right) - 1, \exp\left(\frac{(n_1-1)S_1^2}{\chi^2_{n_1-1,1-\alpha/2}}\right) - 1] \tag{14}$$

and

$$[l_2'', u_2''] = [\exp\left(\frac{(n_2-1)S_2^2}{\chi^2_{n_2-1,\alpha/2}}\right) - 1, \exp\left(\frac{(n_2-1)S_2^2}{\chi^2_{n_2-1,1-\alpha/2}}\right) - 1], \tag{15}$$

where $\chi^2_{n_i-1,1-\alpha/2}$ and $\chi^2_{n_i-1,\alpha/2}$ denote the $100(1-\alpha/2)$-th and $100(\alpha/2)$-th percentiles of the chi-squared distribution with $n_i - 1$ degrees of freedom for $i = 1, 2$.

The lower and upper limits of the modified MOVER confidence interval for $\ln(\tau_1/\tau_2)^2$ are given by

$$L_{MMOVER} = \ln(\hat{\tau}_1^2) - \ln(\hat{\tau}_2^2) - \sqrt{(\ln(\hat{\tau}_1^2) - \ln(l_1''))^2 + (\ln(\hat{\tau}_2^2) - \ln(u_2''))^2} \tag{16}$$

and

$$U_{MMOVER} = \ln(\hat{\tau}_1^2) - \ln(\hat{\tau}_2^2) + \sqrt{(\ln(\hat{\tau}_1^2) - \ln(u_1''))^2 + (\ln(\hat{\tau}_2^2) - \ln(l_2''))^2}, \tag{17}$$

where $\hat{\tau}_1^2 = \exp(S_1^2) - 1$ and $\hat{\tau}_2^2 = \exp(S_2^2) - 1$.

Therefore, the $100(1-\alpha)\%$ modified MOVER confidence interval for ratio of coefficients of variation $\theta$ is defined as

$$CI_{\theta.MMOVER} = [L_{\theta.MMOVER}, U_{\theta.MMOVER}] = [\sqrt{\exp(L_{MMOVER})}, \sqrt{\exp(U_{MMOVER})}]. \tag{18}$$

### 2.1.3 Approximate fiducial confidence interval for ratio of coefficients of variation

The fiducial confidence interval is computed based on a fiducial quantity. The coefficient of variation of log-normal distribution is used the fiducial quantity for $\sigma^2$ only. This is because the coefficient of variation is the function of $\sigma^2$ only. The percentiles of fiducial generalized pivotal quantity for ratio of coefficients of variation is estimated using simulation. To avoid using the simulation, Hasan and Krishnamoorthy (2017) used modified normal based approximation to construct the approximate fiducial confidence interval. Let $s_1^2$ and $s_2^2$ be observed values of $S_1^2$ and $S_2^2$, respectively.

The lower and upper limits of the approximate fiducial confidence interval for $\ln(\tau_1/\tau_2)^2$ are given by

$$L_{AF} = \ln(T_{1;0.5}) - \ln(T_{2;0.5}) - \sqrt{(\ln(T_{1;0.5}) - \ln(T_{1;\alpha/2}))^2 + (\ln(T_{2;0.5}) - \ln(T_{2;1-\alpha/2}))^2} \tag{19}$$

and

$$U_{AF} = \ln(T_{1;0.5}) - \ln(T_{2;0.5}) + \sqrt{(\ln(T_{1;0.5}) - \ln(T_{1;1-\alpha/2}))^2 + (\ln(T_{2;0.5}) - \ln(T_{2;\alpha/2}))^2}, \tag{20}$$

where $T_{i;p} = \exp((n_i-1)S_i^2/\chi^2_{n_i-1,p}) - 1$ and $\chi^2_{n_i-1,p}$ is the $100(p)$-th percentile of the chi-squared distribution with $n_i - 1$ degrees of freedom, respectively.

Therefore, the $100(1-\alpha)\%$ approximate fiducial confidence interval for the ratio of coefficients of variation $\theta$ is defined as

$$CI_{\theta.AF} = [L_{\theta.AF}, U_{\theta.AF}] = [\sqrt{\exp(L_{AF})}, \sqrt{\exp(U_{AF})}]. \tag{21}$$

### 2.2. Bayesian confidence interval for ratio of coefficients of variation

Bayesian confidence interval is constructed using the concept of Bayesian inference. The Bayesian confidence interval uses a prior distribution. This distribution is based on the experimenter's belief and is updated with the sample information. The Bayesian confidence interval derives a posterior probability as a consequence of a prior probability and a likelihood function. Posterior probability is computed by Bayes' theorem. Let $X_1 = \ln(Y_1)$ be the normal distribution with mean $\mu_1$ and variance $\sigma_1^2$. Also, let $X_2 = \ln(Y_2)$ be the normal distribution with mean $\mu_2$ and variance $\sigma_2^2$.The likelihood function for $\mu_1$, $\mu_2$, $\sigma_1^2$ and $\sigma_2^2$ is

$$\begin{aligned} L(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2|data) &\propto \left(\frac{1}{\sigma_1^2}\right)^{n_1/2} \exp\left(-\frac{(n_1-1)s_1^2+n_1(\mu_1-\bar{x}_1)^2}{2\sigma_1^2}\right) \\ &\times \left(\frac{1}{\sigma_2^2}\right)^{n_2/2} \exp\left(-\frac{(n_2-1)s_2^2+n_2(\mu_2-\bar{x}_2)^2}{2\sigma_2^2}\right), \end{aligned} \tag{22}$$

where $i = 1,2$ and $\bar{x}_i$ and $s_i^2$ are the observed values of $\bar{X}_i$ and $S_i^2$, respectively.

Taking the logarithm of the likelihood function, the log-likelihood function is obtained by

$$\begin{aligned} \ln(L) &= -\frac{n_1}{2}\ln(\sigma_1^2) - \frac{(n_1-1)s_1^2+n_1(\mu_1-\bar{x}_1)^2}{2\sigma_1^2} \\ &- \frac{n_2}{2}\ln(\sigma_2^2) - \frac{(n_2-1)s_2^2+n_2(\mu_2-\bar{x}_2)^2}{2\sigma_2^2}. \end{aligned} \tag{23}$$

The second derivatives of log-likelihood function with respect to each parameter are

$$\frac{\partial^2\ln(L)}{\partial\mu_1^2} = -\frac{n_1}{\sigma_1^2} \quad \text{and} \quad \frac{\partial^2\ln(L)}{\partial\mu_2^2} = -\frac{n_2}{\sigma_2^2}, \tag{24}$$

$$\frac{\partial^2\ln(L)}{\partial\mu_1\partial\sigma_1^2} = \frac{n_1(\mu_1-\bar{x}_1)}{(\sigma_1^2)^2} \quad \text{and} \quad \frac{\partial^2\ln(L)}{\partial\mu_2\partial\sigma_2^2} = \frac{n_2(\mu_2-\bar{x}_2)}{(\sigma_2^2)^2}, \tag{25}$$

$$\frac{\partial^2\ln(L)}{(\partial\sigma_1^2)^2} = \frac{n_1}{2}\left(\frac{1}{\sigma_1^2}\right)^2 - \left(\frac{1}{\sigma_1^2}\right)^3((n_1-1)s_1^2+n_1(\mu_1-\bar{x}_1)^2), \tag{26}$$

$$\frac{\partial^2\ln(L)}{(\partial\sigma_2^2)^2} = \frac{n_2}{2}\left(\frac{1}{\sigma_2^2}\right)^2 - \left(\frac{1}{\sigma_2^2}\right)^3((n_2-1)s_2^2+n_2(\mu_2-\bar{x}_2)^2), \tag{27}$$

and

$$\frac{\partial^2\ln(L)}{\partial\sigma_1^2\partial\sigma_2^2} = 0 \quad \text{and} \quad \frac{\partial^2\ln(L)}{\partial\sigma_2^2\partial\sigma_1^2} = 0. \tag{28}$$

The Fisher information matrix is

$$F(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)=\begin{bmatrix} \frac{n_1}{\sigma_1^2} & 0 & 0 & 0 \\ 0 & \frac{n_2}{\sigma_2^2} & 0 & 0 \\ 0 & 0 & \frac{n_1}{2}\left(\frac{1}{\sigma_1^2}\right)^2 & 0 \\ 0 & 0 & 0 & \frac{n_2}{2}\left(\frac{1}{\sigma_2^2}\right)^2 \end{bmatrix}. \tag{29}$$

The Bayesian confidence intervals can be construct based on different choices of prior distributions. This paper is interested in the Jeffreys Independence prior. This prior follows from the Fisher information matrix. According the Fisher information matrix, the Jeffreys Independence prior is

$$p(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)=p(\mu_1,\mu_2)p(\sigma_1^2,\sigma_2^2). \tag{30}$$

The joint prior for the mean is

$$p(\mu_1,\mu_2)\propto\begin{vmatrix} \frac{n_1}{\sigma_1^2} & 0 \\ 0 & \frac{n_2}{\sigma_2^2} \end{vmatrix}^{1/2}. \tag{31}$$

The joint prior for the variance is

$$p(\sigma_1^2,\sigma_2^2)\propto\begin{vmatrix} \frac{n_1}{2}\left(\frac{1}{\sigma_1^2}\right)^2 & 0 \\ 0 & \frac{n_2}{2}\left(\frac{1}{\sigma_2^2}\right)^2 \end{vmatrix}^{1/2}. \tag{32}$$

Therefore, the Jeffreys Independence prior is obtained by

$$p(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)\propto\frac{1}{\sigma_1^2}\left(\frac{1}{\sigma_2^2}\right). \tag{33}$$

The conditional posterior distributions of $\mu_1$ and $\mu_2$ are normal distributions. The conditional posterior distributions are given by

$$\mu_1|\sigma_1^2,x_1\sim N\left(\hat{\mu}_1,\frac{\sigma_1^2}{n_1}\right) \tag{34}$$

and

$$\mu_2|\sigma_2^2,x_2\sim N\left(\hat{\mu}_2,\frac{\sigma_2^2}{n_2}\right). \tag{35}$$

For $\sigma_1^2$ and $\sigma_2^2$, the posterior distributions are the inverse gamma distributions given by

$$\sigma_1^2|x_1\sim IG(\frac{n_1-1}{2},\frac{(n_1-1)s_1^2}{2}) \tag{36}$$

and

$$\sigma_2^2|x_2 \sim IG(\frac{n_2-1}{2}, \frac{(n_2-1)s_2^2}{2}). \tag{37}$$

The posterior distribution of $\ln(\tau_1/\tau_2)^2$ is given by

$$\ln(\theta)^2 = \ln\left(\frac{\tau_1}{\tau_2}\right)^2 = \ln(\exp(\sigma_1^2)-1) - \ln(\exp(\sigma_2^2)-1), \tag{38}$$

where $\sigma_1^2$ and $\sigma_2^2$ are defined in Equation (36) and Equation (37), respectively.

Let $L_{BS}$ and $U_{BS}$ be the lower and upper limits of the shortest $100(1-\alpha)\%$ highest posterior density interval of $\ln(\theta)^2$, respectively. Therefore, the $100(1-\alpha)\%$ Bayesian confidence interval for ratio of coefficients of variation $\theta$ is defined as

$$CI_{\theta.BS} = [L_{\theta.BS}, U_{\theta.BS}] = [\sqrt{\exp(L_{BS})}, \sqrt{\exp(U_{BS})}]. \tag{39}$$

**Algorithm 1**

*Step 1:* Generate $\sigma_i^2|x_i \sim IG(\frac{n_i-1}{2}, \frac{(n_i-1)s_i^2}{2})$, where $i = 1,2$.
*Step 2:* Calculate the value of $\ln(\theta)^2$ as given in Equation (38).
*Step 3:* Repeat the step 1 - step 2 for $q$ times.
*Step 4:* Calculate $L_{BS}$ and $U_{BS}$.
*Step 5:* Calculate $L_{\theta.BS}$ and $U_{\theta.BS}$.

**Algorithm 2**

For a given $n_1$, $n_2$, $\mu_1$, $\mu_2$, $\sigma_1$, $\sigma_2$, and $\theta$.
*Step 1:* Generate $x_1$ from $N(\mu_1, \sigma_1^2)$ and generate $x_2$ from $N(\mu_2, \sigma_2^2)$.
*Step 2:* Calculate $\bar{x}_1$, $\bar{x}_2$, $s_1^2$ and $s_2^2$.
*Step 3:* Construct $CI_{\theta.MOVER(h)} = [L_{\theta.MOVER(h)}, U_{\theta.MOVER(h)}]$.
*Step 4:* Construct $CI_{\theta.MMOVER(h)} = [L_{\theta.MMOVER(h)}, U_{\theta.MMOVER(h)}]$.
*Step 5:* Construct $CI_{\theta.AF(h)} = [L_{\theta.AF(h)}, U_{\theta.AF(h)}]$.
*Step 6:* Construct $CI_{\theta.BS(h)} = [L_{\theta.BS(h)}, U_{\theta.BS(h)}]$.
*Step 7:* If $L_{(h)} \leq \theta \leq U_{(h)}$ set $p_{(h)} = 1$, else $p_{(h)} = 0$.
*Step 8:* Calculate $U_{(h)} - L_{(h)}$.
*Step 9:* Repeat the step 1 - step 8 for a large number of times (say, $M$ times) and calculate coverage probability and average length.

## 3. Results

The MOVER, modified MOVER, approximate fiducial and Bayesian confidence intervals for ratio of coefficients of variation were conducted to compare the performance. The confidence intervals with the coverage probability greater than or equal to the nominal

confidence level of 0.95 and the shortest average length were considered to be the best-performing ones.

Since the log-normal coefficient of variation depends on parameter $\sigma^2$ and does not depend on parameter $\mu$, the population means $\mu_1 = \mu_2 = 1$, the population standard deviations $(\sigma_1, \sigma_2)$ and sample sizes $(n_1, n_2)$ were varied based on Hasan and Krishnamoorthy's (2017) approach. The coverage probabilities and average lengths were estimated for some assumed values of parameters $(\sigma_1, \sigma_2)$ and sample sizes varying from small to moderate. 10,000 random samples were generated using Algorithm 2 for each set of parameters. For the Bayesian confidence interval, 2,500$\ln(\theta)^2$'s were obtained by applying Algorithm 1 for each of the random samples.

The coverage probabilities and average lengths of the four confidence intervals are given in Tables 1 and 2. The MOVER confidence intervals attained coverage probabilities under the nominal confidence level of 0.95 for all sample sizes. Meanwhile, the coverage probabilities of the modified MOVER and approximate fiducial confidence intervals were close to the nominal confidence level of 0.95, but their average lengths were not balanced. The Bayesian confidence intervals provided the best coverage probabilities for all sample sizes and the average lengths were shorter than those of the modified MOVER and approximate fiducial confidence intervals. Overall, the Bayesian confidence intervals are preferable in terms of coverage probability and average length.

Table 1: The CP and AL of 95% two-sided confidence intervals for ratio of coefficients of variation of log-normal distributions as function of parameters

| $(n_1, n_2)$ | Approach | $(\sigma_1, \sigma_2)$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **(0.1,0.3)** | | **(0.3,0.7)** | | **(0.4,1.2)** | | **(0.5,1.6)** | |
| | | **CP** | **AL** | **CP** | **AL** | **CP** | **AL** | **CP** | **AL** |
| (4,8) | MOVER | 0.8639 | 0.6160 | 0.8793 | 0.8242 | 0.9031 | 0.6675 | 0.9208 | 0.6745 |
| | MMOVER | **0.9546** | 1.2912 | **0.9553** | 2.8125 | **0.9589** | 4.4569 | **0.9527** | 104.7483 |
| | AF | **0.9501** | 1.2483 | **0.9501** | 2.6929 | **0.9519** | 4.1770 | 0.9487 | 93.2188 |
| | BS | 0.9485 | 1.1105 | **0.9629** | 1.9912 | **0.9768** | 2.3785 | **0.9777** | 5.6386 |
| (5,10) | MOVER | 0.8882 | 0.5420 | 0.9000 | 0.7301 | 0.9227 | 0.5793 | 0.9243 | 0.5805 |
| | MMOVER | **0.9563** | 0.9144 | **0.9567** | 1.4768 | **0.9544** | 1.3068 | **0.9502** | 1.6590 |
| | AF | **0.9532** | 0.8914 | **0.9525** | 1.4319 | **0.9512** | 1.2577 | 0.9470 | 1.5768 |
| | BS | 0.9499 | 0.8222 | **0.9597** | 1.2783 | **0.9693** | 1.1615 | **0.9715** | 1.3628 |
| (10,5) | MOVER | 0.8848 | 0.7279 | 0.8823 | 1.0082 | 0.8455 | 0.8969 | 0.8156 | 1.0019 |
| | MMOVER | **0.9554** | 0.7778 | **0.9513** | 1.0888 | 0.9472 | 0.8510 | 0.9465 | 0.7785 |
| | AF | **0.9511** | 0.7602 | 0.9485 | 1.0625 | 0.9458 | 0.8302 | 0.9457 | 0.7574 |
| | BS | **0.9558** | 0.8067 | **0.9618** | 1.2256 | **0.9621** | 1.0964 | **0.9613** | 1.1222 |
| (7,12) | MOVER | 0.9080 | 0.4707 | 0.9212 | 0.6313 | 0.9326 | 0.5058 | 0.9315 | 0.4974 |
| | MMOVER | **0.9547** | 0.6445 | **0.9590** | 0.9189 | **0.9517** | 0.7179 | **0.9510** | 0.6634 |
| | AF | **0.9519** | 0.6324 | **0.9562** | 0.8988 | 0.9489 | 0.7012 | 0.9493 | 0.6467 |
| | BS | 0.9494 | 0.6055 | **0.9597** | 0.8679 | **0.9663** | 0.7264 | **0.9699** | 0.7154 |
| (10,10) | MOVER | 0.9226 | 0.4725 | 0.9276 | 0.6410 | 0.9234 | 0.5410 | 0.9030 | 0.5532 |
| | MMOVER | **0.9561** | 0.5521 | **0.9516** | 0.7651 | 0.9465 | 0.6012 | 0.9474 | 0.5374 |
| | AF | **0.9532** | 0.5427 | 0.9485 | 0.7510 | 0.9452 | 0.5911 | 0.9471 | 0.5289 |
| | BS | **0.9510** | 0.5397 | **0.9586** | 0.7695 | **0.9626** | 0.6576 | **0.9614** | 0.6385 |

Table 1: Continued

| $(n_1, n_2)$ | **Approach** | $(\sigma_1, \sigma_2)$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **(0.1,0.3)** | | **(0.3,0.7)** | | **(0.4,1.2)** | | **(0.5,1.6)** | |
| | | **CP** | **AL** | **CP** | **AL** | **CP** | **AL** | **CP** | **AL** |
| (10,15) | MOVER | 0.9222 | 0.4034 | 0.9285 | 0.5368 | 0.9332 | 0.4347 | 0.9308 | 0.4230 |
| | MMOVER | **0.9527** | 0.4884 | **0.9504** | 0.6664 | **0.9506** | 0.5162 | 0.9470 | 0.4560 |
| | AF | 0.9494 | 0.4817 | 0.9488 | 0.6561 | 0.9486 | 0.5088 | 0.9458 | 0.4503 |
| | BS | 0.9464 | 0.4696 | **0.9532** | 0.6488 | **0.9610** | 0.5320 | **0.9598** | 0.5002 |
| (20,20) | MOVER | 0.9358 | 0.3105 | 0.9438 | 0.4202 | 0.9394 | 0.3451 | 0.9252 | 0.3364 |
| | MMOVER | **0.9517** | 0.3327 | **0.9551** | 0.4526 | **0.9536** | 0.3558 | 0.9494 | 0.3147 |
| | AF | 0.9496 | 0.3301 | **0.9534** | 0.4489 | **0.9531** | 0.3537 | 0.9493 | 0.3138 |
| | BS | 0.9489 | 0.3278 | **0.9561** | 0.4511 | **0.9571** | 0.3694 | **0.9594** | 0.3432 |

Table 2: The CP AL of 95% two-sided confidence intervals for ratio of coefficients of variation of log-normal distributions as function of sample sizes

| $(\sigma_1,\sigma_2)$ | Approach | $(n_1,n_2)$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | (5,5) | | (10,10) | | (10,20) | | (30,40) | |
| | | CP | AL | CP | AL | CP | AL | CP | AL |
| (0.1,0.7) | MOVER | 0.8914 | 0.3437 | 0.9202 | 0.2062 | 0.9257 | 0.1552 | 0.9443 | 0.0980 |
| | MMOVER | **0.9526** | 0.4948 | 0.9466 | 0.2367 | **0.9508** | 0.1885 | **0.9504** | 0.1027 |
| | AF | 0.9476 | 0.4740 | 0.9442 | 0.2329 | 0.9491 | 0.1863 | 0.9491 | 0.1022 |
| | BS | **0.9638** | 0.5392 | 0.9496 | 0.2406 | 0.9487 | 0.1832 | 0.9484 | 0.1018 |
| (0.3,0.9) | MOVER | 0.8938 | 0.8596 | 0.9275 | 0.5055 | 0.9283 | 0.3755 | 0.9470 | 0.2370 |
| | MMOVER | **0.9513** | 1.4219 | **0.9541** | 0.5868 | 0.9490 | 0.4683 | **0.9510** | 0.2487 |
| | AF | 0.9466 | 1.3475 | **0.9517** | 0.5767 | 0.9458 | 0.4622 | **0.9501** | 0.2476 |
| | BS | **0.9666** | 1.5980 | **0.9609** | 0.6099 | **0.9513** | 0.4566 | **0.9520** | 0.2478 |
| (0.5,1.4) | MOVER | 0.8833 | 1.0967 | 0.9156 | 0.6134 | 0.9451 | 0.4188 | 0.9467 | 0.2655 |
| | MMOVER | **0.9504** | 2.6600 | 0.9461 | 0.6597 | **0.9571** | 0.5162 | **0.9550** | 0.2699 |
| | AF | 0.9456 | 2.4364 | 0.9445 | 0.6478 | **0.9544** | 0.5093 | **0.9546** | 0.2691 |
| | BS | **0.9724** | 3.2572 | **0.9633** | 0.7464 | **0.9649** | 0.5235 | **0.9575** | 0.2757 |
| (0.3,1.5) | MOVER | 0.8650 | 0.5899 | 0.9014 | 0.3268 | 0.9422 | 0.2166 | 0.9425 | 0.1380 |
| | MMOVER | 0.9473 | 0.7247 | 0.9499 | 0.3076 | **0.9526** | 0.2359 | 0.9486 | 0.1356 |
| | AF | 0.9441 | 0.6858 | 0.9490 | 0.3043 | **0.9513** | 0.2340 | 0.9488 | 0.1355 |
| | BS | **0.9678** | 1.0401 | **0.9596** | 0.3623 | **0.9624** | 0.2500 | **0.9537** | 0.1401 |
| (0.9,1.8) | MOVER | 0.8597 | 2.4916 | 0.8984 | 1.1677 | 0.9478 | 0.7360 | 0.9428 | 0.4259 |
| | MMOVER | **0.9556** | 48337.6800 | **0.9511** | 1.6359 | **0.9533** | 1.3112 | 0.9494 | 0.4250 |
| | AF | **0.9512** | 37293.2300 | **0.9500** | 1.5698 | **0.9513** | 1.2785 | 0.9489 | 0.4235 |
| | BS | **0.9805** | 5024.7520 | **0.9671** | 1.8142 | **0.9685** | 1.1855 | **0.9557** | 0.4429 |
| (0.4,0.5) | MOVER | 0.8929 | 2.0732 | 0.9218 | 1.2401 | 0.9195 | 0.9674 | 0.9427 | 0.5952 |
| | MMOVER | **0.9575** | 4.7748 | **0.9510** | 1.5650 | 0.9481 | 1.3012 | **0.9501** | 0.6378 |
| | AF | **0.9524** | 4.5000 | 0.9485 | 1.5325 | 0.9462 | 1.2837 | 0.9494 | 0.6344 |
| | BS | **0.9669** | 4.0819 | **0.9527** | 1.5125 | 0.9489 | 1.2134 | 0.9483 | 0.6259 |

## 4. Empirical application

PM10 and PM2.5 from haze smog in Chiang Mai and Nan provinces, in the northern of Thailand, have become serious problems with air pollution having serious effects on health and visibility for transportation. The data in Tables 3 and 6 from the Pollution Control Department show PM10 and PM2.5 levels in Chiang Mai and Nan provinces from 24 March 2019 to 17 April 2019. Moreover, PM2.5 levels in Bangkok and Chiang Rai provinces from 24 March 2019 to 17 April 2019 are presented in Table 9. The confidence intervals for the ratio of coefficients of variation were constructed using these data.

### 4.1. Example 1

Using Table 3, the statistics of PM10 pollution are summarized in Table 4. In Table 5, the Akaike Information Criterion values support that the two datasets follow log-normal distributions. These two districts were compared with respect to the coefficient of variation. The 95% two-sided confidence intervals were constructed based on the MOVER, modified MOVER, and approximate fiducial approaches, and then compared with the Bayesian approach.

The ratio of the log-normal coefficients of variation for the Chiang Mai and Nan was $\hat{\theta} = 0.9066$. The confidence intervals based on the MOVER, modified MOVER, and approximate fiducial approaches were $CI_{\theta.MOVER} = [0.6009, 1.3676]$ with an interval length of 0.7667, $CI_{\theta.MMOVER} = [0.5829, 1.3977]$ with an interval length of 0.8148, and $CI_{\theta.AF} = [0.5846, 1.3940]$ with an interval length of 0.8094. Meanwhile, the confidence interval based on the Bayesian approach was $CI_{\theta.BS} = [0.5972, 1.3604]$ with an interval length of 0.7632. These results indicate that all of the confidence intervals contained the true ratio of the coefficients of variation. However, the Bayesian confidence interval provided the shortest length.

### 4.2. Example 2

To assess the PM2.5 level in Chiang Mai and Nan provinces, we used the data in Table 6 for the second analysis and summarized the statistics in Table 7. Using the Akaike Information Criterion values in Table 8, we found that the two PM2.5 samples came from log-normal populations.

The ratio of log-normal coefficients of variation for the Chiang Mai and Nan was $\hat{\theta} = 0.9654$. The confidence intervals for the ratio based on MOVER, modified MOVER, and approximate fiducial approaches were $CI_{\theta.MOVER} = [0.6355, 1.4667]$, $CI_{\theta.MMOVER} = [0.6153, 1.5031]$, and $CI_{\theta.AF} = [0.6171, 1.4988]$ with interval lengths of 0.8312, 0.8878, and 0.8817, respectively. Meanwhile, the confidence interval for the Bayesian approach was $CI_{\theta.BS} = [0.6274, 1.4457]$ with an interval length of 0.8183. The interval length of the Bayesian approach was shorter than the others, thus it more accurately estimated the coefficient of variation ratio for these two log-normal populations.

### 4.3. Example 3

The PM2.5 levels of Bangkok and Chiang Rai provinces in Table 9 were used to construct the confidence intervals for the ratio of coefficients of variation for comparing the dispersion of PM 2.5 with different levels. The statistics and the Akaike Information Criterion values were presented in Table 10 and Table 11, respectively. The result showed that the PM2.5 levels samples came from log-normal distributions.

The ratio of coefficients of variation of log-normal distributions for the Bangkok and Chiang Rai was $\hat{\theta} = 0.5300$. The confidence intervals for the ratio based on MOVER, modified MOVER, and approximate fiducial approaches were$CI_{\theta.MOVER} = [0.3519,0.7984]$, $CI_{\theta.MMOVER} = [0.3401,0.8130]$, and $CI_{\theta.AF} = [0.3411,0.8111]$ with interval lengths of 0.4465, 0.4729, and 0.4700, respectively. Moreover, the confidence interval for the Bayesian approach was $CI_{\theta.BS} = [0.3458,0.7901]$ with an interval length of 0.4443. The Bayesian confidence interval had the shortest interval length.

## 5. Discussion

Nam and Kwon (2017) proposed the MOVER approach for constructing the confidence intervals for the ratio of coefficients of variation of two log-normal distributions, while Hasan and Krishnamoorthy (2017) constructed them based on modified MOVER and approximate fiducial approaches and compared them with the MOVER approach. In this paper, we propose the Bayesian approach for the confidence interval estimation of the ratio of coefficients of variation of log-normal distributions.

## 6. Conclusions

Using the data examples from data set PM10 level and PM2.5 level in the northern Thailand, all approaches were illustrated with real data analysis. The performance of the Bayesian approach was compared to three existing approaches. The performances of the confidence intervals agreed with our simulation studies. Since the coverage probability of the Bayesian confidence interval was better than those of the others, and its average length was also shorter. Therefore, the Bayesian approach is recommended to construct the confidence intervals for the ratio of coefficients of variation of log-normal distributions when the dispersions of PM10 level and PM2.5 level are at the harmful level ($\geq 50\mu g/m^3$).

## Acknowledgments

This research was funded by King Mongkut's University of Technology North Bangkok. Grant No. KMUTNB-63-DRIVE-13.

## References

CÄSELLA, G., BERGER, R. L., (2002). *Statistical Inference*, California:Duxbury.

DONNER, A., ZOU, G. Y., (2002). Interval estimation for a difference between intraclass kappa statistics. *Biometrics*, 58, pp. 209–215.

ZOU, G.Y., DONNER, A., (2008). Construction of confidence limits about effect measures: a general approach. *Statistics in Medicine*, 27, pp. 1693–1702.

FAUPEL-BADGER, J.M., FUHRMAN, B.J., XU, X., FALK, R.T., KEEFER, L.K., VEENSTRA, T.D., HOOVER, R.N., ZIEGLER, R.G., (2010). Comparison of liquid chromatographytandem mass spectrometry, RIA, and ELISA methods for measurement of urinary estrogens. *Cancer Epidemiology Biomarkers & Prevention*, 19, pp. 292–300.

HANNIG, J., LIDONG, E., ABDEL-KARIM, A., IYER, H., (2006). Simultaneous fiducial generalized confidence intervals for ratios of means of lognormal distributions. *Austrian Journal of Statistics*, 35, pp. 261–269.

HARVEY, J., VAN DER MERWE, A. J., (2012). Bayesian confidence intervals for means and variances of lognormal and bivariate lognormal distributions. *Journal of Statistical Planning and Inference*, 142, pp. 1294–1309.

HASAN, M. S., KRISHNAMOORTHY, K., (2017). Improved confidence intervals for the ratio of coefficients of variation of two lognormal distributions. *Journal of Statistical Theory and Applications*, 16, pp. 345–353.

KRISHNAMOORTHY, K., (2016). Modified normal-based approximation to the percentiles of linear combination of independent random variables with applications. *Communications in Statistics - Simulation and Computation*, 45, pp. 2428–2444.

KRISHNAMOORTHY, K., MATHEW, T., (2003). Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. *Journal of Statistical Planning and Inference*, 115, pp. 103–121.

LACEY, L.F., KEENE, O. N., PRITCHARD, J. F., BYE, A., (1997). Common noncompartmental pharmacokinetic variables: are they normally or log-normally distributed? *Journal of Biopharmaceutical Statistics*, 7, pp. 171–178.

LIN, S.H., WANG, R. S., (2013). Modified method on the means for several log-normal distributions. *Journal of Applied Statistics*, 40, pp. 194–208.

MA, Z., CHEN, G., (2018). Bayesian methods for dealing with missing data problems. *Journal of the Korean Statistical Society*, 47, pp. 297-–313.

NAM, J.M., KWON, D., (2017). Inference on the ratio of two coefficients of variation of two lognormal distributions. *Communications in Statistics - Theory and Methods*, 46, pp. 8575–8587.

NIWITPONG, S.-A., (2013). Confidence intervals for coefficient of variation of lognormal distribution with restricted parameter space. *Applied Mathematical Sciences*, 7, pp. 3805–3810.

NG, C.K., (2014). Inference on the common coefficient of variation when populations are lognormal: A simulation-based approach. *Journal of Statistics: Advances in Theory and Applications*, 11, pp. 117–134.

RAO, K. A., D'CUNHA, J.G., (2016). Bayesian inference for median of the lognormal distribution. *Journal of Modern Applied Statistical Methods*, 15, pp. 526–535.

ROYSTON, P., (2001). The Lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Statistica Neerlandica*, 55, pp. 89–104.

SHARMA, M.A., SINGH, J.B., (2010). Use of Probability Distribution in Rainfall Analysis. *New York Science Journal*, 3, pp. 40–49.

TIAN, L., WU, J., (2007). Inferences on the common mean of several log-normal populations: The generalized variable approach. *Biometrical Journal*, 49, pp. 944–951.

THANGJAI, W., NIWITPONG, S.-A., (2019). Confidence intervals for the signal-to-noise ratio and difference of signal-to-noise ratios of log-normal distributions. *Stats*, 2, pp. 164–173.

THANGJAI, W., NIWITPONG, S.-A., NIWITPONG, S., (2016). Simultaneous fiducial generalized confidence intervals for all differences of coefficients of variation of log-normal distributions. *Lecture Notes in Artificial Intelligence*, 9978, pp. 552–561.

TSIM, Y. L., YIP, S. P., TSANG, K. S., LI, K. F., WONG, H. F., (1991). Haematogoy and Serology. In *Annual Report, Honk Kong Medical Technology Association Quality Assurance Programme*, pp. 25–40.

XING, Y. F., XU, Y. H., SHI, M. H., LIAN, Y. X., (2016). The impact of PM2.5 on the human respiratory system. *Journal of Thoracic Disease*, 8, E69–E74.

## APPENDIX

Table 3: PM10 levels in Chiang Mai province and Nan province ($\mu g/m^3$)

| **Chiang Mai** | | | | | **Nan** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 227 | 170 | 164 | 105 | 128 | 224 | 134 | 138 | 148 | 190 |
| 156 | 262 | 167 | 112 | 103 | 145 | 232 | 136 | 144 | 127 |
| 138 | 146 | 166 | 123 | 94 | 114 | 199 | 100 | 155 | 116 |
| 125 | 191 | 142 | 139 | 96 | 107 | 176 | 90 | 178 | 126 |
| 113 | 184 | 117 | 138 | 98 | 80 | 130 | 126 | 254 | |

Source: Pollution Control Department (http://aqmthai.com/aqi.php)

Table 4: Statistics of PM10 levels in Chiang Mai province and Nan province

| **Statistics** | **Chiang Mai** | **Nan** |
|---|---|---|
| $n$ | 25 | 24 |
| $\bar{y}$ | 144.1600 | 148.7083 |
| $s_Y$ | 41.2580 | 44.9662 |
| $\bar{x}$ | 4.9355 | 4.9603 |
| $s_X$ | 0.2665 | 0.2931 |
| $\hat{\tau}$ | 0.2656 | 0.2930 |

Table 5: The minimum Akaike Information Criterion values of PM10 level in Chiang Mai province and Nan province

| **Distribution** | **Chiang Mai** | **Nan** |
|---|---|---|
| Normal | 259.9186 | 253.7713 |
| Log-Normal | **254.5765** | **250.2824** |
| Gamma | 255.8663 | 250.9095 |
| Exponential | 299.5462 | 289.0954 |

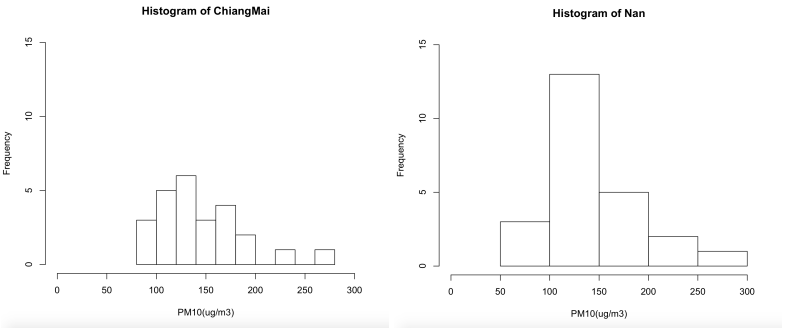



Figure 1: Histogram plots of PM10 level in Chiang Mai province and Nan province

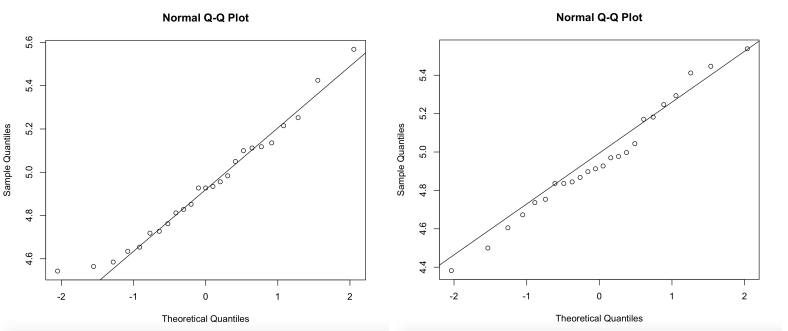



Figure 2: The normal QQ-plots of log-PM10 level in Chiang Mai province and Nan province

Table 6: PM2.5 levels in Chiang Mai province and Nan province ($\mu g/m^3$)

| **Chiang Mai** | | | | | **Nan** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 189 | 129 | 124 | 69 | 92 | 192 | 104 | 111 | 115 | 154 |
| 118 | 213 | 126 | 72 | 68 | 118 | 199 | 107 | 108 | 100 |
| 100 | 109 | 125 | 83 | 64 | 88 | 167 | 73 | 119 | 90 |
| 92 | 147 | 105 | 99 | 66 | 86 | 146 | 61 | 136 | 89 |
| 82 | 145 | 79 | 102 | 62 | 55 | 105 | 96 | 209 | |

Source: Pollution Control Department (http://aqmthai.com/aqi.php)

Table 7: Statistics of PM2.5 levels in Chiang Mai province and Nan province

| **Statistics** | **Chiang Mai** | **Nan** |
|---|---|---|
| $n$ | 25 | 24 |
| $\bar{y}$ | 106.4000 | 117.8333 |
| $s_Y$ | 38.1335 | 41.3718 |
| $\bar{x}$ | 4.6120 | 4.7125 |
| $s_X$ | 0.3324 | 0.3440 |
| $\hat{\tau}$ | 0.3346 | 0.3465 |

Table 8: The minimum Akaike Information Criterion values of PM2.5 level in Chiang Mai province and Nan province

| Distribution | Chiang Mai | Nan |
|---|---|---|
| Normal | 255.9811 | 249.7724 |
| Log-Normal | **249.4643** | **246.0677** |
| Gamma | 250.9027 | 246.5411 |
| Exponential | 284.3603 | 277.9250 |

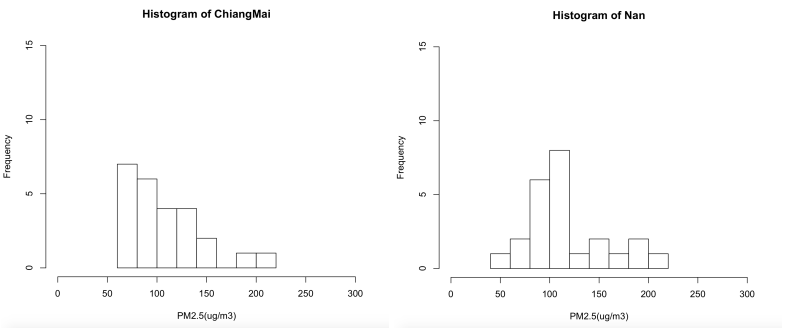



Figure 3: Histogram plots of PM2.5 level in Chiang Mai province and Nan province

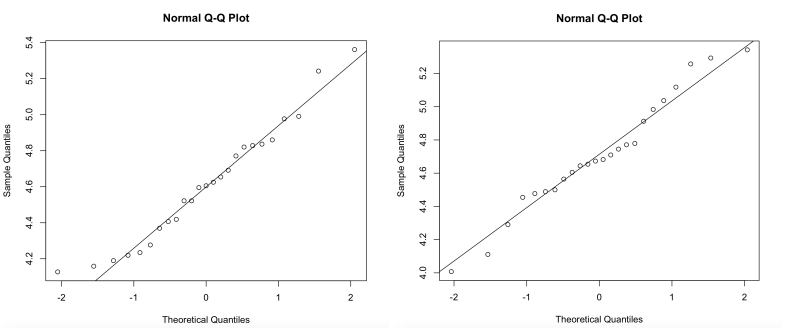



Figure 4: The normal QQ-plots of log-PM2.5 level in Chiang Mai province and Nan province

Table 9: PM2.5 levels in Bangkok province and Chiang Rai province ($\mu g/m^3$)

| Bangkok | | | | | Chiang Rai | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 30 | 19 | 18 | 25 | 19 | 184 | 89 | 109 | 63 | 104 |
| 22 | 19 | 21 | 15 | 14 | 147 | 228 | 77 | 72 | 85 |
| 22 | 23 | 15 | 16 | 14 | 79 | 254 | 77 | 79 | 74 |
| 20 | 19 | 22 | 16 | 15 | 77 | 140 | 83 | 82 | 113 |
| 20 | 23 | 17 | 18 | 13 | 86 | 132 | 82 | 104 | 162 |

Source: Pollution Control Department (http://aqmthai.com/aqi.php)

Table 10: Statistics of PM2.5 levels in Bangkok province and Chiang Rai province

| Statistics | Bangkok | Chiang Rai |
|---|---|---|
| $n$ | 25 | 25 |
| $\bar{y}$ | 19.0000 | 111.2800 |
| $s_Y$ | 3.9791 | 49.8795 |
| $\bar{x}$ | 2.9242 | 4.6361 |
| $s_X$ | 0.2043 | 0.3762 |
| $\hat{\tau}$ | 0.2022 | 0.3815 |

Table 11: The minimum Akaike Information Criterion values of PM2.5 level in Bangkok province and Chiang Rai province

| Distribution | Bangkok | Chiang Rai |
|---|---|---|
| Normal | 142.9793 | 269.4068 |
| Log-Normal | **140.7307** | **256.8566** |
| Gamma | 141.1844 | 260.2494 |
| Exponential | 198.2219 | 286.6025 |

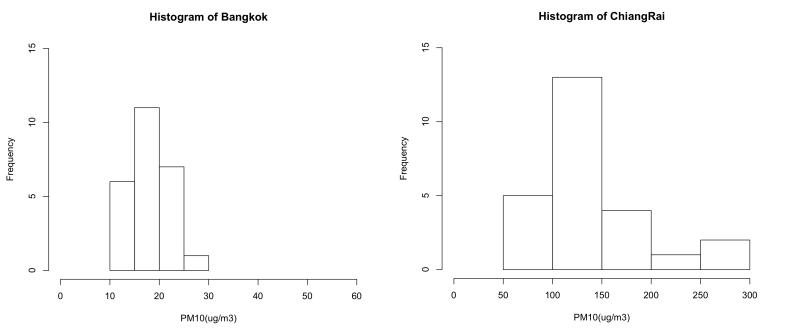



Figure 5: Histogram plots of PM2.5 level in Bangkok province and Chiang Rai province

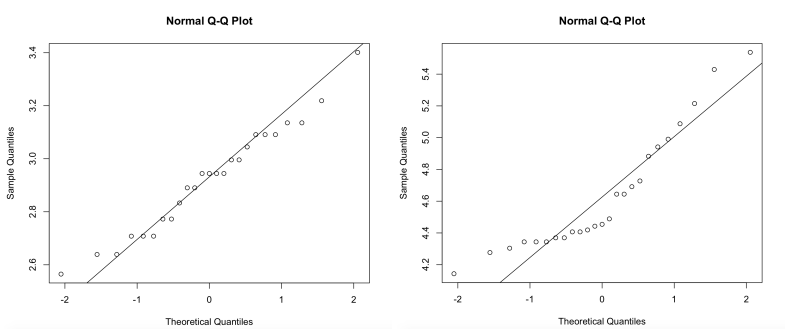



Figure 6: The normal QQ-plots of log-PM2.5 level in Bangkok province and Chiang Rai province

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 61–84, DOI 10.21307/stattrans-2020-055*
Received – 26.03.2020; accepted – 11.05.2020

# A new generalization of the Pareto distribution and its applications

**Ehab M. Almetwally[1], Hanan A. Haj Ahmad[2]**

## ABSTRACT

This paper introduces a new generalization of the Pareto distribution using the Marshall-Olkin generator and the method of alpha power transformation. This new model has several desirable properties appropriate for modelling right skewed data. The Authors demonstrate how the hazard rate function and moments are obtained. Moreover, an estimation for the new model parameters is provided, through the application of the maximum likelihood and maximum product spacings methods, as well as the Bayesian estimation. Approximate confidence intervals are obtained by means of an asymptotic property of the maximum likelihood and maximum product spacings methods, while the Bayes credible intervals are found by using the Monte Carlo Markov Chain method under different loss functions. A simulation analysis is conducted to compare the estimation methods. Finally, the application of the proposed new distribution to three real-data examples is presented and its goodness-of-fit is demonstrated. In addition, comparisons to other models are made in order to prove the efficiency of the distribution in question.

**Key words:** Marshall-Olkin distribution, alpha power transformation, maximum likelihood estimator, maximum product spacings, Bayes estimation, simulation.

## 1. Introduction

Marshall-Olkin (MO) is a well-known distribution, which was generated by Marshall and Olkin (1997). The basic idea in this generator is to add a parameter through which the new distribution will be more flexible and will have many good properties. Many authors used MO to generate new lifetime models, for example Jose and Alice (2001, 2005), Ghitany et al. (2005), Ghitany and Kotz (2007), Jose and Uma (2009), Haj Ahmad et al. (2017), Bdair and Haj Ahmad (2019) and Ahmad and Almetwally (2020). The method of alpha power transformation (APT) class is

[1] Faculty of Business Administration, Delta University of Science and Technology, Egypt. E-mail: ehabxp_2009@hotmail.com.

[2] Department of Basic Science, Preparatory Year Deanship, King Faisal University, Hofuf, Al-Ahsa, 31982, Saudi Arabia, E-mail: hhajahmed@kfu.edu.sa, hananahm1@yahoo.com.

a procedure which makes the lifetime distribution more applicable and rich towards real data analysis. It was first introduced by Mahdavi and Kundu (2017). A new generalization appeared in the literature by doing combination between MO and APT, this was first studied by Nassar et al. (2019), and the new family is called "G-family (MOAP-G). It was noticed that the MOAP-G family is analytically tractable and efficient for real data analysis.

The cumulative distribution function (CDF) of MOAP-G random variable *X* is of the form

$$F_{MOAP}(x;\alpha,\theta)=\begin{cases}\frac{\alpha^{G(x)}-1}{(\alpha-1)[\theta+(1-\theta)(\alpha-1)^{-1}(\alpha^{G(x)}-1)]} & ,\alpha>0,\alpha\neq 1\\ G(x) & ,\alpha=1\end{cases} \quad (1)$$

The corresponding probability density function (pdf)

$$f_{MOAP}(x;\alpha,\theta)=\begin{cases}\frac{\theta\log(\alpha)\alpha^{G(x)}g(x)}{(\alpha-1)[\theta+\frac{1-\theta}{\alpha-1}(\alpha^{G(x)}-1)]^2} & ,\alpha>0,\alpha\neq 1\\ g(x) & ,\alpha=1\end{cases} \quad (2)$$

where *G(x)* is the baseline distribution.

In this paper we will consider Pareto distribution with shape parameter $\lambda$ as a baseline distribution, where the pdf and cdf are respectively as follows:

$$g(x)=\frac{\lambda}{x^{\lambda+1}},x\geq 1 \quad (3)$$

$$G(x)=1-\frac{1}{x^{\lambda}},\ \ x\geq 1 \quad (4)$$

The new generated distribution, namely Marshall-Olkin Alpha Power Pareto (MOAPP), is a lifetime model with three parameters. This distribution has several desirable properties and acts well for modelling right skewed data, it has upside-down bathtub hazard rate and attractive time series representation by which many statistical computations can be easily handled. Real data examples show that MOAPP behaves better than many other generalized Pareto distributions.

The main purpose of this paper is to introduce MOAPP distribution and study some of its statistical properties, which are useful in data modelling. We use statistical inference such as maximum likelihood, maximum product spacings and Bayes estimation methods to perform point estimation. We construct confidence intervals for the unknown parameter as well. A simulation study is conducted to check the performance of the different estimation methods applied in this work This is done by comparing the bias and the mean square error (MSE) for point estimation methods and by using interval length for interval estimation. Finally, we present numerical examples that illustrate the model efficiency.

The rest of this paper is organized as follows: In Section 2 we introduce MOAPP distribution with some of its properties. Classical point estimation methods for the unknown parameters are discussed in Section 3, while in Section 4 the Bayesian

estimation method is considered. In Section 5 interval estimation methods are presented. In Section 6 a simulation study and real-life data analysis are conducted and finally conclusions are given in Section 7.

## 2. Probability Density Function

Let $X$ be a continuous random variable with Marshall-Olkin Alpha Power Pareto distribution (MOAPP), then using Eqs. (1) and (2) and assuming that the baseline distribution $G(x)$ is Pareto distribution given in Eqs. (3) and (4), we obtain the pdf and CDF of (MOAPP) respectively as

$$f_{MOAPP}(x;\alpha,\theta,\lambda) = \frac{\theta\lambda(log\alpha)\alpha^{1-x^{-\lambda}}}{(\alpha-1)x^{\lambda+1}[\theta+(1-\theta)(\alpha-1)^{-1}\left(\alpha^{1-x^{-\lambda}}-1\right)]^2}, \quad x \geq 1, \alpha \neq 1 \tag{5}$$

$$F_{MOAPP}(x;\alpha,\theta,\lambda) = \frac{\alpha^{1-x^{-\lambda}}-1}{(\alpha-1)[\theta+(1-\theta)(\alpha-1)^{-1}\left(\alpha^{1-x^{-\lambda}}-1\right)]}, \; x \geq 1, \alpha \neq 1, \tag{6}$$

In the following subsection we investigate some important properties of MOAPP distribution such as: monotonicity, hazard rate function, series representation, moments and quantiles.

### 2.1. Monotonicity of MOAPP Distribution

The monotonicity of MOAPP distribution is necessary to be investigated for data modelling, many areas such as medical, industrial, engineering and reliability researches need data modelling for prediction of future values and estimation of some unknown or missing variables; hence, in this section we study the monotonicity of MOAPP distribution. We consider the pdf of MOAPP distribution in Eq. (5), and study the monotonicity of this pdf by using the logarithmic function of its pdf. The following lemma illustrates the behaviour of MOAPP distribution for different parameter values, and Figure (1) shows these cases.

***Lemma 1***

The pdf of MOAPP distribution is either decreasing when $0 < \theta < 1$, or upside-down bathtub curve that attains its maximum at some point $x_0 \in [1, \infty)$ when $\theta > 1$ and $\lambda > 1$

***Proof***

Consider the pdf of MOAPP density in Eq. (5), then the derivative of the logarithmic function of pdf with respect to x is

$$\frac{\mathrm{dLog}\, f_{MOAPP}(x;\alpha,\theta,\lambda)}{dx} = \lambda Log(\alpha)x^{-\lambda-1} - \frac{\lambda+1}{x} - 2\frac{\lambda Log(\alpha)\frac{(1-\theta)}{(\alpha-1)}\alpha^{1-x^{-\lambda}}x^{-\lambda-1}}{\left[\theta+\frac{(1-\theta)}{(\alpha-1)}\left(\alpha^{1-x^{-\lambda}}-1\right)\right]}$$

$$\frac{\mathrm{dLog}\, f_{MOAPP}(x;\alpha,\theta,\lambda)}{dx} = \frac{S(x)(\lambda Log(\alpha)-(\lambda+1)x^{\lambda})-2\lambda Log(\alpha)}{s(x)x^{\lambda+1}} \tag{7}$$

where $S(x) = [\frac{(\alpha-1)}{(1-\theta)}\theta\alpha^{x^{-\lambda}-1} + (1-\alpha^{x^{-\lambda}-1})]$. Equating (7) to zero, we obtain the cases:

1- If $0 < \theta < 1$ then $S(x)$ is positive and since $\lambda Log(\alpha) < (\lambda+1)x^{\lambda}$ then the numerator of equation (7) is negative hence the derivative of the logarithmic function of MOAPP is negative, which indicates that the pdf of MOAPP is a decreasing function.
2- If $\theta > 1$ and $\lambda > 1$ then by using Bolzano theorem on the interval [1,∞) there exist a root $x_0 \in [1,\infty)$ of Log $f_{MOAPP}$ hence $f_{MOAPP}$ attains its maximum at $x_0$.

### 2.2. Hazard Rate Function

The hazard rate function or failure rate is important in survival analysis and reliability theory. The hazard rate function for MOAPP distribution is of the form

$$h(x;\alpha,\theta,\lambda) = \frac{\lambda \text{Log}(\alpha) x^{-(\lambda+1)}}{(\alpha^{x^{-\lambda}}-1)(\theta+(1-\theta)(\alpha-1)^{-1}(\alpha^{1-x^{-\lambda}}-1))}, x \geq 1 \quad (8)$$

In order to determine the shape of h(x;α,θ, $\lambda$) it is quite enough to determine the shape of log h(x;α,θ, $\lambda$), as shown in the following lemma.

***Lemma 2***

The hazard rate of MOAPP distribution is either decreasing or upside down curve where the curve is skewed to the right.

***Proof:***

We consider a logarithmic function of the hazard rate given in Eq. (8) and take the first derivative with respect to x so that:

$$\frac{d\log h(x;\alpha,\theta,\lambda)}{dx} = \frac{-(\lambda+1)\left(\alpha^{x^{-\lambda}}-1\right)w(x) + \lambda\log(\alpha)\, x^{-\lambda}[w(x)\alpha^{x^{-\lambda}} - (1-\theta)\alpha(1-\alpha^{-x^{-\lambda}})]}{x\left(\alpha^{x^{-\lambda}}-1\right)w(x)}$$

where $w(x) = \alpha\left(\theta\left(1-\alpha^{-x^{-\lambda}}\right) + \alpha^{-x^{-\lambda}}\right) - 1$. The hazard rate curve may take several shapes according to different parameter values, so we summarize these cases by:

1- If $\theta > 1$ and $\alpha > 1$ then $\alpha^{-x^{-\lambda}} < 1$ and hence $w(x) < 0$, then log h(x; α, θ, $\lambda$) attains its maximum at a certain point $h_0 \in (1,\infty)$ so the hazard rate function is increasing on the interval $(1, h_0)$ and is decreasing $(h_0, \infty)$.
2- If $0 < \theta < 1$ and $0 < \alpha < 1$ then $\alpha^{-x^{-\lambda}} > 1$ hence $w(x) > 0$ and log h(x; α, θ, $\lambda$) is decreasing for all values of x, which indicates a decreasing hazard rate where h(1)= $\frac{\lambda \text{Log}(\alpha)}{(\alpha-1)\theta}$, and h(∞)=0.

Figure 2 illustrated the shape of the hazard rate function for some selected parameters' values.

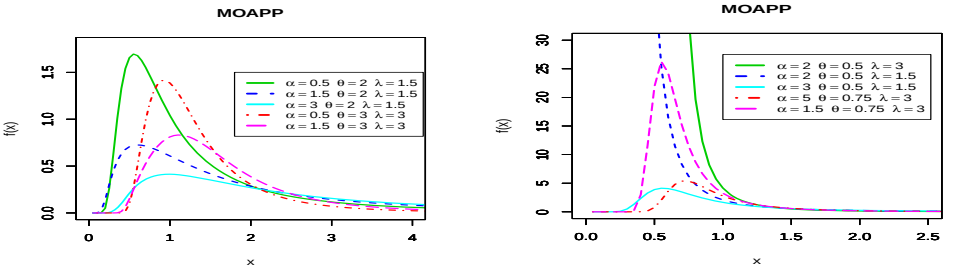



**Figure 1**. pdf of MOAPP under different values of the parameters

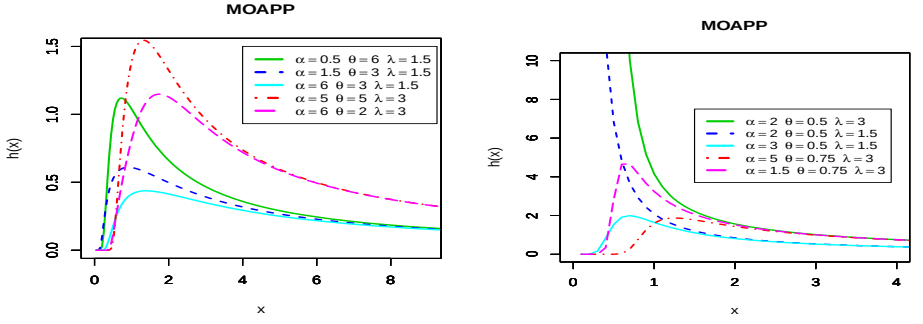



**Figure 2.** Hazard rate function of MOAPP under different values of the parameters

## 2.3. Moments

In order to obtain the moments for MOAPP distribution we use series representation for the pdf that is given in Eq. (5). The generalized binomial expansion will be used for this purpose, hence the MOAPP density can be rewritten as:

$$f_{MOAPP}(x) = \sum_{m=0}^{\infty} \wp_m \, \Omega_{m+1}(Y_m) \tag{9}$$

where $\wp_m =$

$$\begin{cases} \sum_{k=0}^{\infty}\sum_{j=0}^{k}(-1)^j(k+1)\theta(1-\theta)^k\binom{k}{j}\alpha^{k-j}\frac{(\log\alpha)^{m+1}(j+1)^m}{(\alpha-1)^{k+1}(m+1)!}, & 0<\theta<1 \\ \sum_{k=0}^{\infty}\sum_{j=0}^{k}(-1)^j(k+1)(1-\theta^{-1})^k\binom{k}{j}\frac{(\log\alpha)^{m+1}(k+1-j)^m}{\theta(\alpha-1)^{k+1}(m+1)!} & \theta>1 \end{cases},$$

and $\Omega_{m+1}(Y_m) = \frac{\lambda\,(m+1)}{y^{\lambda+1}}(1-\frac{1}{y^{\lambda}})^m$ , y≥1, which is the exponentiated-Pareto distribution with two shape parameters (m+1, $\lambda$).

Eq. (9) represents the MOAPP family density as a linear combination of exponentiated-Pareto density, hence some mathematical properties can be determined from this representation.

The $r^{th}$ moment MOAPP distribution can be computed from

$$E(X^r) = \sum_{m=0}^{\infty} \wp_m \, \mathrm{E}({Y_m}^r),$$

where $\mathrm{E}({Y_m}^r) = \int_1^{\infty} y^r \, \Omega_{m+1}(y)dy = (m+1)B(m+1,1-\frac{r}{\lambda})$, and $B(\alpha,\beta)$ is beta function.

### 2.4. Quantile function

By inverting Equation (6), we have the quantile of MOAPP distribution as follows:

$$x_q = \left(1 - \frac{1}{\ln(\alpha)} \ln\left(1 + \frac{\theta q(\alpha-1)}{1-q(1-\theta)}\right)\right)^{-1/\lambda}; 0 < q < 1 \tag{10}$$

## 3. Classical Point Estimation Methods

In this section we discuss two different classical point estimation methods, namely maximum likelihood estimation and maximum product spacing. Simulation analysis will take place in Section 6 in order to compare between the efficiency of these two methods.

### 3.1. Maximum Likelihood Estimation

The maximum likelihood estimation (MLE) is used in inferential statistics since it has many attractive properties, such as invariance, consistency and normal approximation properties. It depends basically on maximizing the likelihood function of MOAPP distribution. Let $X_1, X_2, \ldots, X_n$ be a random sample from MOAPP distribution, then the log likelihood function for the vector of parameters $\gamma=(\alpha,\theta,\lambda)$ can be expressed by

$$\ell(\gamma) = nLog[\theta\lambda Log(\alpha)] + \left(n - \sum_{i=1}^{n} {x_i}^{-\lambda}\right) Log(\alpha) - nLog(\alpha-1) - (\lambda+1)\sum_{i=1}^{n} Log x_i - 2\sum_{i=1}^{n} Log[\theta + \frac{(1-\theta)}{(\alpha-1)}\left(\alpha^{1-{x_i}^{-\lambda}} - 1\right)] \tag{11}$$

In order to obtain the MLE of the parameters α, θ and λ it is necessary to find the derivative of equation (11) with respect to α, θ and λ respectively.

$$\frac{\partial \ell(\gamma)}{\partial \alpha} = \frac{n + Log(\alpha)(n - \sum_{i=1}^{n} x_i)}{\alpha Log(\alpha)} - \frac{n}{\alpha-1} - 2\sum_{i=1}^{n} \frac{(1-\theta)\alpha^{-{x_i}^{-\lambda}}[(1-\alpha){x_i}^{-\lambda} + \alpha^{{x_i}^{-\lambda}} - 1]}{(\alpha-1)^2[\theta + \frac{(1-\theta)}{(\alpha-1)}\left(\alpha^{1-{x_i}^{-\lambda}} - 1\right)]}$$

$$\frac{\partial \ell(\gamma)}{\partial \theta} = \frac{n}{\theta} - 2\sum_{i=1}^{n} \frac{1 - (\alpha-1)^{-1}(\alpha^{1-{x_i}^{-\lambda}} - 1)}{[\theta + \frac{(1-\theta)}{(\alpha-1)}\left(\alpha^{1-{x_i}^{-\lambda}} - 1\right)]}$$

$$\frac{\partial \ell(\gamma)}{\partial \lambda} = \frac{n}{\lambda} + 2\sum_{i=1}^{n} \frac{(1-\theta)(\alpha-1)^{-1}\alpha^{1-{x_i}^{-\lambda}} {x_i}^{-\lambda} Log(x_i) Log(\alpha)}{[\theta + \frac{(1-\theta)}{(\alpha-1)}\left(\alpha^{1-{x_i}^{-\lambda}} - 1\right)]}$$

The solution for the above normal equations is not in an explicit form, hence the MLEs can be obtained numerically by using Newton or Newton-Raphson methods.

### 3.2. Maximum Product Spacings

The Maximum Product Spacings (MPS) method is a new point estimation method that is considered as an alternative to MLE, see Cheng and Amin (1983). This method was recently used by many authors, see, for example Singh et al. (2014), Singh et al. (2016), and Almetwally and Almongy (2019[a, b]). It was observed that MPS acts better than MLE in many cases. The MPS is defined as:

$$M = (\prod_{i=1}^{n+1} D_i)^{\frac{1}{n+1}},$$

where M is defined as the geometric mean of the product spacings function $D_i$ such that

$$D_1 = F(x_1)$$
$$D_i = F(x_i) - F(x_{i-1}); i = 2, \dots, n$$
$$D_{n+1} = 1 - F(x_n)$$

It is easy to see that $\sum_{i=1}^{n+1} D_i = 1$. The MPS method is based on the observed ordered sample $x_1<\cdots<x_n$ from MOAPP distribution, hence the product spacings function is

$$M(\gamma) = \left\{\frac{\alpha^{1-x_1{}^{-\lambda}}-1}{(\alpha-1)u(x_1)}\left(1-\frac{\alpha^{1-x_n{}^{-\lambda}}-1}{(\alpha-1)u(x_n)}\right)\prod_{i=2}^{n}\left[\frac{\alpha^{1-x_i{}^{-\lambda}}-1}{(\alpha-1)u(x_i)}-\frac{\alpha^{1-x_{i-1}{}^{-\lambda}}-1}{(\alpha-1)u(x_{i-1})}\right]\right\}^{\frac{1}{n+1}},$$

where $u(x_i) = \theta + (1-\theta)(\alpha-1)^{-1}\left(\alpha^{1-x_i{}^{-\lambda}}-1\right)$.

The natural logarithm of the product spacings function is

$$lnM(\gamma) = \frac{1}{n+1}\left\{ln\left(\alpha^{1-x_1{}^{-\lambda}}-1\right) - \ln\left((\alpha-1)u(x_1)\right) + \ln\left(1-\frac{\alpha^{1-x_n{}^{-\lambda}}-1}{(\alpha-1)u(x_n)}\right) + \sum_{i=2}^{n} ln\left[\frac{\alpha^{1-x_i{}^{-\lambda}}-1}{(\alpha-1)u(x_i)}-\frac{\alpha^{1-x_{i-1}{}^{-\lambda}}-1}{(\alpha-1)u(x_{i-1})}\right]\right\}. \quad (12)$$

To obtain the normal equations for the unknown parameters, we differentiate Eq. (12) partially with respect to the vector parameter γ and equate them to zero.

$$\frac{dlnM(\gamma)}{d\alpha} = \frac{1}{n+1}\left\{\frac{(1-x_1{}^{-\lambda})\alpha^{-x_1{}^{-\lambda}}}{\alpha^{1-x_1{}^{-\lambda}}-1} - \frac{u(x_1)+(\alpha-1)u_\alpha(x_1)}{(\alpha-1)u(x_1)} + \frac{(\alpha-1)u(x_n)(1-x_n{}^{-\lambda})\alpha^{-x_n{}^{-\lambda}}-(\alpha^{1-x_n{}^{-\lambda}}-1)[(\alpha-1)u_\alpha(x_n)+u(x_n)]}{((\alpha-1)u(x_n))^2-((\alpha-1)u(x_n)\alpha^{1-x_n{}^{-\lambda}}-1)} + \right.$$

$$\left. \sum_{i=1}^{n} \frac{\frac{(\alpha-1)u(x_i)(1-x_i{}^{-\lambda})\alpha^{-x_i{}^{-\lambda}}-\left(\alpha^{1-x_i{}^{-\lambda}}-1\right)[(\alpha-1)u_\alpha(x_i)+u(x_i)]}{((\alpha-1)u(x_i))^2} - \frac{(\alpha-1)u(x_{i-1})(1-x_{i-1}{}^{-\lambda})\alpha^{1-x_{i-1}{}^{-\lambda}}-(\alpha^{1-x_{i-1}{}^{-\lambda}}-1)[(\alpha-1)u_\alpha(x_{i-1})+u(x_{i-1})]}{((\alpha-1)u(x_{i-1}))^2}}{\frac{\alpha^{1-x_i{}^{-\lambda}}-1}{(\alpha-1)u(x_i)}-\frac{\alpha^{1-x_{i-1}{}^{-\lambda}}-1}{(\alpha-1)u(x_{i-1})}} \right\}$$

$$\frac{dlnM(\gamma)}{d\theta}=\frac{1}{n+1}\left\{\frac{u_\theta(x_1)}{u(x_1)}+\frac{(\alpha^{1-x_n{}^{-\lambda}}-1)u_\theta(x_n)}{u(x_n)((\alpha-1)u(x_n)-(\alpha^{1-x_n{}^{-\lambda}}-1))}+\right.$$

$$\left.\sum_{i=2}^{n}\frac{-\frac{\left(\alpha^{1-x_i{}^{-\lambda}}-1\right)u_\theta(x_i)}{\left(u(x_i)\right)^2}+\frac{(\alpha^{1-x_{i-1}{}^{-\lambda}}-1)u_\theta(x_{i-1})}{(u(x_{i-1}))^2}}{\left[\frac{\alpha^{1-x_i{}^{-\lambda}}-1}{u(x_i)}-\frac{\alpha^{1-x_{i-1}{}^{-\lambda}}-1}{u(x_{i-1})}\right]}\right\},$$

$$\frac{dlnM(\gamma)}{d\lambda}=\frac{1}{n+1}\left\{\frac{Log(\alpha x_1)x_1{}^{-\lambda}}{1-\alpha^{x_1{}^{-\lambda}-1}}-\frac{(\alpha-1)u_\lambda(x_1)}{(\alpha-1)u(x_1)}-\frac{u(x_n)\alpha^{1-x_n{}^{-\lambda}}Log(\alpha x_n)x_n{}^{-\lambda}-\left(\alpha^{1-x_n{}^{-\lambda}}-1\right)u_\lambda(x_n)}{u(x_n)((\alpha-1)u(x_n)-(\alpha^{1-x_n{}^{-\lambda}}-1))}+\right.$$

$$\left.\sum_{i=2}^{n}\frac{\frac{u(x_i)\alpha^{1-x_i{}^{-\lambda}}Log(\alpha x_i)x_i{}^{-\lambda}-\left(\alpha^{1-x_i{}^{-\lambda}}-1\right)u_\lambda(x_i)}{u(x_i)((\alpha-1)u(x_i)-(\alpha^{1-x_i{}^{-\lambda}}-1))}-\frac{u(x_{i-1})\alpha^{1-x_{i-1}{}^{-\lambda}}Log(\alpha x_{i-1})x_{i-1}{}^{-\lambda}-\left(\alpha^{1-x_{i-1}{}^{-\lambda}}-1\right)u_\lambda(x_{i-1})}{u(x_{i-1})((\alpha-1)u(x_{i-1})-(\alpha^{1-x_{i-1}{}^{-\lambda}}-1))}}{\frac{\alpha^{1-x_i{}^{-\lambda}}-1}{(\alpha-1)u(x_i)}-\frac{\alpha^{1-x_{i-1}{}^{-\lambda}}-1}{(\alpha-1)u(x_{i-1})}}\right\}, \tag{13}$$

where $u_\alpha, u_\theta$ and $u_\lambda$ represent the partial derivative of $u(x_i)$ with respect to α, θ and λ respectively. The estimators of γ can be obtained by solving the above system of nonlinear equations numerically, so the MPS of α, θ and λ are denoted by $\hat{\alpha}_{MP}$, $\hat{\theta}_{MP}$ and $\hat{\lambda}_{MP}$ respectively.

## 4. Bayesian Estimation Method

In this section we consider the non-classical method of estimation that is Bayes estimates for the unknown parameters α, θ and λ of MOAPP distribution. The quadratic loss and LINEX loss functions are the assumed loss functions.

In Bayesian method, all parameters are random variables with a certain distribution called prior distribution. If prior information is not available which is usually the case, we need to select a prior distribution. Since the selection of prior distribution plays an important role in estimation of the parameters, our choice for the prior of α,θ and λ are the independent gamma distributions, which are G($a_1$, $b_1$), G($a_2$, $b_2$) and G($a_3$, $b_3$) respectively. Thus, the suggested prior for α, θ and λ can be written as:

$$\pi_1(\alpha)\propto\alpha^{a_1-1}e^{-b_1\alpha},\ \pi_2(\theta)\propto\theta^{a_2-1}e^{-b_2\theta},\ \pi_3(\lambda)\propto\lambda^{a_3-1}e^{-b_3\lambda},$$

respectively, where $a_1$, $a_2$, $a_3$, $b_1$, $b_2$ and $b_3$ are the hyper parameters of prior distributions.

The joint prior of α, θ and λ is

$$k(\alpha,\theta,\lambda)\propto\alpha^{a_1-1}\theta^{a_2-1}\lambda^{a_3-1}e^{-b_1\alpha-b_2\theta-b_3\lambda},\ \alpha,\theta,\lambda,a_1,a_2,a_3,b_1,b_2,b_3>0.$$

The joint posterior of α, θ and λ is given by

$$p(\alpha,\theta,\lambda/\underline{x}) \propto L(\underline{x}/\alpha,\theta,\lambda)k(\alpha,\theta,\lambda),$$

where L($\underline{x}$/α,θ, λ) is the likelihood function of MOAPP distribution. When substituting the likelihood function L($\underline{x}$/α, θ, λ) and the joint prior $k(\alpha,\theta,\lambda)$in the above equation, the joint posterior will be:

$$p\left(\alpha,\theta,\frac{\lambda}{\underline{x}}\right) \propto \alpha^{n-\sum_{i=1}^{n} x_i{}^{-\lambda}+a_1-1}\theta^{n+a_2-1}\lambda^{n+a_3-1}e^{-b_1\alpha-b_2\theta-b_3\lambda}\prod_{i=1}^{n}\frac{1}{\alpha-1}\frac{Log\alpha}{x_i{}^{\lambda+1}}\left(\theta+\frac{(1-\theta)}{\alpha-1}\left(\alpha^{1-x_i{}^{-\lambda}}-1\right)\right)^{-2}$$

$$p(\alpha,\theta,\lambda/\underline{x}) \propto G_{\alpha\backslash\lambda}(n-\textstyle\sum_{i=1}^{n} x_i{}^{-\lambda}+a_1,b_1)G_\theta(n+a_2,b_2)G_\lambda(n+a_3,b_3)e^{\phi(\alpha,\theta,\lambda)},$$

where $\phi(\alpha,\theta,\lambda)=\sum_{i=1}^{n} ln\frac{1}{\alpha-1}\frac{Log\alpha}{x_i{}^{\lambda+1}}-2\ln\left(\theta+\frac{(1-\theta)}{\alpha-1}\left(\alpha^{1-x_i{}^{-\lambda}}-1\right)\right)$

In the case of quadratic loss function, Bayes estimate is the posterior mean, the determination of posterior mean for the purpose of obtaining Bayes estimation of the parameters α, θ and λ, is not easy to obtain unless we use numerical approximation methods.

In the literature, there are several approximation methods available to solve this kind of problem. Here, we consider Monte Carlo Markov Chain (MCMC) approximation method, see Karandikar (2006). This approximation method reduces the ratio of integrals into a whole and produces a single numerical result.

A wide variety of MCMC schemes are available. An important sub-class of MCMC methods are Gibbs sampling and more general Metropolis within Gibbs samplers. Indeed, the MCMC samples may be used to completely summarize the posterior uncertainty about the parameters α, θ and λ, through a kernel estimate of the posterior distribution. This is also true of any function of the parameters.

Therefore, to generate samples from MOAPP distribution, we use the Metropolis-Hastings method (Metropolis et al. (1953) with normal proposal distribution). For details regarding the implementation of the Metropolis-Hasting algorithm, the readers may refer to Robert and Casella (2013) and Almetwally et al. (2018). The full conditional posterior densities of α, θ and λ and the data are given by:

$$\pi(\alpha/\theta,\lambda;\underline{x}) \propto G_{\alpha\backslash\lambda}\left(n-\sum_{i=1}^{n} x_i{}^{-\lambda}+a_1,b_1\right)e^{\phi(\alpha,\theta,\lambda)}$$

$$\pi(\theta/\alpha,\lambda,\underline{x}) = G_\theta(n+a_2,b_2)e^{-2\ln\left(\theta+\frac{(1-\theta)}{\alpha-1}\left(\alpha^{1-x_i{}^{-\lambda}}-1\right)\right)} \tag{14}$$

$$\pi(\lambda/\alpha,\theta,\underline{x}) = G_\lambda(n+a_3,b_3)e^{\phi(\alpha,\theta,\lambda)}$$

To apply the Gibbs technique we need the following algorithm:

(1) Start with initial values $(\alpha_0,\theta_0,\lambda_0)$

(2) Use M-H algorithm to generate posterior sample for α, θ and λ from Eq. (14)

(3) Repeat step 2 (T)-times and obtain $(\alpha_1, \theta_1, \lambda_1)$, $\alpha_2, \theta_2, \lambda_2)$,…,$( \alpha_T, \theta_T, \lambda_T)$
(4) After obtaining the posterior sample, the Bayes estimates of α, θ and λ with respect to quadratic loss function are:

$$\hat{\alpha}^{MC} = [E_\pi(\alpha/x)] \approx (\frac{1}{T-T_0}\sum_{i=1}^{T-T_0}\alpha_i)$$

$$\hat{\theta}^{MC} = [E_\pi(\theta\Big/x)] \approx (\frac{1}{T-T_0}\sum_{i=1}^{T-T_0}\theta_i)$$

$$\hat{\lambda}^{MC} = [E_\pi(\lambda\Big/x)] \approx (\frac{1}{T-T_0}\sum_{i=1}^{T-T_0}\lambda_i)$$

where $T_0$ is the burn-in-period of Markov Chain.

The Bayes estimates of the unknown parameters α, θ and λ under the LINEX loss function can be calculated through the following equation:

$$\gamma_j = \frac{-1}{\upsilon} ln\left(\sum_{i=1}^{L}\frac{e^{-\upsilon\gamma^{(i)}}}{L}\right),$$

where ν reflects the direction and degree of asymmetry, L is number of periods in the MCMC.

## 5. Interval Estimation Methods

In this section we consider three methods of approximate confidence intervals for the parameters of MOAPP distribution. Numerical analysis via simulation is used for comparisons between these methods in Section 6.

### 5.1. Asymptotic confidence Interval for (MLE)

When the sample size is large enough, the normal approximation of the MLE can be used to construct asymptotic confidence intervals for the parameters α, θ and λ. The asymptotic normality of MLE can be stated as $\sqrt{n}(\hat{\gamma}-\gamma) \xrightarrow{d} N_3(0, I^{-1}(\gamma))$, where γ=(α,θ, λ) is a vector of parameters, $\xrightarrow{d}$ denotes convergence in distribution and $I(\gamma)$ is the Fisher information matrix

$$I(\gamma) = -\begin{bmatrix} E(\ell_{\alpha\alpha}) & E(\ell_{\alpha\theta}) & E(\ell_{\alpha\lambda}) \\ E(\ell_{\theta\alpha}) & E(\ell_{\theta\theta}) & E(\ell_{\theta\lambda}) \\ E(\ell_{\lambda\alpha}) & E(\ell_{\lambda\theta}) & E(\ell_{\lambda\lambda}) \end{bmatrix}$$

The expected values of the second derivatives can be found by using some integration techniques. Therefore, the (1- $\zeta$) 100% approximate CIs for α, θ and λ are

$$\hat{\alpha} \pm z_{\frac{\zeta}{2}}\sqrt{\upsilon_{11}}, \hat{\theta} \pm z_{\frac{\zeta}{2}}\sqrt{\upsilon_{22}}, \hat{\lambda} \pm z_{\frac{\zeta}{2}}\sqrt{\upsilon_{33}},$$

respectively, where $\upsilon_{11}, \upsilon_{22}, \upsilon_{33}$ are the entries in the main diagonal of the Fisher matrix $I^{-1}(\gamma)$, and $z_{\frac{\zeta}{2}}$ is the ($\frac{\zeta}{2}$) 100% lower percentile of the standard normal distribution.

### 5.2. Asymptotic Confidence Interval for (MPS)

In this section, we propose the asymptotic confidence intervals using MPS method. As it was mentioned by Cheng and Amin [1979], Ghosh and Jammalamadaka [2001] and Anatolyev and Kosenok [2005], the MPS method also shows asymptotic properties like the maximum likelihood estimator and is asymptotically equivalent to MLE. Therefore, we may propose the asymptotic confidence intervals using MPS. The exact distribution of the MPS cannot be obtained explicitly. Therefore, the asymptotic properties of MPS similar to that of MLE can be used to construct the confidence intervals.

$$J(\gamma) = -\begin{bmatrix} E(M_{\alpha\alpha}) & E(M_{\alpha\theta}) & E(M_{\alpha\lambda}) \\ E(M_{\theta\alpha}) & E(M_{\theta\theta}) & E(M_{\theta\lambda}) \\ E(M_{\lambda\alpha}) & E(M_{\lambda\theta}) & E(M_{\lambda\lambda}) \end{bmatrix}$$

The first derivatives of the product of spacing, i.e. the function M with respect to parameters α, θ and λ, are given by Equation (13), second derivative can be found numerically and hence one can obtain the (1- $\zeta$) 100% asymptotic confidence intervals based on MPS as follows:

$$\hat{\alpha}_{MP} \pm z_{\frac{\zeta}{2}}\sqrt{\omega_{11}}, \hat{\theta}_{MP} \pm z_{\frac{\zeta}{2}}\sqrt{\omega_{22}}\,, \hat{\lambda}_{MP} \pm z_{\frac{\zeta}{2}}\sqrt{\omega_{33}},$$

where $\omega_{11}$, $\omega_{22}$ and $\omega_{33}$ are the diagonal entries of the Fisher matrix $J^{-1}(\gamma)$.

### 5.3. Credible intervals

Using MCMC techniques in Section (4), the Bayes credible intervals of the parameter α, θ and λ can be obtained as follows:

(1) Arrange $\alpha_i$, $\theta_i$ and $\lambda_i$; in ascending order as follow $\alpha_{[1]}, \alpha_{[2]}, \dots, \alpha_{[T]}$, $\theta_{[1]}, \theta_{[2]}, \dots, \theta_{[T]}$ and $\lambda_{[1]}, \lambda_{[2]}, \dots, \lambda_{[T]}$

(2) A two-sided (1- $\zeta$) 100% credible intervals for the unknown parameters α, θ and λ are given by

$$(\alpha_{\left[T\frac{\zeta}{2}\right]}, \alpha_{\left[T_{1-\frac{\zeta}{2}}\right]}), (\theta_{\left[T\frac{\zeta}{2}\right]}, \theta_{\left[T_{1-\frac{\zeta}{2}}\right]}), (\lambda_{\left[T\frac{\zeta}{2}\right]}, \lambda_{\left[T_{1-\frac{\zeta}{2}}\right]}),$$

respectively, where [x] denoted the largest integer less than or equal to x.

## 6. Simulation Study and Data Analysis

### 6.1. Simulation Study

In this section, we consider some experimental results that are produced to see the effectiveness of different point and interval estimation methods. We mainly compare different point estimates in terms of mean squared errors (MSE) and bias values. Efficiency of confidence intervals is compared in terms of average interval length (AIL). Based on the generated data, we compute MLE and MPS estimates using the Newton-Raphson method. Further, we compute Bayes estimates using a Monte Carlo simulation and the MH algorithm under both squared error and LINEX loss functions with v=1.5 by using R language.

We start by building our model with generating all simulation controls. In this stage, we must do the following steps in sequence:

Step 1: Suppose the following values for the parameter vector of MOAPP distribution γ=(α,θ, λ), case 1=(0.5, 0.5, 1.5), case 2=(1.5, 0.5, 1.5), case 3=(3,0.5,1.5), case 4=(0.5,1.5, 1.5), case 5 =(1.5,1.5,1.5) and case 6=(3,1.5,1.5), case 7=(0.5,3,1.5), case 8=(1.5,3,1.5), case 9=(3,3,1.5).

Step 2: Choose sample sizes n =30, 70 and 200.

Step 3: Generate the sample random values of MOAPP distribution by using quantile function$X = \left(1 - \frac{1}{\ln(\alpha)} \ln\left(1 + \frac{\theta U(\alpha-1)}{1-U(1-\theta)}\right)\right)^{-1/\lambda}$, where U is a uniform distribution (0, 1).

Step 4: Solve differential equations for each estimation methods, to obtain the estimators of the parameters for MOAPP distribution, so we calculate α, θ, and λ.

Step 5: Repeat this experiment (L-1) times. In each experiment use the same values of the parameters. It is certain that the values of generating random samples are varying from experiment to experiment even though the sample size (n) does not change.

Finally, we have L-values of bias and MSE. We compute the average biases and average MSE's over 10,000 runs. This number of runs will give the accuracy in the order ±0.01 (see Karian and Dudewicz (1998)). The bias of estimator is equal to $\hat{\gamma} - \gamma$, where $\hat{\gamma}$ is the estimated value of γ, and the mean squared error (MSE) of the estimator is MSE=Mean $(\hat{\gamma} - \gamma)^2$.

The simulated results of point estimation methods are presented in Tables (1) to (3), where the MSE and the bias are given in each cell and it can be pointed out that the MPS and Bayesian methods for estimating the unknown parameters of MOAPP distribution are better than the MLE method, where the MSE value is considered for comparison. We summarize the cases as follows:

1- For $0 < \alpha < 1$, the best point estimation method for estimating $\alpha$ is the Bayesian method under LINEX loss function, while for $\alpha > 1$ the best estimation method is the MPS and Baysian under the SE loss function.

2- For $0 < \theta < 1$, the best point estimation method for $\theta$ is the Bayesian method under LINEX loss function, while for $\theta > 1$ the best estimation method is the MPS and Bayesian under the SE loss function.
3- For all values of $\lambda$ the Bayesian under the SE loss function is the best estimation method.
4- The average bias and MSE decrease as the sample size increases. It verifies the consistency properties of all the estimators.
5- The MLE overestimates α, θ and λ for almost all cases except for case 3, where the MLE underestimates α. It is also noticed that when the sample size n=200, the MLE underestimate α for cases 1, 3, 6 and underestimate θ for case 7, see Table (3).
6- MPS and Bayesian estimation sometimes overestimate the parameters and sometimes underestimate them.

Figure 3 shows the three dimension plots of MSE with different parameters values

For confidence interval estimation of MOAPP parameters α, θ and λ, we observe the 95% confidence intervals (L,U) where L represents the lower bound and U is the upper bound of this interval. Three confidence intervals are considered in simulation analysis, i.e. asymptotic confidence intervals of MLE and MPS, also the credible intervals of Bayesian method under SE and LINEX loss functions. The comparison is conducted depending on the average interval length (AIL), hence the smaller the AIL is the better confidence estimate we observe. The results are reported in Tables (4) to (6) below.

**Table 1.** Bias and MSE for $\alpha, \theta, and\ \lambda$, with n=30

| $\lambda$ | $\theta$ | $\alpha$ | **n=30** | MLE | | MPS | | SE | | LINEX ($v = 1.5$) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Bias | MSE | Bias | MSE | Bias | MSE | Bias | MSE |
| 1.5 | 0.5 | 0.5 | $\alpha$ | 0.1072 | 0.1866 | -0.0160 | 0.0818 | 0.1051 | 0.0994 | 0.0310 | 0.0661 |
| | | | $\theta$ | 0.3212 | 0.4946 | 0.0540 | 0.2012 | 0.0517 | 0.0536 | 0.0071 | 0.0395 |
| | | | $\lambda$ | 0.3134 | 0.9250 | -0.2585 | 0.6378 | -0.0858 | 0.1772 | -0.1893 | 0.1832 |
| | | 1.5 | $\alpha$ | 0.0975 | 0.4482 | 0.0092 | 0.0306 | -0.2242 | 0.3509 | -0.3552 | 0.4101 |
| | | | $\theta$ | 0.2341 | 0.3511 | -0.0125 | 0.1299 | 0.1058 | 0.0816 | 0.0559 | 0.0549 |
| | | | $\lambda$ | 0.2786 | 0.6504 | -0.1677 | 0.4781 | -0.0966 | 0.1989 | -0.1919 | 0.2103 |
| | | 3 | $\alpha$ | -0.0792 | 1.1131 | -0.0032 | 0.0421 | -0.3287 | 0.5870 | -0.5214 | 0.8308 |
| | | | $\theta$ | 0.2925 | 0.4977 | -0.0030 | 0.1425 | 0.1049 | 0.0894 | 0.0561 | 0.0620 |
| | | | $\lambda$ | 0.2350 | 0.4980 | -0.1624 | 0.3824 | -0.0554 | 0.1675 | -0.1436 | 0.1748 |
| | 1.5 | 0.5 | $\alpha$ | 0.3291 | 0.6917 | 0.0265 | 0.1905 | 0.1369 | 0.1079 | 0.0642 | 0.0693 |
| | | | $\theta$ | 0.2813 | 0.7987 | -0.0524 | 0.3616 | -0.1709 | 0.2413 | -0.2738 | 0.2666 |
| | | | $\lambda$ | 0.1274 | 0.4280 | -0.2490 | 0.4257 | -0.1173 | 0.1662 | -0.2064 | 0.1787 |
| | | 1.5 | $\alpha$ | 0.4186 | 1.2902 | -0.0008 | 0.1538 | -0.1472 | 0.2932 | -0.2716 | 0.3255 |
| | | | $\theta$ | 0.3391 | 1.1591 | -0.0960 | 0.5198 | -0.1357 | 0.2656 | -0.2441 | 0.2761 |
| | | | $\lambda$ | 0.1351 | 0.2503 | -0.1453 | 0.2163 | -0.1746 | 0.1315 | -0.2488 | 0.1643 |
| | | 3 | $\alpha$ | 0.2624 | 2.2557 | 0.0121 | 0.1190 | -0.2997 | 0.5282 | -0.4753 | 0.7364 |
| | | | $\theta$ | 0.5562 | 1.8669 | -0.0793 | 0.6735 | -0.1658 | 0.2256 | -0.2638 | 0.2538 |
| | | | $\lambda$ | 0.1207 | 0.2080 | -0.1428 | 0.1813 | -0.1191 | 0.0944 | -0.1806 | 0.1122 |

**Table 1.** Bias and MSE for $\alpha, \theta, and\ \lambda$, with n=30 (cont.)

| $\lambda$ | $\theta$ | $\alpha$ | **n=30** | MLE | | MPS | | SE | | LINEX ($\upsilon = 1.5$) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Bias | MSE | Bias | MSE | Bias | MSE | Bias | MSE |
| 1.5 | 3 | 0.5 | $\alpha$ | 0.4689 | 1.2082 | 0.0355 | 0.2951 | 0.1449 | 0.1438 | 0.0665 | 0.0872 |
| | | | $\theta$ | 0.1613 | 0.9882 | -0.0498 | 0.3326 | -0.3164 | 0.4939 | -0.4915 | 0.7154 |
| | | | $\lambda$ | 0.0897 | 0.2667 | -0.2146 | 0.2902 | -0.0869 | 0.0977 | -0.1582 | 0.1124 |
| | | 1.5 | $\alpha$ | 0.6631 | 3.0228 | -0.1891 | 0.6353 | -0.1309 | 0.3296 | -0.2644 | 0.3409 |
| | | | $\theta$ | 0.4555 | 2.2871 | -0.0747 | 0.5972 | -0.3635 | 0.5986 | -0.5534 | 0.8297 |
| | | | $\lambda$ | 0.0844 | 0.1606 | -0.1612 | 0.1698 | -0.1407 | 0.0853 | -0.1967 | 0.1038 |

**Table 2.** Bias and MSE for $\alpha, \theta, and\ \lambda$, with n=70

| $\lambda$ | $\theta$ | $\alpha$ | n=70 | MLE | | MPS | | SE | | LINEX ($\upsilon = 1.5$) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Bias | MSE | Bias | MSE | Bias | MSE | Bias | MSE |
| 1.5 | 0.5 | 0.5 | $\alpha$ | 0.0336 | 0.1130 | -0.0739 | 0.0776 | 0.0250 | 0.0670 | -0.0041 | 0.0572 |
| | | | $\theta$ | 0.2013 | 0.2553 | 0.0725 | 0.1261 | 0.0400 | 0.0359 | 0.0215 | 0.0304 |
| | | | $\lambda$ | 0.1153 | 0.3558 | -0.2104 | 0.3237 | -0.0639 | 0.1000 | -0.0990 | 0.1095 |
| | | 1.5 | $\alpha$ | 0.0623 | 0.2540 | 0.0138 | 0.0160 | -0.0476 | 0.1202 | -0.0944 | 0.1319 |
| | | | $\theta$ | 0.0885 | 0.0804 | -0.0261 | 0.0491 | 0.0456 | 0.0296 | 0.0289 | 0.0255 |
| | | | $\lambda$ | 0.1287 | 0.2330 | -0.1076 | 0.2087 | -0.0293 | 0.0771 | -0.0618 | 0.0782 |
| | | 3 | $\alpha$ | -0.0249 | 0.6191 | 0.0128 | 0.0235 | -0.0472 | 0.1116 | -0.0870 | 0.1191 |
| | | | $\theta$ | 0.1154 | 0.1198 | -0.0197 | 0.0518 | 0.0161 | 0.0295 | 0.0005 | 0.0265 |
| | | | $\lambda$ | 0.1113 | 0.1867 | -0.0977 | 0.1677 | -0.0338 | 0.0770 | -0.0667 | 0.0822 |
| | 1.5 | 0.5 | $\alpha$ | 0.1695 | 0.3175 | -0.0285 | 0.1061 | 0.0281 | 0.0477 | 0.0027 | 0.0410 |
| | | | $\theta$ | 0.1438 | 0.3712 | 0.0043 | 0.1709 | -0.0248 | 0.0884 | -0.0606 | 0.0921 |
| | | | $\lambda$ | 0.0368 | 0.1866 | -0.1745 | 0.2275 | -0.0403 | 0.0616 | -0.0687 | 0.0652 |
| | | 1.5 | $\alpha$ | 0.2486 | 0.6849 | -0.0010 | 0.0723 | -0.0557 | 0.1385 | -0.1022 | 0.1463 |
| | | | $\theta$ | 0.1372 | 0.4181 | -0.0742 | 0.2391 | -0.0262 | 0.0840 | -0.0630 | 0.0839 |
| | | | $\lambda$ | 0.0646 | 0.1015 | -0.0772 | 0.0965 | -0.0315 | 0.0444 | -0.0537 | 0.0460 |
| | | 3 | $\alpha$ | 0.1262 | 1.3314 | 0.0191 | 0.0541 | -0.0835 | 0.1412 | -0.1363 | 0.1604 |
| | | | $\theta$ | 0.2561 | 0.6573 | -0.0634 | 0.3260 | -0.0548 | 0.0829 | -0.0907 | 0.0905 |
| | | | $\lambda$ | 0.0609 | 0.0867 | -0.0736 | 0.0816 | -0.0377 | 0.0371 | -0.0590 | 0.0397 |
| | | | $\theta$ | 0.2134 | 0.7791 | -0.0560 | 0.2755 | -0.0493 | 0.1043 | -0.0916 | 0.1168 |
| | | | $\lambda$ | 0.0865 | 0.2809 | -0.1488 | 0.2530 | -0.0658 | 0.0681 | -0.0989 | 0.0762 |

**Table 3.** Bias and MSE for $\alpha, \theta, and\ \lambda$, with n=200

| $\lambda$ | $\theta$ | $\alpha$ | **n=200** | MLE | | MPS | | SE | | LINEX ($\upsilon = 1.5$) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Bias | MSE | Bias | MSE | Bias | MSE | Bias | MSE |
| 1.5 | 0.5 | 0.5 | $\alpha$ | -0.0015 | 0.0775 | -0.0905 | 0.0613 | 0.0009 | 0.0125 | -0.0042 | 0.0124 |
| | | | $\theta$ | 0.1046 | 0.0836 | 0.0714 | 0.0588 | 0.0091 | 0.0066 | 0.0058 | 0.0064 |
| | | | $\lambda$ | 0.0222 | 0.1137 | -0.1368 | 0.1294 | -0.0161 | 0.0117 | -0.0210 | 0.0119 |
| | | 1.5 | $\alpha$ | 0.0142 | 0.0877 | 0.0118 | 0.0068 | -0.0160 | 0.0178 | -0.0224 | 0.0182 |
| | | | $\theta$ | 0.0353 | 0.0227 | -0.0186 | 0.0152 | -0.0021 | 0.0049 | -0.0053 | 0.0049 |
| | | | $\lambda$ | 0.0528 | 0.0709 | -0.0515 | 0.0683 | -0.0041 | 0.0117 | -0.0087 | 0.0119 |
| | | 3 | $\alpha$ | -0.0391 | 0.3088 | 0.0124 | 0.0111 | -0.0136 | 0.0120 | -0.0189 | 0.0124 |
| | | | $\theta$ | 0.0460 | 0.0303 | -0.0153 | 0.0155 | -0.0028 | 0.0054 | -0.0058 | 0.0054 |
| | | | $\lambda$ | 0.0470 | 0.0581 | -0.0453 | 0.0551 | -0.0054 | 0.0120 | -0.0103 | 0.0120 |

**Table 3.** Bias and MSE for $\alpha, \theta, and\ \lambda$, with n=200 (cont.)

| $\lambda$ | $\theta$ | $\alpha$ | **n=200** | MLE | | MPS | | SE | | LINEX ($\upsilon = 1.5$) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Bias | MSE | Bias | MSE | Bias | MSE | Bias | MSE |
| 1.5 | 1.5 | 0.5 | $\alpha$ | 0.0811 | 0.1085 | -0.0388 | 0.0462 | -0.0007 | 0.0105 | -0.0051 | 0.0104 |
| | | | $\theta$ | 0.1062 | 0.2893 | 0.0313 | 0.0781 | -0.0098 | 0.0133 | -0.0153 | 0.0133 |
| | | | $\lambda$ | 0.0015 | 0.0617 | -0.0828 | 0.0750 | -0.0065 | 0.0107 | -0.0110 | 0.0110 |
| | | 1.5 | $\alpha$ | 0.2053 | 0.7433 | -0.0043 | 0.0249 | -0.0106 | 0.0184 | -0.0164 | 0.0189 |
| | | | $\theta$ | 0.0514 | 0.1709 | -0.0382 | 0.0765 | -0.0062 | 0.0175 | -0.0120 | 0.0177 |
| | | | $\lambda$ | 0.0244 | 0.0320 | -0.0322 | 0.0321 | -0.0068 | 0.0093 | -0.0111 | 0.0093 |
| | | 3 | $\alpha$ | -0.0239 | 1.1126 | 0.0035 | 0.0187 | -0.0147 | 0.0171 | -0.0205 | 0.0176 |
| | | | $\theta$ | 0.1709 | 0.3354 | -0.0348 | 0.1077 | -0.0227 | 0.0146 | -0.0285 | 0.0151 |
| | | | $\lambda$ | 0.0272 | 0.0289 | -0.0302 | 0.0276 | -0.0167 | 0.0074 | -0.0207 | 0.0077 |
| | 3 | 0.5 | $\alpha$ | 0.0965 | 0.1246 | -0.0138 | 0.0489 | -0.0011 | 0.0108 | -0.0056 | 0.0106 |
| | | | $\theta$ | -0.0013 | 0.2440 | 0.0044 | 0.0374 | -0.0133 | 0.0136 | -0.0191 | 0.0139 |
| | | | $\lambda$ | 0.0164 | 0.0375 | -0.0498 | 0.0434 | -0.0005 | 0.0076 | -0.0047 | 0.0076 |
| | | 1.5 | $\alpha$ | 0.0865 | 0.2790 | -0.0494 | 0.1157 | -0.0120 | 0.0160 | -0.0177 | 0.0163 |
| | | | $\theta$ | 0.0794 | 0.2092 | -0.0320 | 0.1043 | -0.0416 | 0.0195 | -0.0484 | 0.0209 |
| | | | $\lambda$ | 0.0205 | 0.0219 | -0.0281 | 0.0209 | -0.0158 | 0.0063 | -0.0195 | 0.0064 |
| | | | $\theta$ | 0.0329 | 0.5907 | -0.0334 | 0.1001 | -0.0094 | 0.0162 | -0.0155 | 0.0166 |
| | | | $\lambda$ | 0.0312 | 0.0863 | -0.0568 | 0.0813 | -0.0111 | 0.0129 | -0.0162 | 0.0131 |

From Tables (4) to (6) we notice that the (AIL) of the credible intervals under SE and LINEX are smaller than the (AIL) of MLE and MPS in most cases except for some restricted ones.

We can summarize the analysis of the confidence interval estimation in the following points:

1. For $0 < \alpha < 1$, the best interval estimate for $\alpha$ is the Bayesian credible interval under SE and LINEX loss functions, while for $\alpha > 1$ the best interval estimation is the asymptotic interval under the MPS method except for cases 8 and 11, where the Bayesian credible interval under LINEX has the smallest AIL.
2. For $\theta < 3$, the best interval estimate for $\theta$ is the Bayesian credible interval under LINEX loss function, while for $\theta \geq 3$, the best interval estimation is the asymptotic interval under the MPS method and the Bayesian credible interval under the SE loss function.
3. Bayesian credible interval under the LINEX loss function has the smallest AIL for estimating $\lambda$ , and hence it can be considered as the best confidence interval of $\lambda$. For the case 5, the Bayesian credible interval under the SE loss function is preferable to estimate $\lambda$.
4. AIC decreases as the sample size increases.

**Table 4.** 95% confidence intervals and Average Interval length for $\alpha, \theta, and\ \lambda$, with n=30

| $\lambda$ | $\theta$ | $\alpha$ | **30** | MLE | | | MPS | | | SE | | | LINEX($\upsilon$ = 1.5) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | L | U | AIL | L | U | AIL | L | U | AIL | L | U | AIL |
| 1.5 | 0.5 | 0.5 | $\alpha$ | -0.2133 | 1.4278 | 1.6411 | -0.0759 | 1.0439 | 1.1198 | 0.0210 | 1.1892 | 1.1683 | 0.0294 | 1.0326 | 1.0033 |
| | | | $\theta$ | -0.4056 | 2.0480 | 2.4536 | -0.3191 | 1.4271 | 1.7462 | 0.1082 | 0.9952 | 0.8870 | 0.1169 | 0.8972 | 0.7803 |
| | | | $\lambda$ | 0.0303 | 3.5965 | 3.5662 | -0.2403 | 2.7233 | 2.9636 | 0.6044 | 2.2240 | 1.6196 | 0.5564 | 2.0650 | 1.5086 |
| | | 1.5 | $\alpha$ | 0.2986 | 2.8963 | 2.5976 | 1.1668 | 1.8515 | 0.6847 | 0.1985 | 2.3531 | 2.1546 | 0.0978 | 2.1918 | 2.0940 |
| | | | $\theta$ | -0.3332 | 1.8014 | 2.1346 | -0.2188 | 1.1938 | 1.4126 | 0.0845 | 1.1270 | 1.0425 | 0.1088 | 1.0030 | 0.8942 |
| | | | $\lambda$ | 0.2944 | 3.2627 | 2.9683 | 0.0168 | 2.6477 | 2.6309 | 0.5480 | 2.2588 | 1.7108 | 0.4897 | 2.1264 | 1.6367 |
| | | 3 | $\alpha$ | 0.8578 | 4.9839 | 4.1260 | 2.5943 | 3.3992 | 0.8049 | 1.3115 | 4.0312 | 2.7196 | 1.0097 | 3.9475 | 2.9379 |
| | | | $\theta$ | -0.4663 | 2.0514 | 2.5176 | -0.2431 | 1.2372 | 1.4803 | 0.0549 | 1.1549 | 1.1000 | 0.0793 | 1.0329 | 0.9535 |
| | | | $\lambda$ | 0.4303 | 3.0398 | 2.6095 | 0.1675 | 2.5076 | 2.3401 | 0.6478 | 2.2414 | 1.5936 | 0.5849 | 2.1280 | 1.5430 |
| | 1.5 | 0.5 | $\alpha$ | -0.6687 | 2.3269 | 2.9956 | -0.3277 | 1.3808 | 1.7085 | 0.0501 | 1.2236 | 1.1735 | 0.0627 | 1.0658 | 1.0031 |
| | | | $\theta$ | 0.1178 | 3.4447 | 3.3268 | 0.2730 | 2.6223 | 2.3493 | 0.4241 | 2.2340 | 1.8099 | 0.3661 | 2.0863 | 1.7202 |
| | | | $\lambda$ | 0.3690 | 2.8857 | 2.5167 | 0.0684 | 2.4337 | 2.3653 | 0.6155 | 2.1500 | 1.5345 | 0.5688 | 2.0184 | 1.4496 |
| | | 1.5 | $\alpha$ | -0.1520 | 3.9892 | 4.1412 | 0.7303 | 2.2682 | 1.5379 | 0.3290 | 2.3767 | 2.0476 | 0.2425 | 2.2143 | 1.9718 |
| | | | $\theta$ | -0.1647 | 3.8429 | 4.0076 | 0.0028 | 2.8053 | 2.8025 | 0.3874 | 2.3412 | 1.9539 | 0.3417 | 2.1701 | 1.8284 |
| | | | $\lambda$ | 0.6905 | 2.5797 | 1.8892 | 0.4884 | 2.2209 | 1.7325 | 0.7009 | 1.9499 | 1.2489 | 0.6224 | 1.8800 | 1.2576 |
| | | 3 | $\alpha$ | 0.3625 | 6.1622 | 5.7997 | 2.3361 | 3.6882 | 1.3521 | 1.3994 | 4.0013 | 2.6020 | 1.1209 | 3.9286 | 2.8078 |
| | | | $\theta$ | -0.3910 | 4.5035 | 4.8945 | -0.1811 | 3.0224 | 3.2035 | 0.4596 | 2.2088 | 1.7492 | 0.3928 | 2.0797 | 1.6869 |
| | | | $\lambda$ | 0.7583 | 2.4830 | 1.7247 | 0.5705 | 2.1440 | 1.5735 | 0.8245 | 1.9374 | 1.1129 | 0.7651 | 1.8737 | 1.1086 |
| | 3 | 0.5 | $\alpha$ | -0.9806 | 2.9184 | 3.8990 | -0.5275 | 1.5985 | 2.1261 | -0.0438 | 1.3336 | 1.3774 | 0.0013 | 1.1317 | 1.1304 |
| | | | $\theta$ | 1.2377 | 5.0848 | 3.8470 | 1.8234 | 4.0769 | 2.2535 | 1.4506 | 3.9167 | 2.4661 | 1.1560 | 3.8610 | 2.7050 |
| | | | $\lambda$ | 0.5923 | 2.5870 | 1.9946 | 0.3165 | 2.2544 | 1.9379 | 0.8230 | 2.0032 | 1.1802 | 0.7610 | 1.9227 | 1.1617 |
| | | 1.5 | $\alpha$ | -0.9886 | 5.3148 | 6.3033 | -0.2075 | 2.8293 | 3.0368 | 0.2708 | 2.4675 | 2.1967 | 0.2128 | 2.2585 | 2.0457 |
| | | | $\theta$ | 0.6277 | 6.2833 | 5.6557 | 1.4170 | 4.4336 | 3.0166 | 1.2945 | 3.9785 | 2.6840 | 1.0250 | 3.8681 | 2.8431 |
| | | | $\lambda$ | 0.8161 | 2.3526 | 1.5364 | 0.5950 | 2.0826 | 1.4876 | 0.8565 | 1.8621 | 1.0056 | 0.8019 | 1.8046 | 1.0027 |
| | | 3 | $\alpha$ | 0.2547 | 6.8300 | 6.5753 | 1.7246 | 4.0180 | 2.2933 | 1.3338 | 4.0549 | 2.7210 | 1.0840 | 3.9622 | 2.8781 |
| | | | $\theta$ | -0.0412 | 7.4860 | 7.5271 | 0.4755 | 4.9025 | 4.4270 | 1.4397 | 3.8171 | 2.3774 | 1.1806 | 3.7464 | 2.5657 |
| | | | $\lambda$ | 0.8854 | 2.2719 | 1.3865 | 0.7585 | 1.9615 | 1.2030 | 0.9602 | 1.8124 | 0.8522 | 0.9186 | 1.7663 | 0.8477 |
| 3 | | | $\alpha$ | -0.3982 | 7.9265 | 8.3247 | 1.7148 | 4.0117 | 2.2969 | 1.4737 | 4.0449 | 2.5712 | 1.2437 | 3.9261 | 2.6824 |
| | | | $\theta$ | -0.4460 | 8.2602 | 8.7062 | 0.5164 | 4.8636 | 4.3472 | 1.4178 | 3.9410 | 2.5232 | 1.1647 | 3.8338 | 2.6690 |
| | | | $\lambda$ | 1.7433 | 4.6034 | 2.8601 | 1.5358 | 3.9054 | 2.3696 | 1.9628 | 3.6126 | 1.6498 | 1.8242 | 3.5302 | 1.7061 |
| | | 1.5 | $\alpha$ | -1.4085 | 6.2428 | 7.6513 | -0.1771 | 2.8018 | 2.9789 | 0.2497 | 2.5138 | 2.2641 | 0.2016 | 2.3069 | 2.1053 |
| | | | $\theta$ | 0.1217 | 6.8545 | 6.7328 | 1.4359 | 4.4034 | 2.9675 | 1.4646 | 3.7652 | 2.3006 | 1.2311 | 3.6803 | 2.4492 |
| | | | $\lambda$ | 1.6064 | 4.7305 | 3.1241 | 1.2692 | 4.1083 | 2.8392 | 1.7903 | 3.6714 | 1.8810 | 1.6402 | 3.5800 | 1.9398 |

**Table 5.** 95% confidence intervals and Average Interval length for $\alpha, \theta, and\ \lambda$, with n=70

| $\lambda$ | $\theta$ | $\alpha$ | n=70 | MLE | | | MPS | | | SE | | | LINEX ($v = 1.5$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | L | U | AIL | L | U | AIL | L | U | AIL | L | U | AIL |
| 1.5 | 0.5 | 0.5 | $\alpha$ | -0.1221 | 1.1894 | 1.3115 | -0.1005 | 0.9527 | 1.0532 | 0.0188 | 1.0311 | 1.0124 | 0.0262 | 0.9656 | 0.9394 |
| | | | $\theta$ | -0.2076 | 1.6101 | 1.8178 | -0.1093 | 1.2543 | 1.3636 | 0.1760 | 0.9040 | 0.7280 | 0.1817 | 0.8613 | 0.6796 |
| | | | $\lambda$ | 0.4676 | 2.7630 | 2.2954 | 0.2531 | 2.3260 | 2.0729 | 0.8275 | 2.0448 | 1.2173 | 0.7805 | 2.0215 | 1.2410 |
| | | 1.5 | $\alpha$ | 0.5817 | 2.5430 | 1.9613 | 1.2670 | 1.7606 | 0.4936 | 0.7775 | 2.1272 | 1.3497 | 0.7164 | 2.0948 | 1.3784 |
| | | | $\theta$ | 0.0601 | 1.1169 | 1.0568 | 0.0424 | 0.9054 | 0.8630 | 0.2195 | 0.8717 | 0.6522 | 0.2204 | 0.8373 | 0.6170 |
| | | | $\lambda$ | 0.7164 | 2.5410 | 1.8245 | 0.5217 | 2.2631 | 1.7413 | 0.9281 | 2.0133 | 1.0852 | 0.9025 | 1.9740 | 1.0715 |
| | | 3 | $\alpha$ | 1.4329 | 4.5173 | 3.0844 | 2.7130 | 3.3126 | 0.5996 | 2.3029 | 3.6027 | 1.2998 | 2.2568 | 3.5692 | 1.3124 |
| | | | $\theta$ | -0.0246 | 1.2553 | 1.2798 | 0.0356 | 0.9250 | 0.8893 | 0.1800 | 0.8521 | 0.6721 | 0.1808 | 0.8201 | 0.6392 |
| | | | $\lambda$ | 0.7925 | 2.4301 | 1.6376 | 0.6225 | 2.1821 | 1.5596 | 0.9249 | 2.0075 | 1.0826 | 0.8856 | 1.9810 | 1.0954 |
| | 1.5 | 0.5 | $\alpha$ | -0.3842 | 1.7232 | 2.1074 | -0.1648 | 1.1079 | 1.2727 | 0.1026 | 0.9535 | 0.8509 | 0.1047 | 0.9007 | 0.7960 |
| | | | $\theta$ | 0.4829 | 2.8048 | 2.3218 | 0.6937 | 2.3150 | 1.6213 | 0.8930 | 2.0574 | 1.1644 | 0.8552 | 2.0236 | 1.1684 |
| | | | $\lambda$ | 0.6928 | 2.3809 | 1.6882 | 0.4551 | 2.1960 | 1.7409 | 0.9784 | 1.9410 | 0.9626 | 0.9481 | 1.9144 | 0.9664 |
| | | 1.5 | $\alpha$ | 0.2007 | 3.2964 | 3.0957 | 0.9716 | 2.0263 | 1.0547 | 0.7212 | 2.1673 | 1.4461 | 0.6736 | 2.1221 | 1.4485 |
| | | | $\theta$ | 0.3980 | 2.8763 | 2.4782 | 0.4780 | 2.3736 | 1.8957 | 0.9067 | 2.0409 | 1.1342 | 0.8816 | 1.9925 | 1.1109 |
| | | | $\lambda$ | 0.9529 | 2.1763 | 1.2234 | 0.8326 | 2.0130 | 1.1804 | 1.0593 | 1.8778 | 0.8185 | 1.0385 | 1.8541 | 0.8156 |
| | | 3 | $\alpha$ | 0.8771 | 5.3753 | 4.4982 | 2.5645 | 3.4737 | 0.9093 | 2.1967 | 3.6364 | 1.4397 | 2.1237 | 3.6038 | 1.4801 |
| | | | $\theta$ | 0.2477 | 3.2646 | 3.0169 | 0.3239 | 2.5493 | 2.2254 | 0.8900 | 2.0004 | 1.1105 | 0.8455 | 1.9730 | 1.1275 |
| | | | $\lambda$ | 0.9959 | 2.1259 | 1.1300 | 0.8850 | 1.9677 | 1.0827 | 1.0909 | 1.8337 | 0.7428 | 1.0669 | 1.8151 | 0.7481 |
| | 3 | 0.5 | $\alpha$ | -0.3870 | 1.7949 | 2.1819 | -0.2668 | 1.2531 | 1.5200 | 0.0700 | 1.0371 | 0.9671 | 0.0774 | 0.9710 | 0.8936 |
| | | | $\theta$ | 1.9630 | 4.1461 | 2.1831 | 2.3265 | 3.6698 | 1.3434 | 2.2411 | 3.5523 | 1.3113 | 2.1837 | 3.5206 | 1.3369 |
| | | | $\lambda$ | 0.8594 | 2.2130 | 1.3536 | 0.6518 | 2.0811 | 1.4293 | 1.0213 | 1.9079 | 0.8866 | 0.9933 | 1.8881 | 0.8948 |
| | | 1.5 | $\alpha$ | -0.0243 | 3.5307 | 3.5550 | 0.2681 | 2.5019 | 2.2338 | 0.7207 | 2.0742 | 1.3535 | 0.6664 | 2.0365 | 1.3701 |
| | | | $\theta$ | 1.5904 | 4.7831 | 3.1927 | 1.9043 | 3.9939 | 2.0896 | 2.2433 | 3.6152 | 1.3719 | 2.1870 | 3.5810 | 1.3940 |
| | | | $\lambda$ | 1.0468 | 2.0441 | 0.9973 | 0.9438 | 1.9072 | 0.9634 | 1.1532 | 1.7436 | 0.5904 | 1.1356 | 1.7237 | 0.5881 |
| | | | $\theta$ | 1.5340 | 4.8929 | 3.3589 | 1.9207 | 3.9674 | 2.0467 | 2.3237 | 3.5777 | 1.2540 | 2.2615 | 3.5554 | 1.2940 |
| | | | $\lambda$ | 2.0612 | 4.1119 | 2.0507 | 1.9090 | 3.7934 | 1.8844 | 2.4380 | 3.4304 | 0.9924 | 2.3947 | 3.4075 | 1.0128 |

**Table 6.** 95% confidence intervals and Average Interval length for $\alpha, \theta, and\ \lambda$, with n=200

| $\lambda$ | $\theta$ | $\alpha$ | n=200 | MLE | | | MPS | | | SE | | | LINEX ($\upsilon = 1.5$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | L | U | AIL | L | U | AIL | L | U | AIL | L | U | AIL |
| 1.5 | 0.5 | 0.5 | $\alpha$ | -0.0474 | 1.0444 | 1.0918 | -0.0423 | 0.8613 | 0.9036 | 0.2813 | 0.7205 | 0.4392 | 0.2773 | 0.7142 | 0.4369 |
| | | | $\theta$ | 0.0758 | 1.1333 | 1.0575 | 0.1171 | 1.0257 | 0.9086 | 0.3506 | 0.6677 | 0.3170 | 0.3491 | 0.6625 | 0.3134 |
| | | | $\lambda$ | 0.8623 | 2.1822 | 1.3198 | 0.7108 | 2.0155 | 1.3046 | 1.2741 | 1.6937 | 0.4196 | 1.2683 | 1.6898 | 0.4216 |
| | | 1.5 | $\alpha$ | 0.9343 | 2.0941 | 1.1598 | 1.3516 | 1.6720 | 0.3204 | 1.2235 | 1.7445 | 0.5210 | 1.2160 | 1.7392 | 0.5232 |
| | | | $\theta$ | 0.2482 | 0.8224 | 0.5742 | 0.2421 | 0.7207 | 0.4786 | 0.3602 | 0.6356 | 0.2754 | 0.3582 | 0.6313 | 0.2731 |
| | | | $\lambda$ | 1.0410 | 2.0647 | 1.0237 | 0.9461 | 1.9510 | 1.0049 | 1.2832 | 1.7086 | 0.4254 | 1.2777 | 1.7049 | 0.4272 |
| | | 3 | $\alpha$ | 1.8739 | 4.0479 | 2.1739 | 2.8074 | 3.2174 | 0.4099 | 2.7727 | 3.2001 | 0.4273 | 2.7659 | 3.1963 | 0.4304 |
| | | | $\theta$ | 0.2169 | 0.8750 | 0.6581 | 0.2425 | 0.7268 | 0.4843 | 0.3527 | 0.6417 | 0.2890 | 0.3508 | 0.6377 | 0.2869 |
| | | | $\lambda$ | 1.0836 | 2.0104 | 0.9268 | 1.0031 | 1.9064 | 0.9032 | 1.2794 | 1.7098 | 0.4304 | 1.2751 | 1.7044 | 0.4293 |
| | 1.5 | 0.5 | $\alpha$ | -0.0451 | 1.2072 | 1.2524 | 0.0466 | 0.8758 | 0.8292 | 0.2982 | 0.7004 | 0.4022 | 0.2951 | 0.6946 | 0.3996 |
| | | | $\theta$ | 0.5721 | 2.6402 | 2.0681 | 0.9867 | 2.0760 | 1.0893 | 1.2644 | 1.7159 | 0.4516 | 1.2599 | 1.7095 | 0.4495 |
| | | | $\lambda$ | 1.0143 | 1.9886 | 0.9742 | 0.9054 | 1.9290 | 1.0236 | 1.2903 | 1.6968 | 0.4065 | 1.2839 | 1.6940 | 0.4101 |
| | | 1.5 | $\alpha$ | 0.0633 | 3.3473 | 3.2840 | 1.1863 | 1.8051 | 0.6189 | 1.2238 | 1.7549 | 0.5311 | 1.2157 | 1.7515 | 0.5358 |
| | | | $\theta$ | 0.7470 | 2.3557 | 1.6086 | 0.9246 | 1.9990 | 1.0744 | 1.2344 | 1.7532 | 0.5188 | 1.2278 | 1.7482 | 0.5204 |
| | | | $\lambda$ | 1.1770 | 1.8718 | 0.6948 | 1.1222 | 1.8135 | 0.6913 | 1.3046 | 1.6819 | 0.3772 | 1.3006 | 1.6772 | 0.3767 |
| | | 3 | $\alpha$ | 0.9082 | 5.0440 | 4.1358 | 2.7353 | 3.2717 | 0.5364 | 2.7297 | 3.2409 | 0.5112 | 2.7223 | 3.2367 | 0.5145 |
| | | | $\theta$ | 0.5858 | 2.7560 | 2.1702 | 0.8252 | 2.1053 | 1.2801 | 1.2441 | 1.7105 | 0.4665 | 1.2366 | 1.7064 | 0.4698 |
| | | | $\lambda$ | 1.1981 | 1.8563 | 0.6581 | 1.1494 | 1.7902 | 0.6408 | 1.3176 | 1.6489 | 0.3313 | 1.3123 | 1.6464 | 0.3341 |
| | 3 | 0.5 | $\alpha$ | -0.0692 | 1.2622 | 1.3314 | 0.0536 | 0.9188 | 0.8651 | 0.2951 | 0.7027 | 0.4076 | 0.2927 | 0.6961 | 0.4034 |
| | | | $\theta$ | 2.0302 | 3.9673 | 1.9371 | 2.6254 | 3.3834 | 0.7580 | 2.7593 | 3.2140 | 0.4547 | 2.7524 | 3.2095 | 0.4571 |
| | | | $\lambda$ | 1.1380 | 1.8949 | 0.7569 | 1.0534 | 1.8470 | 0.7936 | 1.3280 | 1.6709 | 0.3429 | 1.3247 | 1.6658 | 0.3410 |
| | | 1.5 | $\alpha$ | 0.5647 | 2.6084 | 2.0437 | 0.7906 | 2.1106 | 1.3200 | 1.2407 | 1.7352 | 0.4946 | 1.2341 | 1.7304 | 0.4963 |
| | | | $\theta$ | 2.1961 | 3.9628 | 1.7668 | 2.3378 | 3.5981 | 1.2603 | 2.6968 | 3.2200 | 0.5232 | 2.6839 | 3.2194 | 0.5355 |
| | | | $\lambda$ | 1.2333 | 1.8076 | 0.5743 | 1.1939 | 1.7498 | 0.5559 | 1.3316 | 1.6369 | 0.3053 | 1.3275 | 1.6334 | 0.3059 |
| | | | $\theta$ | 1.5271 | 4.5387 | 3.0116 | 2.3497 | 3.5835 | 1.2338 | 2.7410 | 3.2402 | 0.4993 | 2.7333 | 3.2356 | 0.5023 |
| | | | $\lambda$ | 2.4584 | 3.6040 | 1.1456 | 2.3951 | 3.4913 | 1.0962 | 2.7670 | 3.2109 | 0.4439 | 2.7608 | 3.2069 | 0.4461 |

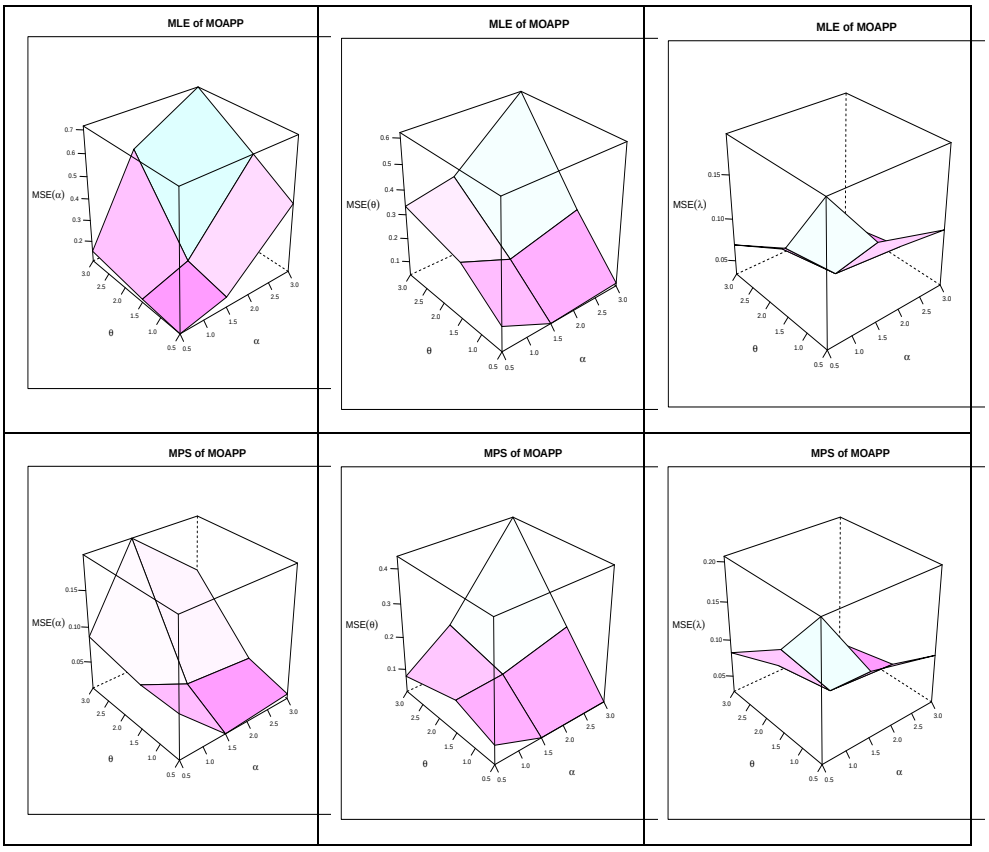



**Figure 3.** MSE for MLE, MPS and Bayes estimation under SE and LINEX Loss Functions for n= 120.

## 6.2. Data Analysis

In this section we take three different examples of real-life data set. The MLEs estimates of the parameters are reported in Tables (9), (10) and (11), then the MOAPP model is compared with other special case models like Pareto type 1, generalized Pareto (GP), and alpha power Pareto (APP). This comparison was conducted using Kolmogorov–Smirnov (KS) distance (D) between the fitted and the empirical distribution functions and the corresponding p-values. Also, Akaike information criterion (AIC) such that AIC=-2 L(γ)+2p, where p is the number of parameters in the model and L is the maximized value of the likelihood function for the model. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Bayesian information criterion (BIC) is also used for comparison between models, where BIC can be defined as: BIC=-2 L(γ)+p ln(n), where n is the sample size. As a model selection criterion, the researcher must choose the model with the minimum BIC value. The MLEs of $\alpha, \theta, and\ \lambda$ are computed numerically using the

function optimal in R statistical package. The values of the KS statistic with p-values, AIC and BIC are reported in Tables (7), (8) and (9).

The first example is from Lawless (1982). The data set consists of failure times or censoring times for 36 appliances subjected to an automated life test. Failures are mainly classified into 18 different modes, although among 33 observed failures only 7 modes are present and only model 6 and 9 appear more than once. We are mainly interested in the failure mode 9. The data are given below:

**Data 1**: 1167, 1925, 1990, 2223, 2400, 2471, 2551, 2568, 2694, 3034, 3112, 3214, 3478, 3504, 4329, 176976, 7846.

**Table 7.** MLE estimation with KS, p-values and different model goodness of fit criterion for data 1

| | $\hat{\alpha}_{MLE}$ | $\hat{\theta}_{MLE}$ | $\hat{\lambda}_{MLE}$ | D | P-value | AIC | BIC |
|---|---|---|---|---|---|---|---|
| P | - | - | 0.12231 | 0.57843 | 6.34E-06 | 385.4278 | 3.86E+02 |
| GP | - | 3341.032 | 0.609273 | 0.33089 | 0.03686 | 334.5826 | 336.2491 |
| APP | 78.74852 | - | 0.257025 | 0.48467 | 0.00034 | 367.7567 | 3.69E+02 |
| MOAPP | 9.00E+07 | 4.60E+07 | 2.57536 | 0.2088 | 0.3941 | 324.6243 | 327.124 |

The second example represents survival times of guinea pigs injected with different amount of tubercle bacilli studied by Bjerkedal [1960]. Guinea pigs are subject to high susceptibility of human tuberculosis, which is one of the causes for choosing this species.

**Table 8.** MLE estimation with KS, p-values and different model goodness of fit criterion for data 2

| | $\hat{\alpha}_{MLE}$ | $\hat{\theta}_{MLE}$ | $\hat{\lambda}_{MLE}$ | D | P-value | AIC | BIC |
|---|---|---|---|---|---|---|---|
| P | - | - | 0.199805 | 0.51093 | 2.2E-16 | 1098.51 | 1.10E+03 |
| GP | - | 237.9788 | 0.393478 | 0.23142 | 0.000895 | 879.3132 | 883.8665 |
| APP | 152.982 | - | 0.446334 | 0.40147 | 1.67E-10 | 1009.978 | 1014.531 |
| MOAPP | 112100 | 322998.1 | 3.011837 | 0.06837 | 0.8894 | 857.2212 | 864.0512 |

The third example is from Almetwally et al. (2019). The data set consists of economic data of 31 observations subjected to a GDP growth of Egypt. The data are given below.

**Table 9.** MLE estimation with KS, p-values and different model goodness of fit criterion for data 3

| | $\hat{\alpha}_{MLE}$ | $\hat{\theta}_{MLE}$ | $\hat{\lambda}_{MLE}$ | D | P-value | AIC | BIC |
|---|---|---|---|---|---|---|---|
| P | - | - | 0.6473 | 0.3956 | 0.0001 | 186.7626 | 188.1966 |
| GP | - | 6.8839 | -0.6607 | 0.2910 | 0.0080 | 149.8744 | 152.7423 |
| APP | 90.2664 | - | 1.3679 | 0.2501 | 0.0340 | 160.3021 | 163.1700 |
| MOAPP | 8.5373 | 153.5946 | 3.7857 | 0.0726 | 0.9927 | 139.4962 | 143.7982 |

When comparing the values of KS statistics of MOAPP and other sub models like Pareto type 1, GP and APP for the two data examples above, we obtain the minimum KS for MOAPP with highest p-values. Also, it can be noticed that the values of AIC and BIC take their minimum values when the distribution is MOAPP. Therefore, this indicates that the MOAPP distribution fits the two sets of data very well and is better than other distributions. This also emphasizes the need of new distributions in managing real-life data. So, in general we can say that the new distribution is superior according to other sub models.

## 7. Conclusions

In this study we have considered MOAPP distribution which has three unknown parameters. This new distribution proved to be more flexible and more appropriate for monotone and right skewed lifetime data, also its hazard rate function can be either a decreasing or upside-down bathtub curve. We estimate the parameters of MOAPP using MLE, MPS and Bayesian method under SE and LINEX loss functions. It is not possible to compare different methods theoretically, so we have used some simulations to compare different estimators. We have compared different estimators mainly with respect to biases and mean squared errors. Confidence intervals are obtained and are compared numerically in terms of interval lengths. The best method for estimating $\alpha$ and $\theta$ is the Bayesian method under the LINEX and SE loss functions depending on the values of $\alpha$ and $\theta$, it is also noticed that the MPS method acts better for estimating $\alpha$ and $\theta$ than the MLE method. The Bayesian method under the SE loss function is the best appropriate method for estimating $\lambda$. Confidence intervals under the MPS method and the Bayesian credible interval are preferable to confidence intervals under the MLE method. Therefore, we recommend the use of the MPS and Bayes estimation methods for practical purposes. The flexibility of this distribution was illustrated in some applications to real data sets, where the new model proves to better fit data than some other sub models.

## References

AHMAD, H. H., ALMETWALLY, E., (2020). Marshall-Olkin Generalized Pareto Distribution: Bayesian and Non Bayesian Estimation. Pakistan Journal of Statistics and Operation Research, 16(1), pp. 21–33; DOI: 10.18187/pjsor.v16i1.2935.

ALMETWALLY, E. M., ALMONGY, H. M. and EL-SHERPIENY, E. A., (2019). Adaptive Type-II Progressive Censoring Schemes based on Maximum Product Spacing with Application of Generalized Rayleigh Distribution. Journal of Data Science, 17(4), pp. 767–793.

ALMETWALLY, E. M., ALMONGY, H. M., ($2019_b$). Estimation method for new Weibull-Pareto distribution: Simulation and application. Journal of Data Science, 17(3), pp. 610–630.

ALMETWALLY, E. M., ALMONGY, H. M., ($2019_a$). Maximum Product Spacing and Bayesian Method for Parameter Estimation for Generalized Power Weibull Distribution under Censoring Scheme. Journal of Data Science, 17(2), pp. 407–444.

ALMETWALLY, E. M., ALMONGY, H. M. and EL SAYED MUBARAK, A., (2018). Bayesian and Maximum Likelihood Estimation for the Weibull Generalized Exponential Distribution Parameters Using Progressive Censoring Schemes. Pakistan Journal of Statistics and Operation Research, 14(4), pp. 853–868.

ANATOLYEV, S., KOSENOK, G., (2005). An Alternative to Maximum Likelihood Based on Spacings, Econometric Theory 21(02), pp. 472–476.

BDAIR, O., HAJ AHMAD, H., (2019). Estimation of The Marshall-Olkin Pareto Distribution Parameters: Comparative Study, Revista Investigacion Operacional, 41(2), forthcoming.

BJERKEDAL, T., (1960). Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli, Am J Hyg, 72(1), pp. 130–148.

CHENG R. C. H.; AMIN, N. A. K., (1979). product-of-spacings estimation with applications to the lognormal distribution, University of Wales IST, Math Report 79–1.

CHENG, R. C. H., AMIN, N. A. K., (1983). Estimating parameters in continuous univariate distributions with a shifted origin, J. Roy. Statist. Soc. Ser. B, 45, pp. 394–403.

GHITANY, M. E., (2005). Marshall-Olkin extended Pareto distribution and its application, International Journal of Applied Mathematics, 18, No. 1, pp. 17–31.

GHITANY, M., E., KOTZ, S., (2007). Reliability Properties of Extended Linear Failure-Rate Distributions, Probability in the Engineering and informational Sciences 21, pp. 441–450.

GHOSH, K., JAMMALAMADAKA, S. R., (2001). A general estimation method using spacings. Journal of Statistical Planning and Inference, 93, pp. 71–82.

HAJ AHMAD, H., BDAIR, O., AHSANULLAH, M., (2017). On Marshall-Olkin Extended Weibull Distribution, Journal of Statistical Theory and Applications, Vol. 16, No. 1, pp 1–17.

JOSE, K. K., ALICE, T., (2001). Marshall-Olkin Generalized Weibull distributions and applications, STARS: int. Journal, 2,1, pp. 1–8.

JOSE, K. K., ALICE, T., (2005). Marshall-Olkin Family of Distributions: Applications in Time series modelling and Reliability, J.C Publications, Palakkad.

JOSE, K. K., UMA, P., (2009). On Marshall-Olkin Mittag-Leffler distributions and processes, Far East Journal of Theoretical Statistics, 28, pp. 189–199.

KARIAN, Z. A., DUDEWICZ, E. J., (1998). Modern statistical, systems, and GPSS simulation, CRC press.

KARANDIKAR, R. L., (2006). On the Markov chain Monte Carlo (MCMC) method. Sadhana, 31(2), pp. 81–104.

LAWLESS, J. F., (1982). Statistical Models and Methods for Lifetime Data. John Wiley & Sons, New York.

MAHDAVI, A., KUNDU, D., (2017). A new method for generating distributions with an application to exponential distribution. Communications in Statistics-Theory and Methods, 46(13), pp. 6543–6557.

MARSHALL, A. W., OLKIN, I., (1997). A New Method for Adding a Parameter to a Family of Distributions with Application to the Exponential and Weibull Families, Biometrika, 84(3), pp. 641–652.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E., (1953). Equation of state calculations by fast computing machines. The journal of chemical physics, 21(6), pp. 1087–1092.

NASSAR, M., KUMAR, D., SANKU DEY, S., GAUSS M. CORDEIRO, G. and AFIFY, A. Z., (2019). The Marshall–Olkin alpha power family of distributions with applications, Journal of Computational and Applied Mathematics, 351, pp. 41–53.

ROBERT, C., CASELLA, G., (2013). Monte Carlo statistical methods. Springer Science & Business Media.

SINGH, U., SINGH, S. K. and SINGH, R. K., (2014). A comparative study of traditional estimation methods and maximum product spacing method in generalized inverted exponential distribution. Journal of Statistics Applications & Probability, 3(2), p. 153.

KUMAR SINGH, R., KUMAR SINGH, S. and SINGH, U., (2016). Maximum product spacings method for the estimation of parameters of generalized inverted exponential distribution under Progressive Type II Censoring. Journal of Statistics and Management Systems, 19(2), pp. 219–245.

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 85–98, DOI 10.21307/stattrans-2020-056*
Received – 13.03.2018; accepted – 11.07.2020

# Some linear regression type ratio exponential estimators for estimating the population mean based on quartile deviation and deciles

**Shakti Prasad**[1]

## ABSTRACT

This paper deals some linear regression type ratio exponential estimators for estimating the population mean using the known values of quartile deviation and deciles of an auxiliary variable in survey sampling. The expressions of the bias and the mean square error of the suggested estimators have been derived. It was compared with the usual mean, usual ratio (Cochran (1977)), Kadilar and Cingi (2004, 2006) and Subzar *et al.* (2017) estimators. After comparison, the condition which makes the suggested estimators more efficient than others is found. To verify the theoretical results, numerical results are performed on two natural population data sets.

**Key words:** Bias, Mean square error (MSE), Auxiliary variable, Relative Efficiency $(\%)$.

## 1. Introduction

Cochran (1977) considered a classical ratio type estimator for the estimation of finite population mean by using auxiliary information when the coefficient of correlation between auxiliary variable $X$ and the estimated variable $Y$ is positive. Sisodia and Dwivedi (1981) utilized the coefficient of variation of the auxiliary variable in survey sampling. Upadhyaya and Singh (1999) modified ratio type estimators using the coefficient of variation and the coefficient of kurtosis of the auxiliary variable. Yan and Tian (2010), Subramani and Kumarapandian (2012 (a), 2012 (b), 2012 (c), 2012 (d)), Swain (2014) and Abid *et al.* (2016 (a), 2016 (b), 2016 (c)) etc, considered a large number of modified ratio estimators using the known values of population parameters of auxiliary variable in survey sampling. Recently Subzar *et al.* (2017) considered new ratio type estimators in simple random sampling by using the conventional location parameters.

The paper is structured as follows. In Section 2, the existing and studied so far linear regression type ratio estimators are presented. In Section 3, newly proposed classes of estimators are formally presented. The properties of the suggested estimators are discussed in Section 4. The theoretical comparisons between the suggested estimators and the other existing estimators are considered in Section 5. A numerical demonstration is conducted in Section 6 to support and verify the theoretical results and some concluding remarks are given in Section 7.

[1]Department of Basic and Applied Science, National Institute of Technology, Arunachal Pradesh, Yupia, Papumpare-791112, India. E-mail: shakti.pd@gmail.com. ORCID: https://orcid.org/0000-0002-7867-7586.

## 2. Brief Description of Some Existing Estimators

Let $y$ and $x$ be denoted by the positively correlated study variable and auxiliary variable respectively. A simple random sample (without replacement) $s_n$ of $n$ units is drawn from a finite population $U = (U_1, U_2, ..., U_N)$ of $N$ units to estimate population mean $\bar{Y}$, which uses the known values of population parameters of auxiliary variable such as quartile deviation and deciles. The following notations have been approached in this work:
$\bar{Y}, \bar{X}$: The population means of the variables $y$ and $x$ respectively.
$S_x^2 = (N-1)^{-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$ : The population variance of the variable $x$.
$S_y^2$ : The population variance of the variable $y$.
$S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})$: The population covariance between the variables $y$ and $x$.
$C_y$ and $C_x$ : The coefficients of variation of the variables $y$ and $x$ respectively.
$\beta_1(x)$ : The population coefficient of skewness of the variable $x$.
$\beta_2(x)$ : The population coefficient of kurtosis of the variable $x$.
$\rho$ : The Pearson correlation coefficient between the variables $y$ and $x$.
$D_i, i = 1,2,...,10$ : Population Deciles.
$QD = \frac{Q_3 - Q_1}{2}$ : Population quartile deviation.
In this section, several ratio type estimators have been considered for estimating the population mean in survey sampling:

### 2.1. Usual Mean Estimator

The estimator of sample mean $\bar{y}$ is derived as $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, which is known as the usual unbiased estimator $\bar{y}$ of population mean $\bar{Y}$. The variance of the sample mean $\bar{y}$, is given by $Var(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2 C_y^2$.

### 2.2. Usual Ratio (Cochran (1977)) Estimator

Cochran (1977) considered the ratio estimator of population mean $\bar{Y}$ as $\bar{y}_{Ratio} = \bar{y}\frac{\bar{X}}{\bar{x}}, (\bar{x} \neq 0)$. The MSE of estimator $\bar{y}_{Ratio}$, is given by
$MSE(\bar{y}_{Ratio}) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left(C_y^2 - 2\rho C_y C_x + C_x^2\right)$.

### 2.3. Kadilar and Cingi (2004) Estimators

Kadilar and Cingi (2004) considered the following ratio estimators for the population mean of the study variable $\bar{Y}$ by using auxiliary information in survey sampling:

$T_{KC(1)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}}\bar{X}$, $T_{KC(2)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}+C_x}(\bar{X}+C_x)$,

$T_{KC(3)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}+\beta_2(x)}(\bar{X}+\beta_2(x))$, $T_{KC(4)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}\beta_2(x)+C_x}(\bar{X}\beta_2(x)+C_x)$,

$T_{KC(5)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}C_x+\beta_2(x)}(\bar{X}C_x+\beta_2(x))$,

where $\hat{\beta} = \frac{s_{yx}}{s_x^2}$, $s_{yx} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$, $s_x^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
The MSEs of the estimators $T_{KC(1)}$, $T_{KC(2)}$, $T_{KC(3)}$, $T_{KC(4)}$ and $T_{KC(5)}$, are given by
$MSE(T_{KC(i)}) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(KC_i^2 C_x^2 + \left(1-\rho^2\right)C_y^2\right)\bar{Y}^2$, where $(i = 1,2,...,5)$,
$KC_1 = 1$, $KC_2 = \frac{\bar{X}}{\bar{X}+C_x}$, $KC_3 = \frac{\bar{X}}{\bar{X}+\beta_2(x)}$, $KC_4 = \frac{\bar{X}\beta_2(x)}{\bar{X}\beta_2(x)+C_x}$, $KC_5 = \frac{C_x\bar{X}}{C_x\bar{X}+\beta_2(x)}$.

### 2.4. Kadilar and Cingi (2006) Estimators

Kadilar and Cingi (2006) considered the following ratio estimators for the population mean of the study variable $\bar{Y}$ by using the coefficient of correlation in survey sampling:

$T_{KC(6)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}+\rho}(\bar{X}+\rho)$, $T_{KC(7)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}C_x+\rho}(\bar{X}C_x+\rho)$,

$T_{KC(8)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}\rho+C_x}(\bar{X}\rho+C_x)$, $T_{KC(9)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}\beta_2(x)+\rho}(\bar{X}\beta_2(x)+\rho)$,

$T_{KC(10)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}\rho+\beta_2(x)}(\bar{X}\rho+\beta_2(x))$.

The MSEs of the estimators $T_{KC(6)}$, $T_{KC(7)}$, $T_{KC(8)}$, $T_{KC(9)}$ and $T_{KC(10)}$, are given by

$MSE(T_{KC(i)}) = \left(\frac{1}{n}-\frac{1}{N}\right)\left(KC_i^2C_x^2+\left(1-\rho^2\right)C_y^2\right)\bar{Y}^2$, where $(i=6,7,...,10)$,

$KC_6 = \frac{\bar{X}}{\bar{X}+\rho}$, $KC_7 = \frac{\bar{X}C_x}{\bar{X}C_x+\rho}$, $KC_8 = \frac{\bar{X}\rho}{\bar{X}\rho+C_x}$, $KC_9 = \frac{\bar{X}\beta_2(x)}{\bar{X}\beta_2(x)+\rho}$, $KC_{10} = \frac{\rho\bar{X}}{\bar{X}\rho+\beta_2(x)}$.

### 2.5. Subzar *et al.* (2017) Estimators

Subzar *et al.* (2017) proposed a new ratio estimators for estimation of the population mean $\bar{Y}$ as

$T_{Smrs(1)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}QD+D_1}(\bar{X}QD+D_1)$, $T_{Smrs(2)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}QD+D_2}(\bar{X}QD+D_2)$,

$T_{Smrs(3)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}QD+D_3}(\bar{X}QD+D_3)$, $T_{Smrs(4)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}QD+D_4}(\bar{X}QD+D_4)$,

$T_{Smrs(5)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}QD+D_5}(\bar{X}QD+D_5)$, $T_{Smrs(6)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}QD+D_6}(\bar{X}QD+D_6)$,

$T_{Smrs(7)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}QD+D_7}(\bar{X}QD+D_7)$, $T_{Smrs(8)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}QD+D_8}(\bar{X}QD+D_8)$,

$T_{Smrs(9)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}QD+D_9}(\bar{X}QD+D_9)$, $T_{Smrs(10)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}QD+D_{10}}(\bar{X}QD+D_{10})$,

$T_{Smrs(11)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}D_1+QD}(\bar{X}D_1+QD)$, $T_{Smrs(12)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}D_2+QD}(\bar{X}D_2+QD)$,

$T_{Smrs(13)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}D_3+QD}(\bar{X}D_3+QD)$, $T_{Smrs(14)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}D_4+QD}(\bar{X}D_4+QD)$,

$T_{Smrs(15)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}D_5+QD}(\bar{X}D_5+QD)$, $T_{Smrs(16)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}D_6+QD}(\bar{X}D_6+QD)$,

$T_{Smrs(17)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}D_7+QD}(\bar{X}D_7+QD)$, $T_{Smrs(18)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}D_8+QD}(\bar{X}D_8+QD)$,

$T_{Smrs(19)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}D_9+QD}(\bar{X}D_9+QD)$, $T_{Smrs(20)} = \frac{\bar{y}+\hat{\beta}(\bar{X}-\bar{x})}{\bar{x}D_{10}+QD}(\bar{X}D_{10}+QD)$.

The MSEs of the estimators $T_{Smrs(i)}$ $(i=1,2,...,20)$, are given by

$MSE(T_{Smrs(i)}) = \left(\frac{1}{n}-\frac{1}{N}\right)\left(\alpha_i^2C_x^2+\left(1-\rho^2\right)C_y^2\right)\bar{Y}^2$, where $(i=1,2,...,20)$,

$\alpha_1 = \frac{QD\bar{X}}{QD\bar{X}+D_1}$, $\alpha_2 = \frac{QD\bar{X}}{QD\bar{X}+D_2}$, $\alpha_3 = \frac{QD\bar{X}}{QD\bar{X}+D_3}$, $\alpha_4 = \frac{QD\bar{X}}{QD\bar{X}+D_4}$, $\alpha_5 = \frac{QD\bar{X}}{QD\bar{X}+D_5}$, $\alpha_6 = \frac{QD\bar{X}}{QD\bar{X}+D_6}$,

$\alpha_7 = \frac{QD\bar{X}}{QD\bar{X}+D_7}$, $\alpha_8 = \frac{QD\bar{X}}{QD\bar{X}+D_8}$, $\alpha_9 = \frac{QD\bar{X}}{QD\bar{X}+D_9}$, $\alpha_{10} = \frac{QD\bar{X}}{QD\bar{X}+D_{10}}$, $\alpha_{11} = \frac{D_1\bar{X}}{D_1\bar{X}+QD}$, $\alpha_{12} = \frac{D_2\bar{X}}{D_2\bar{X}+QD}$,

$\alpha_{13} = \frac{D_3\bar{X}}{D_3\bar{X}+QD}$, $\alpha_{14} = \frac{D_4\bar{X}}{D_4\bar{X}+QD}$, $\alpha_{15} = \frac{D_5\bar{X}}{D_5\bar{X}+QD}$, $\alpha_{16} = \frac{D_6\bar{X}}{D_6\bar{X}+QD}$, $\alpha_{17} = \frac{D_7\bar{X}}{D_7\bar{X}+QD}$, $\alpha_{18} = \frac{D_8\bar{X}}{D_8\bar{X}+QD}$, $\alpha_{19} = \frac{D_9\bar{X}}{D_9\bar{X}+QD}$, $\alpha_{20} = \frac{D_{10}\bar{X}}{D_{10}\bar{X}+QD}$.

Motivated by the work of Subzar *et al.* (2017), we suggest some linear regression type ratio exponential estimators for estimating population mean of the study variable $\bar{Y}$ using quartile deviation and deciles of auxiliary variable in survey sampling. We have also formulated the condition which makes the suggested classes of estimators more efficient than others and have shown that under this condition they are really more efficient than the existing estimators on the basis of numerical results in this literature.

## 3. Mathematical Formulation of Suggested Classes of Linear Regression Type Ratio Exponential Estimators

We suggest two classes (Class A and B) of efficient linear regression type ratio exponential estimators for estimating the population mean $\bar{Y}$ using the known population values of quartile deviation and deciles of auxiliary variable.

### 3.1. The First Suggested Class of Linear Regression Type Ratio Exponential Estimators (Class A)

The first suggested estimators $T_{SP1(j)}(j=1,2,...,10)$ considered the linear regression type ratio exponential estimators for estimating population mean of the study variable $\bar{Y}$ by using the linear combination of known population values of quartile deviation $(QD)$ and deciles $(D_j(j=1,2,...,10))$ of an auxiliary variable in survey sampling:

$$T_{SP1(j)} = \left[\bar{y}+\hat{\beta}\,(\bar{X}-\bar{x})\right]\exp\left[\Phi_j\frac{1-\frac{\bar{x}}{\bar{X}}}{1+\frac{QD\bar{x}+D_j}{QD\bar{X}+D_j}}\right], \tag{1}$$

where $\hat{\beta}=\frac{s_{yx}}{s_x^2}$, $s_{yx}=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$, $s_x^2=(n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$, and $\Phi_j=\frac{QD\bar{X}}{QD\bar{X}+D_j}$,(j = 1, 2, ..., 10).

### 3.2. The Second Suggested Class of Linear Regression Type Ratio Exponential Estimators (Class B)

The second suggested estimators $T_{SP2(j)}(j=1,2,...,10)$ considered the linear regression type ratio exponential estimators for estimating population mean of the study variable $\bar{Y}$ by using the linear combination of known population values of deciles $(D_j(j=1,2,...,10))$ and quartile deviation $(QD)$ of an auxiliary variable in survey sampling:

$$T_{SP2(j)} = \left[\bar{y}+\hat{\beta}\,(\bar{X}-\bar{x})\right]\exp\left[\Psi_j\frac{1-\frac{\bar{x}}{\bar{X}}}{1+\frac{D_j\bar{x}+QD}{D_j\bar{X}+QD}}\right], \tag{2}$$

where $\hat{\beta}=\frac{s_{yx}}{s_x^2}$, $s_{yx}=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$, $s_x^2=(n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$, and $\Psi_j=\frac{D_j\bar{X}}{D_j\bar{X}+QD}$, (j = 1, 2, ..., 10).
We get some members of the family of suggested estimators $T_{SP1(j)}$ and $T_{SP2(j)}$ in Table 1.

## 4. Behaviours of the suggested estimators $T_{SP1(j)}$ and $T_{SP2(j)}$ $(j=1,2,...,10)$

To obtain the bias and mean square error (MSE) of the suggested estimators $T_{SP1(j)}$ and $T_{SP2(j)}$, $(j=1,2,...,10)$ up-to the first order of large sample approximations are derived under the following transformations:
$\bar{y}=\bar{Y}(1+e_0)$, $\bar{x}=\bar{X}(1+e_1)$, $s_{yx}=S_{yx}(1+e_2)$, and $s_x^2=S_x^2(1+e_3)$ such that $E(e_j)=$

$0,\ |e_j|<1\forall\ j=0,1,2,3.,\ E(e_0^2)=(\frac{1}{n}-\frac{1}{N})C_y^2,\ E(e_1^2)=(\frac{1}{n}-\frac{1}{N})C_x^2,\ E(e_2^2)=(\frac{1}{n}-\frac{1}{N})\frac{\mu_{22}}{S_{yx}^2}$,
$E(e_0e_1)=(\frac{1}{n}-\frac{1}{N})\rho_{yx}C_yC_x,\ E(e_1e_2)=(\frac{1}{n}-\frac{1}{N})\frac{\mu_{12}}{\bar{X}S_{yx}},\ E(e_1e_3)=(\frac{1}{n}-\frac{1}{N})\frac{\mu_{03}}{\bar{X}S_x^2}$,
where $\mu_{rs}=E[(y-\bar{Y})^r(x-\bar{X})^s]$, $r$ and $s$ being positive integers.
Under the above transformations, expressing the equations "(1) and (2)" in terms of e's, we have

$$\begin{aligned} T_{SP1(j)} &= \left[\bar{y}+\frac{s_{yx}}{s_x^2}(\bar{X}-\bar{x})\right]\exp\left[\Phi_j\frac{1-\frac{\bar{x}}{\bar{X}}}{1+\frac{QD\bar{x}+D_j}{QD\bar{X}+D_j}}\right], \\ &= \{\bar{Y}(1+e_0)-\bar{X}\beta e_1(1+e_2)(1+e_3)^{-1}\}\exp\left[-\frac{1}{2}\Phi_j e_1\left(1+\frac{1}{2}\Phi_j e_1\right)^{-1}\right], \end{aligned} \quad (3)$$

$$\begin{aligned} T_{SP2(j)} &= \left[\bar{y}+\frac{s_{yx}}{s_x^2}(\bar{X}-\bar{x})\right]\exp\left[\Psi_j\frac{1-\frac{\bar{x}}{\bar{X}}}{1+\frac{D_j\bar{x}+QD}{D_j\bar{X}+QD}}\right], \\ &= \{\bar{Y}(1+e_0)-\bar{X}\beta e_1(1+e_2)(1+e_3)^{-1}\}\exp\left[-\frac{1}{2}\Psi_j e_1\left(1+\frac{1}{2}\Psi_j e_1\right)^{-1}\right], \end{aligned} \quad (4)$$

where $\beta=\frac{S_{yx}}{S_x^2}$.
Expanding the right side of "(3) and (4)", multiplying and neglecting the terms of $e's$ having power greater than two, we get

$$T_{SP1(j)}\cong\bar{Y}\left[1+e_0-\frac{1}{2}\Phi_j e_1+\frac{3}{8}\Phi_j^2e_1^2-\frac{1}{2}\Phi_j e_0e_1-\frac{\bar{X}\beta}{\bar{Y}}\left(e_1-\frac{1}{2}\Phi_j e_1^2+e_1e_2-e_1e_3\right)\right]. \quad (5)$$

$$T_{SP2(j)}\cong\bar{Y}\left[1+e_0-\frac{1}{2}\Psi_j e_1+\frac{3}{8}\Psi_j^2e_1^2-\frac{1}{2}\Psi_j e_0e_1-\frac{\bar{X}\beta}{\bar{Y}}\left(e_1-\frac{1}{2}\Psi_j e_1^2+e_1e_2-e_1e_3\right)\right]. \quad (6)$$

or

$$T_{SP1(j)}-\bar{Y}\cong\bar{Y}\left[e_0-\frac{1}{2}\Phi_j e_1+\frac{3}{8}\Phi_j^2e_1^2-\frac{1}{2}\Phi_j e_0e_1-B\left(e_1-\frac{1}{2}\Phi_j e_1^2+e_1e_2-e_1e_3\right)\right], \quad (7)$$

$$T_{SP2(j)}-\bar{Y}\cong\bar{Y}\left[e_0-\frac{1}{2}\Psi_j e_1+\frac{3}{8}\Psi_j^2e_1^2-\frac{1}{2}\Psi_j e_0e_1-B\left(e_1-\frac{1}{2}\Psi_j e_1^2+e_1e_2-e_1e_3\right)\right], \quad (8)$$

where $B=\rho\frac{C_y}{C_x}$.
Taking expectation of both sides of equations "(7) and (8)", we get the biases of the sug-

gested estimators up-to first order of large approximations as

$$\begin{aligned} Bias[T_{SP1(j)}] &= E\left[T_{SP1(j)}-\bar{Y}\right], \\ &= \bar{Y}E\left[e_0-\frac{1}{2}\Phi_j e_1+\frac{3}{8}\Phi_j^2 e_1^2-\frac{1}{2}\Phi_j e_0 e_1-B\left(e_1-\frac{1}{2}\Phi_j e_1^2+e_1e_2-e_1e_3\right)\right], \\ &= \left(\frac{1}{n}-\frac{1}{N}\right)\bar{Y}\left(\frac{3}{8}\Phi_j^2 C_x^2+\frac{B}{\bar{X}^2C_x}\{\frac{\mu_{03}}{\bar{X}C_x}-\frac{\mu_{12}}{\bar{Y}\rho C_y}\}\right). \end{aligned} \tag{9}$$

$$\begin{aligned} Bias[T_{SP2(j)}] &= E\left[T_{SP2(j)}-\bar{Y}\right], \\ &= \bar{Y}\left[e_0-\frac{1}{2}\Psi_j e_1+\frac{3}{8}\Psi_j^2 e_1^2-\frac{1}{2}\Psi_j e_0 e_1-B\left(e_1-\frac{1}{2}\Psi_j e_1^2+e_1e_2-e_1e_3\right)\right], \\ &= \left(\frac{1}{n}-\frac{1}{N}\right)\bar{Y}\left(\frac{3}{8}\Psi_j^2 C_x^2+\frac{B}{\bar{X}^2C_x}\{\frac{\mu_{03}}{\bar{X}C_x}-\frac{\mu_{12}}{\bar{Y}\rho C_y}\}\right). \end{aligned} \tag{10}$$

Now, after squaring of both sides of equations "(7) and (8)" and neglecting the terms of $e's$ having power of more than two, we have

$$\left[T_{SP1(j)}-\bar{Y}\right]^2=\bar{Y}^2\left[e_0^2+e_1^2\left(\frac{1}{2}\Phi_j+B\right)^2-2e_0e_1\left(\frac{1}{2}\Phi_j+B\right)\right]. \tag{11}$$

$$\left[T_{SP2(j)}-\bar{Y}\right]^2=\bar{Y}^2\left[e_0^2+e_1^2\left(\frac{1}{2}\Psi_j+B\right)^2-2e_0e_1\left(\frac{1}{2}\Psi_j+B\right)\right]. \tag{12}$$

Taking expectation of both sides of equations "(11) and (12)", we get the MSEs of the suggested estimators $T_{SP1(j)}$ and $T_{SP1(j)}$ (where $j=1,2,...,10$) for the first order of large approximations as

$$\begin{aligned} MSE[T_{SP1(j)}] &= E\left[T_{SP1(j)}-\bar{Y}\right]^2, \\ &= \bar{Y}^2E\left[e_0^2+e_1^2\left(\frac{1}{2}\Phi_j+B\right)^2-2e_0e_1\left(\frac{1}{2}\Phi_j+B\right)\right], \\ &= \left(\frac{1}{n}-\frac{1}{N}\right)\bar{Y}^2\left[\frac{1}{4}\Phi_j^2C_x^2+\left(1-\rho^2\right)C_y^2\right]. \end{aligned} \tag{13}$$

$$\begin{aligned} MSE[T_{SP2(j)}] &= E\left[T_{SP2(j)}-\bar{Y}\right]^2, \\ &= \bar{Y}^2E\left[e_0^2+e_1^2\left(\frac{1}{2}\Psi_j+B\right)^2-2e_0e_1\left(\frac{1}{2}\Psi_j+B\right)\right], \\ &= \left(\frac{1}{n}-\frac{1}{N}\right)\bar{Y}^2\left[\frac{1}{4}\Psi_j^2C_x^2+\left(1-\rho^2\right)C_y^2\right]. \end{aligned} \tag{14}$$

## 5. Efficiency Comparisons

In this section, the efficiency conditions for suggested estimators $T_{SP1(j)}$ and $T_{SP2(j)}$ ($j = 1,2,...,10$ ) have been derived algebraically according to the usual mean estimator, usual ratio (Cochran (1977)) estimator, Kadilar and Cingi (2004, 2006) estimators and Subzar *et al.* (2017) estimators.

### 5.1. Comparison with usual mean estimator

**(i)** $Var(\bar{y}) - MSE(T_{SP1(j)}) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left[\rho^2 C_y^2 - \frac{1}{4}\Phi_j^2 C_x^2\right] > 0$, if

$$\frac{\rho C_y}{C_x} > \frac{1}{2}\Phi_j, (j = 1,2,...,10). \tag{15}$$

**(ii)** $Var(\bar{y}) - MSE(T_{SP2(j)}) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left[\rho^2 C_y^2 - \frac{1}{4}\Psi_j^2 C_x^2\right] > 0$, if

$$\frac{\rho C_y}{C_x} > \frac{1}{2}\Psi_j, (j = 1,2,...,10). \tag{16}$$

### 5.2. Comparison with usual ratio estimator

**(i)** $MSE(\bar{y}_{Ratio}) - MSE(T_{SP1(j)}) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left[(\rho C_y - C_x)^2 - \frac{1}{4}\Phi_j^2 C_x^2\right] > 0$, if

$$(\frac{\rho C_y}{C_x} - 1) > \frac{1}{2}\Phi_j, (j = 1,2,...,10). \tag{17}$$

**(ii)** $MSE(\bar{y}_{Ratio}) - MSE(T_{SP2(j)}) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left[(\rho C_y - C_x)^2 - \frac{1}{4}\Psi_j^2 C_x^2\right] > 0$, if

$$(\frac{\rho C_y}{C_x} - 1) > \frac{1}{2}\Psi_j, (j = 1,2,...,10). \tag{18}$$

### 5.3. Comparison with Kadilar and Cingi (2004, 2006) estimators

**(i)** $(MSE(T_{KC(i)})) - MSE(T_{SP1(j)}) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left[KC_i^2 - \frac{1}{4}\Phi_j^2\right] > 0$, if

$$KC_i > \frac{1}{2}\Phi_j, ((i = 1,2,...,10), (j = 1,2,...,10)). \tag{19}$$

**(ii)** $MSE(T_{KC(i)}) - MSE(T_{SP2(j)}) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left[KC_i^2 - \frac{1}{4}\Psi_j^2\right] > 0$, if

$$KC_i > \frac{1}{2}\Psi_j, ((i = 1,2,...,10), (j = 1,2,...,10)). \tag{20}$$

### 5.4. Comparison with Subzar *et al.* (2017) estimators

**(i)** $MSE(T_{Smrs(i)}) - MSE(T_{SP1(j)}) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left[\alpha_i^2 - \frac{1}{4}\Phi_j^2\right] > 0$, if

$$\alpha_i > \frac{1}{2}\Phi_j, ((i = 1,2,...,20), (j = 1,2,...,10)). \tag{21}$$

**(ii)** $MSE(T_{Smrs(i)}) - MSE(T_{SP2(j)}) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left[\alpha_i^2 - \frac{1}{4}\Psi_j^2\right] > 0$, if

$$\alpha_i > \frac{1}{2}\Psi_j, ((i = 1,2,...,20), (j = 1,2,...,10)). \tag{22}$$

From the equations [(15)-(22)], the suggested classes of estimators $T_{SP1(j)}$ and $T_{SP2(j)}$ (where $j = 1,2,...,10$ ) are more efficient than the usual mean estimator, usual ratio (Cochran (1977)) estimator, Kadilar and Cingi (2004, 2006) estimators and Subzar *et al.* (2017) estimators as long as the conditions (15), (16), (17), (18), (19), (20), (21) and (22) are satisfied.

## 6. Numerical Demonstration

In this section, the suggested estimators are compared with respect to the some other existing estimators in this literature. The relative efficiencies $(\%)$ of the suggested estimators $T_{SP1(j)}$ and $T_{SP2(j)}$ (where $j = 1,2,...,10$) with respect to the usual mean estimator, usual ratio (Cochran (1977)) estimator, Kadilar and Cingi (2004, 2006) estimators and Subzar *et al.* (2017) estimators respectively, are computed as follows:

$$RE(ExistingEstimators, SuggestedEstimators) = \frac{MSE(ExistingEstimators)}{MSE(SuggestedEstimators)} \times 100.$$

The values of relative efficiencies $(\%)$ of the suggested estimators are shown in Tables [3-8]. Two different types of natural population data sets from the books (Murty (1967), page 228) and (Singh and Chaudhary (1986), page 177) have been considered to analyse the performance of the suggested estimators over other existing estimators.

## 7. Conclusions

In this paper, two natural population data sets have been considered for different parameters in Table 2. From Tables [3-8], it is found that our suggested classes of estimators are more efficient than the usual mean estimator, usual ratio (Cochran (1977)) estimator, Kadilar and Cingi (2004, 2006) estimators and Subzar *et al.* (2017) estimators. Hence, the performances of the suggested classes of linear regression type ratio exponential estimators are highly justified in numerical demonstration which are shown in Tables [3-8] that may be recommended for further use.

## Acknowledgements

The author is grateful to the referees for their valuable suggestions.

## References

ABID, M., ABBAS, N., RIAZ, M., (2016a). Enhancing the Mean Ratio Estimators for Estimating Population Mean using Non-Conventional Location Parameters, *Revista Colombiana de Estadistica*, 39 (1), pp. 63–79.

ABID, M., ABBAS, N., RIAZ, M., (2016b). Improved Modified Ratio Estimators of Population Mean Based on Deciles, *Chiang Mai Journal of Science*, 43 (1), pp. 1311–1323.

ABID, M., ABBAS, N., SHERWANI, K. A. R., NAZIR, Z., H., (2016c). Improved Ratio Estimators for the Population Mean using Non-Conventional Measures of Dispersion, *Pakistan Journal of Statistics and Operation Research*, 12 (2), pp. 353-367.

COCHRAN, W. G., (1977). Sampling Techniques, *3rd edn. Wiley and Sons*.

KADILAR, C., CINGI, H., (2004). Ratio Estimators in Simple Random Sampling, *Applied Mathematics and Computaton*, 151, pp. 893-902.

KADILAR, C., CINGI, H., (2006). An Improvement in Estimating the Population Mean by using the Correlation Coefficient, *Hacettepe Journal of Mathematics and Statistics*, 35 (1), pp. 103–109.

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

SINGH, D., CHAUDHARY, F. S., (1986). Theory and Analysis of Sample Survey Designs, *1st edn. New Age International Publisher*, India .

SISODIA, B. V. S., DWIVEDI, V. K., (1981). A Modified Ratio Estimator using Coefficient of Variation of Auxiliary Variable, *Journal of the Indian Society of Agricultural Statistics*, 33 (1), pp. 13–18.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2012a). Estimation of Population Mean using Co-efficient of Variation and Median of an Auxiliary Variable, *International Journal of Probability and Statistics*, (1), pp. 111–118.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2012b). Estimation of Population Mean using known Median and Co-efficient of Skewness, *American Journal of Mathematics and Statistics*, 2, 101–107.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2012c). Modified Ratio Estimators using known Median and Co-efficient of kurtosis, *American Journal of Mathematics and Statistics*, 2, pp. 95–100.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2012d). A Class of Modified Ratio Estimators using Deciles of an Auxiliary Variable, *International Journal of Statistical Application*, 2, pp. 101–107.

SUBZAR, M., MAQBOOL, S., RAJA, T. A., SHABEER, M., (2017). A New Ratio Estimators for Estimation of Population Mean using Conventional Location Parameters, *World Applied Sciences Journal*, 35 (3), pp. 377–384.

SWAIN, A. K. P. C., (2014). An Improved Ratio Type Estimator of Finite Population in Sample Surveys, *Revista Investigación Operacional*, 35 (1), pp. 49–57.

UPADHYAYA, L. N., SINGH, H. P., (1999). Use of Transformed Auxiliary Variable in Estimating the Finite Population Mean, *Biometrical Journal*, 41 (5), pp. 627–636.

YAN, Z., TIAN, B., (2010). Ratio Method to the Mean Estimation Using Coefficient of Skewness of Auxiliary Variable, *ICICA 2010, Part 11, CCIS* 106, pp. 103–110.

Table 1: Some members of the suggested class of linear regression ratio type exponential estimators of Class A ($T_{SP1(j)}(j=1,2,...,10)$) and Class B ( $T_{SP2(j)}(j=1,2,...,10)$) respectively.

| The First Suggested Estimators (Class A) | $\Phi_i$ |
|---|---|
| $T_{SP1(1)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Phi_1 \frac{(1-\bar{x}/\bar{X})}{1+((QD\bar{x}+D_1)/(QD\bar{X}+D_1))}\right]$ | $\Phi_1 = QD\bar{X}/(QD\bar{X}+D_1)$ |
| $T_{SP1(2)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Phi_2 \frac{(1-\bar{x}/\bar{X})}{1+((QD\bar{x}+D_2)/(QD\bar{X}+D_2))}\right]$ | $\Phi_2 = QD\bar{X}/(QD\bar{X}+D_2)$ |
| $T_{SP1(3)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Phi_3 \frac{(1-\bar{x}/\bar{X})}{1+((QD\bar{x}+D_3)/(QD\bar{X}+D_3))}\right]$ | $\Phi_3 = QD\bar{X}/(QD\bar{X}+D_3)$ |
| $T_{SP1(4)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Phi_4 \frac{(1-\bar{x}/\bar{X})}{1+((QD\bar{x}+D_4)/(QD\bar{X}+D_4))}\right]$ | $\Phi_4 = QD\bar{X}/(QD\bar{X}+D_4)$ |
| $T_{SP1(5)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Phi_5 \frac{(1-\bar{x}/\bar{X})}{1+((QD\bar{x}+D_5)/(QD\bar{X}+D_5))}\right]$ | $\Phi_5 = QD\bar{X}/(QD\bar{X}+D_5)$ |
| $T_{SP1(6)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Phi_6 \frac{(1-\bar{x}/\bar{X})}{1+((QD\bar{x}+D_6)/(QD\bar{X}+D_6))}\right]$ | $\Phi_6 = QD\bar{X}/(QD\bar{X}+D_6)$ |
| $T_{SP1(7)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Phi_7 \frac{(1-\bar{x}/\bar{X})}{1+((QD\bar{x}+D_7)/(QD\bar{X}+D_7))}\right]$ | $\Phi_7 = QD\bar{X}/(QD\bar{X}+D_7)$ |
| $T_{SP1(8)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Phi_8 \frac{(1-\bar{x}/\bar{X})}{1+((QD\bar{x}+D_8)/(QD\bar{X}+D_8))}\right]$ | $\Phi_8 = QD\bar{X}/(QD\bar{X}+D_8)$ |
| $T_{SP1(9)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Phi_9 \frac{(1-\bar{x}/\bar{X})}{1+((QD\bar{x}+D_9)/(QD\bar{X}+D_9))}\right]$ | $\Phi_9 = QD\bar{X}/(QD\bar{X}+D_9)$ |
| $T_{SP1(10)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Phi_{10} \frac{(1-\bar{x}/\bar{X})}{1+((QD\bar{x}+D_{10})/(QD\bar{X}+D_{10}))}\right]$ | $\Phi_{10} = QD\bar{X}/(QD\bar{X}+D_{10})$ |
| The Second Suggested Estimators (Class B) | $\Psi_i$ |
| $T_{SP2(1)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Psi_1 \frac{(1-\bar{x}/\bar{X})}{1+((D_1\bar{x}+QD)/(D_1\bar{X}+QD))}\right]$ | $\Psi_1 = D_1\bar{X}/(D_1\bar{X}+QD)$ |
| $T_{SP2(2)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Psi_2 \frac{(1-\bar{x}/\bar{X})}{1+((D_2\bar{x}+QD)/(D_2\bar{X}+QD))}\right]$ | $\Psi_2 = D_2\bar{X}/(D_2\bar{X}+QD)$ |
| $T_{SP2(3)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Psi_3 \frac{(1-\bar{x}/\bar{X})}{1+((D_3\bar{x}+QD)/(D_3\bar{X}+QD))}\right]$ | $\Psi_3 = D_3\bar{X}/(D_3\bar{X}+QD)$ |
| $T_{SP2(4)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Psi_4 \frac{(1-\bar{x}/\bar{X})}{1+((D_4\bar{x}+QD)/(D_4\bar{X}+QD))}\right]$ | $\Psi_4 = D_4\bar{X}/(D_4\bar{X}+QD)$ |
| $T_{SP2(5)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Psi_5 \frac{(1-\bar{x}/\bar{X})}{1+((D_5\bar{x}+QD)/(D_5\bar{X}+QD))}\right]$ | $\Psi_5 = D_5\bar{X}/(D_5\bar{X}+QD)$ |
| $T_{SP2(6)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Psi_6 \frac{(1-\bar{x}/\bar{X})}{1+((D_6\bar{x}+QD)/(D_6\bar{X}+QD))}\right]$ | $\Psi_6 = D_6\bar{X}/(D_6\bar{X}+QD)$ |
| $T_{SP2(7)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Psi_7 \frac{(1-\bar{x}/\bar{X})}{1+((D_7\bar{x}+QD)/(D_7\bar{X}+QD))}\right]$ | $\Psi_7 = D_7\bar{X}/(D_7\bar{X}+QD)$ |
| $T_{SP2(8)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Psi_8 \frac{(1-\bar{x}/\bar{X})}{1+((D_8\bar{x}+QD)/(D_8\bar{X}+QD))}\right]$ | $\Psi_8 = D_8\bar{X}/(D_8\bar{X}+QD)$ |
| $T_{SP2(9)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Psi_9 \frac{(1-\bar{x}/\bar{X})}{1+((D_9\bar{x}+QD)/(D_9\bar{X}+QD))}\right]$ | $\Psi_9 = D_9\bar{X}/(D_9\bar{X}+QD)$ |
| $T_{SP2(10)} = \left[\bar{y}+\hat{\beta}(\bar{X}-\bar{x})\right] \exp\left[\Psi_{10} \frac{(1-\bar{x}/\bar{X})}{1+((D_{10}\bar{x}+QD)/(D_{10}\bar{X}+QD))}\right]$ | $\Psi_{10} = D_{10}\bar{X}/(D_{10}\bar{X}+QD)$ |

Table 2: Parameters of two natural population data sets

| Population A<br>Murthy (1967), page 228 | | Population B<br>Singh and Chaudhary (1986), page 177 | |
|---|---|---|---|
| $N = 80$ | $n = 20$ | $N = 34$ | $n = 20$ |
| $\bar{Y} = 5182.637$ | $\bar{X} = 1126.463$ | $\bar{Y} = 856.4117$ | $\bar{X} = 199.4412$ |
| $\rho = 0.941$ | $S_y = 1835.659$ | $\rho = 0.4453$ | $S_y = 733.1407$ |
| $C_y = 0.354$ | $S_x = 845.610$ | $C_y = 0.8561$ | $S_x = 150.2150$ |
| $C_x = 0.751$ | $\beta_2(x) = -0.063$ | $C_x = 0.7531$ | $\beta_2(x) = 1.0445$ |
| $\beta_1(x) = 1.050$ | $D_1 = 360$ | $\beta_1(x) = 1.1823$ | $D_1 = 60.60$ |
| $D_2 = 460$ | $D_3 = 590$ | $D_2 = 83.00$ | $D_3 = 102.70$ |
| $D_4 = 670$ | $D_5 = 750$ | $D_4 = 111.20$ | $D_5 = 142.50$ |
| $D_6 = 850$ | $D_7 = 1480$ | $D_6 = 210.20$ | $D_7 = 264.50$ |
| $D_8 = 1810$ | $D_9 = 2500$ | $D_8 = 304.40$ | $D_9 = 373.20$ |
| $D_{10} = 3480$ | $QD = 588.125$ | $D_{10} = 634.00$ | $QD = 80.25$ |

Table 3: Relative efficiencies (%) of the suggested estimators $T_{SP1(j)}$ and $T_{SP2(j)}$ $(j = 1,2,...,10)$ over the existing estimators $T_{KC(1)}$, $T_{KC(2)}$, $T_{KC(3)}$, $T_{KC(4)}$, $T_{KC(5)}$, $T_{KC(6)}$, $T_{KC(7)}$, $T_{KC(8)}$, $T_{KC(9)}$ and $T_{KC(10)}$ respectively for Population A.

| *Estimators* | $T_{KC(1)}$ | $T_{KC(2)}$ | $T_{KC(3)}$ | $T_{KC(4)}$ | $T_{KC(5)}$ | $T_{KC(6)}$ | $T_{KC(7)}$ | $T_{KC(8)}$ | $T_{KC(9)}$ | $T_{KC(10)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| The First Suggested Estimators (Class A) | | | | | | | | | | |
| $T_{SP1(1)}$ | 372.654 | 372.170 | 372.694 | 380.469 | 372.708 | 372.047 | 371.847 | 372.139 | 382.486 | 372.697 |
| $T_{SP1(2)}$ | 372.756 | 372.272 | 372.797 | 380.573 | 372.810 | 372.149 | 371.949 | 372.241 | 382.591 | 372.799 |
| $T_{SP1(3)}$ | 372.889 | 372.404 | 372.929 | 380.709 | 372.943 | 372.282 | 372.081 | 372.374 | 382.727 | 372.932 |
| $T_{SP1(4)}$ | 372.970 | 372.486 | 373.011 | 380.792 | 373.024 | 372.363 | 372.162 | 372.455 | 382.811 | 373.013 |
| $T_{SP1(5)}$ | 373.052 | 372.567 | 373.093 | 380.875 | 373.106 | 372.445 | 372.244 | 372.537 | 382.895 | 373.095 |
| $T_{SP1(6)}$ | 373.154 | 372.669 | 373.195 | 380.980 | 373.208 | 372.547 | 372.346 | 372.639 | 383.000 | 373.197 |
| $T_{SP1(7)}$ | 373.797 | 373.312 | 373.838 | 381.637 | 373.852 | 373.189 | 372.988 | 373.281 | 383.660 | 373.841 |
| $T_{SP1(8)}$ | 374.134 | 373.648 | 374.175 | 381.981 | 374.189 | 373.526 | 373.324 | 373.618 | 384.006 | 374.178 |
| $T_{SP1(9)}$ | 374.840 | 374.353 | 374.881 | 382.701 | 374.894 | 374.230 | 374.028 | 374.322 | 384.730 | 374.883 |
| $T_{SP1(10)}$ | 375.842 | 375.354 | 375.883 | 383.724 | 375.897 | 375.231 | 375.028 | 375.323 | 385.759 | 375.886 |
| The Second Suggested Estimators (Class B) | | | | | | | | | | |
| $T_{SP2(1)}$ | 373.267 | 372.782 | 373.308 | 381.095 | 373.321 | 372.660 | 372.459 | 372.752 | 383.116 | 373.310 |
| $T_{SP2(2)}$ | 373.054 | 372.569 | 373.095 | 380.877 | 373.108 | 372.447 | 372.246 | 372.539 | 382.897 | 373.097 |
| $T_{SP2(3)}$ | 372.885 | 372.400 | 372.925 | 380.705 | 372.939 | 372.278 | 372.077 | 372.370 | 382.723 | 372.928 |
| $T_{SP2(4)}$ | 372.813 | 372.329 | 372.854 | 380.632 | 372.867 | 372.207 | 372.006 | 372.299 | 382.650 | 372.857 |
| $T_{SP2(5)}$ | 372.757 | 372.273 | 372.798 | 380.575 | 372.811 | 372.151 | 371.950 | 372.243 | 382.592 | 372.800 |
| $T_{SP2(6)}$ | 372.702 | 372.218 | 372.742 | 380.518 | 372.756 | 372.095 | 371.895 | 372.187 | 382.535 | 372.745 |
| $T_{SP2(7)}$ | 372.525 | 372.041 | 372.566 | 380.338 | 372.579 | 371.919 | 371.718 | 372.011 | 382.354 | 372.568 |
| $T_{SP2(8)}$ | 372.482 | 371.998 | 372.522 | 380.293 | 372.536 | 371.875 | 371.675 | 371.967 | 382.309 | 372.525 |
| $T_{SP2(9)}$ | 372.428 | 371.944 | 372.468 | 380.238 | 372.482 | 371.822 | 371.621 | 371.914 | 382.254 | 372.471 |
| $T_{SP2(10)}$ | 372.388 | 371.904 | 372.429 | 380.198 | 372.442 | 371.782 | 371.581 | 371.874 | 382.213 | 372.431 |

Table 4: Relative efficiencies (%) of the suggested estimators $T_{SP1(j)}$ and $T_{SP2(j)}$ $(j = 1,2,...,10)$ over the existing estimators $T_{Smrs(i)}$ $(i = 1,2,...,10)$ respectively for Population A.

| *Estimators* | $T_{Smrs(1)}$ | $T_{Smrs(2)}$ | $T_{Smrs(3)}$ | $T_{Smrs(4)}$ | $T_{Smrs(5)}$ | $T_{Smrs(6)}$ | $T_{Smrs(7)}$ | $T_{Smrs(8)}$ | $T_{Smrs(9)}$ | $T_{Smrs(10)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| The First Suggested Estimators (Class A) | | | | | | | | | | |
| $T_{SP1(1)}$ | 372.259 | 372.150 | 372.007 | 371.920 | 371.832 | 371.723 | 371.036 | 370.676 | 369.927 | 368.866 |
| $T_{SP1(2)}$ | 372.361 | 372.252 | 372.109 | 372.022 | 371.934 | 371.825 | 371.137 | 370.778 | 370.028 | 368.967 |
| $T_{SP1(3)}$ | 372.494 | 372.384 | 372.242 | 372.154 | 372.067 | 371.957 | 371.269 | 370.910 | 370.160 | 369.098 |
| $T_{SP1(4)}$ | 372.575 | 372.466 | 372.323 | 372.236 | 372.148 | 372.039 | 371.351 | 370.991 | 370.241 | 369.179 |
| $T_{SP1(5)}$ | 372.657 | 372.547 | 372.405 | 372.317 | 372.230 | 372.120 | 371.432 | 371.072 | 370.322 | 369.26 |
| $T_{SP1(6)}$ | 372.759 | 372.649 | 372.507 | 372.419 | 372.331 | 372.222 | 371.534 | 371.174 | 370.423 | 369.361 |
| $T_{SP1(7)}$ | 373.401 | 373.292 | 373.149 | 373.061 | 372.973 | 372.864 | 372.174 | 371.814 | 371.062 | 369.998 |
| $T_{SP1(8)}$ | 373.738 | 373.628 | 373.485 | 373.398 | 373.310 | 373.200 | 372.510 | 372.149 | 371.396 | 370.331 |
| $T_{SP1(9)}$ | 374.443 | 374.333 | 374.190 | 374.102 | 374.014 | 373.904 | 373.212 | 372.851 | 372.097 | 371.030 |
| $T_{SP1(10)}$ | 375.444 | 375.334 | 375.190 | 375.102 | 375.014 | 374.904 | 374.210 | 373.848 | 373.092 | 372.022 |
| The Second Suggested Estimators (Class B) | | | | | | | | | | |
| $T_{SP2(1)}$ | 372.872 | 372.762 | 372.620 | 372.532 | 372.444 | 372.335 | 371.646 | 371.286 | 370.535 | 369.473 |
| $T_{SP2(2)}$ | 372.659 | 372.549 | 372.407 | 372.319 | 372.232 | 372.122 | 371.434 | 371.074 | 370.324 | 369.262 |
| $T_{SP2(3)}$ | 372.490 | 372.380 | 372.238 | 372.150 | 372.063 | 371.953 | 371.265 | 370.906 | 370.156 | 369.094 |
| $T_{SP2(4)}$ | 372.419 | 372.309 | 372.167 | 372.079 | 371.992 | 371.882 | 371.194 | 370.835 | 370.085 | 369.024 |
| $T_{SP2(5)}$ | 372.362 | 372.253 | 372.111 | 372.023 | 371.935 | 371.826 | 371.138 | 370.779 | 370.029 | 368.968 |
| $T_{SP2(6)}$ | 372.307 | 372.198 | 372.055 | 371.968 | 371.880 | 371.771 | 371.083 | 370.724 | 369.974 | 368.913 |
| $T_{SP2(7)}$ | 372.130 | 372.021 | 371.879 | 371.791 | 371.704 | 371.595 | 370.907 | 370.548 | 369.799 | 368.738 |
| $T_{SP2(8)}$ | 372.087 | 371.978 | 371.835 | 371.748 | 371.660 | 371.551 | 370.864 | 370.505 | 369.756 | 368.695 |
| $T_{SP2(9)}$ | 372.033 | 371.924 | 371.782 | 371.694 | 371.607 | 371.498 | 370.810 | 370.451 | 369.702 | 368.642 |
| $T_{SP2(10)}$ | 371.994 | 371.884 | 371.742 | 371.655 | 371.567 | 371.458 | 370.771 | 370.412 | 369.663 | 368.603 |

Table 5: Relative efficiencies (%) of the suggested estimators $T_{SP1(j)}$ and $T_{SP2(j)}$ $(j = 1,2,...,10)$ over the existing estimators $T_{Smrs(i)}$ $(i = 11,12,...,20)$ respectively for Population A.

| *Estimators* | $T_{Smrs(11)}$ | $T_{Smrs(12)}$ | $T_{Smrs(13)}$ | $T_{Smrs(14)}$ | $T_{Smrs(15)}$ | $T_{Smrs(16)}$ | $T_{Smrs(17)}$ | $T_{Smrs(18)}$ | $T_{Smrs(19)}$ | $T_{Smrs(20)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| The First Suggested Estimators (Class A) | | | | | | | | | | |
| $T_{SP1(1)}$ | 371.602 | 371.830 | 372.012 | 372.088 | 372.148 | 372.208 | 372.398 | 372.444 | 372.502 | 372.545 |
| $T_{SP1(2)}$ | 371.704 | 371.932 | 372.113 | 372.190 | 372.250 | 372.310 | 372.500 | 372.546 | 372.604 | 372.647 |
| $T_{SP1(3)}$ | 371.836 | 372.065 | 372.246 | 372.323 | 372.383 | 372.442 | 372.632 | 372.679 | 372.737 | 372.779 |
| $T_{SP1(4)}$ | 371.918 | 372.146 | 372.327 | 372.404 | 372.464 | 372.524 | 372.714 | 372.760 | 372.818 | 372.861 |
| $T_{SP1(5)}$ | 371.999 | 372.227 | 372.409 | 372.486 | 372.546 | 372.605 | 372.795 | 372.842 | 372.900 | 372.943 |
| $T_{SP1(6)}$ | 372.101 | 372.329 | 372.511 | 372.587 | 372.648 | 372.707 | 372.897 | 372.944 | 373.002 | 373.045 |
| $T_{SP1(7)}$ | 372.742 | 372.971 | 373.153 | 373.230 | 373.290 | 373.350 | 373.540 | 373.587 | 373.645 | 373.688 |
| $T_{SP1(8)}$ | 373.078 | 373.308 | 373.490 | 373.567 | 373.627 | 373.687 | 373.877 | 373.924 | 373.982 | 374.025 |
| $T_{SP1(9)}$ | 373.782 | 374.011 | 374.194 | 374.271 | 374.331 | 374.391 | 374.582 | 374.629 | 374.687 | 374.730 |
| $T_{SP1(10)}$ | 374.781 | 375.012 | 375.194 | 375.272 | 375.332 | 375.392 | 375.584 | 375.631 | 375.689 | 375.732 |
| The Second Suggested Estimators (Class B) | | | | | | | | | | |
| $T_{SP2(1)}$ | 372.214 | 372.442 | 372.624 | 372.700 | 372.761 | 372.820 | 373.010 | 373.057 | 373.115 | 373.158 |
| $T_{SP2(2)}$ | 372.001 | 372.229 | 372.411 | 372.488 | 372.548 | 372.607 | 372.797 | 372.844 | 372.902 | 372.945 |
| $T_{SP2(3)}$ | 371.832 | 372.061 | 372.242 | 372.319 | 372.379 | 372.438 | 372.628 | 372.675 | 372.733 | 372.776 |
| $T_{SP2(4)}$ | 371.761 | 371.989 | 372.171 | 372.247 | 372.308 | 372.367 | 372.557 | 372.604 | 372.661 | 372.704 |
| $T_{SP2(5)}$ | 371.705 | 371.933 | 372.115 | 372.191 | 372.252 | 372.311 | 372.501 | 372.547 | 372.605 | 372.648 |
| $T_{SP2(6)}$ | 371.650 | 371.878 | 372.059 | 372.136 | 372.196 | 372.256 | 372.445 | 372.492 | 372.550 | 372.593 |
| $T_{SP2(7)}$ | 371.474 | 371.702 | 371.883 | 371.959 | 372.020 | 372.079 | 372.269 | 372.315 | 372.373 | 372.416 |
| $T_{SP2(8)}$ | 371.430 | 371.658 | 371.839 | 371.916 | 371.976 | 372.036 | 372.225 | 372.272 | 372.330 | 372.373 |
| $T_{SP2(9)}$ | 371.377 | 371.605 | 371.786 | 371.862 | 371.923 | 371.982 | 372.172 | 372.218 | 372.276 | 372.319 |
| $T_{SP2(10)}$ | 371.337 | 371.565 | 371.746 | 371.823 | 371.883 | 371.942 | 372.132 | 372.179 | 372.236 | 372.279 |

Table 6: Relative efficiencies (%) of the suggested estimators $T_{SP1(j)}$ and $T_{SP2(j)}$ $(j = 1,2,...,10)$ over the existing estimators $T_{KC(1)}$, $T_{KC(2)}$, $T_{KC(3)}$, $T_{KC(4)}$, $T_{KC(5)}$, $T_{KC(6)}$, $T_{KC(7)}$, $T_{KC(8)}$, $T_{KC(9)}$ and $T_{KC(10)}$ respectively for Population B.

| *Estimators* | $T_{KC(1)}$ | $T_{KC(2)}$ | $T_{KC(3)}$ | $T_{KC(4)}$ | $T_{KC(5)}$ | $T_{KC(6)}$ | $T_{KC(7)}$ | $T_{KC(8)}$ | $T_{KC(9)}$ | $T_{KC(10)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| The First Suggested Estimators (Class A) | | | | | | | | | | |
| $T_{SP1(1)}$ | 158.552 | 157.968 | 157.743 | 157.992 | 157.481 | 158.206 | 158.093 | 157.248 | 158.221 | 156.753 |
| $T_{SP1(2)}$ | 158.638 | 158.053 | 157.828 | 158.077 | 157.565 | 158.291 | 158.178 | 157.333 | 158.306 | 156.837 |
| $T_{SP1(3)}$ | 158.713 | 158.127 | 157.902 | 158.152 | 157.640 | 158.366 | 158.252 | 157.407 | 158.380 | 156.911 |
| $T_{SP1(4)}$ | 158.745 | 158.159 | 157.934 | 158.184 | 157.672 | 158.398 | 158.284 | 157.439 | 158.412 | 156.943 |
| $T_{SP1(5)}$ | 158.863 | 158.277 | 158.052 | 158.302 | 157.789 | 158.516 | 158.402 | 157.556 | 158.531 | 157.060 |
| $T_{SP1(6)}$ | 159.117 | 158.530 | 158.305 | 158.555 | 158.041 | 158.769 | 158.656 | 157.808 | 158.784 | 157.311 |
| $T_{SP1(7)}$ | 159.319 | 158.732 | 158.506 | 158.756 | 158.242 | 158.971 | 158.857 | 158.009 | 158.986 | 157.511 |
| $T_{SP1(8)}$ | 159.467 | 158.879 | 158.653 | 158.903 | 158.389 | 159.118 | 159.004 | 158.155 | 159.133 | 157.656 |
| $T_{SP1(9)}$ | 159.719 | 159.130 | 158.904 | 159.155 | 158.639 | 159.370 | 159.256 | 158.406 | 159.385 | 157.906 |
| $T_{SP1(10)}$ | 160.655 | 160.062 | 159.835 | 160.088 | 159.569 | 160.304 | 160.189 | 159.334 | 160.319 | 158.831 |
| The Second Suggested Estimators (Class B) | | | | | | | | | | |
| $T_{SP2(1)}$ | 158.726 | 158.141 | 157.916 | 158.165 | 157.653 | 158.379 | 158.266 | 157.421 | 158.394 | 156.924 |
| $T_{SP2(2)}$ | 158.617 | 158.032 | 157.808 | 158.057 | 157.545 | 158.270 | 158.157 | 157.313 | 158.285 | 156.817 |
| $T_{SP2(3)}$ | 158.560 | 157.976 | 157.751 | 158.000 | 157.489 | 158.214 | 158.101 | 157.256 | 158.229 | 156.760 |
| $T_{SP2(4)}$ | 158.542 | 157.957 | 157.733 | 157.982 | 157.470 | 158.196 | 158.083 | 157.238 | 158.210 | 156.742 |
| $T_{SP2(5)}$ | 158.494 | 157.909 | 157.685 | 157.934 | 157.422 | 158.147 | 158.034 | 157.190 | 158.162 | 156.694 |
| $T_{SP2(6)}$ | 158.438 | 157.854 | 157.629 | 157.878 | 157.367 | 158.092 | 157.979 | 157.135 | 158.106 | 156.639 |
| $T_{SP2(7)}$ | 158.414 | 157.830 | 157.605 | 157.854 | 157.343 | 158.068 | 157.955 | 157.111 | 158.082 | 156.615 |
| $T_{SP2(8)}$ | 158.402 | 157.817 | 157.593 | 157.842 | 157.331 | 158.055 | 157.942 | 157.099 | 158.070 | 156.603 |
| $T_{SP2(9)}$ | 158.387 | 157.802 | 157.578 | 157.827 | 157.316 | 158.040 | 157.927 | 157.084 | 158.055 | 156.589 |
| $T_{SP2(10)}$ | 158.359 | 157.775 | 157.551 | 157.800 | 157.289 | 158.013 | 157.900 | 157.057 | 158.028 | 156.562 |

Table 7: Relative efficiencies (%) of the suggested estimators $T_{SP1(j)}$ and $T_{SP2(j)}$ $(j = 1,2,...,10)$ over the existing estimators $T_{Smrs(i)}$ $(i = 1,2,...,10)$ respectively for Population B.

| *Estimators* | $T_{Smrs(1)}$ | $T_{Smrs(2)}$ | $T_{Smrs(3)}$ | $T_{Smrs(4)}$ | $T_{Smrs(5)}$ | $T_{Smrs(6)}$ | $T_{Smrs(7)}$ | $T_{Smrs(8)}$ | $T_{Smrs(9)}$ | $T_{Smrs(10)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| The First Suggested Estimators (Class A) | | | | | | | | | | |
| $T_{SP1(1)}$ | 157.966 | 157.751 | 157.563 | 157.482 | 157.184 | 156.547 | 156.041 | 155.673 | 155.044 | 152.731 |
| $T_{SP1(2)}$ | 158.051 | 157.836 | 157.647 | 157.566 | 157.269 | 156.631 | 156.125 | 155.756 | 155.127 | 152.813 |
| $T_{SP1(3)}$ | 158.126 | 157.91 | 157.722 | 157.641 | 157.343 | 156.705 | 156.199 | 155.83 | 155.201 | 152.885 |
| $T_{SP1(4)}$ | 158.158 | 157.942 | 157.754 | 157.673 | 157.375 | 156.736 | 156.23 | 155.862 | 155.232 | 152.916 |
| $T_{SP1(5)}$ | 158.276 | 158.06 | 157.871 | 157.79 | 157.492 | 156.853 | 156.347 | 155.978 | 155.348 | 153.03 |
| $T_{SP1(6)}$ | 158.529 | 158.313 | 158.124 | 158.042 | 157.744 | 157.104 | 156.597 | 156.227 | 155.596 | 153.275 |
| $T_{SP1(7)}$ | 158.73 | 158.514 | 158.325 | 158.243 | 157.944 | 157.304 | 156.796 | 156.426 | 155.794 | 153.47 |
| $T_{SP1(8)}$ | 158.877 | 158.661 | 158.471 | 158.39 | 158.09 | 157.449 | 156.941 | 156.57 | 155.938 | 153.612 |
| $T_{SP1(9)}$ | 159.129 | 158.912 | 158.722 | 158.64 | 158.341 | 157.699 | 157.189 | 156.818 | 156.185 | 153.855 |
| $T_{SP1(10)}$ | 160.061 | 159.843 | 159.652 | 159.57 | 159.269 | 158.623 | 158.11 | 157.737 | 157.1 | 154.756 |
| The Second Suggested Estimators (Class B) | | | | | | | | | | |
| $T_{SP2(1)}$ | 158.139 | 157.924 | 157.735 | 157.654 | 157.356 | 156.718 | 156.212 | 155.843 | 155.214 | 152.898 |
| $T_{SP2(2)}$ | 158.031 | 157.815 | 157.627 | 157.546 | 157.248 | 156.61 | 156.105 | 155.736 | 155.107 | 152.793 |
| $T_{SP2(3)}$ | 157.974 | 157.759 | 157.571 | 157.489 | 157.192 | 156.554 | 156.049 | 155.681 | 155.052 | 152.739 |
| $T_{SP2(4)}$ | 157.956 | 157.741 | 157.552 | 157.471 | 157.174 | 156.536 | 156.031 | 155.663 | 155.034 | 152.721 |
| $T_{SP2(5)}$ | 157.907 | 157.692 | 157.504 | 157.423 | 157.126 | 156.488 | 155.983 | 155.615 | 154.986 | 152.674 |
| $T_{SP2(6)}$ | 157.852 | 157.637 | 157.449 | 157.368 | 157.071 | 156.434 | 155.928 | 155.56 | 154.932 | 152.621 |
| $T_{SP2(7)}$ | 157.828 | 157.613 | 157.425 | 157.344 | 157.047 | 156.41 | 155.905 | 155.537 | 154.908 | 152.598 |
| $T_{SP2(8)}$ | 157.816 | 157.601 | 157.413 | 157.332 | 157.035 | 156.398 | 155.893 | 155.525 | 154.896 | 152.586 |
| $T_{SP2(9)}$ | 157.801 | 157.586 | 157.398 | 157.317 | 157.02 | 156.383 | 155.878 | 155.51 | 154.882 | 152.571 |
| $T_{SP2(10)}$ | 157.774 | 157.559 | 157.371 | 157.29 | 156.993 | 156.356 | 155.851 | 155.483 | 154.855 | 152.545 |

Table 8: Relative efficiencies (%) of the suggested estimators $T_{SP1(j)}$ and $T_{SP2(j)}$ $(j = 1,2,...,10)$ over the existing estimators $T_{Smrs(i)}$ $(i = 11,12,...,20)$, respectively for Population B.

| *Estimators* | $T_{Smrs(11)}$ | $T_{Smrs(12)}$ | $T_{Smrs(13)}$ | $T_{Smrs(14)}$ | $T_{Smrs(15)}$ | $T_{Smrs(16)}$ | $T_{Smrs(17)}$ | $T_{Smrs(18)}$ | $T_{Smrs(19)}$ | $T_{Smrs(20)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| The First Suggested Estimators (Class A) | | | | | | | | | | |
| $T_{SP1(1)}$ | 157.529 | 157.803 | 157.946 | 157.992 | 158.115 | 158.255 | 158.316 | 158.347 | 158.385 | 158.454 |
| $T_{SP1(2)}$ | 157.613 | 157.888 | 158.031 | 158.077 | 158.2 | 158.34 | 158.401 | 158.432 | 158.47 | 158.539 |
| $T_{SP1(3)}$ | 157.688 | 157.962 | 158.105 | 158.151 | 158.274 | 158.415 | 158.476 | 158.507 | 158.545 | 158.614 |
| $T_{SP1(4)}$ | 157.72 | 157.994 | 158.137 | 158.184 | 158.306 | 158.447 | 158.508 | 158.539 | 158.577 | 158.646 |
| $T_{SP1(5)}$ | 157.837 | 158.112 | 158.255 | 158.301 | 158.424 | 158.565 | 158.626 | 158.657 | 158.695 | 158.764 |
| $T_{SP1(6)}$ | 158.09 | 158.365 | 158.508 | 158.555 | 158.678 | 158.819 | 158.88 | 158.911 | 158.949 | 159.018 |
| $T_{SP1(7)}$ | 158.29 | 158.566 | 158.71 | 158.756 | 158.879 | 159.02 | 159.082 | 159.113 | 159.151 | 159.220 |
| $T_{SP1(8)}$ | 158.437 | 158.713 | 158.857 | 158.903 | 159.026 | 159.168 | 159.229 | 159.26 | 159.298 | 159.367 |
| $T_{SP1(9)}$ | 158.688 | 158.964 | 159.108 | 159.155 | 159.278 | 159.42 | 159.481 | 159.512 | 159.55 | 159.620 |
| $T_{SP1(10)}$ | 159.618 | 159.896 | 160.04 | 160.087 | 160.211 | 160.354 | 160.416 | 160.447 | 160.485 | 160.555 |
| The Second Suggested Estimators (Class B) | | | | | | | | | | |
| $T_{SP2(1)}$ | 157.701 | 157.976 | 158.119 | 158.165 | 158.288 | 158.428 | 158.489 | 158.52 | 158.558 | 158.627 |
| $T_{SP2(2)}$ | 157.593 | 157.867 | 158.01 | 158.056 | 158.179 | 158.32 | 158.381 | 158.412 | 158.449 | 158.518 |
| $T_{SP2(3)}$ | 157.536 | 157.811 | 157.954 | 158 | 158.123 | 158.263 | 158.324 | 158.355 | 158.393 | 158.462 |
| $T_{SP2(4)}$ | 157.518 | 157.793 | 157.936 | 157.982 | 158.104 | 158.245 | 158.306 | 158.337 | 158.375 | 158.443 |
| $T_{SP2(5)}$ | 157.47 | 157.744 | 157.887 | 157.933 | 158.056 | 158.196 | 158.257 | 158.288 | 158.326 | 158.395 |
| $T_{SP2(6)}$ | 157.415 | 157.689 | 157.832 | 157.878 | 158 | 158.141 | 158.202 | 158.233 | 158.27 | 158.339 |
| $T_{SP2(7)}$ | 157.391 | 157.665 | 157.808 | 157.854 | 157.976 | 158.117 | 158.178 | 158.209 | 158.246 | 158.315 |
| $T_{SP2(8)}$ | 157.379 | 157.653 | 157.796 | 157.842 | 157.964 | 158.105 | 158.165 | 158.196 | 158.234 | 158.303 |
| $T_{SP2(9)}$ | 157.364 | 157.638 | 157.781 | 157.827 | 157.949 | 158.09 | 158.151 | 158.181 | 158.219 | 158.288 |
| $T_{SP2(10)}$ | 157.337 | 157.611 | 157.753 | 157.8 | 157.922 | 158.062 | 158.123 | 158.154 | 158.192 | 158.261 |

*STATISTICS IN TRANSITION new series, December 2020*

*Vol. 21, No. 5, pp. 99–118, DOI 10.21307/stattrans-2020-057*

Received – 26.03.2020; accepted – 14.05.2020

# Modelling bid-ask spread conditional distributions using hierarchical correlation reconstruction

**Jarosław Duda[1], Henryk Gurgul[2], Robert Syrek[3]**

## ABSTRACT

While we would like to predict exact values, the information available, being incomplete, is rarely sufficient - usually allowing only conditional probability distributions to be predicted. This article discusses hierarchical correlation reconstruction (HCR) methodology for such a prediction using the example of bid-ask spreads (usually unavailable), but here predicted from more accessible data like closing price, volume, high/low price and returns. Using HCR methodology, as in copula theory, we first normalized marginal distributions so that they were nearly uniform. Then we modelled joint densities as linear combinations of orthonormal polynomials, obtaining their decomposition into mixed moments. Then we modelled each moment of the predicted variable separately as a linear combination of mixed moments of known variables using least squares linear regression. By combining these predicted moments, we obtained the predicted density as a polynomial, for which we can e.g. calculate the expected value, but also the variance to determine the uncertainty of the prediction, or we can use the entire distribution for, e.g. more accurate further calculations or generating random values. 10-fold cross-validation log-likelihood tests were conducted for 22 DAX companies, leading to very accurate predictions, especially when individual models were used for each company, as significant differences were found between their behaviours. An additional advantage of using this methodology is that it is computationally inexpensive; estimating and evaluating a model with hundreds of parameters and thousands of data points by means of this methodology takes only a second on a computer.

**Key words:** machine learning, conditional distribution, bid-ask spread, liquidity.

**JEL Classification:** C49, C58, G15.

## 1. Introduction

Liquidity is one of the key measures of financial market quality. The notion liquidity denotes a desirable function that should reflect a well-organized financial market. By liquid market we understand a market for which there exists a prompt and secure channel between the supply and demand of assets accompanied by low transaction costs. Providing a rigorous scientific definition of market liquidity happens to be a challenging aim. Liquidity is the main index of the health of a given stock market and the condition of the associated investment industry, using funds from this stock market. It is clear that more active trading leads

[1]Jagiellonian University, Poland. E-mail: jaroslaw.duda@uj.edu.pl.
ORCID: https://orcid.org/0000-0001-9559-809X.

[2]AGH University of Science and Technology, Poland. E-mail: henryk.gurgul@gmail.com.
ORCID: https://orcid.org/0000-0002-6192-2995.

[3]Jagiellonian University, Poland. E-mail: robert.syrek@uj.edu.pl.
ORCID: https://orcid.org/0000-0002-8212-8248.

to lower trading costs, more intensive flows of information and more activity concerning the relevant stocks displayed by potential investors. It is worth mentioning the role of speculators who can significantly increase the liquidity of the market, but may not necessarily have a positive impact on it.

In some recent contributions the definitions of market liquidity are based on the bid-ask spread and an estimation of its components. However, the difference between bid and ask quotes for an asset provides a liquidity measure with respect to a dealer market. Not to a broker market. Nevertheless, it is possible to compute approximations that replicate the difference between bid and ask quotes even in broker markets. Therefore, intradaily measures of liquidity can describe the main feature of a market, such as the arrival of new information in the hands of market participants. There are several definitions of liquidity. In each study on liquidity the initial goal is to formulate a definition of liquidity and justify it. The notion liquidity is related on the one hand to the transaction time - i.e. the duration of transactions, and on the other to transaction costs, understood as the price paid by investors for the supply of liquidity.

The common definition widely used by both researchers and market participants states that an asset is liquid if it can be sold quickly at a minimal cost. This definition of liquidity for a particular asset can be generalized for the whole market. A similar definition can also be applied to the stock market as a whole. In this sense, a market is liquid if it is possible to buy and sell assets at a minimal cost without a significant delay from the placement of the order. When assessing the liquidity of the stock market, in relation to incurring the lowest transaction costs, it is also important to take into consideration other elements than the size of the spread, which affect the cost of concluding buy/sell transactions, such as commissions and exchange fees; or taxation on capital gains; market volatility. However, in this contribution we focus on the spread which reflects to some extent the listed factors.

In the literature different measures of asset liquidity are known. These measures of liquidity take into account various alternative elements of the measurement approach. Some measures focus on the trading volume while other indices are based on the execution-cost relation of liquidity. The measures related to volume information reflect the price impact of transactions. After combining them into scalar measures they denote the liquidity on the whole market. However, the indices based on execution costs enable the properties of an asset to be evaluated. This is possible by analyzing the cost paid to the market maker (dealer or specialist) for matching the supply and demand.

The value added of this study is twofold. First of all, in order to find the characteristics of the future bid-ask spread we use a new methodology that has not been used for a financial time series before. Secondly, on the basis of empirical data from the German stock index DAX we have confirmed the advantages of this approach.

The most important conclusions concerning liquidity are based on the bid-ask spread and its variations. We aim to use our hierarchical correlation reconstruction (HCR) methodology in spread bid-ask description and forecasting. A more detailed outline of the advantages of this new methodology is at the end of the next Section. The content of the paper is organized as follows. In the next Section, the literature overview is presented. The third Section includes data and methodology. In the fourth Section, the empirical results are presented. The last Section provides conclusions.

## 2. Literature review

The pioneer in estimation of bid-ask spread, most often used measure of liquidity, was Roll Roll (1984). The model derived by Roll has been very useful tool of bid-ask description since the mid-eighties. The followers tried to improve and extend this approach. In Roll's model the spread is approximated based on return autocovariance.

According to Butler et al. (2005), lower liquidity implies higher transaction costs if the share capital increases. Moreover, a higher return on equity, or cost of equity, is expected.

Lesmond et al. (1999) belong to the first researchers who tested the quality of measures of stock liquidity. The contributors compared them based on different stocks. Bid-ask spread was used as a benchmark measure. Armitage et al. (2014) in contribution based on empirical data for Ukraine (2005-2006) found that the proportion of nontrading days, the proportion of zero-return days, stock volatility, and measure of Amihud (2002) exhibit high correlations with this spread. In conclusion the contributors stated that these indicators are good enough to measure liquidity for Ukraine. The findings of Armitage et al. (2014) regarding turnover are in line with those of Lesmond et al. (1999) for other emerging markets. In addition, they found that the proportion of zero-return days is a better measure for emerging markets than for developed markets.

In their studies of the determinants of the cost of trading, Armitage et al. (2014), Stoll (2000), Naik and Yadav (2003) and Gajewski and Gresse (2007) proved that the effective bid-ask spreads mentioned above depend on stock liquidity. Stock liquidity was measured by the number of non-trading days per year and the average number of trades per day. It turned out that higher liquidity stocks had narrower bid-ask spreads, as assumed. In the opinion of these and other scholars these effective spreads are related to the risk of the stock. The last is measured by return volatility. The more risky stocks exhibit usually wider bid-ask spreads. However, the opposite relationship between cost and trade size was observed for dealership markets like the London Stock Exchange (LSE) and NASDAQ. Some results are not consistent, e.g. on the basis of the data for the LSE, Reiss and Werner (1996) demonstrated that larger trades (but not to large) receive better prices. However, for unusually large orders this empirical observation is not true. Hansch et al. (1999) reported that on the LSE the price rise in relation to this spread is smallest for small trades, larger for medium-sized trades and largest for large ones. Huang and Stoll (1996) calculated that the mean spread for small trades amounts to almost 20 cents but for large trades it is smaller approximately by 30-35 percent. They discovered an asymmetry in the cost of trading between buyer- and seller-initiated trades. In addition, the authors analysing the company spreads on NYSE and NASDAQ in their paper, found out that spreads on NASDAQ are higher than on NYSE.

Chan and Lakonishok (1993) claim that in a portfolio for sale the number of stocks is limited. They try to convince the readers that the decision to sell must not convey negative information. On the contrary, according to the authors purchases are usually implied by firm specific information which is available.

Stoll (2000) conjectured that the spread depends on some factors related to a stock's liquidity and risk. On the basis of data from the USA, he performed a panel regression of this spread using five determinants as explanatory variables, namely trading volume, the

number of trades per day, free float, return variance and stock price. The models fit the data well since all explanatory variables are significant and the determination coefficient is over 0.6.

Naik and Yadav (2003) conducted similar research and obtained interesting results for the London Stock Exchange. Unfortunately, their results are not in line with later findings reported in Gajewski and Gresse (2007), who used data from Euronext Paris and the London Stock Exchange. As a new explanatory variable they included the imbalance between purchase and sale orders. They established that trading volume, return variance, and order imbalance were significant and exhibited the expected signs. However, free float, stock price, and the number of trades per day turned out not to be significant.

In some research the bid-ask spread is used as a measure of stock market liquidity employed in market microstructure studies. In Christie and Schultz (1994); Huang and Stoll (1996); Bessembinder (2003) the bid-ask spread is used to conduct inter-market comparisons of trading costs. The efficiency of rules and regulations aimed at reducing the cost of trading can be proven by checking the rules and regulations and their impact on the bid-ask spread.

In a more recent study Chen et al. (2017) proposed a non-parametric method to estimate the spread on the basis of the Roll (1984)) model. A further development can be found in Abdi and Ranaldo (2017), who incorporate the Corwin and Schultz (2012) model into the Roll model to derive a new estimator.

In the next part of this paper we shall focus on scarce bid-ask spreads, predicted on the basis of data which is more accessible, such as closing price, volume, high/low price, returns. Very preliminary results of this paper are in unpublished working paper by Duda et al. (2019).

In our calculations, we use hierarchical correlation reconstruction (HCR) methodology: each moment of the predicted variable is independently modelled as a linear combination of mixed moments of the variables used, then they are finally combined into the predicted (conditional) probability distribution. A basic use of predicting the entire distribution is to predict a value, e.g. as its expected value, additionally also estimating the uncertainty from its variance. Another use may be to handle more sophisticated situations such as a binomial distribution with two (or more) separate maxima: when predicting the expected value might not be a good choice (it may have a much lower density), a better prediction might be, e.g. one of the maxima, or may be both: providing a prediction as an alternative of two (or more) possibilities.

We can also use the entire predicted density, e.g. for a more accurate additional calculation, estimating the quantiles, or generating random values. HCR methodology combines the advantages of classical statistics and machine learning. While the former allows for well controlled and interpretable but relatively small (rough) models/descriptions, machine learning allows for very accurate descriptions using huge models, but usually lacks uniqueness of solution, control and interpretability of coefficients, and often is computationally costly. HCR allows one to work on huge models obtained from (unique) least-squares optimization, using well interpretable coefficients: as mixed moments of variables, starting, e.g. with moments of single variables and the correlation coefficients. The results for 22 DAX companies seem to be promising, especially using individual models for each company. An

additional advantage of this methodology is that it is computationally inexpensive; such complex models for these data can be estimated and evaluated in a second.

## 3. Data set and basic concepts

This Section discusses the data set and reminds one of the standard concepts, to be used for describing the methodology used in the next Section.

### 3.1. Data set and variables

Daily data for DAX companies from the 1999-2013 period were used (source in Acknowledgment); they were selected as they have at least 2000 data points: Deutsche Telekom AG (DTE), Daimler AG (DAI), SAP SE (SAP), Siemens AG (SIE), Deutsche Post AG (DPW), Allianz SE (ALV), BMW AG St (BMW), Infineon Technologies AG (IFX), Volkswagen AG Vz (VOW3), Fresenius SE & Co. KGaA (FRE), Henkel AG & Co. KGaA Vz (HNK3), Continental (CON), Merck KGaA (MRK), Münchener Rück AG (MUV2), Deutsche Börse AG (DB1), Deutsche Lufthansa AG (LHA), Fresenius Medical Care AG & Co. KGaA St (FME), Deutsche Bank AG (DBK), Fresenius Medical Care AG & Co. KGaA St (HEI), RWE AG St (RWE), Beiersdorf Aktiengesellschaft (BEI), Thyssenkrupp AG (TKA).

The basic set of variables is $P$ - closing price, $V$ - volume, $R$ - return, $H,L$ - high/low price. However, it turned out that trying to exploit dependence on $R$ and $L$ alone did improve evaluation, hence finally the basic model considered: '123' uses only $P$ as '1'-st variable, $V$ as '2'-nd variable and normalized $(H-L)/P$ as '3'-rd variable. It might be worth noting that the paper presents average spreads on the German stock market in question. This type of data is also applied in the cited references.

### 3.2. Bid-ask spread and some of its standard predictors

Bid-ask spread is the difference between the lowest asking price (*ask*, offered by a seller) and the highest bid price (*bid*, offered by a buyer). While this value is important because it is a main measure of market quality (Mestel et al. (2018); Gurgul and Machno (2017)), this information is usually publicly unavailable. Therefore, there is an interest in being able to predict this value on the basis of other, more accessible data.

At this point, one can present an important account that the smaller the spread, the more efficiently the market operates, and its liquidity understood by the volume of trading in securities also increases indirectly (Roll (1984)).

We consider bid-ask spread as a standard measure of liquidity. More specifically, we work on relative quoted spread, which is normalized by dividing by midpoint $(ask+bid)/2$: $S = \frac{ask-bid}{(ask+bid)/2}$.

Simple examples of its predictors based on the 5 basic variables are *AMI* (Amihud (2002); Fong et al. (2017)), *HLR* (Będowska-Sójka and Echaust (2019); Gurgul and Syrek (2019)):

$$AMI = \ln\left(1+\frac{|R|}{P\cdot V}\right) \qquad HLR = 2\frac{H-L}{H+L} \tag{1}$$

They are intended for a simpler task than that discussed: to predict values, while here we want to predict entire conditional probability distributions. We can reduce the predicted probability distributions into predicted values, e.g. as the expected value, median, or positions of maxima (especially for multimodal distributions). Fig. 1 presents comparisons using such predictions reduced with the expected value.

However, in practice such a prediction is often further processed through several functions, generally $E(f(X)) \neq f(E(X))$ for nonlinear, hence it is more accurate to process the probability distribution (e.g. on a lattice) through the functions before, e.g. taking the expected value.

### 3.3. Normalization to nearly uniform marginal distributions

Like in copula theory, in HCR methodology it is convenient to initially normalize all variables to nearly uniform marginal distributions in $[0,1]$, hence below we shall only work on such normalized variables, which beside usually better prediction also allows for better presentation of evaluation: e.g. density without prediction is 1, log-likelihood is 0.

This standard normalization requires estimation of the cumulative distribution function (CDF), individually for each variable, and this CDF function to be applied to the original values. Finally, having a prediction we can go back to the original variable using $\text{CDF}^{-1}$, for example as in the original Duda and Szulc (2018) article, although for simplicity we omit this step here - working only on normalized variables. Also, $AMI, HLR$ predictions underwent such normalization for the purpose of Fig. 1 visual performance comparison - which means that a perfect predictor would give a diagonal plot.

The empirical distribution function (EDF) was used for this normalization here: for each variable its $n$ observed values are sorted, then $i$-th value in such an order obtains $(i-0.5)/n$ normalized value. Hence, values become their estimated quantiles this way, a difference of two normalized values describes the percentage of population between these two values.

Having predicted density for normalized variable, we can transform it to the original variable, e.g. by discretizing this density to probability distribution on a $\{(i-0.5)/n\}_{i=1,\ldots,n}$ lattice, and assigning probability of its $i$-th position to $i$-th ordered original value. For simplicity it is omitted in this article.

### 3.4. Evaluation: log-likelihood with 10-fold cross-validation

The most standard evaluation of probability distributions is log-likelihood as in ML estimation: the average (natural) logarithm of the (predicted) density in the actually observed value. Hence, we will use this evaluation here.

Working on variables normalized to $\rho \approx 1$ marginal distributions, without prediction they would have practically zero log-likelihood. This allows to imagine the gains from predictions as an averaged improvement over this $\rho \approx 1$, as in Fig. 2. For example, the best observed log-likelihood $\approx 1$ corresponds to $\approx \exp(1) \approx 2.7$ density: 2.7 times as good as without the prediction, the same as if we could squeeze a $[0,1]$ range 2.7 times to a 0.37 wide range. Sorting the predicted densities into the actually observed values, we can obtain additional information regarding the distribution of prediction, as presented in this Figure.

Here, we predict the conditional density - denoted as $\rho(Y = y|X = x)$ for the density of $Y$ predicted on the basis of the known value of $X$. Hence its evaluation can be seen as an estimation of $E_{XY}(\ln(\rho(Y|X))$, which is minus conditional entropy $-H(Y|X)$. While it is unknown here, random variables have some concrete value of conditional entropy - we can

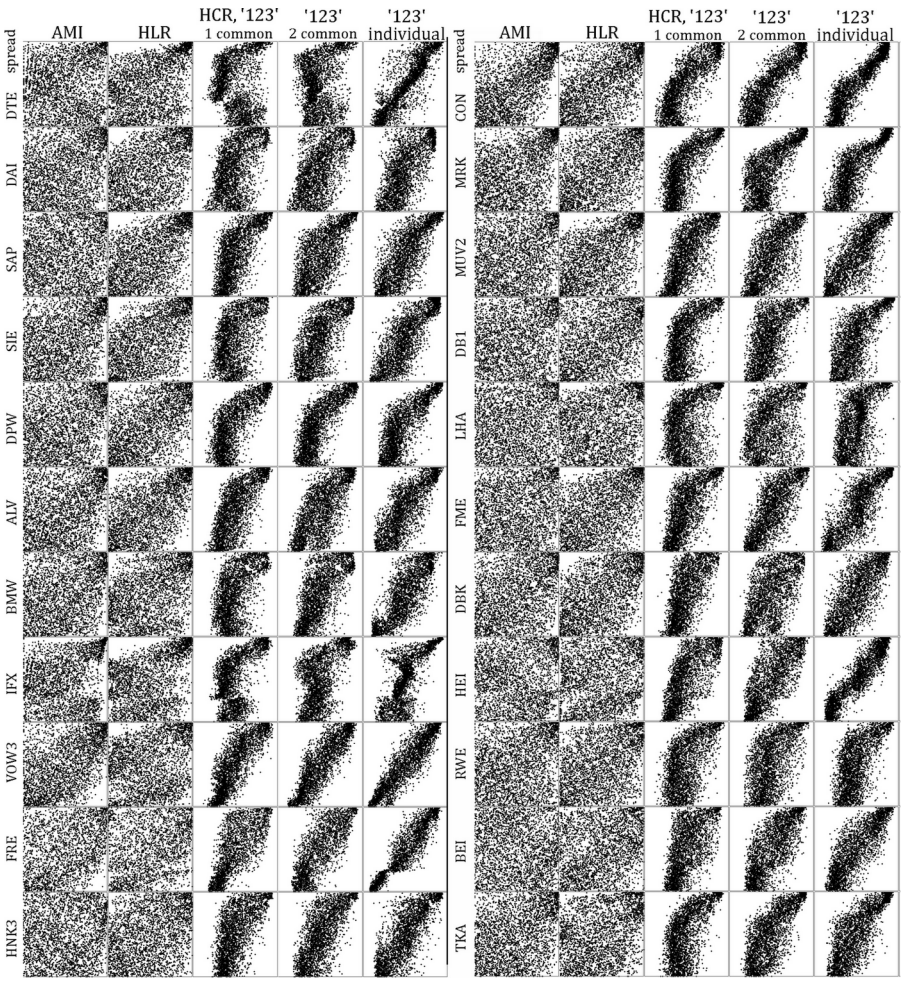



Figure 1: Comparison of spread predictors on data set for visual evaluation: a perfect predictor would give a diagonal scatter plot, a completely useless one would give a uniform distribution. All variables are normalized to nearly uniform marginal distributions, including outcomes of standard methods: *AMI*, *HLR*. The following 3 columns use the expected values of predicted densities from the discussed '123' model (using $P,V,(H-L)/P$ variables, $8 \cdot 53 = 424$ coefficients). The "1 common" column uses one model for all, "2 common" groups companies into two subsets and uses one of two models (as in Fig. 7, using models comL, comR from Fig. 6). The last column uses models individually optimized for each company.

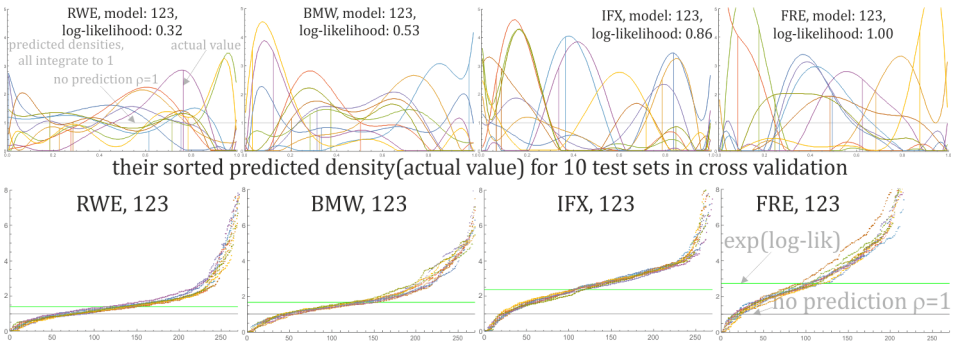



Figure 2: Top: examples of predicted conditional densities. Bottom: evaluation of such a prediction. While log-likelihood only provides averaged $\ln(\rho(y^i|x^i))$, sorted $\rho(y^i|x^i)$ values are presented here, allowing to additionally see, e.g. how frequently such prediction is below $\rho = 1$ threshold of using no prediction. Colours denote one of 10 rounds of 10-fold cross-validation, visualizing dependence of randomly splitting into the training and test set.

hopefully try to approach it with better and better models.

Here, we are focusing on large models that use hundreds of coefficients, estimated from thousands of data points. Hence we need to be careful not to overfit: represent only behaviour which indeed generalizes - is not just a statistical artefact of the training set. Machine learning also builds large models, usually evaluating them using cross-validation: a randomly split data set into a training and test set, the training set is used to build the model, then the test (or validation) set is used to evaluate this model.

However, this evaluation depends on the random splitting into the training and test set. Standard 10-fold cross-validation is used here to weaken this random effect: the data set is randomly split into 10 nearly equal size subsets, the evaluation is an average from 10 cross-validations: using successive subsets as the test set and the remaining ones as the training set. However, a scale $\approx 0.01$ randomness of such an evaluation is still observed, hence for log-likelihoods only two digits after the comma are presented.

## 4. The HCR-based methodology used

This Section discusses the methodology used, which is an expansion of the one used in Duda and Szulc (2018). To predict conditional distribution $\rho(Y|X)$ we decompose $X$ and $Y$ variables into mixed moments and model separately each moment of $Y$ using least-squares linear regression of moments of $X$, then combine them into the predicted conditional probability distribution of $Y$.

### 4.1. Decomposing joint distribution into mixed moments

After normalizing the marginal distributions of all variables to nearly uniform on $[0,1]$, for $d$ variables their joint distribution on $[0,1]^d$ would also be nearly uniform if they were statistically independent. Distortion from uniform joint distribution corresponds to statistical

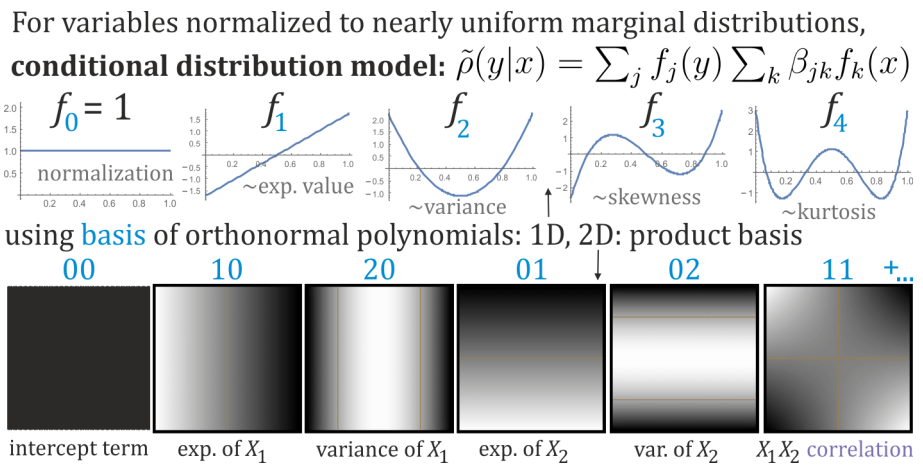



Figure 3: General concept, some first functions of the 1 and 2 dimensional basis of orthornormal polynomials used ($f_{j_1 j_2}(x) = f_{j_1}(x_1) f_{j_2}(x_2)$), and application example. For simplicity we assume working on variables normalized to nearly uniform marginal densities on $[0,1]$. We would like to model distortion from this uniform distribution for the predicted variable $Y$ on the basis of the context $X$: as a linear combination, e.g. of orthornormal polynomials here, for which coefficients have similar interpretation as moments/cumulants: $a_1$ shifts right/left like the expected value, $a_2$ increases/decreases the probability of extreme values as variance, etc.

dependencies between these variables - we would like to model and exploit it.

In HCR we model it as just a linear combination using an orthornormal basis, e.g. of polynomials, which gives the coefficients a similar interpretation as moments and mixed moments: the dependencies between moments for multiple variables. In Fig. 3 the general concept of the HCR methodology is presented.

The first orthonormal ($\int_0^1 f_i(x) f_j(x) dx = \delta_{ij}$) polynomials (rescaled Legendre) for $[0,1]$ are $f_0 = 1$ and $f_1, f_2, f_3, f_4$ correspondingly (plotted in Fig. 3):

$$\sqrt{3}(2x-1), \sqrt{5}(6x^2-6x+1), \sqrt{7}(20x^3-30x^2+12x-1), 3(70x^4-140x^3+90x^2-20x+1)$$

We could alternatively use, e.g. $1, \sqrt{2}\cos(\pi x k)$ for $k \geq 1$ orthonormal basis. However, experimentally this usually leads to inferior evaluation.

Decomposing density $\rho(x) = \sum_j a_j f_j(x)$, we need $a_0 = 1$ normalization to integrate to 1. Due to orthogonality, $\int_0^1 f_j(x) dx = 0$ for $j > 0$, hence the following coefficients do not affect normalization. As we can see in their plots in Fig. 3, positive $a_1$ shifts density toward right - acting analogously as the expected value. Positive $a_2$ increases the probability of extreme values at the cost of central values - analogously as variance. Skewness-like higher order asymmetry is brought by $a_3$ and so on - we can intuitively interpret these coefficients as moments (cumulants). This is only an approximation, but useful for interpreting these models.

In multiple dimensions we can use the product basis:

$$f_j(x) = f_{j_1}(x_1) \cdot \ldots \cdot f_{j_d}(x_d) \qquad \text{for } j = (j_1, \ldots, j_d) \tag{2}$$

leading to a model of joint distribution:

$$\rho(x) = \sum_{j \in B} f_j(x) = \sum_{j \in B} a_j f_{j_1}(x_1) \cdot \ldots \cdot f_{j_d}(x_d) \tag{3}$$

where $B \subset \mathbb{N}^d$ is the basis of the mixed moments we are using for our modelling. It is required that it contains $(0, \ldots, 0)$ for normalization. Besides, there is freedom in choosing this basis, which allows one to hierarchically decompose the statistical dependencies of multiple variables into mixed moments: describing marginal distribution first, then pairwise dependencies, and so on for dependencies of growing numbers of variables.

Fig. 3 contains the first 5 functions of such a product basis for $d = 2$ variables: $f_{00}$ corresponds to normalization and requires $a_{00} = 1$. The coefficients of $f_{10}$, $f_{20}$ describe the expected value and the variance of the first variable, $f_{01}$ and $f_{02}$ analogously of the second. Then we can start including moment dependencies, starting with $a_{11}$, which determines the decrease/increase in the expected value of one variable with the growth in the expected value of the second variable - analogously to the correlation coefficient. We also have dependencies between higher moments, such as asymmetric $a_{12}$, which relates the expected value of the first variable and the variance of the second.

And analogously for more variables, e.g. $a_{010010}$ describes the correlation between the 2nd and 5th out of 6 variables. Finally, we can hierarchically decompose the statistical dependencies between multiple variables into their mixed moments. However, to completely describe the general joint distribution, we would need $B = \mathbb{N}^d$ infinite number of mixed moments for complete expansion - for practical modelling we need to choose the finite basis $B$ of moments to focus on.

### 4.2. Estimation using least squares linear regression

Having a data sample $\mathscr{X}$, we would like to estimate such mixed moments as coefficients for the linear combination of an orthonomal basis of functions, e.g. polynomials. Smoothing the sample using kernel density estimation, finding a linear combination which minimizes the square distance to such a smoothed sample, and performing limit to zero width of the kernel used, we obtain a convenient and inexpensive MSE estimation Duda (2018): independently for each coefficient $j$ as just the average over the data set of value of the corresponding function:

$$a_j = \frac{1}{|\mathscr{X}|} \sum_{x \in \mathscr{X}} f_j(x) \tag{4}$$

We could use this model for predicting conditional distribution: substitute the known variables to the modelled joint distribution, after normalization obtaining the (conditional) density of the unknown variables.

However, for the bid-ask spread prediction problem, a slightly better evaluation was obtained using the generalizing alternative approach of Duda and Szulc (2018), which allows

one to additionally exploit subtle variable dependencies, hence we will focus on this.

Specifically, to model $\rho(Y=y|X=x)$, let us use separate bases of (mixed) moments: $B_X$ for $X$, $B_Y$ for $Y$, and model relations between them. While more sophisticated models could be considered for such relations including neural networks, for simplicity and interpretability we focus on linear models here, treating $f_j(x)$ as interpretable features:

$$\rho(y|x) = \sum_{j\in B_Y} f_j(y) a_j(x) \qquad \text{for} \qquad a_j(x) = \sum_{k\in B_X} \beta_{jk} f_k(x) \tag{5}$$

hence the model is defined by the $|B_Y| \times |B_X|$ matrix $\beta$.

It allows for good interpretability: $\beta_{jk}$ coefficient is linear contribution of $k$-th mixed moment of $X$ to $j$-th (mixed) moment of $Y$. We focus on one-dimensional $Y$, but the formalism allows one to analogously predict density for multidimensional $Y$.

To find the $\beta$ we use least-squares optimization here - it is very inexpensive, can be used independently for each $j \in B_Y$ thanks to the use of an orthonormal basis, and intuitively it is a proper heuristic: least-squares optimization estimates the mean - exactly as we would like for coefficient estimation (4). However, this is not necessarily the optimal choice - it might also be worth exploring more sophisticated ways.

This least-squares optimization has to be performed separately for each $j \in B_Y$. Denoting $\beta_{j\cdot} = (\beta_{jk})_{k\in B_X}$ as a coefficient vector for $j$-th moment and $\mathscr{D} = \{(y^i, x^i)\}_{i=1..n}$ as (e.g. training) data set of $(y,x)$ pairs:

$$\beta_{j\cdot} = \text{argmin}_v \sum_{(y,x)\in\mathscr{D}} \left( \sum_{k\in B_X} f_k(x) v_k - f_j(y) \right)^2 = \text{argmin}_v \left\| Mv - b^j \right\|^2$$

$$\text{for } M = [f_k(x^i)]_{i=1..n, k\in B_X}, \qquad b^j = (f_j(y^i))_{i=1..n}$$

matrix $M$ and vector $b^j$ for $j \in B_Y$. This least-squares optimization has a unique solution:

$$\beta_{j\cdot} = (M^T M)^{-1} M^T b^j \tag{6}$$

Separately calculated for each $j \in B_Y$, leading to the entire model as $\beta$ matrix with $\beta_{j\cdot}$ rows.

### 4.3. Applying the model, enforcing nonnegativity

We can apply the found model $\beta$ to (e.g. test) data points as in (5), obtaining the predicted conditional density for $y$ on $[0,1]$ as a polynomial. However, sometimes it can drop below 0, so let us refer to it as $\tilde{\rho}$ and then enforce the non-negativity required for densities:

$$\tilde{\rho}(y|x) = \sum_{j\in B_Y} f_j(y) \sum_{k\in B_X} \beta_{jk} f_k(x) \tag{7}$$

This polynomial always integrates to 1. However, it can occasionally be below zero, which should be interpreted as corresponding to a low positive density. This interpretation to non-negative density $\rho$ is referred to as calibration, and can be optimized on the basis of the data

set. For simplicity only the following was used:

$$\rho(y|x) = \max\left(\tilde{\rho}(y|x), 0.03\right)/N \tag{8}$$

where $N$ normalization factor is chosen to integrate to 1: $N = \int_0^1 \max\left(\tilde{\rho}(y|x), 0.03\right) dy$. The 0.03 threshold was experimentally chosen as a compromise for the data set used, its tuning can slightly improve evaluation.

### 4.4. Basic basis selection

The optimal choice of the basis is a difficult open question. As the basic choice the combinatorial family was used:

$$\mathscr{B}((m_1,\ldots,m_d),s,r) := \left\{ j \in \mathbb{N}^d : \forall_i j_i \leq m_i,\ \sum_{i=1}^{d} j_i \leq s,\ \sum_{i=1}^{d} \operatorname{sgn}(j_i) \leq r \right\} \tag{9}$$

where $m_i$ chooses how many first moments to use for $i$-th variable, $s$ bounds the sum of used moments (and formally the degree of the corresponding polynomial), $r$ bounds the number of nonzero $j_i$: to include the dependencies of up to $r$ variables.

For example the '123' model infers 8 moments $B_Y = \mathscr{B}((8),8,1)$ from 3 variables using a compromise: $B_X = \mathscr{B}((4,4,4),5,3)$ of size $|B_X| = 53$ basis, directly written, e.g. in Fig. 6.

### 4.5. '123' model using basic variables

The initial plan for this article was to improve prediction from standard models: $AMI, HLR$, trying to predict the conditional distribution of spread from their values using the methodology under discussion. However, the results were disappointing, especially for $AMI$, as we can see in Fig. 1.

Therefore, we decided to use the original variables $(P,V,L,H,R)$ instead, which turned out to lead to essentially better predictions. A search for parameters using $\mathscr{B}$ basic basis selection (9) was performed manually to maximize the averaged log-likelihood in 10-fold cross-validation. This search finally leads to $B_X = \mathscr{B}((4,4,4),5,3)$ basis for only 3 variables: $P, V, (H-L)/P$ to predict up to the 8-th moment of $Y$. Surprisingly, adding dependence on $R$ and $L$ alone worsened the evaluation - their dependence did not generalize from training to test sets, hence they are not used in the final model.

The top of Fig. 2 contains examples of conditional densities predicted. The predicted $\tilde{\rho}(y|x^i) = \sum_j f_j(y) \sum_k \beta_{jk} f_k(x^i)$ polynomial for $i$-th data point undergoes $\rho = \max(\tilde{\rho}, 0.03)/N$ to remove negative densities, and normalization to integrate to $1 = \int_0^1 \rho(y|x)dy$. Each diagram contains 10 example predictions, vertical lines show the actual values $(y^i, \rho(y^i|x^i))$: the higher the better prediction, without prediction all would have height 1. Companies were chosen to present prediction examples of various evaluation levels. The best ones predict mainly narrow unimodal distributions in line with the actual values, although weaker ones can usually only predict wide often multimodal distributions. We can see rapid growths at the ends - they are likely artefacts of using polynomials, their additional removal might

improve prediction. The bottom part presents their sorted predicted densities in the actual values $\{\rho(y^i|x^i)\}_i$, with marked gray $\rho = 1$ line of using no prediction and green exp(log-likelihood) line corresponding to average improvement over no prediction. The points are of different colours denoting one of 10 rounds of 10-fold cross-validation.

Integration required for normalization is relatively costly to compute, especially in higher dimensions, hence for efficient calculation the predicted polynomial $\tilde{\rho}$ was discretized here into 100 values on a $((i-0.5)/100)_{i=1,\ldots,100}$ lattice, which corresponds to approximating the density with a piecewise constant function on length 1/100 subranges. Then $\max(\cdot, 0.03)$ was applied, and division by the sum for normalization. Finally, the density in discretized $\lceil 100y^i \rceil/100$ position was used as $\rho(y^i|x^i)$ in the log-likelihood evaluation.

In Figure 4 the results of cross-validation are presented. Model '123' denotes using the three basic variables: where '1' denotes the closing price ($P$), '2' volume ($V$), and '3' the difference between high and low price normalized by dividing by the closing price: $(H-L)/P$. The last column presents the averaged evaluation for using common model for all data. We can also see that there are large differences between companies, hence we will mostly focus on building individual models for each company. The three lowest dots correspond to predicting from single variable, then evaluation grows when adding information from succeeding variables.

Copulas are a general, well-established method of modelling multivariate distribution. In higher dimensions r-vines are a flexible class of mulivariate distributions. This type of copulas allows for flexible modelling of asymmetric and nonlinear dependence patterns Gurgul and Machno (2016). For comparison purposes we estimated such models and it turns out that on average log-likelihoods for individual model from copulas were smaller than from HCR. In Figure 4 points denoted by "123vc" correspond to results from r-vines. On average, log-likelihood for individual HCR models was 0.603, while for vine-copulas it was 0.366, getting better representation of complex behaviour thanks of allowing for high parametric models. HCR also has much less expensive estimation (least squares regression of moments), and interpretation of the found parameters as moment dependencies.

While the optimal choice of the basis seems a difficult open problem, an exhaustive search over all subsets is rather impractically costly, Figure 5 presents some heuristic approaches. The $\mathscr{B}$ family seems generally a good start, e.g. to successively modify some its parameter by one as long as improvement is observed. In this Figure we can see a large improvement while the number of predicted moments rises up to $\approx 7$, which suggests that the complexity of the conditional distributions for this problem requires this degree of polynomial in order to be described properly. This Figure also contains trials of using different orders of some first mixed moments. The selective removal, which is presented there, seems a reasonable optimization: for each mixed moment from $B_X$ calculate the evaluation when it is removed, finally remove the one that leads to the best evaluation, and so on as long as the evaluation improves.

Examples of $\beta$ matrix are visualized in Fig. 6 for $|B_X| = 53$, $|B_Y| = 1+8$. Trying to split all companies into subsets of similar behaviour, as visualized in tree Fig. 7, splitting into two subsets we obtain the comL and comR models - correspondingly for the left (DPW, BEI, HNK3, FME, SAP, DB1, RWE, FRE, HEI, DTE, IFX) and right (DAI, SIE, TKA, CON, MRC, LHA, VOW3, MUV2, ALV, BMW, DBK) subtree of this tree. Then individual

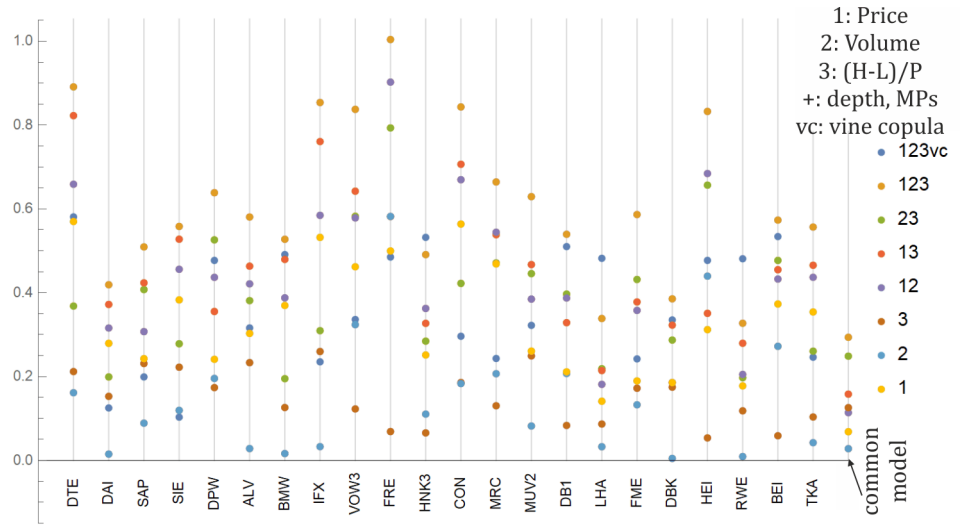



Figure 4: Log-likelihoods from 10-fold cross-validation for individual models for companies using various types of information. We can see individual behaviour of companies and growth of prediction evaluation while adding information from succeeding variables. The "123vc" points correspond to vine copulas using the same evaluation: for HCR average log-likelihood for individual models was 0.603, while for vine-copulas it was 0.366, additionally requiring $\approx 100\times$ more computational time.

models for 5 selected companies were presented. The rows correspond to the predicted moments of $Y$, as linear combinations of mixed moments of $X$ corresponding to columns. Row zero has always only 000 nonzero coefficient equal to 1 for normalization. The next row describes the prediction of the expected value, the next one of variance and so on. In the top model, common for all companies, we can, e.g. see large positive $001 \to 1$ coefficient: the spread increases with the growth of $H - L$, negative $010 \to 1$: the spread decreases with growth of $V$, and negative $011 \to 2$: variance of spread decreases for correlated $V$ and $H - L$. Blue $100 \to 3$ for FRE denotes a reduction in skewness of spread with growth of price. Generally, we can see rather individual behaviour for different companies, starting with $100 \to 1$ analogous to the price-spread correlation, which seems the main dividing factor between comL and comR companies.

### 4.6. Individual vs common models, universality

A natural question is how helpful for prediction a given variable is - Fig. 4 presents some answers by calculating the log-likelihood also for models using only some of the variables. We can see different companies can have very different behaviour here, e.g. for some $V$ is helpful (volume and spread are correlated), for some it is not. Fig. 6 shows that they can even display the opposite behaviour: e.g. for $100 \to 1$ dependence on price.

It is a general lesson that while we would like predictors to be nice simple formulas, the reality might be much more complicated - the models found here are the results of the

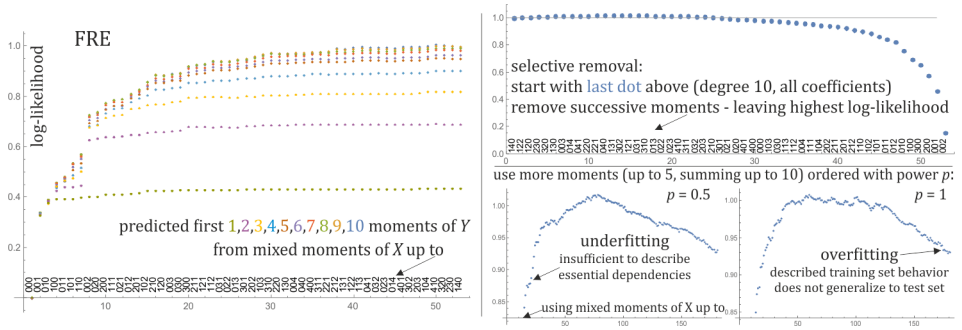



Figure 5: Left: Optimizing the basis and model size using the example of the company FRE and $B_X = \mathscr{B}((4,4,4),5,3)$ size 53 basis of mixed moments from '123' model. Log-likelihoods for predicting the first $1\ldots10$ moments (denoted by colours) using some first of mixed moments (sorted lexicographically) of 3 $X$ variables: $P,V,(H-L)/P$. We can see that we should predict $\approx 8$ moments, higher moments are necessary to represent more complex distributions. Top right: selective removal of successive mixed moments to maximize log-likelihood - we can see that we can slightly improve evaluation this way, additionally reducing the model size. However, it requires individual optimization for each company. Bottom right: analogously as top, but using size 181 larger $B_X = \mathscr{B}((5,5,5),10,3)$, also trying different orders of mixed moments: accordingly to $\sum_i (j_i)^p$. While using all such mixed moments clearly leads to overfitting, selectively using some of the first ones can lead to slightly improved evaluation.

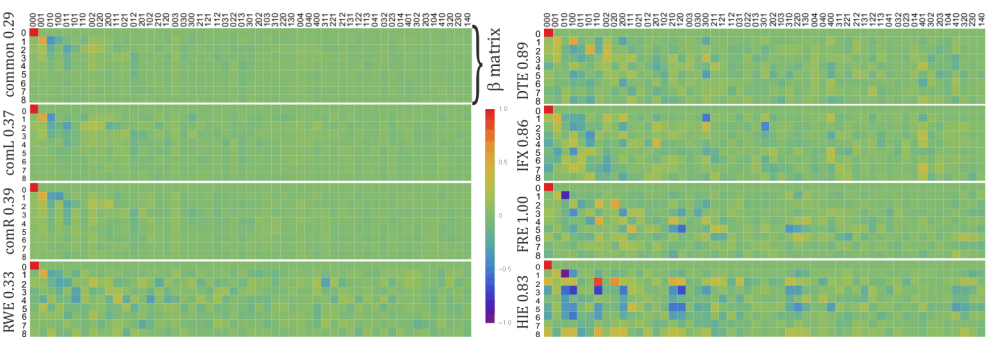



Figure 6: Visualized coefficients of '123' models ($9 \times 53$ matrix $\beta$ for $\rho(y|x) = \sum_j f_j(y) \sum_k \beta_{jk} f_k(x)$) for $(P,V,H-L)$ variables, the numbers above the names are log-likelihoods. The 'common' is the model built for all the data combined - it presents general trends. The 'comL' and 'comR' models are for the left (DPW, BEI, HNK3, FME, SAP, DB1, RWE, FRE, HEI, DTE, IFX) and right (DAI, SIE, TKA, CON, MRC, LHA, VOW3, MUV2, ALV, BMW, DBK) subtree in Fig. 7 - we can see that these subsets of companies mainly differ by $100 \to 1$ coefficient corresponding to correlation between price and spread.

cultures of traders of the stocks of individual companies, which can essentially vary between companies.

Therefore, to obtain the most accurate predictions we should build individual models

for each company. Furthermore, a specific behaviour of a given company can additionally evolve in time - which could be exploited, e.g. by building separate models for shorter time periods, or using adaptive least-squares linear regression Duda (2019), and this is planned for future investigation.

However, building such models requires training data, which in the case of variables like bid-ask spread might be difficult to access. Hence, it is also important to search for universality - e.g. try to guess a model for a company for which we lack such data, on the basis of the available information for other companies. This generally seems a very difficult problem, Fig. 7 shows that even having all the data, using the common model for multiple companies we should expect a large evaluation drop. For example, we can see that the behaviour of DTE completely disagrees with the common model for all.

As we can see in this tree Figure, the use of common model situation improves if we can cluster companies into groups of similar behaviour - results are also presented for splitting companies into just two groups with separate models (comL, comR in Fig. 6), also visually leading to slightly better predictions as we can see comparing the 3rd and 4th column in Fig. 1. The heights of the names show the evaluation of using an individual model for a given company, orange dots show the successive reduction of log-likelihood for a given company while using common models for subsets that grow according to this tree. The lowest dots correspond to the use of one common model for all (common in Fig. 6) we can see that it is worse than zero only for DTE (we get zero when using no prediction at all). Splitting companies into a left and right subtree and using separate two models for them (comL and comR in Fig. 6), we essentially obtain a better prediction (one dot up). The tree structure was calculated by combining subsets to maximize (log-likelihood of common model / average log-likelihood of individual models) - grouping companies into pairs and then further, up to a single common model for all. The positions of lines represent such grouped companies: a light-gray line their averaged log-likelihoods of individual models, dark-gray line their log-likelihood for a common model. The difference between these two lines represent a loss while using the common model. The common models are fixed hence there is no cross-validation (CV) used, which artificially improves performance, for example for the first dot of FME corresponding to the common model with HNK - making it above CV individual model, generally suggesting large time inhomogeneities - to be included in future adaptive models.

## 5. Conclusions and further work

A general methodology has been presented for extracting and exploiting complex statistical dependencies between multiple variables in an inexpensive and interpretable way for predicting conditional probability distributions, using the example of the difficult problem of predicting bid-ask spreads from more accessible information. This expands the approach of Duda and Szulc (2018) by inferring from mixed moments, and searching for a basis in large spaces of possibilities.

Figure 1 presents a comparison between it and standard methods when using only the expected value from such predicted conditional density. A perfect predictor would lead to diagonal scatter plot, standard methods provide rather a noise instead, while the predictions

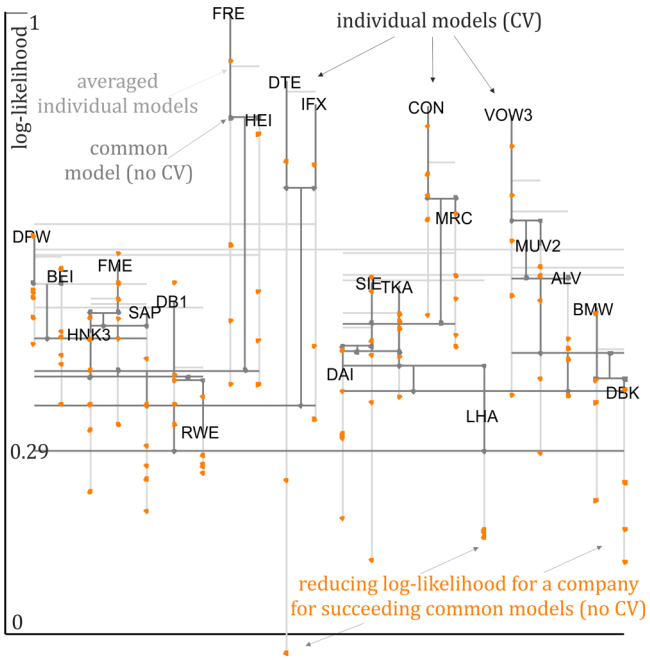



Figure 7: Visualization of optimized hierarchical grouping and evaluation loss while using common models for multiple companies, the height denotes log-likelihoods. It was constructed by starting with individual models, then successively joining subsets of companies leading to lowest loss of evaluation while using a common model for them.

from the approaches discussed indeed often resemble diagonal plot, especially when using individual models. The predicted conditional probability density provides much more information than the value alone: e.g. it allows one to additionally estimate the uncertainty of such a prediction as value, or provide prediction for multimodal densities, or it allows random values to be generated, e.g. for Monte-Carlo simulations, or just provides the entire density for accurate considerations especially if transforming such random variables through some further nonlinear functions.

There are many directions for further development of this relatively new general methodology, for example:

- Optimal choice of the basis is a difficult problem, which should be automatized especially for a larger number of variables - selecting from the basis of orthonormal polynomials discussed, or maybe automatically optimizing a completely different basis on the basis of a data set.
- Large differences between the behaviours of individual companies have been observed - raising difficult questions regarding how to optimize for common behaviour, optimize models on the basis of an incomplete information, etc. Additionally, such behaviour has probably also time inhomogeneity - the models should evolve in time,

requiring adaptive models to improve performance, where the problem of data availability becomes even more crucial.

- These models rapidly grow with the number of variables, which requires some modifications for exploiting high dimensional information - like extracting features from these variables, e.g. as averages, dimensionality reduction like PCA, etc.
- We have predicted the conditional distributions for one-dimensional variables, but the methodology was introduced to be more general: predicting for multidimensional $Y$ should be just a matter of using proper $B_Y$, which is planned to be tested in the future.
- The densities predicted as polynomials often have rapid growths at the ends of $[0, 1]$ - their removal might improve performance.
- A linear relation was assumed between moments with least-squares optimization, which is inexpensive and has good interpretability, but is not necessarily optimal - one could consider, e.g. using neural networks instead, and optimizing criteria closer to the log-likelihood of final predictions.
- In the light of the Epps effect we can see the dependence of stock return cross-correlations on the data sampling frequency, i.e. for high-resolution data the cross-correlations are significantly smaller than their asymptotic value as observed for daily data. One should check the performance of HCR with respect to the data sampling frequency.
- The share of algorithmic trading in the market is growing. The HCR method may be helpful in the forecast of quoted and effective bid-ask spread regressed on the share of algorithmic trading in the market.
- A comparison of the results of bid-ask spread modelling and forecasting using HCR methodology with respect to the microstructure of stock markets in particular countries, their size and the level of development.

## Acknowledgement

We would like to thank the Editors of this journal, and anonymous Referees for their valuable comments on earlier versions of the paper.

Henryk Gurgul thanks Professor Roland Mestel for providing the bid-ask data from the data bank ”Finance Research Graz Data Services” and Professor Erik Theissen and Stefan Scharnowski from Mannheim for providing data from the ”Market Microstructure Database”. Henryk Gurgul was financed by AGH University of Science and Technology in Krakow (institutional subsidy for maintaining Research Capacity Grant 16.16.200/396.)

## References

ABDI, F., RANALDO, A., (2017). A simple estimation of bid-ask spreads from daily close, high, and low prices. The Review of Financial Studies 30(12), pp. 4437–4480

AMIHUD, Y., (2002). Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 5(1), pp. 31–56

ARMITAGE, S, BRZESZCZYŃSKI, J., SERDYUK, A., (2014). Liquidity measures and cost of trading in an illiquid market. Journal of Emerging Market Finance 13(2), pp. 155–196

BĘDOWSKA-SÓJKA, B., ECHAUST, K., (2019). Commonality in Liquidity Indices: The Emerging European Stock Markets. Systems 7(2), pp. 7–24

BESSEMBINDER, H., (2003). Issues in assessing trade execution costs. Journal of Financial Markets 6(3), pp. 233–257

BUTLER, A. W., GRULLON, G., WESTON, J. P., (2005). Stock market liquidity and the cost of issuing equity. Journal of Financial and Quantitative Analysis 40(2), pp. 331–348

CHAN, L. K., LAKONISHOK, J., (1993). Institutional trades and intraday stock price behavior. Journal of Financial Economics 33(2), pp. 173–199

CHEN, X., LINTON, O., YI, Y., (2017). Semiparametric identification of the bid–ask spread in extended Roll models. Journal of Econometrics 200(2), pp. 312–325

CHRISTIE, W. G., SCHULTZ, P. H., (1994). Why do NASDAQ market makers avoid odd-eighth quotes? The Journal of Finance 49(5), pp. 1813–1840

CORWIN, S. A., SCHULTZ, P., (2012). A simple way to estimate bid-ask spreads from daily high and low prices. The Journal of Finance 67(2), pp. 719–760

DUDA, J., (2018). Exploiting statistical dependencies of time series with hierarchical correlation reconstruction. arXiv preprint arXiv:180704119

DUDA, J., (2019). Parametric context adaptive Laplace distribution for multimedia compression. arXiv preprint arXiv:190603238

DUDA, J., SZULC, A., (2018). Credibility evaluation of income data with hierarchical correlation reconstruction. arXiv preprint arXiv:181208040

DUDA, J., SYREK, R., GURGUL, H., (2019). Modelling bid-ask spread conditional distributions using hierarchical correlation reconstruction. arXiv preprint arXiv:191102361

FONG, K. Y., HOLDEN, C. W., TRZCINKA, C. A., (2017). What are the best liquidity proxies for global research? Review of Finance 21(4), pp. 1355–1401

GAJEWSKI, J. F., GRESSE, C., (2007). Centralised order books versus hybrid order books: A paired comparison of trading costs on NSC (Euronext Paris) and SETS (London Stock Exchange). Journal of Banking & Finance 31(9), pp. 2906–2924

GURGUL, H., MACHNO, A., (2016). Modeling dependence structure among European markets and among Asian-Pacific markets: a regime switching regular vine copula approach. Central European Journal of Operations Research 24(3), pp. 763–786

GURGUL, H., MACHNO, A., (2017). The impact of asynchronous trading on Epps effect on Warsaw Stock Exchange. Central European Journal of Operations Research 25(2), pp. 287–301

GURGUL, H., SYREK, R., (2019). Dependence Structure of Volatility and Illiquidity on Vienna and Warsaw Stock Exchanges. Finance a Uver: Czech Journal of Economics & Finance 69(3), pp. 298–321

HANSCH, O., NAIK, N. Y., VISWANATHAN, S., (1999). Preferencing, internalization, best execution, and dealer profits. The Journal of Finance 54(5), pp. 1799–1828

HUANG, R. D., STOLL, H. R., (1996). Dealer versus auction markets: A paired comparison of execution costs on NASDAQ and the NYSE. Journal of Financial Economics 41(3), pp. 313–357

LESMOND, D. A., OGDEN, J. P., TRZCINKA, C. A., (1999). A new estimate of transaction costs. The Review of Financial Studies 12(5), pp. 1113–1141

MESTEL, R., MURG, M., THEISSEN, E., (2018). Algorithmic trading and liquidity: Long term evidence from Austria. Finance Research Letters 26, pp. 198–203

NAIK, N. Y., YADAV, P. K., (2003). Trading costs of public investors with obligatory and voluntary market-making: Evidence from market reforms. In: EFA 2003 Annual Conference Paper, 408

REISS, P. C., WERNER, I. M., (1996). Transaction costs in dealer markets: Evidence from the London Stock Exchange. In: The Industrial Organization and Regulation of the Securities Industry, University of Chicago Press

ROLL, R., (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. The Journal of Finance 39(4), pp. 1127–1139

STOLL, H. R., (2000) Presidential address: friction. The Journal of Finance 55(4), pp. 1479–1514

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 119–132, DOI 10.21307/stattrans-2020-058*
Received – 24.03.2019; accepted – 17.06.2020

# Unbiased estimator modeling in unrelated dichotomous randomized response

**Adetola Adedamola Adediran[1], Femi Barnabas Adebola[2], Olusegun Sunday Ewemooje[3]**

## ABSTRACT

The unrelated design has been shown to improve the efficiency of a randomized response method and reduces respondents' suspicion. In the light of this, the paper proposes a new Unrelated Randomized Response Model constructed by incorporating an unrelated question into the alternative unbiased estimator in the dichotomous randomized response model proposed by Ewemooje in 2019. An unbiased estimate and variance of the model are thus obtained. The variance of the proposed model decreases as the proportion of the sensitive attribute π_A and the unrelated attribute π_U increases, in contrast to the earlier Ewemooje model, whose variance increases as the proportion of the sensitive attribute increases. The relative efficiency of the proposed model over the earlier Ewemooje model decreases as π_U increases when $0.1 \leq \pi\_A \leq 0.3$ and increases as π_U increases when $0.35 \leq \pi\_A \leq 0.45$. Application of the proposed model also revealed its efficiency over the direct method in estimating the prevalence of examination malpractices among university students;the direct method gave an estimate of 19.0%, compared to the proposed method's estimate of 23.0%. Hence, the proposed model is more efficient than the direct method and the earlier Ewemooje model as the proportion of people belonging to the sensitive attribute increases.

**Key words:** dichotomous, relative efficiency, sensitive attribute.

## 1. Introduction

One of the problems in a survey is non-response; this is referred to as failure of getting the required information from a respondent. Non-response reduces the sample size as some respondents do not give the needed information and thereby making the

[1] Department of Statistics, Federal University of Technology Akure, Nigeria. E-mail: aaadediran@futa.edu.ng. ORCID: https://orcid.org/0000-0003-3176-7872.
[2] Department of Statistics, Federal University of Technology Akure, Nigeria. E-mail: fbadebola@futa.edu.ng. ORCID: https://orcid.org/0000-0001-7790-1331.
[3] Department of Statistics, Federal University of Technology Akure, Nigeria. E-mail: osewemooje@futa.edu.ng. ORCID: https://orcid.org/0000-0003-3236-6018.

accuracy of the estimate to be compromised. Obtaining information about sensitive attributes lead to non-response or false response as participants in the sample may give false response or decide not to give an answer for diverse reasons. In order to reduce error due to this non-response bias, Warner in 1965 developed the Randomized Response Model (RRM) for estimating the proportion of people that belong to a sensitive attribute.

Quite a number of authors have reviewed and expanded the work of Warner, including Horvitz *et al.* (1967) Unrelated Question Design, Greenberg *et al.* (1969) Unrelated Question Design with known distribution, Mangat and Singh (1990) Randomized Response Model (RRM), Hussain-Shabbir (2007) Dichotomous Randomized Response Model (DRRM), Adebola and Adepetun (2011), Tripartite Randomized Response Model (TRRM), Ewemooje (2017) Equal Probabilities of Protection, Adebola *et al.* (2017) Hybrid Tripartite Randomized Response Technique, Ewemooje *et al.* (2019a) Dichotomous Randomized Response Technique, Ewemooje *et al.* (2018) Stratified Hybrid Tripartite Randomized Response Technique. Also, Yu *et al.* (2008) worked on the Crosswise Model (CM) and Triangular Model (TM) while Fox *et al.* (2019) proposed Generalized Linear Mixed Models for Randomized Responses (GLMRR), among others.

To test the applicability of the RRM; Jann *et al.* (2012) applied a modified RRM (Crosswise Model by Yu *et al.,* 2008) to elicit information on plagiarism among German and Swiss students. They found out that RRM elicited more socially undesirable answers than direct questioning. Ewemooje *et al.,* (2017) also used Improved Randomized Response Technique for two sensitive attributes (IRRT2) to show that RRM performs better that Direct Method of questioning (DM) by estimating prevalence of induced abortion and multiple sexual partners. Cobo *et al.,* (2016) used RRM to investigate cannabis use by Spanish University students and then compared the result with DM. Their results revealed that RRM increases the response rate for cannabis use and that it is an efficient method. Furthermore, Ewemooje *et al.,* (2019b) measured substance use disorder prevalence using RRM and DM; their findings showed that RRM estimated the disorder better with lower error than DM. Conversely, Hoglinger and Jann (2018) evaluated the variability of several variants of RRM and the crosswise model by comparing the respondents' self-reports on cheating in dice games to actual cheating behaviour; their result showed that the RRM fails to reduce the level of misinterpreting compared to DM and none of the RRMs evaluated outperformed the conventional DM.

Therefore, in this work we consider dichotomous randomized response design in the presence of unrelated questions; the estimator and variance are obtained and compared with the Dichotomous Randomized Response Model by Ewemooje *et al.,* (2019a) using relative efficiency. Also, to verify more-is-better assumption, the proposed method and Direct Method (DM) were applied to the same subpopulation in a survey.

## 2. Dichotomous Randomized Response Model by Ewemooje *et al.,* (2019a)

In their model, respondents were asked sensitive question directly, if he/she responds "yes" then he/she is not allowed to use the randomized device while if "no", he/she is required to use the randomized device. Two randomized devices were used each consisting of two questions with different selection probabilities. A simple random sample with replacement sampling was adopted in their selection of the sample of size, n with α and β as any two positive real numbers such that $q = \frac{\alpha}{\alpha+\beta}$ is the probability of using the first randomized device and $1- q = \frac{\beta}{\alpha+\beta}$ is the probability of using the second randomized device.

If all respond truthfully, their population proportion of "yes" answers is given by:

$$P(yes) = \theta_1 = \pi + \frac{\alpha}{\alpha+\beta}(1-P_1)(1-\pi) + \frac{\alpha}{\alpha+\beta}(1-P_2)(1-\pi) \quad (1)$$

where $P_1$ is the probability of the sensitive attribute in randomized devices $R_1$ and $P_2$ is the probability of the sensitive attribute in randomized devices $R_2$.

This yielded an unbiased estimate of the population proportion as:

$$\hat{\pi} = \frac{\hat{\boldsymbol{\theta}}_1(\alpha+\beta) - P_2\alpha - P_1\beta}{P_1\alpha + P_2\beta} \quad (2)$$

The variance of their estimate was given as

$$V(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(P_2\alpha+P_1\beta)}{n(P_1\alpha+P_2\beta)^2} \quad (3)$$

## 3. Proposed Model

In sampling a finite population, the simple random sample with replacement was used to obtain the sample size of respondents who respond to sensitive questions using Randomized Response Model. Sensitive question was asked directly from the respondents. If "yes" answer is obtained, he/she does not need to use the randomized device but if he/she answers "no", then he/she uses the randomized device. The two randomized devices $R_1$ and $R_2$ consists of two unrelated questions (the sensitive question A in which the interviewer is interested in with probability P, and non-sensitive attribute question B that is unrelated to the sensitive question A with probability, 1-P) each. Say:

Sensitive question: "do you belong to a sensitive attribute A?"

Non-sensitive question: "do you love soccer?"

Two responses were considered for each of the two unrelated questions: "yes" and "no", where α and β are positive real numbers such that $q = \frac{\alpha}{\alpha+\beta}, \alpha \neq \beta$ is the

probability of using $R_1$ and $1-q=\frac{\beta}{\alpha+\beta}, \alpha \neq \beta$ is the probability of using $R_2$ with preset probabilities $P_1$ and $P_2$ respectively for each of the devices.

Let $\pi_A$ be the true proportion of people that belongs to the sensitive attribute and $\pi_u$, the proportion of people that belongs to the unrelated non-sensitive attribute. If all respond truthfully as the devices provide protection for respondents, the population proportion of "yes" answers is given by:

$$P(yes)=\theta=\pi_A+\frac{\alpha}{\alpha+\beta}[P_1\pi_A+(1-P_1)\pi_U]+\frac{\beta}{\alpha+\beta}[P_2\pi_A+(1-P_2)\pi_U] \quad (4)$$

where $P_1$ is the probability of the sensitive attribute in randomized devices $R_1$ while $P_2$ is the probability of the sensitive attribute in randomized devices $R_2$.

Solving equation (4) further yield the estimate of the population proportion of the sensitive attribute

$$\hat{\pi}_A=\frac{\hat{\theta}(\alpha+\beta)-\pi_U((\alpha+\beta)-\alpha P_1-\beta P_2)}{(\alpha+\beta+\alpha P_1+\beta P_2)} \quad (5)$$

where $\hat{\theta}=n_0/n$, $n_0$ is the number of respondents that answered "yes" to sensitive question while $n$ is the sample size.

The proposed estimator, $\hat{\pi}_A$, is an unbiased estimator of the population parameter $\pi_A$.

### 3.1. Variance Estimation

The variance of the model is obtained as follows:

$$v(\hat{\pi}_A)=v\left(\frac{\hat{\theta}(\alpha+\beta)-\pi_U((\alpha+\beta)-\alpha P_1-\beta P_2)}{(\alpha+\beta+\alpha P_1+\beta P_2)}\right)$$

$$v(\hat{\pi}_A)=\frac{(\alpha+\beta)^2 v(\hat{\theta})}{(\alpha+\beta+\alpha P_1+\beta P_2)^2} \quad (6)$$

where $v(\hat{\theta})=\frac{\theta(1-\theta)}{n}$

recall that $\theta=\left(\frac{\alpha\pi_A+\beta\pi_A+\alpha P_1\pi_A+\beta P_2\pi_A+\alpha(1-P_1)\pi_U+\beta(1-P_2)\pi_U}{\alpha+\beta}\right)$, substituting this in equation (6), the variance of the proposed unbiased estimator is given as:

$$v(\hat{\pi}_A)=\frac{\pi_A\{(\alpha+\beta)-\pi_A(\alpha+\beta+\alpha P_1+\beta P_2)\}}{n(\alpha+\beta+\alpha P_1+\beta P_2)}+\frac{\pi_U(\alpha+\beta-\alpha P_1-\beta P_2)(\alpha+\beta-2\pi_A(\alpha+\beta+\alpha P_1+\beta P_2))}{n(\alpha+\beta+\alpha P_1+\beta P_2)^2} \quad (7)$$

Therefore, the variance of the proposed unbiased estimator can be estimated using:

$$\hat{v}(\hat{\pi}_A)=\frac{\hat{\pi}_A\{(\alpha+\beta)-\hat{\pi}_A(\alpha+\beta+\alpha P_1+\beta P_2)\}}{(n-1)(\alpha+\beta+\alpha P_1+\beta P_2)}+\frac{\pi_U(\alpha+\beta-\alpha P_1-\beta P_2)(\alpha+\beta-2\hat{\pi}_A(\alpha+\beta+\alpha P_1+\beta P_2))}{(n-1)(\alpha+\beta+\alpha P_1+\beta P_2)^2} \quad (8)$$

## 4. Efficiency Comparison

The proposed model will be more efficient than the conventional one if the condition for the relative efficiency holds:

$$\text{RE} = \frac{\text{variance of conventional model}}{\text{variance of proposed model}} > 1$$

The relative efficiency of the proposed model over the conventional model were gotten for varying sample sizes (n), varying probabilities $P_1$ and $P_2$ of using the randomized devices at different values of $\pi_A$ and $\pi_U$.

The comparison between the proposed estimator and Ewemooje *et al.* (2019a) estimator at different sample sizes in Table 1 shows that the proposed estimator is approximately ten (10) times more efficient than that due to Ewemooje *et al.* (2019a). As the sample size increases from 50 to 500, the variances due to Ewemooje *et al.* (2019a) estimator reduces from 0.0053 to 0.0005 while the proposed estimator reduces from 0.0005 to 0.0001. Therefore, as the sample sizes increases the variability reduces, this implies consistency of the two models.

Considering a constant sample size at varying probabilities of selecting the randomized device, the variances due to Ewemooje *et al.* (2019a) estimator increases from 0.00131 to 0.00138, the proposed estimator increases from 0.00018 to 0.00022 while the relative efficiency reduces from 7.089 to 6.227 as shown in Table 2.

**Table 1.** Relative efficiency comparison between the proposed model and Ewemooje *et al.* (2019a) model when $\pi_A = 0.5$; $\pi_U = 0.5$; $P_1 = 0.5$; $P_2 = 0.5$; $\alpha = 25$; $\beta = 35$ for varying sample sizes (n).

| n | $\pi_A$ | $\pi_U$ | $P_1$ | $P_2$ | α | β | $v(\hat{\pi})$ | $v(\hat{\pi}_A)$ | RE |
|---|---|---|---|---|---|---|---|---|---|
| 50 | 0.5 | 0.5 | 0.5 | 0.5 | 25 | 35 | 0.005333 | 0.000537 | 9.931034 |
| 100 | 0.5 | 0.5 | 0.5 | 0.5 | 25 | 35 | 0.002667 | 0.000269 | 9.931034 |
| 150 | 0.5 | 0.5 | 0.5 | 0.5 | 25 | 35 | 0.001778 | 0.000179 | 9.931034 |
| 200 | 0.5 | 0.5 | 0.5 | 0.5 | 25 | 35 | 0.001333 | 0.000134 | 9.931034 |
| 250 | 0.5 | 0.5 | 0.5 | 0.5 | 25 | 35 | 0.001067 | 0.000107 | 9.931034 |
| 300 | 0.5 | 0.5 | 0.5 | 0.5 | 25 | 35 | 0.000889 | 0.0000895 | 9.931034 |
| 350 | 0.5 | 0.5 | 0.5 | 0.5 | 25 | 35 | 0.000762 | 0.0000767 | 9.931034 |
| 400 | 0.5 | 0.5 | 0.5 | 0.5 | 25 | 35 | 0.000667 | 0.0000671 | 9.931034 |
| 450 | 0.5 | 0.5 | 0.5 | 0.5 | 25 | 35 | 0.000593 | 0.0000597 | 9.931034 |
| 500 | 0.5 | 0.5 | 0.5 | 0.5 | 25 | 35 | 0.000533 | 0.0000537 | 9.931034 |

**Table 2.** Relative efficiency comparison between the proposed model and Ewemooje *et al.* (2019a) model when $\pi_A = 0.5$; $\pi_U = 0.5$; $\alpha = 25$; $\beta = 35$; n=200 for varying $P_1$ and $P_2$

| n | $\pi_A$ | $\pi_U$ | $P_1$ | $P_2$ | α | β | $v(\hat{\pi})$ | $v(\hat{\pi}_A)$ | RE |
|---|---|---|---|---|---|---|---|---|---|
| 200 | 0.5 | 0.5 | 0.1 | 0.9 | 25 | 35 | 0.001306 | 0.000184 | 7.089034 |
| 200 | 0.5 | 0.5 | 0.2 | 0.8 | 25 | 35 | 0.001312 | 0.000188 | 6.966797 |
| 200 | 0.5 | 0.5 | 0.3 | 0.7 | 25 | 35 | 0.001318 | 0.000193 | 6.848074 |
| 200 | 0.5 | 0.5 | 0.4 | 0.6 | 25 | 35 | 0.001325 | 0.000197 | 6.733115 |
| 200 | 0.5 | 0.5 | 0.5 | 0.5 | 25 | 35 | 0.001333 | 0.000201 | 6.622212 |
| 200 | 0.5 | 0.5 | 0.6 | 0.4 | 25 | 35 | 0.001342 | 0.000206 | 6.515705 |
| 200 | 0.5 | 0.5 | 0.7 | 0.3 | 25 | 35 | 0.001352 | 0.000211 | 6.413994 |
| 200 | 0.5 | 0.5 | 0.8 | 0.2 | 25 | 35 | 0.001363 | 0.000216 | 6.317551 |
| 200 | 0.5 | 0.5 | 0.9 | 0.1 | 25 | 35 | 0.001376 | 0.000221 | 6.226937 |

Table 3 shows that for varying $\pi_A$ and $\pi_U$, $P_1 = 0.3$; $P_2 = 0.7$, the variance of the Ewemooje *et al.* (2019a) model increases at all values of $\pi_A$ while the variance of the proposed model increases as $\pi_U$ increases when $0.1 \leq \pi_A \leq 0.3$ and decreases as $\pi_U$ increases when $0.35 \leq \pi_A \leq 0.45$. The relative efficiency of the proposed model over Ewemooje *et al.* (2019a) reduces as $\pi_U$ increases when $0.1 \leq \pi_A \leq 0.3$ and increases as $\pi_U$ increases when $0.35 \leq \pi_A \leq 0.45$. However, as the sensitive character, $\pi_A$ increases, the relative efficiency increases with the values ranging from 1.0135 to 21.4409. The relative efficiency (RE) is greater than 1 for $\pi_A = 0.1$ when $0.1 \leq \pi_U \leq 0.4$, RE greater than 1 for $\pi_A = 0.15$ when $0.1 \leq \pi_U \leq 0.7$ and RE greater than 1 when $0.2 \leq \pi_A \leq 0.45$ at all values of $\pi_U$. This shows that the proposed model is more efficient than the Ewemooje *et al.* (2019a) model as the proportion of people belonging to the sensitive attribute increases.

In Table 4, the probability of selecting the sensitive attribute was increased to 0.4 i.e. $P_1 = 0.4$ while $P_2 = 0.6$. The relative efficiency of the proposed model over Ewemooje *et al.* (2019a) also reduces as $\pi_U$ increases when $0.1 \leq \pi_A \leq 0.3$ and increases as $\pi_U$ increases when $0.35 \leq \pi_A \leq 0.45$. The relative efficiencies range between 1. 0284 and 18.8538. This shows that there is increase in efficiency as $P_1$ increases.

**Table 3.** Relative efficiency comparison between the proposed model and Ewemooje *et al.* (2019a) model when $P_1 = 0.3$; $P_2 = 0.7$; $\alpha = 25$; $\beta = 35$; n=200 for varying $\pi_A$ and $\pi_U$.

| $\pi_A$ | $\pi_U$ | $v(\hat{\pi})$ | $v(\hat{\pi}_A)$ | RE | $\pi_A$ | $\pi_U$ | $v(\hat{\pi})$ | $v(\hat{\pi}_A)$ | RE |
|---|---|---|---|---|---|---|---|---|---|
| | 0.1 | 0.000573 | 0.000345 | 1.662303 | | 0.1 | 0.001146 | 0.000536 | 2.137366 |
| | 0.2 | 0.000573 | 0.000413 | 1.387377 | | 0.2 | 0.001146 | 0.000543 | 2.108094 |
| | 0.3 | 0.000573 | 0.000481 | 1.191302 | | 0.3 | 0.001146 | 0.000551 | 2.080862 |
| | 0.4 | 0.000573 | 0.000549 | 1.044416 | | 0.4 | 0.001146 | 0.000557 | 2.055543 |
| 0.1 | 0.5 | 0.000573 | 0.000616 | 0.930275 | 0.3 | 0.5 | 0.001146 | 0.000564 | 2.032025 |
| | 0.6 | 0.000573 | 0.000683 | 0.839031 | | 0.6 | 0.001146 | 0.00057 | 2.010205 |
| | 0.7 | 0.000573 | 0.00075 | 0.764424 | | 0.7 | 0.001146 | 0.000576 | 1.989992 |
| | 0.8 | 0.000573 | 0.000816 | 0.702286 | | 0.8 | 0.001146 | 0.000581 | 1.971302 |
| | 0.9 | 0.000573 | 0.000882 | 0.649735 | | 0.9 | 0.001146 | 0.000586 | 1.954063 |
| | 1 | 0.000573 | 0.000948 | 0.604711 | | 1 | 0.001146 | 0.000591 | 1.938206 |
| | 0.1 | 0.000754 | 0.00043 | 1.752585 | | 0.1 | 0.001226 | 0.000521 | 2.352242 |
| | 0.2 | 0.000754 | 0.000483 | 1.559987 | | 0.2 | 0.001226 | 0.000514 | 2.387847 |
| | 0.3 | 0.000754 | 0.000536 | 1.406395 | | 0.3 | 0.001226 | 0.000505 | 2.426134 |
| | 0.4 | 0.000754 | 0.000588 | 1.281057 | | 0.4 | 0.001226 | 0.000497 | 2.46731 |
| 0.15 | 0.5 | 0.000754 | 0.00064 | 1.176839 | 0.35 | 0.5 | 0.001226 | 0.000488 | 2.511608 |
| | 0.6 | 0.000754 | 0.000692 | 1.088822 | | 0.6 | 0.001226 | 0.000479 | 2.559291 |
| | 0.7 | 0.000754 | 0.000744 | 1.013506 | | 0.7 | 0.001226 | 0.00047 | 2.610657 |
| | 0.8 | 0.000754 | 0.000795 | 0.948329 | | 0.8 | 0.001226 | 0.00046 | 2.666043 |
| | 0.9 | 0.000754 | 0.000846 | 0.891377 | | 0.9 | 0.001226 | 0.00045 | 2.725833 |
| | 1 | 0.000754 | 0.000896 | 0.841189 | | 1 | 0.001226 | 0.000439 | 2.790465 |
| | 0.1 | 0.000909 | 0.00049 | 1.854419 | | 0.1 | 0.001282 | 0.000482 | 2.661543 |
| | 0.2 | 0.000909 | 0.000528 | 1.721451 | | 0.2 | 0.001282 | 0.000459 | 2.79495 |
| | 0.3 | 0.000909 | 0.000566 | 1.607214 | | 0.3 | 0.001282 | 0.000435 | 2.944671 |
| | 0.4 | 0.000909 | 0.000603 | 1.508023 | | 0.4 | 0.001282 | 0.000412 | 3.113841 |
| 0.2 | 0.5 | 0.000909 | 0.00064 | 1.421098 | 0.4 | 0.5 | 0.001282 | 0.000388 | 3.306451 |
| | 0.6 | 0.000909 | 0.000676 | 1.344305 | | 0.6 | 0.001282 | 0.000363 | 3.52767 |
| | 0.7 | 0.000909 | 0.000713 | 1.275977 | | 0.7 | 0.001282 | 0.000339 | 3.784304 |
| | 0.8 | 0.000909 | 0.000749 | 1.214796 | | 0.8 | 0.001282 | 0.000314 | 4.085509 |
| | 0.9 | 0.000909 | 0.000784 | 1.159703 | | 0.9 | 0.001282 | 0.000288 | 4.443899 |
| | 1 | 0.000909 | 0.000819 | 1.109838 | | 1 | 0.001282 | 0.000263 | 4.877343 |
| | 0.1 | 0.00104 | 0.000526 | 1.978355 | | 0.1 | 0.001313 | 0.000417 | 3.147851 |
| | 0.2 | 0.00104 | 0.000548 | 1.896602 | | 0.2 | 0.001313 | 0.000379 | 3.465365 |
| | 0.3 | 0.00104 | 0.000571 | 1.822394 | | 0.3 | 0.001313 | 0.00034 | 3.857866 |
| | 0.4 | 0.00104 | 0.000593 | 1.754752 | | 0.4 | 0.001313 | 0.000301 | 4.355411 |
| 0.25 | 0.5 | 0.00104 | 0.000614 | 1.692864 | 0.45 | 0.5 | 0.001313 | 0.000262 | 5.006603 |
| | 0.6 | 0.00104 | 0.000636 | 1.636043 | | 0.6 | 0.001313 | 0.000223 | 5.895497 |
| | 0.7 | 0.00104 | 0.000657 | 1.583709 | | 0.7 | 0.001313 | 0.000183 | 7.181136 |
| | 0.8 | 0.00104 | 0.000677 | 1.53537 | | 0.8 | 0.001313 | 0.000143 | 9.205181 |
| | 0.9 | 0.00104 | 0.000698 | 1.490601 | | 0.9 | 0.001313 | 0.000102 | 12.85955 |
| | 1 | 0.00104 | 0.000718 | 1.449035 | | 1 | 0.001313 | 0.0000612 | 21.44086 |

**Table 4.** Relative efficiency comparison between the proposed model and Ewemooje *et al.* (2019a) model when $P_1$= 0.4; $P_2$= 0.6; α= 25; β= 35; n=200 for varying $\pi_A$ and $\pi_U$

| $\pi_A$ | $\pi_U$ | $v(\hat{\pi})$ | $v(\hat{\pi}_A)$ | RE | $\pi_A$ | $\pi_U$ | $v(\hat{\pi})$ | $v(\hat{\pi}_A)$ | RE |
|---|---|---|---|---|---|---|---|---|---|
| | 0.1 | 0.000586 | 0.000353 | 1.660952 | | 0.1 | 0.001156 | 0.000548 | 2.107674 |
| | 0.2 | 0.000586 | 0.000425 | 1.377199 | | 0.2 | 0.001156 | 0.000557 | 2.073896 |
| | 0.3 | 0.000586 | 0.000498 | 1.177079 | | 0.3 | 0.001156 | 0.000566 | 2.042447 |
| | 0.4 | 0.000586 | 0.00057 | 1.02837 | | 0.4 | 0.001156 | 0.000574 | 2.013165 |
| 0.1 | 0.5 | 0.000586 | 0.000641 | 0.913521 | 0.3 | 0.5 | 0.001156 | 0.000582 | 1.985906 |
| | 0.6 | 0.000586 | 0.000713 | 0.822151 | | 0.6 | 0.001156 | 0.000589 | 1.960538 |
| | 0.7 | 0.000586 | 0.000783 | 0.747731 | | 0.7 | 0.001156 | 0.000597 | 1.936947 |
| | 0.8 | 0.000586 | 0.000854 | 0.685946 | | 0.8 | 0.001156 | 0.000603 | 1.915028 |
| | 0.9 | 0.000586 | 0.000924 | 0.633833 | | 0.9 | 0.001156 | 0.00061 | 1.894686 |
| | 1 | 0.000586 | 0.000994 | 0.589287 | | 1 | 0.001156 | 0.000616 | 1.875838 |
| | 0.1 | 0.000766 | 0.000439 | 1.74396 | | 0.1 | 0.001236 | 0.000535 | 2.310815 |
| | 0.2 | 0.000766 | 0.000496 | 1.544414 | | 0.2 | 0.001236 | 0.000528 | 2.341486 |
| | 0.3 | 0.000766 | 0.000552 | 1.386723 | | 0.3 | 0.001236 | 0.00052 | 2.37458 |
| | 0.4 | 0.000766 | 0.000608 | 1.258975 | | 0.4 | 0.001236 | 0.000513 | 2.410267 |
| 0.15 | 0.5 | 0.000766 | 0.000664 | 1.153387 | 0.35 | 0.5 | 0.001236 | 0.000505 | 2.448743 |
| | 0.6 | 0.000766 | 0.000719 | 1.064657 | | 0.6 | 0.001236 | 0.000496 | 2.490223 |
| | 0.7 | 0.000766 | 0.000774 | 0.989051 | | 0.7 | 0.001236 | 0.000487 | 2.534952 |
| | 0.8 | 0.000766 | 0.000829 | 0.923862 | | 0.8 | 0.001236 | 0.000478 | 2.583208 |
| | 0.9 | 0.000766 | 0.000883 | 0.867078 | | 0.9 | 0.001236 | 0.000469 | 2.635303 |
| | 1 | 0.000766 | 0.000937 | 0.817176 | | 1 | 0.001236 | 0.000459 | 2.691595 |
| | 0.1 | 0.000921 | 0.0005 | 1.839616 | | 0.1 | 0.001291 | 0.000496 | 2.601386 |
| | 0.2 | 0.000921 | 0.000541 | 1.700959 | | 0.2 | 0.001291 | 0.000473 | 2.727499 |
| | 0.3 | 0.000921 | 0.000582 | 1.582692 | | 0.3 | 0.001291 | 0.00045 | 2.868693 |
| | 0.4 | 0.000921 | 0.000622 | 1.480635 | | 0.4 | 0.001291 | 0.000426 | 3.027789 |
| 0.2 | 0.5 | 0.000921 | 0.000662 | 1.391678 | 0.4 | 0.5 | 0.001291 | 0.000402 | 3.208359 |
| | 0.6 | 0.000921 | 0.000701 | 1.313461 | | 0.6 | 0.001291 | 0.000378 | 3.414995 |
| | 0.7 | 0.000921 | 0.00074 | 1.244157 | | 0.7 | 0.001291 | 0.000353 | 3.653698 |
| | 0.8 | 0.000921 | 0.000779 | 1.182331 | | 0.8 | 0.001291 | 0.000328 | 3.932472 |
| | 0.9 | 0.000921 | 0.000817 | 1.126842 | | 0.9 | 0.001291 | 0.000303 | 4.262223 |
| | 1 | 0.000921 | 0.000855 | 1.076768 | | 1 | 0.001291 | 0.000277 | 4.658218 |
| | 0.1 | 0.001051 | 0.000537 | 1.956943 | | 0.1 | 0.00132 | 0.000432 | 3.053186 |
| | 0.2 | 0.001051 | 0.000562 | 1.870325 | | 0.2 | 0.00132 | 0.000394 | 3.354703 |
| | 0.3 | 0.001051 | 0.000586 | 1.79212 | | 0.3 | 0.00132 | 0.000354 | 3.725976 |
| | 0.4 | 0.001051 | 0.00061 | 1.72118 | | 0.4 | 0.00132 | 0.000315 | 4.194316 |
| 0.25 | 0.5 | 0.001051 | 0.000634 | 1.656556 | 0.45 | 0.5 | 0.00132 | 0.000275 | 4.80343 |
| | 0.6 | 0.001051 | 0.000658 | 1.59746 | | 0.6 | 0.00132 | 0.000235 | 5.627905 |
| | 0.7 | 0.001051 | 0.000681 | 1.543228 | | 0.7 | 0.00132 | 0.000194 | 6.80632 |
| | 0.8 | 0.001051 | 0.000704 | 1.4933 | | 0.8 | 0.00132 | 0.000153 | 8.628625 |
| | 0.9 | 0.001051 | 0.000726 | 1.447199 | | 0.9 | 0.00132 | 0.000112 | 11.82045 |
| | 1 | 0.001051 | 0.000748 | 1.404517 | | 1 | 0.00132 | 0.00007 | 18.85377 |

Table 5 shows that as $P_1 = P_2 = 0.5$, the variance of the Ewemooje *et al.* (2019a) model increases at all values of $\pi_A$ from 0.00040 to 0.00133 while the variance of the proposed model decreases as $\pi_A$ increases with values ranging from 0.00108 to 0.00008.

The relative efficiency of the proposed model over Ewemooje *et al.* (2019a) also reduces as $\pi_U$ increases when $0.1 \leq \pi_A \leq 0.3$ and increases as $\pi_U$ increases when $0.35 \leq \pi_A \leq 0.45$. However, as the sensitive character $\pi_A$ increases, an appreciable increase is noticed in the values of the relative efficiency ranging from 1.0144 to 16.5954.

**Table 5.** Relative efficiency comparison between the proposed model and Ewemooje *et al.* (2019a) model when $P_1 = 0.5$; $P_2 = 0.5$; $\alpha = 25$; $\beta = 35$; n=200 for varying $\pi_A$ and $\pi_U$

| π_A | π_U | v(π^) | v(π^_A ) | RE | π_A | π_U | v(π^) | v(π^_A ) | RE |
|---|---|---|---|---|---|---|---|---|---|
| | 0.1 | 0.0006 | 0.000361 | 1.662391 | | 0.1 | 0.001167 | 0.000561 | 2.079894 |
| | 0.2 | 0.0006 | 0.000438 | 1.3694 | | 0.2 | 0.001167 | 0.000571 | 2.041478 |
| | 0.3 | 0.0006 | 0.000515 | 1.165049 | | 0.3 | 0.001167 | 0.000582 | 2.005731 |
| | 0.4 | 0.0006 | 0.000591 | 1.014402 | | 0.4 | 0.001167 | 0.000591 | 1.972448 |
| 0.1 | 0.5 | 0.0006 | 0.000668 | 0.898752 | 0.3 | 0.5 | 0.001167 | 0.000601 | 1.941448 |
| | 0.6 | 0.0006 | 0.000743 | 0.807175 | | 0.6 | 0.001167 | 0.00061 | 1.912568 |
| | 0.7 | 0.0006 | 0.000819 | 0.732866 | | 0.7 | 0.001167 | 0.000619 | 1.885663 |
| | 0.8 | 0.0006 | 0.000894 | 0.671363 | | 0.8 | 0.001167 | 0.000627 | 1.860602 |
| | 0.9 | 0.0006 | 0.000968 | 0.619621 | | 0.9 | 0.001167 | 0.000635 | 1.83727 |
| | 1 | 0.0006 | 0.001043 | 0.575488 | | 1 | 0.001167 | 0.000643 | 1.815562 |
| | 0.1 | 0.000779 | 0.000448 | 1.737559 | | 0.1 | 0.001246 | 0.000548 | 2.271653 |
| | 0.2 | 0.000779 | 0.000509 | 1.530835 | | 0.2 | 0.001246 | 0.000542 | 2.297251 |
| | 0.3 | 0.000779 | 0.000569 | 1.36896 | | 0.3 | 0.001246 | 0.000536 | 2.325039 |
| | 0.4 | 0.000779 | 0.000629 | 1.238775 | | 0.4 | 0.001246 | 0.000529 | 2.355155 |
| 0.15 | 0.5 | 0.000779 | 0.000688 | 1.131809 | 0.35 | 0.5 | 0.001246 | 0.000522 | 2.387755 |
| | 0.6 | 0.000779 | 0.000748 | 1.042363 | | 0.6 | 0.001246 | 0.000514 | 2.423015 |
| | 0.7 | 0.000779 | 0.000806 | 0.966464 | | 0.7 | 0.001246 | 0.000506 | 2.46113 |
| | 0.8 | 0.000779 | 0.000865 | 0.901253 | | 0.8 | 0.001246 | 0.000498 | 2.502325 |
| | 0.9 | 0.000779 | 0.000923 | 0.844625 | | 0.9 | 0.001246 | 0.000489 | 2.546848 |
| | 1 | 0.000779 | 0.00098 | 0.794993 | | 1 | 0.001246 | 0.00048 | 2.594986 |
| | 0.1 | 0.000933 | 0.000511 | 1.826749 | | 0.1 | 0.0013 | 0.000511 | 2.5444 |
| | 0.2 | 0.000933 | 0.000555 | 1.682243 | | 0.2 | 0.0013 | 0.000488 | 2.663126 |
| | 0.3 | 0.000933 | 0.000598 | 1.559889 | | 0.3 | 0.0013 | 0.000465 | 2.795699 |
| | 0.4 | 0.000933 | 0.000641 | 1.454965 | | 0.4 | 0.0013 | 0.000441 | 2.944631 |
| 0.2 | 0.5 | 0.000933 | 0.000684 | 1.364005 | 0.4 | 0.5 | 0.0013 | 0.000418 | 3.113082 |
| | 0.6 | 0.000933 | 0.000727 | 1.284404 | | 0.6 | 0.0013 | 0.000393 | 3.305085 |
| | 0.7 | 0.000933 | 0.000769 | 1.214165 | | 0.7 | 0.0013 | 0.000369 | 3.525866 |
| | 0.8 | 0.000933 | 0.00081 | 1.151737 | | 0.8 | 0.0013 | 0.000344 | 3.782328 |
| | 0.9 | 0.000933 | 0.000852 | 1.09589 | | 0.9 | 0.0013 | 0.000318 | 4.08377 |
| | 1 | 0.000933 | 0.000893 | 1.045643 | | 1 | 0.0013 | 0.000293 | 4.443038 |
| | 0.1 | 0.001063 | 0.000548 | 1.937363 | | 0.1 | 0.001329 | 0.000448 | 2.964072 |
| | 0.2 | 0.001063 | 0.000576 | 1.845746 | | 0.2 | 0.001329 | 0.000409 | 3.249943 |
| | 0.3 | 0.001063 | 0.000603 | 1.763485 | | 0.3 | 0.001329 | 0.000369 | 3.600451 |
| | 0.4 | 0.001063 | 0.000629 | 1.689239 | | 0.4 | 0.001329 | 0.000329 | 4.040248 |
| 0.25 | 0.5 | 0.001063 | 0.000655 | 1.621908 | 0.45 | 0.5 | 0.001329 | 0.000288 | 4.608347 |
| | 0.6 | 0.001063 | 0.000681 | 1.560588 | | 0.6 | 0.001329 | 0.000248 | 5.37037 |
| | 0.7 | 0.001063 | 0.000706 | 1.504523 | | 0.7 | 0.001329 | 0.000206 | 6.445891 |
| | 0.8 | 0.001063 | 0.000731 | 1.453083 | | 0.8 | 0.001329 | 0.000165 | 8.078222 |
| | 0.9 | 0.001063 | 0.000756 | 1.405733 | | 0.9 | 0.001329 | 0.000123 | 10.85034 |
| | 1 | 0.001063 | 0.00078 | 1.362018 | | 1 | 0.001329 | 0.00008 | 16.59538 |

As the probability of selecting the sensitive attribute was increased to $P_1 = 0.6$ and $P_2 = 0.4$. The relative efficiency of the proposed model over Ewemooje *et al.* (2019a) increases with each value of $\pi_U$ as $\pi_A$ increases when $0.1 \leq \pi_A \leq 0.3$ and decreases

when $0.35 \leq \pi_A \leq 0.45$. The variance of the Ewemooje *et al.* (2019a) model increases at all values of $\pi_A$ from 0.00041 to 0.00134, the variance of the proposed model decreases as $\pi_A$ increases with values ranging from 0.00104 to 0.00009 while the relative efficiencies range between 1.0026 and 14.6381 (see Table 6).

**Table 6.** Relative efficiency comparison between the proposed model and Ewemooje *et al.* (2019a) model when $P_1 = 0.6$; $P_2 = 0.4$; $\alpha = 25$; $\beta = 35$; $n=200$ for varying $\pi_A$ and $\pi_U$.

| $\pi_A$ | $\pi_U$ | $v(\hat{\pi})$ | $v(\hat{\pi}_A)$ | RE | $\pi_A$ | $\pi_U$ | $v(\hat{\pi})$ | $v(\hat{\pi}_A)$ | RE |
|---|---|---|---|---|---|---|---|---|---|
| | 0.1 | 0.000616 | 0.000369 | 1.666953 | | 0.1 | 0.001179 | 0.000574 | 2.05419 |
| | 0.2 | 0.000616 | 0.000451 | 1.364207 | | 0.2 | 0.001179 | 0.000586 | 2.010996 |
| | 0.3 | 0.000616 | 0.000533 | 1.155374 | | 0.3 | 0.001179 | 0.000598 | 1.970869 |
| | 0.4 | 0.000616 | 0.000614 | 1.002628 | | 0.4 | 0.001179 | 0.00061 | 1.933553 |
| 0.1 | 0.5 | 0.000616 | 0.000695 | 0.886052 | 0.3 | 0.5 | 0.001179 | 0.000621 | 1.898819 |
| | 0.6 | 0.000616 | 0.000776 | 0.794162 | | 0.6 | 0.001179 | 0.000632 | 1.866467 |
| | 0.7 | 0.000616 | 0.000856 | 0.719869 | | 0.7 | 0.001179 | 0.000642 | 1.836318 |
| | 0.8 | 0.000616 | 0.000935 | 0.658563 | | 0.8 | 0.001179 | 0.000652 | 1.808212 |
| | 0.9 | 0.000616 | 0.001014 | 0.607113 | | 0.9 | 0.001179 | 0.000662 | 1.782007 |
| | 1 | 0.000616 | 0.001093 | 0.563322 | | 1 | 0.001179 | 0.000671 | 1.757575 |
| | 0.1 | 0.000794 | 0.000458 | 1.733651 | | 0.1 | 0.001257 | 0.000563 | 2.234883 |
| | 0.2 | 0.000794 | 0.000523 | 1.519457 | | 0.2 | 0.001257 | 0.000557 | 2.255279 |
| | 0.3 | 0.000794 | 0.000587 | 1.353272 | | 0.3 | 0.001257 | 0.000552 | 2.277664 |
| | 0.4 | 0.000794 | 0.000651 | 1.220588 | | 0.4 | 0.001257 | 0.000546 | 2.302146 |
| 0.15 | 0.5 | 0.000794 | 0.000714 | 1.112209 | 0.35 | 0.5 | 0.001257 | 0.00054 | 2.328848 |
| | 0.6 | 0.000794 | 0.000777 | 1.022022 | | 0.6 | 0.001257 | 0.000533 | 2.357906 |
| | 0.7 | 0.000794 | 0.00084 | 0.945804 | | 0.7 | 0.001257 | 0.000526 | 2.389473 |
| | 0.8 | 0.000794 | 0.000902 | 0.880547 | | 0.8 | 0.001257 | 0.000519 | 2.423725 |
| | 0.9 | 0.000794 | 0.000964 | 0.824049 | | 0.9 | 0.001257 | 0.000511 | 2.460857 |
| | 1 | 0.000794 | 0.001025 | 0.774659 | | 1 | 0.001257 | 0.000503 | 2.501089 |
| | 0.1 | 0.000947 | 0.000522 | 1.816041 | | 0.1 | 0.001311 | 0.000526 | 2.490643 |
| | 0.2 | 0.000947 | 0.000569 | 1.66549 | | 0.2 | 0.001311 | 0.000504 | 2.601916 |
| | 0.3 | 0.000947 | 0.000616 | 1.538968 | | 0.3 | 0.001311 | 0.000481 | 2.725814 |
| | 0.4 | 0.000947 | 0.000662 | 1.431157 | | 0.4 | 0.001311 | 0.000458 | 2.864549 |
| 0.2 | 0.5 | 0.000947 | 0.000708 | 1.338201 | 0.4 | 0.5 | 0.001311 | 0.000434 | 3.020886 |
| | 0.6 | 0.000947 | 0.000754 | 1.257237 | | 0.6 | 0.001311 | 0.00041 | 3.198326 |
| | 0.7 | 0.000947 | 0.000799 | 1.186092 | | 0.7 | 0.001311 | 0.000385 | 3.401361 |
| | 0.8 | 0.000947 | 0.000844 | 1.123088 | | 0.8 | 0.001311 | 0.00036 | 3.635866 |
| | 0.9 | 0.000947 | 0.000888 | 1.06691 | | 0.9 | 0.001311 | 0.000335 | 3.90966 |
| | 1 | 0.000947 | 0.000932 | 1.016511 | | 1 | 0.001311 | 0.00031 | 4.233393 |
| | 0.1 | 0.001076 | 0.00056 | 1.919806 | | 0.1 | 0.001339 | 0.000465 | 2.880362 |
| | 0.2 | 0.001076 | 0.00059 | 1.823035 | | 0.2 | 0.001339 | 0.000425 | 3.150967 |
| | 0.3 | 0.001076 | 0.000619 | 1.736648 | | 0.3 | 0.001339 | 0.000385 | 3.481228 |
| | 0.4 | 0.001076 | 0.000648 | 1.659079 | | 0.4 | 0.001339 | 0.000344 | 3.893252 |
| 0.25 | 0.5 | 0.001076 | 0.000677 | 1.589061 | 0.45 | 0.5 | 0.001339 | 0.000303 | 4.421607 |
| | 0.6 | 0.001076 | 0.000705 | 1.52556 | | 0.6 | 0.001339 | 0.000261 | 5.123548 |
| | 0.7 | 0.001076 | 0.000733 | 1.467722 | | 0.7 | 0.001339 | 0.000219 | 6.10128 |
| | 0.8 | 0.001076 | 0.00076 | 1.414838 | | 0.8 | 0.001339 | 0.000177 | 7.556836 |
| | 0.9 | 0.001076 | 0.000787 | 1.366311 | | 0.9 | 0.001339 | 0.000135 | 9.953349 |
| | 1 | 0.001076 | 0.000814 | 1.321637 | | 1 | 0.001339 | 0.0000915 | 14.63817 |

Table 7 shows that for varying $\pi_A$ and $\pi_U$, $P_1 = 0.7$; $P_2 = 0.3$, the variance of the Ewemooje *et al.* (2019a) model increases at all values of $\pi_A$ from 0.00043 to 0.00135 while the variance of the proposed model also increases as $\pi_U$ increases when $0.1 \leq$

$\pi_A \leq 0.3$ and decreases as $\pi_U$ increases when $0.35 \leq \pi_A \leq 0.45$. The relative efficiency of the proposed model over Ewemooje *et al.* (2019a) shows that as the sensitive character $\pi_A$ increases, the relative efficiency increases with the values ranging from 1.0037 to 12.9490.

**Table 7.** Relative efficiency comparison between the proposed model and Ewemooje *et al.* (2019a) model when $P_1 = 0.7$; $P_2 = 0.3$; $\alpha = 25$; $\beta = 35$; n=200 for varying $\pi_A$ and $\pi_U$.

| $\pi_A$ | $\pi_U$ | $v(\hat{\pi})$ | $v(\hat{\pi}_A)$ | RE | $\pi_A$ | $\pi_U$ | $v(\hat{\pi})$ | $v(\hat{\pi}_A)$ | RE |
|---|---|---|---|---|---|---|---|---|---|
| | 0.1 | 0.000634 | 0.000378 | 1.67503 | | 0.1 | 0.001193 | 0.000587 | 2.030752 |
| | 0.2 | 0.000634 | 0.000465 | 1.361891 | | 0.2 | 0.001193 | 0.000602 | 1.982633 |
| | 0.3 | 0.000634 | 0.000552 | 1.148251 | | 0.3 | 0.001193 | 0.000615 | 1.938042 |
| | 0.4 | 0.000634 | 0.000638 | 0.993193 | | 0.4 | 0.001193 | 0.000629 | 1.896659 |
| 0.1 | 0.5 | 0.000634 | 0.000724 | 0.875529 | 0.3 | 0.5 | 0.001193 | 0.000642 | 1.8582 |
| | 0.6 | 0.000634 | 0.000809 | 0.783192 | | 0.6 | 0.001193 | 0.000655 | 1.822421 |
| | 0.7 | 0.000634 | 0.000894 | 0.7088 | | 0.7 | 0.001193 | 0.000667 | 1.789101 |
| | 0.8 | 0.000634 | 0.000979 | 0.647589 | | 0.8 | 0.001193 | 0.000679 | 1.758048 |
| | 0.9 | 0.000634 | 0.001063 | 0.596341 | | 0.9 | 0.001193 | 0.00069 | 1.729089 |
| | 1 | 0.000634 | 0.001146 | 0.552808 | | 1 | 0.001193 | 0.000701 | 1.702072 |
| | 0.1 | 0.000811 | 0.000468 | 1.73255 | | 0.1 | 0.00127 | 0.000577 | 2.200659 |
| | 0.2 | 0.000811 | 0.000537 | 1.510522 | | 0.2 | 0.00127 | 0.000573 | 2.215729 |
| | 0.3 | 0.000811 | 0.000605 | 1.33985 | | 0.3 | 0.00127 | 0.000569 | 2.232628 |
| | 0.4 | 0.000811 | 0.000673 | 1.20457 | | 0.4 | 0.00127 | 0.000564 | 2.251434 |
| 0.15 | 0.5 | 0.000811 | 0.000741 | 1.094713 | 0.35 | 0.5 | 0.00127 | 0.000559 | 2.272238 |
| | 0.6 | 0.000811 | 0.000808 | 1.003731 | | 0.6 | 0.00127 | 0.000553 | 2.295143 |
| | 0.7 | 0.000811 | 0.000875 | 0.927151 | | 0.7 | 0.00127 | 0.000547 | 2.320265 |
| | 0.8 | 0.000811 | 0.000941 | 0.861805 | | 0.8 | 0.00127 | 0.000541 | 2.347734 |
| | 0.9 | 0.000811 | 0.001007 | 0.805395 | | 0.9 | 0.00127 | 0.000534 | 2.377699 |
| | 1 | 0.000811 | 0.001072 | 0.756207 | | 1 | 0.00127 | 0.000527 | 2.410328 |
| | 0.1 | 0.000963 | 0.000533 | 1.807757 | | 0.1 | 0.001322 | 0.000542 | 2.440203 |
| | 0.2 | 0.000963 | 0.000583 | 1.650922 | | 0.2 | 0.001322 | 0.00052 | 2.54398 |
| | 0.3 | 0.000963 | 0.000634 | 1.520119 | | 0.3 | 0.001322 | 0.000497 | 2.659183 |
| | 0.4 | 0.000963 | 0.000683 | 1.409372 | | 0.4 | 0.001322 | 0.000474 | 2.787741 |
| 0.2 | 0.5 | 0.000963 | 0.000733 | 1.314408 | 0.4 | 0.5 | 0.001322 | 0.000451 | 2.932044 |
| | 0.6 | 0.000963 | 0.000782 | 1.232084 | | 0.6 | 0.001322 | 0.000427 | 3.095093 |
| | 0.7 | 0.000963 | 0.00083 | 1.160041 | | 0.7 | 0.001322 | 0.000403 | 3.280704 |
| | 0.8 | 0.000963 | 0.000879 | 1.096473 | | 0.8 | 0.001322 | 0.000379 | 3.493806 |
| | 0.9 | 0.000963 | 0.000926 | 1.039974 | | 0.9 | 0.001322 | 0.000354 | 3.740884 |
| | 1 | 0.000963 | 0.000974 | 0.989432 | | 1 | 0.001322 | 0.000328 | 4.030638 |
| | 0.1 | 0.001091 | 0.000573 | 1.904498 | | 0.1 | 0.00135 | 0.000482 | 2.801958 |
| | 0.2 | 0.001091 | 0.000605 | 1.802396 | | 0.2 | 0.00135 | 0.000441 | 3.057695 |
| | 0.3 | 0.001091 | 0.000637 | 1.711794 | | 0.3 | 0.00135 | 0.000401 | 3.368273 |
| | 0.4 | 0.001091 | 0.000669 | 1.630872 | | 0.4 | 0.00135 | 0.00036 | 3.753384 |
| 0.25 | 0.5 | 0.001091 | 0.0007 | 1.558175 | 0.45 | 0.5 | 0.00135 | 0.000318 | 4.243433 |
| | 0.6 | 0.001091 | 0.000731 | 1.492525 | | 0.6 | 0.00135 | 0.000276 | 4.887965 |
| | 0.7 | 0.001091 | 0.000761 | 1.432962 | | 0.7 | 0.00135 | 0.000234 | 5.773544 |
| | 0.8 | 0.001091 | 0.000791 | 1.37869 | | 0.8 | 0.00135 | 0.000191 | 7.066269 |
| | 0.9 | 0.001091 | 0.000821 | 1.329049 | | 0.9 | 0.00135 | 0.000148 | 9.13035 |
| | 1 | 0.001091 | 0.00085 | 1.283481 | | 1 | 0.00135 | 0.000104 | 12.94899 |

## 5. Application of the Proposed Model

The proposed method and direct method (DM) were used in collecting information on examination malpractices prevalence among students at a Nigerian university. Two hundred (200) instruments were administered using two decks of cards consisting of the sensitive question *"have you ever been involved in examination malpractices?"* and unrelated question "*do you love soccer*?" as the randomized devices. The respondents were given proper education on how to use the randomized devices with appropriate demonstration. They were also assured of confidentiality by ensuring that responses given cannot be traced to respondents; hence, they willingly participated in the survey. The respondents were directly (DM) asked the sensitive question *"have you ever been involved in examination malpractices?"*. If "yes" answer is obtained, he/she does not use the randomized device but if he/she answers "no", then he/she is instructed to choose one of the two decks of cards at random and then respond accordingly without revealing question answered to the interviewer. The two randomized devices $R_1$ and $R_2$ consist of two unrelated questions (the sensitive question with probability $P_1 = 0.7$, and unrelated question with probability $1\text{-}P_1 = 0.3$ for $R_1$ while $P_2 = 0.3$ and $P_2 = 0.7$ for $R_2$).

The age distribution of the sampled respondents ranges between 16 and 29 years with the age group 20–24 years having the higher percentage of 58.0% and about three-quarters of them are male (74.0%). The true proportion of respondents who answered "yes" to the unrelated question ($\pi_U$) "*do you love soccer*?" is 0.45. The estimate of examination malpractices prevalence and their associated coefficient of variation (CV) are presented in Table 8. The DM estimate prevalence of examination malpractices at 19.0% compared to 23.0% for the proposed method. The standard error associated with DM is 0.028 (CV = 14.6%) while the proposed model is 0.026 (CV = 11.5%).

However, contrary to what was reported by Jann et al., (2012) where Crosswise Model (CM) produced higher estimate with higher standard error, the proposed method produced higher estimate with lower standard error as against the DM. Hence, the proposed model performs better than the DM in line with earlier works of Jann *et al.* (2012), Ewemooje *et al.,* (2017), Cobo *et al.,* (2016) and Ewemooje *et al.,* (2019b).

**Table 8.** Comparative analysis of the proposed model versus the direct method

| Method | $\pi_A$ | $\mathbf{V}(\hat{\pi}_A)$ | $\mathbf{S.E}(\hat{\pi}_A)$ | $\mathbf{C.V}(\hat{\pi}_A)$ |
|---|---|---|---|---|
| Direct Method | 0.19 | 0.00077 | 0.028 | 14.6% |
| Proposed Model | 0.23 | 0.00070 | 0.026 | 11.5% |

## 6. Conclusion

The unrelated design has been shown to improve efficiency of a randomized response method and to reduce distrust of the respondents; hence, we proposed a new Randomized Response Model (RRM) which consists of the unrelated questions in dichotomous randomized response model. To ensure better efficiency, the proportion of the sensitive attribute must be at least 0.2 and not greater than 0.5. The variance of the proposed model decreases as the proportion of the sensitive attribute $\pi_A$ and unrelated attribute $\pi_U$ increases as against the Ewemooje *et al.* (2019a) model, which increases as the proportion of the sensitive attribute increases. The relative efficiency of the proposed model over Ewemooje *et al.* (2019a) reduces as $\pi_U$ increases when $0.05 \leq \pi_A \leq 0.3$ and increases as $\pi_U$ increases when $0.35 \leq \pi_A \leq 0.45$. Also, as the sample size increases from 50 to 500, the relative efficiency of the proposed model stood at 9.93 while as $P_1$ increases and $P_2$ decreases, the relative efficiency reduces from 7.09 to 6.23. Application of the proposed model also revealed it efficiency over the direct method in estimating the prevalence of examination malpractices among university students. The direct method estimated the prevalence of examination malpractices among university students at 19.0% while the proposed method estimated it at 23.0%. Hence, the proposed model is shown to be more efficient than the direct method and Ewemooje *et al.* (2019a) model as the proportion of people belonging to the sensitive attribute increases.

## References

ADEBOLA, F. B., ADEPETUN, A. O., (2011). A new Tripartite Randomized Response Technique. Journal of the Nigerian Association of Mathematical Physics, 19, pp. 119–122.

ADEBOLA, F. B., ADEDIRAN, A. A., EWEMOOJE, O. S., (2017). Hybrid tripartite randomized response technique. Communications in Statistics-Theory and Methods, 46(23), pp. 11756–11763; https://doi.org/10.1080/03610926.2016.1277760.

COBO, B., RUEDA, M. M., LÓPEZ –TORRECILLAS, F., (2016). Application of randomized response techniques for investigating cannabis use by Spanish university students, International Journal of Methods in Psychiatric Research; https://doi.org/10.1002/mpr.1517.

EWEMOOJE, O. S., (2017). Estimating two sensitive characters with equal probabilities of protection, Cogent Mathematics, 4, pp. 1–14.

EWEMOOJE, O. S., ADEBOLA, F. B., AMAHIA, G. N., (2019a). Alternative Unbiased Estimator in Dichotomous Randomized Response Technique, Communication in

Mathematics and Statistics, 7(4), pp. 383–400; https://doi.org/10.1007/s40304-018-0145-x.

EWEMOOJE, O. S., ADEBOLA, F. B., ADEDIRAN, A. A., (2018). A Stratified Hybrid Tripartite Randomized Response Technique, Gazi University Journal of Science, 31(4), pp. 1246–1266.

EWEMOOJE O. S., ADEBOLA, F. B., AWOGBEMILA, A. T., (2019b). Substance Use disorder prevalence using tripartite randomized technique, 14(1), pp. 55–61. https://dx.doi.org/10.4314/njtr.v14i1.8.

EWEMOOJE, O. S., AMAHIA, G. N., ADEBOLA, F. B., (2017). Estimating prevalence of induced abortion and multiple sexual partners using improved randomized response technique for two sensitive attributes, Communications in Statistics: Case Studies, Data Analysis and Applications, 3(1-2), pp. 21–28; https://doi.org/10.1080/23737484.2017.1398057.

FOX, J-P., VEEN, D., KLOTZKE K., (2019). Generalized Linear Mixed Models for Randomized Responses, Methodology (2019), 15(1), pp. 1–18; https://doi.org/10.1027/1614-2241/a000153.

GREENBERG, B., ABUL-ELA, A., SIMMONS, W., HORVITZ, D., (1969). The Unrelated question Randomized Response Technique: Theoretical framework, Journal of American Statistical Association, 64, pp. 529–539.

HOGLINGER, M., JANN, B., (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model, PLoS ONE 13(8): e0201770; https://doi.org/10.1371/journal.pone.0201770.

HORVITZ, D. G., SHAH, B. V., SIMMONS, W. R., (1967). The unrelated question Randomized Response Technique, Proceedings of social statistical section American Statistical Association, pp. 65–72.

HUSSAIN, Z., SHABBIR, J., (2007). Randomized use of Warner's randomized response Technique, Interstat: April #7.http://interstat.statjournals.net/INDEX/Apro7.html

JANN, B., JERKE, J., KRUMPAL, I., (2012). Asking sensitive question using the crosswise model, an experimental survey measuring plagiarism. Public Opinion Quarterly, 76, pp. 32–49; https://doi.org/10.1093/poq/nfr036.

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

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

YU, J.-W.., TIAN, G.-L., TANG, M.-L., (2008). Two new models for survey sampling with sensitive characteristic; design and analysis. Metrika, 67, pp. 251–263; https://doi.org/10.1007/s00184-007-0131-x.

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 133–149, DOI 10.21307/stattrans-2020-059*
Received – 28.03.2018; accepted – 04.05.2020

# A Bayesian analysis of complete multiple breaks in a panel autoregressive (CMB-PAR(1)) time series model

**Varun Agiwal[1,2], Jitendra Kumar[2], Dahud Kehinde Shangodoyin[3]**

## ABSTRACT

Most economic time series, such as GDP, real exchange rate and banking series are irregular by nature as they may be affected by a variety of discrepancies, including political changes, policy reforms, import-export market instability, etc. When such changes entail serious consequences for time series modelling, various researchers manage this problem by applying a structural break. Thus, the aim of this paper is to develop a generalised structural break time series model. The paper discusses a panel autoregressive model with multiple breaks present in all parameters, i.e. in the autoregressive coefficient and mean and error variance, which is a generalisation of various sub-models. The Bayesian approach is applied to estimate the model parameters and to obtain the highest posterior density interval. Strong evidence is observed to support the Bayes estimator and then it is compared with the maximum likelihood estimator. A simulation experiment is conducted and an empirical application on the SARRC association's GDP per capita time series is used to illustrate the performance of the proposed model. This model is also extended to a temporary shift model.

**Key words:** panel autoregressive model, structural break, MCMC, posterior probability.

## 1. Introduction

When modelling any time series, one may identify characteristics of series such as stationarity, seasonality, outliers, linear trend, structural breaks, etc., and then produce a good forecast for making a better conclusion. If there is an unexpected shift in time series, then this may occur due to outlier(s) or structural break(s). In the structural break, mainly any or all model parameters are affected for a particular time interval, which may have different inferences. These break points may split time series into two or multiple parts. If at multiple time points, which are identified in terms of change on

[1] Department of Community Medicine, Jawaharlal Nehru Medical College, Ajmer, India. E-mail: varunagiwal.stats@gmail.com. ORCID: https://orcid.org/0000-0003-1955-8832
[2] Department of Statistics, Central University of Rajasthan, Bandersindri, Ajmer, India. E-mail: vjitendrav@gmail.com. ORCID: https://orcid.org/0000-0003-4473-4148
[3] Department of Statistics, University of Botswana, Gaborone, Botswana. E-mail: shangodoyink@gamil.com. ORCID: https://orcid.org/0000-0002-0449-9510

model parameters, the series changes temporarily or permanently, then the model must be analysed in such a way that it gives better explanation and prediction. Handling of such time series received importance by several researchers, who made inference about the break and show its impact in real applications. The problem of estimation and testing of change points in the linear model was proposed by Bai and Perron (1998) and then extended into multiple breaks in a multiple regression model. Altissimo and Corradi (2003) considered a nonlinear process which has dependent and heterogeneous observations and contained a break in the mean component. They proposed an estimator for the detection and estimation of the number of breaks and applied for weekly Eurodollar interest rate. Jin *et al.* (2013) addressed the problem of multiple breaks in piecewise stationary AR process and detected the breaks by the penalized model selection approach. Topal *et al.* (2016) compared various detection techniques of multiple break points in artificially modified time series and applied to vine sprout length data as well as mercury injection capillary pressure curve. Jibrin *et al.* (2015) modelled an AR fractionally integrated moving average process and used Bayes information criterion to study the structural breaks in crude oil prices of Brent and WTI series.

The consequences of the structural break under Bayesian approach is studied by several researchers, see Albert and Chib (1993), Bai (2010), Kumar et al. (2012), Eo (2012) and Maheu and Song (2018). Further on, Chin *et al.* (2016) combined both robust-jump volatility estimator and a structural break heterogeneous autoregressive (HAR) model to battle the structural break in stock market volatility modelling and added the empirical literature of high-frequency volatility analysis by using modified HAR models and robust-jump volatility estimators. Yamamoto (2016) considered a simple modification in EM confidence set proposed by Elliott and Muller (2007) in a linear regression model having a single structural break and achieved a shorter confidence set than the EM method. Baltagi *et al.* (2016) considered both cross sectional dependence and a structural break in Pesaran (2006) heterogeneous panels and applied least square and common correlated effects estimators to estimate the change points. Pestova and Pesta (2017) constructed an estimator for a break in panel mean without a boundary condition, which was also consistent in no break situation and demonstrated in non-life insurance application. Meligkotsiduo *et al.* (2017) suggested a Bayesian approach to detect stationarity from AR(p) model with multiple breaks in mean, variance and autoregressive coefficients. To determine the marginal likelihood and posterior probability for comparing models, filtering recursions algorithm is used in the structural break model. Hwang and Shin (2017) proposed a sequential test for detecting mean breaks that allow long memory errors. The proposed test is consistent with asymptotic normal distribution and produced an unbiased break estimate as compared with Bai and Perron (1998) biased estimates.

Many studies have also been carried out on a structural break in the panel data model in reference to testing for unit root hypothesis, break point detection, estimation, etc. Karavias and Tzavalis (2017) studied the asymptotic properties of least squares based fixed-T panel unit root tests of panel AR(1) model considering a structural break in the deterministic components and obtained the limiting distribution which is dependent on the break date and time. Chen and Huang (2018) considered a non-parametric method to analyse the consistence of changing parameters and developed two types of consistent tests to check the stability of model parameter in time varying interaction panel model. Okui and Wang (2018) established a new model which allows a common structural change in the coefficients, while the number of breaks, break points, and the size of breaks are different across groups. They also obtained a hybrid estimation procedure under grouped fixed effects and an adaptive group, fused in panel data model with heterogeneous structural breaks. Bardwell *et al.* (2019) developed an approach to detect the change point in panel data model that pools the information across time series and come up with the most recent break points in multiple series at the same time point.

This paper is an extension of Agiwal *et al.* (2018), which discussed the panel autoregressive time series model of order one (PAR(1)) with a break in mean and error variance. This model does not allow a change on autoregressive parameter. However, it may also have multiple breaks so a PAR(1) time series model with multiple breaks is explored in the present study that considers a break in autoregressive coefficient also. As this allows breaks on all parameters of the model including coefficients, mean and error variance. Therefore, this is termed as a complete multiple breaks panel autoregressive time series model of order one (CMB-PAR(1)). A Bayesian analysis of the proposed model has been carried out to estimate the parameters under both symmetric and asymmetric loss functions and then compared with MLE through both simulation and empirical study. This paper has also discussed the temporary shift model, where a change occurs in the parameter for a short time interval, then it comes to the original structure. This model is a particular form of CMB-PAR(1) model with two break points.

## 2. Model and Assumptions

Let $\{y_{it}\}$ be a PAR(1) time series model having multiple structural breaks and break points in each panel that are assumed to be same and known. Due to multiple breaks, the structure of PAR(1) model may be shifted temporarily or permanently depending on the situation. If all parameters are instable permanently for assumed time intervals,

at that time structure of the series also shifted permanently. Let there be $B$ break points, then permanently shifted PAR(1) model (PS-PAR(1)) is

$$y_{it}=\begin{cases}\rho_1 y_{i,t-1}+(1-\rho_1)\mu_{i1}+\sigma_1\varepsilon_{it} & T_0<t\le T_1\\ \vdots & \\ \rho_j y_{i,t-1}+(1-\rho_j)\mu_{ij}+\sigma_j\varepsilon_{it} & T_{j-1}<t\le T_j\\ \vdots & \\ \rho_{B+1}y_{i,t-1}+(1-\rho_{B+1})\mu_{i,B+1}+\sigma_{B+1}\varepsilon_{it} & T_B<t\le T_{B+1}\end{cases} \tag{1}$$

There are several practical situations where a change occurs on a model for a temporary period, i.e. a change in the series only for a particular time interval and later on it comes back to the original model/process. Such a model is called a temporary shift (TS) model. So, this type of series contains only two breaks to observe the short term changes in the model parameters. In that situation, temporary shift PAR(1) model (TS-PAR(1)) is expressed as

$$y_{it}=\begin{cases}\rho_1 y_{i,t-1}+(1-\rho_1)\mu_{i1}+\sigma_1\varepsilon_{it} & 1\le t\le T_1\\ \rho_2 y_{i,t-1}+(1-\rho_2)\mu_{i2}+\sigma_2\varepsilon_{it} & T_1<t\le T_2\\ \rho_1 y_{i,t-1}+(1-\rho_1)\mu_{i1}+\sigma_1\varepsilon_{it} & T_2<t\le T\end{cases} \tag{2}$$

where $\{y_{it}, t=1,2,\ldots,T; i=1,2\ldots,n\}$ is a sequence of observations which contains $n$ cross-sectional units recorded at $T$ time period between $T_0=0$ to $T_{B+1}=T$. The error term $\varepsilon_{it}$ is a sequence of an independently distributed normal random variable with mean zero and variance $\sigma_j^2$ for $j^{th}$ break point. Models (1) and (2) are complete multiple structural breaks PAR(1) models (CMB-PAR(1)), which contain breaks in autoregressive coefficient, mean as well as error variance. The likelihood function for the observed data under model (1) is

$$L(\Theta\mid y)=(2\pi)^{-\frac{nT}{2}}\prod_{j=1}^{B+1}\left(\sigma_j^{-n(T_j-T_{j-1})}\right)\exp\left[-\frac{1}{2}\sum_{j=1}^{B+1}\left\{\frac{1}{\sigma_j^2}\sum_{i=1}^{n}\sum_{t=T_{j-1}}^{T_j}\left(y_{it}-\rho_j y_{i,t-1}-(1-\rho_j)\mu_{ij}\right)^2\right\}\right] \tag{3}$$

where $\Theta=\left\{\left(\mu_{ij},\sigma_j^2,\rho_j\right),\ \forall\ i=1,2,\ldots,n;\ j=1,2,\ldots,B\right\}$.

Similarly for model (2), the likelihood function is

$$L(\mu_{i1},\mu_{i2},\sigma_1^2,\sigma_2^2,\rho_1,\rho_2\mid y)=(2\pi)^{-\frac{nT}{2}}\sigma_1^{-n(T+T_1-T_2)}\sigma_2^{-n(T_2-T_1)}\exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n}\sum_{t\ne T_1+1}^{T_2}\left(y_{it}-\rho_1 y_{i,t-1}-(1-\rho_1)\mu_{i1}\right)^2\right.$$
$$\left.-\frac{1}{2\sigma_2^2}\sum_{i=1}^{n}\sum_{t=T_1+1}^{T_2}\left(y_{it}-\rho_2 y_{i,t-1}-(1-\rho_2)\mu_{i2}\right)^2\right] \tag{4}$$

## 3. Bayesian Analysis

In Bayesian inference, the current sample information is incorporated within the available prior information because the prior distribution gives additional information about the unknown parameters that are useful to improve further inference. For Bayesian estimation, prior distribution is required to obtain the estimator for unknown parameters. If enough information about the parameter is available then it is better to incorporate the informative prior, otherwise non-informative prior is considered. In general, normal and inverse gamma distributions are the most often used conjugate priors for intercept ($\mu_{ij}$) and error variance ( $\sigma_j^2$ ) parameters in various time series model (see Meligkotsidou et al. (2017)). For autoregressive coefficient, non-informative prior as a uniform distribution is considered that provides little information related to the proposed model. Therefore, we assume $\mu_{ij}$ parameter is conditionally independent and other parameters are mutually independent, having the form as

$$\pi\left(\mu_{ij} \mid \sigma_j^2\right) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[ -\frac{1}{2\sigma_j^2}\left(\mu_{ij} - \gamma_{ij}\right)^2 \right]; \quad \mu_{ij}, \gamma_{ij} \in \Re, \sigma_j > 0$$

$$\pi\left(\sigma_j^2\right) = \frac{d_j^{c_j}}{\Gamma c_j}\left(\sigma_j^2\right)^{-c_j-1} \exp\left[ -\frac{d_j}{\sigma_j^2} \right]; \quad c_j, d_j > 0$$

$$\pi\left(\rho_j\right) = \frac{1}{1-l_j}; \quad l_j > -1, -1 < \rho_j < 1$$

Then, the joint prior distribution for $\Theta = \left\{\left(\mu_{ij}, \sigma_j^2, \rho_j\right), \forall\, i = 1,2,...,n;\ j = 1,2,...,B\right\}$ is given as

$$\pi(\Theta) = (2\pi)^{-\frac{n(B+1)}{2}} \prod_{j=1}^{B+1}\left( \frac{d_j^{c_j}}{\Gamma c_j (1-l_j)}\left(\sigma_j^2\right)^{-c_j-1-\frac{n}{2}} \right) \exp\left[ -\sum_{j=1}^{B+1}\frac{1}{\sigma_j^2}\left\{ d_j + \frac{1}{2}\sum_{i=1}^{n}\left(\mu_{ij} - \gamma_{ij}\right)^2 \right\} \right] \quad (5)$$

Without loss of generality, one may know that prior distributions accurately describe the nature of the parameter and assist correctly to find the best estimator. The joint posterior distribution of PS-PAR(1) model obtained from the likelihood function given in equation (3) with incorporating the joint prior distribution given in equation (5) is expressed as

$$\pi(\Theta \mid y) = K_P (2\pi)^{-\frac{n(T+B+1)}{2}} \prod_{j=1}^{B+1}\left( \frac{d_j^{c_j}}{\Gamma(c_j)(1-l_j)}\left(\sigma_j^2\right)^{-\left[\frac{n\left(T_j - T_{j-1}+1\right)}{2} + c_j + 1\right]} \right) \exp\left[ -\frac{1}{2}\sum_{j=1}^{B+1}\left\{ \frac{1}{\sigma_j^2}\left( \sum_{i=1}^{n}\sum_{t=T_{j-1}}^{T_j}\left(y_{it} - \rho_j y_{i,t-1}\right.\right.\right.\right.$$

$$\left.\left.\left.\left. -\left(1-\rho_j\right)\mu_{ij}\right)^2 + \sum_{i=1}^{n}\left(\mu_{ij} - \gamma_{ij}\right)^2 + 2d_j \right)\right\}\right] \quad (6)$$

where $K_P$ is the normalizing constant. Using equation (6), the Bayesian estimator is obtained but due to complexity in expression under different loss functions, a numerical technique is used to solve the posterior distribution. So, we use MCMC sampler technique to generate posterior samples. For this, we obtain the form of conditional posterior distributions for PS-PAR(1) model as given by (see Gilks *et al.* (1995), page 75–76)

$$\pi\left(\mu_{ij} \mid \rho_j, \sigma_j^2, y\right) \sim N\left(\frac{B_{ij}}{A_j}, \frac{\sigma_j^2}{A_j}\right) \tag{7}$$

$$\pi\left(\sigma_j^2 \mid \rho_j, \mu_{ij}, y\right) \sim IG\left(\frac{n\left(T_j - T_{j-1}\right)+n}{2} + c_j, C_j\right) \tag{8}$$

$$\pi\left(\rho_j \mid \mu_{ij}, \sigma_j^2, y\right) \sim TN\left(\frac{E_j}{D_j}, \frac{\sigma_j^2}{D_j}, l_j, 1\right) \tag{9}$$

where

$$A_j = \left(1-\rho_j\right)^2\left(T_j - T_{j-1}\right)+1$$

$$B_{ij} = \left(1-\rho_j\right)\sum_{t=T_{j-1}+1}^{T_j}\left(y_{it} - \rho_j y_{i,t-1}\right)+\gamma_{ij}$$

$$C_j = d_j + \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{t=T_{j-1}}^{T_j}\left(y_{it} - \rho_j y_{i,t-1} - \left(1-\rho_j\right)\mu_{ij}\right)^2 + \left(\mu_{ij} - \gamma_{ij}\right)^2\right)$$

$$D_j = \sum_{i=1}^{n}\sum_{t=T_{j-1}+1}^{T_j}\left(y_{it-1} - \mu_{ij}\right)^2$$

$$E_j = \sum_{i=1}^{n}\sum_{t=T_{j-1}+1}^{T_j}\left(y_{it} - \mu_{ij}\right)\left(y_{i,t-1} - \mu_{ij}\right)$$

A temporary shifted model contains only two break points so that joint prior distribution has parameters $\Theta = \left\{\left(\mu_{i1}, \mu_{i2}, \sigma_1^2, \sigma_2^2, \rho_1, \rho_2\right), \forall\, i = 1,2,..., n\right\}$. Then, posterior distribution for the given likelihood function is obtained as

$$\pi(\Theta \mid y) = K_T \frac{(2\pi)^{-\frac{n(T+2)}{2}} d_1^{c_1} d_2^{c_2}}{\Gamma(c_1)\Gamma(c_2)(1-l_1)(1-l_2)} \left(\sigma_1^2\right)^{-\left[\frac{n(T+T_1-T_2+1)}{2}+c_1+1\right]} \left(\sigma_2^2\right)^{-\left[\frac{n(T_2-T_1+1)}{2}+c_2+1\right]}$$
$$\exp\left[-\frac{1}{2\sigma_1^2}\left\{\sum_{i=1}^{n}\sum_{t\neq T_1+1}^{T_2}\left(y_{it}-\rho_1 y_{i,t-1}-(1-\rho_1)\mu_{i1}\right)^2+\sum_{i=1}^{n}(\mu_{i1}-\gamma_{i1})^2+2d_1\right\}\right]$$
$$\exp\left[-\frac{1}{2\sigma_2^2}\left\{\sum_{i=1}^{n}\sum_{t=T_1+1}^{T_2}\left(y_{it}-\rho_2 y_{i,t-1}-(1-\rho_2)\mu_{i2}\right)^2+\sum_{i=1}^{n}(\mu_{i2}-\gamma_{i2})^2+2d_2\right\}\right] \tag{10}$$

where $K_T$ is the normalizing constant. For computing the conditional posterior distribution, one may integrate equation (10) with respect to other parameters and get the expression. The expressions of conditional posterior distribution for various parameters are (see Gilks *et al.* (1995), page 75-76)

$$\pi\left(\mu_{i1} \mid \rho_1, \sigma_1^2, y\right) \sim N\left(\frac{(1-\rho_1)\sum_{t\neq T_1+1}^{T_2}\left(y_{it}-\rho_1 y_{i,t-1}\right)+\gamma_{i1}}{(1-\rho_1)^2(T+T_1-T_2)+1}, \frac{\sigma_1^2}{(1-\rho_1)^2(T+T_1-T_2)+1}\right) \tag{11}$$

$$\pi\left(\sigma_1^2 \mid \rho_1, \mu_{i1}, y\right) \sim IG\left(\frac{n(T+T_1-T_2+1)}{2}+c_1, d_1+\frac{1}{2}\sum_{i=1}^{n}\left(\sum_{t\neq T_1+1}^{T_2}\left(y_{it}-\rho_1 y_{i,t-1}-(1-\rho_1)\mu_{i1}\right)^2+(\mu_{i1}-\gamma_{i1})^2\right)\right) \tag{12}$$

$$\pi\left(\rho_1 \mid \mu_{i1}, \sigma_1^2, y\right) \sim TN\left(\frac{\sum_{i=1}^{n}\sum_{t\neq T_1+1}^{T_2}(y_{it}-\mu_{i1})(y_{i,t-1}-\mu_{i1})}{\sum_{i=1}^{n}\sum_{t\neq T_1+1}^{T_2}(y_{it-1}-\mu_{i1})^2}, \frac{\sigma_1^2}{\sum_{i=1}^{n}\sum_{t\neq T_1+1}^{T_2}(y_{it-1}-\mu_{i1})^2}, l_1, 1\right) \tag{13}$$

$$\pi\left(\mu_{i2} \mid \rho_2, \sigma_2^2, y\right) \sim N\left(\frac{(1-\rho_2)\sum_{t=T_1+1}^{T_2}\left(y_{it}-\rho_2 y_{i,t-1}\right)+\gamma_{i2}}{(1-\rho_2)^2(T_2-T_1)+1}, \frac{\sigma_2^2}{(1-\rho_2)^2(T_2-T_1)+1}\right) \tag{14}$$

$$\pi\left(\sigma_2^2 \mid \rho_2, \mu_{i2}, y\right) \sim IG\left(\frac{n(T_2-T_1+1)}{2}+c_2, d_2+\frac{1}{2}\sum_{i=1}^{n}\left(\sum_{t=T_1+1}^{T_2}\left(y_{it}-\rho_2 y_{i,t-1}-(1-\rho_2)\mu_{i2}\right)^2 + (\mu_{i2}-\gamma_{i2})^2\right)\right) \tag{15}$$

$$\pi\left(\rho_2 \mid \mu_{i2}, \sigma_2^2, y\right) \sim TN\left(\frac{\sum_{i=1}^{n}\sum_{t=T_1+1}^{T_2}\left(y_{it}-\mu_{i2}\right)\left(y_{i,t-1}-\mu_{i2}\right)}{\sum_{i=1}^{n}\sum_{t=T_1+1}^{T_2}\left(y_{it-1}-\mu_{i2}\right)^2}, \frac{\sigma_2^2}{\sum_{i=1}^{n}\sum_{t=T_1+1}^{T_2}\left(y_{it-1}-\mu_{i2}\right)^2}, l_2, 1\right) \tag{16}$$

For getting better estimator form the conditional posterior distribution, a suitable loss function is generally adopted. The commonly used loss function is squared error (symmetric) loss function (SELF) that takes equal magnitude due to over and under-estimation and another one is entropy (asymmetric) loss function (ELF). The Bayes estimator and its posterior risk for both loss functions are described below:

| Loss Function | Bayes Estimator | Posterior Risk |
|---|---|---|
| $\text{SELF} = \left(\theta - \hat{\theta}\right)^2$ | $E(\theta \mid y)$ | $Var(\theta \mid y)$ |
| $\text{ELF} = \left[\frac{\hat{\theta}}{\theta} - \ln\frac{\hat{\theta}}{\theta} - 1\right]$ | $\left(E\left(\theta^{-1} \mid y\right)\right)^{-1}$ | $E(\ln\theta \mid x) - \ln\left(E\left(\theta^{-1} \mid x\right)\right)$ |

It is obvious that the form of the posterior distribution will not be tractable and the computation of its respective Bayes estimator under different loss functions will not be analytically obtained. Consequently, one can choose stochastic simulation procedures, namely, the Gibbs and Metropolis samplers (Gilks *et al.*, 1995) to generate samples from the posterior distributions. Then, compute Bayes estimates of the parameters and their

corresponding interval. This study utilizes the following steps to obtain the posterior samples using Gibbs sampling algorithm:

1. Starting with initial values $\rho_j^{(0)}, \mu_{ij}^{(0)}, \left(\sigma_j^2\right)^{(0)}$ and set k=1
2. Generate $\mu_{ij}^{(k)}$ from conditional posterior density $\pi\left(\mu_{ij}^{(k)} \mid \rho^{(k-1)}, \left(\sigma_j^2\right)^{(k-1)}, \underline{y}\right)$.
3. Generate $\left(\sigma_j^2\right)^{(k)}$ from conditional posterior density $\pi\left(\left(\sigma_j^2\right)^{(k)} \mid \rho^{(k-1)}, \mu_{ij}^{(k-1)}, \underline{y}\right)$.
4. Generate $\rho_j^{(k)}$ from conditional posterior density $\pi\left(\rho_j^{(k)} \mid \mu_{ij}^{(k-1)}, \left(\sigma_j^2\right)^{(k-1)}, \underline{y}\right)$.
5. Set k=k+1
6. Repeat steps 3-6, P times and record the sequence of observations of parameters.
7. Obtained Bayes estimate under different loss functions.

## 4. Simulation Study

To investigate and compare the performance of the various proposed estimators, a simulation study is conducted to observe the behaviour of the proposed models for various values of true parameters. For generating a series of sample size 1000, consider the following series size T =200 with different break points combination {(T/4, T/2); (T/2, 3T/4); (T/4, 3T/4)} for a set of true value: $(y_{10}, y_{20}, y_{30}) = (20, 40, 60)$; $(\rho_1, \rho_2, \rho_3) = (0.8, 0.85, 0.9)$; $\left(\sigma_1^2, \sigma_2^2, \sigma_3^2\right) = (2, 3, 4)$; $(\mu_{11}, \mu_{12}, \mu_{13}) = (10, 35, 65)$; $(\mu_{21}, \mu_{22}, \mu_{23}) = (15, 40, 70)$ and $(\mu_{31}, \mu_{32}, \mu_{33}) = (20, 45, 75)$. For numerical purpose, hyper parameters are to be known in normal and inverse gamma prior. We have taken $c_j$ = 0.01, $d_j$ = 1 for all break points and normal prior mean is equal to average of the generated series at $(T_{j-1}, T_j)$ break interval with parallel variance given in disturbances term. For simulation experiment, each pair of break point series is generated based on 10,000 replications. The generated samples are obtained using an iterative procedure of Gibbs sampling algorithm and get the estimates. We mainly compare the performances of the Bayes estimator with the maximum likelihood estimator (MLE) by calculating average absolute biases (AB) and mean squared error (MSE). A confidence interval (CI) of MLE and highest posterior density (HPD) interval of the Bayes estimator are also computed. Tables 1-6 report the MSE, AB and confidence/HPD interval of all parameters present in both permanent and temporary shifted models.

**Table-1**. MSE, AB and CI/HPD of $\mu$ parameter under PS-PAR(1) model

| $T_B$ | Estimator | $\mu_{11}$ | | | $\mu_{12}$ | | | $\mu_{13}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | AB | CI/HPD | MSE | AB | CI/HPD | MSE | AB | CI/HPD |
| (T/4,T/2) | MLE | 1.0097 | 0.7992 | (8.3520,11.6731) | 2.8189 | 1.3439 | (32.1162,37.6330) | 4.1938 | 1.6321 | (61.4577,68.2851) |
| | SELF | 0.4467 | 0.5310 | (8.8963,11.1026) | 0.7323 | 0.6838 | (33.5560,36.3717) | 1.0261 | 0.8037 | (63.2851,66.6140) |
| | ELF | 0.4587 | 0.5385 | (8.8201,11.0390) | 0.7386 | 0.6865 | (33.5144,36.3355) | 1.0315 | 0.8057 | (63.2497,66.5852) |
| (T/4,3T/4) | MLE | 1.0214 | 0.8077 | (8.3261,11.6969) | 1.3548 | 0.9268 | (33.0093,36.8435) | 6.2002 | 2.4134 | (59.7821,69.6432) |
| | SELF | 0.4480 | 0.5349 | (8.8875,11.1057) | 0.6333 | 0.6338 | (34.6719,36.3143) | 0.9074 | 0.7529 | (63.3898,66.4942) |
| | ELF | 0.4619 | 0.5440 | (8.8113,11.0439) | 0.6345 | 0.6340 | (34.6440,36.2917) | 0.9193 | 0.7584 | (63.3398,66.4532) |
| (T/2, 3T/4) | MLE | 0.5090 | 0.5702 | (8.7913,11.1464) | 2.8941 | 1.3561 | (32.0412,.37.5400) | 6.0916 | 2.4069 | (59.7758,69.7308) |
| | SELF | 0.3249 | 0.4553 | (9.0410,10.9053) | 0.7501 | 0.6878 | (33.5291,36.3639) | 0.8998 | 0.7517 | (63.4293,66.5347) |
| | ELF | 0.3303 | 0.4589 | (8.9923,10.8737) | 0.7589 | 0.6916 | (33.4820,36.3237) | 0.9088 | 0.7561 | (63.3835,66.4915) |
| $T_B$ | Estimator | $\mu_{21}$ | | | $\mu_{22}$ | | | $\mu_{23}$ | | |
| | | MSE | AB | CI/HPD | MSE | AB | CI/HPD | MSE | AB | CI/HPD |
| (T/4,T/2) | MLE | 1.0579 | 0.8203 | (13.3809,16.7410) | 2.9550 | 1.3646 | (37.0322,42.7015) | 4.2959 | 1.6558 | (66.4581,73.2309) |
| | SELF | 0.4524 | 0.5361 | (13.9168,16.1238) | 0.7642 | 0.6949 | (38.5207,41.3918) | 1.0474 | 0.8145 | (68.2761,71.6419) |
| | ELF | 0.4560 | 0.5387 | (13.8672,16.0811) | 0.7689 | 0.6967 | (38.4815,41.3583) | 1.0519 | 0.8162 | (68.2468,71.6148) |
| (T/4,3T/4) | MLE | 1.0466 | 0.8161 | (13.3834,16.7071) | 1.4273 | 0.9550 | (37.9952,41.9061) | 6.1377 | 2.4031 | (64.7961,74.6504) |
| | SELF | 0.4472 | 0.5327 | (13.9129,16.1081) | 0.6651 | 0.6507 | (38.6354,41.3231) | 0.8987 | 0.7514 | (68.3781,71.5061) |
| | ELF | 0.4525 | 0.5356 | (13.8627,16.0628) | 0.6670 | 0.6517 | (38.6153,41.2989) | 0.9076 | 0.7550 | (68.3328,71.4683) |
| (T/2, 3T/4) | MLE | 0.5177 | 0.5759 | (13.8422,16.2097) | 2.8491 | 1.3429 | (37.0584,42.6770) | 9.3331 | 2.4443 | (64.6742,74.5916) |
| | SELF | 0.3274 | 0.4569 | (14.0637,15.9513) | 0.7371 | 0.6835 | (38.5278,41.3831) | 0.9208 | 0.7632 | (68.3538,71.4903) |
| | ELF | 0.3291 | 0.4575 | (14.0380,15.9231) | 0.7416 | 0.6849 | (38.4910,41.3515) | 0.9317 | 0.7677 | (68.3074,71.4531) |
| $T_B$ | Estimator | $\mu_{31}$ | | | $\mu_{32}$ | | | $\mu_{33}$ | | |
| | | MSE | AB | CI/HPD | MSE | AB | CI/HPD | MSE | AB | CI/HPD |
| (T/4,T/2) | MLE | 1.1202 | 0.8484 | (18.3352,21.8126) | 2.8274 | 1.3391 | (42.1176,47.5853) | 4.1250 | 1.6185 | (71.4899,78.1837) |
| | SELF | 0.4531 | 0.5383 | (18.9068,21.1262) | 0.7377 | 0.6808 | (43.5558,46.3537) | 1.0014 | 0.7949 | (73.2988,76.5888) |
| | ELF | 0.4555 | 0.5395 | (18.8665,21.0915) | 0.7431 | 0.6831 | (43.5204,46.3205) | 1.0052 | 0.7963 | (73.2724,76.5637) |
| (T/4,3T/4) | MLE | 1.1247 | 0.8477 | (18.3371,21.8030) | 1.4071 | 0.9446 | (43.0046,46.8866) | 9.4117 | 2.4297 | (69.6114,79.7753) |
| | SELF | 0.4522 | 0.5374 | (18.9187,21.1211) | 0.6536 | 0.6443 | (43.6595,46.3211) | 0.9259 | 0.7596 | (73.3484,76.5294) |
| | ELF | 0.4546 | 0.5394 | (18.8757,21.0888) | 0.6544 | 0.6443 | (43.6383,46.2961) | 0.9353 | 0.7637 | (73.3094,76.4956) |
| (T/2, 3T/4) | MLE | 0.5293 | 0.5790 | (18.8400,21.2299) | 2.8747 | 1.3513 | (42.0962,47.6055) | 9.0290 | 2.3882 | (69.6876,79.5348) |
| | SELF | 0.3290 | 0.4569 | (19.0554,20.9419) | 0.7446 | 0.6863 | (43.5311,46.5311) | 0.8870 | 0.7422 | (73.3785,76.4407) |
| | ELF | 0.3310 | 0.4585 | (19.0332,20.9224) | 0.7500 | 0.6889 | (43.4978,46.3552) | 0.8973 | 0.7468 | (73.3380,76.4061) |

**Table-2**. MSE, AB and CI/HPD of $\rho$ parameter under PS-PAR(1) model

| $T_B$ | Estimator | $\rho_1$ | | | $\rho_2$ | | | $\rho_3$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | AB | CI/HPD | MSE | AB | CI/HPD | MSE | AB | CI/HPD |
| (T/4,T/2) | MLE | 3.86E-04 | 0.0155 | (0.7617,0.8231) | 6.05E-04 | 0.0191 | (0.8009,0.8735) | 3.45E-04 | 0.0144 | (0.8623,0.9156) |
| | SELF | 3.08E-04 | 0.0140 | (0.7775, 0.8152) | 3.85E-04 | 0.0156 | (0.8226,0.8665) | 2.37E-04 | 0.0120 | (0.8897,0.9088) |
| | ELF | 3.11E-04 | 0.0141 | (0.7710,0.8149) | 3.91E-04 | 0.0157 | (0.8220,0.8662) | 2.40E-04 | 0.0121 | (0.8893,0.9086) |
| (T/4,3T/4) | MLE | 3.70E-04 | 0.0153 | (0.7635,0.8238) | 4.40E-04 | 0.0165 | (0.8018,0.8718) | 5.46E-04 | 0.0183 | (0.8537,0.9293) |
| | SELF | 3.01E-04 | 0.0139 | (0.7786,0.8159) | 3.42E-04 | 0.0147 | (0.8250,0.8653) | 2.72E-04 | 0.0130 | (0.8776,0.9099) |
| | ELF | 3.04E-04 | 0.0140 | (0.7782,0.8156) | 3.46E-04 | 0.0147 | (0.8245,0.8650) | 2.76E-04 | 0.0131 | (0.8772,0.9097) |
| (T/2, 3T/4) | MLE | 3.12E-04 | 0.0140 | (0.7660,0.8211) | 6.39E-04 | 0.0197 | (0.8000,0.8741) | 5.55E-04 | 0.0182 | (0.8527,0.9188) |
| | SELF | 2.75E-04 | 0.0132 | (0.7801,0.8139) | 4.06E-04 | 0.0160 | (0.8221,0.8669) | 2.73E-04 | 0.0129 | (0.8775,0.9198) |
| | ELF | 2.77E-04 | 0.0133 | (0.7796,0.8136) | 4.12E-04 | 0.0162 | (0.8215,0.8665) | 2.77E-04 | 0.0130 | (0.8771,0.9196) |

**Table-3**. MSE, AB and CI/HPD of $\sigma^2$ parameter under PS-PAR(1) model

| $T_B$ | Estimator | $\sigma_1^2$ | | | $\sigma_2^2$ | | | $\sigma_3^2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | AB | CI/HPD | MSE | AB | CI/HPD | MSE | AB | CI/HPD |
| (T/4,T/2) | MLE | 0.0553 | 0.1871 | (1.6443,2.4148) | 0.1313 | 0.2881 | (2.4492,3.6385) | 0.1075 | 0.2624 | (3.4804,4.5571) |
| | SELF | 0.0526 | 0.1829 | (1.8313,2.2798) | 0.1257 | 0.2830 | (2.6303,3.3918) | 0.1052 | 0.2598 | (3.7656,4.3302) |
| | ELF | 0.0524 | 0.1824 | (1.9096,2.2477) | 0.1249 | 0.2825 | (2.6976,3.3445) | 0.1052 | 0.2600 | (3.7420,4.3005) |
| (T/4,3T/4) | MLE | 0.0561 | 0.1885 | (1.6373,2.4258) | 0.0623 | 0.2001 | (2.6040,3.4266) | 0.2228 | 0.3725 | (3.2728,4.8284) |
| | SELF | 0.0534 | 0.1846 | (1.7258,2.2931) | 0.0609 | 0.1982 | (2.7886,3.3007) | 0.2150 | 0.3676 | (3.4610,4.6897) |
| | ELF | 0.0530 | 0.1848 | (1.7034,2.2607) | 0.0612 | 0.1988 | (2.7713,3.2783) | 0.2125 | 0.3677 | (3.4182,4.6233) |
| (T/2, 3T/4) | MLE | 0.0278 | 0.1317 | (1.7418,2.2914) | 0.1246 | 0.2800 | (2.4502,3.6277) | 0.2198 | 0.3721 | (3.2543,4.7904) |
| | SELF | 0.0269 | 0.1299 | (1.8322,2.1748) | 0.1190 | 0.2738 | (2.6383,3.3839) | 0.2114 | 0.3654 | (3.5519,4.5576) |
| | ELF | 0.0269 | 0.1300 | (1.8209,2.1597) | 0.1182 | 0.2739 | (2.6046,3.3360) | 0.2098 | 0.3652 | (3.5085,4.4927) |

**Table-4**. MSE, AB and CI/HPD of $\mu$ parameter under TS-PAR(1) model

| $T_B$ | Estimator | $\mu_{11}$ | | | $\mu_{12}$ | | |
|---|---|---|---|---|---|---|---|
| | | MSE | AB | CI/HPD | MSE | AB | CI/HPD |
| (T/4,T/2) | MLE | 0.3484 | 0.4701 | (9.0470,10.9694) | 9.6882 | 2.4927 | (59.7725,70.0058) |
| | SELF | 0.2544 | 0.4018 | (9.1776,10.6238) | 0.8789 | 0.7509 | (63.4696,66.5473) |
| | ELF | 0.2569 | 0.4041 | (9.1451,10.7981) | 0.8860 | 0.7538 | (63.4210,66.5083) |
| (T/4,3T/4) | /.MLE | 0.5264 | 0.5769 | (8.8041,11.1993) | 4.3461 | 1.6701 | (61.4640,68.2793) |
| | SELF | 0.3320 | 0.4578 | (9.0383,10.9339) | 1.0267 | 0.8107 | (63.2910,66.6277) |
| | ELF | 0.3380 | 0.4618 | (8.9952,10.9010) | 1.0311 | 0.8126 | (63.2588,66.6000) |
| (T/2, 3T/4) | MLE | 0.3564 | 0.4763 | (9.0041,10.9704) | 9.5505 | 2.4672 | (59.5605,69.8348) |
| | SELF | 0.2601 | 0.4070 | (9.1457,10.8226) | 0.8662 | 0.7403 | (63.3814,66.5010) |
| | ELF | 0.2627 | 0.4092 | (9.1124,10.7953) | 0.8746 | 0.7442 | (63.3389,66.4576) |
| $T_B$ | Estimator | $\mu_{21}$ | | | $\mu_{22}$ | | |
| | | MSE | AB | CI/HPD | MSE | AB | CI/HPD |
| (T/4,T/2) | MLE | 0.3389 | 0.4640 | (14.0701,15.9820) | 9.0765 | 2.4004 | (64.7812,74.5734) |
| | SELF | 0.2460 | 0.3951 | (14.2963,15.6310) | 0.8259 | 0.7231 | (68.4579,71.4092) |
| | ELF | 0.2470 | 0.3958 | (14.2751,15.6134) | 0.8342 | 0.7265 | (68.4169,71.3705) |
| (T/4,3T/4) | MLE | 0.5396 | 0.5827 | (13.8021,16.2317) | 4.2236 | 1.6387 | (66.5093,73.2619) |
| | SELF | 0.3371 | 0.4609 | (14.0440,15.9567) | 1.0094 | 0.7992 | (68.2958,71.6303) |
| | ELF | 0.3397 | 0.4627 | (14.0114,15.9304) | 1.0123 | 0.8003 | (68.2650,71.5994) |
| (T/2, 3T/4) | MLE | 0.3463 | 0.4674 | (14.0265,15.9676) | 8.9987 | 2.3973 | (64.9211,74.7977) |
| | SELF | 0.2516 | 0.3983 | (14.1652,15.8195) | 0.8139 | 0.7184 | (68.5092,71.4686) |
| | ELF | 0.2528 | 0.3989 | (14.1437,15.8033) | 0.8190 | 0.7211 | (68.4693,71.4312) |
| $T_B$ | Estimator | $\mu_{31}$ | | | $\mu_{32}$ | | |
| | | MSE | AB | CI/HPD | MSE | AB | CI/HPD |
| (T/4,T/2) | MLE | 0.3410 | 0.4622 | (19.0455,20.9890) | 9.1295 | 2.4025 | (69.7670,79.6721) |
| | SELF | 0.2465 | 0.3930 | (19.1830,20.7309) | 0.8263 | 0.7213 | (73.4572,76.4584) |
| | ELF | 0.2470 | 0.3934 | (19.1673,20.7180) | 0.8335 | 0.7245 | (73.4118,76.4259) |
| (T/4,3T/4) | MLE | 0.5579 | 0.5957 | (18.7864,21.2446) | 4.1963 | 1.6336 | (71.5202,78.2553) |
| | SELF | 0.3452 | 0.4683 | (19.0392,20.9707) | 0.9911 | 0.7936 | (73.3388,76.6031) |
| | ELF | 0.3464 | 0.4691 | (19.0184,20.9520) | 0.9944 | 0.7948 | (73.3101,76.5762) |
| (T/2, 3T/4) | MLE | 0.3559 | 0.4773 | (19.0131,20.9763) | 9.7078 | 2.4736 | (69.6701,79.8414) |
| | SELF | 0.2579 | 0.4063 | (19.1466,20.8161) | 0.8784 | 0.7447 | (73.4009,76.4541) |
| | ELF | 0.2590 | 0.4071 | (19.1316,20.8015) | 0.8865 | 0.7481 | (73.3617,76.4218) |

**Table-5**. MSE, AB and CI/HPD of $\rho$ parameter under TS-PAR(1) model

| $T_B$ | Estimator | $\rho_1$ | | | $\rho_2$ | | |
|---|---|---|---|---|---|---|---|
| | | MSE | AB | CI/HPD | MSE | AB | CI/HPD |
| (T/4,T/2) | MLE | 6.84E-05 | 0.0066 | (0.7852,0.8120) | 1.44E-04 | 0.0095 | (0.8765,0.9140) |
| | SELF | 6.50E-05 | 0.0064 | (0.7863,0.8127) | 8.61E-05 | 0.0074 | (0.8828,0.9031) |
| | ELF | 6.51E-05 | 0.0064 | (0.7863,0.8126) | 8.66E-05 | 0.0074 | (0.8827,0.9030) |
| (T/4,3T/4) | MLE | 7.53E-05 | 0.0068 | (0.7842,0.8125) | 1.06E-04 | 0.0081 | (0.8795,0.9115) |
| | SELF | 6.94E-05 | 0.0066 | (0.7865,0.8109) | 8.22E-05 | 0.0072 | (0.8828,0.9109) |
| | ELF | 6.96E-05 | 0.0066 | (0.7864.0.8108) | 8.27E-05 | 0.0072 | (0.8827,0.9100) |
| T/2, 3T/4 | MLE | 6.78E-05 | 0.0065 | (0.7851,0.8122) | 1.46E-04 | 0.0094 | (0.8758,0.9138) |
| | SELF | 6.48E-05 | 0.0064 | (0.7861,0.8130) | 8.51E-05 | 0.0072 | (0.8831,0.9128) |
| | ELF | 6.49E-05 | 0.0064 | (0.7860,0.8129) | 8.56E-05 | 0.0072 | (0.8830,0.9128) |

**Table-6**. MSE, AB and CI/HPD of $\sigma^2$ parameter under TS-PAR(1) model

| $T_B$ | Estimator | $\sigma_1^2$ | | | $\sigma_2^2$ | | |
|---|---|---|---|---|---|---|---|
| | | MSE | AB | CI/HPD | MSE | AB | CI/HPD |
| (T/4,T/2) | MLE | 0.0179 | 0.1060 | (1.7948,2.2324) | 0.2226 | 0.3779 | (3.2909,4.8390) |
| | SELF | 0.0113 | 0.0810 | (1.8216,2.1787) | 0.1358 | 0.2834 | (3.5064,4.5448) |
| | ELF | 0.0113 | 0.0811 | (1.9124,2.1674) | 0.1355 | 0.2842 | (4.5515,4.4714) |
| (T/4,3T/4) | MLE | 0.0266 | 0.1292 | (1.7484,2.2860) | 0.1069 | 0.2602 | (3.4820,4.5781) |
| | SELF | 0.0163 | 0.0978 | (1.8903,2.2176) | 0.0673 | 0.1976 | (3.6646,4.3453) |
| | ELF | 0.0165 | 0.0987 | (1.8756,2.1985) | 0.0678 | 0.1988 | (3.6351,4.3069) |
| T/2, 3T/4 | MLE | 0.0174 | 0.1051 | (1.7963,2.2275) | 0.2146 | 0.3666 | (3.2875,4.8103) |
| | SELF | 0.0111 | 0.0804 | (1.8269,2.1748) | 0.1319 | 0.2756 | (3.4099,4.6227) |
| | ELF | 0.0112 | 0.0809 | (1.8176,2.1635) | 0.1330 | 0.2794 | (3.3519,4.5439) |

For the simulation study, we observed that both PS-PAR(1) and TS-PAR(1) models are having minimum AB and average MSE when estimated through the Bayesian estimator as compared to MLE. It is also observed that there is a considerable difference in AB and MSE in respective sets of break points on both models with complete and temporary shifts. We observe the same performance of the Bayes estimates under both symmetric and asymmetric loss functions and approximately same magnitude in terms of their MSE and AB.

## 5. Real Data Analysis

An empirical application is the way of analysis of real data to get the applicability of the proposed model. There are sufficient studies that show a change on economic series due to a change on economic policy, trade strategy, market fluctuation, etc. For example, present scenario of India is making several policies specially demonetization, good and service tax (GST), which may be improving the economic condition in the future. For analysis purpose, we have taken annual series of gross domestic product (GDP) per capita of South Asian Association for Regional Cooperation (SAARC) countries over the period from 1981 to 2016. Due to restrictions in data availability, it was not possible to include the economy series of Afghanistan as it is available since

2002. GDP per capita determines the growth of the economy of a country and compares it with its trading participant countries as well as applies it in better economic analysis and policy-making in the future. Over the world, SAARC association has a common cultural background and shared political experience and decides five areas namely agriculture, rural development, telecommunications, meteorology, health and population activities, where economic prosperity is the best achieved. The purpose is to investigate whether the presence of break point(s) in GDP per capita series may be varying due to a change in all model parameters or not and then find the estimates of the parameter for the best fitted model. For better understanding, we require a strongly balanced panel that has multiple breaks at the same time point. For this, it is natural to determine the number and location of structural breaks, which is developed by Zeileis *et al.* (2002). The most preferred break point(s) and its location for GDP per capita series for all countries are summarized in Table 7.

**Table-7**. Number of breaks and its location for GDP series of SAARC countries

| Country | Number of Breaks | $T_1$ | $T_2$ |
|---|---|---|---|
| Bangladesh | 1 | 2008 | - |
| Bhutan | 2 | 1997 | 2008 |
| India | 2 | 1997 | 2008 |
| Maldives | 2 | 1997 | 2008 |
| Nepal | 1 | 2008 | - |
| Pakistan | 2 | 1992 | 2008 |
| Sri Lanka | 2 | 1994 | 2008 |

Results reported in Table 7 indicate that the break arises mostly in 1997 and 2008. These break points occur when Asian financial crisis and Global financial crisis happened. These financial crises were analysed by various researchers from both theoretical and application point of view. To study the PS-PAR(1) model, assembly Bhutan, India and Maldives as a panel, which has similar break points $T_B = (1997, 2008)$ and compute the estimated values of the proposed model. To check the validity of the proposed PS-PAR(1) model to the other change point models which have a break in lesser number of parameter(s), i.e. incomplete multiple breaks PAR(1) models. For GDP per capita series, we verify the applicability of PS-PAR(1) model using Akaike information criterion (AIC) and Bayesian information criterion (BIC). The AIC and BIC values are based on the likelihood function, which needs to be determined by Bayesian estimators. The mathematical formula for the calculation of AIC and BIC is

$$AIC = -2\log L\left(\hat{\Theta} \mid y\right) + 2K$$

$$BIC = -2\log L\left(\hat{\Theta} \mid y\right) + K\log(nT)$$

where $L\left(\hat{\Theta} \mid y\right)$ is the likelihood of the PS-PAR(1) model given the data when it is evaluated at the Bayesian estimator of $\Theta$ for 1000 iterations and K is the number of

estimated parameters in the proposed model. The results are obtained by taking the average of all values of AIC and BIC.

Table 8 records the AIC and BIC values for each model as per the break presence in the parameters. From Table 8, one can observe that the PS-PAR(1) model with a break present in autoregressive coefficient, mean and error variance having minimum AIC and BIC value with other permanent shifted models at breaks (1997, 2008). Hence, the PS-PAR(1) model is well fitted for the GDP series. We also verify the result based on the Bayes factor. The Bayes factor is the ratio of posterior probability under null and alternative hypothesis. Higher values of the Bayes factor lead to rejection of null hypothesis. This shows that series is well fitted from the alternative model, i.e. proposed model. Hence, Table 9 records the value of Bayes factor (BF) to take decision about the best fitted model. This table shows that there is a strong evidence to support the presence of breaks in all parameters as Bayes factor is so much high to reject the null hypothesis. Overall, we conclude that PS-PAR(1) model is well fitted for the GDP series at breaks (1997, 2008).

**Table-8**. Selection the parameter(s) shifting in PS-PAR(1) model using information criterion

| Model | Break in Parameter(s) | -logL | AIC | BIC |
|---|---|---|---|---|
| PAR($\rho_j$, $\mu_{ij}$, $\sigma_j$) | AR coefficient, mean & error variance | 205.8028 | 441.6056 | 481.8375 |
| PAR($\rho_j$, $\mu_{ij}$, $\sigma$) | AR coefficient & mean | 250.9187 | 527.8375 | 562.7052 |
| PAR($\rho_j$, $\mu_i$, $\sigma_j$) | AR coefficient & error variance | 221.6518 | 461.3035 | 485.4427 |
| PAR($\rho$, $\mu_{ij}$, $\sigma_j$) | Mean & error variance | 471.7729 | 969.5459 | 1004.414 |
| PAR($\rho_j$, $\mu_i$, $\sigma$) | AR coefficient | 232.5329 | 479.0658 | 497.8407 |
| PAR($\rho$, $\mu_{ij}$, $\sigma$) | Mean | 8.35E+28 | 1.67E+29 | 1.67E+29 |
| PAR($\rho$, $\mu_i$, $\sigma_j$) | Error variance | 2.15E+30 | 4.30E+30 | 4.30E+30 |
| PAR($\rho$, $\mu_i$, $\sigma$) | - | 7.19E+30 | 1.44E+31 | 1.44E+31 |

**Table-9**. Model selection using Bayes factor when alternative hypothesis ($H_1$) considers multiple breaks in all parameters

| Model | Null hypothesis ($H_0$ ) consider breaks in | BF | Evidence against $H_0$ |
|---|---|---|---|
| PAR($\rho_j$, $\mu_{ij}$, $\sigma$) | AR coefficient & mean | 1.13E+34 | Very Strong |
| PAR($\rho_j$, $\mu_i$, $\sigma_j$) | AR coefficient & error variance | 6.26E+13 | Very Strong |
| PAR($\rho$, $\mu_{ij}$, $\sigma_j$) | Mean & error variance | 1.20E+11 | Very Strong |
| PAR($\rho_j$, $\mu_i$, $\sigma$) | AR coefficient | 1.00E+38 | Very Strong |
| PAR($\rho$, $\mu_{ij}$, $\sigma$) | Mean | 9.69E+29 | Very Strong |
| PAR($\rho$, $\mu_i$, $\sigma_j$) | Error variance | 3.09E+20 | Very Strong |
| PAR($\rho$, $\mu_i$, $\sigma$) | - | 4.60E+30 | Very Strong |

After identifying the best suitable model, the estimated value of the maximum likelihood and Bayesian estimators of PS-PAR(1) model parameters are summarized in Table 10.

**Table-10**. MLE and Bayes estimates based on GDP series using PS-PAR(1) model

| Parameter | MLE | SELF | ELF |
|---|---|---|---|
| $\rho_1$ | 9.54E-01 | 9.97E-01 | 9.97E-01 |
| $\rho_2$ | 9.61E-01 | 9.66E-01 | 9.66E-01 |
| $\rho_3$ | 9.78E-01 | 9.29E-01 | 9.29E-01 |
| $\mu_{11}$ | 3.15E+02 | 4.65E+02 | 4.38E+02 |
| $\mu_{21}$ | 2.80E+02 | 3.41E+02 | 2.85E+02 |
| $\mu_{31}$ | 2.97E+02 | 1.17E+03 | 1.16E+03 |
| $\mu_{12}$ | 5.54E+02 | 1.01E+03 | 1.80E+03 |
| $\mu_{22}$ | 2.94E+02 | 6.33E+02 | 4.98E+02 |
| $\mu_{32}$ | 1.93E+03 | 3.84E+03 | 3.81E+03 |
| $\mu_{13}$ | 8.07E+03 | 2.34E+03 | 2.27E+03 |
| $\mu_{23}$ | 4.72E+03 | 1.50E+03 | 1.33E+03 |
| $\mu_{33}$ | 3.80E+04 | 9.70E+03 | 1.00E+04 |
| $\sigma_1^2$ | 2.45E+03 | 7.42E+04 | 7.84E+04 |
| $\sigma_2^2$ | 1.03E+04 | 1.24E+05 | 2.21E+05 |
| $\sigma_3^2$ | 9.95E+03 | 1.16E+05 | 1.78E+05 |

## 6. Conclusion

There is a sufficient literature on the time series model with a structural break, which allows a break on mean and variance, but the present paper has extended the frontier of knowledge in a PAR(1) model, which allows a break on all parameters of the model at multiple time points, and carried out the Bayesian analysis. Sometimes, changes on parameters are temporary, so the model with a temporary shift is also discussed. It recorded better results in a simulation study. An empirical application on GDP per capita time series of SARRC association is applied to PS-PAR(1) model and it is observed that both Asian and World financial crises have affected the GDP series of SAARC countries due to a break in all parameters permanently and the same may be applied in other areas like insurance, agriculture, administrative, crime, etc. The result may be extended for other structural break models with non-normal error and time trend.

## References

AGIWAL, V., KUMAR, J., SHANGODOYIN, D. K., (2018). A Bayesian inference of multiple structural breaks in mean and error variance in panel AR(1) model. *Statistics in Transition*, 19(1), pp. 7–23.

ALBERT, J. H., CHIB, S., (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business & Economic Statistics*, 11(1), pp. 1–15.

ALTISSIMO, F., CORRADI, V., (2003). Strong rules for detecting the number of breaks in a time series. *Journal of Econometrics*, 117(2), pp. 207–244.

BAI, J., (2010). Common breaks in means and variances for panel data. *Journal of Econometrics,* 157(1), pp. 78–92.

BAI, J., PERRON, P., (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1), pp. 47–78.

BALTAGI, B. H., FENG, Q., KAO, C., (2016). Estimation of heterogeneous panels with structural breaks. *Journal of Econometrics*, 191(1), pp. 176–195.

BARDWELL, L., FEARNHEAD, P., ECKLEY, I. A., SMITH, S., SPOTT, M., (2019). Most recent change point detection in panel data. *Technometrics*, 61(1), pp. 88–98.

CHEN, B., HUANG, L., (2018). Nonparametric testing for smooth structural changes in panel data models. *Journal of Econometrics*, 202(2), pp. 245–267.

CHIN, W. C., LEE, M. C., YAP, G. L. C., (2016). Heterogeneous autoregressive model with structural break using nearest neighbor truncation volatility estimators for DAX. *SpringerPlus*, 5, pp. 1–13.

ELLIOTT, G., MULLER, U. K., (2007). Confidence sets for the date of a single break in linear time series regressions. *Journal of Econometrics*, 141, pp. 1196–1218.

EO, Y., (2012). Bayesian inference about the types of structural breaks when there are different breaks in many parameters. Available at SSRN: http://ssrn. com/abstract, volume 2011825.

GILKS, W. R., RICHARDSON, S., SPIEGELHALTER, D., (1995*). Markov chain Monte Carlo in practice*. Chapman and Hall/CRC.

HWANG, E., SHIN, D.W., (2017). Stationary bootstrapping for structural break tests for a heterogeneous autoregressive model. *Communications for Statistical Applications and Methods*, 24(4), pp. 367–382.

JIBRIN, S. A., MUSA, Y., ZUBAIR, U. A., SAIDU, A. S., (2015). ARFIMA modelling and investigation of structural break (s) in West Texas Intermediate and Brent series. *CBN Journal of Applied Statistics*, 6(2), pp. 59–79.

JIN, B., SHI, X., WU, Y., (2013). A novel and fast methodology for simultaneous multiple structural break estimation and variable selection for nonstationary time series models. *Statistics and Computing*, 23(2), pp. 1–11.

KARAVIAS, Y., TZAVALIS, E., (2017). Local power of panel unit root tests allowing for structural breaks. *Econometric Reviews*, 36(10), pp. 1123–1156.

KUMAR, R., KUMAR, J., CHATURVEDI, A., (2012). Bayesian unit root test for time series models with structural break in variance. *Journal of Economics and Econometrics*, 55(1), pp. 75–86.

MAHEU, J. M., SONG, Y., (2018). An efficient Bayesian approach to multiple structural change in multivariate time series. *Journal of Applied Econometrics*, 33(2), pp. 251–270.

MELIGKOTSIDOU, L., TZAVALIS, E., VRONTOS, I. D., (2017). On Bayesian analysis and unit root testing for autoregressive models in the presence of multiple structural breaks. *Econometrics and Statistics,* 4, pp. 70–90.

OKUI, R., WANG, W., (2018). Heterogeneous structural breaks in panel data models. Available at SSRN: https://ssrn.com/abstract=3031689 or http://dx.doi.org/10.2139/ssrn.3031689.

PESARAN, M. H., (2006). Estimation and inference in large heterogeneous panels with multifactor error structure. *Econometrica*, 74, pp. 967–1012.

PESTOVA, B., PESTA, M., (2017). Change point estimation in panel data without boundary issue. *Risks*, *5*(1), pp.1–22.

TOPÁL, D., MATYASOVSZKYT, I., KERN, Z., HATVANI, I. G., (2016). Detecting breakpoints in artificially modified and real-life time series using three state-of-the-art methods. *Open Geosciences*, 8(1), pp. 78–98.

YAMAMOTO, Y., (2016). A modified confidence set for the structural break date in linear regression models. *Econometric Reviews*, pp. 1–26.

ZEILEIS, F., LEISCH, K., H., KLEIBER, C., (2002). Strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software,* 7, pp. 1–38.

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 151–178, DOI 10.21307/stattrans-2020-060*
Received – 17.09.2020 ; accepted – 07.11.2020

# Comparison of selected tests for univariate normality based on measures of moments

**Czesław Domański**[1], **Piotr Szczepocki**[2]

## ABSTRACT

Univariate normality tests are typically classified into tests based on empirical distribution, moments, regression and correlation, and other. In this paper, power comparisons of nine normality tests based on measures of moments via the Monte Carlo simulations is extensively examined. The effects on power of the sample size, significance level, and on a number of alternative distributions are investigated. None of the considered tests proved uniformly most powerful for all types of alternative distributions. However, the most powerful tests for different shape departures from normality (symmetric short-tailed, symmetric long-tailed or asymmetric) are indicated.

**Key words:** normality tests, Monte Carlo simulation, power of test.

## 1. Introduction

The normality of the data assumption is one of the most commonly found in statistical studies, especially in econometric models and generally in research on applied economics. It is well known that departures from normality may lead to substantial inaccuracy of estimation procedures and incorrect inference. Popular graphical methods (Q–Q plot, histogram or box plot) are unable to provide formal conclusive evidence that the normal assumption holds. Therefore, formal statistical tests are required to conclude the normality of the data.

The problem of testing normality has gained considerable importance and has led to the development of a large number of goodness-of-fit tests to detect departures from normality. Comprehensive descriptions and power comparisons of such tests have been the focus of attention of many previous works (for the newest research see: Thadewald and Büning, 2007, Romão, Delgado and Costa, 2010, Yap and Sim, 2011, Wijekularathna, Manage and Scariano, 2019). Although the referred comparison studies have been appearing over the years, there are fewer works that compare only normality tests based on the measures of the moments. The more recent ones, Domański (2010) and Domański and Jędrzejczak (2016), do not include several interesting and more recently developed tests. This class of tests is very broad, and among other encompasses one of the most popular econometric normality test (the Jarque–Bera test) and a number of new tests based on robust estimates of the moments. Furthermore, these tests offer also clear interpretation of results, which may be very useful for users: when normality is rejected, one also obtains information on

[1]Department of Statistical Methods, The Faculty of Economics and Sociology, University of Lodz, Poland. E-mail: czedoman@uni.lodz.pl. ORCID: https://orcid.org/0000-0001-6144-6231
[2]Department of Statistical Methods, The Faculty of Economics and Sociology, University of Lodz, Poland. E-mail: piotr.szczepocki@uni.lodz.pl.ORCID: https://orcid.org/0000-0001-8377-3831.

the sample: the distribution may be skewed to the left/right and/or long (or short) tailed. A further comparison of such normality tests can, therefore, be considered to be of foremost interest.

In Section 2, we present the procedures for normality tests considered in this study. The Monte Carlo simulation methodology for comparisons of the power of the normality tests and results are discussed in Section 3. Finally, a conclusion is given in Section 4.

## 2. Tests for Normality

In this article, we assume that we have a random sample $X_1, X_2, ..., X_n$ of independently and identically distributed random variables from a continuous univariate distribution with an unknown probability density function $f(x, \theta)$, where $\theta = (\theta_1, \theta_2, ..., \theta_k)$ is a vector of real-valued parameters. We test normality of this sample by verifying a composite null hypothesis:

$$H_0 : f(x;\theta) \in N(x;\mu,\sigma)$$

against the alternative:

$$H_1 : f(x;\theta) \notin N(x;\mu,\sigma)$$

where $N(x;\mu,\sigma)$ is a class of normal distributions with mean $\mu$ standard deviation $\sigma$, and probability density function given by

$$g(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

Let a random variable $X$ be distributed with mean $\mu$ and standard deviation $\sigma$. Then, the third (skewness) and fourth (kurtosis) standardized moment central moments (provided they exist) are respectively given by:

$$\sqrt{\beta_1} = \frac{E(X-\mu)^3}{\left[E(X-\mu)^2\right]^{3/2}} = \frac{E(X-\mu)^3}{\sigma^3} \tag{1}$$

and

$$\beta_2 = \frac{E(X-\mu)^4}{\left[E(X-\mu)^2\right]^2} = \frac{E(X-\mu)^3}{\sigma^4}. \tag{2}$$

These measures of probabilistic distribution are sometimes referred to as Pearson's moment coefficient of skewness and kurtosis.

Skewness is a measure of symmetry about the mean of a probability density. Kurtosis is a measure of the peakness of a probability density. For the normal distribution $\sqrt{\beta_1} = 0$ and $\beta_2 = 3$. However, there are also non-normal distributions that are symmetric (e.g. t-Student) or have kurtosis equal to three (e.g. the Tukey distribution with parameter $\lambda = 0.135$). Furthermore, testing only skewness, when kurtosis is uncontrolled, may lead to incorrect conclusions. This is often the case of testing skewness in financial returns, for which kurtosis is significantly higher than in the case of normal distribution (Piontek, 2007).

Therefore, usually for the normality testing both skewness and kurtosis are involved. Such normality tests are often referred to as 'omnibus', because they are able to detect deviations from normality due to either skewness or kurtosis. In this study we only compare omnibus tests due to their convenience for practitioners: clear interpretation of results.

Empirical counterparts of the skewness and kurtosis are respectively given by

$$\sqrt{b_1} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S}\right)^3, \tag{3}$$

and

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S}\right)^4, \tag{4}$$

where $\bar{X} = 1/n\sum_{i=1}^{n} X_i$ is mean and $S = \sqrt{1/n\sum_{i=1}^{n}(X_i - \bar{X})^2}$ is standard devation. A number of transformations and alternative measures of skewness and kurtosis are the basis for the considered univariate normality tests presented below.

### 2.1. The D'Agostino–Pearson $K^2$ test

D'Agostino and Pearson (1973) proposed the test statistic $K^2$ that combines normalizing transformations of sample skewness and kurtosis.

The transformation of sample skewness $\sqrt{b_1}$ is based on Johnson's $S_U$ transformation (Johanson, 1949) and is given by

$$Z(\sqrt{b_1}) = \frac{\ln\left(Y/c + \sqrt{(Y/c+)^2+1}\right)}{\sqrt{\ln(w)}}, \tag{5}$$

where

$$Y = \sqrt{b_1}\sqrt{\frac{(n+1)(n+3)}{6(n-2)}}, \qquad w^2 = -1 + \sqrt{2\gamma_2 - 1},$$

$$\gamma_2 = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}, \qquad c = \sqrt{\frac{2}{(w^2-1)}}.$$

D'Agostino and Pearson (1973) gave only percentage points of the distribution of transformation of $b_2$ under normal distribution. Anscombe and Glynn (1983) proposed similar transformation for sample kurtosis $b_2$ by fitting a linear function of the reciprocal of a chi-squared variable and then using the Wilson-Hilferty transformation (Wilson and Hilferty, 1931). The transformed sample kurtosis from Anscombe and Glynn (1983) is given by

$$Z(\sqrt{b_2}) = \left[\left(1 - \frac{2}{9A}\right) - \sqrt{\frac{1-2/A}{1+y\sqrt{2/(A-4)}}}\right]\sqrt{\frac{9A}{2}}, \tag{6}$$

where

$$y = \frac{b_2 - 3(n-1)/(n+1)}{24n(n-2)(n-3)/\left[(n+1)^2(n+3)(n+5)\right]},$$

$$A = 6 + \frac{8}{\sqrt{\gamma_1}}\left(\frac{2}{\sqrt{\gamma_1}} + \sqrt{1+\frac{4}{\gamma_1}}\right),$$

$$\sqrt{\gamma_1} = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}.$$

The test statistic $K^2 = \left[Z(\sqrt{b_1})\right]^2 + \left[Z(b_2)\right]^2$ that combines D'Agostino and Pearson's transformation of sample skewness (5) and Anscombe and Glynn's transformation of sample kurtosis (6) follows approximately chi-squared distributed with two degrees of freedom as the sum of squares of two asymptotically independent standardized normals (D'Agostino, Belanger and D'Agostino, 1990).

### 2.2. The Jarque–Bera test

The Jarque–Bera test is one of the most popular goodness-of-fit test in the field of econometrics. Although it was first proposed by Bowman and Shenton (1975), it is mostly known from the work of Jarque and Bera (1987). The test Statistic $JB$ is based on sample skewness and kurtosis and is defined as

$$JB = n\left(\frac{(b_1^{1/2})^2}{6} + \frac{(b_2-3)^2}{24}\right). \tag{7}$$

This test statistic is derived from the fact that, under normality, the asymptotic means of $b_1^{1/2}$ and $b_2$ are 0 and 3, and the asymptotic variances are $6/n$ and $24/n$, and finally the asymptotic covariance is zero. Thus, $JB$ statistic is the sum of squares of two asymptotically independent standardized normals and has approximately chi-squared distribution with two degrees of freedom. However, the statistics $b_1^{1/2}$ and $b_2$ are not independently distributed and the sample kurtosis approaches normality very slowly. Thus, asymptotic critical values are strongly not recommended.

Jarque and Bera (1987) also proved that if the alternative distributions are in the Pearson family, $JB$ statistic is the corresponding Lagrange multiplier test (also known as Rao's score test) for normality.

### 2.3. The Urzùa test

Urzùa (1996) proposed a modification of the Jarque–Bera test called the adjusted Lagrange multiplier test by standardizing the sample skewness and kurtosis in the formula of $JB$ statistics in the following way

$$ALM = n\left(\frac{(b_1^{1/2})^2}{c_1} + \frac{(b_2-c_2)^2}{c_3}\right), \tag{8}$$

where

$$c_1 = \frac{6(n-2)}{(n+1)(n+3)}, \qquad c_2 = \frac{3(n-1)}{(n+1)}, \qquad c_3 = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

The idea of this modification is to use, instead of the asymptotic means and variances of the standardized third and fourth moments, their exact counterparts. On the basis of Fisher (1930) $k-$statistics, Urzùa showed that under normality, the exact mean and variance of $b_1^{1/2}$ are 0 and $c_1$, and the exact mean and variance of $b_2$ are $c_2$ and $c_3$.

On the basis of asymptotical distributions of *ALM* statistic, the hypothesis of normality is rejected at some significance level if the value of statistic exceeds critical value of a chi-squared distribution with two degrees of freedom. This modification of *JB* statistic behaves much better for small- and medium-size samples, than the original statistic when one uses asymptotical tables of critical values (Urzùa, 1996). However, in the case of Monte Carlo simulated critical values, Thadewald and Büning (2007) reported no improvement of power to the classical *JB* test.

### 2.4. The Doornik–Hansen test

Doornik and Hansen (2008) introduced another modification of the Jarque–Bera test for which the transformation creates statistics that are much closer to standard normal than in original *JB* statistic. Statistic of Doornik-Hansen test is given by

$$DH = \left[Z(\sqrt{b_1})\right]^2 + [z_2]^2, \tag{9}$$

in which they proposed to use the transformed sample skewness $Z(\sqrt{b_1})$ according to equation (5) and sample kurtosis is transformed to a chi-squared distribution with non-integer degrees of freedom, which is then translated into standard normal using the Wilson–Hilferty transformation

$$z_2 = \left[\left(\frac{\xi}{2a}\right)^{\frac{1}{3}} - 1 + \frac{1}{9a}\right]\sqrt{9a}, \tag{10}$$

where

$$\xi = (b_2 - 1 - b_1)2k, \qquad k = \frac{(n+5)(n+7)(n^3+37n^2+11n-313)}{12(n-3)(n+1)(n^2+15n-4)},$$

$$a = \frac{(n+5)(n+7)\left[(n^2+27n-70)+b_1(n-7)(n^2+2n-5)\right]}{6(n-3)(n+1)(n^2+15n-4)}.$$

The formulae (10) break down for $n \leqslant 7$. The DH statistic is also approximately chi-squared distribution with two degrees of freedom. However, because of its fast coverage, *DH* statistic does not require simulated quantiles of distribution under null hypothesis (Doornik and Hansen, 2008).

### 2.5. The Gel–Gastwirth test

Gel and Gastwirth (2008) proposed a modification of *JB* that uses a robust estimate of the dispersion, the average absolute deviation from the sample median (MAAD), instead of the second order central moment $m_2$. MAAD is defined by

$$\mathrm{MAAD} = \frac{\sqrt{\pi/2}}{n} \sum_{i=1}^{n} |X_i - \mathrm{med}_F| , \tag{11}$$

where $\mathrm{med}_F$ is the sample median. Robust dispersion measure is used, due to the fact that sample moments are known to be sensitive to outliers, and the sample variance is even more affected by outliers than the mean (Gel and Gastwirth, 2008). Thus, *RJB* statistic performs better than *JB* statistics in the case of long-tailed distributions (Gel and Gastwirth, 2008). However, in the case of short-tailed distribution robust measures of the dispersion may not be necessary.

The test statistic is given by

$$RJB = \frac{n}{6} \left( \frac{m_3}{\mathrm{MAAD}^3} \right)^2 + \frac{n}{64} \left( \frac{m_4}{\mathrm{MAAD}^4} - 3 \right)^2 . \tag{12}$$

Gel and Gastwirth (2008) also proved that under the null hypothesis of normality, the *RJB* test statistic asymptotically follows the chi-square distribution with two degrees of freedom. However, similarly to *JB*, for small and moderate samples the Monte Carlo simulated critical values are more preferable than asymptotic chi-squared distribution values.

### 2.6. The Bontemps-Meddahi tests

Bontemps and Meddahi (2005) proposed a family of normality tests developed on the basis of generalized method of moments approach and Hermite polynomials. The family of test statistics is given by

$$BM_{3-\rho} = \sum_{k=3}^{\rho} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} H_k \left( \frac{x_i - \bar{x}}{s} \right) \right)^2 , \tag{13}$$

where $H_k$ is the $k-$th order normalized Hermite polynomial. The considered moment conditions in the Bontemps-Meddahi tests are based on the Stein equation (Stein, 1972). The important property of the Stein equation is that, the expectation of the considered function is zero by construction. Bontemps and Meddahi (2005) showed that special examples of this equation correspond to the zero mean of any Hermite polynomial. The family of the Bontemps-Meddahi tests asymptotically follows the chi-square distribution with $\rho - 2$ degrees of freedom. The *JB* statistic almost coincides with $BM_{3-4}$. The only difference is that in *JB* test, the variance is estimated by $S = 1/n \sum_{i=1}^{n} (X_i - \bar{X})^2$ while in the Hermite case it is estimated by $1/(n-1) \sum_{i=1}^{n} (X_i - \bar{X})^2$. In the presented study we use the Bontemps-Meddahi test termed $BM_{3-6}$, because tests based on Hermite polynomials of degree at least seven do not provide gain in power (Bontemps and Meddahi, 2005).

### 2.7. The Hosking test

Hosking (1990) proposed to use L-moments, linear combinations of the order statistics, instead of classic central moments in order to obtain more powerful test in case of long-tailed distributions. L-moments are less affected by sample variability, and thus more robust to outliers.

Based on the second, third and fourth sample L-moments, which correspond to the second, third and fourth central moments, Hosking (1990) introduced new measures of skewness and kurtosis, termed L-skewness $\tau_3$ and L-kurtosis $\tau_4$ defined as

$$\tau_3 = \frac{l_3}{l_2}, \qquad \tau_4 = \frac{l_4}{l_2} \tag{14}$$

where $l_r$ are order sample L-moment that can be estimated by

$$l_r = \sum_{k=0}^{r-1} p^*_{r-1,k} b_k, \tag{15}$$

where

$$p^*_{r-1,k} = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k}, \qquad b_k = \frac{1}{n} \sum_{i=1}^{n} \frac{(i-1)(i-2)\cdots(i-k)}{(n-1)(n-2)\cdots(n-k)}.$$

Hosking (1990) proposed to test normality by the following statistic

$$T_{Lmom} = \frac{\tau_3 - \mu_{\tau_3}}{\mathrm{var}(\tau_3)} + \frac{\tau_4 - \mu_{\tau_4}}{\mathrm{var}(\tau_4)}, \tag{16}$$

where values of means $(\mu_{\tau_3}, \mu_{\tau_4})$ and variances $(\mathrm{var}(\tau_3), \mathrm{var}(\tau_4))$ of L-skewness $\tau_3$ and L-kurtosis $\tau_4$ may be obtained by simulation. The $T_{Lmom}$ is approximately chi-squared distribution with two degrees of freedom.

### 2.8. The Brys-Hubert-Struyf & Bonett-Seier test

The Brys-Hubert-Struyf & Bonett-Seier test $T_{MC-LR} - T_w$ is omnibus test for normality proposed in (Romão, Delgado and Costa, 2010) as combination of two tests: the Bonett-Seier test (Bonett and Seier, 2002) and the Brys–Hubert–Struyf *MC˘LR* test (Brys, Hubert and Struyf, 2007). The former is a kurtosis associated test, the latter is a skewness-based test. The statistic of the Bonett-Seier test is defined as

$$T_w = \frac{(\hat{\omega} - 3)\sqrt{n+2}}{3.54}, \tag{17}$$

where

$$\hat{\omega} = 13.29 \left[ \ln \sqrt{m_2} - \ln \left( \frac{1}{n} \sum_{i=1}^{n} |X_i - \bar{X}| \right) \right],$$

in which $m_2$ is a sample second central moment. Statistics $T_w$ approximately follows a standard normal distribution, and consequently null hypothesis is rejected for both small and large values of $T_w$. The Bonett-Seier statistic is a simple transformation of Geary's measure of kurtosis (Geary, 1936) , which is defined as $\tau/\sigma$, where $\tau = E(|X-\mu|)$. After transformation (17), like its Pearson's counterpart (given by equation (4)), Geary's measure of kurtosis equals 3 under normality and increases without bound with increasing leptokurtosis.

The Brys–Hubert–Struyf MC–LR $T_{MC-LR}$ test is given by

$$T_{MC-LR} = n(v-\mathrm{v})'V^{-1}(v-\mathrm{v}), \tag{18}$$

where $v$ is the vector of robust measures of skewness $[MC, LMC, RMC]'$, and $\nu$, $V$ are estimates based on the distribution under null hypothesis. In the case of normal distribution $\nu$, $V$ are given by

$$\nu = [0, 0.199, 0.199]', \qquad V = \begin{bmatrix} 1.25 & 0.323 & -0.323 \\ 0.323 & 2.62 & -0.0123 \\ -0.323 & -0.0123 & 2.62 \end{bmatrix}.$$

$T_{MC-LR}$ statistic approximately follows the chi-square distribution with three degrees of freedom.

The first element of vector of $v$ is medcouple, proposed in Brys et al. (2004), defined as

$$MC = \underset{X_{(i)} \le \mathrm{med}_F \le X_{(j)}}{\mathrm{med}} h\left(X_{(i)}, X_{(j)}\right), \tag{19}$$

where med is the median, $h$ is the kernel function given by

$$h\left(X_{(i)}, X_{(j)}\right) = \frac{\left(X_{(j)} - \mathrm{med}_F\right) - \left(\mathrm{med}_F - X_{(i)}\right)}{X_{(i)} - X_{(j)}.}$$

Medcouple is a robust skewness measure bounded by $[-1, 1]$.

The two other elements of vector $v$ are the left medcouple (LMC) and the right medcouple (RMC), the left and right tail weight measure, proposed in Brys et al. (2006). LMC and RMC are respectively defined as

$$LMC = -MC(x < \mathrm{med}_F), \qquad RMC = MC(x > \mathrm{med}_F). \tag{20}$$

Like medcouple, they are robust against outlying values. These three measures have great advantage that can be computed at any distribution, even when finite moments do not exist (Brys, Hubert and Struyf, 2007).

The joint test $T_{MC-LR} - T_w$ proposed by Romão, Delgado and Costa (2010) is based on the assumption that individual tests can be considered independent. This assumption was positively verified in simulation study of 200,000 samples of size 100 drawn from a standard normal distribution. In order to control the overall type I error at the nominal level $\alpha$, the normality hypothesis of the data is rejected for the joint test when rejection is obtained for

either one of the two individual tests for a significance level of $\alpha/2$ (Romão, Delgado and Costa, 2010).

### 2.9. Desgagnéa and Lafaye de Micheaux test

Desgagnéa and Lafaye de Micheaux (2018) has recently proposed new alternatives to the classical Pearson's measures of skewness and kurtosis, which they termed 2nd-power skewness and kurtosis. They used them to build two tests of normality. First test $X^a_{APD}$ can be derived as the Lagrange multiplier test on the asymmetric power distribution (APD) class, introduced by Komunjer (2007). This class of distribution is a generalization of the generalized power distribution (GPD) (also known as the generalized error distribution (GED)), which is symmetric, to a broader class that includes asymmetric distributions. The APD class encompasses all GPD distributions (i.e. the Laplace distribution, normal distribution) and asymmetric distributions (i.e. asymmetric Laplace distribution, split normal distribution).

The basis of this test are 2nd-power skewness $B_2$ and 2nd-power kurtosis $K_2$, which are defined as

$$B_2 = \frac{1}{n}\sum_{i=1}^{n} Z_i^2 \operatorname{sign}(Z_i), \qquad \text{and} \qquad K_2 = \frac{1}{n}\sum_{i=1}^{n} Z_i^2 \ln(|Z_i|), \tag{21}$$

where $Z_i = (X_i - \bar{X})/S$. This sample statistics are analogous to 2nd-power skewness and kurtosis for a random variable X, which are defined as $E(Z^2\operatorname{sign}(Z))$ and $E(Z^2\ln(Z))$, respectively.

The $X^a_{APD}$ statistics is defined as

$$X^a_{APD} = \frac{nB_2^2}{3-8/\pi} + \frac{n\left(K_2 - (2-\ln 2-\gamma)/2\right)^2}{(3\pi^2-28)/8}, \tag{22}$$

where $\gamma$ is the Euler–Mascheroni constant. The $X^a_{APD}$ is approximately chi-squared distribution with two degrees of freedom as a sum of squares of two independent standard normals. However, $X^a_{APD}$ has rather poor small sample properties (just as $JB$ statistic). Thus, Desgagnéa and Lafaye de Micheaux (2018) proposed the second statistic $X_{APD}$ defined as

$$X_{APD} = Z^2(B_2) + Z^2(K_2 - B_2), \tag{23}$$

where

$$Z(B_2) = \sqrt{\frac{nB_2^2}{(3-8/\pi)(1-1.9/n)}}$$

is transformed 2nd-power skewness, and

$$Z(K_2 - B_2) = \frac{\sqrt{n}\left[(K_2 - B_2^2)^{1/3} - ((2-\ln 2-\gamma)/2)^{1/3}(1-1.026/n)\right]}{\sqrt{((2-\ln 2-\gamma)/2)^{-4/3}(3\pi^2-28)(1-2.25/n^{0.8})/72}}$$

is transformed 2nd-power net kurtosis. Under the null hypothesis $X_{APD}$ is, with high numerical precision, approximately distributed as chi-squared distribution with two degrees of

freedom, for all sample sizes with at least 10 observations. This is a rare and desirable characteristic for normality test statistic based on measures of the moments. In the simulation study we use only $X_{APD}$ statistics.

## 3. Simulation study

In our simulation study we considered three levels of significance: $\alpha = 0.01$, 0.05 and 0.10, and five different sample sizes: $n$ = 10, 20, 50, 100, 500. First, appropriate critical values were obtained for each test based on 100,000 simulated samples from a standard normal distribution. We decided to use empirical rather than approximated limit distributions, because many preavious studies emphasized that in the case of Jarque-Berra test and their modifications chi-squared distribution approximation of the limit distribution did not work well, even for large sample sizes (Thadewald and Büning, 2007 and Romão, Delgado and Costa, 2010).

In order to investigate the power of the various tests a total of 10,000 samples of the appropriate size were drawn from each of 15 different non-normal distributions. These distributions are categorized as symmetric short-tailed, symmetric long-tailed and asymmetric in shape (the same categories were considered by Farrell and Rogers-Stewart, 2006). The choice of shape category is based on the values of Pearson's measures of the skewness and kurtosis of the distribution given by the formulas (1) and (2). Specifically, asymmetric distributions have $\sqrt{\beta_1} \neq 0$, symmetric short-tailed $\sqrt{\beta_1} = 0$ and $\beta_2 < 3$ and symmetric long-tailed $\sqrt{\beta_1} = 0$ and $\beta_2 > 3$.

Tables 1, 2, 3 presents results for the first category of alternative distributions, namely symmetrical short-tailed distributions, respectively for three levels of significance: $\alpha = 0.01$, 0.05 and 0.10. Distributions are ordered from the distribution with the lowest kurtosis (the most distinct from normal), to the distribution with the highest kurtosis (the closest to normal). The average power across all short-tailed distributions is presented in Table 4. Firstly, power of normality test for this group of distributions is not sufficient. Especially, at significance level $\alpha = 0.01$ and small samples sizes (below 50), all considered tests perform very poorly. When the significance level and/or sample size increase tests become more powerful. For the smallest sample sizes $X_{APD}$ statistics seems to perform best for most alternative distributions. For moderate and big samples $K^2$ achieves good power for alternative distributions with low kurtosis, but for distributions with kurtosis more close to normal $T_{MC-LR} - T_w$ statistics performs even better. On the basis of average results, $T_{Lmom}$ tests perform fairly well for moderate and big samples, too. The results show that for symmetrical short-tailed distributions the popular $JB$ statistic performs poorly. From modifications of this statistics DH seems to be the best. It performs quite well for all sample sizes.

Table 1: Empirical power results for symmetrical short-tailed distributions ($\alpha = 0.01$).

| Alternative | $n$ | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K^2$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR} - T_w$ | $X_{APD}$ |
| Uniform (a=0,b=1) | 10 | 0.24 | 0.24 | 0.3 | 1.72 | 0.21 | 0.26 | 0.35 | 0.67 | **3.44** |
| $\sqrt{\beta_1} = 0, \beta_2 = 1.8$ | 20 | 1.32 | 0 | 0 | 1.44 | 0 | 0.03 | 2.81 | 1.02 | **5.62** |
| | 50 | **47.5** | 0 | 0 | 8.8 | 0 | 0 | 41.96 | 16.57 | 35.26 |
| | 100 | **96.72** | 0 | 0 | 62.14 | 0 | 1 | 90.25 | 55.82 | 84.82 |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Tukey (l=0.8) | 10 | 0.33 | 0.33 | 0.3 | 1.97 | 0.29 | 0.33 | 0.35 | 0.74 | **3.46** |
| $\sqrt{\beta_1} = 0, \beta_2 = 1.86$ | 20 | 1.44 | 0.01 | 0 | 1.52 | 0 | 0 | 2.81 | 0.97 | **6.1** |
| | 50 | **46.94** | 0 | 0 | 8.41 | 0 | 0 | 41.01 | 16.25 | 33.71 |
| | 100 | **97.03** | 0 | 0 | 61.76 | 0 | 1.29 | 90.47 | 56.12 | 84.77 |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Tukey (l=0.3) | 10 | 0.39 | 0.4 | 0.51 | 0.73 | 0.37 | 0.39 | 0.47 | 0.92 | **0.99** |
| $\sqrt{\beta_1} = 0, \beta_2 = 2.41$ | 20 | 0.13 | 0.08 | 0 | 0.3 | 0.07 | 0.09 | 0.21 | 0.7 | **0.58** |
| | 50 | 1.07 | 0 | 0 | 0.13 | 0 | 0 | 0.97 | **1.28** | 1.14 |
| | 100 | **4.46** | 0 | 0 | 0.33 | 0 | 0 | 3.05 | 2.25 | 2.86 |
| | 500 | **78.89** | 10.99 | 8.8 | 43.61 | 2.33 | 0.52 | 0.34 | 28.93 | 57.35 |
| Beta (a=5,b=5) | 10 | 0.51 | 0.51 | 0.5 | 0.72 | 0.51 | 0.5 | 0.6 | 0.88 | **0.91** |
| $\sqrt{\beta_1} = 0, \beta_2 = 2.53$ | 20 | 0.14 | 0.13 | 0.4 | 0.29 | 0.13 | 0.16 | 0.28 | **0.79** | 0.42 |
| | 50 | 0.37 | 0 | 0 | 0.11 | 0.01 | 0 | 0.59 | **0.9** | 0.64 |
| | 100 | **1.97** | 0 | 0 | 0.26 | 0 | 0 | 1.72 | 1.41 | 1.54 |
| | 500 | **34.41** | 1.25 | 0.9 | 10.86 | 0.18 | 0.02 | 100 | 10.96 | 22.58 |
| Beta (a=10,b=10) | 10 | 0.7 | 0.68 | 0.7 | 0.87 | 0.72 | 0.72 | 0.7 | **0.97** | 0.85 |
| $\sqrt{\beta_1} = 0, \beta_2 = 2.73$ | 20 | 0.29 | 0.3 | 1 | 0.47 | 0.34 | 0.3 | 0.49 | **0.71** | 0.49 |
| | 50 | 0.4 | 0.18 | 0 | 0.39 | 0.2 | 0.19 | 0.62 | **0.75** | 0.52 |
| | 100 | 0.65 | 0.06 | 0 | 0.25 | 0.08 | 0.09 | 0.71 | **0.88** | 0.61 |
| | 500 | **3.66** | 0.12 | 0.1 | 1 | 0.03 | 0 | 100 | 2.39 | 3.14 |

Table 2: Empirical power results for symmetrical short-tailed distributions ($\alpha = 0.05$).

| Alternative | n | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K^2$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR} - T_w$ | $X_{APD}$ |
| Uniform (a=0,b=1) | 10 | 2.43 | 1.86 | 1.6 | 6.58 | 1.67 | 2.33 | 5.7 | 4.48 | **9.75** |
| $\sqrt{\beta_1} = 0, \beta_2 = 1.8$ | 20 | 13.3 | 0.21 | 0.1 | 10.08 | 0.18 | 3.65 | **20.18** | 8.45 | 19.58 |
| | 50 | **77.79** | 1.16 | 0 | 45.96 | 0.02 | 44.53 | 70.41 | 38.82 | 62.68 |
| | 100 | **99.68** | 75.63 | 53.6 | 95.32 | 1.14 | 93.76 | 97.76 | 80.01 | 96.52 |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Tukey (l=0.8) | 10 | 1.86 | 1.56 | 1.6 | 6.45 | 1.56 | 2.27 | 5.45 | 4.45 | **9.53** |
| $\sqrt{\beta_1} = 0, \beta_2 = 1.86$ | 20 | 12.85 | 0.25 | 0.1 | 9.33 | 0.24 | 3.43 | **19.11** | 7.95 | 18.44 |
| | 50 | **78.32** | 1.06 | 0 | 45.74 | 0.03 | 44.21 | 70.17 | 38.43 | 62.45 |
| | 100 | **99.76** | 75.16 | 53.6 | 95.49 | 1.04 | 93.89 | 97.75 | 79.62 | 96.53 |
| | 500 | 99.99 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Tukey (l=0.3) | 10 | 2.68 | 2.73 | 2.4 | 3.46 | 2.74 | 2.97 | 3.32 | **4.39** | 4.26 |
| $\sqrt{\beta_1} = 0, \beta_2 = 2.41$ | 20 | 2.01 | 1.01 | 1 | 1.77 | 1.12 | 1.36 | 2.9 | **4.21** | 3.6 |
| | 50 | **6.4** | 0.17 | 0.2 | 1.94 | 0.18 | 1.44 | 6.11 | 6.19 | 6.25 |
| | 100 | **17.34** | 0.46 | 0 | 5.17 | 0.03 | 3.87 | 13.23 | 10.24 | 13.1 |
| | 500 | **94.85** | 80.93 | 76.7 | 84.53 | 59.18 | 62.26 | 22.64 | 54.74 | 82.1 |
| Beta (a=5,b=5) | 10 | 3.48 | 3.39 | 3.2 | 3.52 | 3.26 | 3.3 | 3.8 | **4.68** | 4.31 |
| $\sqrt{\beta_1} = 0, \beta_2 = 2.53$ | 20 | 2.18 | 1.55 | 1.9 | 2.04 | 1.63 | 1.97 | 3.3 | **4.35** | 3.41 |
| | 50 | 3.8 | 0.48 | 0.2 | 1.54 | 0.5 | 1.16 | 4.01 | **4.9** | 3.71 |
| | 100 | **8.77** | 0.32 | 0.36 | 2.87 | 0.18 | 1.66 | 7.86 | 7.18 | 7.69 |
| | 500 | 65.12 | 38.12 | 33.6 | 44.66 | 21.69 | 22.26 | **100** | 28.03 | 48.96 |
| Beta (a=10,b=10) | 10 | 3.96 | 3.95 | 3.9 | 4.56 | 4.23 | 4.07 | 4.45 | **4.77** | 4.72 |
| $\sqrt{\beta_1} = 0, \beta_2 = 2.73$ | 20 | 3.08 | 2.84 | 3 | 3.39 | 2.82 | 3.23 | 3.84 | **4.71** | 3.93 |
| | 50 | 3.37 | 1.73 | 1.7 | 2.62 | 1.56 | 2.33 | 4.22 | **4.99** | 3.87 |
| | 100 | 3.82 | 0.91 | 0.9 | 2.33 | 0.88 | 1.57 | 4.52 | **5.36** | 4.16 |
| | 500 | 15.67 | 6 | 4.3 | 8.68 | 2.99 | 2.6 | **100** | 8.44 | 12.76 |

Table 3: Empirical power results for symmetrical short-tailed distributions ($\alpha = 0.1$).

| Alternative | $n$ | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K^2$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR}-T_w$ | $X_{APD}$ |
| Uniform (a=0,b=1) | 10 | 8.02 | 4.71 | 4 | 11.45 | 3.61 | 8.05 | 13.71 | 8.91 | **15.01** |
| $\sqrt{\beta_1}=0, \beta_2=1.8$ | 20 | 27.79 | 2.97 | 0.4 | 20.8 | 0.76 | 22.29 | **34.08** | 16.87 | 30.53 |
| | 50 | **88.96** | 50.18 | 26.6 | 68.86 | 0.1 | 75.85 | 82.45 | 54.33 | 76.84 |
| | 100 | **99.93** | 98.66 | 97.1 | 98.96 | 74.52 | 99.11 | 99.15 | 88.59 | 98.67 |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Tukey (l=0.8) | 10 | 8.36 | 4.98 | 4 | 11.88 | 3.94 | 8.5 | 14.13 | 9.59 | **15.8** |
| $\sqrt{\beta_1}=0, \beta_2=1.86$ | 20 | 27.67 | 3.45 | 0.4 | 21.35 | 0.85 | 22.75 | **34.44** | 17.02 | 30.69 |
| | 50 | **88.64** | 50.27 | 56.11 | 68.72 | 0.17 | 74.76 | 81.96 | 53.54 | 76.19 |
| | 100 | **99.88** | 98.46 | 97.1 | 98.77 | 73.63 | 98.85 | 98.87 | 87.71 | 98.4 |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Tukey (l=0.3) | 10 | 6.46 | 5.88 | 6.1 | 7.07 | 5.95 | 6.64 | 7.81 | **9.61** | 8.26 |
| $\sqrt{\beta_1}=0, \beta_2=2.41$ | 20 | 6.06 | 3.41 | 3.2 | 5.08 | 3.02 | 5.97 | 8.09 | **9.31** | 8 |
| | 50 | **13.07** | 2.15 | 1.4 | 6.34 | 0.83 | 7.92 | 12.62 | 12.24 | 11.92 |
| | 100 | 30.21 | 10.37 | 5.7 | 14.86 | 1.27 | 16.36 | **24.41** | 17.56 | 24.39 |
| | 500 | **98.02** | 93.97 | 92.4 | 93.83 | 83.65 | 85.09 | 86.61 | 68.58 | 90.22 |
| Beta (a=5,b=5) | 10 | 7.21 | 6.81 | 7.3 | 8.02 | 6.91 | 7.24 | 8.17 | **8.87** | 8.82 |
| $\sqrt{\beta_1}=0, \beta_2=2.53$ | 20 | 6.29 | 4.73 | 3.9 | 5.81 | 4.29 | 6.66 | 7.8 | **9.26** | 7.78 |
| | 50 | 8.43 | 2.84 | 1.5 | 5.13 | 1.7 | 6.01 | 9.82 | **10.72** | 8.95 |
| | 100 | **16.64** | 5.7 | 2.6 | 8.52 | 1.37 | 9.01 | 15.22 | 13.03 | 14.44 |
| | 500 | 78.06 | 63.62 | 60.5 | 63.89 | 48.5 | 48.3 | **100** | 40.19 | 63.25 |
| Beta (a=10,b=10) | 10 | 8.44 | 8.54 | 7.8 | 8.94 | 8.4 | 8.94 | 8.96 | **9.99** | 9.21 |
| $\sqrt{\beta_1}=0, \beta_2=2.73$ | 20 | 7.86 | 7.28 | 6.7 | 7.55 | 6.98 | 7.92 | 8.45 | **9.46** | 8.66 |
| | 50 | 7.3 | 4.97 | 4.7 | 5.84 | 4.79 | 6.23 | 8.67 | **10** | 7.76 |
| | 100 | 9.18 | 4.93 | 3.6 | 6.13 | 3.35 | 6.19 | 10.16 | **10.6** | 9.33 |
| | 500 | **26.61** | 17.69 | 14.2 | 19.18 | 12.41 | 11.39 | 11.0 | 16.08 | 22.92 |

Table 4: Average Power for symmetrical short-tailed distributions.

| $\alpha$ | $n$ | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K^2$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR}-T_w$ | $X_{APD}$ |
| 0.01 | 10 | 0.43 | 0.43 | 0.46 | 1.20 | 0.42 | 0.44 | 0.49 | 0.84 | **1.93** |
| | 20 | 0.66 | 0.10 | 0.28 | 0.80 | 0.11 | 0.12 | 1.32 | 0.84 | **2.64** |
| | 50 | **19.26** | 0.04 | 0.00 | 3.57 | 0.04 | 0.04 | 17.03 | 7.15 | 14.25 |
| | 100 | **40.17** | 0.01 | 0.00 | 24.95 | 0.02 | 0.48 | 37.24 | 23.30 | 34.92 |
| | 500 | 63.39 | 42.47 | 41.96 | 51.09 | 40.51 | 40.11 | **80.07** | 48.46 | 56.61 |
| 0.05 | 10 | 2.88 | 2.70 | 2.54 | 4.91 | 2.69 | 2.99 | 4.54 | 4.55 | **6.51** |
| | 20 | 6.68 | 1.17 | 1.22 | 5.32 | 1.20 | 2.73 | **9.87** | 5.93 | 9.79 |
| | 50 | **33.94** | 0.92 | 0.42 | 19.56 | 0.46 | 18.73 | 30.98 | 18.67 | 27.79 |
| | 100 | **45.87** | 30.50 | 21.69 | 40.24 | 0.65 | 38.95 | 44.22 | 36.48 | 43.60 |
| | 500 | 75.13 | 65.01 | 62.92 | 67.57 | 56.77 | 57.42 | **84.53** | 58.24 | 68.76 |
| 0.1 | 10 | 7.70 | 6.18 | 5.84 | 9.47 | 5.76 | 7.87 | 10.56 | 9.39 | **11.42** |
| | 20 | 15.13 | 4.37 | 2.92 | 12.12 | 3.18 | 13.12 | **18.57** | 12.38 | 17.13 |
| | 50 | **41.28** | 22.08 | 18.06 | 30.98 | 1.52 | 34.15 | 39.10 | 28.17 | 36.33 |
| | 100 | **51.17** | 43.62 | 41.22 | 45.45 | 30.83 | 45.90 | 49.56 | 43.50 | 49.05 |
| | 500 | **80.54** | 75.06 | 73.42 | 75.38 | 68.91 | 68.96 | 79.52 | 64.97 | 75.28 |

Results for the second category of alternative distributions, symmetrical long-tailed distributions, are presented in Tables 5, 6, 7. Distributions are ordered from the distributions with the highest kurtosis (the most distinct from normal), to the distribution with the lowest kurtosis (the closest to normal). The average power across all long-tailed distributions is presented in Table 8. For this group of distribution, normality tests perform better than for the short-tailed distributions, but for small sample size results are still not very impressive. On the basis of average results, the *RJB* statistic outperforms other tests for almost all sample sizes. It is not surprisingly, bearing in mind that this test is based on the robust estimate of the dispersion. However, when one takes a closer look at particular alternative distributions, one may see that for the distribution with kurtosis closer to three also D'Agostino–Pearson $K^2$ test performs well. Contrary to the short-tailed distribution, *JB* statistic has quite good power properties, even better than its modifications (apart from *RJB*).

Table 5: Empirical power results for symmetrical long-tailed distributions ($\alpha = 0.01$).

| Alternative | $n$ | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K^2$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR}-T_w$ | $X_{APD}$ |
| Laplace (mu=0,b=1) | 10 | 6.78 | 6.75 | 7.25 | 6.07 | 7.63 | 7.02 | **7.89** | 3.64 | 5.17 |
| $\sqrt{\beta_1}=0, \beta_2=6$ | 20 | 13.68 | 14.37 | 14.59 | 16.18 | **18.09** | 16.15 | 17.31 | 11.2 | 15.83 |
| | 50 | 28.58 | 33.84 | 30.55 | 34.76 | **43.96** | 33.56 | 43.08 | 35.39 | 41.95 |
| | 100 | 48.96 | 60.25 | 51.85 | 62.62 | 74.35 | 54.43 | **74.61** | 69.81 | 73.89 |
| | 500 | 99.82 | 99.96 | 99.06 | 99.96 | **100** | 99.76 | 5.43 | **100** | **100** |
| Logistic (mu=0,sigma=1) | 10 | 2.6 | 2.6 | 2.27 | 2.14 | 2.66 | 2.65 | **2.78** | 1.62 | 1.93 |
| $\sqrt{\beta_1}=0, \beta_2=4.4$ | 20 | 5.19 | 5.37 | 5.1 | 5.27 | 5.46 | **5.51** | 4.58 | 3.01 | 4.6 |
| | 50 | 10.06 | **11.71** | 9.18 | 11.7 | 12.83 | 11.38 | 8.4 | 5.96 | 11.4 |
| | 100 | 16.54 | 21.12 | 20.87 | 21.61 | **23.49** | 18.67 | 14.17 | 11.31 | 20.19 |
| | 500 | 67.32 | 76.3 | 63.96 | 76.75 | **80.31** | 57.87 | 2.04 | 60.4 | 77.76 |
| Student-t (df=10) | 10 | 2.05 | 2.02 | 1.1 | 1.65 | 2.07 | 2.02 | **2.1** | 1.33 | 1.39 |
| $\sqrt{\beta_1}=0, \beta_2=4$ | 20 | 4.08 | 4.14 | 3.25 | 4.03 | **4.25** | 4.22 | 3.38 | 2.59 | 3.7 |
| | 50 | 7.92 | 9.09 | 6.89 | 8.8 | **9.55** | 9.03 | 5.59 | 3.88 | 7.99 |
| | 100 | 13.27 | 16.22 | 15.94 | 15.88 | **16.65** | 14.63 | 8.28 | 6.93 | 13.79 |
| | 500 | 49.63 | 58.58 | 48.09 | 58.55 | **60.22** | 43.6 | 1.85 | 32.93 | 53.43 |
| Student-t (df=20) | 10 | 1.26 | 1.26 | 1.19 | 1.07 | 1.4 | 1.3 | **1.46** | 1.09 | 1.06 |
| $\sqrt{\beta_1}=0, \beta_2=3.375$ | 20 | 2.23 | 2.18 | 1.84 | 2.24 | **2.41** | 2.32 | 2.11 | 1.6 | 2.05 |
| | 50 | 3.4 | 3.92 | 2.88 | 3.41 | 3.84 | **3.95** | 2.36 | 1.68 | 3.1 |
| | 100 | 4.76 | 5.48 | 4.11 | 5.08 | **5.52** | 5.3 | 2.8 | 1.97 | 4.21 |
| | 500 | 12.99 | 16.89 | 16.13 | 16.58 | **17** | 12.51 | 1.33 | 5.35 | 11.93 |
| Student-t (df=30) | 10 | 1.19 | 1.2 | 1.1 | 1.18 | 1.22 | 1.19 | **1.39** | 1.11 | 1.06 |
| $\sqrt{\beta_1}=0, \beta_2=3.23$ | 20 | 1.59 | 1.56 | 1.49 | **1.68** | 1.66 | 1.57 | 1.38 | 1.33 | 1.5 |
| | 50 | 2.36 | 2.58 | 2.02 | 2.42 | 2.5 | **2.59** | 1.86 | 1.42 | 2.1 |
| | 100 | 3.09 | **3.52** | 2.79 | 3.07 | 3.49 | 3.27 | 1.91 | 1.58 | 2.74 |
| | 500 | 5.94 | 8.01 | 7.9 | 7.43 | **8.02** | 6.3 | 1.16 | 2.5 | 5.45 |

Table 6: Empirical power results for symmetrical long-tailed distributions ($\alpha = 0.05$).

| Alternative | $n$ | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K^2$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR}-T_w$ | $X_{APD}$ |
| Laplace (mu=0,b=1) | 10 | 17.78 | 17.41 | 17.88 | 18.65 | **20.31** | 17.84 | 19.59 | 10.13 | 16.81 |
| $\sqrt{\beta_1}=0, \beta_2=6$ | 20 | 28.6 | 30.02 | 29.81 | 31.74 | **35.6** | 32.65 | 33.11 | 20.29 | 31.12 |
| | 50 | 49.1 | 55.81 | 52.95 | 56.81 | **65.67** | 59.69 | 61.34 | 50.92 | 61.1 |
| | 100 | 72.44 | 79.84 | 79.92 | 80.74 | **89.12** | 84.12 | 87.6 | 82.73 | 87.09 |
| | 500 | 99.98 | 99.98 | **100** | 99.98 | **100** | **100** | **100** | **100** | **100** |
| Logistic (mu=0,sigma=1) | 10 | 9.61 | 9.51 | 7.26 | 8.47 | **9.79** | 9.45 | 8.85 | 6.49 | 7.81 |
| $\sqrt{\beta_1}=0, \beta_2=4.4$ | 20 | 13.77 | 14.55 | 12.27 | 14.02 | **14.94** | 14.17 | 12.29 | 8.17 | 12.42 |
| | 50 | 22.82 | 26.48 | 20.97 | 25.76 | **27.79** | 26.36 | 19.86 | 13.78 | 23.38 |
| | 100 | 33.51 | 39.63 | 31.98 | 38.41 | **42.39** | 39.37 | 29.41 | 22.16 | 36.32 |
| | 500 | 84.56 | 89.04 | 83.77 | 88.88 | **91.41** | 87.18 | 79.9 | 75.51 | 88.98 |
| Student-t (df=10) | 10 | **8.84** | 8.69 | 7.34 | 8.13 | **8.84** | 8.77 | 8.21 | 5.73 | 7.51 |
| $\sqrt{\beta_1}=0, \beta_2=4$ | 20 | 11.67 | 12.05 | 9.78 | 11.74 | **12.37** | 11.73 | 9.89 | 7.14 | 10.46 |
| | 50 | 18.07 | **20.86** | 15.14 | 19.71 | 21.3 | 20.4 | 14.5 | 10.65 | 17.35 |
| | 100 | 25.87 | 30.47 | 22.96 | 29.17 | **32.09** | 29.82 | 19.94 | 15.39 | 25.94 |
| | 500 | 68.91 | 75.09 | 67.18 | 74.33 | **76.3** | 71.74 | 71.7 | 49.77 | 70.31 |
| Student-t (df=20) | 10 | **6.79** | 6.76 | 6.21 | 6.42 | 6.75 | 6.6 | 6.44 | 5.33 | 6.27 |
| $\sqrt{\beta_1}=0, \beta_2=3.375$ | 20 | 7.53 | 7.68 | 8.78 | 7.69 | **7.8** | 7.57 | 6.96 | 5.45 | 7.2 |
| | 50 | 9.76 | 11.03 | 10.85 | 10.81 | **11.37** | 10.84 | 8.25 | 6.89 | 9.46 |
| | 100 | 12.82 | 15.3 | 14.34 | 14.27 | **15.69** | 14.93 | 9.84 | 7.81 | 12.22 |
| | 500 | 27.59 | 33.33 | 33.11 | 31.91 | **33.64** | 31.11 | 6.14 | 14.3 | 26.19 |
| Student-t (df=30) | 10 | **6.06** | 6.01 | 5.52 | 5.47 | 5.78 | 5.89 | 5.56 | 5.19 | 5.47 |
| $\sqrt{\beta_1}=0, \beta_2=3.23$ | 20 | **6.87** | 6.84 | 6.66 | 6.75 | 6.64 | 6.65 | 6.38 | 5.89 | 6.53 |
| | 50 | 8.45 | **9.11** | 7.45 | 8.36 | 8.97 | 8.81 | 6.87 | 5.8 | 7.78 |
| | 100 | 9.83 | **11.22** | 10.78 | 10.52 | 11.37 | 11.01 | 7.55 | 6.67 | 9.01 |
| | 500 | 16.56 | 20.51 | 19.87 | 19.05 | **20.52** | 19.77 | 5.41 | 8.76 | 15.39 |

Table 7: Empirical power results for symmetrical long-tailed distributions ($\alpha = 0.1$).

| Alternative | $n$ | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K^2$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR}-T_w$ | $X_{APD}$ |
| Laplace (mu=0,b=1) | 10 | 26.27 | 26.68 | 28.4 | 27.14 | **29.13** | 26.62 | 26.77 | 14.94 | 25.23 |
| $\sqrt{\beta_1}=0, \beta_2=6$ | 20 | 38.91 | 41.07 | 42.21 | 42.32 | **47.49** | 41.21 | 42.44 | 27.4 | 41.93 |
| | 50 | 60.4 | 65.4 | 63.49 | 66.96 | **75.8** | 68.95 | 70.95 | 59.36 | 71.01 |
| | 100 | 81.62 | 86.26 | 86.07 | 86.87 | **93.13** | 89.76 | 91.72 | 87.47 | 91.49 |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Logistic (mu=0,sigma=1) | 10 | 15.5 | 15.7 | 13.77 | 15.55 | **16.34** | 15.15 | 14.97 | 10.94 | 14.32 |
| $\sqrt{\beta_1}=0, \beta_2=4.4$ | 20 | 20.99 | 22.19 | 23.03 | 21.88 | **23.49** | 21.34 | 19.62 | 14.46 | 20.61 |
| | 50 | 31.14 | 34.57 | 36.15 | 34.59 | **37.44** | 33.58 | 27.54 | 20.43 | 31.17 |
| | 100 | 43.03 | 48.55 | 49.44 | 48.39 | **52.66** | 49.28 | 39.6 | 30.7 | 45.74 |
| | 500 | 89.97 | 92.15 | 88.62 | 92.39 | **94.26** | 92.78 | 14.61 | 82.51 | 92.75 |
| Student-t (df=10) | 10 | 14.7 | 14.85 | 12.37 | 13.84 | **14.89** | 14.3 | 13.35 | 10.3 | 12.6 |
| $\sqrt{\beta_1}=0, \beta_2=4$ | 20 | 18.69 | 19.6 | 15.94 | 19.06 | **20.39** | 18.66 | 17.3 | 13.13 | 17.65 |
| | 50 | 26.42 | **28.96** | 22.84 | 28.56 | 30.8 | 28.19 | 22.43 | 17.11 | 25.75 |
| | 100 | 35.53 | 39.91 | 31.58 | 39.49 | **42.58** | 39.6 | 28.82 | 22.32 | 35.33 |
| | 500 | 77.97 | 81.46 | 73.47 | 81.44 | **82.82** | 81.47 | 13.64 | 59.23 | 78.38 |
| Student-t (df=20) | 10 | 12.43 | 12.62 | 11.22 | 11.95 | **12.74** | 12.18 | 11.53 | 10.55 | 11.68 |
| $\sqrt{\beta_1}=0, \beta_2=3.375$ | 20 | 13.28 | 13.97 | 12.26 | 13.51 | **14.34** | 13.59 | 12.77 | 10.95 | 12.91 |
| | 50 | 15.82 | 17.6 | 15.15 | 17.46 | **18.3** | 16.54 | 13.88 | 11.88 | 15.35 |
| | 100 | 19.84 | 22.25 | 21.3 | 21.83 | **23.33** | 21.78 | 15.82 | 13.52 | 18.98 |
| | 500 | 36.61 | 40.86 | 40.29 | 40.76 | **42.11** | 41.29 | 11.9 | 21.48 | 34.71 |
| Student-t (df=30) | 10 | **11.16** | 10.88 | 10.47 | 10.63 | 11.06 | 10.71 | 10.5 | 10.47 | 10.4 |
| $\sqrt{\beta_1}=0, \beta_2=3.23$ | 20 | **12.91** | 12.94 | 11.53 | 12.6 | 12.75 | 12.32 | 12.03 | 10.34 | 12.17 |
| | 50 | 13.97 | 14.98 | 14.89 | 14.58 | **15.4** | 14.22 | 12.64 | 11.03 | 13.83 |
| | 100 | 16.42 | 18.04 | 17.1 | 17.76 | **18.83** | 17.88 | 13.83 | 11.65 | 15.84 |
| | 500 | 23.86 | 26.97 | 27.43 | 26.6 | 28.16 | **28.67** | 10.73 | 14.52 | 21.97 |

Table 8: Average Power for symmetrical long-tailed distribution.

| $\alpha$ | $n$ | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $n$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR}-T_w$ | $X_{APD}$ |
| 0.01 | 10 | 2.78 | 2.77 | 2.58 | 2.42 | 3.00 | 2.84 | **3.12** | 1.76 | 2.12 |
| | 20 | 5.35 | 5.52 | 5.25 | 5.88 | **6.37** | 5.95 | 5.75 | 3.95 | 5.54 |
| | 50 | 10.46 | 12.23 | 10.30 | 12.22 | **14.54** | 12.10 | 12.26 | 9.67 | 13.31 |
| | 100 | 17.32 | 21.32 | 19.11 | 21.65 | **24.70** | 19.26 | 20.35 | 18.32 | 22.96 |
| | 500 | 47.14 | 51.95 | 47.03 | 51.85 | **53.11** | 44.01 | 2.36 | 40.24 | 49.71 |
| 0.05 | 10 | 9.82 | 9.68 | 8.84 | 9.43 | **10.29** | 9.71 | 9.73 | 6.57 | 8.77 |
| | 20 | 13.69 | 14.23 | 13.46 | 14.39 | **15.47** | 14.55 | 13.73 | 9.39 | 13.55 |
| | 50 | 21.64 | 24.66 | 21.47 | 24.29 | **27.02** | 25.22 | 22.16 | 17.61 | 23.81 |
| | 100 | 30.89 | 35.29 | 32.00 | 34.62 | **38.13** | 35.85 | 30.87 | 26.95 | 34.12 |
| | 500 | 59.52 | 63.59 | 60.79 | 62.83 | **64.37** | 61.96 | 52.63 | 49.67 | 60.17 |
| 0.1 | 10 | 16.01 | 16.15 | 15.25 | 15.82 | **16.83** | 15.79 | 15.42 | 11.44 | 14.85 |
| | 20 | 20.96 | 21.95 | 20.99 | 21.87 | **23.69** | 21.42 | 20.83 | 15.26 | 21.05 |
| | 50 | 29.55 | 32.30 | 30.50 | 32.43 | **35.55** | 32.30 | 29.49 | 23.96 | 31.42 |
| | 100 | 39.29 | 43.00 | 41.10 | 42.87 | **46.11** | 43.66 | 37.96 | 33.13 | 41.48 |
| | 500 | 65.68 | 68.29 | 65.96 | 68.24 | **69.47** | 68.84 | 30.18 | 55.55 | 65.56 |

The results for the last category of alternative distributions, asymmetric distributions, are presented in Tables 9, 10, 11. First three distributions are skewed to the right (ordered from the highest skewness to the most close to zero), and the rest two distributions are left-skewed (one with low skewness and one with close to zero). The average results power across all asymmetric distributions is presented in Table 12. For asymmetric distributions normality tests have much more power than in case of symmetric distributions. The results do not show one particular test that outperforms the rest. The results vary widely depending on the type of asymmetry, sample size and significance level. For lognormal distribution (strongly right-skewed) $BM_{3-6}$ and $T_{Lmom}$ perform the best. From modifications of $JB$ statistic, $DH$ also performs well. However, for big sample sizes (100 and 500) almost all statistics have 100% power. For distributions with weaker right asymmetry $BM_{3-6}$ and $DH$ are the most powerful tests. As far as distributions skewed to the left are concerned, $T_{Lmom}$, $X_{APD}$ and $BM_{3-6}$ perform the best. Contrary to the left asymmetry, $DH$ test is not better than the standard $JB$ test.

Table 9: Empirical power results for asymmetric distributions ($\alpha = 0.01$).

| Alternative | $n$ | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K^2$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR} - T_w$ | $X_{APD}$ |
| Lognormal (logmean=0,logsd=1) | 10 | 28.03 | 28.19 | 29.67 | **34.12** | 28.42 | 29.96 | 29.03 | 7.86 | 28.56 |
| $\sqrt{\beta_1} = 6.18, \beta_2 = 5.22$ | 20 | 59.66 | 58.57 | 51.78 | **75.48** | 60.17 | 66.11 | 74.95 | 28.88 | 72.18 |
| | 50 | 96.97 | 96.58 | 96.98 | 99.82 | 96.44 | 97.32 | **99.9** | 81.31 | 99.77 |
| | 100 | 99.99 | 99.99 | **100** | **100** | 99.99 | 99.99 | **100** | 99.3 | **100** |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Gumbel (mu=1,sigma=1) | 10 | 5.19 | 5.2 | 5.23 | 4.72 | 5.07 | **5.32** | 4.56 | 1.86 | 3.93 |
| $\sqrt{\beta_1} = 1.14, \beta_2 = 5.4$ | 20 | 13.05 | 12.68 | 12.12 | 11.91 | 11.98 | **13.35** | 12.18 | 3.79 | 11.71 |
| | 50 | 37.17 | 34.85 | 39.12 | 40.46 | 33.79 | 31.44 | **41.93** | 9.31 | 41.83 |
| | 100 | 71.42 | 67.65 | 65.56 | **80.97** | 66.12 | 55.62 | 80.18 | 17.91 | 80.55 |
| | 500 | **100** | **100** | **100** | **100** | **100** | 99.99 | **100** | 88.42 | **100** |
| Chi-squared (df=20) | 10 | 1.92 | 1.92 | 1.8 | 1.67 | 1.88 | **2** | 1.7 | 1.07 | 1.51 |
| $\sqrt{\beta_1} = 0.63, \beta_2 = 3.6$ | 20 | 4.54 | 4.42 | 4.25 | 3.58 | 3.98 | **4.55** | 3.79 | 1.4 | 3.81 |
| | 50 | 12.14 | 11.2 | 12.1 | 11.3 | 10.5 | 9.35 | 12.45 | 2.09 | **12.57** |
| | 100 | 26.75 | 23.85 | 29.23 | 31.15 | 22.42 | 16.44 | 30.72 | 3.09 | **31.37** |
| | 500 | 99.16 | 98.69 | 98.8 | 99.74 | 98.36 | 82.91 | **100** | 18.58 | 99.56 |
| Weibull (shape=1,scale=1) | 10 | 15.58 | 15.74 | 15.79 | 18.16 | 14.72 | **16.56** | 14.51 | 2.97 | 14.34 |
| $\sqrt{\beta_1} = -2.85, \beta_2 = 6$ | 20 | 36.05 | 35.01 | 35.56 | **51.54** | 34.1 | 40.76 | 50.59 | 9.71 | 47.35 |
| | 50 | 81.53 | 79.42 | 8.33 | 97.42 | 77.54 | 81.33 | **98.41** | 43.35 | 96.79 |
| | 100 | 99.59 | 99.43 | **100** | **100** | 98.54 | 98.84 | **100** | 85.74 | **100** |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Beta (a=10,b=5) | 10 | 1.05 | **1.07** | 1.05 | 1.05 | 0.98 | **1.07** | 0.97 | 0.93 | 1.04 |
| $\sqrt{\beta_1} = -0.33, \beta_2 = 2.82$ | 20 | 1.12 | 1.05 | 1.13 | 0.94 | 0.98 | 1.1 | 1.11 | 0.76 | **1.21** |
| | 50 | 1.64 | 1.25 | 1.32 | 1.7 | 1.15 | 1.11 | **2.78** | 1.2 | 2.4 |
| | 100 | 3.21 | 1.89 | 2.59 | 5.03 | 1.81 | 1.16 | **7.21** | 1.45 | 6.41 |
| | 500 | 61.14 | 47.14 | 48.1 | 79.87 | 39.99 | 5.05 | **100** | 6.77 | 75.83 |

Table 10: Empirical power results for asymmetric distributions ($\alpha = 0.05$).

| Alternative | $n$ | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K^2$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR}-T_w$ | $X_{APD}$ |
| Lognormal (logmean=0,logsd=1) | 10 | 46.59 | 50.03 | 50.09 | 51.3 | 47.31 | **54.21** | 49.18 | 14.55 | 50.7 |
| $\sqrt{\beta_1}=6.18, \beta_2=5.22$ | 20 | 78.84 | 82.11 | 83.13 | 88.6 | 78.4 | 87.96 | **90.14** | 40.9 | 88.16 |
| | 50 | 99.58 | 99.77 | 99.79 | 99.95 | 99.29 | 99.88 | **99.98** | 90.4 | 99.92 |
| | 100 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | 99.78 | **100** |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Gumbel (mu=1,sigma=1) | 10 | 15.18 | 15.67 | 15.47 | 12.81 | 14.23 | **16.22** | 12.92 | 5.93 | 12.74 |
| $\sqrt{\beta_1}=1.14, \beta_2=5.4$ | 20 | 27.5 | 28.7 | 28.82 | 26.18 | 26.07 | **29.99** | 27.15 | 8.45 | 26.68 |
| | 50 | 58.49 | 62.06 | 62.55 | 65 | 56.83 | 62.19 | 63.47 | 17.87 | **64.02** |
| | 100 | 88.44 | 90.53 | 90.44 | **93.63** | 86.83 | 89.21 | 91.8 | 31.85 | 92.55 |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | 95.62 | **100** |
| Chi-squared (df=20) | 10 | 9.03 | 9.24 | 9.17 | 7.82 | 8.99 | **9.44** | 8.25 | 4.83 | 8.06 |
| $\sqrt{\beta_1}=0.63, \beta_2=3.6$ | 20 | 13.38 | 13.56 | 13.58 | 11.93 | 12.44 | **13.82** | 12.22 | 5.43 | 12.57 |
| | 50 | 26.96 | 28.42 | 28.6 | **28.79** | 25.04 | 28.16 | 27.35 | 7.64 | 28.16 |
| | 100 | 50.69 | 53.27 | 52.47 | **58.46** | 47.79 | 49.58 | 54.21 | 9.47 | 56.14 |
| | 500 | 99.95 | 99.96 | **100** | **100** | 99.95 | 99.62 | **100** | 36.75 | 99.94 |
| Weibull (shape=1,scale=1) | 10 | 30.27 | 33.02 | 34.28 | 34.36 | 30.05 | **36.53** | 31.86 | 7.11 | 33.9 |
| $\sqrt{\beta_1}=-2.85, \beta_2=6$ | 20 | 57.74 | 62.34 | 63.85 | 73.64 | 55.47 | 71.49 | **77.38** | 17.7 | 72.76 |
| | 50 | 95.36 | 97.86 | 97.93 | 99.63 | 93.29 | 99.02 | **99.75** | 64.34 | 99.45 |
| | 100 | **100** | **100** | **100** | **100** | 99.98 | **100** | **100** | 94.65 | **100** |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Beta (a=10,b=5) | 10 | 4.92 | 4.94 | 5.67 | 5.18 | 4.95 | 5.27 | 5.17 | 4.86 | **5.73** |
| $\sqrt{\beta_1}=-0.33, \beta_2=2.82$ | 20 | 5.76 | 5.53 | 5.32 | 5.17 | 4.74 | 5.85 | 6.1 | 4.15 | **6.37** |
| | 50 | 8.87 | 8.04 | 8.55 | 9.24 | 6.15 | 8.15 | **11.01** | 5.4 | 10.35 |
| | 100 | 15.71 | 13.48 | 15.19 | 20.57 | 10.31 | 13.54 | 21.52 | 6.4 | **21.55** |
| | 500 | 92.18 | 91.79 | 92.79 | 95.47 | 88.27 | 77.92 | **100** | 18.51 | 92.31 |

Table 11: Empirical power results for asymmetric distributions ($\alpha = 0.1$).

| Alternative | $n$ | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K^2$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR} - T_w$ | $X_{APD}$ |
| Lognormal (logmean=0,logsd=1) | 10 | 58.41 | 64.63 | 62.95 | 62.3 | 58.35 | **67.68** | 61.86 | 19.56 | 62.47 |
| $\sqrt{\beta_1} = 6.18, \beta_2 = 5.22$ | 20 | 86.56 | 92.12 | 92.15 | 93.38 | 86.66 | **93.99** | 94.34 | 50.08 | 93.31 |
| | 50 | 99.94 | 99.99 | **100** | **100** | 99.82 | **100** | **100** | 93.72 | 99.99 |
| | 100 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | 99.89 | **100** |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Gumbel (mu=1,sigma=1) | 10 | 22.7 | 24.21 | 23.79 | 20.66 | 22.54 | **24.39** | 20.93 | 10.51 | 21.25 |
| $\sqrt{\beta_1} = 1.14, \beta_2 = 5.4$ | 20 | 37.46 | 41.26 | 41.23 | 37.63 | 37.25 | **41.77** | 37.79 | 14.32 | 38.18 |
| | 50 | 69.92 | 75.73 | 75.45 | **75.8** | 69.93 | 74.47 | 73.31 | 25.53 | 74.18 |
| | 100 | 94.41 | 96.18 | 96.24 | **96.8** | 94.5 | 95.21 | 95.56 | 41.23 | 95.96 |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | 97.5 | **100** |
| Chi-squared (df=20) | 10 | 14.43 | 14.89 | 14.6 | 13.46 | 14.04 | **14.95** | 13 | 9.78 | 13.42 |
| $\sqrt{\beta_1} = 0.63, \beta_2 = 3.6$ | 20 | 20.49 | 22.54 | 22.53 | 19.81 | 20.71 | **22.59** | 20.17 | 10.45 | 20.65 |
| | 50 | 38.17 | **42.94** | 41.51 | 41.5 | 37.87 | 40.94 | 39.35 | 13.29 | 40.52 |
| | 100 | 65.83 | 70.92 | 70.75 | **71.66** | 65.41 | 66.67 | 67.54 | 16.5 | 69.47 |
| | 500 | 99.99 | 99.99 | 99.99 | 99.99 | 99.99 | 99.96 | **100** | 50 | 99.95 |
| Weibull (shape=1,scale=1) | 10 | 41.83 | 48.58 | 57.59 | 45.36 | 41.31 | **52.26** | 45.64 | 11.87 | 46.29 |
| $\sqrt{\beta_1} = -2.85, \beta_2 = 6$ | 20 | 70.42 | 80.12 | 80.55 | 83.32 | 69.43 | 84.29 | **86.4** | 26.37 | 82.75 |
| | 50 | 99.19 | 99.79 | 99.77 | 99.91 | 98.12 | 99.83 | **99.94** | 73.9 | 99.89 |
| | 100 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | 96.77 | **100** |
| | 500 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Beta (a=10,b=5) | 10 | 9.52 | 9.78 | 9.69 | 9.59 | 9.12 | 9.89 | 9.65 | 9.46 | **10.1** |
| $\sqrt{\beta_1} = -0.33, \beta_2 = 2.82$ | 20 | 11.07 | 10.82 | 10.37 | 10.35 | 9.86 | **12.1** | 11.82 | 9.54 | 11.81 |
| | 50 | 16.8 | 16.17 | 16.25 | 16.54 | 12.95 | 16.99 | **19.27** | 10.66 | 18.44 |
| | 100 | 29.52 | 30.31 | 31.1 | 34.1 | 22.9 | 28.58 | 34.2 | 12.03 | **34.25** |
| | 500 | 97.11 | 97.41 | 98.23 | 98.24 | 96.41 | 92.39 | **100** | 28.01 | 96.23 |

Table 12: Average Power for asymmetric distributions.

| $\alpha$ | $n$ | Goodness-of-fit tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K^2$ | $JB$ | $ALM$ | $DH$ | $RJB$ | $BM_{3-6}$ | $T_{Lmom}$ | $T_{MC-LR}-T_w$ | $X_{APD}$ |
| 0.01 | 10 | 10.35 | 10.42 | 10.71 | **11.94** | 10.21 | 10.98 | 10.15 | 2.94 | 9.88 |
| | 20 | 22.88 | 22.35 | 20.97 | **28.69** | 22.24 | 25.17 | 28.52 | 8.91 | 27.25 |
| | 50 | 45.89 | 44.66 | 31.57 | 50.14 | 43.88 | 44.11 | **51.09** | 27.45 | 50.67 |
| | 100 | 60.19 | 58.56 | 59.48 | 63.43 | 57.78 | 54.41 | 63.62 | 41.50 | **63.67** |
| | 500 | 92.06 | 89.17 | 89.38 | 95.92 | 87.67 | 77.59 | **100.00** | 62.75 | 95.08 |
| 0.05 | 10 | 21.20 | 22.58 | 22.94 | 22.29 | 21.11 | **24.33** | 21.48 | 7.46 | 22.23 |
| | 20 | 36.64 | 38.45 | 38.94 | 41.10 | 35.42 | 41.82 | **42.60** | 15.33 | 41.31 |
| | 50 | 57.85 | 59.23 | 59.48 | **60.52** | 56.12 | 59.48 | 60.31 | 37.13 | 60.38 |
| | 100 | 70.97 | 71.46 | 71.62 | **74.53** | 68.98 | 70.47 | 73.51 | 48.43 | 74.05 |
| | 500 | 98.43 | 98.35 | 98.56 | 99.09 | 97.64 | 95.51 | **100.00** | 70.18 | 98.45 |
| 0.1 | 10 | 29.38 | 32.42 | 33.72 | 30.27 | 29.07 | **33.83** | 30.22 | 12.24 | 30.71 |
| | 20 | 45.20 | 49.37 | 49.37 | 48.90 | 44.78 | **50.95** | 50.10 | 22.15 | 49.34 |
| | 50 | 64.80 | **66.92** | 66.60 | 66.75 | 63.74 | 66.45 | 66.37 | 43.42 | 66.60 |
| | 100 | 77.95 | 79.48 | 79.62 | **80.51** | 76.56 | 78.09 | 79.46 | 53.28 | 79.94 |
| | 500 | 99.42 | 99.48 | 99.64 | 99.65 | 99.28 | 98.47 | **100.00** | 75.10 | 99.24 |

## 4. Conclusions

In this study, we performed a comprehensive investigation of nine tests for normality based on measures of the moments. In a simulation study we focused on a different forms of shape departure from normality such as symmetric short-tailed, symmetric long-tailed, or asymmetric. None of the tests considered in this study is uniformly most powerful for all types of alternative distributions, sample sizes and significance levels considered. If the distribution is symmetric and short-tailed two test are the most powerful, Desgagnéa and Lafaye de Micheaux's $X_{APD}$ test and D'Agostino and Pearson's $K^2$ . Gel and Gastwirth's *RJB* test is one of the most powerful tests for normality based on measures of the moments across a wide array of symmetrical and long-tailed alternative distributions. For the last category of alternative distributions, asymmetric distributions, it is difficult to distinguish one test. Bontemps-Meddahi's $BM_{3-6}$ for right-skewed distributions and Hosking's $T_{Lmom}$ for left-skewed perform fairly well.

The *JB* test performs well for symmetric distributions with long tails and for slightly skewed distributions with long tails. However, the power of the *JB* test is very poor for distributions with short tails. As Thadewald and Büning (2007) reported the Urzùa test has no improvement of power to the classical *JB* test in the case of Monte Carlo simulated critical values. Gel and Gastwirth modification of *JB* that uses a robust estimate of the dispersion seems to be the best modification in the case distributions with long tails and Doornik–Hansen modification in the case of short-tailed distributions.

Finally, the authors would like to indicate two tests that have quite reasonable power for all alternative distributions and have advantage of being very closely approximated by chi-squared distribution with two degrees of freedom. These two test are the Doornik–Hansen test and the Desgagnéa and Lafaye de Micheaux test. As a concluding remark, practitioners should carefully act when graphical techniques such as histogram or moment statistics suggest that the sample comes from symmetric distribution. In this case, normality tests do not perform well for small sample sizes (below 50), especially when symmetry is accompanied by short tail of distribution.

## Acknowledgement

This paper was presented on the MSA 2019 conference, 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

BONETT, D., SEIER, E., (2002). A test of normality with high uniform power. *Computational Statistics & Data Analysis*, 40(3), pp. 435–445.

BONTEMPS, C., MEDDAHI, N., (2005). Testing Normality: A GMM Approach. *Journal of Econometrics*, 124(1), pp. 149–186. doi:10.1016/j.jeconom.2004.02.014.

BOWMAN, K. O., SHENTON, L. R., (1975). Omnibus test contours for departures from normality based on b1 and b2. *Biometrika*, 62(2), pp. 243–250.

BRYS, G., HUBERT, M. and STRUYF, A., (2006). Robust measures of tail weight. *Computational Statistics & Data Analysis*, 50(3), pp. 733–759. doi:10.1016/j.csda.2004.09.012.

BRYS, G., HUBERT, M. and STRUYF, A., (2004). A Robust Measure of Skewness. *Journal of Computational and Graphical Statistics*, 13(4), pp. 996–1017. doi:10.1198/106186004x12632.

BRYS, G., HUBERT, M. and STRUYF, A., (2007). Goodness-of-fit tests based on a robust measure of skewness. *Computational Statistics*, 23(3), pp. 429–442. doi:10.1007/s00180-007-0083-7.

D'AGOSTINO, R. B., PEARSON, E. S., (1973). Tests for Departure From Normality. Empirical Results for the Distributions of b2 and p b1. *Biometrika*, 60, pp. 613–622. doi:10.1093/biomet/60. 3.613.

D'AGOSTINO, R., BELANGER, A. and D'AGOSTINO, R. JR., (1990). A Suggestion for Using Powerful and Informative Tests of Normality. *The American Statistician*, 44(4), p. 316. doi:10.2307/2684359.

DOMAŃSKI, C. K., (2010). Uwagi o testach Jarque-Bera, *Przegląd Statystyczny*, 57(4), pp. 19–26.

DOMAŃSKI, C. K., JĘDRZEJCZAK, A., (2017). Consistency Tests Based on Moments. *Annales Universitatis Mariae Curie-Skłodowska, sectio H, Oeconomia*, 50(4), pp. 89–99. doi:dx.doi.org/10.17951/h.2016.50.4.89.

DOORNIK, J. A., HANSEN, H., (2008). An Omnibus Test for Univariate and Multivariate Normality. *Oxford Bulletin of Economics and Statistics*, 70, pp. 927–939. doi:10.1111/j.1468-0084. 2008.00537.x.

FARRELL, P. J., ROGERS-STEWART, K., (2006). Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation*, 76(9), pp. 803–816. doi:10.1080/10629360500109023.

FISHER, R. A., (1930), The moments of the distribution for normal samples of measures of departure from normality, *Proceedings of the Royal Statistical Society A*, 130, pp. 16–28.

GEARY, R., (1936). Moments of the Ratio of the Mean Deviation to the Standard Deviation for Normal Samples. *Biometrika*, 28(3/4), p. 295. doi:10.2307/2333953.

GEL, Y. R., GASTWIRTH, J. L., (2008). A Robust Modification of the Jarque-Bera Test of Normality. *Economics Letters*, 99(1), pp. 30–32. doi:10.1016/j.econlet.2007.05.022.

HOSKING, J. R. M., (1990). L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1), pp. 105–124.

JARQUE, C. M., BERA, A. K., (1987). A Test for Normality of Observations and Regression Residuals. International Statistical Review. *Revue International de Statistique*, 55(2), pp. 163–172. doi:10.2307/1403192.

JOHNSON, N. L., (1949). Bivariate Distributions Based on Simple Translation Systems . *Biometrika*. 36 (3/4), pp. 297–304. doi:10.1093/biomet/36.3-4.297.

KOMUNJER, I., (2007). Asymmetric power distribution: Theory and applications to risk measurement. *Journal of Applied Econometrics*, 22(5), pp. 891–921. doi:10.1002/jae.961.

PIONTEK, K., (2007). Pomiar i testowanie skośności rozkładów stóp zwrotu instrumentów finansowych. *Prace Naukowe Akademii Ekonomicznej we Wrocławiu. Taksonomia*, 14, pp. 122–130.(in Polish).

ROMÃO, X., DELGADO, R. and COSTA, A., (2010). An empirical power comparison of univariate goodness-of-fit tests for normality. *Journal of Statistical Computation and Simulation*, 80(5), pp. 545–591. doi:10.1080/00949650902740824.

THADEWALD, T., BÜNING, H., (2007). Jarque–Bera Test and its Competitors for Testing Normality – A Power Comparison. *Journal of Applied Statistics*, 34(1), pp. 87–105. doi:10.1080/02664760600994539.

STEIN, C., (1972). A bound for the error in the normal approximation to the distribution of a sum of dependant random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematics, Statistics and Probability*, Vol. 2, pp. 583–602.

URZÙA, C., (1996). On the correct use of omnibus tests for normality. *Economics Letters* 53, pp. 247–251.doi:10.1016/S0165-1765(96)00923-8.

WIJEKULARATHNA, D. K., MANAGE, A. B. W. and SCARIANO, S. M. (2019). Power analysis of several normality tests: A Monte Carlo simulation study. *Communications in Statistics - Simulation and Computation*, pp. 1–17. doi:10.1080/03610918.2019.1658780.

WILSON E. B., HILFERTY M. M., (1931). The Distribution of Chi-square. *Proceedings of the National Academy of Sciences of the United States of America* Vol. 17, No. 12, pp. 684–688. doi:10.1073/pnas.17.12.684.

YAP, B. W., SIM, C. H., (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12), pp. 2141–2155.

ZGHOUL, A. A., (2010). A Goodness of Fit Test for Normality Based on the Empirical Moment Generating Function. *Communications in Statistics – Simulation and Computation*, 39(6), pp. 1292–1304.

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 179–191, DOI 10.21307/stattrans-2020-061*
Received – 30.09.2020; accepted – 25.11.2020

# Predicting Polish transport industry equilibrium characteristics as an inverse problem: An Entropy Econometrics Model

**Second Bwanakare[1], Marek Cierpiał-Wolan[2]**

## ABSTRACT

The business environment dynamics is governed by a high degree of uncertainty and risk; consequently, in a majority of cases investors face serious difficulties when making business decisions. Additionally, when detailed statistical information relating to industry is missing, any decisions may become a matter of highly risky conjectures.

The present article proposes a simultaneous equation model based on the entropy econometrics estimator for recovering some key industrial subsector long-term equilibrium characteristics in the situation where only sparse, insufficient statistical information is available (e.g. only aggregated data on the whole industry).

The model is applied to the transportation equipment manufacturing industry in Poland, which is composed of eight sub-sectors. As a result of the above procedure, an observation has been made that all firms from different sub-sectors have to increase their steady-state concentration ratios, while the highest concentration corresponds to the lowest increase in profitability. The model outputs conform to the market tendency in this sector and should lead to further applications of the NCEE methodology in business activity on a worldwide scale.

**Key words:** transport industry, inverse problem, econometrics, non-extensive entropy econometrics.

## 1. Introduction

One of the most important areas of services is transport, which largely affects the economic development of each country. Not only is it an instrument for the exchange of goods and services but also an important factor in GDP growth and it also influences the development of other sectors of the national economy. It is worth emphasizing that the production of transport equipment is an extremely important determinant of

[1] University of Information Technology and Management (WSIZ), Rzeszow, Poland. E-mail: sbwanakare@wsiz.rzeszow.pl. ORCID: https://orcid.org/0000-0003-0574-1302.

[2] Statistical Office in Rzeszów, University of Rzeszów, Rzeszów, Poland. E-mail: M.Cierpiał-Wolan@stat.gov.pl. ORCID: https://orcid.org/0000-0003-2672-3234.

transport development[3]. In the European Union, three groups of countries can be distinguished in this respect. The first group includes countries where the average share of the production of transport equipment in the global production in 2012–2018 ranged from 7 to about 12 percent (Slovakia, the Czech Republic, Hungary and Germany). The second group, which includes Romania, Sweden, Poland, Slovenia and Spain, covers countries where this share amounts to around 4 percent, and the remaining countries do not exceed the 3 percent share.

In Poland, in the entities included in the production of transport equipment, after the decline in dynamics in 2012, a successive increase was observed in subsequent years, including the highest in 2015 (by 11.1%). In the last two years of the analysed period, the growth rate of global production slightly slowed down and was lower than the total. Both the pace and the volatility of dynamics in production entities for transport was shaped mainly by the results achieved by entities producing motor vehicles, trailers and semi-trailers, excluding motorcycles. The share of this division's revenues accounts for approximately 90% of production revenues for transport. The remaining production showed significant fluctuations in dynamics. After a period of growth in 2012–2015, in the next two years, global production in this division decreased, while the last year of the analysed period brought a significant increase (by 19.9%).

What is also interesting is the fact that the global production in Poland calculated for entities employing more than 10 people in the years 2012–2016 brought a stable growth (in the range of 1.5% -3.4%). Both 2017 and 2018 saw acceleration in the growth rate of global production.

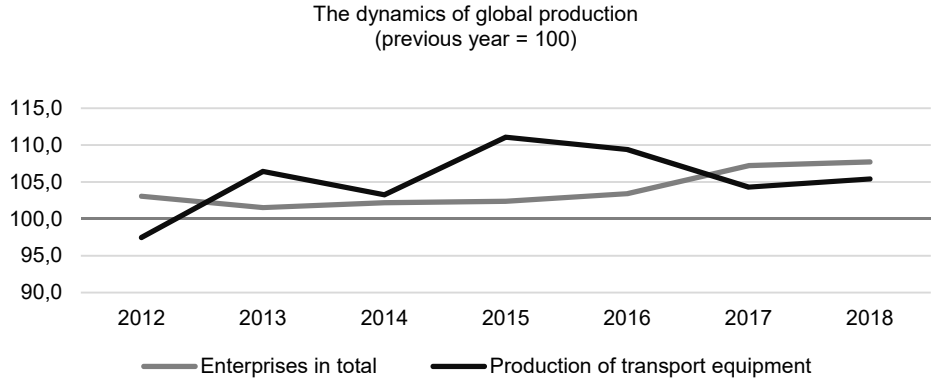



**Figure 1.** Dynamics of global production

[3] Production of transport equipment consists of two divisions: production of motor vehicles, trailers and semi-trailers, excluding motorcycles (29) and production of other transport equipment (30).

The present paper applies the power-law (PL)-related cross-entropy econometrics (PL-CEE) methodology for recovering the main optimal equilibrium characteristics of the Polish transport industry sub-sectors. In this paper, we will reply to the business question concerning, among others, the optimal number of subsector firms of the transport manufacturing industry consistent with steady-state industry configuration given industry initial characteristic conditions.

In this kind of the problem, we are dealing with ill-behaved inverse problems, suggesting that we want to recover a larger number of model parameters than there are associated data point observations known with uncertainty in this study. As documented in recent publications, the Tsallis power-law (PL)-based non-extensive cross-entropy econometrics (NCEE) approach better deals, conceptually, with such complex non-linear inverse problems. NCEE is based upon the q-generalized Kullback-Leibler (K-L) information divergence criterion function under constraining information characterized by the Bayesian information processing optimal rule. Thus, we consider PL-related non-extensive entropy will remain valuable even in the case of low-frequency series since the outputs provided by Gaussian law correspond to the limiting case of the Tsallis entropy when the Tsallis q-parameter equals unity.

## 2. Modelling the Polish Industry of Transport

For decades, statistical and mathematical tools to handle ill-posed inverse problem systems have been sought in diverse fields–model parameter estimation, medical imaging, modelling in the life sciences, oil and mineral deposit exploration, shape optimization, etc. For more about inverse problems, see, e.g. Tikhonov regularization theory, Gibbs-Shannon-Jaynes. Interestingly enough, a non-particular hypothesis is required while applying the PL-CEE model in contrast with the traditional econometrics techniques, which generally impose a large number of not always realistic hypotheses on the model.

Contrary to many other fields, the management or economics science, in general, have neglected the link between phenomena and power-law (Gabaix X., September 2008) characterizing complex systems within the class of Levy's processes. In light of recent literature, the amplitude and frequency of socio-economic fluctuations are not considered to substantially diverge from many other extreme events, natural or human-related, once they are explained at the same time (or space) scale. Y. Ikeda and W. Souma (2008) have worked on an international comparison of labour productivity distribution for manufacturing and non-manufacturing firms. A power-law distribution in terms of firms and sector productivity has been found in the US and Japan data. Testing Gibrat's law of proportionate effect, Fujiwara et al. (2004) have found, among other things, that the upper-tail of the distribution of the firm size can

be fitted with a power-law (Pareto–Zipf law). According to many studies (e.g. Bottazzi G. et al., 2007; Champernowne D. G., June 1953), a large array of economic laws take the form of PL, in particular, macroeconomic scaling laws, distribution of income, wealth, the size of cities and firms, and the distribution of financial variables such as returns and trading volume. In a recent monograph publication (2019), the author has proposed a theorem linking low-frequency time series socio-economic phenomena–and thus input-output one period systems–with PL distribution. The above citations are not exhaustive.

The PL-CEE is a precious device for econometric modelling even in the case of low-frequency series since outputs provided by the Gibbs-Shannon entropy approach correspond to the Tsallis entropy limiting case of Tsallis q-parameter equal unity. What is more, many complex phenomena involve the long-range correlations which can continuously be seen when data are time (space) scale-aggregated. This could be because of the interaction between the functional relationships describing the phenomena involved and the inheritance properties of power-law (PL). Thus, delimiting the threshold values for a PL (Levy's stable process) transition plausibly towards the Gaussian structure as a function of data frequency level is difficult since each phenomenon may display its rate of convergence towards the central theorem limit attractor. Consequently, a systematic application of the Shannon-Gibbs entropy approach, even based on annual data, could lead to unstable and misleading results. Inversely, since non-extensive Tsallis entropy generalizes the exponential family of laws, it should fit well with high or low-frequency series. In particular, Mantegna R. N. and Stanley H. E. (1999) have studied the dynamics of a general system composed of interacting units each with a complex internal structure comprising many subunits, where they grow in a multiplicative way over 20 years. They found a system following a PL distribution. This is similar to the present case study, where we deal with an industrial sector composed of sub-sectors within which a large number of economic agents provide complex business activities for a given period.

Following the above reasoning, the present study based on non-extensive entropy econometrics extends Shannon-Gibbs maximum entropy econometrics to non-ergodic systems. As in statistical physics, socio-economic random events should display two types of stochastic behaviour: ergodic and non-ergodic. Whenever isolated in a closed space, ergodic systems dynamically visit with equal probability all the allowed micro-states (Tsallis, 2009). This is the case for Gibbs-Shannon entropy. Next, since all events are independent or quasi-independent (locally dependent) and equally probable, this means that the above entropy is a linear, positive function of the number of possible states – thus of new data, and then is extensive. In reverse, as a consequence of possible multi-level correlation between system microstates, non-ergodic systems are characterized by entropy which is no longer a linear, positive function of the number

of possible states, and then non-extensive. An important fact to be noticed here is the connection between information theory and a Gaussian variable. This connection results from the fact that a Gaussian variable displays the largest entropy among all random variables of equal variance.

***Q-generalized Cross-entropy for Inverse Problem Solution***

As already said above, the model to be estimated displays more unknown parameters than observed data point observations. In this section, we recall the definition of an ill-posed inverse problem and present a PL-related cross-entropy model in the context of the proposed model. In essence, the canonical ill-posed inverse problem as the one we deal with in this paper can be formally presented as follows:

$$X(\zeta) = \int_D g(Y)h(Y,\zeta)dY + b(\zeta) \tag{1}$$

$X$ : means the observed matrix of updated priors, e.g. the prior data matrix in Table 1,

$Y$ : designates the unknown matrix of the Polish optimal, long-run subsector transport industry configuration to be later estimated,

D: defines the Hilbert support space of the model,

$g$ : is the transformation kernel linking measures $X$ and Y,

$b$ : explains random errors.

This is a basic model which consists in solving an integral equation of the first kind. As said in (Bwanakare, 2014), inverse problem recovery finds application in various fields of science, particularly in the context of Optimal Control Theory. Among different techniques proposed for solving this type of problems, the Tikhonov related regularization theory remains the most applied besides the Gibbs-Shannon-Jaynes maximum (cross) entropy principle and the ill-posed stationary first-order Markov process, in which the operator $g$ can be seen as a generalized transition matrix while $X$ and $Y$ as the Markov states. The contribution of this paper consists in extending the application of the non-extensive cross entropy formalism to search for global regularity—consistent with the maximum (non-extensive) entropy principle—while yielding the smoothest reconstructions of the Polish optimal subsector transport industry configuration, given initial conditions to be presented in the next paragraphs, according to the Jaynes approach.

Next, we follow recent works applying the non-extensive entropy econometrics and define a q-Tsallis-Kullback-Leibler dual entropy criterion function for forecasting the Polish optimal subsector transport industry configuration, as follows:

$$Min\, H_q\left(p // p^0\right) \equiv \lambda \sum p_{kjm} \frac{\left[\frac{p_{kjm}}{p^o{}_{kjm}}\right]^{q-1} - 1}{q-1} + (1-\lambda) \sum \mu_{k\bullet s} \frac{\left[\mu_{j\bullet s}/\mu^o{}_{j\bullet s}\right]^{q-1} - 1}{q-1} \tag{2}$$

Subject to

$$\sum\nolimits_j P_j = 1 \text{ with} \tag{3}$$

$$\sum\nolimits_{s>2\ldots S}^{S} \mu_{j\bullet s} = 1 \tag{4}$$

$$\Omega(\text{Nj,Xj,Yj}) = \text{Cij} \tag{5}$$

where:

$X_j$ : transport industry subsector average costs,

$Y_j$ : transport industry subsector average production,

$N_j$: number of entreprises in a given subsector,

$p_{ikm}$ : probability distribution on the support space point m defining the parameter k in the equation i

$\mu_{j\bullet s}$ is the random error probability on subsector accounts defined on a support space s,

$\lambda$: weight on parameters in the criterion function.

The system of equations explained in the equation 5 will be explained later in the next section. Nevertheless, the main fact to underscore is that the system stands for an inverse problem, suggesting that the model presents more parameters to estimate than the observation points.

There exist a few types of constraining forms defining expectations in Tsallis statistics. In the above model we apply the normalized Tsallis-Mendes-Plastino (TMP) constraints (also known as q-averages or an escort distribution) to the reparametrized parameters; see, e.g. Golan (1996) of the system of equations (equ. 5).

The form of the TMP is as follows:

$$\left\langle y_q \right\rangle = \sum_i \frac{{p_i}^q}{\sum_i {p_i}^q} y_i \tag{6}$$

The real $q$ stands for the Tsallis parameter, whose value varies between 1 and 5/3, suggesting the case of phenomena evolving within the Gaussian basin of attraction. If the q-Tsallis parameter recovers the value 1, we get the PL limiting Gaussian case already alluded to in the Introduction section.

Above, $H_q\left(p // p^0\right)$ is nonlinear and measures the relative (cross-entropy) entropy in the model. The symbol // is a "distance metric[4]" of divergence information. We need to find the minimum divergence between the prior $p^0$ and the posterior $p$ (equ. 2) while the imposed restrictions (equ. 3–5) must be fulfilled. For more information about cross-entropy interpretation; see, e.g. Golan (1996), Bwanakare (2014). As far as the parameter confidence area is concerned, we send interested readers, e.g. to the work (Bwanakare, 2019). Finally, it would be worthwhile to summarize below the main steps to be followed while applying the proposed cross-entropy approach:

a) fixing the phenomenon to be modelled, its explicative variables, plus its mathematical form,
b) collecting sample data,
c) setting up parameter support space points for each parameter and for the random component. The support space points are defined over the potential existence area of the parameters,
d) setting up the initial values for each parameter. These values should reflect the highest knowledge about each parameter,
e) building a program code linking all the information provided in Steps *a* through *d*. The main part of this program is of the optimization formulated as follows.

Minimizing the weighted divergence between unknown posterior and prior probabilistic of a non-extensive Tsallis entropy functional subject to the next Bayesian[5] restrictions:

- Moment formulation in the form of econometric model equations according to Steps *c* and *d* above.
- The random component is formulated according to Step *c*.
- The regular conditions must sum the probability space points of each parameter up to unity.

[4] However, note that K-L divergence is not a true metric since it is not symmetric and does not satisfy the triangle inequality.

[5] For the relationship between the Bayesian and maximum entropy parameter parameterization; see, e.g. (Golan A., 1996).

## 3. The proposed business model and input data

The system of equations Ω(Nj,Xj,Yj) in equ. (5) relates to the below econometric model (equ. 7–11), which stands for the main constraining component of the cross-entropy system (equ. 2–5).

The econometric model is as follows:

$$N_{jt} = \alpha_{0j} + \lambda_{1j} N_{jt-1} - \alpha_{2j} V_{jt} + \varepsilon_{1j} \quad (7)$$

$$X_{jt} = \beta_{0j} + \lambda_{2j} X_{jt-1} + \beta_{2j} Y_{jt} + \varepsilon_{2j} \quad (8)$$

$$Y_{jt} = \delta_{0j} + \lambda_{3j} Y_{jt-1} - \delta_{2j} X_{jt} + \varepsilon_{3j} \quad (9)$$

$$V_{jt} = X_{jt}/Y_{jt} \quad (10)$$

$$X_{jt}/Y_{jt} \leq 1 \quad (11)$$

$N_{jt}$: number of firms of the j subsector of the Polish transport industry for the period t,

$X_{jt}$: inputs of firms in the subsector j of the Polish transport industry for the period t,

$Y_{jt}$: outputs of firms in the subsector j of the Polish transport industry for the period t,

$V_{jt}$: level of technology of firms in the subsector j of the Polish transport industry for the period t,

$\varepsilon_j$: common random term in the j subsectors of the transport Polish industry.

In the above model, we deal with four interconnected simultaneous dynamic equations of which one equation forms a deterministic relation. Reasoning trough traditional econometrics, we may set the assumption of a component random error reflecting individual behaviour of each of the 8 subsectors and a correlated random error affecting the whole sector. Thus, each of the three first equations accounts for 18 unknown parameters to be estimated, suggesting 54 parameters for the whole model based on data from one period of time (2018). This expectation model describes a partial adjustment of each of the three equations $N_{jt}$, $X_{jt}$, $Y_{jt}$ of which the expected values $N^{ex}_j$, $X^{ex}_j$, $Y^{ex}_j$ have to be determined through the estimated parameters of the above equation system. The expected number $N^{ex}_j$ (in the steady state) of firms in the different 8 subsectors of the Polish transport industry will depend on the present technological coefficient $V_j$, which explains the relation between input and output. The $X^{ex}_j$ is a response of the present level of output while producers base their future output on the present level of input. Let us recall below basic aspects of a partial adjustment model. Formally, let $y_t *$ be the unknown, targeted level of $y_t$:

$$y_t * = \alpha + \beta_i x_{it} + \varepsilon_t, \text{ t=1..T, i=1..K} \quad (12)$$

And a progressive adjustment equation:

$$y_t - y_{t-1} = (1 - \lambda)(y_t * - y_{t-1}, with (0<\lambda<-1) \quad (13)$$

Solving the second equation for $y_t$ and inserting the first expression for $y_t$, we finally obtain the next equation (Koyck, 1954):

$$y_t = \alpha' + \beta'_i x_{it} + \lambda y_{t-1} + \varepsilon'_t \quad (14)$$

Since this form is linear in parameters and disturbance non-auto correlated (Greene , 2011), the LS estimator will generate consistent and efficient estimates.

Estimated parameters $\alpha'$ and $\beta'_i$ are the short-run multipliers. To obtain the long-run effect, one transforms $\alpha = \alpha'/(1-\lambda)$ and $\beta_i = \beta'_i/(1-\lambda)$. The long-run disturbance estimates become $\varepsilon_t = \varepsilon'_t/(1-\lambda)$. We then retrieve estimates of equation (12) explaining the targeted value of $y_t^*$. Next, for $\lambda$ equal to zero, we may have to deal with a pragmatic agent who prefers to pay the whole attention on the present environment, thereby ignoring information of the past. In the present study, the short-run and long-run effects are estimated and presented below in Tables 2 and 3.

On the epistemological side, the particular advantage of the model is to link the generalized maximum entropy principle with the Bayes optimal information processing rule trough an econometric model embedded in the system as a constraining structure explained as model moments. Finally, Table 1 presents the priors and data used to estimate the model.

**Table 1.** Some key parameters of the Polish current transport industry Subsectors in 2018

| Transport industry Subsectors (Manufacture of) | number of firms($N_j$) | Average gross output/(1000*$N_j$) | Average intermediary Input/1000*$N_j$ | Ratio input-output (V) |
|---|---|---|---|---|
| Manufacture of motor vehicles | 35 | 2002 | 1402 | 0.7 |
| Manufacture of bodies (coachwork) for motor vehicles; manufacture of trailers and semi-trailers | 96 | 77 | 50 | 0.649 |
| Manufacture of parts and accessories for motor vehicles | 323 | 301 | 213 | 0.708 |
| Building of ships and boats | 61 | 68 | 48 | 0.702 |
| Manufacture of railway locomotives and rolling stock | 25 | 322 | 222 | 0.,692 |
| Manufacture of air and spacecraft and related machinery | 41 | 241 | 149 | 0.619 |
| Manufacture of military fighting vehicles | 4 | 238 | 101 | 0.422 |
| Manufacture of transport equipment n.e.c. | 38 | 64 | 38 | 0.599 |

Source: Own based on Statistics Poland data.

## 4. Outputs and discussion

This section presents the outputs of the proposed model. The computations of the NCEE model were carried out with the GAMS (General Algebraic Modelling System) code. Table 2 presents the new post-entropy outputs of the industry subsector steady state optimal configuration resulting from the NCEE model. Given priors presented in Table 1 and the formal model explained in equ. 2-11, we note a long-run potential increase of about 99% of the total number of firms. The new structure represents the expected steady-state configuration of the Polish transport industry Subsectors. Changes between the present (Table 1) and the expected structure is displayed in Table 2. Precisely, this table provides information on the subsector percent changes of the number of firms, inputs and outputs between initial data (inputs) and model outputs (posteriors) explaining future equilibrium firm activity. One can observe the highest number increase rate in the sub-sector of military fighting vehicles (775%) and a slight decrease in the sub-sector of motor vehicles (-3%).

Next, the same table displays the ratio of input-output change (%). It reveals the long-run equilibrium subsector average profitability change rate (%) (or subsector value added change rate), given the initial conditions and the model formulation presented in equ. 2–11. This ratio is obtained as a difference between the output change (%) and the input change (%) for a given subsector. We notice that this ratio seems to decrease in the long-run and this decrease to be globally proportional to sub-sectorial firm concentration, as shown through column 1 and 2 of Table 2.

As far as the model interval confidence is concerned (see Table 2), we observe a global cross-entropy norm I(m) of around 0.368. This index compounds the parameter cross-entropy norm of around 0.391 and the error term index of around 0.125. These two values will depend on the value level of the weight $\lambda$ in the criterion function (equ. 2). The higher value of this parameter tends to increase the parameter precision while worsening the prediction level of the model through the error component. Finally, as the cross-entropy index varies between zero and one, its higher value suggests weaker discrimination of the model (data) against the prior. Its value closer to zero, in the contrary, reveals a higher significance of the model in terms of discriminating against the prior. Table 3 presents estimate values of the model and model parameter inference index value I(m). For instance, based on the global cross-entropy norm displayed in Table 3, one can say that the model has discriminated in favour of the posterior (the proposed model outputs) for approximately 63.2% (1- Global cross entropy norm I(m)). Readers interested in information theory statistical inference can find details on the subject, e.g. in Golan et al. (1996), or for the non-extensive entropy, e.g. in Bwanakare (2014).

Finally, as presented in Table 2, we notice that all the transport subsectors but one have increased their firm concentration ratio, while the highest increase corresponds to the lowest increase in profitability.

**Table 2.** The post-entropy expected steady-state configuration of the Polish transport industry Subsectors

| Sub-sectors of the Polish transport industry | Change of number of firms in % | Ratio input output change (%) | Change of inputs in % | Change of outputs in % |
|---|---|---|---|---|
| Motor vehicles | -3.0 | 2.398 | 10 | 13 |
| Bodies (coachwork) for motor vehicles; trailers and semi-trailers | 96.0 | -20.646 | 45 | 20 |
| Parts and accessories for motor vehicles | 85.0 | -3.123 | 4 | 1 |
| Building of ships and boats | 115.0 | -29.814 | 6 | -18 |
| Railway locomotives and rolling stock | 168 | 1.279 | 21 | 23 |
| Air and spacecraft and related machinery | 134 | -16.139 | 27 | 9 |
| Military fighting vehicles | 775 | -80.679 | 37 | -24 |
| Transport equipment n.e.c. | 137 | -51.218 | 26 | -17 |
| Total number of enterprises of all subsectors | 1239 | | | |
| Average input | 2536 | | | |
| Average output | 3606 | | | |
| q-Tsallis parameter | 1.000 | | | |
| Global cross-entropy norm | 0.368 | | | |
| Parameter cross-entropy norm | 0.391 | | | |

Source: Own work.

**Table 3.** Model parameter estimates and statistical inference

| | $\alpha_{0j}$ | $\lambda_{1j}$ | $\alpha_{2j}$ | $\beta_{0j}$ | $\lambda_{2j}$ | $\beta_{2j}$ | $\delta_{0j}$ | $\lambda_{3j}$ | $\delta_{2j}$ |
|---|---|---|---|---|---|---|---|---|---|
| Estimate values | 0.879 | 0.835 | -0.067 | -0.882 | 0.449 | -0.899 | -0.708 | 0.099 | 0.696 |
| Normed index I(p) | 0.349 | 0.418 | 0.900 | 0.353 | 0.867 | 0.295 | 0.637 | 0.872 | 0.626 |

Source: Own work.

Table 3 presents model system parameter estimates and their normalized statistical precision I(p). As already explained, I(p) close to unity means that prior and posterior parameters are identical, which suggests that the model (new data, in Bayesian interpretation) is no pertinent. In terms of entropic formalism, no entropy reduction is reached through the incorporated econometric model system (equ. 2–11). If we adopt the rule of thumb presented in Golan et al. (1996), all parameters are more or less significant as all precision indices I(p) are lower than 0.99.

## 5. Concluding remarks

The proposed model aimed at predicting the subsector's most plausible, long-run financial configuration of firms consistent with current information on their inputs, outputs and structure through a generalized maximum entropy principle. It consisted of minimizing information divergence between unknown posteriors related to industry subsector main characteristics and corresponding priors and model initial data. This model proposed NCEE as a recent approach for solving complex inverse problems. As revealed through the model outputs, the long-run change of different profitability ratios is diversified while the highest increase in firm concentration corresponds to the lowest increase in profitability. The model could be developed to take into account recent theoretical developments in management and economics. Based on the above outputs, we can now expect a further dynamic development of the transport industry in Poland evidenced by sub sector firm concentration. This phenomenon significantly leads to the increase in firm competition while negatively impacting on different ratios of profitability as reported in this paper.

## References

ABE, S., BAGCI, G. B., (2004). arXiv:cond-mat/0404253.

BORGES, E. P., (2004). Physica A, 334, 255.

BOTTAZZI, G. et al., (2007). *Invariances and Diversities in the Patterns of Industrial Evolution: Some Evidence from Italian Manufacturing Industries*, Springer, Small Business Economics, 29, pp. 137–159.

BWANAKARE, S., (2014). *Entropy*, 16, pp. 2713–2728.

BWANAKARE, S., (2019). *Non-Extensive Entropy Econometrics for Low Frequency Series. National Accounts-Based Inverse Problems*. Berlin, Boston: De Gruyter, Web. Retrieved 8 Mar. 2020, from https://www.degruyter.com/view/product/506062.

CHAMPERNOWNE, D. G., (1953). The Economic Journal, Vol. 63, No. 250, pp. 318–351.

GABAIX, X., (2008). Power Laws in Economics and Finance. Retrieved: http://www.nber.org/papers/ w14299.

GOLAN, A., JUDGE, G. and MILLER, D., (1996). Wiley in Chichester, England.

GOLAN, A., PERLOFF, J. M., (2001). University of California, Berkeley, USA.

GUS, http://old.stat.gov.pl/gus/publikacje_a_z_PLK_HTML.htm.

JAYNES, E. T., (1994). Washington University.

MANTEGNA, R. N., STANLEY, H. E., (1999). Cambridge University Press, Cambridge.

TIKHONOV, A. N., ARSENIN, V. I., (1977). John Wiley & Sons.

VENKATESAN, R. C., PLASTINO, A., (2011). *arXiv:1102.1025v3*.

TSALLIS, C., MENDES, R. S. and PLASTINO, A. R., (1998). Physica A., Statistical Mechanics and its Applications, North-Holland.

TSALLIS, C., (2009). Springer, Berlin.

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 193–206, DOI 10.21307/stattrans-2020-062*
Received – 08.09.2020; accepted – 07.11.2020

# Flow management system for maximising business revenue and profitability

**Piotr Zawada**[1], **Włodzimierz Okrasa**[2], **Jack Warchalowski**[3]

## ABSTRACT

Most for-profit organisations must constantly improve their business strategies and approaches to remain competitive. Many of them choose to embark on Lean or Six Sigma journeys with the intention of maximising productivity and increasing sales. Despite a significant progress in the development of the Big 3 Improvement Methodologies (Lean, Six Sigma, Theory of Constraints – TOC), many manufacturers are still involved in ineffective operations, resulting in longer-than-desired lead times, late deliveries, high inventories and considerable operational costs. All of these business errors seriously challenge the company's competitiveness. The aim of the paper is to demonstrate the importance of effective analysis of maintaining the appropriate level of inventory in gaining a competitive advantage of the company using the company's key resources in the competitive struggle on the market while conducting continuous reporting of reasons for not achieving the assumed business goals, and using the principles of the economy of bandwidth in order to maximize the profitability.

**Key words:** inventory, improvement of profitability, economy, management.

## 1. Introduction

In order to stay competitive and to maximize productivity while increasing sales, many organizations need to continuously improve using innovative approaches, such as Lean or Six Sigma journeys (Mason et al., 2015; de Freitas J., 2017). Sometimes their efforts do not bring the expected results and consume a lot of time and money (Babiceanu and Seker, 2016). Moreover, according to a recent survey, 74% of companies claim to be adopting Lean Thinking Methodology but only 24% claim any kind of positive results. Proponents of this approach believe that one of the most effective way to improve manufacturing business revenue and profitability is to

[1] Cardinal Stefan Wyszynski University in Warsaw, Poland. E-mail: piotrek.za@poczta.fm. ORCID: https://orcid.org/0000-0003-2817-9578.
[2] Cardinal Stefan Wyszynski University in Warsaw and Statistics Poland, Poland. E-mail: W.Okrasa@stat.gov.pl. ORCID: https://orcid.org/0000-0001-6443-480X.
[3] CMS Montera Canada, Canada. E-mail: jack.warchalowski@cmsmontera.com.

implement a Flow Management System (FMS) approach. This approach utilizes all 3 improvement methodologies focused by TOC.

The FMS consists of four key components:

1) define inventory position and levels and create a pull-based replenishment signal,
2) identify production streams in the plant, schedule only key resources and reinforce schedule attainment as the primary measure,
3) drive plant-wide continuous improvement process based on the main reasons the schedule is not achieved,
4) base key market and product profitability decisions on the Throughput Economics approach.

Each of the constitutive elements of this approach is discussed in the following sections, together with an indication of its suitability to the problem stated in the title. That is, how it might work for maximizing business revenue and profitability. The structure of the article is as follows. The next section characterizes the first component, devoted to the issue of assuring a balance between sales goals, production plans and the storage time of components taking into account the customer needs. The specification of main positive consequences of the FSM implementation in this context concludes this section. The third section discusses strategic aspects of the production streams and key resources within the schedule attainment reinforcement as the primary measure, along with the issue of adequacy and compatibility of the activities undertaken in the area of production and the required competences of human capital. Drum-Buffer-Rope (DBR) approach – a production planning and execution methodology – which is an integral part of FMS, makes it possible to implement a Continuous Improvement process in the plant. The next section continues the above issues based on a supposition that the main reasons the objectives of DDR have not been achieved can be identified – a Continuous Improvement process that uses Pareto Diagrams comprises reasons hindering the flow through the plant. In the fifth section, the **co**st-per-unit – the most popular analysis process and paradigm of traditional business decision making - which allows managers to use the concept of gross margin to evaluate business opportunities is put under critical review due to its potential distortions and shortcomings. Therefore, the use of a Throughput Economics (TE) based approach, as a part of the FMS approach, taking into account relative product and market profitability, is being recommended along with empirical evidence for its support. The last section summarizes positive effects of using the four-component FMS approach, with focus on the improvement in operational and financial performance of the organizations implementing it.

***Define inventory positions and levels and create a pull-based replenishment signal***

Stable production systems must, in their assumptions, answer the questions of how to build a balance between sales goals, production plans and the storage time of components necessary for implementation, in many cases very unstable customer orders. This purpose is served by the described system of a competitive advantage use, based on the identification of customer needs and matching production processes with the highest degree of security in the implementation of serial production. FMS focuses first on defining all inventory requirements, utilizing a TOC-based Demand Driven Replenishment (DDR) sizing algorithm, to set up targets for key Finished Goods, Raw Materials and Sub-Assembly items. These inventory buffers break supply chain dependence between unreliable supplier deliveries, variable customer demand and the plant, providing significant stability for the manufacturing operation. Once inventory buffers are in place, a pull-based replenishment signal, in combination with other customer demand, creates the basis for generating the plant load. More stable plant load creates larger production butch for key resources, minimizes their set up requirements, increases overall plant throughput and often reduces manpower. In addition, improved ability to more often make to stock vs. to variable customer demand increase finished goods availability, improves customer service levels and leads to increase in sales. Overall, DDR results in a significant positive impact on business profitability by often reducing operational expenses and driving sales increase at the same time. In addition, DDR, most of the time, results in lower overall inventory levels and / or increased inventory turns. For many years, make-to-order (MTO) has been the preferred approach for manufacturers to use to determine when and in what quantity to make their products. In addition, to help manufacturers buy their required raw materials, they relied on Materials Requirement Planning (MRP) systems.

While the appeal of MTO seems obvious, only make the required quantity of a product once the customer has ordered, the negative side effects are numerous. First, in many manufacturing environments, the customers' order is often their best guess of what they need (their own forecast). Too often customers change either the quantity or the due date of the order. These order changes often force manufacturers to expedite and / or create excess finished goods inventory. Most MTO manufacturers store higher than desired levels of MTO inventory. Second, following an MTO approach often results in periods of high demand (in excess of capacity) and low demand, making it challenging to properly utilize the plant workforce. Both of these issues lead to increased costs through excessive overtime, expediting and even quality mistakes. Finally, following an MTO strategy extends lead time as the product needs to be manufactured after the order is received. This lead time is extended even more when the manufacturer has a backlog of orders. And of course, longer lead times lead to more order changes – a self-reinforcing negative loop.

The appeal of MRP system support is also easy to understand. Explode the customer orders through the Bill of Materials (BOM) and buy only the amount needed to make the customer orders. The allure of minimal inventory and high raw material availability makes the idea of MRP very compelling. However, reality is often very different. Many manufacturers buy items with lead times greater than the lead time they must offer their customers. This issue forces manufactures to feed their MRP systems with forecasted orders. Since it is impossible to accurately forecast at the individual SKU level, the demand changes inflicted on the plant, as the actual orders vary from the forecasted orders, leads to material shortages, expediting, "stealing" and overstock.

MRP is not only trying to help buyers bring in the precise quantity of material, it is also trying to do that at just the right time – not too early or too late. As a result, MRP limits a manufacturer's flexibility. If the customer order changes or a supplier is late with delivering raw materials, manufacturers are often unable to pivot and build something else – as the materials needed to change the schedule are also not available yet. Quality problems in the plant only magnify these issues. While it is true that MRP systems often provide functionality for safety stock and/or min/max inventory level management, these levels are rarely maintained frequently enough to reflect the current often highly variable environment – leading to too much safety stock of some items and not enough of others.

The primary reason that using MTO and MRP leads to all sorts of problems is that manufacturing environments are characterized by high amounts of variability. Variability in the sales orders (dates and quantities), supplier performance (dates, quantities and quality), bill of material and inventory accuracy, and production performance (dates, quantities and quality). MTO and MRP assume low levels of variability and increase a manufacturer's dependence on good, stable performance. As most manufacturers' environments are far from being stable, MTO and MRP too often fail.

FMS, utilizing a DDR approach, minimizes system dependence by positioning inventory in key supply chain points (i.e. Raw Materials, Finished Goods, customer locations, etc.), provides better protection from on-going disturbances, monitors sources of system variability and allows the entire system function at a higher performance level.

## 2. Identify production streams in the plant, schedule only key resources and reinforce schedule attainment as the primary measure

At the beginning of this part of the study, it is worth considering how necessary it is to ask the following question: "Is it worth to use the strategy to limit your resources to the level which is the most difficult or the most expensive to obtain?". Perhaps, the most important thing is to answer the question "Are the activities undertaken in the area of production accompanied by the required competences to our human capital?"

In the further part of the study, these analyses will be accompanied by the presentation of a solution that minimizes the effects of incorrect production planning. Every plant can be divided into production flow streams. Drum-Buffer-Rope (DBR), a TOC production planning and execution methodology, is used to schedule each production flow stream within the plant and ensures timely production execution. Then, while measuring schedule attainment of each critical resource in a production stream, the reasons and plant locations that most often hinder the flow are tracked and recorded.

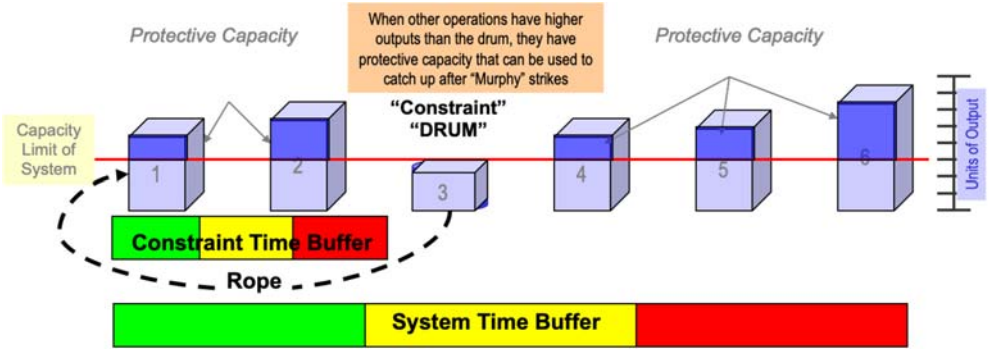



**Figure 1.** Drum-Buffer-Rope System

Source: Own work.

### *The Drum is usually the Constraint – for every flow stream in the plant*

The drum/constraint is usually the machine restraining your overall throughput. Most of the operations have one constraint (machine or department) for every flow stream in the plant, but sometimes in some plants the drum can follow the product mix changes. In more of a continuous flow operation the constraint is usually located in one place and does not move often with a product mix.

By definition, a constraint can be any resource with customer demand larger than its effective capacity. For every hour lost on this constraint we lose an hour for the entire operation. By the same token, gaining an hour of output at the constraint, increases output for the entire plant. It is also important to realize that every time we elevate the output of the constraint above the capacity level of another resource the constraint/drum will be moved to another plant location. Usually, this is not a preferable direction since it creates an immediate need to redesign the entire production planning and scheduling process as well as manpower management in the plant.

The constraint is called a Drum because it creates the pace for a given flow stream. The speed of the flow stream or its production rate is equal to the throughput of its drum resource. The book "The Goal" by Eli Goldratt was the first one describing this concept. The production schedule is normally set for the drum resource and made

visible to the rest of the flow stream. Schedule attainment is closely monitored and evaluated on the shift by shift basis.

From the continuous improvement perspective, improvements of non-bottlenecks have no effect on the overall plant output. Attempts to utilize non-bottlenecks to a 100% capacity (and often above 80%) drive WIP (Work in Process) up, reduce overall system throughput, make the entire system unstable and cause constraint to move from place to place – a wondering bottleneck phenomenon.

The wondering bottleneck phenomenon can be observed most often in the plants with balanced capacity. Balanced capacity means that available effective capacity at every production resource in a flow stream is closely matched to a market demand and each other. The DBR diagram shown in Figure 1 above would show balanced capacity if all work centres from 1 to 6 had the capacity of 3 units. In some plants, with highly variable product mix, balanced production line decreases potential throughput and increases costs – contrary to its original intent.

One of the most strategic question, for any manufacturing operation, that needs to be answered is – where should we strategically position our constraint resource? Sometimes, before resolving this dilemma, it helps to define where we do not want the drum/constraint to be positioned. By definition, non-bottlenecks must have excess capacity. Normally, we will need at least 20% excess capacity at non-constraints to keep the system stable. This is called a PROTECTIVE capacity. Anything above is Excess, and often should be eliminated (good use of Lean Manufacturing techniques to deal with "Muda").

Market availability of resources to acquire, their price, strategic fit are some of the factors that may help you answer the question where and where not to locate your constraint or non-constraint. You certainly do not want your protective capacity to be difficult to find or expensive to buy. Therefore, most probably you want the drum resource to be the more expensive one to get or the hardest to find. You quickly realize that it is the resource that represents your core competence or the reason why you built your business.

Unlike a commonplace definition of inventory buffer, the DBR system Buffer is articulated in the units of time. DBR system buffer is the amount of work expressed in time (hours, days, etc.) before the constraint work centre. The rope mechanism controls the amount of work released to the flow stream by choking its introduction according to the buffer size. By collecting a buffer of work to in front of the constraint machine, we can guarantee the constraint does not starve and never stops. In any given flow stream, the drum is the only machine where maximized (100%) utilization is desirable and beneficial to the system performance.

Buffer's main objective is to mitigate variability of the system. In a traditional Drum Buffer Rope system there are 2 buffers – one protecting the constraint and one for the entire system (flow stream). The role of constraint buffer (before the constraint) is to protect the constraint itself while the system buffer protects the shipping/due date.

All buffers (time and inventory) are divided into three main zones. Colours red, yellow and green designate main buffer sections. In the FMS system two more colours indicate a stock out (black) and too much inventory (blue) - above and beyond its target size indicated by top of the green zone.

Mentioned earlier, the target 20% protective capacity is just a starting point. In order to define what is the optimum level of protective capacity, you need to collect time buffer statistics data. When the buffer's red zone gets penetrated more than 5% to 10% of the time, you will need to create additional protective capacity at least at one if not more non-constraint resources. In case you do not experience any red zone penetrations, your level of protective capacity is most probably excessive, and you can successfully reduce the buffer size (its duration) generating new improvement opportunities. In general, the more variability in the system, the more protective capacity you need. Applying Six Sigma and Lean Manufacturing (Raj and Attri, 2010) techniques can greatly help create process stability within the DBR framework (Mithun et al., 20202; Albliwi , 2014; Alhuraish, 2017).

Figure 1 above shows the first buffer (constraint time buffer) buffering the Drum/Constraint from the variability of upstream resources. Generating this buffer statistics will help size capacity requirements of resources 1 and 2. Recording daily reasons for buffer penetration into red will enable you to determine which machine will need additional protective capacity.

The second buffer (system buffer) is buffering the shipping schedule (due date). The main reason for the system buffer is to absorb the overall system variability – especially after the drum resource. Any order commit date is always an estimate and often wrong because of embedded system variability. Therefore, we need a mechanism that will allow us to mitigate variability impact and provide ability to determine capacity requirements for all flow stream resources. The system buffer allows us to accomplish these objectives.

Accomplishing the above objectives is critical especially from the perspective of ensuring protective capacity and avoiding the wondering bottleneck phenomenon. Not being able to successfully manage this sometimes delicate balance will lower your system overall Throughput.

Required protective capacity could be gained by applying lean manufacturing tools like set-up reduction techniques, using Statistical Process Control (SPC) to control process variability, staggering lunch and other breaks, creating cross functional/trained production teams or by simply buying more capacity if necessary. However, in order to maximize business profitability, creating protective capacity where needed and increasing effective drum capacity without spending money is preferred.

DBR approach is an integral part of FMS and is a prerequisite to enable a Continuous Improvement process in the plant.

## 3. Drive continuous improvement process based on the main reasons the Drum schedule is not achieved

One of the main assumptions made in this study is that there is a need to develop a universal method thanks to which it would be possible to design production processes in such a way that they are carried out in the most effective manner taking into account the high profitability rate of the type of conducted activity. FMS creates a Continuous Improvement process that uses Pareto Diagrams comprised of reasons hindering the flow through the plant. It prioritizes plant-wide improvement opportunities and reduces system variability in a quick and systematic way (Figure 2 below). The Flow Issue Reporting (FIR) Pareto contains all reported reasons why the schedule attainment was not possible on every scheduled shift. The issues may include mechanical breakdowns, but also shortages, quality, longer than expected set-ups, absenteeism, material handling, etc. It is critical that FIR process is clearly communicated, enforced and reviewed at the end of every shift across the plant.

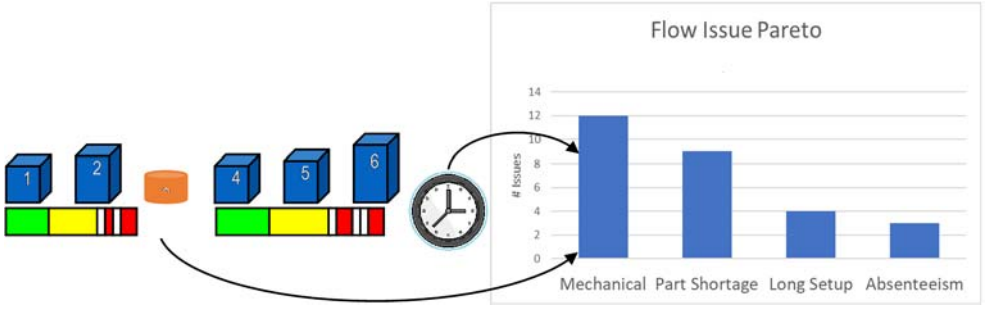



**Figure 2.** Flow Issue Reporting Process

Source: Own work.

Once the FIR process is in place and improvement opportunities are known, Lean Thinking and Six Sigma principles and tools are used to remove obstacles and create operational improvements (Antony and Banuelas, 2002; Costa et al., 2017). Continuous Improvement team (Kaizen team) meets on a regular basis (often weekly) and decides when and where Lean Thinking and 6 Sigma tools are applied based on the Pareto information. Based on FIR driven priorities, plant performance drastically improves, throughput goes up, service levels increase, and productivity and revenue are maximized.

Once plant performance is stabilized, by breaking dependence (inventory buffers) and having FIR based continuous improvement process in place to reduce process and flow variability, the business is in a much better position to turn its improvement focus towards increasing business profitability through changing product profitability decisions.

## 4. Base key market, customer and product profitability decisions on the Throughput Economics approach

Cost-per-unit, the world's most popular analysis process, is a devastating and flawed paradigm of traditional business decision-making. Regardless, many organizations still attempt to align their understanding of profitable markets/products with their manufacturing operation's performance using this approach. The cost-per-unit approach supports a simple process for decision-making as it allows managers to use the concept of gross margin or contribution margin to evaluate business opportunities. That is what makes it very popular. However, many managers are aware of the potential distortions and shortcomings of the cost-per-unit approach.

As an example of product profitability decisions consider a company that produces only 2 products: A and B (shown below – Figure 3). It is a 24hr operation, with a labour cost of $10/hr and Operating Expenses of $5,000 per week.

- Product A is produced from Raw Materials costing $14 per unit. This product must be processed on Machine 1, Machine 3 and a Final Assembly operation at the rates specified below. Product A is sold at a price of $50 each and its demand is 100 units per week.
- Product B is produced from Raw Materials costing $12 per unit and requires Machine 1, Machine 2 and the Final Assembly at the rates also specified below. Product B is sold for $60 each and its demand is 50 units per week.

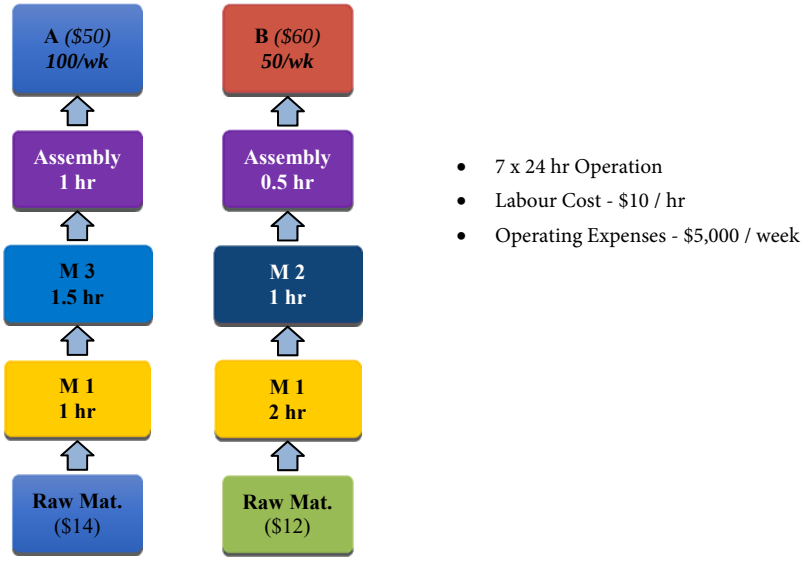



**Figure 3.** Product Profitability Example

Source: Own work

**Table 1.** Profit Margin Analysis

| | A | B |
|---|---|---|
| Price ($) | 50.0 | 60.0 |
| Material ($) | 14.0 | 12.0 |
| Production Time (hr) | 3.5 | 3.5 |
| Labour Cost / hr | 10.0 | 10.0 |
| | | |
| Material Margin ($) | 36.0 | 48.0 |
| Production Cost ($) | 35.0 | 35.0 |
| Profit Margin ($) | 1.0 | 13.0 |

Source: Own work.

Which product makes more money for the company? Profit Margin (Sales less Material Costs less Labour) Analysis shown in Table 1 above clearly demonstrates that Product B with a profit margin of $13/unit is the most profitable product *(overhead allocations were omitted to simplify the discussion).*

However, in order to judge how to maximize profitability of this company, we first have to decide which product to prioritize in production, since we do not have enough machine capacity to satisfy the market demand for both products (168 hours available per week in a 7x24 operation vs. 200 hours of M 1 required to produce both A and B – see Table 2 below).

**Table 2.** Machine Capacity Analysis

| | Demand (pcs) | M 1 (hr) | M 2 (hr) | M 3 (hr) | Assembly (hr) | Total (hr) |
|---|---|---|---|---|---|---|
| A | 100 | 100 | | 150 | 100 | 350 |
| B | 50 | 100 | 50 | | 25 | 175 |
| Total | | 200 | 50 | 150 | 125 | 525 |

Source: Own work.

The business logic suggests that in order to maximize business profitability we should first produce the product with the highest Profit Margin (Product B) and then use the remaining capacity for the other product (Product A). In order to demonstrate the profit impact on the company overall, we need to calculate the net profit associated with producing all B and some A. Since the demand for B is 50 units per week, we need 100 hours (50 x 2 hours) of production capacity for B, leaving only 68 hours open for Product A. This scenario leads to the company generating $4,848 of Material Margin ($ Throughput) every week as demonstrated in the Table 3 below.

**Table 3.** Material Margin Analysis (prioritize B)

| | Demand (pcs) | M 1 (hr) | M 2 (hr) | M 3 (hr) | Assembly (hr) | Price/pc ($) | Revenue ($) | Material ($) | Throughput ($) |
|---|---|---|---|---|---|---|---|---|---|
| A | 68 | 68 | | 102 | 68 | 50.0 | 3,400.0 | 952.0 | 2,448.0 |
| B | 50 | 100 | 50 | | 25 | 60.0 | 3,000.0 | 600.0 | 2,400.0 |
| Total | | 168 | 50 | 102 | 93 | | 6,400.0 | 1,552.0 | 4,848.0 |

Source: Own work.

Considering that Operating Expenses for the company are $5,000 per week the resulting business Net Profit is negative $152. This means that maximizing production of the highest Profit Margin Product (B) and satisfying its full demand of 50 units per week will lead to a weekly **profit loss of $152**.

Another interesting question in front of us is to find out what the business profitability would look like if we decided to prioritize product A – the product with a substantially lower profit margin. Under this scenario we can satisfy the entire demand for product A (100 pcs) and dedicate the rest of M1 capacity to product B production. Based on 68 hrs. available we can only produce 34 pcs of product B (2hrs per piece on M1). This scenario leads to the company generating $5,232 of Material Margin every week as demonstrated in the Table 4 below.

**Table 4**. Material Margin Analysis (prioritize A)

| | Demand (pcs) | M 1 (hr) | M 2 (hr) | M 3 (hr) | Assembly (hr) | Price/pc ($) | Revenue ($) | Material ($) | Throughput ($) |
|---|---|---|---|---|---|---|---|---|---|
| A | 100 | 100 | | 150 | 100 | 50 zł | 5,000 zł | 1,400 zł | 3,600 zł |
| B | 34 | 68 | 34 | | 17 | 60 zł | 2,040 zł | 408 zł | 1,632 zł |
| Total | | 168 | 34 | 150 | 117 | | 7,040 zł | 1,808 zł | 5,232 zł |

Source: Own work.

Considering that Operating Expenses for the company are $5,000 per week the resulting business Net Profit is positive $232. This means that maximizing production of the lowest Profit Margin Product (A) and satisfying its full demand of 100 units per week will lead to a weekly **profit gain of $232.**

How is this possible? Did not the decision to focus on the product with a lower profit / contribution margin just lead to the company generating more net profit? What is going on here? In this example, we have decided to challenge the widely held belief that contribution margin of a product is the best indication of a company's profitability. In most situations it is not. In fact, ***Contribution and/or Profit Margin is a totally arbitrary and completely misleading indicator.***

This conclusion has very large implications on several important sales and marketing decisions, such as: which markets to focus on, which business to accept, how to develop new products that maximize profit, and with which customers to further develop long term relationships.

The profitability of a product, and its associated impact on the net profit of a business, cannot and should not be measured using profit margin. Therefore, what is an acceptable substitute?

FMS uses an alternative approach to understand relative product and market profitability – a Throughput Economics (TE) based approach. Using the same example, we can clearly establish that Machine 1 sets the pace for the entire operation and should be considered a critical resource/drum (i.e. it has the least capacity). Profitability is best determined by calculating the rate of dollar contribution of each product on this critical resource as explained in the DBR section. This is measured by taking the difference between a product's sales price and its totally variable cost (mainly raw materials) and dividing it by the production rate on the critical resource – Table 5 below.

**Table 5**. Throughput Velocity

| | A | B |
|---|---|---|
| Price ($) | 50.0 | 60.0 |
| Material ($) | 14.0 | 12.0 |
| Throughout ($) | 36.0 | 48.0 |
| Drum production time (hr) | 1.0 | 2.0 |
| Throughout Velocity ($/hr) | 36.0 | 24.0 |
| | | |
| Profit Margin ($) | 1.0 | 13.0 |

Source: Own work.

In all instances, this measurement of product profitability is perfectly aligned with a business' net profit – higher TV for Product A and higher net profit impact.

The TE-based approach with its Throughput Velocity (TV) indicator, has significant implications on plant performance, market focus, pricing evaluation and new product development strategies. Manufacturing businesses need to understand their products' Throughput Velocity (TV) if they are to maximize profit in these challenging times. Using profit margin analysis to accept or reject new business will unavoidably lead to too many wrong decisions – allowing your competitors to take more of your business.

Some of the strategic questions the new process answers include:

- Which market segments are the most profitable?
- Which products make the company the most profit?
- Is it truly possible for some products to lose money?
- How should investment and make vs. buy decisions be analysed?
- At what price should we accept an order?
- How to align your operating costs and plant capacity with market demand?
- On what products to focus its R&D effort?

## 5. Summary and conclusions

The presented arguments and examples of comparing the results achievable under 'new' and 'old' approaches, prove the sense of replacing the classic profit margin-based strategy by an approach to profitability based on calculating the rate of dollar contribution of each product on this critical resource, as it was explained in the DBR section. Since Contribution and/or Profit Margin is a totally arbitrary and completely misleading indicator, the new approach which takes into account the difference between a product's sales price and its totally variable cost (and dividing it by the production rate on the critical resource) seems to provide a tool needed to deal with several problems concerning plant performance, market focus, pricing evaluation and new product development strategies.

Using the four key components of FMS, organizations can significantly improve operational and financial performance. Most companies that successfully implement FMS obtain the following benefits:

- Improved flow and reduced operating costs because of their new TOC/Constraints' Management scheduling tools.
- Increased sales from pricing decisions driven by 80/20 TE-based methodology.
- Released working capital by improved inventory turns as a result of DDR.
- Maximized throughput from a stable plant protected from system variability by the DBR-based operations management approach.
- Increased shareholder value.

In addition, financial benefits often include:

- Throughput/Sales increase of 20%-30%.
- Inventory reduction of up to 50%.
- Lead time reduction of approximately 50%.
- On time delivery improvement up to 99%.
- EBITDA percent of sales increase by approximately 10%.

The application of the discussed solution is not free from the costs incurred by the entrepreneur in the initial period of implementation. Nevertheless, the financial effects related to the implementation of the assumptions of the production management system referred to in the article are disproportionately high compared to the expenditure incurred, which has been repeatedly checked in implementation in business in Poland, Canada and many other countries around the world.

## References

ANTONY, J., BANUELAS, R., (2002). Key ingredients for the effective implementation of Six Sigma program, Measuring Business Excellence, 6(4).

ALBLIWI, S., ANTONY, J., LIM, S., VAN DER WIELE, T., (2014). Critical failure factors of lean Six Sigma: A systematic literature review, International Journal of Quality and Reliability Management, 31(9).

ALHURAISH, I., ROBLEDO, CH., KOBI, A., (2017). A comparative exploration of lean manufacturing and six sigma in terms of their critical success factors, Journal of Cleaner Production 164/2017.

BABICEANU, R., SEKER, R., (2016). Big Data and virtualization for manufacturing cyberphysical systems: a survey of the current status and future outlook, Computers in Industry 81/2016.

COSTA, T., SILVA, F. J. G., PINTO FERREIRA, L., (2017). Improve the extrusion process in tire production using Six Sigma methodology, Procedia Manufacturing 13/2017.

DE FREITAS, J., COSTA, H., FERRAZ, F., (2017). Impacts of Lean Six Sigma over organizational sustainability: A survey study, Journal of Cleaner Production 156/2017.

MASON, S., NICOLAY, C., DARZI, A., (2015). The use of Lean and Six Sigma methodologies in surgery: A systematic review, The Surgeon 13/2015,

MITHUN ALI, S., HOSSEN, A., MAHTAB, Z., KABIR, G., KUMAR, S., ADNAN, Z., (2020). Barriers to lean six sigma implementation in the supply chain: An ISM model, Computers & Industrial Engineering 149/2020.

RAJ, T., ATTRI, R., (2010). Quantifying barriers to implementing Total Quality Management (TQM), European Journal of Industrial Engineering 4(3)/2010.

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 207–211,*

# About the Authors

**Adebola Femi Barnabas**, PhD, is a Fellow at Royal Statistical Society (RSS), UK. He is currently an Associate Professor of Statistics at the Department of Statistics, Federal University of Technology, Akure, where he served as the pioneer Postgraduate Coordinator and later Head of Department. His major research interest includes sample survey design and its applications, population and health studies. He has graduated many Masters' degree and PhD students. He has over 30 articles published both at national and international peer-review journals and conferences to his credit. He is currently the Head of Sample Survey and its Applications Research Team.

**Adediran Adetola Adedamola** is an assistant lecturer at the Department of Statistics, Federal University of Technology, Akure, Nigeria. Her research interest lies in survey sampling techniques, design and analysis of experiments. She is an emerging scholar, who has published both in national and international peer-review journals and conferences.

**Agiwal Varun** is working as a statistician and lecturer in the Department of Community Medicine, Jawaharlal Medical College, Ajmer, India. He has published 15 research papers in national and international peer-reviewed journals. His main area of interest includes linear and non-linear time series, distribution theory and Bayesian inference.

**Almetwally Ehab M.** is an assistant lecturer at the Faculty of Business Administration, Delta University for Science and Technology. He is a PhD student at the faculty of graduate studies for statistical research, Cairo University, Egypt. He earned an MSc degree in Statistics in 2019 from the faculty of graduate studies for statistical research – Cairo University. He got a bachelor of statistics in 2016 from the faculty of commerce at Zagazig University. He has published over 40 research papers in international/national journals and conferences. His research interests are in the areas of distributions, Bayesian statistics, statistical inference, bivariate distributions, censoring samples, ranked set samples and R statistical package.

**Arshad Rana Muhammad Imran** is currently an Assistant Professor in the Department of Statistics, Govt. S.E College Bahawalpur, Pakistan. He has recently received a PhD from IUB under the supervision of Dr. M. H. Tahir. He has eight publications in his credit.

**Bwanakare Second** is an Associate Professor at the faculty of Management of the University of Information Technology and Management in Rzeszow, Poland. He is an associate researcher at the RIME lab (University of Lille) and a consultant in methodology and analysis at the Statistical Office in Rzeszow. He has belonged to the Corpus of international experts at the Polish Accreditation Committee since 2013. He is also a member of ISI and FENS (Poland). His research interests focus on econometrics and statistics and, in particular, the non-extensive entropy econometrics approach and its generalization to non-linear complex systems. He is the author of two monographs and many articles in international journals.

**Chesneau Christophe** is currently an Assistant Professor in the Department of Mathematics, LMNO, University of Caen Normandie, France. He received PhD in the field of applied mathematics with a speciality in statistics, at LPMA, University Paris 6, France. He is working in the areas of statistical inference, nonparametric statistics, integer-valued time series and data analysis. He has over one hundred publications in his credit.

**Cierpiał-Wolan Marek** is an Assistant Professor at the Department of Quantitative Methods, Institute of Economics and Finance, University of Rzeszow. Moreover, he is the Director of the Statistical Office in Rzeszow. Editor-in-Chief of the scientific journal The Polish Statistician. Author of about 90 national and international publications including journals and monographs. Manager and expert in international research projects. Organizer and active participant of numerous scientific conferences.

**Domański Czesław** is a Full Professor at the Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz. His research interests are: tests based on runs theory and order statistics, (multivariate) normality tests and non-classical methods of statistical inference. Currently, he is a member of the Scientific Statistical Council of the President of Statistics Poland, main council of the Polish Statistical Association and Committee on Statistics and Econometrics at the Polish Academy of Sciences.

**Duda Jarosław** is an Assistant Professor at the Institute of Computer Science, Faculty of Mathematics and Computer Science, Jagiellonian University in Cracow. He has education in computer science (PhD), physics (PhD) and mathematics (MSc). He is known from introducing Asymmetric Numeral Systems, which are used in products of e.g. Apple, Facebook and Google. His main areas of interest include information theory, statistical modelling and machine learning.

**Ewemooje Olusegun Sunday**, PhD, is a lecturer at the Department of Statistics, Federal University of Technology, Akure, Nigeria, and also a Graduate Statistician at Royal Statistical Society (RSS), UK. He is currently a Statistical Consultant; Project and Client Manager with Statistics Without Borders (SWB). He was a Postdoctoral Research

Fellow at the Population and Health Research Entity, North-West University (Mafikeng Campus), South Africa. His research interest includes sample survey designs and applications, health and population studies, demography and environmental statistics. He has published many articles both in national and international peer-review journals and conferences.

**Gurgul Henryk** is a Professor at the AGH University of Science and Technology in Krakow. His research is focused on financial econometrics, economic growth, macroeconomics and multisectoral input-output models. He has been a visiting professor or lecturer at universities in Austria, Finland, Germany, Italy, Spain and Slovenia. He is an active member of editorial boards of international journals: Central European Journal of Operations Research, Managing Global Transiotion and international scientific societies, including the International Statistical Institute (Elected Member). Professor Gurgul is a scientific secretary of the Economic Commission of the Kraków Branch of the Polish Academy of Arts and Sciences (PAU) and a member of the Presidium of Commission for Economics and Statistics of the Krakow Branch of the Polish Academy of Sciences (PAN).

**Haj Ahmad Hanan A.** received her PhD degree in 2009 from University of Jordan, in Mathematical Statistics. She worked as an Assistant Professor in the Mathematics Department in Hail University in Saudi Arabia from 2009 until 2019. Now, she works in King Faisal University in Basic science department (Saudi Arabia). Her research interests are in the distribution theory, goodness of fit, statistical inference and censoring samples. She did several research papers and participated in virtual conferences and recently has been acting as a visiting professor at the Diponegoro University in Indonesia.

**Jamal Farrukh** is currently an Assistant Professor in the Department of Statistics, the Islamia University of Bahawalpur (IUB), Punjab. He worked as a lecturer in Government S.A. postgraduate College in 2012 to 2020, and a statistical officer in the Agriculture Department from 2007 to 2012. He received MSc and MPhil degrees in Statistics from the Islamia University of Bahawalpur (IU), Pakistan in 2003 and 2006. He received a PhD from IUB under the supervision of Dr. M. H. Tahir. He has 118 publications in his credit.

**Kumar Jitendra** is working as an Associate Professor in the Department of Statistics, Central University of Rajasthan, Bandersindri, Ajmer, India. Before joining Central University of Rajasthan, involved in CUB, Patna, IDRBT, Hyderabad and SHIAU, Allahabad. His area of interest is time series, outliers, crime statistics, policy process reengineering and big data.

**Niwitpong Suparat** is an Associate Professor at the Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok,

Thailand. Her main areas of interest include: confidence interval, Bayesian method and prediction interval of parameter of interest.

**Okrasa Włodzimierz** is a Professor and a Head of the Research Methods and Evaluation Unit at the Institute of Sociology, Cardinal Stefan Wyszynski University (UKSW). He serves as an Advisor to the President of Statistics Poland and an Editor-in-Chief of the Statistics in Transition new series. He was teaching and researching in Polish and American universities and was an ASA Senior Research Fellow at the US Bureau of Labor Statistics (1990–1991), a Program Director in the Social Science Research Council, NY (1991–1993), and then worked for the World Bank in Washington, DC (1994–2000), analysing poverty and implementing new household surveys in several 'countries in transition'. He next headed the Social Sciences Unit at the European Science Foundation (2000–2003, in Strasbourg). Elected member of the International Statistical Institute. Author of numerous publications, including books and articles in reputed international journals.

**Prasad Shakti** is an Assistant Professor at the Department of Basic and Applied Science (Maths) in National Institute of Technology, Arunachal Pradesh, India. His research interests are sample surveys, statistical inference and missing data analysis. He has published over twenty research papers in the reputed international/national journals and conferences.

**Shangodoyin Dahud Kehinde**, is a Professor at the Department of Statistics, University of Botswan. His area of interest is data mining in health, population, education and agriculture, econometrics, Bayesian modelling, multivariate analysis and time series analysis.

**Syrek Robert** (PhD) is an Assistant Professor at the Institute of Economics, Finance and Management, Faculty of Management and Social Communication at the Jagiellonian University in Cracow. His main areas of interest include: time series analysis and forecasting, financial econometrics and modelling the dependence structures of financial time series (especially using copula functions).

**Szczepocki Piotr** is an Assistant Professor at the Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz. His main areas of interest include Sequential Monte Carlo methods and volatility modelling.

**Tahir Muhammad Hussain** is currently a Professor of Statistics at the Islamia University of Bahawalpur (IUB), Pakistan. He received MSc and PhD in Statistics in 1990 and 2010 from IUB. He has been teaching in the Department of Statistics (IUB) since 1992. His current research interests include generalized classes of distributions and their special models, compounded and cure rate models. Dr. Tahir has supervised over 60 MPhil students, supervising five PhD students and has over 70 international publications in his credit.

**Thangjai Warisa** is a lecturer at the Department of Statistics, Faculty of Science, Ramkhamhaeng University, Thailand. She is working on confidence intervals for parameter of interest.

**Warchalowski Jack** is the CEO of CMS Montera, a North American technology company, helping manufacturers improve operations and inventory by providing TOC software and contract services. Prior to CMS, he was the head of operations for a high-tech manufacturer, Ernst & Young management consultant, and a project engineer with Babcock & Wilcox. He holds an MBA degree from the Wilfrid Laurier University, Canada and a Bachelor of Applied Science in Mechanical Engineering from the University of Waterloo, Canada. He is an active member and a frequent speaker in many industry associations.

**Wywiał Janusz L.** is a Full Professor of Economics with a specialization in Statistics and Econometrics. He is the head of the Department of Statistics, Econometrics and Mathematics in University of Economics in Katowice. Professor Wywiał's interests are focused on survey sampling, sampling designs and schemes, testing statistical hypotheses, applications of statistics in financial auditing and econometric prediction. Professor Wywiał has published over 110 research papers in national and international journals. He has also published eight monographs and nine textbooks. He is a member of scientific professional bodies.

**Zawada Piotr** is a University Professor of Sociology. His research interests are social economy, innovation management, organizational sociology, human resource management. Professor Zawada has published 68 research papers in international/national journals and conferences. He has also published five books/monographs. Professor Zawada is an active member of cluster organizations. He is also a PARP . He is professionally associated with the Rzeszów Regional Development Agency, where he carries out international and social economy projects

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 213–216,*

# Acknowledgements to Reviewers

The Editor and Editorial Board of the Statistics in Transition new series wish to thank the following persons who served from 1 December 2019 to 30 November 2020 as peer-reviewers of manuscripts for the Statistics in Transition new series – Volume 21, Numbers 1–5, including the Special Issue Number 4; the authors` work has benefited from their feedback.

**Abdullah Norli Anida,** University of Malaya, Malaysia

**Adegoke Nurudeen,** Federal University of Technology Akure, Nigeria

**Alkasadi Najla,** University of Aden, Yemen

**Almongy Hisham Mohamed,** Mansoura University, Egypt,

**Andrzejczak Karol,** Poznan University of Technology, Poland

**Bayoud Husam,** Fahad Bin Sultan University–Tabuk, Saudi Arabia

**Beaumont Jean-François,** Statistics Canada, Canada

**Beck Krzysztof,** Lazarski University, Poland

**Bhoughal Sunil,** University of Jammu, India

**Bouza Carlos,** Universidad de La Habana, Cuba

**Böhning Dankmar,** Southampton Statistical Sciences Research Institute, University of Southampton, United Kingdom

**Chaudhuri Arijit,** Indian Statistical Institute, Kolkata, India

**Chaudhuri Sanjay,** National University of Singapore, Singapore

**Chesneau Christophe,** University of Caen, France

**Chudziak Jacek,** University of Rzeszów, Poland

**de Waal Ton,** Tilburg University, The Netherlands

**Dehnel Grażyna,** University of Economics and Business, Poland

**Dihidar Kajal,** Indian Statistical Institute, Kolkata, India

**Domański Czesław,** University of Lodz, Poland

**Donehower Gretchen,** University of California at Berkeley, the USA

**Drechsler Jörg,** Institute for Employment Research in Nürnberg, Germany

**Echaust Krzysztof,** Poznań University of Economics and Business, Poland

**Eftekharian Abbas,** University of Hormozgan, Iran

**Ercan İlker,** Uludag University, Turkey

**Filip Dariusz,** University of Cardinal Stefan Wyszyński in Warsaw, Poland

**Gadde Srinivasa Rao,** School of Mathematical Sciences, The University of Dodoma, Tanzania

**Giusti Catherina,** University of Pisa, Italy

**Ganczarek-Gamrot Alicja,** University of Economics in Katowice, Poland

**Gerasymenko Sergiy,** University in Poznań & WSB University in Chorzów, Poland

**Gurgul Henryk,** AGH University of Science and Technology, Poland

**Hadaś-Dyduch Monika,** University of Economics in Katowice, Poland

**Hall Michael,** University of New Zealand, New Zealand

**Hušková Marie,** Charles University in Prague, Czech Republic

**Jajuga Krzysztof,** University of Economics, Wroclaw, Poland

**Jamal Farrukh,** Govt. S.A Postgraduate College Dera Nawab Sahib, Pakistan

**Jankiewicz Jacek,** Poznan University of Economics and Bussines, Poland

**Jibrin Sanusi Alhaji,** Kano University of Science and Technology, Nigeria

**Jurek Witold,** University of Technology, Poland

**Kalton Graham,** Westat, USA

**Khan Jahangir Sabbir,** Aligarh Muslim University, India

**Klein Ingo,** Friedrich-Alexander-University of Erlangen-Nürnberg, Germany

**Koç Haydar,** Çankırı Karatekin University, Turkey

**Kogut-Jaworska Małgorzata,** University of Szczecin, Poland

**Komornicki Tomasz,** Polish Academy of Sciences, Poland

**Kończak Grzegorz,** University of Economics in Katowice, Poland

**Kordalska Aleksandra,** Gdansk University of Technology, Poland

**Kovtun Natalia,** Taras Shevchenko National University of Kyiv, Ukraine

**Kowalczyk Barbara,** Warsaw School of Economics, Poland

**Kozłowski Arkadiusz,** University of Gdansk, Poland

**Krzyśko Mirosław,** The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Poland

**Kubacki Jan,** Statistical Office in Łódź, Poland

**Souza Luciano,** Universidade Federal Rural de Pernambuco, Brazil

**Spreeuw Jaap,** Cass Business School, City, University of London, United Kingdom

**Szreder Mirosław,** University of Gdansk, Poland

**Szymkowiak Marcin,** Poznań University of Economics and Business, Poland

**Tahir Muhammad Hussain,** The Islamia University of Bahawalpur, Pakistan

**Tarczyński Waldemar,** University of Szczecin, Poland

**Traat Imbi,** University of Tartu, Estonia,

**Vance Eric,** University of Colorado Boulder, USA

**Veen Duco,** Utrecht University, The Netherlands

**Voskoboynikov Ilya B.,** National Research University Higher School of Economics, Russia

**Wywiał Janusz,** University of Economics in Katowice, Poland

**Yousof Haitham,** Benha University, Egyp

**Zamanezade Ehsan,** University of Isahan, Iran

**Zeghdoudi Halim,** Badji-Mokhtar University, Algieria

**Zeman Kryštof,** Austrian Academy of Sciences, Austria

*STATISTICS IN TRANSITION new series, December 2020*
*Vol. 21, No. 5, pp. 217–222,*

# Index of Authors, Volume 21, 2020

# GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

## Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page*: http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) 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).