Volume 21, Number 3, September 2020 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

EDITOR

Włodzimierz Okrasa, *University of Cardinal Stefan Wyszyński, Warsaw and Statistics Poland, e-mail: w.okrasa@stat.gov.pl; phone number +48 22 — 608 30 66*

ASSOCIATE EDITORS

Arup Banerji	*The World Bank, Washington, USA*	Ralf Münnich	*University of Trier, Germany*
Mischa V. Belkindas	*Open Data Watch, Washington D.C., USA*	Oleksandr H. Osaulenko	*National Academy of Statistics, Accounting and Audit, Kiev, Ukraine*
Sanjay Chaudhuri	*National University of Singapore, Singapore*	Viera Pacáková	*University of Pardubice, Czech Republic*
Eugeniusz Gatnar	*National Bank of Poland, Poland*	Tomasz Panek	*Warsaw School of Economics, Poland*
Krzysztof Jajuga	*Wrocław University of Economics, Wrocław, Poland*	Mirosław Pawlak	*University of Manitoba, Winnipeg, Canada*
Marianna Kotzeva	*EC, Eurostat, Luxembourg*	Mirosław Szreder	*University of Gdańsk, Poland*
Marcin Kozak	*University of Information Technology and Management in Rzeszów, Poland*	Imbi Traat	*University of Tartu, Estonia*
Danute Krapavickaite	*Institute of Mathematics and Informatics, Vilnius, Lithuania*	Vijay Verma	*Siena University, Siena, Italy*
Janis Lapiņš	*Statistics Department, Bank of Latvia, Riga, Latvia*	Vergil Voineagu	*National Commission for Statistics, Bucharest, Romania*
Risto Lehtonen	*University of Helsinki, Finland*	Gabriella Vukovich	*Hungarian Central Statistical Office, Hungary*
Achille Lemmi	*Siena University, Siena, Italy*	Jacek Wesołowski	*Central Statistical Office of Poland, and Warsaw University of Technology, Warsaw, Poland*
Andrzej Młodak	*Statistical Office Poznań, Poland*	Guillaume Wunsch	*Université Catholique de Louvain, Louvain-la-Neuve, Belgium*
Colm A, O'Muircheartaigh	*University of Chicago, Chicago, USA*	Zhanjun Xing	*Shandong University, China*

EDITORIAL BOARD

Dominik Rozkrut (Co-Chairman) *Statistics Poland, Poland*
Waldemar Tarczyński (Co-Chairman) *University of Szczecin, Poland*
Czesław Domański *University of Łódź, Poland*
Malay Ghosh *University of Florida, USA*
Graham Kalton *Westat, USA*
Mirosław Krzyśko *Adam Mickiewicz University in Poznań, Poland*
Partha Lahiri *University of Maryland, USA*
Danny Pfeffermann *Central Bureau of Statistics, Israel*
Carl-Erik Särndal *Statistics Sweden, Sweden*
Janusz L. Wywiał *University of Economics in Katowice, Poland*

FOUNDER/FORMER EDITOR

Jan Kordos *Warsaw School of Economics, Poland*

EDITORIAL OFFICE **ISSN 1234-7655**

Scientific Secretary
Marek Cierpiał-Wolan, *Statistical Office in Rzeszów, Poland, e-mail: m.cierpial-wolan@stat.gov.pl*
Secretary
Patryk Barszcz, *Statistics Poland, Poland, e-mail: p.barszcz@stat.gov.pl, phone number +48 22 — 608 33 66*
Technical Assistant
Rajmund Litkowiec, *Statistical Office in 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, September 2020
Vol. 21, No. 3, pp. III–VI

From the Editor

A compilation of eleven articles by twenty-four authors from nine countries demonstrates the journal's growing variety, at least from a geographic point of view. Diversity is also a feature of the content of this volume, addressing a variety of methodological and substantive issues, ranging from estimation, statistical distributions and data modelling to econometric analyses and statistical characteristics of the national economic performance.

This issue opens with **Jiani Yin's** and **Balgobin Nandram's** article ***A Bayesian Small Area Model with Dirichlet Processes on the Responses***. The authors begin by observing that typically survey data have responses with gaps, outliers and ties, and the distributions of the responses might be skewed. Usually, in small area estimation, predictive inference is done using a two-stage Bayesian model with normality at both levels (responses and area means). This is the Scott-Smith (S-S) model and it may not be robust against these features. Another model that can be used to provide a more robust structure is the two-stage Dirichlet process mixture (DPM) model, which has independent normal distributions on the responses, which, however does not accommodate gaps, outliers and ties in the survey data directly. This is the problem tackled in this paper using a two-stage non-parametric Bayesian model with several independent Dirichlet processes. This model has a Gaussian (normal) distribution on the area means, and is called the DPG model. Therefore, the DPM model and the DPG model are essentially the opposite of each other and they are both different from the S-S model. Of the three models, the DPG model turns out to be the best one for accommodating the features of the survey data. For Bayesian predictive inference, we need to integrate two data sets, one with the responses and other with area sizes. The body mass index application - which is integrated with the census data - and the simulation study used to compare the three models (S-S, DPM, DPG) showed that the DPG model may be preferred.

Mirosław Krzyśko and **Łukasz Smaga** in the article ***Measuring and Testing Mutual Dependence of Multivariate Functional Data*** consider new measures of mutual dependence between multiple multivariate random processes representing multidimensional functional data. In the case of two processes, the extension of functional distance correlation is used by selecting appropriate weight function in the weighted distance between characteristic functions of joint and marginal distributions. For multiple random processes, two measures are sums of squared measures for

pairwise dependence. The dependence measures are zero if and only if the random processes are mutually independent. This property is used to construct permutation tests for mutual independence of random processes. The finite sample properties of these tests are investigated in simulation studies. The use of the tests and the results of simulation studies are illustrated with an example based on real data.

In the paper ***Detection of Outliers in Univariate Circular Data by Means of the Outlier Local Factor*** by **Ali H. Abuzaid** discussed is the problem of outlier detection in univariate circular data, which become an object of increased interest over the last decade. New numerical and graphical methods were developed for samples from different circular probability distributions. The main drawback of the existing methods is, however, that they are distribution-based and ignore the problem of multiple outliers. The local outlier factor (LOF) is a density-based method for detecting outliers in multivariate data and it depends on the local density of every k nearest neighbours. The aim of this paper is to extend the application of the LOF to the detection of possible outliers in circular samples, where the angles of circular data are represented in two Cartesian coordinates and treated as bivariate data. The performance of the LOF is compared against other existing numerical methods by means of a simulation based on the power of a test and the proportion of correct detection. The LOF performance is compatible with the best existing discordancy tests while outperforming other tests. The level of the LOF performance is directly related to the contamination and concentration parameters while having an inverse relationship with the sample size. In order to illustrate the process, the LOF and other existing discordancy tests are applied to detect possible outliers in two common real circular data sets.

Rama Shanker and **Kamlesh Kumar Shukla** in the paper ***A New Quasi Sujatha Distribution*** propose a new quasi Sujatha distribution (NQSD), of which the following are particular cases: the Sujatha distribution devised by Shanker (2016), the size-biased Lindley distribution, and the exponential distribution. Its moments and moments-based measures are derived and discussed. Statistical properties, including the hazard rate and mean residual life functions, stochastic ordering, mean deviations, Bonferroni and Lorenz curves and stress-strength reliability are also analysed. The method of moments and the method of maximum likelihood estimations are discussed for estimating parameters of the proposed distribution. A numerical example is presented to test its goodness of fit, which is then compared with other one-parameter and two-parameter lifetime distributions.

The article ***Power Size-Biased Two-Parameter Akash Distribution*** by **Khaldoon Alhyasat, Ibrahim Kamarulzaman, Amer Ibrahim Al-Omari and Mohd Aftar Abu Bakar** presents the two-parameter Akash distribution generalized to size-biased two-parameter Akash distribution (SBTPAD). A further modification to SBTPAD is introduced, creating the power size-biased two-parameter Akash distribution

(PSBTPAD). Several statistical properties of PSBTPAD distribution are proved. These properties include the following: moments, coefficient of variation, coefficient of skewness, coefficient of kurtosis, the maximum likelihood estimation of the distribution parameters, and finally order statistics. Moreover, plots of the density and distribution functions of PSBTPAD are presented and a reliability analysis is considered. The Rényi entropy of PSBTPAD is proved and the application of real data is discussed.

Amal S. Hassan, Salwa M. Assar, Ahmed M. Abdelghaffar in the article ***Statistical Properties and Estimation of Power-Transmuted Inverse Rayleigh Distribution*** constructed a three-parameter continuous distribution using a power transformation related to the Transmuted Inverse Rayleigh (TIR) distribution. A comprehensive account of the statistical properties is provided, including the following: the quantile function, moments, incomplete moments, mean residual life function and Rényi entropy. Three classical procedures for estimating population parameters are analysed. A simulation study is provided to compare the performance of different estimates. Finally, a real data application is used to illustrate the usefulness of the recommended distribution in modelling real data.

In the article ***Generalised Odd Frechet Family of Distributions: Properties and Applications*** by **Shahdie Marganpoor, Vahid Ranjbar, Morad Alizadeh,** and **Kamel Abdollahnezhad** a new distribution called Generalized Odd Fréchet (GOF) distribution is presented and its properties explored. Some structural properties of the proposed distribution, including the shapes of the hazard rate function, moments, conditional moments, moment generating function, skewness, and kurtosis are presented. Mean deviations, Lorenz and Bonferroni curves, Rényi entropy, and the distribution of order statistics are given. The maximum likelihood estimation technique is used to estimate the model parameters. Finally, applications of the model to a real data set is presented to illustrate the usefulness of the proposed distribution.

Oleksandr H. Osaulenko, Taisiia Bondaruk, and **Liudmyla Momotiuk** in the paper ***Ukraine's State Regulation of the Economic Development of Territories in the Context of Budgetary Decentralisation*** are discussing the theoretical and methodological foundations of Ukraine's state legislation regulating the economic development of territories in the context of budget decentralization. They also describe the transformation of the public administration system necessitated by the above-mentioned phenomenon. The authors reflect also on the basic methods by which the state can regulate the activity of local self-government bodies: the legislative regulation and the administrative regulation, which provides rules and instructions that determine the relations between central and local authorities. They conduct a systematic analysis of state regulations which support the local self-governments' activity focusing especially on those problems that have not been solved yet during the ongoing reform.

Also, statistical estimations of the phenomena relating to the process of producing state legislation regulating the economic development of territories in the context of budgetary decentralization are provided.

Marta Marszałek's paper ***The Unobserved Economy – Invisible Production in Households. The Household Production Satellite Account and the National Time Transfer Accounts*** starts with observation that not only monetary value or economic products create welfare, but non-monetary components should also be included in the System of National Accounts. Although household production is registered in official statistics, the main part of it (nearly 75-80 percent) of the total home production remains beyond GDP. The Household Production Satellite Account (HHSA) is a macroeconomic analysis covering both market and non-market home production. The National Time Transfer Accounts (NTTA) is, next to HHSA, an analysis aimed to register and observe the directions of transfers and to present the recipients and givers of home production. Regular estimations provided by the HHSA and NTTA may prove to be a valuable supporting tool to national accounts, pension systems, or social policy as they provide a great deal of macroeconomic information regarding households, their economic and living conditions and well-being.

In the paper by **Bilal Ahmed Para** and **Tariq Rashid Jan,** ***Poisson Weighted Ishita Distribution: A Model for the Analysis of Over-Dispersed Medical Count Data*** a new over-dispersed discrete probability model is introduced by compounding the Poisson distribution with the weighted Ishita distribution. The statistical properties of the newly introduced distribution have been derived and discussed. Parameter estimation has been done with the application of the maximum likelihood method of estimation, followed by the Monte Carlo simulation procedure to examine the suitability of the ML estimators. In order to verify the applicability of the proposed distribution, a real-life set of data from the medical field has been analysed for modelling a count data set representing epileptic seizure counts.

The section Research Communicate contains just one paper by **Sebastian Wójcik,** entitled ***Through a Random Route to the Goal: Theoretical Background and Application of the Method in Tourism Surveying in Poland***. This paper is motivated by the shortcomings of traditional methods of surveying small or rare population and the lack of mathematical foundation of some recently available approaches. The author proposes estimators of parameters related to Random Route Sampling (RRS), along with their basic properties. A formula for the Horvitz-Thompson estimator weights is given and a case of a tourism-related survey conducted in Poland is discussed.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 3, pp. VII

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, September 2020
Vol. 21, No. 3, pp. IX–X

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, September 2020
Vol. 21, No. 3, pp. 1–19, DOI 10.21307/stattrans-2020-041
Received – 30.01.2020; accepted – 03.08.2020

A Bayesian Small Area Model with Dirichlet Processes on the Responses

Jiani Yin[1], **Balgobin Nandram**[2]

ABSTRACT

Typically survey data have responses with gaps, outliers and ties, and the distributions of the responses might be skewed. Usually, in small area estimation, predictive inference is done using a two-stage Bayesian model with normality at both levels (responses and area means). This is the Scott-Smith (S-S) model and it may not be robust against these features. Another model that can be used to provide a more robust structure is the two-stage Dirichlet process mixture (DPM) model, which has independent normal distributions on the responses and a single Dirichlet process on the area means. However, this model does not accommodate gaps, outliers and ties in the survey data directly. Because this DPM model has a normal distribution on the responses, it is unlikely to be realized in practice, and this is the problem we tackle in this paper. Therefore, we propose a two-stage non-parametric Bayesian model with several independent Dirichlet processes at the first stage that represents the data, thereby accommodating some of the difficulties with survey data and permitting a more robust predictive inference. This model has a Gaussian (normal) distribution on the area means, and so we call it the DPG model. Therefore, the DPM model and the DPG model are essentially the opposite of each other and they are both different from the S-S model. Among the three models, the DPG model gives us the best head-start to accommodate the features of the survey data. For Bayesian predictive inference, we need to integrate two data sets, one with the responses and other with area sizes. An application on body mass index, which is integrated with census data, and a simulation study are used to compare the three models (S-S, DPM, DPG); we show that the DPG model might be preferred.

Key words: Bayesian computation, bootstrap, predictive inference, robust modeling, computational and model diagnostics, survey data.

1. Introduction

There are many methods in the current statistical literature for making inferences based on samples selected from a finite population. The most widely used approach is design-based inference, which is nonparametric but requires large sample sizes. Model-based inference for survey sampling has been proposed as an alternative to the design-based theory, and this is particularly useful for small area estimation (Rao and Molina 2015) when there are sparse data from many areas. We consider the simplest version of a small area model, and we show how to robustify it to fit survey responses with gaps, outliers and ties.

[1]Takeda Pharmaceuticals. USA. E-mail: jianiyin@gmail.com. ORCID: https://orcid.org/0000-0002-5007-2833.
[2]Worcester Polytechnic Institute. USA. E-mail: balnan@wpi.edu. ORCID: https://orcid.org/0000-0002-3204-0301.

Figure 1: Dot plots of body mass index (BMI) for thirty-five areas (counties)

Generally, in a unit-level model, the responses from each area might have a distribution with a mean and a variance. The variance is usually taken constant over areas, but the mean varies over the areas. Sometimes each mean is written as global constant plus a random effect, different over areas. The area random effects or means share a common distribution allowing a borrowing of strength adaptively across areas (sample sizes are generally different). Complete pooling is generally a bad idea, because there is usually heterogeneity across areas. A degree of heterogeneity can be accommodated using covariates, but while useful covariates are particularly important in any analysis, this is not enough because there will

still be heterogeneity across areas. So there is a need to model area means, for example, to provide a small area model. Here, we consider continuous responses from a number of small areas using a unit-level model.

Our application is on body mass index (BMI), a continuous variable used to measure lifestyle. (Because of how survey data are collected, the BMI data can be discrete and there will be gaps, outliers and ties.) We use data from the 35 largest counties (areas) with at least 500,000 people from the third National Health and Nutrition Examination Survey (NHANES III), a survey conducted during the period October 1988 through September 1994. We do not have access to data from other smaller counties. In fact, these are the BMI data from the same 35 counties we analyzed in Nandram and Choi (2005, 2010) and in other places too numerous to mention. However, we have used data for adults who are older than 20 years because these data have very few nonresponse, rather than for children younger than 19 years, because our current study is not about nonresponse. We use BMI data where the BMI values are given up to the first decimal place. Dot plots of the data from 35 counties are shown in Figure 1. There are three things we observe in these data. First, there are ties because several adults have the same BMI values. This is clear because an adult BMI value is some value from about 18.0 kgm^{-2} to about 40.0 kgm^{-2} (one decimal place). Second, there are gaps (i.e. no BMI values between two adjacent values) and this is especially true in the right extreme areas of the dot plots. Third, there are outliers, which occur mostly in the right tails of the dot plots, thereby showing some right skewness with outliers. Therefore, it is clear that these BMI values do not follow normal distributions; a kernel density estimator will hide these features in the data. The data have natural gaps (e.g. there are no values in between 20.1 and 20.2) and ties (e.g. several values at say 20.1); these will exist in the population as well. This is why we model the gaps, outliers and ties in these data. We note that there are some demographic variables such as age, race and sex, which we do not study here, but we discuss in the concluding section how to incorporate covariates into our models.

Our goal is to predict the finite population mean, 85^{th} (overweight) and 95^{th} (obese) finite population percentiles of BMI for all eligible adults from each county. The sample from each area is at least about 100; we have a small area problem because these sample sizes are about a 0.01% of the population. Our problem is how to take care of the gaps, outliers and ties in the BMI data. To this end, we use two-stage Bayesian models with one model having a component that addresses directly these non-standard features in the responses. To do Bayesian predictive inference for the finite population quantities, we also need a data set with the population sizes of the areas (counties). To achieve this end, we integrate the NHANES BMI data with the population counts from the US 1990 Census.

Let y_{ij} denote the value for the j^{th} unit within the i^{th} area, $i = 1,\ldots,\ell, j = 1,\ldots,N_i$. Throughout, we assume that y_{ij}, $i = 1,\ldots,\ell, j = 1,\ldots,n_i$, are the samples from i^{th} area and are observed, and $y_{ij}, j = n_i+1,\ldots,N_i$ are not observed. Inference is required for the finite population mean or a finite population percentile. For example, the finite population mean of the i^{th} area is $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i, i = 1,\ldots,\ell$. We use Bayesian predictive inference that requires specification of parametric distributions. Moreover, to help protect against posterior impropriety, we use non-informative (vague) independent priors, which are proper, for all hyper-parameters. Specifically, we have used Cauchy priors for location parameters

(e.g. Gelman, Jakulin, Pittau and Su 2008) and shrinkage priors for non-negative parameters (e.g. Nandram and Yin 2016 a, b and the references therein).

Scott and Smith (1969) introduced the basic two-stage model for cluster sampling, but the same model has been used for small areas. The difference is that in small area estimation, we are interested in inference about the population of each small area, but in cluster sampling, we are interested in all sub-populations combined into a single one. Nandram, Toto and Choi (2011) has given a Bayesian analysis of this model. However, the use of models raises the question of the robustness of the inference to possible model mis-specification. Again, in particular, survey data tend to have gaps, outliers and ties and we need to remedy this defect. A generalization to include covariates is the model of Battese, Harter and Fuller (1988) in non-Bayesian survey sampling; for a full Bayesian formulation, see Toto and Nandram (2010) and Molina, Nandram and Rao (2014); but this is not our key issue here.

The Bayesian Scott-Smith (S-S) model, as formulated by Nandram, Toto and Choi (2011), is

$$y_{ij}|\mu_i,\sigma^2 \overset{ind}{\sim} N\left(\mu_i,\sigma^2\right), \quad j=1,\ldots,N_i, \tag{1}$$

$$\mu_i \mid \theta,\sigma^2,\rho \overset{ind}{\sim} N\left(\theta,\frac{\rho}{1-\rho}\sigma^2\right), \quad i=1,\ldots,\ell, \tag{2}$$

$$\pi(\theta,\sigma^2,\rho) = \frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}, \tag{3}$$

where $-\infty<\theta<\infty$, $\sigma^2>0$, $0\le\rho\le 1$. Here, ρ is the intra-cluster correlation. It is worth noting that we have taken $\rho\sim\text{Uniform}(0,1)$, θ to have a standard Cauchy distribution and σ^2 to have a shrinkage distribution (i.e. $f(2,2)$ distribution), all independent. Here, we have used vague proper priors on all parameters.

Suppose we have written $\mu_i \mid \theta,\delta^2 \overset{ind}{\sim} \text{Normal}(\theta,\delta^2)$ and define $\rho=\delta^2/(\delta^2+\sigma^2)$, then we will get $\delta^2=\frac{\rho}{1-\rho}\sigma^2$. Clearly, $0\le\rho\le 1$ and this makes one variance component bounded instead of two unbounded ones, σ^2 and δ^2. This simplifies the computations by permitting a random sampler, which requires no monitoring, rather than a Gibbs sampler, which requires monitoring; see Appendix A.

Another standard model that relaxes some parametric assumptions is the Dirichlet process mixture (DPM) model,

$$y_{ij}|\mu_i,\sigma^2 \overset{ind}{\sim} \text{Normal}(\mu_i,\sigma^2), \quad j=1,\ldots,N_i, \tag{4}$$

$$\mu_i|G \sim G, \quad i=1,\ldots,\ell,$$

$$G \mid \theta,\sigma^2,\gamma,\rho \sim \text{DP}\left\{\gamma,\text{Normal}(\theta,\frac{\rho}{1-\rho}\sigma^2)\right\}, \tag{5}$$

$$\pi(\theta,\sigma^2,\gamma,\rho) = \frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}\frac{1}{(1+\gamma)^2}, \tag{6}$$

where $-\infty<\theta<\infty$, $\sigma^2>0,\gamma>0$, $0\le\rho\le 1$, and γ is called the concentration parameter; see Ferguson (1973) for a definition of the Dirichlet process (DP) and Lo (1984), who extended the DP to DPM. Here, in this formulation the S-S model is a baseline model;

the DPM model is centred on the S-S model and γ controls how close DPM model gets to the S-S model. Here, G is a random distribution function, discrete with probability one, and had distribution $DP(\cdot,\cdot)$. Escobar and West (1995) proposed a simple (not necessarily efficient) algorithm by integrating out the random distribution function in the model. Kalli, Griffin and Walker (2011) suggested slice-efficient samplers, an improved slice sampling scheme that we use in our work, and it is based on the stick-breaking construction of Sethuraman (1994); see Appendix B. Nandram and Choi (2004) and Polettini (2017) have applications on small area estimation, but they did not use the slice-efficient sampler of Kalli, Griffin and Walker (2011).

However, this DPM model does not address our main concern. It does not model the responses non-parametrically to take care of gaps, outliers and ties in the survey data in general, not just BMI data. It models ties among the μ_i, thereby clustering the μ_i. Indeed, this is the strength of the Dirichlet process prior. In reality, we want to do the opposite. That is, we want to have independent Dirichlet processes on the responses and possibly a normal distribution on the random effects. This is the key issue we address in this paper, and we will call this model the DPG model (G refers to the normal assumption on the μ_i). However, the DPM model gives a good sense of how to proceed to meet our requirement.

The plan of the rest of the paper is as follows. In Section 2, we discuss the DPG model with independent Dirichlet processes on the responses. In Section 2.1, we discuss the methodology and inferences. In Section 2.2, we discuss the prediction for a finite population quantity using a data integration. In Section 3, we compare the three models (S-S, DPM, DPG). Specifically, in Section 3.1, we discuss an illustrative example on the body mass index (BMI) data and in Section 3.2 a small simulation study. In Section 4, we present our conclusion and two important extensions.

2. DPG Model, Computations and Prediction

In this section, we describe the DPG model that has independent Dirichlet processes on the responses and a normal distribution on the area means. This robustifies the S-S model in the opposite direction to the DPM, our novel contribution. In Section 2.1, we describe the DPG model, in Section 2.2, we describe how to draw samples from it, and in Section 2.3, we show how to do the prediction.

2.1. DPG Model

Using DPs in the first level and a parametric distribution as prior gives us,

$$\begin{aligned} y_{ij}|G_i &\overset{ind}{\sim} G_i, \quad j=1,\ldots,N_i, \\ G_i|\mu_i,\alpha_i,\sigma^2 &\overset{ind}{\sim} \mathrm{DP}\{\alpha_i,\mathrm{Normal}(\mu_i,\sigma^2)\}, \quad i=1,\ldots,\ell, \\ \mu_i|\rho i.e.\theta,\sigma^2,\rho &\overset{iid}{\sim} \mathrm{Normal}(\theta,\frac{\rho}{1-\rho}\sigma^2). \end{aligned} \tag{7}$$

A full Bayesian model can be obtained by adding prior distributions. We use proper non-informative priors,

$$\pi(\alpha_i) = \frac{1}{(\alpha_i+1)^2}, \qquad \alpha_i > 0, \quad i = 1,\ldots,\ell, \tag{8}$$

$$\pi(\theta,\sigma^2,\rho) = \frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}, \quad -\infty<\theta<\infty, 0<\sigma^2<\infty, 0\le\rho\le 1, \tag{9}$$

with independence. Here, (7), (8) and (9) define the DPG model. Note that the concentration parameters α_i are not included in the S-S model or the DPM model. It is not sensible to assume that the α_i are identically distributed, because they can be very different.

We give a brief comparison of the three models and how they are related. The S-S model is a special case of the DPG model and the DPM model, and both are centred on the S-S model. This occurs when the α_i are large for the DPG model and when γ is large for the DPM model. The DPM model is actually the opposite of the DPG model with normal distribution for the data in each area and a DP prior on the area means. In the DPG model, each area has a distinct DP (i.e. ℓ DPs with different μ_i and α_i) and there is pooling across areas because the μ_i share an effect and σ^2 is common.

We look at the sampling process for the DPG model. When we integrate out the random probability measure (Blackwell and MacQueen, 1973), we get

$$f(\underset{\sim}{y}_i \mid \mu_i,\sigma^2,\alpha_i) = \frac{1}{\sigma}\phi(\frac{y_{i1}-\mu_i}{\sigma})$$

$$\times\prod_{k=2}^{n_i}\left\{\frac{k-1}{\alpha_i+k-1}\frac{\sum_{j=1}^{k-1}\delta_{y_{ij}}(y_{ik})}{k-1}+\frac{\alpha_i}{\alpha_i+k-1}\frac{1}{\sigma}\phi(\frac{y_{ik}-\mu_i}{\sigma})\right\}, \tag{10}$$

where $\delta_a(y)$ means that y is a point mass at a and $\phi(.)$ is the standard normal density. Therefore, in each area we are mixing the distributions in (10) using normal mixing distributions in the DPG model. The DPM is different being a Dirichlet process mixture of normals. The DPM model actually produces ties among the random effects (clustering), its major strength, but it does not model gaps, outliers, ties and possibly skewness among the responses. By putting DPs on the responses in different areas, we are actually taking a head-start on the data, because they accommodate the gaps, ties and outliers in the data; see Figure 1. It is important to note that $\delta_{y_{ij}}(y_{ik})$ is a statement that for each i, y_{ik} is a point mass at $y_{ij}, j = 1,\ldots,k-1$. That is, for the i^{th} area, y_{ik} can be the same as y_{ij} with nonzero probability and this is crucial in our new model. Therefore, equation (10) is the key to how we attempt to accommodate gaps, outliers and ties, particularly ties, in the data. The DPG model is attractive even if there are a few ties (or no ties at all) because the data may have heavy tails where the normal distribution is not appropriate (true for the BMI data).

2.2. Computations

Letting $\underset{\sim}{\psi} = \{\underset{\sim}{\mu}, \theta, \sigma^2, \rho\}$ and $\underset{\sim}{\alpha} = \{\alpha_1, \ldots, \alpha_\ell\}$, it is easy to get a sample from the joint posterior density of $(\underset{\sim}{\psi}, \underset{\sim}{\alpha})$, and therefore inference under the DPG model can be easily performed.

The posterior densities of the α_i are independent of the other parameters $\underset{\sim}{\psi}$ in the model, conditioning on only the distinct values. Let k_i denote the number of distinct values for each area in the observed data, $\underset{\sim}{k} = \{k_i, i = 1, \ldots, \ell\}$ be the vector of k_i, $y^*_{i1}, \ldots, y^*_{ik_i}$ be the k_i distinct sample values for each i and $\underset{\sim}{y}^* = \{y^*_{i1}, \ldots, y^*_{ik_i}, i = 1, \ldots, \ell\}$ be the vector of y^*_{ij}. Thus, the joint posterior density is

$$\pi(\underset{\sim}{\alpha}, \underset{\sim}{\psi} \mid \underset{\sim}{k}, \underset{\sim}{y}^*) = \left[\prod_{i=1}^{\ell} \pi(\alpha_i \mid k_i)\right] \pi(\underset{\sim}{\psi} \mid \underset{\sim}{y}^*), \tag{11}$$

where $\pi(\alpha_i|k_i) \propto \pi(k_i \mid \alpha_i)\pi(\alpha_i)$. For the parameters $\underset{\sim}{\psi}$, we have

$$\begin{aligned} y^*_{ij}|\mu_i &\overset{ind}{\sim} N\left(\mu_i, \sigma^2\right), \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, k_i, \\ \mu_i &\overset{iid}{\sim} N\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \\ \pi(\theta, \sigma^2, \rho) &= \frac{1}{\pi(1+\theta^2)} \frac{1}{(1+\sigma^2)^2}, -\infty < \theta < \infty, 0 < \sigma^2 < \infty, 0 \le \rho \le 1. \end{aligned} \tag{12}$$

Therefore, the algorithm for the DPG model is

Step 1 : For each i, $i = 1, \ldots, \ell$, draw α_i from $\pi(\alpha_i|k_i) \propto \alpha^{k_i} \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i+n_i)} \frac{1}{(\alpha_i+1)^2}$; see Antoniak (1974).

Step 2: Draw $\underset{\sim}{\psi}$ from the parametric model (12), which is easy to fit; see Appendix A for the S-S model.

Step 1 is easily realized using the grid method (Nandam and Yin 2016 a,b). Step 2 is accomplished using a random sampler together with the sampling importance resampling (SIR) algorithm. Therefore, samples can be drawn from the DPG model using a random sampler rather than a Gibbs sampler (as in the DPM, Markov chain samplers need monitoring).

2.3. Prediction for the Finite Population

We have a simple random sample of size n_i from a finite population of size N_i, $i = 1, \ldots, \ell$. Let $y_{i1}, \ldots, y_{in_i}$ denote the sampled values. We want to predict $y_{in_i+1}, \ldots, y_{iN_i}$, the nonsampled values, and obtain the predictive distribution and the prediction interval for any finite population quantity (e.g. $\bar{Y}_i$ for the i^{th} area). Prediction under the S-S model and the DPM model is straightforward.

For the DPG model, the sampling process is

$$\begin{aligned} y_{ij}|G_i &\overset{ind}{\sim} G_i, \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \\ G_i|\mu_i &\overset{ind}{\sim} \mathrm{DP}\{\alpha_i, G_0(\mu_i)\}. \end{aligned}$$

Predictive inference for the DPG model simply uses the generalized Polya urn scheme (Blackwell and MacQueen 1973) for each i, since all areas are independent (see Nandram and Yin 2016 a,b). Once the nonsampled $y_{ij}, j = n_i+1,\ldots,N_i, i = 1,\ldots,\ell$, are obtained, one can now calculate any finite population quantity of interest. Here, we are interested in the finite population mean, the 85^{th} percentile (overweight individuals) and the 95^{th} percentile (obese individuals). The N_i are assumed known, and they can be obtained from a census.

Binder (1982), a very nice paper, frustrated with the bootstrap method (discussed later) that does not produce values different from the sample values, introduced the Dirichlet process into finite population sampling. We note that when prediction is done using any of the three models, including the DPG model, new values different from the samples will be generated. For the S-S model and the DPM model, this will happen with probability one, but for the DPG model with just a positive probability. For the DPG model, because the nonsample values are generated from the generalized Polya urn scheme, values already sampled can be repeated. However, for the DPG model, as the prediction proceeds in an order for a long run (population sizes are large here), the α_i will be dominated, thereby making the process draw more and more values that have already been drawn as in "the rich gets richer scheme".

Letting $f_i = \frac{n_i}{N_i}, i = 1,\ldots,\ell$, denote the sample fractions, the finite population mean is the composite, $\bar{Y}_i = f_i\bar{y}_{i,s} + (1-f_i)\bar{Y}_{i,ns}$, where $\bar{y}_{i,s} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, the mean of the sample values, and $\bar{Y}_{i,ns} = \frac{1}{N_i-n_i}\sum_{j=n_i+1}^{N_i} y_{ij}$, the mean of the non-sample values. To obtain the percentiles, one simply sorts all the data (sample values and predicted non-sample values) in increasing order. Then, for the 85^{th} percentile, pick the value at $.85N_i$ (nearest integer) position, and for the 95^{th} percentile, pick the value at $.95N_i$ (nearest integer) position.

It is worth noting that it is easy to estimate the finite population mean; it is more difficult to estimate the two percentiles because they are in the right tail of the posterior distributions. It is interesting that in finite population mean, the sample mean, $\bar{y}_{i,s}$, is constant a posteriori but $\bar{y}_{i,ns}$ is dynamic (i.e. changes with the iterations). However, when the finite population percentiles are estimated, all the sample values and the predicted values are ordered at each iteration (i.e. the actual positions of the sample values in the ordering will change). Therefore, computation of the finite population percentiles at each iteration takes more time than the finite population mean.

3. Empirical Studies

In this section, we compare the three models (S-S model, DPM model and DPG model). Specifically, in Section 3.1, we describe an application on body mass index (BMI) data, and in Section 3.2, we present a small simulation study.

3.1. Application to Body Mass Index Data

As described in the introduction, we use the example on BMI data for illustration. Since the predictive inference for the overweight and obese population is very important, the heavy tail of the distribution cannot be ignored. Thus, we cannot automatically use the S-S model

nor the DPM model to accommodate the gaps, outliers and ties in the BMI data; see Figure 1. A more robust assumption on the responses, such as the DPG model, needs to be considered.

For the DPM, we ran $10,000$ MCMC iterations, used $5,000$ as a "burn in" and thinned every 5^{th} to obtain $1,000$ converged posterior samples. We have monitored the parameters, σ^2, θ, δ^2 and γ, for the DPM model. The Geweke test of stationarity gives p-values of .483, .414, .459, 0.620 respectively; therefore the iterates pass the test of stationarity. The effective sample sizes are 1000, 1000, 698, 1084 respectively, thereby showing that the iterates form an efficient sample. Numerical summaries such as trace plots and auto-correlation plots (not shown) indicate that the MCMC chains converge and mix well, and a 'random sample' is obtained from the joint posterior density. To get samples from the S-S model and the DPG model, we do not need a Gibbs sampler; a random sampler suffices and monitoring of a Gibbs sampler is not needed.

As a comparison, we also use the Bayesian bootstrap to do prediction in each county individually without borrowing across counties. This will allow us to see how much improvement we can have over direct estimation. Note that for each area (county), all sample sizes are over 100. Here, we describe the Bayesian bootstrap (see Rubin 1981 for more details). Momentarily we consider a single subscript (drop subscript i), so that we have $y_1,\ldots,y_n$ (sample values) from an area and we need to predict $y_{n+1},\ldots,y_N$ (nonsample values), where N is the population size of this area. First, we find the distinct values among $y_1,\ldots,y_n$ and we assume that there are d distinct values, denoted by $y_1^*,\ldots,y_d^*$. Let n_j denote the number of times the j^{th} value occurs in the sample. In the bootstrap it is assumed that only $y_1^*,\ldots,y_d^*$ can occur, and let N_j denote the number of times the j^{th} distinct value occurs in the population; the N_j are unknown. The Bayesian bootstrap has the following model,

$$\underset{\sim}{n} \mid \underset{\sim}{p} \sim \text{Multinomial}(n, \underset{\sim}{p}), \underset{\sim}{p} \sim \text{Dirichlet}(\underset{\sim}{0}),$$

where the improper Haldane's prior is used. Then, the posterior density of $\underset{\sim}{p}$ is

$$\underset{\sim}{p} \mid \underset{\sim}{n} \sim \text{Dirichlet}(\underset{\sim}{n}),$$

which is proper. The Bayesian bootstrap has the following steps,

1. Sample $\underset{\sim}{p} \mid \underset{\sim}{n} \sim \text{Dirichlet}(\underset{\sim}{n})$;
2. Sample $(N_1-n_1,\ldots,N_d-n_d) \mid \underset{\sim}{p},\underset{\sim}{n} \sim \text{Multinomial}(N-n,\underset{\sim}{p})$;
3. Repeat (1) and (2) a large number of times.

We have repeated the bootstrap procedure 1000 times. At each repetition, for the nonsamples, y_j^* occurs $N_j - n_j$ times, $j = 1,\ldots,d$; so we have got the entire population with y_j^* occurring N_j times with 1000 repetitions. It is worth noting that the Bayesian bootstrap is different from the DPG model (one-level DP model) when it is applied to an individual area because while the Bayesian bootstrap cannot produce new values, the DPG model can do so.

The 85^{th} and 95^{th} percentiles are also important and the methodology is essentially the same. We perform the predictive inference of the population mean, 85^{th} and 95^{th} percentiles

for each area using the three models (S-S, DPM, DPG). We have compared the DPG model to the S-S model, the DPM model and Bayesian bootstrap. We have computed summary statistics, posterior mean (PM), posterior standard deviation (PSD), and coefficient of variation (CV $= 100\times$ PSD/PM), as a measure of reliability.

We have looked at the five-number summaries (Min, Q1, Med, Q3, Max) of the shrinkage coefficients over the areas; for example, see Appendix A. For the S-S model and the DPM model, these are virtually the same $(.48,\ .55,\ .58,\ .60,\ .84)$ but there is only a small difference from the DPG model, which has $(.52,\ .59,\ .62,\ .64,\ .87)$. These numbers indicate that there is comparable and moderate pooling for all three models.

In Table 1, as a further measure of shrinkage, we have presented five-number summaries over the areas of $PB = (PM - B)/B$, where B is the posterior mean from the Bayesian bootstrap method; recall that the bootstrap is a method to obtain the finite population quantities for each area separately (no pooling). We observe that for the finite population mean, the five-number summaries are virtually the same with 50% negative PBs and 50% positive PBs. The three models are almost the same for the finite population 85^{th} percentile; Min is negative for all three models but they are different. They differ for the finite population 95^{th} percentile; virtually all the PBs are positive under the DPG model, but 25% are positive under the S-S and DPM models. Therefore, there is some evidence that the DPG model is more responsive to the gaps, outliers and ties in the BMI data. The assumption of independent normal responses in the S-S and DPM models is overly restrictive, especially when we get out into the tails of the BMI data.

Table 1: Five-number summaries of $PB = (PM - B)/B$ of the finite population mean, 85^{th} percentile and 95^{th} percentile for BMI data by three models (S-S, DPM, DPG)

	Mean					85^{th} Percentile					95^{th} Percentile				
Model	Min	Q1	Med	Q3	Max	Min	Q1	Med	Q3	Max	Min	Q1	Med	Q3	Max
S-S	-0.02	-0.01	0.00	0.01	0.02	-0.05	0.00	0.01	0.02	0.05	-0.11	-0.04	-0.01	0.00	0.04
DPM	-0.02	-0.01	0.00	0.01	0.02	-0.05	0.00	0.01	0.03	0.05	-0.11	-0.04	-0.01	0.00	0.04
DPG	-0.02	-0.01	0.00	0.01	0.02	-0.02	0.01	0.02	0.03	0.05	-0.05	0.00	0.01	0.02	0.04

NOTE: Min=Minimum; Q1= 1st quartile; Med=median; Q3=3rd quartile; Max= Maximum.

In Table 2, as a measure of reliability, we present the five-number summaries of the coefficient of variation (CV $= 100\times$PSD/PM) over the areas. Overall these are very good for all finite population quantities and models (including the bootstrap) although under the bootstrap these CVs should be a bit bigger because the bootstrap generally underestimates variability. For the finite population mean, the CVs from the three models are mostly similar and those under the S-S, DPM and DPG models are mostly smaller than the bootstrap. For the finite population 85^{th} percentile and the finite population 95^{th} percentile, the S-S model and DPM model are similar, but their CVs are mostly to the left of those of the bootstrap. However, the five-number summaries of DPG model for estimating the finite population 95^{th} percentile are to the right of those of the S-S and DPM models, but still to the left of the bootstrap. Nevertheless, all three models appear to show good reliability.

Table 2: Five-number summaries of coefficient of variation (CV = 100×PSD/PM), of the finite population mean, 85th percentile and 95th percentile for BMI data by three models (S-S, DPM , DPG) and Bayesian bootstrap (Boot)

	Mean					85th Percentile					95th Percentile				
Model	Min	Q1	Med	Q3	Max	Min	Q1	Med	Q3	Max	Min	Q1	Med	Q3	Max
S-S	0.62	1.20	1.26	1.37	1.57	0.59	1.11	1.19	1.25	1.48	0.60	1.12	1.18	1.27	1.46
DPM	0.66	1.28	1.36	1.45	1.60	0.60	1.20	1.29	1.32	1.52	0.63	1.21	1.25	1.30	1.53
DPG	0.61	1.18	1.23	1.30	1.55	1.13	1.60	1.84	2.24	2.65	1.85	2.19	2.44	2.54	3.97
Boot	0.61	1.34	1.49	1.62	1.97	1.19	2.04	2.67	3.04	4.49	2.17	2.99	3.58	4.10	7.39

NOTE: Min=Minimum; Q1= 1st quartile; Med=median; Q3=3rd quartile; Max= Maximum, Boot=Bootsrap.

We have looked at plots (not shown) of the posterior densities of the finite population mean, 85^{th} and 95^{th} percentiles for the three models (S-S, DPM and DPG) and Bayesian bootstrap for the 35 areas of BMI data. For the population mean, most parts of the density under the S-S, DPM and DPG models are similar, the DPG model has slightly smaller variation. Plots of the estimated densities of the population 85^{th} and 95^{th} percentiles under the DPG model are not smooth and the estimated densities of the population 85^{th} and 95^{th} percentiles under the S-S and DPM models are similar. Because the BMI data have some gaps, ties and outliers in the right tails, the estimations given by parametric models may be incorrect. Thus, based on a belief that the parametric model is too restrictive, we prefer the analysis based on the nonparametric DPG model.

Finally, we compare predictive inference of the finite population mean, 85^{th} and 95^{th} percentile for each area by the three models (S-S, DPM and DPG). We use three plots (not shown), which contain posterior means with credible bands versus direct estimates for BMI data. The posterior means are very similar under the S-S, DPM and DPG models and the predictive inferences of the population percentile are similar under the S-S and DPM models. For the finite population mean, the points (plot not shown) are all roughly on a straight line crossing the 45-degree straight line with slightly smaller slope, as it is should be. For the 85^{th} percentile, the points (plot not shown) are a little bit more spread out. However, as expected, the DPG model tends to have higher predictions (closer to the 45-degree straight line) of the population percentiles with similar credible bands when it is compared to the other two models. We suspect that S-S and DPM model might underestimate the 85^{th} and 95^{th} population percentiles when the data are right skewed. Without the restrictive parametric assumptions, the DPG model tends to provide less biased estimation with similar variation comparing to the other candidate models, thereby showing a distinct advantage of the DPG model; see Figure 2 for the finite population 95^{th} percentile. We investigate this issue in a small simulation study.

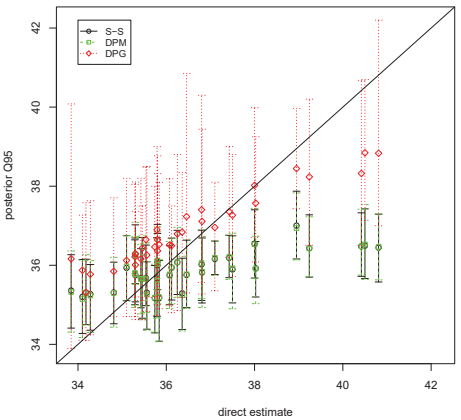


Figure 2: Comparison for body mass index (BMI) data (posterior means with credible bands versus direct estimates): the predictive inference of the finite population 95^{th} percentile for each county under the three models (S-S, DPM, DPG)

3.2. Simulation Study

We conduct a small simulation study. We choose $\ell = 50$ and the sample sizes, n_i, for 50 areas. The sample sizes are 35 for each of the first 10 areas, 50 for each of the second 10 areas, 100 for each of the third 10 areas, 200 for each of the fourth 10 areas and 500 for each of the last 10 areas. Then, the population sizes are selected as $N_i = 100n_i, i = 1,\ldots,\ell$. These are comparable to the BMI data. For convenience, we have taken $\theta = 0.0$, $\sigma^2 = 0.01$, $\delta^2 = 0.04$, thereby making $\rho = 0.8$. For the concentration parameters of the Dirichlet processes, we have selected $\gamma = 0.5$, and $\alpha_i \overset{ind}{\sim} 0.5 + \text{Beta}(5,5), i = 1,\ldots,\ell$. These choices allow us to have data similar to the BMI data with some flexibility to get gaps, outliers and ties when data are simulated from the DPG model.

We have simulated the entire finite population separately under the three models. This is done the same way under each model separately. For example, under the S-S model, because we have set θ, σ^2 and ρ, we have generated $\mu_1,\ldots,\mu_\ell$ from (2) and for the i^{th} area, we have generated $y_{ij}, j = 1,\ldots,N_i$ from (1). Therefore, we have all three finite population quantities. Given the parameters, because the observations are independent and identically distributed within each area, we simply take the first n_i values as the sample. In the case of the DPG model, the population values are exchangeable and so we still take the first n_i values as our sample.

When data are generated from the S-S model and the DPM model, there could be gaps and outliers in different areas. We note, in particular, there will be no ties because two data values cannot be the same (this happens with probability zero). Of course, the data from distinct areas will show some differences. However, as we have explained in this paper, when data are generated from the DPG model, there will be gaps, outliers and ties because

the data values are generated via the Polya urn scheme. By the nature of a DP, two values can be the same with nonzero probability; so there can be ties.

We fit the three models to the simulated data in exactly the same manner as for the NHANES-III BMI data. When the DPM model was fit, the Geweke tests show stationarity and the effective sample sizes are comparable to 1000. We have looked at plots (not shown) of the posterior means with 95% credible bands and true population means for the simulated S-S, DPM and DPG data under three different models (S-S, DPM and DPG models). The values are close to the true population mean. We also use absolute bias (AB) and posterior root mean squared error (PRMSE) to compare the models. We know the true value of the finite population quantities, denoted by T. Then, $AB = \mid PM - T \mid$ and $PRMSE = \sqrt{(PM-T)^2 + PSD^2}$. We compute these quantities for each of the fifty counties for the finite population mean and the 85^{th} and 95^{th} finite population percentiles, and respectively we average them. We present AB and PRMSE in Table 3; note that the entries in the table must be divided by $10,000$.

First, consider the finite population mean in Table 3 (a). We observe that AB is always too large when the DPM model is fit to any of the three simulated data sets. The S-S model and DPG model show comparable AB, much smaller than those for DPM. The PRMSEs under the DPG model are larger than those from S-S model and the DPM model (first two rows) by about 7% (marginal) but they are not larger than that of the DPM model when data are generated from the S-S model (0.01272 vs. 0.01008). The DPG is almost always better when data are generated from it; there is only a minor difference for AB under the S-S model and the DPG model (0.0001409 vs. 0.0001484).

Second, consider the finite population 85^{th} percentile in Table 3 (b). When data are generated from the S-S model, the S-S model and the DPG model are comparable and better than the DPM model. When data are generated from the DPM model, the three models are comparable with the PRMSE under the DPG model slightly higher than the other two models. When data are generated from the DPG model, the DPG model is a clear winner by far.

Third, consider the finite population 95^{th} percentile in Table 3 (c). When data are generated from the S-S model, the S-S model and the DPG model are comparable and better than the DPM model. When data are generated from the DPM model, the three models are more comparable. When data are generated from the DPG model, the DPG model is enormously better than the S-S model and DPM model.

When data are generated from the DPG model, in terms of AB and PRMSE, it performs much better than the S-S model and the DPM model for all three finite population quantities. This is strong evidence that when there are gaps, outliers and ties, the DPG model is the best. It is risky to use the S-S model or the DPM model for such data. The DPG model does not have to do better for data that are generated from the S-S model or the DPM model. By drawing a dot plot, one can see clearly which model is appropriate; data generated from the DPG model will have gaps, outliers and ties. Therefore, it is safe to conclude that the DPG model will perform better for data like the BMI data; see Figure 1.

Table 3: Comparison of absolute bias (AB) and posterior root mean squared error (PRMSE) of the finite population mean, 85^{th} percentile and 95^{th} percentile for each simulated data by three models (S-S, DPM and DPG) averaged over areas

	(a) Mean					
	S-S		DPM		DPG	
Data	AB	PRMSE	AB	PRMSE	AB	PRMSE
S-S	6.172	94.21	87.32	127.2	6.536	100.8
DPM	6.169	93.82	43.27	85.44	6.32	100.70
DPG	1.409	42.88	28.95	64.55	1.484	27.64
	(b) 85^{th} Percentile					
	S-S		DPM		DPG	
Data	AB	PRMSE	AB	PRMSE	AB	PRMSE
S-S	70.21	130.7	111.2	155.1	77.39	137.4
DPM	69.93	133.9	75.49	123.4	78.38	141.3
DPG	379.0	385.9	384.1	394.6	18.05	40.0
	(c) 95^{th} Percentile					
	S-S		DPM		DPG	
Data	AB	PRMSE	AB	PRMSE	AB	PRMSE
S-S	120.6	182.1	150.3	203.5	133.8	188.4
DPM	104.0	168.9	114.9	166.0	118.9	176.1
DPG	550.2	556.3	555.3	563.7	35.5	101.0

NOTE: Each row gives a model that generates the data and each column gives a model that is fit to the simulated data. The same three data sets are used in (a), (b) and (c). [The numbers in the table must be multiplied by 10^{-4}.]

4. Concluding Remarks and Future Work

If the parametric distribution assumption does not hold, the model is mis-specified and the inference may be invalid. The Bayesian nonparametric methods are motivated by the desire to avoid overly restrictive assumptions. We believe that our DPG model, which has independent Dirichlet processes on the responses and a normal distribution on the area means, can accommodate survey responses with gaps, outliers and ties reasonably well.

Our illustration using the BMI data in our novel DPG model is a step forward. Our simulation shows the advantage of the DPG model when the finite population mean and the 85^{th} and 95^{th} finite population percentiles are being estimated. In the illustrative example on BMI data, it is interesting that Bayesian predictive inference can be performed using a data integration because the area sizes are available from the 1990 census. In future, we can adjust the DPG model to include a DP prior on the area means, rather than a normal distribution (Nandram and Yin 2019).

For future work, we may also include covariates in the DPG model in a manner in which Battese, Harter and Fuller (1988) actually extended the model of Scott and Smith (1969) to include covariates. The two-stage nonparametric alternative of the DPG model

with p covariates and an intercept, $\mathbf{x_{ij}} = (\mathbf{1}, \mathbf{x'_{ij}}^{(\mathbf{0})})'$, is

$$\begin{aligned}
y_{ij} - \mathbf{x'_{ij}}^{(\mathbf{0})}\beta^{(\mathbf{0})} | \mathbf{G_i} &\overset{ind}{\sim} G_i, \quad i = 1,\ldots,\ell, \quad j = 1,\ldots,N_i, \\
G_i | \beta_{0i} &\overset{ind}{\sim} \mathrm{DP}\left\{\alpha_i, \mathrm{Normal}(\beta_{0i}, \sigma^2)\right\}, \\
\beta_{0i} \quad | \, \theta, \sigma^2, \rho &\overset{ind}{\sim} \mathrm{Normal}(\theta, \frac{\rho}{1-\rho}\sigma^2), \\
\pi(\alpha_i) &= \frac{1}{(\alpha_i+1)^2}, \qquad \alpha_i > 0, \quad i = 1,\ldots,\ell, \\
\pi(\beta^{(\mathbf{0})}, \theta, \boldsymbol{\sigma^2}, \rho) &\propto \frac{1}{1+\theta^2}\frac{1}{(1+\sigma^2)^2},
\end{aligned}$$

$-\infty < \theta, \beta_s^{(0)} < \infty, s = 1,\ldots,p, 0 < \sigma^2 < \infty, 0 \le \rho \le 1$, where ρ is the intra-cluster correlation, $\mathbf{x_{ij}}^{(\mathbf{0})}$ and $\beta^{(\mathbf{0})}$ denote $\mathbf{x_{ij}}$ and β without the intercepts. Note also that a priori the α_i are independent and there is a flat prior on $\beta^{(\mathbf{0})}$. This is how we can incorporate demographic variables (age, race and sex) for the BMI data from NHANES III.

In many complex surveys, there are also survey weights; this is also true for NHANES III. We may include the survey weights in the model using a normalized composite likelihood. However, if the survey weights for the nonsampled values are unknown, it is not obvious how to perform predictive inference under the model. One solution may be to use surrogate sampling (Nandram 2007).

Acknowledgements

This research was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram). The authors thank Professor Wlodzimierz Okrasa for his invitation and encouragement.

REFERENCES

ANTONIAK, C. E., (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2 (6), pp. 1152–1174.

BATTESE, G. E., HARTER, R. M. and FULLER, W. A., (1988). An Error-components Model for Prediction of County Crop Areas using Survey and Satellite Data. *Journal of the American Statistical Association*, 83 (401), pp. 28–36.

BINDER, D. A., (1982). Non-parametric Bayesian Models for Samples from Finite Populations. *Journal of the Royal Statistical Society*, Series B, 44 (3), pp. 388–393.

BLACKWELL, D., MACQUEEN, J. B., (1973). Ferguson Distributions via Polya Urn Schemes. *The Annals of Statistics*, 1 (2), pp. 353–355.

ESCOBAR, M. D., WEST, M., (1995). Bayesian Density Estimation and Inference using Mixtures. *Journal of the American Statistical Association*, 90 (430), pp. 577–588.

FERGUSON, T. S., (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1 (2), pp. 209–230.

GELMAN, A., JAKULIN, A., PITTAU, M. P. and SU, Y-S., (2008). A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *Annals of Applied Statistics*, 2 (4), pp. 1360–1383.

KALLI, M., GRIFFIN, J. E. and WALKER, S. G., (2011). Slice Sampling Mixture Models. *Statistics and Computing*, 21 (1), pp. 83–105.

LO, A. Y., (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, 12 (1), pp. 351–357.

MOLINA, I., NANDRAM, B. and RAO, J. N. K., (2014). Small Area Estimation of General Parameters with Application to Poverty Indicators: A Hierarchical Bayes Approach. *The Annals of Applied Statistics*, 8 (2), pp. 852–885.

NANDRAM, B., CHOI, J. W., (2010). Bayesian Analysis of Body Mass Index Data from Small Domains Under Nonignorable Nonresponse and Selection. *Journal of the American Statistical Association*, 105 (489), pp. 120–135.

NANDRAM, B., CHOI, J. W., (2005). Hierarchical Bayesian Nonignorable Nonresponse Regression Models for Small Areas: An Application to the NHANES Data. *Survey Methodology*, 31 (1), pp. 73–84.

NANDRAM, B., CHOI, J. W., (2004). Nonparametric Bayesian Analysis of a Proportion for a Small Area under Nonignorable Nonresponse. *Journal of Nonparametric Statistics*, 16 (6), pp. 821–839.

NANDRAM, B., (2007). Bayesian Predictive Inference under Informative Sampling via Surrogate Samples. *In Bayesian Statistics and Its Applications*, edited by S.K. Upadhyay, U.Singh and D.K. Dey, Anamaya New Delhi, pp. 356–374.

NANDRAM, B., TOTO, M. C. S. and CHOI, J. W., (2011). A Bayesian Benchmarking of the Scott-Smith Model for Small Areas. *Journal of Statistical Computation and Simulation*, 81 (11), pp. 1593–1608.

NANDRAM, B., YIN, J., (2016a). Bayesian Predictive Inference under a Dirichlet Process with Sensitivity to the Normal Baseline. *Statistical Methodology*, 28, pp. 1–17.

NANDRAM, B., YIN, J., (2016b). A Nonparametric Bayesian Prediction Interval for a Finite Population Mean. *Journal of Statistical Computation and Simulation*, 86 (16), pp. 3141–3157.

NANDRAM, B., YIN, J., (2019). Hierarchical Bayesian Models for Small Areas With Dirichlet Processes. *JSM Proceedings, pp. 2594–2613, Survey Research Methods Section. Alexandria, VA: American Statistical Association*.

POLETTINI, S., (2017). A Generalized Semiparametric Bayesian Fay-Herriot Model for Small Area Estimation Shrinking Both Means and Variances. *Bayesian Analysis*, 12 (3), pp. 729–752.

RAO, J. N. K., MOLINA, I., (2015). *Small Area Estimation*, John Wiley & Sons, NY.

RUBIN, D. B., (1981). The Bayesian Bootstrap. *The Annals of Statistics*, 9 (1), pp. 130–134.

SETHURAMAN, J., (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4, pp. 639–650.

SCOTT, A., SMITH, T. M. F., (1969). Estimation in Multi-Stage Surveys. *Journal of the American Statistical Association*, 64 (327), pp. 830–840.

TOTO, M. C. S., NANDRAM, B., (2010). A Bayesian Predictive Inference for Small Area Means Incorporating Covariates and Sampling Weights. *Journal of Statistical Planning and Inference*, 140, pp. 2963–2979.

APPENDICES

A: Fitting the S-S Model

Let $\underset{\sim}{y} = (\underset{\sim}{y}_s, \underset{\sim}{y}_{ns})$, where $\underset{\sim}{y}_s = \{y_{ij}, i = 1,\ldots,\ell, j = 1,\ldots,n_i\}$ is the vector of observed values and $\underset{\sim}{y}_{ns} = \{y_{ij}, i = 1,\ldots,\ell, j = n_i+1,\ldots,N_i\}$ vector of unobserved values. First, define the sample means and sample variances, $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ and $s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2, n_i > 1, i = 1,\ldots,\ell$. Second, let $\lambda_i = \frac{n_i}{n_i+(1-\rho)/\rho}, i = 1,\ldots,\ell$, $\tilde{y} = \sum_{i=1}^{\ell}\lambda_i\bar{y}_i/\sum_{i=1}^{\ell}\lambda_i$, and $A_1 = \frac{1-\rho}{\rho}\sum_{i=1}^{\ell}\lambda_i(\tilde{y}-\bar{y}_i)^2 + \sum_{i=1}^{\ell}(n_i-1)s_i^2$. Here, the λ_i are *shrinkage* coefficients.
Then, using Bayes' theorem, the joint posterior density of $\underset{\sim}{\mu}, \theta, \sigma^2, \rho$ is

$$\begin{aligned}\pi(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y}_s) \;\propto\; & \left(\frac{1}{\sigma^2}\right)^{(n+\ell)/2}\left(\frac{1-\rho}{\rho}\right)^{\ell/2}\exp\Bigg\{-\frac{1}{2\sigma^2}\Bigg\{\sum_{i=1}^{\ell}\Big\{(n_i-1)s_i^2 \\ & + \left(n_i+\frac{1-\rho}{\rho}\right)(\mu_i-[\lambda_i\bar{y}_i+(1-\lambda_i)\theta])^2 \\ & + \lambda_i\left(\frac{1-\rho}{\rho}\right)(\bar{y}_i-\theta)^2\Big\}\Bigg\}\Bigg\}\times\frac{1}{(1+\sigma^2)^2}\times\frac{1}{\pi(1+\theta^2)}. \quad \text{(A.1)}\end{aligned}$$

We use a simple method called the sampling importance resampling (SIR) algorithm to draw from the posterior distribution $\pi(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y}_s)$ in (A.1). That is, we take a sample of draws from a proposal density $\pi_a(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y}_s)$, then use these draws to produce a sample from $\pi(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y}_s)$. As a well-known result, one would need $\pi(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y}_s)/\pi_a(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y}_s)$ to be uniformly bounded in its parameters. A reasonable approximation to the joint posterior density (A.1) and one from which it is easy to draw samples will suffice. We use the same likelihoods (1) and (2) in the two-level normal model together with an improper prior $\pi(\theta,\sigma^2,\rho) \propto \frac{1}{\sigma^2}, -\infty < \theta < \infty, 0 < \sigma^2 < \infty, 0 \le \rho \le 1$ as a Bayesian model from which we use the posterior density as a proposal density,

$$\begin{aligned}\pi_a(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y}_s) \;\propto\; & \pi_a(\underset{\sim}{\mu}|\theta,\sigma^2,\rho,\underset{\sim}{y}_s)\pi_a(\theta|\sigma^2,\rho,\underset{\sim}{y}_s)\pi_a(\sigma^2|\rho,\underset{\sim}{y}_s)\pi_a(\rho|\underset{\sim}{y}_s) \quad \text{(A.2)} \\ \propto\; & \prod_{i=1}^{\ell} N\left[\mu_i;\lambda_i\bar{y}_i+(1-\lambda_i)\theta,(1-\lambda_i)\frac{\rho}{1-\rho}\sigma^2\right] \\ \times\; & N\left(\theta;\tilde{y},\frac{\sigma^2\rho}{\sum_{i=1}^{\ell}\lambda_i(1-\rho)}\right)\times \mathrm{IG}\left[\sigma^2;(n-1)/2,A_1/2\right] \\ \times\; & \frac{\Gamma[(n-1)/2]}{(A_1/2)^{(n-1)/2}}\prod_{i=1}^{\ell}(1-\lambda_i)^{1/2}\left[\frac{\rho}{\sum_{i=1}^{\ell}\lambda_i(1-\rho)}\right]^{1/2}.\end{aligned}$$

Note that $\pi(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y}_s)/\pi_a(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y}_s) = \frac{1}{\pi(1+\theta^2)}\frac{\sigma^2}{(1+\sigma^2)^2} \le \frac{1}{\pi}$ (uniformly bounded as required). We draw a sample from the approximate joint posterior density (A.2) by first drawing a sample from $\pi_a(\rho|\underset{\sim}{y}_s)$ using the grid method and continue using the multiplication rule of probability. The algorithm works fine because the sub-sampling weights are nearly uniform.

B: Fitting the DPM Model

Kalli, Griffin and Walker (2011) suggested slice-efficient samplers, and it is based on the stick-breaking algorithm (Sethuraman 1994). Letting $G = \sum_{s=1}^{\infty} \pi_s \delta_{\mu_s^*}$, where

$$\pi_1 = \beta_1, \quad \pi_s = \beta_s \prod_{j=1}^{s-1}(1-\beta_j), \quad \beta_s \overset{iid}{\sim} \text{Beta}(1,\gamma), \quad \mu_s^* \overset{iid}{\sim} G_0,$$

and G_0 is a baseline distribution. Here, for convenience, we will use a short-hand notation for the formulas below, $h(y_{ij};\mu_i) = \text{Normal}_{y_{ij}}(\mu_i, \sigma^2), j = 1,\ldots,n_i, i = 1,\ldots,\ell$ and so $h(\underset{\sim}{y_i};\mu_i) = \prod_{j=1}^{n_i}\{\text{Normal}_{y_{ij}}(\mu_i,\sigma^2)\}, i = 1,\ldots,\ell$. Also, we use $g(\mu_i) = \text{Normal}_{\mu_i}(\theta, \frac{\rho}{1-\rho}\sigma^2)$, $i = 1,\ldots,\ell$.

The idea is to introduce latent variables $\{u_1,u_2,\ldots,u_\ell\}$, which allows us to sample a finite number of variables at each iteration. One can introduce further latent variables, $\{d_1,d_2,\ldots,d_\ell\}$ that indicate the components of the mixture from which observations are to be taken to give a general class of slice samplers,

$$f(\underset{\sim}{y_i}, u_i, d_i|\pi,\mu^*) = \mathbf{1}(u_i < \xi_{d_i})\pi_{d_i}/\xi_{d_i} h(\underset{\sim}{y_i};\mu_{d_i}^*),$$

where $\xi_1,\xi_2,\ldots$ is any positive sequence. Typically, the sequence will be deterministic decreasing sequence. In our computation, we use $\xi_s = (1-\kappa)\kappa^{s-1}$ where the tuning constant κ is between 0 and 1; other choices are possible. Let $K = \max_{i=1}^{\ell}(K_i)$, where K_i is the largest integer t such that $\xi_t > u_i$.

Specifically, for our DPM model, the joint posterior distribution is proportional to

$$\pi(\theta,\sigma^2,\rho,\gamma)\prod_{s=1}^{K}\text{Beta}(\beta_s;1,\gamma)g_0(\mu_s^*)\prod_{i=1}^{\ell}\mathbf{1}(u_i<\xi_{d_i})\pi_{d_i}/\xi_{d_i}h(\underset{\sim}{y_i};\mu_{d_i}^*).$$

The variables $\{(\mu_s^*,\beta_s), s = 1,2,\ldots,K; (d_i,u_i), i = 1,\ldots,\ell\}$ need to be sampled at each iteration. The Gibbs sampler is obtained by drawing samples, each in turn, from the conditional posterior distributions, (a) $\pi(u_i|\ldots) \propto \mathbf{1}(0 < u_i < \xi_{d_i})$; (b) $\pi(\mu_s^*|\ldots) \propto g_0(\mu_s^*)\prod_{\{i|d_i=s\}} h(\underset{\sim}{y_i};\mu_s^*)$; (c) $\pi(\beta_s|\ldots) \propto \text{Beta}(a_s,b_s)$, where $a_s = 1+\sum_{i=1}^{\ell}\mathbf{1}(d_i=s)$ and $b_s = \gamma+\sum_{i=1}^{\ell}\mathbf{1}(d_i>s)$; (d) $P(d_i = r|\ldots) \propto \mathbf{1}(r:\xi_r > u_i)\pi_r/\xi_r h(\underset{\sim}{y_i};\mu_r^*)$, $r = 1,\ldots,K$. The other parameters are included in the Gibbs sampler, and the grid method is used to draw some of them (e.g. γ).

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 3, pp. 21–37, DOI 10.21307/stattrans-2020-042
Received – 17.12.2019; accepted – 11.05.2020

Measuring and Testing Mutual Dependence of Multivariate Functional Data

Mirosław Krzyśko[1], **Łukasz Smaga**[2]

ABSTRACT

This paper considers new measures of mutual dependence between multiple multivariate random processes representing multidimensional functional data. In the case of two processes, the extension of functional distance correlation is used by selecting appropriate weight function in the weighted distance between characteristic functions of joint and marginal distributions. For multiple random processes, two measures are sums of squared measures for pairwise dependence. The dependence measures are zero if and only if the random processes are mutually independent. This property is used to construct permutation tests for mutual independence of random processes. The finite sample properties of these tests are investigated in simulation studies. The use of the tests and the results of simulation studies are illustrated with an example based on real data.

Key words: characteristic function, dependence measure, distance covariance, multivariate functional data, permutation method, test of independence..

1. Introduction

In recent years, statistical methods for analysing data expressed as functions or curves have received much attention. Such data are called functional data, which can be univariate and multivariate, and appear in many application domains as, for instance, chemometrics, economics, medicine, meteorology. For analysis of such data (i.e. the so called functional data analysis), there is currently a wide spectrum of models and methods as, for example, clustering and classification, functional principal component analysis, hypothesis testing, regression models. For an overview, we refer to the following monographs and recent review papers: Ramsay and Silverman (2005), Ferrary and Vieu (2006), Horváth and Kokoszka (2012), Zhang (2013), Kokoszka and Reimherr (2017) and Cuevas (2014), Wang et al. (2016) respectively.

This paper addresses the correlation analysis and testing independence for functional data in both univariate and multivariate cases. For functional time series, independence testing was considered by Horváth and Rice (2015). We would like to explore the association between two or more sets of functional variables. For two multivariate variables, the canonical correlation in the framework of canonical correlation analysis (CCA) was first proposed for this problem by Hotelling (1936). For functional data, this method was

[1]Interfaculty Institute of Mathematics and Statistics, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Poland. E-mail: mkrzysko@amu.edu.pl. ORCID: https://orcid.org/0000-0001-8075-4432

[2]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland. E-mail: ls@amu.edu.pl. ORCID: https://orcid.org/0000-0002-2442-8816.

extended by Leurgans et al. (1993), He et al. (2004), Krzyśko and Waszak (2013) and Krzyśko and Smaga (2019). Unfortunately, the association viewed by canonical correlation is not a global measurement, since the intensity of the relationship is expressed component by component (see, Górecki et al., 2017, for more details). This was one of the reasons for constructing other association measures. The two very popular of them are the ρV coefficient by Escoufier (1970, 1973) and the distance correlation *dCor* coefficient proposed in Székely et al. (2007). Their functional extensions were investigated in Górecki et al. (2016, 2017). Moreover, Górecki et al. (2019) used the functional distance correlation coefficient, among others, to construct variable selection procedures for classification of functional data. Unfortunately, the ρV coefficient may not detect non-linear dependence between two sets of variables, and it is difficult to evaluate the magnitude of the relationship just by considering its value. In these directions, the distance correlation coefficient seems to perform better and, moreover, (under mild conditions) it is equal to zero if and only if the random vectors are independent, which is not true for the ρV coefficient in general. Recently, Chen et al. (2019) proposed other distance-based coefficients with similar properties to distance correlation coefficient, which can even result in more powerful test for independence of two random vectors. In this paper, we adapt their results to a functional data framework by defining the functional version of their coefficient using a basis function representation of functional observations. In contrast to Górecki et al. (2016), we allow non-orthogonal basis making our results more general. In particular, we redefine the functional distance correlation coefficient in more generality.

The above considerations concern the case of two sets of variables only. Sometimes, there is a need of measure association or test independence of more than two sets of features. In this direction, very good results were obtained by Jin and Matteson (2018) in the case of multivariate data. They proposed a few methods, but the best of them are two procedures based on sums of squared distance covariance coefficients. Thus, in this paper, we extend these methods for functional data using also the functional versions of coefficients by Chen et al. (2019).

The remainder of this paper is organized as follows. In Section 2, we propose permutation tests of independence and dependence measures of multiple random processes. The finite sample properties of the testing procedures are investigated in simulation studies in Section 3. In Section 4, the real data example is presented. Finally, Section 5 is the summary of our work.

2. Methodology

In this section, we first present the basis representation of functional data, which is a kind of dimension reduction method. Then, using this representation and characteristic function apparatus, we propose tests of independence and dependence measures of two random processes. Finally, we extend these results for more than two processes.

2.1. Basis representation of functional data

Let $\mathbf{X} = (X_1, \ldots, X_p)^\top$ be a random process belonging to the Hilbert space $L_2^p(I)$ of p-dimensional vectors of square integrable functions defined on the interval $I = [a,b]$, $a,b \in \mathbb{R}$. This space is endowed with the following inner product:

$$\langle \mathbf{f}, \mathbf{g} \rangle_p = \int_I \mathbf{f}^\top(t)\mathbf{g}(t)dt$$

for $\mathbf{f}, \mathbf{g} \in L_2^p(I)$. For $i = 1, \ldots, p$, let $\{\phi_{ij}\}_{j=1}^\infty$ be basis in $L_2^1(I)$. Then each element of $L_2^1(I)$ can be represented as an infinite linear combination of basis functions. Such representation is difficult to apply in practice. Moreover, only a number of the first coefficients in this representation is usually the largest and the most important (Ramsey and Silverman, 2005). Therefore, we assume that each component of the process $\mathbf{X}$ can be represented by a finite number of basis functions, i.e.

$$X_i(t) = \sum_{j=1}^{B_i} \alpha_{ij}\phi_{ij}(t), \tag{1}$$

for $t \in I$ and $i = 1, \ldots, p$. The linear combination of basis functions in the right hand side of equality (1) will be called the basis representation of the process X_i.

The choice of the basis is usually not very crucial. However, some suggestions for this subject can be found in the literature (see, for example, Horváth and Kokoszka, 2012). The value of B_i determine the degree of smoothness of the basis representation, i.e. small value cause more smoothness. This value can be chosen deterministically or taking into account the problem at hand or using the Bayesian Information Criterion (BIC). The coefficients α_{ij} are usually estimated by the least squares method. For details about the practical construction of the basis representation, see, for example, Krzyśko and Waszak (2013).

Finally, let us introduce the following matrix form of the basis representation of a random process $\mathbf{X}$. Let

$$\alpha = (\alpha_{11}, \ldots, \alpha_{1B_1}, \ldots, \alpha_{p1}, \ldots, \alpha_{pB_p})^\top$$

and

$$\Phi(t) = \operatorname{diag}\left(\phi_1^\top(t), \ldots, \phi_p^\top(t)\right)$$

is the block diagonal matrix of

$$\phi_i^\top(t) = (\phi_{i1}(t), \ldots, \phi_{iB_i}(t)),$$

for $i = 1, \ldots, p$. Then the representation (1) can be expressed as follows:

$$\mathbf{X}(t) = \Phi(t)\alpha$$

which can be seen as the basis representation of the process $\mathbf{X}$. This means that the process $\mathbf{X}$ belongs to the finite dimensional subspace, say $\mathscr{L}_2^p(I)$, of the space $L_2^p(I)$.

2.2. Two sets of functional data

Assume that $\mathbf{X}$ and $\mathbf{Y}$ are two random processes belonging to the Hilbert spaces $L_2^p(I_1)$ and $L_2^q(I_2)$ respectively, where $I_1=[a_1,b_1]$ and $I_2=[a_2,b_2]$, $a_1,b_1,a_2,b_2\in\mathbb{R}$. We would like to test the following hypotheses:

$$H_0:\mathbf{X},\mathbf{Y}\text{ are independent vs. }H_1:\mathbf{X},\mathbf{Y}\text{ are dependent}$$

and in the case of rejecting the null hypothesis, to measure the correlation between the processes $\mathbf{X}$ and $\mathbf{Y}$. For this purpose, we use the concept of characteristic function. Namely, (roughly speaking) we want to use the fact that the null hypothesis H_0 is equivalent to the equality of the characteristic function of the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ with the product of the characteristic functions of the distributions of $\mathbf{X}$ and $\mathbf{Y}$.

Let us first recall the definition of the characteristic function of a random process (Bosq, 2000, p. 37) in our framework. The characteristic functions of the processes $\mathbf{X}$ and $\mathbf{Y}$ are as follows:

$$\varphi_{\mathbf{X}}(\mathbf{u})=E(\exp(i\langle\mathbf{u},\mathbf{X}\rangle_p)),\quad \varphi_{\mathbf{Y}}(\mathbf{v})=E(\exp(i\langle\mathbf{v},\mathbf{Y}\rangle_q))$$

for $\mathbf{u}\in L_2^p(I_1)$ and $\mathbf{v}\in L_2^q(I_2)$, where $i^2=-1$. (Of course, we assume that for all $\mathbf{u}\in L_2^p(I_1)$ the integral $\langle\mathbf{u},\mathbf{X}\rangle_p$ converges for almost all realizations of $\mathbf{X}$, and the same applies to $\mathbf{v}\in L_2^q(I_2)$ and $\mathbf{Y}$.) Then the joint characteristic function of the pair of processes $\mathbf{X}$ and $\mathbf{Y}$ is of the form

$$\varphi_{\mathbf{X},\mathbf{Y}}(\mathbf{u},\mathbf{v})=E(\exp(i\langle\mathbf{u},\mathbf{X}\rangle_p+i\langle\mathbf{v},\mathbf{Y}\rangle_q)).$$

The next step is to combine these definitions with the basis representation of the processes $\mathbf{X}$ and $\mathbf{Y}$ (see Section 2.1). Suppose that $\mathbf{X}\in\mathscr{L}_2^p(I_1)$ and $\mathbf{Y}\in\mathscr{L}_2^q(I_2)$ and

$$\mathbf{X}(t)=\Phi_1(t)\alpha,\quad \mathbf{Y}(s)=\Phi_2(s)\beta,$$

where $\Phi_1(t)=\operatorname{diag}(\boldsymbol{\phi}_{11}^\top(t),\ldots,\boldsymbol{\phi}_{1p}^\top(t))$, $\Phi_2(s)=\operatorname{diag}(\boldsymbol{\phi}_{21}^\top(s),\ldots,\boldsymbol{\phi}_{2q}^\top(s))$, $\alpha\in\mathbb{R}^{K_x}$ and $\beta\in\mathbb{R}^{K_y}$ are random vectors, $K_x=B_1^x+\cdots+B_p^x$ and $K_y=B_1^y+\cdots+B_q^y$. Moreover, we assume that the functions $\mathbf{u}\in\mathscr{L}_2^p(I_1)$ and $\mathbf{v}\in\mathscr{L}_2^q(I_2)$, and they are represented as follows:

$$\mathbf{u}(t)=\Phi_1(t)\gamma,\quad \mathbf{v}(s)=\Phi_2(s)\delta,$$

where $\gamma\in\mathbb{R}^{K_x}$ and $\delta\in\mathbb{R}^{K_y}$ are constant vectors. Then we have

$$\langle\mathbf{u},\mathbf{X}\rangle_p=\int_{I_1}\mathbf{u}^\top(t)\mathbf{X}(t)dt=\gamma^\top\int_{I_1}\Phi_1^\top(t)\Phi_1(t)dt\,\alpha=\gamma^\top\mathbf{J}_{\Phi_1}\alpha,$$

where $\mathbf{J}_{\Phi_1}=\operatorname{diag}(\mathbf{J}_{\boldsymbol{\phi}_{11}},\ldots,\mathbf{J}_{\boldsymbol{\phi}_{1p}})$ and $\mathbf{J}_{\boldsymbol{\phi}_{1i}}=\int_{I_i}\boldsymbol{\phi}_{1i}(t)\boldsymbol{\phi}_{1i}^\top(t)dt$ is the $B_i^x\times B_i^x$ cross product matrix, $i=1,\ldots,p$. Analogously, we obtain $\langle\mathbf{v},\mathbf{Y}\rangle_q=\delta^\top\mathbf{J}_{\Phi_2}\beta$. Therefore, the characteristic functions of the random processes $\mathbf{X}$ and $\mathbf{Y}$ are the characteristic functions of the random vectors $\mathbf{J}_{\Phi_1}\alpha$ and $\mathbf{J}_{\Phi_2}\beta$, i.e.

$$\varphi_{\mathbf{X}}(\mathbf{u})=E(\exp(i\gamma^\top\mathbf{J}_{\Phi_1}\alpha))=\varphi_{\mathbf{J}_{\Phi_1}\alpha}(\gamma),\quad \varphi_{\mathbf{Y}}(\mathbf{v})=E(\exp(i\delta^\top\mathbf{J}_{\Phi_2}\beta))=\varphi_{\mathbf{J}_{\Phi_2}\beta}(\delta).$$

Furthermore, the joint characteristic function of random processes $\mathbf{X}$ and $\mathbf{Y}$ is the joint characteristic function of random vectors $\mathbf{J}_{\Phi_1}\alpha$ and $\mathbf{J}_{\Phi_2}\beta$, i.e.

$$\varphi_{\mathbf{X},\mathbf{Y}}(\mathbf{u},\mathbf{v}) = E(\exp(i\gamma^\top \mathbf{J}_{\Phi_1}\alpha + i\delta^\top \mathbf{J}_{\Phi_2}\beta)) = \varphi_{\mathbf{J}_{\Phi_1}\alpha,\mathbf{J}_{\Phi_2}\beta}(\gamma,\delta).$$

These relations imply that for our purpose, we can use the distance methods for random vectors, which are based on

$$D_w = \int_{\mathbb{R}^{K_x+K_y}} |\varphi_{\mathbf{J}_{\Phi_1}\alpha,\mathbf{J}_{\Phi_2}\beta}(\gamma,\delta) - \varphi_{\mathbf{J}_{\Phi_1}\alpha}(\gamma)\varphi_{\mathbf{J}_{\Phi_2}\beta}(\delta)|^2 w(\gamma,\delta)d\gamma d\delta,$$

where $|z|$ is the modulus of $z \in \mathbb{C}$, and w is a weight function, which is positive almost everywhere. Different choices of the function w may result in plenty different methods. In the following, we consider two of them, which seem to be meaningful.

The most famous method of this kind was proposed by Székely et al. (2007). Górecki et al. (2016) used their methodology and considered the following functional distance covariance of random processes $\mathbf{X}$ and $\mathbf{Y}$:

$$FdCov(\mathbf{X},\mathbf{Y}) = dCov(\mathbf{J}_{\Phi_1}\alpha,\mathbf{J}_{\Phi_2}\beta) = \mathscr{V}_{\mathbf{J}_{\Phi_1}\alpha,\mathbf{J}_{\Phi_2}\beta} = \sqrt{D_{w_0}},$$

where

$$w_0(\gamma,\delta) = \frac{1}{C_{K_x}C_{K_y}\|\gamma\|_{K_x}^{K_x+1}\|\delta\|_{K_y}^{K_y+1}},$$

and

$$C_l = \frac{\pi^{(l+1)/2}}{\Gamma((l+1)/2)}$$

and $\|\cdot\|_l$ is the standard Euclidean norm in $\mathbb{R}^l$. The functional distance correlation between random processes $\mathbf{X}$ and $\mathbf{Y}$ is defined as follows:

$$FdCor(\mathbf{X},\mathbf{Y}) = \frac{FdCov(\mathbf{X},\mathbf{Y})}{\sqrt{FdCov(\mathbf{X},\mathbf{X})FdCov(\mathbf{Y},\mathbf{Y})}},$$

when $FdCov(\mathbf{X},\mathbf{X})$ and $FdCov(\mathbf{Y},\mathbf{Y})$ are positive, otherwise $FdCor(\mathbf{X},\mathbf{Y}) = 0$. Note that Górecki et al. (2016) used orthonormal basis, which implies the matrices $\mathbf{J}_{\Phi_1}$ and $\mathbf{J}_{\Phi_2}$ are identity matrices. Thus the above definition is a bit more general. For distributions with finite first moments, $FdCor(\mathbf{X},\mathbf{Y}) \in [0,1]$ and $FdCor(\mathbf{X},\mathbf{Y}) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent. The distance covariance by Székely et al. (2007) is implemented in the R package energy (R Core Team, 2019; Rizzo and Székely, 2019), which can be also used to calculate the functional distance covariance.

Recently, Chen et al. (2019) proposed other choice of weight function, which resulted in a kind of generalization of distance covariance. Namely, their weight functions are products of density functions. Let us now describe the details. Similarly as Székely et al. (2007), assume that the weight function $w(\gamma,\delta) = w_{K_x}(\gamma)w_{K_y}(\delta)$, where w_{K_x} and w_{K_y} are functions defined in the corresponding dimensions. This considerably simplifies expressions without

giving up much generality. Let f_{K_x} and f_{K_y} be densities and let φ_{K_x} and φ_{K_y} be characteristic functions of $K_x \times 1$ and $K_y \times 1$ random vectors respectively. Chen et al. (2019) proved that when the densities f_{K_x} and f_{K_y} are positive with probability 1, then taking $w_{K_x} = f_{K_x}$ and $w_{K_y} = f_{K_y}$, D_w is as follows:

$$\begin{aligned} D_f =& E\left(Re\{\varphi_{K_x}(\mathbf{J}_{\Phi_1}(\alpha-\alpha_1))\}Re\{\varphi_{K_y}(\mathbf{J}_{\Phi_2}(\beta-\beta_1))\}\right) \\ &+ E\left(Re\{\varphi_{K_x}(\mathbf{J}_{\Phi_1}(\alpha-\alpha_1))\}\right)E\left(Re\{\varphi_{K_y}(\mathbf{J}_{\Phi_2}(\beta-\beta_1))\}\right) \\ &- 2E\left(Re\{\varphi_{K_x}(\mathbf{J}_{\Phi_1}(\alpha-\alpha_1))\}Re\{\varphi_{K_y}(\mathbf{J}_{\Phi_2}(\beta-\beta_2))\}\right), \end{aligned}$$

where $Re(z)$ denotes the real part of $z \in \mathbb{C}$, $\alpha_1 \overset{d}{=} \alpha$, $\beta_m \overset{d}{=} \beta$ for $m = 1,2$ and $\overset{d}{=}$ stands for equality in distribution. Moreover, D_f is equal to zero if and only if $\mathbf{J}_{\Phi_1}\alpha$ and $\mathbf{J}_{\Phi_2}\beta$ are mutually independent, and it is strictly positive otherwise.

There are many possible choices of the densities f_{K_x} and f_{K_y}. To greatly simplify D_f, the densities of spherical stable distributions can be used. The characteristic function of a spherical stable distribution with exponent $\alpha \in (0,2]$ is $\varphi_\alpha(\mathbf{t}) = \exp(-\|\mathbf{t}\|^\alpha)$. For $\alpha = 1$ and $\alpha = 2$, we have the multivariate standard Cauchy and normal distributions respectively. Further details about spherical stable distributions can be found in Zolotarev (1981) and Nolan (2013). A recent application of spherical stable distributions in the change-point methods to multivariate time-series can be found in Hlávka et al. (2020). When f_{K_x} and f_{K_y} are the densities of spherical stable distributions with the same exponent α, D_f can be written as

$$\begin{aligned} D_\alpha =& E\left(\exp(-(\|\mathbf{J}_{\Phi_1}(\alpha-\alpha_1)\|^\alpha + \|\mathbf{J}_{\Phi_2}(\beta-\beta_1)\|^\alpha))\right) \\ &+ E\left(\exp(-\|\mathbf{J}_{\Phi_1}(\alpha-\alpha_1)\|^\alpha)\right)E\left(\exp(-\|\mathbf{J}_{\Phi_2}(\beta-\beta_1)\|^\alpha)\right) \\ &- 2E\left(\exp(-(\|\mathbf{J}_{\Phi_1}(\alpha-\alpha_1)\|^\alpha + \|\mathbf{J}_{\Phi_2}(\beta-\beta_2)\|^\alpha))\right). \end{aligned}$$

Thus, we can define the functional distance covariance and correlation with exponent α of random processes $\mathbf{X}$ and $\mathbf{Y}$ as

$$FdCov_\alpha(\mathbf{X},\mathbf{Y}) = \sqrt{D_\alpha}, \quad FdCor_\alpha(\mathbf{X},\mathbf{Y}) = \frac{FdCov_\alpha(\mathbf{X},\mathbf{Y})}{\sqrt{FdCov_\alpha(\mathbf{X},\mathbf{X})FdCov_\alpha(\mathbf{Y},\mathbf{Y})}}$$

respectively. Similarly to $FdCor(\mathbf{X},\mathbf{Y})$, $FdCor_\alpha(\mathbf{X},\mathbf{Y}) \in [0,1]$ and $FdCor_\alpha(\mathbf{X},\mathbf{Y}) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent.

In practice, $FdCor(\mathbf{X},\mathbf{Y})$ and $FdCor_\alpha(\mathbf{X},\mathbf{Y})$ have to be estimated. Assume that $\mathbf{X}_1,\ldots,\mathbf{X}_n$ and $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ are independent realizations of random processes $\mathbf{X}$ and $\mathbf{Y}$ respectively. Let $\mathbf{X}_i(t) = \Phi_1(t)\alpha_i$ and $\mathbf{Y}_i(s) = \Phi_2(s)\beta_i$, $i = 1,\ldots,n$ be the basis representations of the observations. The estimator of $FdCor(\mathbf{X},\mathbf{Y})$, say $\widehat{FdCor}(\mathbf{X},\mathbf{Y})$, was derived in Górecki et al.

(2016), so we omit it to save space. The estimator of $FdCov^2_\alpha(\mathbf{X},\mathbf{Y})$ is as follows:

$$\begin{aligned}\widehat{FdCov}^2_\alpha(\mathbf{X},\mathbf{Y}) = &\frac{1}{n^2}\sum_{1\le j,k\le n}\exp\left(-(\|\mathbf{J}_{\Phi_1}(\alpha_j-\alpha_k)\|^\alpha+\|\mathbf{J}_{\Phi_2}(\beta_j-\beta_k)\|^\alpha)\right)\\ &+\frac{1}{n^4}\sum_{1\le j,k\le n}\exp\left(-\|\mathbf{J}_{\Phi_1}(\alpha_j-\alpha_k)\|^\alpha\right)\sum_{1\le j,k\le n}\exp\left(-\|\mathbf{J}_{\Phi_2}(\beta_j-\beta_k)\|^\alpha\right)\\ &-\frac{2}{n^3}\sum_{1\le j,k,l\le n}\exp\left(-(\|\mathbf{J}_{\Phi_1}(\alpha_j-\alpha_k)\|^\alpha+\|\mathbf{J}_{\Phi_2}(\beta_j-\beta_l)\|^\alpha)\right).\end{aligned}$$

To sum up, both $FdCor(\mathbf{X},\mathbf{Y})$ and $FdCor_\alpha(\mathbf{X},\mathbf{Y})$ can be used as measures of dependence of random processes $\mathbf{X}$ and $\mathbf{Y}$. Moreover, since they both are equal to zero if and only if the processes $\mathbf{X}$ and $\mathbf{Y}$ are independent, testing the null hypothesis H_0 is equivalent to testing $H_0^{dCor}: FdCor(\mathbf{X},\mathbf{Y})=0$ or $H_0^{\alpha}: FdCor_\alpha(\mathbf{X},\mathbf{Y})=0$. For testing these hypotheses, we propose permutation tests based on test statistics $\widehat{FdCor}(\mathbf{X},\mathbf{Y})$ and $\widehat{FdCor}_\alpha(\mathbf{X},\mathbf{Y})$, because the asymptotic null distributions of $n\widehat{FdCor}(\mathbf{X},\mathbf{Y})$ and $n\widehat{FdCor}_\alpha(\mathbf{X},\mathbf{Y})$ are complicated and not distribution free and the convergence rate may be slow (see Székely et al., 2007; Chen et al., 2019). In the permutation method, the test statistics are recalculated many times with the permutation samples $\mathbf{X}_1,\ldots,\mathbf{X}_n,\mathbf{Y}_{\pi(1)},\ldots,\mathbf{Y}_{\pi(n)}$, where a permutation π is uniformly chosen from the symmetric group $\mathscr{S}_n$, the set of all $n!$ permutations of $(1,\ldots,n)$.

In the next section, we show how the above results can be extended for measuring and testing mutual dependence of more than two random processes.

2.3. Multiple sets of functional data

Let $\mathbf{X}_1,\ldots,\mathbf{X}_d$ be d random processes belonging to $L_2^{p_1}(I_1),\ldots,L_2^{p_d}(I_d)$ respectively, where $I_l=[a_l,b_l]$, $a_l,b_l\in\mathbb{R}$, $l=1,\ldots,d$. Of interest is to test the following hypotheses

$$H_0:\mathbf{X}_1,\ldots,\mathbf{X}_d \text{ are independent vs. } H_1:\mathbf{X}_1,\ldots,\mathbf{X}_d \text{ are dependent}$$

and in the case of rejecting the null hypothesis, to measure the correlation between the processes $\mathbf{X}_1,\ldots,\mathbf{X}_d$.

The methods based on characteristic functions of Section 2.2 can be extended for case $d>2$. Namely, for random vectors, this was recently done by Jin and Matteson (2018), whose results could be directly applied to functional data in much the same way as presented in Section 2.2. However, such tests may not perform well as was already shown in Jin and Matteson (2018) for random vectors. Fortunately, they also proposed some alternatives to these methods, which have better finite sample properties. Therefore, we are limited only to these alternative methods, which are asymmetric and symmetric measures of mutual dependence to capture mutual dependence via aggregating pairwise dependence.

Assume that $\mathbf{X}_l\in\mathscr{L}_2^{p_l}(I_l)$ and we have the following basis representation of the processes $\mathbf{X}_l$:

$$\mathbf{X}_l(t_l)=\Phi_l(t_l)\alpha_l,$$

where $t_l \in I_l$ and $\alpha_l \in \mathbb{R}^{K_l}$ are random vectors, $l = 1,\ldots,d$. Let

$$\alpha_{c^+} = \left(\alpha_{c+1}^\top,\ldots,\alpha_d^\top\right)^\top, \quad c = 1,\ldots,d-1,$$
$$\alpha_{-c} = \left(\alpha_1^\top,\ldots,\alpha_{c-1}^\top,\alpha_{c+1}^\top,\ldots,\alpha_d^\top\right)^\top, \quad c = 1,\ldots,d.$$

Thus, α_{c^+} denotes the subset of processes on the right of α_c, while α_{-c} denotes the subset of processes except α_c. Let Cor be a dependence measure for two random vectors such that it is equal to zero if and only if the random vectors are independent. Then the asymmetric and symmetric measures of mutual dependence of random vectors $\alpha_1,\ldots,\alpha_d$ are defined by

$$R(\alpha_1,\ldots,\alpha_d) = \frac{1}{d-1}\sum_{c=1}^{d-1} Cor^2(\alpha_c,\alpha_{c^+}), \quad S(\alpha_1,\ldots,\alpha_d) = \frac{1}{d}\sum_{c=1}^{d} Cor^2(\alpha_c,\alpha_{-c}).$$

Under mild condition, Jin and Matteson (2018) showed that

$$R(\alpha_1,\ldots,\alpha_d) \in [0,\infty), \quad S(\alpha_1,\ldots,\alpha_d) \in [0,\infty)$$

and

$$R(\alpha_1,\ldots,\alpha_d) = 0, \quad S(\alpha_1,\ldots,\alpha_d) = 0$$

if and only if $\alpha_1,\ldots,\alpha_d$ are mutually independent.

In the framework of functional data, we can use $FdCor$ or $FdCor_\alpha$ as Cor above. Then for testing the null hypothesis H_0, we can verify

$$H_0^R : R(\alpha_1,\ldots,\alpha_d) = 0 \text{ or } H_0^S : S(\alpha_1,\ldots,\alpha_d) = 0.$$

For these purposes, we use permutation tests based on the following test statistics being estimators of R and S:

$$\hat{R} = \frac{1}{d-1}\sum_{c=1}^{d-1} \widehat{FdCor}^2(\mathbf{X}_c,\mathbf{X}_{c^+}), \quad \hat{S} = \frac{1}{d}\sum_{c=1}^{d} \widehat{FdCor}^2(\mathbf{X}_c,\mathbf{X}_{-c})$$

or

$$\hat{R}_\alpha = \frac{1}{d-1}\sum_{c=1}^{d-1} \widehat{FdCor}_\alpha^2(\mathbf{X}_c,\mathbf{X}_{c^+}), \quad \hat{S}_\alpha = \frac{1}{d}\sum_{c=1}^{d} \widehat{FdCor}_\alpha^2(\mathbf{X}_c,\mathbf{X}_{-c}).$$

The pooled permutation sample is constructed by separately permuting the samples corresponding to processes $\mathbf{X}_2,\ldots,\mathbf{X}_d$. More precisely, when

$$\mathbf{X}_{11},\ldots,\mathbf{X}_{1n},\mathbf{X}_{21},\ldots,\mathbf{X}_{2n},\ldots,\mathbf{X}_{d1},\ldots,\mathbf{X}_{dn}$$

are the observations, the pooled permutation sample is as follows:

$$\mathbf{X}_{11},\ldots,\mathbf{X}_{1n},\mathbf{X}_{2\pi_1(1)},\ldots,\mathbf{X}_{2\pi_1(n)},\ldots,\mathbf{X}_{d\pi_{d-1}(1)},\ldots,\mathbf{X}_{d\pi_{d-1}(n)},$$

where the permutations $\pi_1, \ldots, \pi_{d-1}$ are uniformly chosen from the symmetric group $\mathscr{S}_n$.

Under appropriate conditions, we have that $\hat{R}$, $\hat{S}$, $\hat{R}_\alpha$ and $\hat{S}_\alpha$ belong to the interval $[0,1]$. Therefore, they can be used as measures of dependence of random processes $\mathbf{X}_1, \ldots, \mathbf{X}_d$.

3. Simulation studies

In this section, the finite sample behaviour of the permutation tests $\hat{R}$, $\hat{S}$, $\hat{R}_\alpha$ and $\hat{S}_\alpha$ for $\alpha = 0.1, 0.5, 1, 1.5, 2$ is determined in simulation studies. We investigate both the control of the type I error and power of the tests.

3.1. Simulation experiments

We set the number of observations $n = 15$ and investigate $d = 3$ random processes $\mathbf{X}_1 = (X_{11}, X_{12})^\top$, $\mathbf{X}_2 = (X_{21}, X_{22})^\top$, $\mathbf{X}_3 = (X_{31}, X_{32})^\top$ with dimensions $p_1 = p_2 = p_3 = 2$. The functional observations corresponding to these processes are generated in the following three models:

Model 1. They are represented by their values in an equally spaced grid of 50 points $t_{1,1} = t_{2,1} = t_{3,1} = 0, \ldots, t_{1,50} = t_{2,50} = t_{3,50} = 1$ in $I_1 = I_2 = I_3 = [0,1]$, which are generated in the following way:

$$\begin{bmatrix} \mathbf{X}_{1r}(t_{1,u}) \\ \mathbf{X}_{2r}(t_{2,u}) \\ \mathbf{X}_{3r}(t_{3,u}) \end{bmatrix} = \begin{bmatrix} \Phi_1(t_{1,u}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Phi_2(t_{2,u}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Phi_3(t_{3,u}) \end{bmatrix} \begin{bmatrix} \alpha_{1,r} \\ \alpha_{2,r} \\ \alpha_{3,r} \end{bmatrix} + \varepsilon_{r,u},$$

where $r = 1, \ldots, n$, $u = 1, \ldots, 50$, the matrices Φ_l are as in Section 2 and contain the Fourier basis functions only and $B_i^l = 5$, $i = 1,2$, $l = 1,2,3$, $(\alpha_{1,r}^\top, \alpha_{2,r}^\top, \alpha_{3,r}^\top)^\top$ are 30-dimensional random vectors, and $\varepsilon_{r,u}^\top = (\varepsilon_{r,u,1}, \ldots, \varepsilon_{r,u,6})$ are the measurement errors such that $\varepsilon_{r,u,v} \sim N(0, 0.025 a_{r,v})$ and $a_{r,v}$ is the range of the v-th row of the following matrix:

$$\begin{bmatrix} \Phi_1(t_{1,1})\alpha_{1,r} & \ldots & \Phi_1(t_{1,50})\alpha_{1,r} \\ \Phi_2(t_{2,1})\alpha_{2,r} & \ldots & \Phi_2(t_{2,50})\alpha_{2,r} \\ \Phi_3(t_{3,1})\alpha_{3,r} & \ldots & \Phi_3(t_{3,50})\alpha_{3,r} \end{bmatrix}.$$

The random vectors $(\alpha_{1,r}^\top, \alpha_{2,r}^\top, \alpha_{3,r}^\top)^\top$ are generated as $Z_r \Sigma_\rho^{1/2}$, where $\Sigma_\rho = (1-\rho)\mathbf{I}_{30} + \rho \mathbf{1}_{30}\mathbf{1}_{30}^\top$, $\rho = 0, 0.1$, $\mathbf{I}_a$ is the $a \times a$ identity matrix, $\mathbf{1}_a$ is the $a \times 1$ vector of ones, and Z_r are 30×1 random vectors with iid coordinates from the following distributions: the standard normal distribution N, the Student t-distribution t_3 with three degrees of freedom, the Fisher-Snedecor distribution $F_{1,5}$ with 1 and 5 degrees of freedom, the standard Cauchy distribution C, the log-normal distribution LN. When $\rho = 0$, the null hypothesis about independence is true and we study the type I error of tests, while for $\rho = 0.1$, the alternative holds and we investigate their power. Note that for Cauchy distribution C, the expected value does not exist, but this distribution was among others considered in similar simulations of Chen et al. (2019), so we also use it.

Model 2. First, for each $t \in \{0.04, 0.08, \ldots, 1\}$, the observations for $X_{li}(t)$, $l = 1,2$, $i = 1,2$ are generated as independent random variables of normal distribution $N(0, 0.25)$ or other non-normal distributions considered in Model 1. Then, for the same t, $X_{3i}(t) = \rho X_{1i}(t) + \varepsilon_i(t)$ for $i = 1,2$, where $\varepsilon_i(t)$ are independent random variables of normal distribution $N(0, 0.1)$. We set $\rho = 0, 0.5$ and then the null, alternative hypothesis is true respectively.

Model 3. This model is similar to Model 2, but here we consider non-linear dependence instead of linear dependence. More precisely, we set $X_{3i}(t) = X_{1i}^{\rho}(t) + \varepsilon_i(t)$ and $\rho = 2,3$. For both values of ρ, the alternative hypothesis holds.

The test statistics are calculated using the Fourier basis with $B_i^l = 5$, $i = 1,2$, $l = 1,2,3$. We use the least squares method to estimate the coefficients of the basis representation of generated functional data. The empirical sizes and powers (resp. p-values) of the permutation tests were estimated in 500 simulation runs (resp. 1,000 permutation samples). For simplicity, the significance level is set to 5%. The simulation experiments as well as real data example of Section 4 were conducted in the R program (R Core Team, 2019).

3.2. Simulation results

The empirical sizes and powers of the permutation tests obtained in Models 1-3 are presented in Tables 1-3 respectively. Let us now discuss these simulation results.

The empirical sizes of all tests obtained in Models 1-2 (Tables 1-2 with $\rho = 0$) are usually very close to the level of significance of 5%. However, we can observe that the testing procedures $\widehat{FdCor}_{\alpha}$ with larger α (i.e. $\alpha = 1.5, 2$) tend to highly over-reject the null hypothesis in the case of Cauchy distribution C in Model 1. It seems that this can be explained by non-existence of the first moment of this distribution. Thus, the permutation tests seem to control the type I error level, except possibly tests based on $\widehat{FdCor}_{1.5}$ and $\widehat{FdCor}_{2}$.

In Model 1, all three processes $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ are equally correlated, which is a similar scenario to that considered in Jin and Matteson (2018) for random vectors. Then both methods R and S perform very similarly in terms of size control and power. On the other hand, in Model 2, the processes $\mathbf{X}_1$ and $\mathbf{X}_3$ are correlated (when $\rho > 0$), and they are uncorrelated to process $\mathbf{X}_2$. Such setting was not considered by Jin and Matteson (2018). In this case, the testing procedures $\hat{S}$ and $\hat{S}_{\alpha}$ are much more powerful than the tests $\hat{R}$ and $\hat{R}_{\alpha}$ respectively. This perhaps can be explained by that the S method considers more comparisons between processes $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ than the R method. In the case of Model 3, the processes $\mathbf{X}_1$ and $\mathbf{X}_3$ are non-linearly dependent (quadratically [$\rho = 2$] or cubically [$\rho = 3$]), and they are uncorrelated to process $\mathbf{X}_2$. Here, the comparison between methods R and S is more complicated and depends on the distribution of the data as well as the test statistic used. For tests $\widehat{FdCor}$ and $\widehat{FdCor}_{0.1}$, the methods R and S have similar empirical powers in most cases. For the other estimators (i.e. $\widehat{FdCor}_{\alpha}$, $\alpha = 0.5, 1, 1.5, 2$), the method R is usually more powerful than the method S. However, there are some exceptions, for example, under normal distribution N and cubic dependence or under Student distribution t_3 and quadratic dependence, the reverse is true.

Table 1: Empirical sizes ($\rho = 0$) and powers ($\rho = 0.1$) (as percentages) of all tests obtained in Model 1.

Distr.	ρ	Method	$\widehat{FdCor}$	$\widehat{FdCor}_{0.1}$	$\widehat{FdCor}_{0.5}$	$\widehat{FdCor}_{1}$	$\widehat{FdCor}_{1.5}$	$\widehat{FdCor}_{2}$
N	0	R	4.2	5.0	5.0	5.8	5.0	7.0
		S	4.4	5.4	5.4	6.4	6.4	4.6
	0.1	R	33.8	29.8	25.2	9.0	6.2	5.8
		S	35.2	30.8	25.2	9.8	5.2	4.0
t_3	0	R	5.6	5.2	5.2	4.6	4.4	5.2
		S	6.0	4.8	4.6	5.2	4.4	9.4
	0.1	R	26.6	24.0	19.2	5.2	4.8	4.8
		S	24.8	22.2	18.0	8.0	5.8	7.6
$F_{1,5}$	0	R	5.0	6.2	6.2	4.2	5.0	5.8
		S	5.0	4.8	5.0	3.4	4.6	7.0
	0.1	R	26.8	35.4	29.2	12.2	11.0	14.2
		S	24.4	30.6	29.8	10.8	8.8	14.8
C	0	R	4.4	4.6	5.6	3.0	19.4	80.4
		S	5.0	4.2	6.2	4.8	24.8	83.6
	0.1	R	43.0	65.8	36.4	11.8	35.0	88.8
		S	37.6	57.6	26.6	9.6	40.8	90.0
LN	0	R	4.2	5.0	4.8	4.6	4.8	4.8
		S	4.0	4.0	4.0	7.2	6.8	4.8
	0.1	R	24.8	29.6	26.4	10.6	8.2	7.4
		S	21.2	25.2	25.0	10.8	6.8	6.4

We can observe that the empirical powers of the tests $\widehat{FdCor}_{\alpha}$ usually decrease with the increasing α. There are only few exceptions (e.g. Model 3 and normal distribution N), but in these cases, the power loss between the most and the least powerful tests is not so large as in the remaining ones. Thus, among the tests $\widehat{FdCor}_{\alpha}$, the test $\widehat{FdCor}_{0.1}$ (i.e. with small α) is the most powerful in most scenarios.

In Models 1-2 and in Model 3 with normal distribution N, the tests $\widehat{FdCor}_{\alpha}$ with small α (e.g. $\alpha = 0.1$) are usually comparable with tests $\widehat{FdCor}$ in terms of power. Nevertheless, in some cases (e.g. under Fisher-Snedecor distribution $F_{1,5}$, Cauchy distribution C and the log-normal distribution LN), the tests $\widehat{FdCor}_{0.1}$ may have greater power than the tests $\widehat{FdCor}$. In Model 3 and non-normal distributions, the testing procedures $\widehat{FdCor}_{0.1}$ are much more powerful than the tests $\widehat{FdCor}$.

To sum up, the permutation test $\hat{S}_{0.1}$ seems to perform best. It maintains the type I error level very well and has power, which is greater than or comparable to power of the other tests considered. This test is followed by testing procedure $\hat{S}$. The test $\hat{S}_{0.1}$ overcomes the test $\hat{S}$ especially in the case of non-linear dependence.

Table 2: Empirical sizes ($\rho = 0$) and powers ($\rho = 0.5$) (as percentages) of all tests obtained in Model 2.

Distr.	ρ	Method	$\widehat{FdCor}$	$\widehat{FdCor}_{0.1}$	$\widehat{FdCor}_{0.5}$	$\widehat{FdCor}_{1}$	$\widehat{FdCor}_{1.5}$	$\widehat{FdCor}_{2}$
N	0	R	4.4	4.6	5.4	4.8	4.8	5.0
		S	4.0	4.8	4.6	4.6	3.8	4.2
	0.5	R	50.4	45.2	44.2	48.4	49.2	50.0
		S	77.6	69.2	70.6	76.0	80.6	82.2
t_3	0	R	5.6	5.4	5.8	5.8	6.4	5.8
		S	5.4	4.2	4.4	4.0	3.8	3.6
	0.5	R	48.2	47.2	41.8	36.8	33.0	29.0
		S	87.0	82.0	84.6	82.0	70.0	50.4
$F_{1,5}$	0	R	5.4	5.6	6.0	5.4	6.2	5.6
		S	4.6	5.8	6.0	6.4	6.2	4.4
	0.5	R	41.4	45.4	36.6	24.8	17.0	11.0
		S	67.4	67.8	72.6	56.8	23.0	9.6
C	0	R	5.0	5.2	6.2	3.8	5.2	4.8
		S	4.0	4.8	6.6	6.2	6.8	6.0
	0.5	R	34.4	44.0	31.8	8.2	6.2	5.2
		S	46.4	61.6	54.8	15.8	9.6	14.4
LN	0	R	4.6	4.0	4.0	3.8	4.2	4.2
		S	4.0	3.8	3.8	4.0	4.4	4.6
	0.5	R	48.8	46.4	41.0	35.8	30.0	25.6
		S	81.8	75.6	77.8	74.4	54.6	26.4

4. Real data example

In this section, we illustrate the use of the dependence measures and tests of independence for functional data proposed in Section 2 and the simulation results of Section 3. For this purpose, we consider the famous Canadian weather data, which are available in the R package fda (Ramsay et al., 2018).

The Canadian weather data contain the daily temperature and precipitation records of 35 Canadian weather stations averaged over 1960 to 1994 for 365 days. The raw temperature and precipitation curves for 35 weather stations are presented in Figure 1. Thus, we have $n = 35$ observations of two random processes ($d = 2$) representing temperature and precipitation. These functional observations are discretized in 365 time points. For illustrative purposes, we would like to measure dependence and test independence of temperature and precipitation treated as functional data. From Figure 1, (rather positive) correlation between temperature and precipitation may be observed. More precisely, weather stations with large temperature are also characterized by higher precipitation (dashed lines). In contrast, weather stations with lower temperature record lower rainfall (solid lines). To theoretically confirm this relationship, we use the methods described in Section 2 in the following.

Table 3: Empirical powers (as percentages) of all tests obtained in Model 3.

Distr.	ρ	Method	$\widehat{FdCor}$	$\widehat{FdCor}_{0.1}$	$\widehat{FdCor}_{0.5}$	$\widehat{FdCor}_{1}$	$\widehat{FdCor}_{1.5}$	$\widehat{FdCor}_{2}$
N	2	R	6.8	7.0	7.6	6.4	6.2	6.4
		S	6.6	5.8	6.6	6.4	6.0	7.0
	3	R	11.4	9.6	10.2	10.2	11.8	13.2
		S	15.8	13.8	14.6	16.0	17.0	17.8
t_3	2	R	54.4	59.2	53.8	26.4	12.8	7.4
		S	52.0	56.0	53.6	30.2	15.8	8.8
	3	R	68.8	84.0	69.2	27.6	14.0	8.6
		S	69.0	82.8	64.0	6.8	3.6	3.4
$F_{1,5}$	2	R	91.0	98.8	98.4	46.6	16.8	12.6
		S	84.0	96.8	95.6	6.4	3.4	3.2
	3	R	75.6	97.2	68.6	19.0	11.2	15.0
		S	71.2	95.0	8.0	5.2	5.4	5.0
C	2	R	74.6	95.8	64.4	23.0	34.4	52.6
		S	58.8	89.6	11.6	5.8	10.0	24.2
	3	R	67.2	98.0	54.0	29.0	40.4	60.8
		S	61.4	94.0	5.6	4.8	8.8	27.2
LN	2	R	98.0	99.6	100.0	71.0	28.0	11.6
		S	96.6	99.4	99.6	24.6	6.6	5.4
	3	R	85.4	98.0	79.4	30.4	6.8	6.6
		S	82.6	95.6	23.6	6.8	5.6	6.4

We use the permutation tests $\widehat{FdCor}$ and $\widehat{FdCor}_\alpha$ with $\alpha = 0.1, 0.5, 1, 1.5, 2$ and $1,000$ permutation samples. For the basis representation of the weather data, we use the Fourier basis with different size (i.e. $B_1^l = 3, 5, \ldots, 15$ for $l = 1, 2$) and the least squares method to estimate coefficients. The Fourier basis is recommended for periodical data (see, for example, Horváth and Kokoszka, 2012), so it is sensible for temperature and precipitation data, since they have annual cycles.

The results of statistical analysis are depicted in Table 4. We observe quite big values of correlation coefficients, especially $\widehat{FdCor}_\alpha$'s. Moreover, these values seem to not depend on the basis size. The same is true for p-values of the tests $\widehat{FdCor}$ and $\widehat{FdCor}_\alpha$ with $\alpha = 0.1, 0.5$. However, this is not true for the remaining testing procedures. This follows from the fact that the tests $\widehat{FdCor}$ and $\widehat{FdCor}_\alpha$ with small α are more robust to increasing dimension than the tests $\widehat{FdCor}_\alpha$ with moderate and large α. This was observed for random vectors in simulation studies in Chen et al. (2019) and moves to the case of functional data. Moreover, the p-values of the tests $\widehat{FdCor}_\alpha$ usually increase with the increasing α. Finally, the tests $\widehat{FdCor}$ and $\widehat{FdCor}_\alpha$ with $\alpha = 0.1, 0.5$ reject the null hypothesis at level of significance of 5%, in contrast to the remaining tests. These confirm the simulation results of Section 3, since the tests $\widehat{FdCor}$ and $\widehat{FdCor}_\alpha$ with small α were observed there to be

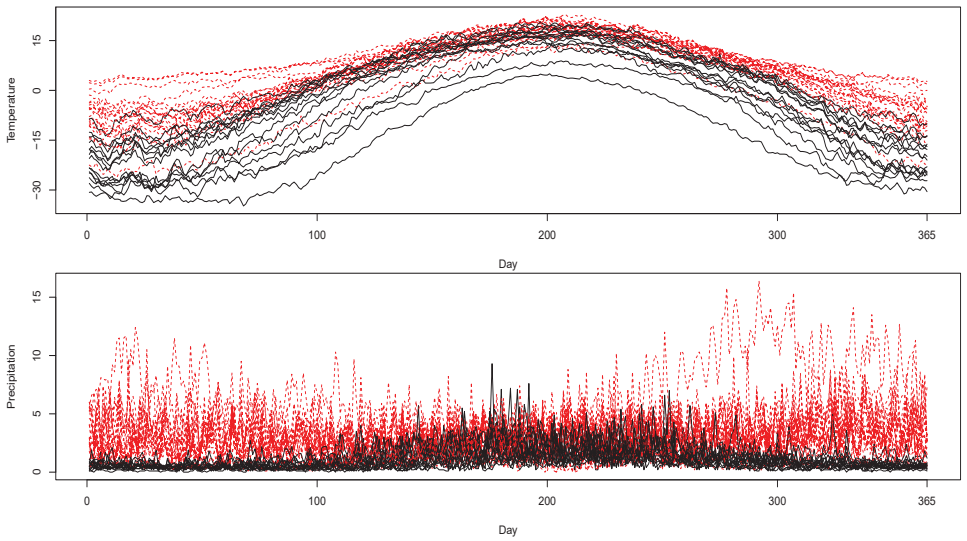


Figure 1: Temperature and precipitation for Canadian weather stations.

Table 4: Functional coefficients of correlation (FCor) and p-values of tests of independence for temperature and precipitation for Canadian weather stations.

	B_1^l	$\widehat{FdCor}$	$\widehat{FdCor}_{0.1}$	$\widehat{FdCor}_{0.5}$	$\widehat{FdCor}_1$	$\widehat{FdCor}_{1.5}$	$\widehat{FdCor}_2$
FCor	3	0.7379	0.9921	0.9825	0.9907	0.9914	0.9909
	5	0.7436	0.9935	0.9895	0.9975	0.9985	0.9988
	7	0.7449	0.9939	0.9913	0.9989	0.9998	0.9999
	9	0.7461	0.9942	0.9924	0.9993	0.9999	0.9999
	11	0.7464	0.9943	0.9932	0.9995	0.9999	0.9999
	13	0.7466	0.9944	0.9937	0.9996	0.9999	0.9999
	15	0.7468	0.9946	0.9942	0.9997	0.9999	0.9999
p-value	3	0.001	0.000	0.037	0.250	0.279	0.319
	5	0.001	0.000	0.015	0.339	0.352	0.306
	7	0.001	0.000	0.008	0.341	0.429	0.422
	9	0.001	0.000	0.002	0.228	0.407	0.411
	11	0.001	0.000	0.003	0.350	0.410	0.413
	13	0.001	0.000	0.006	0.360	0.353	0.325
	15	0.001	0.000	0.006	0.352	0.402	0.452

more powerful than the tests $\widehat{FdCor}_\alpha$ with moderate and large α. For these reasons, we should reject the null hypothesis about independence and conclude that there is a relationship between temperature and precipitation recorded in Canadian weather stations.

5. Conclusions

We have proposed new measures of mutual dependence for two or more sets of univariate and multivariate functional data. Our construction is based on the equivalence to mutual independence through characteristic functions and the basis function representation of the functional observations. Then, the problem is reduced to random vectors of basis expansion coefficients. We do not assume that the basis is orthogonal in contrast to the previous literature. For two sets of functional data, we follow the idea of functional distance correlation and construct functional versions of coefficients by Chen et al. (2019) indexed by hyperparameter $\alpha \in (0,2]$. In the case of more than two sets of functional data, we use the coefficients for pairs of sets and aggregate them by sums of their squares adapting the asymmetric and symmetric methods by Jin and Matteson (2018) to functional data framework. Simulation studies and real data example suggest that permutation tests based on new functional coefficients with small α and symmetric method perform best in terms of size control and power.

REFERENCES

BOSQ, D., (2000). *Linear Processes in Function Spaces. Theory and Applications*, Springer.

CHEN, F., MEINTANIS, S. G., ZHU, L., (2019). On some Characterizations and Multidimensional Criteria for Testing Homogeneity, Symmetry and Independence. *Journal of Multivariate Analysis*, 173, pp. 125–144.

CUEVAS, A., (2014). A Partial Overview of the Theory of Statistics with Functional Data. *Journal of Statistical Planning and Inference*, 147, pp. 1–23.

ESCOUFIER, Y., (1970). Echantillonnage dans une population de variables aléatoires réelles. Ph.D thesis, Université des Sciences et Techniques du Languedoc, Montpellier.

ESCOUFIER, Y., (1973). Le Traitement des Variables Vectorielles. *Biometrics*, 29, pp. 751–760.

FERRATY, F., VIEU, P., (2006). *Nonparametric Functional Data Analysis: Theory and Practice*, Springer: New York.

GÓRECKI, T., KRZYŚKO, M., RATAJCZAK, W., WOŁYŃSKI, W., (2016). An Extension of the Classical Distance Correlation Coefficient for Multivariate Functional Data with Applications. *Statistics in Transition new series*, 17, pp. 449–466.

GÓRECKI, T., KRZYŚKO, M., WOŁYŃSKI, W., (2017). Correlation Analysis for Multivariate Functional Data. *Studies in Classification, Data Analysis, and Knowledge Organization: Data Science*, pp. 243–258.

GÓRECKI, T., KRZYŚKO, M., WOŁYŃSKI, W., (2019). Variable Selection in Multivariate Functional Data Classification. *Statistics in Transition new series*, 20, pp. 123–138.

HE, G., MÜLLER, H. G., WANG, J. L., (2004). Methods of Canonical Analysis for Functional Data. *Journal of Statistical Planning and Inference*, 122, pp. 141–159.

HLÁVKA, Z., HUŠKOVÁ, M., MEINTANIS, S. G., (2020). Change-Point Methods for Multivariate Time-Series: Paired Vectorial Observations. *Statistical Papers*, DOI: https://doi.org/10.1007/s00362-020-01175-3.

HORVÁTH, L., KOKOSZKA, P., (2012). *Inference for Functional Data with Applications*, Springer.

HORVÁTH, L., RICE, G., (2015). Testing for Independence between Functional Time Series. *Journal of Econometrics*, 189, pp. 371–382.

HOTELLING, H., (1936). Relation Between Two Sets of Variables. *Biometrika*, 28, pp. 321–377.

JIN, Z., MATTESON, D. S., (2018). Generalizing Distance Covariance to Measure and Test Multivariate Mutual Dependence via Complete and Incomplete V-statistics. *Journal of Multivariate Analysis*, 168, pp. 304–322.

KOKOSZKA, P., REIMHERR, M., (2017). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC.

KRZYŚKO, M., SMAGA, Ł., (2019). Robust Estimation in Canonical Correlation Analysis for Multivariate Functional Data. *Hacettepe Journal of Mathematics and Statistics*, 48, pp. 521–535.

KRZYŚKO, M., WASZAK, Ł., (2013). Canonical Correlation Analysis for Functional Data. *Biometrical Letters*, 50, pp. 95–105.

LEURGANS, S. E., MOYEED, R. A., SILVERMAN, B. W., (1993). Canonical Correlation Analysis when the Data are Curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55, pp. 725–740.

NOLAN, J. P., (2013). Multivariate Elliptically Contoured Stable Distributions: Theory and Estimation. *Computational Statistics*, 28, pp. 2067–2089.

R CORE TEAM (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

RAMSAY, J. O., SILVERMAN, B. W. (2005). *Functional Data Analysis*, Second Edition, Springer.

RAMSAY, J. O., WICKHAM, H., GRAVES, S., HOOKER, G., (2018). fda: Functional Data Analysis. R package version 2.4.8. https://CRAN.R-project.org/ package=fda

RIZZO, M. L., SZÉKELY, G. J., (2019). energy: E-Statistics: Multivariate Inference via the Energy of Data. R package version 1.7-6.
https://CRAN.R-project.org/package=energy

SZÉKELY, G. J., RIZZO, M. L., BAKIROV, N. K., (2007). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, 35, pp. 2769–2794.

WANG, J. L., CHIOU, J. M., Müller, H. G., (2016). Functional Data Analysis. *Annual Review of Statistics and Its Application*, 3, pp. 257–295.

ZHANG, J. T., (2013). *Analysis of Variance for Functional Data*. Chapman & Hall: London.

ZOLOTAREV, V. M., (1981). Integral Transformations of Distributions and Estimates of Parameters of Multidimensional Spherically Symmetric Stable Laws. In *Contributions to Probability: A Collection of Papers Dedicated to Eugene Lukacs*, Academic Press, New York-London, pp. 283–305.

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 3 pp. 39–51, DOI 10.21307/stattrans-2020-043
Received – 20.01.2020; accepted – 10.06.2020

Detection of Outliers in Univariate Circular Data by Means of the Outlier Local Factor (LOF)

Ali H. Abuzaid[1]

ABSTRACT

The problem of outlier detection in univariate circular data was the object of increased interest over the last decade. New numerical and graphical methods were developed for samples from different circular probability distributions. The main drawback of the existing methods is, however, that they are distribution-based and ignore the problem of multiple outliers.

The local outlier factor (LOF) is a density-based method for detecting outliers in multivariate data and it depends on the local density of every k nearest neighbours.

The aim of this paper is to extend the application of the LOF to the detection of possible outliers in circular samples, where the angles of circular data are represented in two Cartesian coordinates and treated as bivariate data. The performance of the LOF is compared against other existing numerical methods by means of a simulation based on the power of a test and the proportion of correct detection. The LOF performance is compatible with the best existing discordancy tests, while outperforming other tests. The level of the LOF performance is directly related to the contamination and concentration parameters, while having an inverse relationship with the sample size.

In order to illustrate the process, the LOF and other existing discordancy tests are applied to detect possible outliers in two common real circular datasets.

Key words: discordancy, distance, multiple outliers, neighbours, spacing theory.

1. Introduction

The analyses of directions in xy-plane is more convenient to be considered as circular data, which are distributed on a unit circle circumference, measured by degrees or radians and belonging to $[0^\circ, 360^\circ)$ or $[0, 2\pi)$, respectively.

In the context of circular data, due to its closed bounded range property, then considering an outlier as an extreme value is no longer valid, where the extreme value

[1] Department of Mathematics, Al Azhar University – Gaza, Palestine. E-mail: a.abuzaid@alazhar.edu.ps. ORCID: https://orcid.org/ 0000-0002-6680-7371.

is defined as a point with the maximum circular deviation from the mean direction. Thus, the problem of outliers in circular data needs special discordancy tests. Collett (1980) proposed four tests of discordancy for circular samples. The past decade has seen a renewed interest in the detection and classification of outliers in the univariate circular data, either numerically (see Abuzaid et al. 2009, Mohamed 2016, Sidik et al. 2019) or graphically (Abuzaid et al. 2013). Recently, the problem of outliers in circular regression and functional relationship models has been well investigated (see, Satari et al. 2014, Alkasady et al. 2019).

Existing methods of outlier-detection in univariate circular data have some drawbacks: firstly, they are distribution-based methods, which rely on certain probabilistic distributional assumptions, where the cut-off points are needed for any combination of distribution parameters. Secondly, they were built for single outlier detection, and did not address the masking effect or multiple outliers. Lastly, they consider outlying as a binary property (i.e. either the angle is an outlier or not).

In geometrics, for a given angle θ with corresponding coordinates (x,y) on the unit circle, these coordinates are obtained as $x = \cos\theta$ and $y = \sin\theta$. Thus, treating the associated coordinates instead of the angle will allow us to use the available methods of outlier-detection in multivariate linear data. One of these methods is the local outlier factor (*LOF*), which is a density-based method. It computes the outlying factor of every point in a dataset based on its average distance to its k nearest neighbours. Furthermore, the outlier factor estimates the degree the suspected point is being outlying (Breunig et al. 2000). Recently, Abuziad (2020) has extended the concept of density-based local outliers to the medical multivariate circular data based on circular distances.

This article considers the LOF method, which is widely used in the multivariate analysis and available in most of statistical software programs as an alternative method of outlier-detection in univariate circular data, regardless of the probability distribution. The rest of this article is organized as follows: Section 2 reviews the main methods for outlier-detection in univariate circular data. Section 3 introduces the LOF in the circular data context. A comparative power of performance of available methods is presented in Section 4. For illustration, Section 5 analyses two real circular datasets.

2. Tests of discordancy in univariate circular data

Let $\theta_1,\ldots,\theta_n$ be a random sample from a circular variable, and the resultant length, $R = \sqrt{\left(\sum_{i=1}^{n} cos\,\theta_i\right)^2 + \left(\sum_{i=1}^{n} sin\,\theta_i\right)^2}$. The interest is to test the null hypothesis that θ_r, where

$1 \le r \le n$, is not an outlier. The following subsections review five discordancy tests to identify outliers.

2.1. M statistic

Mardia (1975) proposed a statistic based on the effect of removing the jth angle on the resultant length R, given by $M = \frac{R_r - R + 1}{n - R}$, where $R_r = max_j \left\{ R_{(-j)} \right\}$ and $R_{(-j)}$ is the resultant length after excluding the jth angle. The asymptotic distribution of M statistic is approximated by the standard normal distribution for large values of the concentration parameter (Collett, 1980).

2.2. C statistic

Collett (1980) proposed an alternative test of discordancy based on the mean resultant length, $\bar{R} = \frac{R}{n}$ and defined as $C = max_j \frac{\bar{R}_{(-j)} - \bar{R}}{\bar{R}}$, where $\bar{R}_{(-j)}$ is the mean resultant length after excluding the jth angle.

2.3. D statistic

The third statistic was derived by Collett (1980) based on the relative arc lengths between the ordered angles such as $\theta_{(1)} \le \theta_{(2)} \le ... \le \theta_{(n)}$. The arc length between consecutive angles is defined by $T_j = \theta_{(j+1)} - \theta_{(j)}, \; j = 1,...,n-1$ and $T_n = 2\pi - \theta_{(n)} + \theta_{(1)}$. The test statistic is given by $D_j = \frac{T_j}{T_{j-1}}, \; j = 1,...,n$. It corresponds to the greatest arc containing a single angle, θ_r, which is obtained by $D_r = \frac{T_r}{T_{r-1}}$. The $min\left\{ D_r, D_r^{-1} \right\}$ is considered because statistic D_r is a two-tailed statistic.

2.4. A statistic

Abuzaid et al. (2009) proposed a test statistic based on the summation of all circular distances from the angle θ_r to all other angles θ_j; $d_r = \sum_{j=1}^{n} 1 - cos\left(\theta_j - \theta_r \right)$ for $j, \; r = 1,...,n$. The test statistic is given by $max_r \left\{ \frac{d_r}{2(n-1)} \right\}, r = 1,...,n$. The approximated distribution of the A statistic was discussed in Abuzaid et al. (2012).

2.5. G statistic

Mohamed et al. (2016) extended the theory of arc length, which was used in D statistic to the spacing theory. The statistic is defined based on the a-step spacing where $a = 1,2,3,...,$ for the jth ordered angle as $G_{aj} = \theta_{(j+a)} - \theta_{(j)}$, for $j = 1,...,n-a$ and $G_{aj} = 2\pi - \theta_{(j)} + \theta_{(j+a)-n}$, for $j = (n+1)-a,(n+2)-a,...,n$. Then the test statistics is defined as $G_a = max_j \{G_{aj}\}$.

To identify possible outliers in circular samples, the previous five test statistics have to exceed a pre-determined cut-off points which have been obtained via simulation under the assumptions that the circular data come from certain distribution with known sample size and parameters. The cut-off points and power of performance for the five statistics have been obtained for von Mises distribution and wrapped normal distribution (Sidik et al. 2019), while only the associated values of cut-off points for the first four statistics were obtained for the wrapped Cauchy distribution (Abuzaid et al. 2015) and Cardioid distribution (Das and Gogoi, 2015).

3. Local outlier factor (LOF) for univariate circular data

Breunig et al. (2000) proposed a density-based method for detecting outliers in multivariate data. It depends on the local density of every k nearest neighbours. It is the so-called a local outlier factor (*LOF*), and it does not consider the outlier as a binary property, where it assigns a factor for each point to indicate its outlying degree. The term "local" is derived from the fact that the value of the factor for a point θ depends on how that point is isolated with respect to the surrounding neighbourhood. A higher *LOF* value reflects more sparse neighbourhoods and represents an outlier point, while lower value of *LOF* reflects more dense neighbourhoods and represents a normal point.

LOF for a point θ is obtained by computing its average distance to its k nearest neighbours, then the distance is normalized by computing the average distance of each of those neighbours to their k nearest neighbours. The set of the following definitions explains the *LOF* algorithm.

1) Distance $d(\theta,\phi)$ between any two angles θ and ϕ:

Let θ and ϕ be two angles in a univariate circular dataset, φ, with coordinates of (x_θ, y_θ) and (x_ϕ, y_ϕ), respectively. Then, the distance between any two angles θ and ϕ is obtained by

$$d(\theta,\phi) = \sqrt{(x_\phi - x_\theta)^2 + (y_\phi - y_\theta)^2}.$$

2) k-distance of an angle θ:

For any positive integer k, the k-distance of an angle $\theta \in \varphi$ is denoted by $dist_k(\theta)$ and it is defined as the distance $d(\theta,\phi)$ between an angle θ and an angle $\phi \in \varphi$. It represents the k-th nearest neighbourhoods of an angle θ, where there is at least k angles such that $d(\theta,\nu) \le d(\theta,\phi)$ and at most $k-1$ angles, such that $d(\theta,\nu) < d(\theta,\phi)$, where ν is an angle and $\nu \in \varphi \setminus \{\theta\}$.

3) k-distance neighborhood of a point θ:

It contains every point whose distance from θ is not greater than the k-distance. It is defined as $N_k(\theta) = \{\phi \mid dist(\theta,\phi) \le dist_k(\theta)\}$ and it could be greater than k, where multiple points have the same distance.

4) Reachability distance from angle θ to angle ϕ:

For all close angles θ's to an angle ϕ, it is expected that there is a statistical fluctuation of $dist(\theta,\phi)$, which can be significantly reduced by defining the reachability distance as

$$reach\_dist_k(\theta,\phi) = max\{dist_k(\theta), d(\theta,\phi)\}.$$

The higher the value of k, the more similar the reachability distances for angles within the same neighbourhood.

5) Local reachability density of an angle θ:

It is the inverse of the average reachability distance based on the k nearest neighbours of an angle θ, and it is defined as

$$lrd_k(\theta) = \frac{N_k(\theta)}{\sum_{\phi \in N_k(\theta)} reach\_dist_k(\theta,\phi)}.$$

6) Local outlier factor of an angle θ:

It estimates the degree to which an angle θ is called an outlier, and it is defined as

$$LoF_k(\theta) = \frac{\sum_{\phi \in N_k(\theta)} \frac{lrd_k(\phi)}{lrd_k(\theta)}}{N_k(\theta)}.$$

It is the average of the ratio of the local reachability density of θ and those of θ's k-nearest neighbours.

The minimum number of neighbour angles to determine the density, which the so-called *MinPts*, and its effect on changing the values of the *LOF* was discussed by Breunig et al. (2000). They concluded that the *MinPts* can be between two and $n-1$, and suggested it to be at least 10 to remove unwanted statistical fluctuations. Furthermore, the angle is considered as an outlier if its *LOF* value is significantly greater than one.

In general A and G_a statistics have outperform other statistics (Sidik et al. 2019). Therefore, the following section will investigate the performance of the A statistic and *LOF* via simulation.

4. Power of performance

The performance of discordancy tests is evaluated by three measures, namely power function; $P1 = 1-\beta$ where β is the probability of type-II error, $P3$ which is the probability of identifying a contaminated value as an outlier when it is in fact an extreme value, and the probability of wrongly identifying a good observation as discordant, which is denoted by $P1-P3$ (Barnett and Lewis, 1984).

To obtain the three measures of performance, the following settings are considered in this simulation study, which were conducted based on 2000 random samples generated from two different circular distributions; namely the von Mises distribution with mean μ and concentration parameter κ; denoted as $vM(\mu,\kappa)$, and the wrapped Cauchy distribution with mean μ and concentration parameter ρ; denoted as $WC(\mu,\rho)$. Without loss of generality, the mean direction of generated samples from both distributions were fixed equal to zero. Five different sample sizes, namely $n = 20, 50, 70, 100$ and 150 were generated.

The considered values of concentration parameters are $\kappa = 0.5, 2, 5, 7$ and $\rho = 0.2, 0.4, 0.6, 0.8, 0.99$ for samples generated from von Mises and wrapped Cauchy distributions, respectively.

The samples are generated in such a way that $n-1$ of the observations come from the distribution, i.e. $vM\left(0,\kappa\right)$ or $WC\left(0,\rho\right)$, and the remaining one observation comes from $vM\left(\lambda\pi,\kappa\right)$ or $WC\left(\lambda\pi,\rho\right)$, respectively, where λ is the degree of contamination and $0\leq\lambda\leq1$. Then A statistic and *LOF* are calculated as given in Sections 2 and 3, respectively, where the value of k is fixed as the rounded up median for each sample size.

Figure 1 shows that *LOF* and A statistic are compatible in the case of samples from von Mises distribution, while *LOF* outperforms the A statistic in the case of samples from wrapped Cauchy distribution. The full results of the simulation study can be requested from the author. Simulation results show that two measures of performance, namely $P1$ and $P3$ are almost the same, thus the values of $P1-P3$ are always close to zero. The performance of outlier-detection methods is highly dependent on the circular distribution. In general, the methods of outlier-detection for samples from von Mises distribution perform significantly better than the case of wrapped Cauchy distribution. This may referred to the heavy tailed property of wrapped Cauchy distribution.

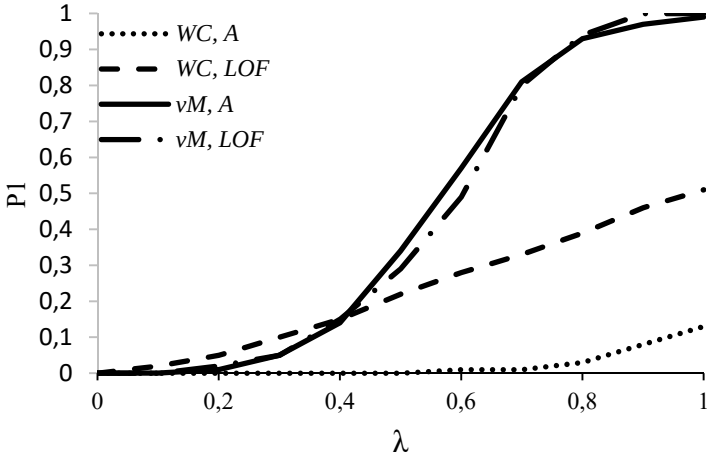


Figure 1. Performance of A statistic and *LOF*, for n =50, ρ =0.8 and κ=5

The performance of the *LOF* method has a direct relationship with the concentration parameter of circular sample as partially shown in Figure 2, while it has an inverse relationship with the sample size as shown in Figure 3. Moreover, in all considered cases, the performance has a direct relationship with the degree of contamination, λ.

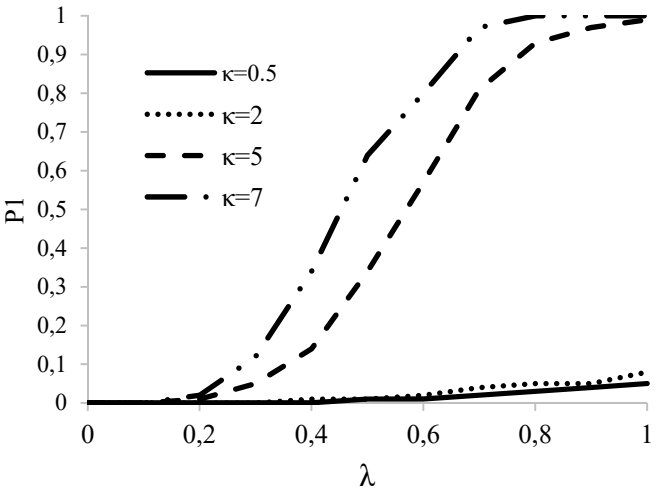


Figure 2. Performance of *LOF*, for samples of size n =50 from von Mises distribution

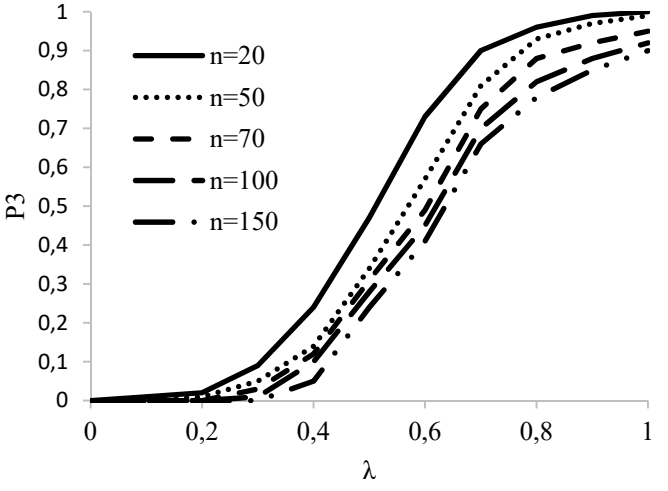


Figure 3. Performance of *LOF*, for samples with concentration parameter κ=5 from von Mises distribution

5. Practical examples

For illustration purposes, this section revisits two common circular datasets, which have been analysed in the context of outliers in circular data.

5.1. Frogs Data

Directions taken by 14 frogs after 30 hours of enclosure within a dark environmental chamber (Fergusion et al. 1967) are illustrated in Figure 4. The circular mean direction is 146.104° and the estimated concentration parameter is 2.18.

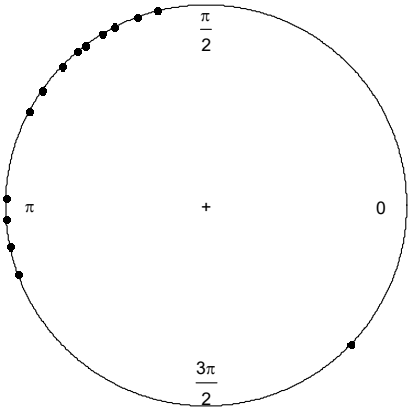


Figure 4. Circular plot of the frogs' directions, (n =14)

The results of applying seven discordancy tests on frogs' directions show that all tests except C statistic are consistent on identifying observation number 14 with value 316° (5.515 radians) as an outlier, which is apparent in Figure 4.

Table 1. Results of outlier-detection tests for frogs' directions, (n =14)

Statistic	*Statistic value*	*Observation*	*Cut-off point*	*Decision*
C	0.182	14	0.20	Not outlier
D	0.78	14	0.74	Outlier
M	0.52	14	0.50	Outlier
A	0.92	14	0.83	Outlier
G_1	2.03	14	1.69	Outlier
G_2	2.16	14	2.05	Outlier
LOF	1.88	14	1	Outlier

The values of *LOF* at k =10 are presented in Figure 5. It is shown that the *LOF* for all observations except the observation number 14 is close to one, which means that they are closed to each other and have similar density as their neighbours, while the *LOF* value of the observation number 14 is 1.88, which reveals that it has lower density than its neighbours.

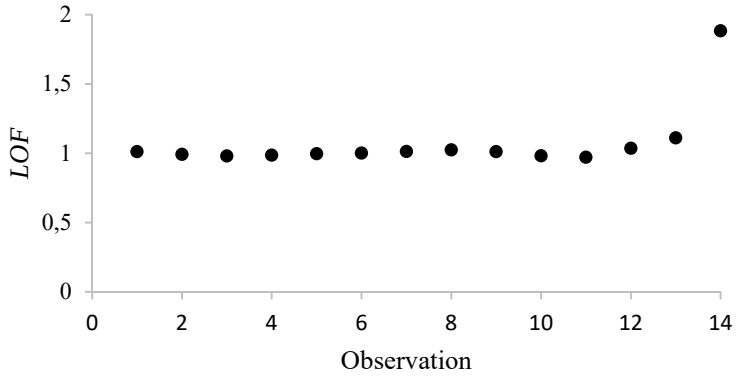


Figure 5. LOF for frogs' directions, k =10

5.2. Eye Data

Mohamed et al. (2016) considered the angle of posterior corneal curvature for 23 glaucoma patients as presented in Figure 6. The circular mean direction is 92° (1.61 radians) and the estimated concentration parameter is 6.84.

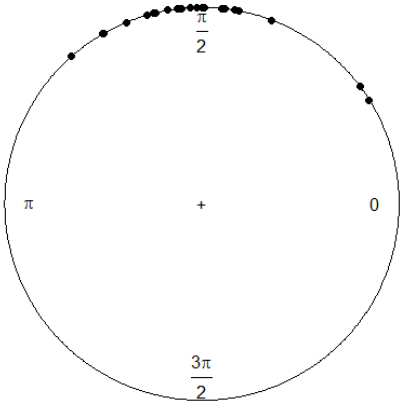


Figure 6. Circular plot of posterior corneal curvature, (n =23)

The results of applying discordancy tests on eye data show that only M statistic, G_2 and LOF at k =17 identified the observation number 17 as an outlier. Moreover, only G_2 and LOF identified the observation number 10 as an outlier. This reveals the weakness of most existing outlier-detection methods in the case of multiple outliers, which are apparently outliers from Figure 6.

Table 2. Results of outliers detection tests for eye dataset, (n =23)

Statistic	*Statistic value*	*Observation*	*Cut-off point*	*Decision*
C	0.02	17	0.03	Not outlier
D	0.04	17	0.18	Not outlier
M	0.31	17	0.12	Outlier
A	0.28	17	0.32	Not outlier
G_2	0.78	17	0.67	Outlier
G_2	0.68	10	0.67	Outlier
LOF	2.10	17	1	Outlier
LOF	1.97	10	1	Outlier

The values of LOF at k =17 are presented in Figure 7. It is shown that the LOF for all observations except the observation numbers 17 and 10 is close to one, which means that they are closed to each other and have similar density as their neighbours. On the other hand, the LOF values for the observation numbers 17 and 10 are 2.10 and 1.97, respectively, which reveals that they have lower density than their neighbours. Furthermore, the observation number 23 has a slightly high value of LOF and equals 1.36.

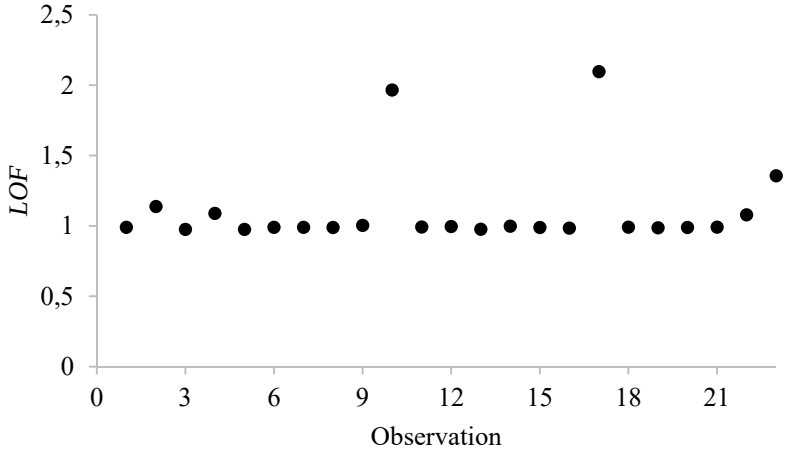


Figure 7. *LOF* for eye data, k =17

6. Conclusions

The presentation of angles in circular data as pairs of Cartesian coordinates gives a chance to use the *LOF* method for outlier detection. The *LOF* is a density-based method compared to the existing distribution-based methods. Furthermore, it does not

consider being an outlier as a binary property, while it gives the degree of being outlying.

The *LOF* performance is compatible with A test and then it outperforms the other tests of discordancy. The performance of the *LOF* has a direct relationship with the degree of contamination and concentration parameter, while it has an inverse relationship with the sample size.

The two considered practical examples illustrated the ability of *LOF* in dealing with multiple outliers compared to other existing outlier-detection methods.

The findings of this article pave the way to detect outliers in multivariate circular samples, either by representing their variates into pair coordinates, or by defining possible circular distances.

REFERENCES

ABUZAID, A. H. (2020). Identifying density-based local outliers in medical multivariate circular data. Statistics in Medicine. Vol. 39 (210), pp. 2793–2798. https://doi.org/10.1002/sim.8576.

ABUZAID, A. H., EL-HANJOURI, M. M., and KULLAB, M. M., (2015). On Discordance Tests for the Wrapped Cauchy Distribution, *Open Journal of Statistics*, Vol. 5 (4), pp. 245–253.

ABUZAID, A. H., MOHAMED, I. B., and HUSSIN, A. G., (2012). Boxplot for Circular Variables. *Computational Statistics*, Vol. 27 (3), pp. 381–392.

ABUZAID, A. H., RAMBLI, A., and HUSSIN, A.G., (2012). Statistics for a New Test of Discordance in Circular Data. *Communications in Statistics - Simulation and Computation*, Vol. 41 (10), pp. 1882–1890.

ABUZAID, A. M., MOHAMED, I. B., and HUSSIN, A. G., (2009). A new test of discordancy in circular data. *Communications in Statistics-Simulation and Computation,* Vol. 38(4), pp. 682–691.

ALKASADI, N. A., IBRAHIM, S. ABUZAID, A. H., and YUSOFF. M. I., (2019). Outliers Detection in Multiple Circular Regression Models Using *DFFITc* Statistic, *Sains Malaysiana*, Vol. 48 (7), pp. 1557–1563.

BARNETT, V. AND LEWIS, T., (1984). *Outliers in Statistical Data*. 2nd ed., John Wiley & Sons, Chichesters.

BREUNIG, M. M., KRIEGEL, H.-P., NG, R. T., and SANDER, JÖ., (2000). LOF: identifying density-based local outliers. *ACM sigmod record*, pp. 93–104.

COLLETT, D. (1980). Outliers in circular data. *Applied Statistics,* Vol. 29 (1), pp. 50–57.

DAS, M. K., GOGOI, B., (2015). Procedures of Outlier Detection in Cardioid Distribution, *Assam Statistical Review*, Vol. 29 (1), pp. 31–45.

FERGUSION, D. E., LANDRETH, H. F., MCKEOWN, J. P., (1967). Sun Compass Orientation of the Northern Cricket Frog. Acris Crepitans. *Animal Behavior,* Vol. 15, pp. 45–53.

MARDIA, K. V., (1975). Statistics of directional data. *Journal of the Royal Statistical Society, Series* B, Vol. 37, 349–393.

MOHAMED, I. B., RAMBLI, A., KHALIDDIN, N., and IBRAHIM, A. I. N., (2016). A New Discordancy Test in Circular Data Using Spacing's Theory, *Communications in Statistics - Simulation and Computation*, Vol. 45 (8), pp. 2904–2916.

SATARI, S. Z., HUSSIN, A. G, ZUBAIRI, Y. Z., and HASSAN, S. F., (2014). A New Functional Relationship Model For Circular Variables. *Pakistan Journal of Statistics*, Vol. 30 (3), pp. 397–410.

SIDIK, M. I., RAMBLI, A., MAHMUD, Z., REDZUAN, R. S., and SHAHRI, N., (2019). The Identification of Outliers in Wrapped Normal Data By Using Ga Statistics. *International Journal of Innovative Technology and Exploring Engineering*, Vol. 8 (4S), pp. 181–189.

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 3 pp. 53–71, DOI 10.21307/stattrans-2020-044
Received – 30.11.2017; accepted – 20.04.2020

A New Quasi Sujatha Distribution

Rama Shanker[1], Kamlesh Kumar Shukla[2]

ABSTRACT

The aim of this paper is to introduce a new quasi Sujatha distribution (NQSD), of which the following are particular cases: the Sujatha distribution devised by Shanker (2016 a), the size-biased Lindley distribution, and the exponential distribution. Its moments and moments-based measures are derived and discussed. Statistical properties, including the hazard rate and mean residual life functions, stochastic ordering, mean deviations, Bonferroni and Lorenz curves and stress-strength reliability are also analysed. The method of moments and the method of maximum likelihood estimations is discussed for estimating parameters of the proposed distribution. A numerical example is presented to test its goodness of fit, which is then compared with other one-parameter and two-parameter lifetime distributions.

Key words: Sujatha distribution, quasi Sujatha distribution, moments, reliability properties, stochastic ordering, stress-strength reliability, estimation of parameters, goodness of fit.

1. Introduction

The Sujatha distribution, introduced by Shanker (2016 a), is defined by its probability density function (pdf) and cumulative distribution function

$$f_1(x;\theta) = \frac{\theta^3}{\theta^2+\theta+2}\left(1+x+x^2\right)e^{-\theta x} \quad ; x>0,\ \theta>0\,. \tag{1.1}$$

$$F_1(x;\theta) = 1-\left[1+\frac{\theta x\left(\theta x+\theta+2\right)}{\theta^2+\theta+2}\right]e^{-\theta x};\quad x>0,\theta>0 \tag{1.2}$$

This distribution has been introduced for modelling lifetime data from engineering and biomedical science and it has been shown by Shanker (2016a) that it gives better fit

[1] Department of Statistics, Assam University, Silchar, India. E-mail: shankerrama2009@gmail.com. ORCID: https://orcid.org/0000-0002-5002-8904.
[2] Department of Statistics, Mainefhi College of Science, Asmara, Eritrea. E-mail: kkshukla222gmail.com. ORCID: https://orcid.org/0000-0001-5064-5569.

than both exponential and Lindley (1958) distributions. It is a convex combination of exponential (θ), gamma $(2,\theta)$ and gamma $(3,\theta)$ distributions.

The first four moments about origin of the Sujatha distribution (1.1) are obtained as

$$\mu_1' = \frac{\theta^2+2\theta+6}{\theta\left(\theta^2+\theta+2\right)}, \qquad \mu_2' = \frac{2\left(\theta^2+3\theta+12\right)}{\theta^2\left(\theta^2+\theta+2\right)},$$

$$\mu_3' = \frac{6\left(\theta^2+4\theta+20\right)}{\theta^3\left(\theta^2+\theta+2\right)}, \qquad \mu_4' = \frac{24\left(\theta^2+5\theta+30\right)}{\theta^4\left(\theta^2+\theta+2\right)}$$

The central moments of the Sujatha distribution (1.1) are obtained as

$$\mu_2 = \frac{\theta^4+4\theta^3+18\theta^2+12\theta+12}{\theta^2\left(\theta^2+\theta+2\right)^2}$$

$$\mu_3 = \frac{2\left(\theta^6+6\theta^5+36\theta^4+44\theta^3+54\theta^2+36\theta+24\right)}{\theta^3\left(\theta^2+\theta+2\right)^3}$$

$$\mu_4 = \frac{3\left(3\theta^8+24\theta^7+172\theta^6+376\theta^5+736\theta^4+864\theta^3+912\theta^2+480\theta+240\right)}{\theta^4\left(\theta^2+\theta+2\right)^4}.$$

Shanker (2016a) studied some of its important properties including skewness, kurtosis, index of dispersion, hazard rate function, mean residual life function, stochastic ordering, mean deviations, Bonferroni and Lorenz curves and stress-strength reliability. Shanker (2016a) discussed the estimation of parameter using maximum likelihood estimation and discussed the applications of the Sujatha distribution for modelling lifetime data from engineering and biomedical sciences. Shanker (2016b) has also obtained a Poisson mixture of the Sujatha distribution named "Poisson-Sujatha distribution (PSD)" and discussed its various properties, estimation of parameter and applications for counts data. Further, Shanker and Hagos (2016a, 2015) have obtained the size-biased and zero-truncated version of PSD, discussed their statistical properties, estimation of their parameter, and applications for modelling data which structurally excludes zero-counts. Shanker and Hagos (2016b) have a detailed and critical study on applications of zero-truncated Poisson, Poisson-Lindley and Poisson-Sujatha distributions.

Recently Shanker (2016 c) has introduced a quasi Sujatha distribution (QSD) having pdf and cdf

$$f_2\left(x;\theta,\alpha\right) = \frac{\theta^2}{\alpha\,\theta+\theta+2}\left(\alpha+\theta\,x+\theta\,x^2\right)e^{-\theta x}\ ;x>0,\ \theta>0\ \alpha>0 \qquad (1.3)$$

$$F_2(x;\theta,\alpha)=1-\left[1+\frac{\theta x(\theta x+\theta+2)}{\alpha\theta+\theta+2}\right]e^{-\theta x};\quad x>0,\theta>0,\alpha>0 \tag{1.4}$$

It can be easily verified that the Sujatha distribution and the size-biased Lindley distribution (SBLD) are particular cases of QSD at $\alpha=\theta$ and $\alpha=0$, respectively. Also, if $\alpha\to\infty$, QSD reduces to exponential distribution. Shanker (2016c) has studied its various mathematical and statistical properties including coefficient of variation, skewness, kurtosis, index of dispersion, hazard rate function, mean residual life function, stochastic ordering, mean deviations, Bonferroni and Lorenz curves and stress-strength reliability. The estimation of the parameters using both the maximum likelihood estimation and the method of moments has also been discussed and the goodness of fit has been discussed with a real lifetime data and compared with several well-known distributions.

The main motivation for searching a new two-parameter quasi Sujatha distribution (NQSD) is that the Sujatha distribution is a particular case of QSD whereas both Sujatha and exponential distributions are particular cases of NQSD and hence it is expected and hoped that NQSD will provide a better fit than QSD, Sujatha and exponential distributions.

In this paper, a new two-parameter quasi Sujatha distribution (NQSD) of which one-parameter Sujatha distribution introduced by Shanker (2016a) and exponential distribution are particular cases, has been proposed. Its raw moments and central moments have been obtained and coefficients of variation, skewness, kurtosis and index of dispersion have been discussed. Some of its important statistical properties including hazard rate function, mean residual life function, stochastic ordering, mean deviations, Bonferroni and Lorenz curves, and stress-strength reliability have also been discussed. The estimation of the parameters has been discussed using both the method of moments and the maximum likelihood estimation. A numerical example has been given to test the goodness of fit of the distribution and the fit has been compared with other well-known one-parameter and two-parameter lifetime distributions.

2. A New Quasi Sujatha Distribution

A two-parameter new quasi Sujatha distribution (NQSD) having parameters θ and α is defined by its pdf and cdf

$$f_3(x;\theta,\alpha)=\frac{\theta^3}{\theta^3+\alpha\theta+2\alpha}\left(\theta+\alpha x+\alpha x^2\right)e^{-\theta x}\ ;x>0,\ \theta>0\ \alpha>0. \tag{2.1}$$

$$F_3(x;\theta,\alpha)=1-\left[1+\frac{\alpha\theta x(\theta x+\theta+2)}{\theta^3+\alpha\theta+2\alpha}\right]e^{-\theta x};\quad x>0,\theta>0,\alpha>0 \tag{2.2}$$

It can be easily verified that the Sujatha distribution and exponential distribution are particular cases of NQSD at $\alpha = \theta$ and $\alpha = 0$ respectively. Like QSD, if $\alpha \to \infty$, NQSD reduces to exponential distribution. Further, it can be easily shown that NQSD (2.1) is a convex combination of exponential (θ) gamma $(2,\theta)$ and gamma $(3,\theta)$ distributions. We have

$$f_3(x;\theta,\alpha) = p_1\, g_1(x;\theta) + p_2\, g_2(x;\theta,2) + (1-p_1-p_2)\, g_3(x;\theta,3) \qquad (2.3)$$

where

$$p_1 = \frac{\theta^3}{\theta^3+\alpha\theta+2\alpha} \quad \text{and} \quad p_2 = \frac{\alpha\theta}{\theta^3+\alpha\theta+2\alpha}$$

$$g_1(x;\theta) = \theta e^{-\theta x}; x>0, \theta>0$$

$$g_2(x;\theta,2) = \frac{\theta^2}{\Gamma(2)} e^{-\theta x} x^{2-1}; x>0, \theta>0$$

$$g_3(x;\theta,3) = \frac{\theta^3}{\Gamma(3)} e^{-\theta x} x^{3-1}; x>0, \theta>0\,.$$

Graphs of the pdf of NQSD for varying values of parameters θ and α have been presented in Figure 1. Graphs of the cdf of NQSD for varying values of parameters θ and α have also been presented in Figure 2.

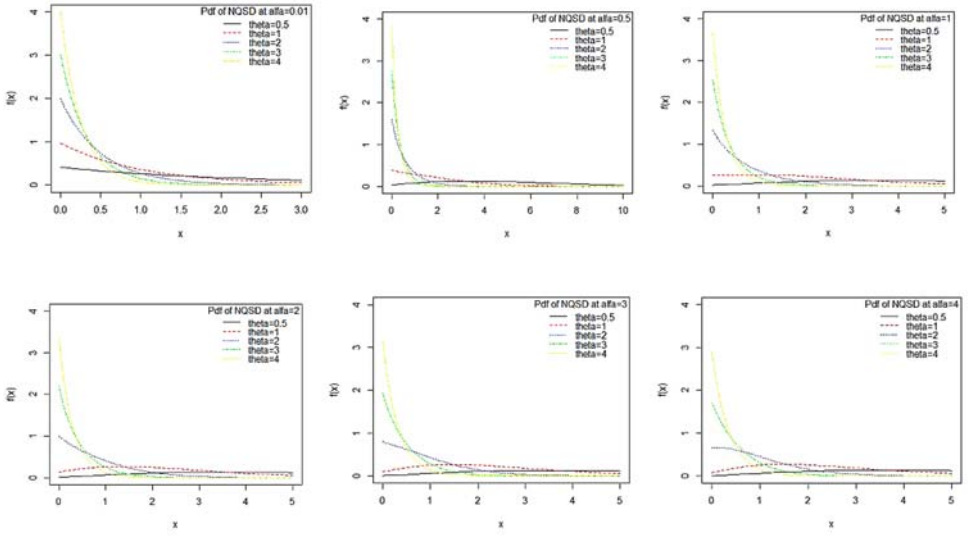


Figure 1. Graphs of the pdf of NQSD for varying values of parameters θ and α

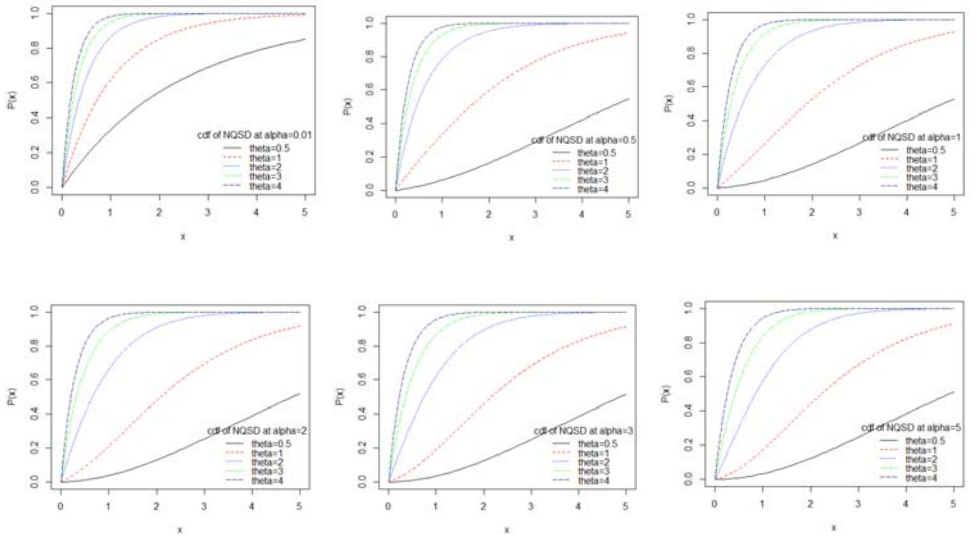


Figure 2. Graphs of the cdf of NQSD for varying values of parameters θ and α

3. Moments and Related Measures

Using the mixture representation (2.3), the r th moment about origin of NQSD (2.1) can be obtained as

$$\mu_r' = \frac{r!\left[\theta^3 + (r+1)\alpha\theta + (r+1)(r+2)\alpha\right]}{\theta^r\left(\theta^3 + \alpha\theta + 2\alpha\right)}; r = 1,2,3,... \tag{3.1}$$

The first four moments about origin of NQSD are thus obtained as

$$\mu_1' = \frac{\theta^3 + 2\alpha\theta + 6\alpha}{\theta\left(\theta^3 + \alpha\theta + 2\alpha\right)}, \qquad \mu_2' = \frac{2\left(\theta^3 + 3\alpha\theta + 12\alpha\right)}{\theta^2\left(\theta^3 + \alpha\theta + 2\alpha\right)},$$

$$\mu_3' = \frac{6\left(\theta^3 + 4\alpha\theta + 20\alpha\right)}{\theta^3\left(\theta^3 + \alpha\theta + 2\alpha\right)}, \qquad \mu_4' = \frac{24\left(\theta^3 + 5\alpha\theta + 30\alpha\right)}{\theta^4\left(\theta^3 + \alpha\theta + 2\alpha\right)}$$

Using the relationship between central moments and moments about origin, central moments of NQSD are obtained as

$$\mu_2 = \frac{\theta^2\left(\alpha^2 + 4\alpha + 2\right) + 16\theta\alpha + 12(\theta+1)}{\theta^2\left(\alpha\theta + \theta + 2\right)^2}$$

$$\mu_3 = \frac{2\left[\theta^3\left(\alpha^3+6\alpha^2+6\alpha+2\right)+6\theta^2\left(5\alpha^2+7\alpha+3\right)+36\theta\alpha+12\left(3\theta+2\right)\right]}{\theta^3\left(\alpha\theta+\theta+2\right)^3}$$

$$\mu_4 = \frac{3\left[\begin{array}{l}\theta^4\left(3\alpha^4+24\alpha^3+44\alpha^2+32\alpha+8\right)+8\theta^3\left(16\alpha^3+43\alpha^2+40\alpha+12\right)\\ +24\theta^2\left(17\alpha^2+32\alpha+14\right)+576\theta\alpha+240\left(2\theta+1\right)\end{array}\right]}{\theta^4\left(\alpha\theta+\theta+2\right)^4}$$

The coefficient of variation $\left(C.V\right)$, coefficient of skewness $\left(\sqrt{\beta_1}\right)$, coefficient of kurtosis $\left(\beta_2\right)$ and index of dispersion $\left(\gamma\right)$ of NQSD are given by

$$C.V = \frac{\sigma}{\mu_1'} = \frac{\sqrt{\theta^2\left(\alpha^2+4\alpha+2\right)+16\theta\alpha+12\left(\theta+1\right)}}{\alpha\theta+2\theta+6}$$

$$\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{2\left[\theta^3\left(\alpha^3+6\alpha^2+6\alpha+2\right)+6\theta^2\left(5\alpha^2+7\alpha+3\right)+36\theta\alpha+12\left(3\theta+2\right)\right]}{\left[\theta^2\left(\alpha^2+4\alpha+2\right)+16\theta\alpha+12\left(\theta+1\right)\right]^{3/2}}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} = \frac{3\left[\begin{array}{l}\theta^4\left(3\alpha^4+24\alpha^3+44\alpha^2+32\alpha+8\right)+8\theta^3\left(16\alpha^3+43\alpha^2+40\alpha+12\right)\\ +24\theta^2\left(17\alpha^2+32\alpha+14\right)+576\theta\alpha+240\left(2\theta+1\right)\end{array}\right]}{\left[\theta^2\left(\alpha^2+4\alpha+2\right)+16\theta\alpha+12\left(\theta+1\right)\right]^2}$$

$$\gamma = \frac{\sigma^2}{\mu_1'} = \frac{\theta^2\left(\alpha^2+4\alpha+2\right)+16\theta\alpha+12\left(\theta+1\right)}{\theta\left(\alpha\theta+\theta+2\right)\left(\alpha\theta+2\theta+6\right)}.$$

Note that at $\alpha=\theta$ and $\alpha=0$, these statistical constants reduce to the corresponding statistical constants of Sujatha and exponential distributions. Graphs for C.V, $\sqrt{\beta_1}$, β_2 and γ for varying values of parameters θ and α have been drawn and presented in Figure 3.

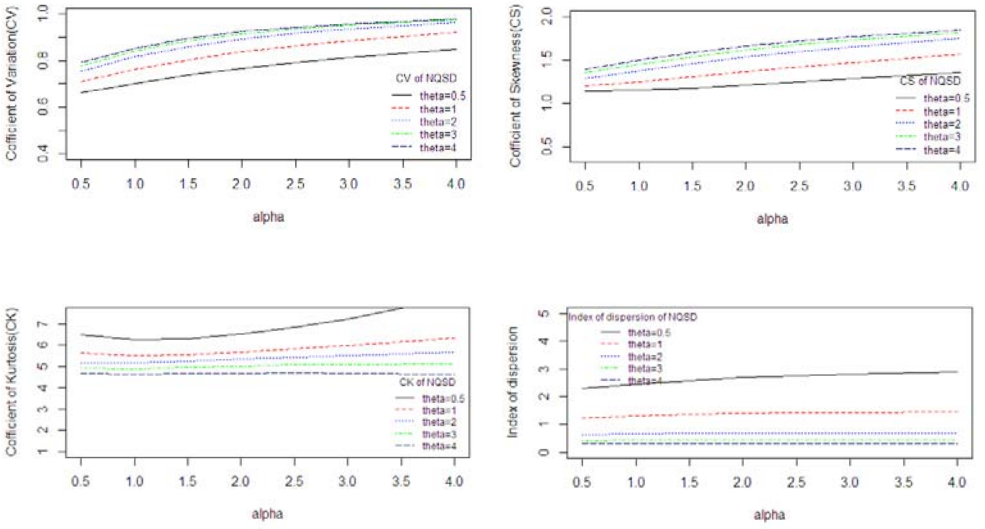


Figure 3. Graphs of C.V, C.S, C.K and I.D of NQSD for varying values of parameters θ and α

4. Hazard Rate Function and Mean Residual Life Function

The hazard rate function (also known as the failure rate function) and the mean residual life function of a continuous random variable X having pdf and cdf $f(x)$ and $F(x)$ are respectively defined as

$$h(x) = \lim_{\Delta x \to 0} \frac{P(X < x + \Delta x \mid X > x)}{\Delta x} = \frac{f(x)}{1 - F(x)} \tag{4.1}$$

and

$$m(x) = E[X - x \mid X > x] = \frac{1}{1 - F(x)} \int_x^{\infty} [1 - F(t)]\, dt \tag{4.2}$$

The corresponding $h(x)$ and $m(x)$ of NQSD (2.1) are obtained as

$$h(x) = \frac{\theta^3 (\theta + \alpha x + \alpha x^2)}{\alpha\theta x(\theta x + \theta + 2) + (\theta^3 + \alpha\theta + 2\alpha)} \tag{4.3}$$

and

$$m(x)=\frac{1}{\left[\alpha\theta x(\theta x+\theta+2)+\left(\theta^3+\alpha\theta+2\alpha\right)\right]e^{-\theta x}}\int_x^\infty\left[\alpha\theta t(\theta t+\theta+2)+\left(\theta^3+\alpha\theta+2\alpha\right)\right]e^{-\theta t}dt$$

$$=\frac{\alpha\theta x(\theta x+\theta+4)+\left(\theta^3+2\alpha\theta+6\alpha\right)}{\theta\left[\alpha\theta x(\theta x+\theta+2)+\left(\theta^3+\alpha\theta+2\alpha\right)\right]} \tag{4.4}$$

It is obvious that $h(0)=\dfrac{\theta^4}{\theta^3+\alpha\theta+2\alpha}=f(0)$ and $m(0)=\dfrac{\theta^3+2\alpha\theta+6\alpha}{\theta\left(\theta^3+\alpha\theta+2\alpha\right)}=\mu_1'$.

Graphs of $h(x)$ of NQSD for varying values of parameters θ and α are presented in Figure 4, whereas graphs of $m(x)$ of NQSD for varying values of parameters θ and α are presented in Figure 5. Graphs of $h(x)$ are either monotonically increasing or decreasing for varying values of parameters. Graphs of $m(x)$ are monotonically decreasing for varying values of parameters.

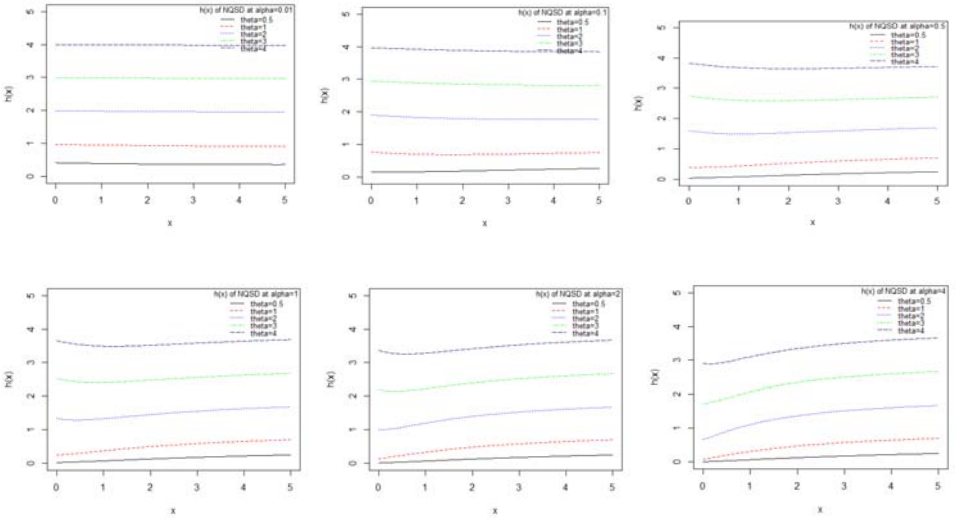


Figure 4. Graphs of $h(x)$ of NQSD for varying values of parameters θ and α

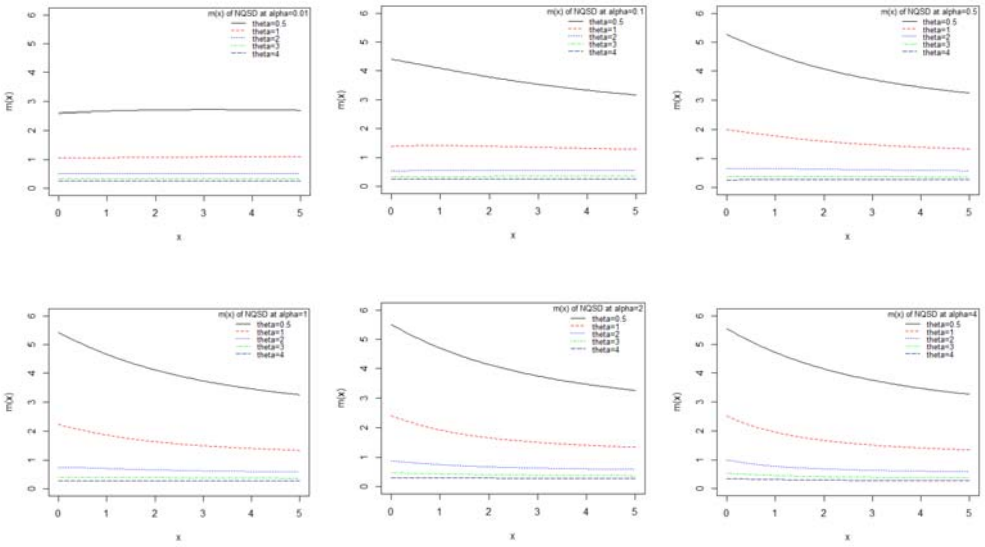


Figure 5. Graphs of $m(x)$ of NQSD for varying values of parameters θ and α

5. Stochastic Orderings

Stochastic ordering of positive continuous random variables is an important tool for judging their comparative behaviour. A random variable X is said to be smaller than a random variable Y in the:

(i) stochastic order $(X \leq_{st} Y)$ if $F_X(x) \geq F_Y(x)$ for all x

(ii) hazard rate order $(X \leq_{hr} Y)$ if $h_X(x) \geq h_Y(x)$ for all x

(iii) mean residual life order $(X \leq_{mrl} Y)$ if $m_X(x) \leq m_Y(x)$ for all x

(iv) likelihood ratio order $(X \leq_{lr} Y)$ if $\dfrac{f_X(x)}{f_Y(x)}$ decreases in x.

The following results due to Shaked and Shanthikumar (1994) are well known for establishing stochastic ordering of continuous distributions

$$X \leq_{lr} Y \Rightarrow X \leq_{hr} Y \Rightarrow X \leq_{mrl} Y$$
$$\Downarrow$$
$$X \leq_{st} Y$$

The NQSD is ordered with respect to the strongest 'likelihood ratio ordering' as established in the following theorem:

Theorem: Let $X \sim \text{NQSD}(\theta_1, \alpha_1)$ and $Y \sim \text{NQSD}(\theta_2, \alpha_2)$. If $\alpha_1 = \alpha_2$ and $\theta_1 > \theta_2$ or $\theta_1 = \theta_2$ and $\alpha_1 < \alpha_2$, then $X \leq_{lr} Y$ and hence $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$.

Proof: We have

$$\frac{f_X(x;\theta_1,\alpha_1)}{f_Y(x;\theta_2,\alpha_2)} = \frac{\theta_1^3\left(\theta_2^3 + \alpha_2\theta_2 + 2\alpha_2\right)}{\theta_2^3\left(\theta_1^3 + \alpha_1\theta_1 + 2\alpha_1\right)}\left(\frac{\theta_1 + \alpha_1 x + \alpha_1 x^2}{\theta_2 + \alpha_2 x + \alpha_2 x^2}\right)e^{-(\theta_1-\theta_2)x} \quad ; \; x > 0$$

Now

$$\ln\frac{f_X(x;\theta_1,\alpha_1)}{f_Y(x;\theta_2,\alpha_2)} = \ln\left[\frac{\theta_1^3\left(\theta_2^3 + \alpha_2\theta_2 + 2\alpha_2\right)}{\theta_2^3\left(\theta_1^3 + \alpha_1\theta_1 + 2\alpha_1\right)}\right] + \ln\left(\frac{\theta_1 + \alpha_1 x + \alpha_1 x^2}{\theta_2 + \alpha_2 x + \alpha_2 x^2}\right) - (\theta_1 - \theta_2)x .$$

This gives $$\frac{d}{dx}\ln\frac{f_X(x;\theta_1,\alpha_1)}{f_Y(x;\theta_2,\alpha_2)} = \frac{(\alpha_1\theta_2 - \alpha_2\theta_1) + 2(\alpha_1\theta_2 - \alpha_2\theta_1)x}{(\theta_1 + \alpha_1 x + \alpha_1 x^2)(\theta_2 + \alpha_2 x + \alpha_2 x^2)} - (\theta_1 - \theta_2).$$

Thus, if $\alpha_1 = \alpha_2$ and $\theta_1 > \theta_2$ or $\theta_1 = \theta_2$ and $\alpha_1 < \alpha_2$, $\frac{d}{dx}\ln\frac{f_X(x;\theta_1,\alpha_1)}{f_Y(x;\theta_2,\alpha_2)} < 0$. This means that $X \leq_{lr} Y$ and hence $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$. This shows flexibility of NQSD over the Sujatha distribution introduced by Shanker (2016 a) and exponential distributions.

6. Mean Deviations

The amount of scatter in a population is measured to some extent by the totality of deviations usually from the mean and the median, known as the mean deviation about the mean and the mean deviation about the median, and is defined by

$\delta_1(X) = \int_0^\infty |x - \mu| f(x)\,dx$ and $\delta_2(X) = \int_0^\infty |x - M| f(x)\,dx$, respectively, where $\mu = E(X)$ and $M = \text{Median}(X)$. The measures $\delta_1(X)$ and $\delta_2(X)$ can be calculated using the following simplified relationships

$$\delta_1(X) = \int_0^\mu (\mu - x) f(x)\,dx + \int_\mu^\infty (x - \mu) f(x)\,dx$$

$$= \mu F(\mu) - \int_0^{\mu} x\, f(x)\, dx - \mu\left[1 - F(\mu)\right] + \int_{\mu}^{\infty} x\, f(x)\, dx$$

$$= 2\mu F(\mu) - 2\mu + 2\int_{\mu}^{\infty} x\, f(x)\, dx$$

$$= 2\mu F(\mu) - 2\int_0^{\mu} x\, f(x)\, dx \tag{6.1}$$

and

$$\delta_2(X) = \int_0^{M} (M - x) f(x)\, dx + \int_{M}^{\infty} (x - M) f(x)\, dx$$

$$= M\, F(M) - \int_0^{M} x\, f(x)\, dx - M\left[1 - F(M)\right] + \int_{M}^{\infty} x\, f(x)\, dx$$

$$= -\mu + 2\int_{M}^{\infty} x\, f(x)\, dx$$

$$= \mu - 2\int_0^{M} x\, f(x)\, dx \tag{6.2}$$

Using pdf (2.1) and expression for the mean of NQSD, we get

$$\int_0^{\mu} x\, f_3(x;\theta,\alpha)\, dx = \mu - \frac{\left\{\begin{array}{l}\mu\theta^4 + \left(\alpha\mu^3 + \alpha\mu^2 + 1\right)\theta^3 + \left(3\alpha\mu^2 + 2\alpha\mu\right)\theta^2 \\ + \left(6\alpha\mu + 2\alpha\right)\theta + 6\alpha\end{array}\right\} e^{-\theta\mu}}{\theta\left(\theta^3 + \alpha\theta + 2\alpha\right)} \tag{6.3}$$

$$\int_0^{M} x\, f_3(x;\theta,\alpha)\, dx = \mu - \frac{\left\{\begin{array}{l}M\theta^4 + \left(\alpha M^3 + \alpha M^2 + 1\right)\theta^3 + \left(3\alpha M^2 + 2\alpha M\right)\theta^2 \\ + \left(6\alpha M + 2\alpha\right)\theta + 6\alpha\end{array}\right\} e^{-\theta M}}{\theta\left(\theta^3 + \alpha\theta + 2\alpha\right)} \tag{6.4}$$

Using expressions from (6.1), (6.2), (6.3) and (6.4), the mean deviation about mean $\delta_1(X)$ and the mean deviation about median $\delta_2(X)$ of NQSD are obtained as

$$\delta_1(X) = \frac{2\left\{\theta^3 + \mu(\mu+1)\alpha\theta^2 + 2(2\mu+1)\alpha\theta + 6\alpha\right\} e^{-\theta\mu}}{\theta\left(\theta^3 + \alpha\theta + 2\alpha\right)} \tag{6.5}$$

$$\delta_2(X)=\frac{2\left\{\begin{matrix}M\theta^4+\left(\alpha M^3+\alpha M^2+1\right)\theta^3+\left(3\alpha M^2+2\alpha M\right)\theta^2\\+\left(6\alpha M+2\alpha\right)\theta+6\alpha\end{matrix}\right\}e^{-\theta M}}{\theta\left(\theta^3+\alpha\theta+2\alpha\right)}-\mu \quad (6.6)$$

7. Bonferroni and Lorenz Curves

The Bonferroni and Lorenz curves (Bonferroni, 1930) and Bonferroni and Gini indices have applications not only in economics to study income and poverty, but also in other fields like reliability, demography, insurance and medicine. The Bonferroni and Lorenz curves are defined as

$$B(p)=\frac{1}{p\mu}\int_0^q x\,f(x)\,dx=\frac{1}{p\mu}\left[\int_0^\infty x\,f(x)\,dx-\int_q^\infty x\,f(x)\,dx\right]=\frac{1}{p\mu}\left[\mu-\int_q^\infty x\,f(x)\,dx\right] \quad (7.1)$$

and

$$L(p)=\frac{1}{\mu}\int_0^q x\,f(x)\,dx=\frac{1}{\mu}\left[\int_0^\infty x\,f(x)\,dx-\int_q^\infty x\,f(x)\,dx\right]=\frac{1}{\mu}\left[\mu-\int_q^\infty x\,f(x)\,dx\right] \quad (7.2)$$

respectively or equivalently

$$B(p)=\frac{1}{p\mu}\int_0^p F^{-1}(x)\,dx \quad (7.3)$$

and

$$L(p)=\frac{1}{\mu}\int_0^p F^{-1}(x)\,dx \quad (7.4)$$

respectively, where $\mu=E(X)$ and $q=F^{-1}(p)$.

The Bonferroni and Gini indices are thus defined as

$$B=1-\int_0^1 B(p)\,dp \quad (7.5)$$

and

$$G=1-2\int_0^1 L(p)\,dp \quad (7.6)$$

respectively.

Using pdf of NQSD (2.1), we get

$$\int_q^\infty x\, f_3(x;\theta,\alpha)\,dx = \frac{\left\{q\theta^4+\left(\alpha q^3+\alpha q^2+1\right)\theta^3+\left(3\alpha q^3+2\alpha q\right)\theta^2+\left(6\alpha q+2\alpha\right)\theta+6\alpha\right\}e^{-\theta q}}{\theta\left(\theta^3+\alpha\theta+2\alpha\right)} \tag{7.7}$$

Now, using equation (7.7) in (7.1) and (7.2), we get

$$B(p)=\frac{1}{p}\left[1-\frac{\left\{\begin{array}{l}q\theta^4+\left(\alpha q^3+\alpha q^2+1\right)\theta^3+\left(3\alpha q^3+2\alpha q\right)\theta^2\\+\left(6\alpha q+2\alpha\right)\theta+6\alpha\end{array}\right\}e^{-\theta q}}{\theta^3+2\alpha\theta+6\alpha}\right] \tag{7.8}$$

and $$L(p)=1-\frac{\left\{\begin{array}{l}q\theta^4+\left(\alpha q^3+\alpha q^2+1\right)\theta^3+\left(3\alpha q^3+2\alpha q\right)\theta^2\\+\left(6\alpha q+2\alpha\right)\theta+6\alpha\end{array}\right\}e^{-\theta q}}{\theta^3+2\alpha\theta+6\alpha} \tag{7.9}$$

Now, using equations (7.8) and (7.9) in (7.5) and (7.6), the Bonferroni and Gini indices of QSD are thus obtained as

$$B=1-\frac{\left\{\begin{array}{l}q\theta^4+\left(\alpha q^3+\alpha q^2+1\right)\theta^3+\left(3\alpha q^3+2\alpha q\right)\theta^2\\+\left(6\alpha q+2\alpha\right)\theta+6\alpha\end{array}\right\}e^{-\theta q}}{\theta^3+2\alpha\theta+6\alpha} \tag{7.10}$$

$$G=\frac{2\left\{\begin{array}{l}q\theta^4+\left(\alpha q^3+\alpha q^2+1\right)\theta^3+\left(3\alpha q^3+2\alpha q\right)\theta^2\\+\left(6\alpha q+2\alpha\right)\theta+6\alpha\end{array}\right\}e^{-\theta q}}{\theta^3+2\alpha\theta+6\alpha}-1 \tag{8.11}$$

8. Stress-Strength Reliability

The stress-strength reliability describes the life of a component which has random strength X that is subjected to a random stress Y. When the stress applied to it exceeds the strength, the component fails instantly and the component will function satisfactorily until $X > Y$. Therefore, $R = P(Y < X)$ is a measure of component

reliability and in the statistical literature it is known as stress-strength parameter. It has wide applications in almost all areas of knowledge especially in biomedical sciences and engineering.

Let X and Y be independent strength and stress random variables having NQSD (2.1) with parameter (θ_1, α_1) and (θ_2, α_2) respectively. Then, the stress-strength reliability R of NQSD can be obtained as

$$R = P(Y < X) = \int_0^\infty P(Y < X \mid X = x) f_X(x)\, dx$$

$$= \int_0^\infty f_3(x;\theta_1,\alpha_1)\, F_3(x;\theta_2,\alpha_2)\, dx$$

$$= 1 - \frac{\theta_1^3 \left[\begin{array}{l} 24\alpha_1\alpha_2\theta_2^2 + 6\{\alpha_1\alpha_2\theta_2^2 + \alpha_1\alpha_2\theta_2(\theta_2+2)\}(\theta_1+\theta_2) \\ +2\{\alpha_2\theta_1\theta_2^2 + \alpha_1\alpha_2\theta_2(\theta_2+2) + \alpha_1(\theta_2^3+\alpha_2\theta_2+2\alpha_2)\}(\theta_1+\theta_2)^2 \\ +\{\alpha_2\theta_1\theta_2(\theta_2+2) + \alpha_1(\theta_2^3+\alpha_2\theta_2+2\alpha_2)\}(\theta_1+\theta_2)^3 \\ +\theta_1(\theta_2^3+\alpha_2\theta_2+2\alpha_2)(\theta_1+\theta_2)^4 \end{array}\right]}{(\theta_1^3+\alpha_1\theta_1+2\alpha_1)(\theta_2^3+\alpha_2\theta_2+2\alpha_2)(\theta_1+\theta_2)^5}.$$

It can be easily verified that the above expression reduces to the corresponding expression for the Sujatha distribution and exponential distribution at $(\alpha_1 = \theta_1,\ \alpha_2 = \theta_2)$ and $(\alpha_1 = \alpha_2 = 0)$.

9. Estimation of Parameters

9.1. Method of Moments Estimates (MOME)

Since NQSD (2.1) has two parameters to be estimated, the first two moments about origin are required to estimate its parameters. Equating the population mean to the sample mean, we have

$$\bar{x} = \frac{\theta^3+2\alpha\theta+6\alpha}{\theta(\theta^3+\alpha\theta+2\alpha)} = \frac{\theta^3+\alpha\theta+2\alpha}{\theta(\theta^3+\alpha\theta+2\alpha)} + \frac{\alpha(\theta+4)}{\theta(\theta^3+\alpha\theta+2\alpha)}$$

$$\bar{x} = \frac{1}{\theta} + \frac{\alpha(\theta+4)}{\theta(\theta^3+\alpha\theta+2\alpha)}$$

$$\frac{\theta\overline{x}-1}{\theta+4}=\frac{\alpha}{\theta^3+\alpha\theta+2\alpha} \tag{9.1.1}$$

Again, replacing the second population moment with the corresponding sample moment, we have

$$m_2{}' = \frac{2\left(\theta^3+3\alpha\theta+12\alpha\right)}{\theta^2\left(\theta^3+\alpha\theta+2\alpha\right)} = \frac{2\left(\theta^3+\alpha\theta+2\alpha\right)}{\theta^2\left(\theta^3+\alpha\theta+2\alpha\right)} + \frac{4\alpha\left(\theta+5\right)}{\theta^2\left(\theta^3+\alpha\theta+2\alpha\right)}$$

$$= \frac{2}{\theta^2} + \frac{4\alpha\left(\theta+5\right)}{\theta^2\left(\theta^3+\alpha\theta+2\alpha\right)}$$

$$\frac{m_2{}'\theta^2-2}{4\left(\theta+5\right)}=\frac{\alpha}{\theta^3+\alpha\theta+2\alpha} \tag{9.1.2}$$

Equations (9.1.1) and (9.1.2) give the following cubic equation in θ

$$m_2{}'\theta^3+4\left(m_2{}'-\overline{x}\right)\theta^2-2\left(10\overline{x}-1\right)\theta+12=0 \tag{9.1.3}$$

Solving equation (9.1.3) using any iterative methods such as the Newton-Raphson method, the Regula-Falsi method or the Bisection method, the method of moments estimate (MOME) $\tilde{\theta}$ of θ can be obtained, and substituting the value of $\tilde{\theta}$ in equation (9.1.1) MOME $\tilde{\alpha}$ of α can be obtained as

$$\tilde{\alpha}=\frac{\left(1-\tilde{\theta}\,\overline{x}\right)\left(\tilde{\theta}\right)^3}{\overline{x}\left(\tilde{\theta}\right)^2+2\left(\overline{x}-1\right)\tilde{\theta}-6} \tag{9.1.4}$$

9.2. Maximum Likelihood Estimates (MLE)

Let $\left(x_1, x_2, x_3, \ldots, x_n\right)$ be a random sample from NQSD (2.1). The likelihood function L of (2.1) is given by

$$L=\left(\frac{\theta^3}{\theta^3+\alpha\theta+2\alpha}\right)^n \prod_{i=1}^{n}\left(\theta+\alpha x_i+\alpha x_i{}^2\right)e^{-n\theta\overline{x}}$$

The natural log likelihood function is thus obtained as

$$\ln L = n\ln\left(\frac{\theta^3}{\theta^3+\alpha\theta+2\alpha}\right)+\sum_{i=1}^{n}\ln\left(\theta+\alpha x_i+\alpha x_i{}^2\right)-n\theta\overline{x}$$

The maximum likelihood estimates (MLEs) $\hat{\theta}$ and $\hat{\alpha}$ of θ and α are then the solutions of the following log likelihood equations:

$$\frac{\partial \ln L}{\partial \theta} = \frac{3n}{\theta} - \frac{n\left(3\theta^2+\alpha\right)}{\theta^3+\alpha\theta+2\alpha} + \sum_{i=1}^{n} \frac{1}{\theta+\alpha x_i+\alpha x_i^{\,2}} - n\,\overline{x} = 0$$

$$\frac{\partial \ln L}{\partial \alpha} = -\frac{n\left(\theta+2\right)}{\theta^3+\alpha\theta+2\alpha} + \sum_{i=1}^{n} \frac{x_i+x_i^{\,2}}{\theta+\alpha x_i+\alpha x_i^{\,2}} = 0\,,$$

where $\overline{x}$ is the sample mean.

These two natural log likelihood equations do not seem to be solved directly because they are not in closed forms. However, Fisher's scoring method can be applied to solve these equations. We have

$$\frac{\partial^2 \ln L}{\partial \theta^2} = -\frac{3n}{\theta^2} + \frac{n\left(3\theta^4+\alpha^2-12\alpha\theta\right)}{\left(\theta^3+\alpha\theta+2\alpha\right)^2} - \sum_{i=1}^{n} \frac{1}{\left(\theta+\alpha x_i+\alpha\, x_i^{\,2}\right)^2}$$

$$\frac{\partial^2 \ln L}{\partial \alpha^2} = \frac{n\left(\theta+2\right)^2}{\left(\theta^3+\alpha\theta+2\alpha\right)^2} - \sum_{i=1}^{n} \frac{\left(x_i+x_i^{\,2}\right)^2}{\left(\theta+\alpha x_i+\alpha x_i^{\,2}\right)^2}$$

$$\frac{\partial^2 \ln L}{\partial \theta\,\partial \alpha} = -\frac{2n\theta^2\left(\theta+3\right)}{\left(\theta^3+\alpha\theta+2\alpha\right)^2} - \sum_{i=1}^{n} \frac{\left(x_i+x_i^{\,2}\right)}{\left(\theta+\alpha x_i+\alpha\, x_i^{\,2}\right)^2}$$

The following equations can be solved for MLEs $\hat{\theta}$ and $\hat{\alpha}$ of θ and α of NQSD

$$\left[\begin{array}{cc} \dfrac{\partial^2 \ln L}{\partial \theta^2} & \dfrac{\partial^2 \ln L}{\partial \theta\,\partial \alpha} \\ \dfrac{\partial^2 \ln L}{\partial \theta\,\partial \alpha} & \dfrac{\partial^2 \ln L}{\partial \alpha^2} \end{array}\right]_{\substack{\hat{\theta}=\theta_0 \\ \hat{\alpha}=\alpha_0}} \left[\begin{array}{c} \hat{\theta}-\theta_0 \\ \hat{\alpha}-\alpha_0 \end{array}\right] = \left[\begin{array}{c} \dfrac{\partial \ln L}{\partial \theta} \\ \dfrac{\partial \ln L}{\partial \alpha} \end{array}\right]_{\substack{\hat{\theta}=\theta_0 \\ \hat{\alpha}=\alpha_0}}$$

where θ_0 and α_0 are the initial values of θ and α, respectively, as given by the method of moments. These equations are solved iteratively until sufficiently close values of $\hat{\theta}$ and $\hat{\alpha}$ are obtained.

10. An Illustrative Example

A numerical example of real lifetime data has been presented to test the goodness of fit of NQSD over other one-parameter and two-parameter lifetime distribution.

The following data represent the tensile strength, measured in GPa, of 69 carbon fibres tested under tension at gauge lengths of 20mm, available in Bader and Priest (1982)

1.312	1.314	1.479	1.552	1.700	1.803	1.861	1.865	1.944	1.958	1.966	1.997
2.006	2.021	2.027	2.055	2.063	2.098	2.140	2.179	2.224	2.240	2.253	2.270
2.272	2.274	2.301	2.301	2.359	2.382	2.382	2.426	2.434	2.435	2.478	2.490
2.511	2.514	2.535	2.554	2.566	2.570	2.586	2.629	2.633	2.642	2.648	2.684
2.697	2.726	2.770	2.773	2.800	2.809	2.818	2.821	2.848	2.880	2.954	3.012
3.067	3.084	3.090	3.096	3.128	3.233	3.433	3.585	3.585			

For this data set, NQSD has been fitted along with one-parameter exponential, Lindley and Sujatha distributions and two-parameter QSD. The ML estimates of parameters, values of $-2\ln L$, AIC (Akaike Information Criterion), AICC (Akaike Information Criterion Corrected), BIC (Bayesian Information Criterion) and K-S Statistic (Kolmogorov-Smirnov Statistic) for the considered data set have been computed and presented in Table 1. The formulae for computing AIC, AICC, BIC and K-S Statistic (Kolmogorov-Smirnov Statistic) are as follows:

$AIC = -2\ln L + 2k$, $AICC = AIC + \dfrac{2k(k+1)}{(n-k-1)}$, $BIC = -2\ln L + k\ln n$, and $K\text{-}S = \sup_x \left|F_n(x) - F_0(x)\right|$, where k is the number of parameters involved in the respective distributions, n is the sample size and $F_n(x)$ is the empirical distribution function. The distribution corresponding to the lower values of $-2\ln L$, AIC, AICC, BIC and K-S statistic is the best fit distribution.

Table 1: MLE's, - 2ln L, AIC, AICC, BIC, and K-S of the fitted distributions

Distributions	ML Estimates		$-2\ln L$	AIC	AICC	BIC	KS
	$\hat{\theta}$	$\hat{\alpha}$					
NQSD	1.0693	40.01604	199.36	205.36	205.54	205.36	0.332
QSD	0.44259	87.0494	264.72	268.72	268.90	270.72	0.448
Sujatha	0.93611	------	221.60	223.60	223.66	224.60	0.364
Lindley	0.65450	-------	238.38	240.38	240.44	241.37	0.401
Exponential	0.40794	--------	261.73	263.73	263.79	264.73	0.448

It is obvious from the above table that NQSD is the best distribution among the considered distributions for modelling the considered lifetime data from engineering. Therefore, NQSD can be one of the important lifetime distributions for lifetime data from engineering.

11. Concluding Remarks

A two-parameter new quasi Sujatha distribution (NQSD), which includes one-parameter Sujatha distribution and exponential distribution as particular cases, has been proposed and studied. Its mathematical properties including moments, coefficient of variation, skewness, kurtosis, index of dispersion, hazard rate function, mean residual life function, stochastic ordering, mean deviations, Bonferroni and Lorenz curves, and stress-strength reliability have been discussed. The method of moments and the method of maximum likelihood estimation have also been discussed for estimating its parameters. Finally, a numerical example of real lifetime data set has been presented to test the goodness of fit of NQSD over one-parameter exponential, Lindley, Sujatha distributions and two-parameter QSD.

Acknowledgements

Authors are grateful to the Editor-in-Chief of the Journal and anonymous reviewers for their minor comments which were really fruitful.

REFERENCES

BADER, M.G., PRIEST, A. M., (1982). Statistical aspects of fiber and bundle strength in hybrid composites, In: hayashi, T., Kawata, K. Umekawa, S. (Eds.), Progress in Science in Engineering Composites, ICCM-IV, Tokyo, pp. 1129–1136.

BONFERRONI, C. E., (1930). Elementi di Statistca generale, Seeber, Firenze.

GUPTA, R. D., KUNDU, D., (1999). Generalized Exponential Distribution, Australian & New Zealand Journal of Statistics, 41(2), pp. 173–188.

LINDLEY, D.V., (1958). Fiducial distributions and Bayes' theorem, Journal of the Royal Statistical Society, Series B, 20, pp. 102–107.

SHAKED, M., SHANTHIKUMAR, J. G., (1994). Stochastic Orders and Their Applications, Academic Press, New York.

SHANKER, R., (2016a). Sujatha distribution and Its Applications, Statistics in Transition new series, 17 (3), pp. 1–20.

SHANKER, R., (2016b). The discrete Poisson-Sujatha distribution, International Journal of Probability and Statistics, 5(1), pp. 1–9.

SHANKER, R., (2016c). A Quasi Sujatha Distribution "International Journal of Probability and Statistics, 5(4), pp. 89–100.

SHANKER, R., MISHRA, A., (2013). A Quasi Lindley Distribution, African Journal of Mathematics and Computer Science Research (AJMCSR), 6(4), pp. 64–71.

SHANKER, R., HAGOS, F., (2015). Zero-truncated Poisson-Sujatha distribution with Applications, Journal of Ethiopian Statistical Association, 24, pp. 55–63.

SHANKER, R., HAGOS, F., (2016a). Size-biased Poisson-Sujatha distribution with Applications, American Journal of Mathematics and Statistics, 6(4), pp. 145–154.

SHANKER, R., HAGOS, F., (2016b). On Zero-truncation of Poisson, Poisson-Lindley and Poisson-Sujatha distributions and their Applications, Biometrics and Biostatistics International Journal, 3(5), pp. 1–13.

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 3 pp. 73–91, DOI 10.21307/stattrans-2020-045
Received – 08.07.2019; accepted – 21.04.2020

Power Size Biased Two-Parameter Akash Distribution

Khaldoon Alhyasat[1], Ibrahim Kamarulzaman[2], Amer Ibrahim Al-Omari[3], Mohd Aftar Abu Bakar[4]

ABSTRACT

In this paper, the two-parameter Akash distribution is generalized to size-biased two-parameter Akash distribution (SBTPAD). A further modification to SBTPAD is introduced, creating the power size-biased two-parameter Akash distribution (PSBTPAD). Several statistical properties of PSBTPAD distribution are proved. These properties include the following: moments, coefficient of variation, coefficient of skewness, coefficient of kurtosis, the maximum likelihood estimation of the distribution parameters, and finally order statistics. Moreover, plots of the density and distribution functions of PSBTPAD are presented and a reliability analysis is considered. The Rényi entropy of PSBTPAD is proved and the application of real data is discussed.

Mathematics Subject Classification: 62E10, 62F15.

Key words: Akash distribution, two-parameter Akash distribution, size-biased distribution, moments, coefficient of variation, coefficient of skewness, coefficient of kurtosis, maximum likelihood estimation, entropy.

1. Introduction

Recently, it has been noted that there has been an increasing interest in suggesting new flexible distributions for explaining and fitting data in different fields of science such as medicine, pharmacy, environment and so on. Many authors have introduced several types of new flexible distributions such as weighted distributions. The weighted distributions are quite flexible for model specification and data interpretation.

[1] School of Mathematical Sciences, Faculty of Science and Technology, University Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia. E-mail: p89225@siswa.ukm.edu.my.

[2] School of Mathematical Sciences, Faculty of Science and Technology, University Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia.

[3] School of Mathematical Sciences, Faculty of Science and Technology, University Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia. E-mail: alomari_amer@yahoo.com. ORCID: https://orcid.org/0000-0002-6901-8263.

[4] Department of Mathematics, Faculty of Science, Al al-Bayt University, Mafraq, Jordan. E-mail: aftar@ukm.edu.my. ORCID: https://orcid.org/0000-0002-3009-6168.

Fisher (1934) was the first who introduced the concept of weighted distributions. He studied how the verification methods can affect the form of the distribution of recorded observations. Also, see Rao (1965), Patil and Rao (1978), Gupta and Keating (1986), Gupta and Kirmani (1990), and (Oluyede 1999).

For a non-negative continuous random variable Y with probability density function (pdf) $f(y)$, the pdf of the weighted random variable Y_w is defined as

$$f_w(y) = \frac{w(y) f(y)}{E[w(y)]} = \frac{w(y) f(y)}{\mu_w}, \tag{1}$$

where $w(y)$ is a non-negative weight function. A special case of Equation (1) arises when the weight function is $w(y) = y^\beta$. In this case the distribution is known as a size-biased distribution of order β with pdf given by

$$f_\beta(y) = \frac{y^\beta f(y)}{\int y^\beta f(y) dy},$$

where for $\beta = 1$ or 2, the resulting are known as the length-biased and area-biased distributions, respectively.

Saghir et al. (2017) proposed several weighted distributions. A size biased Ishita distribution is introduced by Al-Omari et al. (2019) as a generalization of the Ishita distribution. Haq et al. (2017) proposed Marshall-Olkin length-biased exponential distribution. Al-Omari and Alsmairan (2019) suggested a length-biased Suja distribution as a modification of the Suja distribution, which is suggested by Shanker (2017).

Shanker (2015) suggested a one-parameter Akash distribution (AD). Then, Shanker and Shukla (2017) generalized the AD to suggest a two-parameter Akash distribution (TPAD) with pdf given by

$$f(y;\theta,\alpha) = \frac{\theta^3}{\alpha\theta^2 + 2}\left(\alpha + y^2\right)e^{-\theta y},\ y > 0, \theta, \alpha > 0, \tag{2}$$

and a cumulative distribution function (cdf) defined as

$$F(y;\theta,\alpha) = 1 - \left[1 + \frac{\theta y(\theta y + 2)}{\alpha\theta^2 + 2}\right]e^{-\theta y},\ y > 0, \theta, \alpha > 0. \tag{3}$$

The mean of TPAD is given by $E(Y) = \mu = \frac{\alpha\theta^2 + 6}{\theta(\alpha\theta^2 + 2)}$.

Abebe and Shanker (2018) suggested a discrete Akash distribution. Shanker et al. (2018) proposed a two-parameter Poisson-Akash distribution. Shanker et al. (2016) considered Poisson-Akash distribution. Shanker et al. (2018) proposed a generalized Akash distribution. Tesfalem et al. (2019) suggested a weighted Quasi Akash

distribution. Shanker (2016) suggested Qausi Akash distribution. Shanker and Shukla (2017) introduced the power Akash distribution.

The main objective of this study is to add a more flexibility distribution for fitting real data in the field. This paper is organized as follows: in Section 2, the pdf and the cdf of SBTPAD and PSBTPAD are presented as well as the shapes of the distribution are illustrated for various parameters. In Section 3 we present some statistical properties of the PSBTPAD, including the rth moment, mean, variance, coefficients of variation, skewness and kurtosis. Also, some simulations results are presented to illustrate these properties. The maximum likelihood estimators of the distribution parameters are derived in Section 4. The distributions of order statistics and reliability analysis are introduced in Section 5. An application of real data set is presented in Section 6 for illustration. Finally, the main results and some conclusions are provided in Section 7.

2. Suggested distributions

This section presents the pdf and cdf of the suggested distributions. A random variable Y is said to have a size biased two-parameter Akash distribution (SBTPAD) if its probability density function is given by

$$f_{SBTPAD}(y;\theta,\alpha)=\frac{\theta^4 y\left(\alpha+y^2\right)}{\alpha\theta^2+6}e^{-\theta y},\ y>0,\alpha,\theta>0, \tag{4}$$

and a cumulative distribution function is in the form

$$F_{SBTPAD}(y;\theta,\alpha)=1-\frac{6+\theta\left[\theta^2 y(\alpha+y^2)+\theta(\alpha+3y^2)+6y\right]}{\alpha\theta^2+6}e^{-\theta y}. \tag{5}$$

It is easy to derive the pdf given in Equation (4) by utilizing Equations (1) and the pdf of the TPAD given in (2), with the mean of the TPAD.

In this paper we modified the SBTPAD to a power size biased two-parameter Akash distribution (PSBTPAD) Taking the power transformation $X=Y^{1/\beta}$ in (4) a pdf of a random variable X can be defined as

$$f_{PSBTPAD}(x;\theta,\alpha,\beta)=\frac{\beta\theta^4}{\alpha\theta^2+6}x^{2\beta-1}\left(\alpha+x^{2\beta}\right)e^{-\theta x^{\beta}},\ x>0,\alpha,\theta,\beta>0. \tag{6}$$

We would call the density in (6) as the power size biased two-parameter Akash distribution (PSBTPAD). It is easy to prove that $\int_0^{\infty} f(x;\theta,\alpha,\beta)dx=1$.

Shukla and Shanker (2018) proposed a power Ishita distribution. Ghitany et al. (2013) introduced power Lindley distribution. Al-Omari et al. (2019) proposed a power length-biased Suja distribution. The corresponding pdf of the PSBTPAD is

$$F_{PSBTPAD}(x;\theta,\alpha,\beta)=1-\frac{\alpha\theta^2\Gamma\left(2,x^{\beta}\theta\right)+\Gamma\left(4,x^{\beta}\theta\right)}{\alpha\theta^2+6},\quad x>0,\theta,\alpha,\beta>0, \tag{7}$$

where $\Gamma(n+1,z)=n!e^{-z}\sum_{r=0}^{n}\frac{z^r}{r!}$ is the incomplete Gamma function. The lower incomplete gamma function is $\Gamma(\alpha,x)=\int_0^x t^{\alpha-1}e^{-t}dt$.

Figures 1 and 2 illustrate the shape of the pdf and cdf of the PSBTPAD for various values of the distribution parameters.

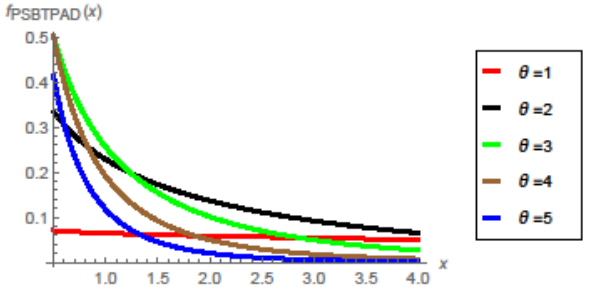


Figure 1. The pdf of PSBTPAD random variable X for $\theta=1,2,3,4,5$, $\alpha=1.7$ and $\beta=0.5$

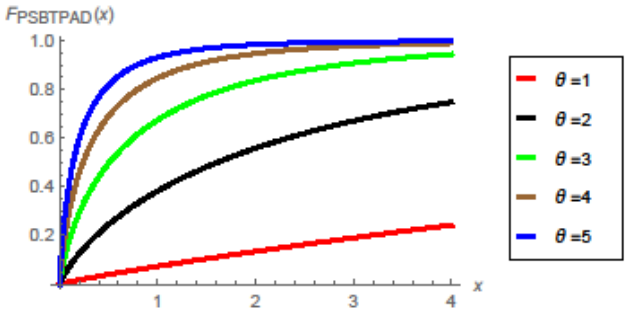


Figure 2. The cdf of PSBTPAD random variable X for $\theta=1,2,3,4,5$, $\alpha=1.7$ and $\beta=0.5$

Based on Figure 1, it can be seen that the PSBTPAD is asymmetric and skewed to the right.

3. Statistical properties

This section presents the *r*th moment, mean, variance, coefficients of variation, skewness and kurtosis of the PSBTPAD. Also, some simulations for these properties are provided.

3.1. Moments of the PSBTPAD

Theorem 2: Let $X \sim f_{PSBTPAD}(x;\theta,\alpha,\beta)$, then the rth moment of X about the origin is

$$E(X^r) = \frac{\theta^{-\frac{r}{\beta}}\Gamma\left(\frac{r}{\beta}+2\right)\left[\beta^2\left(\alpha\theta^2+6\right)+r^2+5\beta r\right]}{\beta^2\left(\alpha\theta^2+6\right)}, \tag{8}$$

for $2\beta + r > 0, \theta > 0, \beta > 0, r = 1,2,3,..$

Proof: By the expectation definition of the *r*th moment we have

$$\begin{aligned}
\mu^r_{PSBTPAD} &= E(X^r) = \int_0^\infty x^r \frac{\beta\theta^4}{\alpha\theta^2+6} x^{2\beta-1}\left(\alpha + x^{2\beta}\right)e^{-\theta x^\beta}dx \\
&= \frac{\beta\theta^4}{\alpha\theta^2+6}\left[\alpha\int_0^\infty x^{2\beta+r-1}e^{-\theta x^\beta}dx + \int_0^\infty x^{4\beta+r-1}e^{-\theta x^\beta}dx\right] \\
&= \frac{\beta\theta^4}{\alpha\theta^2+6}\left[\frac{\alpha\theta^{-\frac{r}{\beta}-2}\Gamma\left(\frac{r}{\beta}+2\right)}{\beta} + \frac{\theta^{-\frac{r}{\beta}-4}\Gamma\left(\frac{r}{\beta}+4\right)}{\beta}\right] \\
&= \theta^{-\frac{r}{\beta}}\Gamma\left(\frac{r}{\beta}+2\right)\left[\frac{\beta^2(\alpha\theta^2+6)+r^2+5r\beta}{\beta^2(\alpha\theta^2+6)}\right].
\end{aligned}$$

Based on Equation (8), it is simple to deduce the first, second, third and fourth moments of the BTPAD, respectively, as

$$E(X) = \frac{\theta^{-1/\beta}\Gamma\left(2+\frac{1}{\beta}\right)\left(\beta(\Psi+5)+1\right)}{\beta^2\left(\alpha\theta^2+6\right)},$$

$$E\left(X^2\right) = \frac{\theta^{-2/\beta}\Gamma\left(2+\frac{2}{\beta}\right)\left(\beta(\Psi+10)+4\right)}{\beta^2\left(\alpha\theta^2+6\right)},$$

$$E\left(X^3\right) = \frac{\theta^{-3/\beta}\Gamma\left(2+\frac{3}{\beta}\right)\left(\beta(\Psi+15)+9\right)}{\beta^2\left(\alpha\theta^2+6\right)},$$

$$E\left(X^4\right) = \frac{\theta^{-4/\beta}\Gamma\left(2+\frac{4}{\beta}\right)\left(\beta(\Psi+20)+16\right)}{\beta^2\left(\alpha\theta^2+6\right)},$$

where $\Psi = \beta\left(\alpha\theta^2 + 6\right)$. Hence, the variance of PSBTPAD is given by

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{\theta^{-2/\beta}\left[\beta\Psi\Gamma\left(2+\frac{2}{\beta}\right)\left(\beta(\Psi+10)+4\right) - \Gamma\left(2+\frac{1}{\beta}\right)^2\left(\beta(\Psi+5)+1\right)^2\right]}{\beta^2\Psi^2} \quad (9)$$

3.2. The coefficient of skewness

The coefficient of skewness determines the degree of skewness of SBTPAD. It is given by:

$$Sk_{PSBTPAD} = \frac{\theta^{-3/\beta}\left[\beta^2\Psi^2\Gamma\left(2+\frac{3}{\beta}\right)\left(\beta(\Psi+15)+9\right) - 3\beta\Psi\Phi\Gamma\left(2+\frac{2}{\beta}\right)\left(\beta(\Psi+10)+4\right) + 2\Phi^3\right]}{\beta^3\Psi^3\left[\frac{\theta^{-2/\beta}\left(\beta\Psi\Gamma\left(2+\frac{2}{\beta}\right)\left(\beta(\Psi+10)+4\right) - \Phi^2\right)}{\beta^2\Psi^2}\right]^{3/2}}, \quad (10)$$

where $\Psi = \beta\left(\alpha\theta^2 + 6\right)$ and $\Phi = \Gamma\left(2+\frac{1}{\beta}\right)\left(\beta(\Psi+5)+1\right)$

3.3. The coefficient of kurtosis

The coefficient of kurtosis measures the flatness of the distribution. The coefficient of kurtosis for PSBTPAD is defined as

$$Ku_{PSBTPAD} = \frac{\left\{\begin{array}{l}\beta^3\Psi^3\Gamma\left(2+\frac{4}{\beta}\right)\left(\beta(\Psi+20)+16\right) + 6\beta\,\Phi^2\Lambda \\ -3\Phi^4 - 4\beta^2\Psi^2\Phi\Gamma\left(2+\frac{3}{\beta}\right)\left(\beta(\Psi+15)+9\right)\end{array}\right\}}{\left(\Phi^2 - \beta\Lambda\right)^2} \quad (11)$$

where $\Psi = \beta\left(\alpha\theta^2 + 6\right)$, and $\Phi = \Gamma\left(2+\frac{1}{\beta}\right)\left(\beta(\Psi+5)+1\right)$,

$\Lambda = \Psi\Gamma\left(2+\frac{2}{\beta}\right)\left(\beta(\Psi+10)+4\right)$.

3.4. The coefficient of variation

The coefficient of variation of the PSBTPAD is given by

$$Cv_{PSBTPAD} = \frac{\beta\Psi\theta^{1/\beta}\sqrt{\dfrac{\theta^{-2/\beta}\left(\beta\Psi\Gamma\left(2+\dfrac{2}{\beta}\right)\left(\beta\left(\Psi+10\right)+4\right)-\Phi^2\right)}{\beta^4\left(\alpha\theta^2+6\right)^2}}}{\Phi}, \quad (12)$$

where $\theta>0, \beta>0.$

Theorem 2: Let $X \sim f_{PSBTPAD}(x;\theta,\alpha,\beta)$, then the harmonic mean of X is

$$H(\theta,\alpha,\beta) = \frac{\beta\Psi}{\theta^{1/\beta}\Gamma\left(2-\dfrac{1}{\beta}\right)\left[\beta\left(\Psi-5\right)+1\right]}, \theta>0, \beta>\frac{1}{2}. \quad (13)$$

To investigate the behaviour of these measures, we calculate some values of $\mu_{PSBTPAD}, \sigma_{PSBTPAD}$, $Cv_{PSBTPAD}$, $Sk_{PSBTPAD}$ and $Ku_{PSBTPAD}$ of the PSBTPAD for ($\theta=5$, $\beta=3$), ($\theta=5, \beta=7$), for various values of α and the results are presented in Tables 1 and 2, respectively.

Table 1. The mean, variance, coefficients of variation, skewness and kurtosis for the SBTPAD distribution for some values of α with $\theta=5$ and $\beta=3$

α	$\mu_{PSBTPAD}$	$\sigma_{PSBTPAD}$	$Cv_{PSBTPAD}$	$Sk_{PSBTPAD}$	$Ku_{PSBTPAD}$
1	0.736221	0.187730	0.254991	0.080619	2.79987
1.1	0.733241	0.186927	0.254933	0.085470	2.80863
1.2	0.730675	0.186195	0.254826	0.089115	2.81627
1.3	0.728442	0.185527	0.254690	0.091850	2.82290
1.4	0.726482	0.184916	0.254536	0.093889	2.82868
1.5	0.724746	0.184356	0.254373	0.095393	2.83371
1.6	0.723200	0.183842	0.254206	0.096480	2.83809
1.7	0.721813	0.183368	0.254039	0.097241	2.84192
1.8	0.720562	0.182931	0.253873	0.097745	2.84528
1.9	0.719427	0.182527	0.253711	0.098047	2.84822
2	0.718395	0.182151	0.253553	0.098188	2.85080
2.1	0.717450	0.181802	0.253400	0.098202	2.85308
2.2	0.716583	0.181477	0.253253	0.098113	2.85509
2.3	0.715784	0.181173	0.253111	0.097944	2.85687
2.4	0.715045	0.180888	0.252974	0.097711	2.85844
2.5	0.714361	0.180621	0.252843	0.097426	2.85984
2.6	0.713725	0.180370	0.252717	0.097102	2.86107
2.7	0.713132	0.180134	0.252596	0.096747	2.86218
2.8	0.712578	0.179912	0.252480	0.096368	2.86315

Table 2. The mean, variance, coefficients of variation, skewness and kurtosis for the SBTPAD distribution for some values of α with $\theta=5$ and $\beta=7$

α	$\mu_{PSBTPAD}$	$\sigma_{PSBTPAD}$	$Cv_{PSBTPAD}$	$Sk_{PSBTPAD}$	$Ku_{PSBTPAD}$
1	0.869614	0.098748	0.113554	-0.35893	3.12901
1.1	0.868112	0.098522	0.113490	-0.35521	3.13229
1.2	0.866818	0.098309	0.113413	-0.35250	3.13574
1.3	0.865692	0.098109	0.113330	-0.35056	3.13918
1.4	0.864704	0.097922	0.113244	-0.34918	3.14253
1.5	0.863829	0.097748	0.113157	-0.34824	3.14572
1.6	0.863049	0.097587	0.113072	-0.34761	3.14874
1.7	0.862350	0.097436	0.112989	-0.34725	3.15158
1.8	0.861719	0.097296	0.112909	-0.34707	3.15424
1.9	0.861147	0.097165	0.112832	-0.34705	3.15673
2	0.860626	0.097042	0.112758	-0.34714	3.15906
2.1	0.860150	0.096928	0.112687	-0.34732	3.16123
2.2	0.859713	0.096821	0.112620	-0.34757	3.16326
2.3	0.859310	0.096720	0.112555	-0.34787	3.16515
2.4	0.858938	0.096625	0.112494	-0.34821	3.16693
2.5	0.858593	0.096536	0.112435	-0.34859	3.16859
2.6	0.858272	0.096452	0.112380	-0.34899	3.17015
2.7	0.857973	0.096373	0.112326	-0.34940	3.17162
2.8	0.857693	0.096298	0.112275	-0.34982	3.17300

From Tables 1- 3 we can conclude the following:

1. For fixed values of α, the values of $\mu_{PSBTPAD}$ and $Ku_{PSBTPAD}$ of the PSBTPAD decrease as the values of β increase.
2. The $Cv_{PSBTPAD}$ values are about 0.25 when $\theta=5$ and $\beta=3$, and it is about 0.11 when $\theta=5$ and $\beta=7$.
3. The $Sk_{PSBTPAD}$ values are about 0.098 for all the parameter values in Table 1 and about -0.35 in for the parameters in Table 2. This indicates that the shape of the PSBTPAD depends on the parameter values.

4. Maximum likelihood estimation

Let $X_1, X_2, ..., X_n$ be a random sample of size n from PSBTPAD with parameters $\alpha>0$, $\beta>0$ and $\theta>0$. The maximum likelihood estimators for the parameters of PSBTPAD can be derived based on the likelihood function as

$$L(\theta,\alpha,\beta)=\prod_{i=1}^{n}\frac{\beta\theta^4}{\alpha\theta^2+6}x_i^{2\beta-1}\left(\alpha+x_i^{2\beta}\right)\mathrm{e}^{-\theta x_i^{\beta}}$$
$$=\left(\frac{\beta\theta^4}{\alpha\theta^2+6}\right)^n\prod_{i=1}^{n}x_i^{2\beta-1}\left(\alpha+x_i^{2\beta}\right)e^{\sum_{i=1}^{n}-\theta x_i^{\beta}}.$$

Then, the log likelihood function is given by

$$\ln L(\theta,\alpha,\beta)=\ln\left[\left(\frac{\beta\theta^4}{\alpha\theta^2+6}\right)^n\prod_{i=1}^{n}x_i^{2\beta-1}\left(\alpha+x_i^{2\beta}\right)e^{\sum_{i=1}^{n}-\theta x_i^{\beta}}\right]$$
$$=4n\ln(\theta)+n\ln(\beta)-n\ln\left(\alpha\theta^2+6\right)+\sum_{i=1}^{n}\ln\left(x_i^{2\beta-1}\right)+\sum_{i=1}^{n}\ln\left(\alpha+x_i^{2\beta}\right)-\sum_{i=1}^{n}\theta x_i^{\beta}. \quad (14)$$

Take the derivative of Equation (14) with respect to θ, α and β, respectively, as

$$\frac{\partial lnL(\theta,\alpha,\beta)}{\partial\theta}=\frac{4n}{\theta}-\frac{2n\alpha\theta}{\alpha\theta^2+6}-\sum_{i=1}^{n}x_i^{\beta}, \quad (15)$$

$$\frac{\partial\ln L(\theta,\alpha,\beta)}{\partial\alpha}=\frac{n}{\beta}-\frac{n\theta^2}{\alpha\theta^2+6}+\sum_{i=1}^{n}\frac{1}{\alpha+x_i^{2\beta}}, \quad (16)$$

and

$$\frac{\partial\ln L(\theta,\alpha,\beta)}{\partial\beta}=\frac{n}{\beta}+\sum_{i=1}^{n}\frac{2\ln(x_i)}{x_i^{2\beta-1}}+\sum_{i=1}^{n}\frac{2\ln(x_i)}{\alpha+x_i^{2\beta}}-\theta x_i^{\beta}\ln(x_i). \quad (17)$$

Since there is no closed form solutions for the above system of equations, the MLEs of the PSBTPAD parameters $\alpha,\theta,$ and β denoted as $\hat{\alpha},\hat{\theta}$ and $\hat{\beta}$, respectively, can be obtained by solving the equations $\frac{\partial lnL(\theta,\alpha,\beta)}{\partial\theta}=0$, $\frac{\partial lnL(\theta,\alpha,\beta)}{\partial\alpha}=0$, $\frac{\partial lnL(\theta,\alpha,\beta)}{\partial\beta}=0$ numerically.

5. Order statistics and reliability analysis

Let $X_1,X_2,...X_m$ be a random sample of size m from the power size biased two-parameter Akash distribution. Also, let $X_{(1:m)},X_{(2:m)},...,X_{(m:m)}$ denote the corresponding order statistics of the sample. The probability density function of the *i*th order statistic $X_{(i:m)}$ for $1\le i\le m$ is

$$f_{(i:m)}(x)=\frac{m!}{(i-1)(m-i)}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{m-i}f(x). \quad (18)$$

By substituting the pdf and cdf of the PSBTPAD in Equation (18), the pdf of $X_{(i:m)}$ is given by

$$f_{(i:m)}(x;\alpha,\theta,\beta)=\frac{\beta\theta^4 m!x^{2\beta-1}\left(\alpha+x^{2\beta}\right)e^{-\theta x^{\beta}}}{\left(\alpha\theta^2+6\right)\Gamma(i)\Gamma(-i+m+1)}\mathrm{H}, \tag{19}$$

where $\mathrm{H}=\left(1-\frac{\alpha\theta^2\Gamma\left(2,x^{\beta}\theta\right)-\Gamma\left(4,x^{\beta}\theta\right)}{\alpha\theta^2+6}\right)^{i-1}\left(\frac{\alpha\theta^2\Gamma\left(2,x^{\beta}\theta\right)+\Gamma\left(4,x^{\beta}\theta\right)}{\alpha\theta^2+6}\right)^{m-i}$.

Based on Equation (19) the pdfs of smallest order statistic, $X_{(1:m)}$ and largest order statistic, $X_{(m:m)}$, are respectively, given by

$$f_{(1:m)}(x;\alpha,\theta,\beta)=\frac{\beta\theta^4 m\,x^{2\beta-1}\left(\alpha+x^{2\beta}\right)e^{-\theta x^{\beta}}\left[\alpha\theta^2\Gamma\left(2,x^{\beta}\theta\right)+\Gamma\left(4,x^{\beta}\theta\right)\right]^{m-1}}{\left(\alpha\theta^2+6\right)^{m}}, \tag{20}$$

and

$$f_{(m:m)}(x;\alpha,\theta,\beta)=\frac{\beta\theta^4 m\,x^{2\beta-1}\left(\alpha+x^{2\beta}\right)e^{-\theta x^{\beta}}\left[\alpha\theta^2\left(1-\Gamma\left(2,x^{\beta}\theta\right)\right)-\Gamma\left(4,x^{\beta}\theta\right)+6\right]^{m-1}}{\left(\alpha\theta^2+6\right)^{m}} \tag{21}$$

The reliability and hazard rate functions of the PSBTPAD random variable are given by

$$\begin{aligned}R_{PSBTPAD}(x;\alpha,\theta,\beta) &= 1-F_{PSBTPAD}(x;\alpha,\theta,\beta)\\ &=\frac{\alpha\theta^2\Gamma\left(2,x^{\beta}\theta\right)+\Gamma\left(4,x^{\beta}\theta\right)}{\alpha\theta^2+6},\end{aligned} \tag{22}$$

$$\begin{aligned}H_{PSBTPAD}(x;\alpha,\theta,\beta) &= \frac{f_{PSBTPAD}(x;\alpha,\theta,\beta)}{1-F_{PSBTPAD}(x;\alpha,\theta,\beta)}\\ &=\frac{\beta\theta^4 x^{2\beta-1}\left(\alpha+x^{2\beta}\right)Exp\left(-\theta x^{\beta}\right)}{\alpha\theta^2\Gamma\left(2,x^{\beta}\theta\right)+\Gamma\left(4,x^{\beta}\theta\right)}.\end{aligned} \tag{23}$$

Figure (3) shows the reliability and hazard rate functions of the PSBTPAD with $\theta=1,2,3,4,5$, $\alpha=1.7$ and $\beta=0.5$.

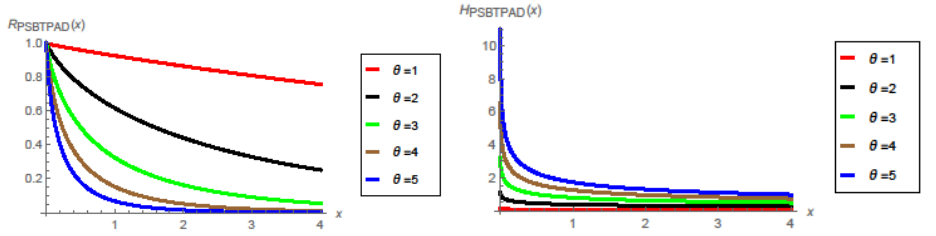


Figure 3. The reliability and hazard rate functions PSBTPAD for $\theta = 1,2,3,4,5$, $\alpha = 1.7$ and $\beta = 0.5$.

Figure (3) shows that the plots of the reliability and hazard rate functions of the PSBTAD are decreasing functions.

The reversed hazard rate and odds functions of the PSBTAD, respectively, are defined as

$$RH_{PSBTPAD}(x;\alpha,\theta,\beta) = \frac{f_{PSBTPAD}(x;\alpha,\theta,\beta)}{F_{PSBTPAD}(x;\alpha,\theta,\beta)} = \frac{\beta\theta^4 x^{2\beta-1}\left(\alpha + x^{2\beta}\right)Exp\left(-\theta x^{\beta}\right)}{\alpha\theta^2 + 6 - \alpha\theta^2\Gamma\left(2, x^{\beta}\theta\right) + \Gamma\left(4, x^{\beta}\theta\right)}, \tag{23}$$

and

$$O_{PSBTPAD}(x;\alpha,\theta,\beta) = \frac{F_{PSBTPAD}(x;\alpha,\theta,\beta)}{1 - F_{PSBTPAD}(x;\alpha,\theta,\beta)} = \frac{\alpha\theta^2 + 6 - \alpha\theta^2\Gamma\left(2, x^{\beta}\theta\right) - \Gamma\left(4, x^{\beta}\theta\right)}{\alpha\theta^2\Gamma\left(2, x^{\beta}\theta\right) + \Gamma\left(4, x^{\beta}\theta\right)}. \tag{24}$$

Figure (4) represents the reversed hazard and odds functions of the PSBTPAD distribution with $\theta = 1,2,3,4,5$, $\alpha = 1.7$ and $\beta = 0.5$.

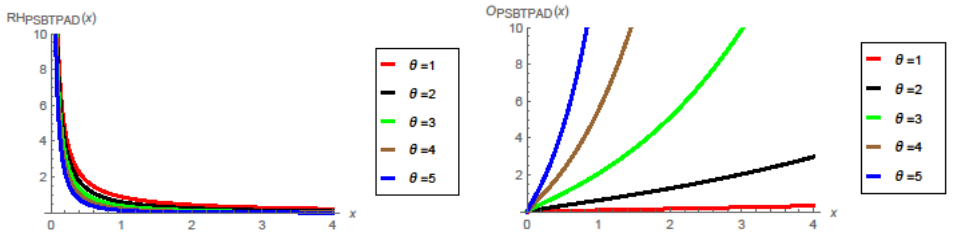


Figure **4.** The reversed hazard and odds functions of the PSBTPAD for $\theta = 1,2,3,4,5$, $\alpha = 1.7$ and $\beta = 0.5$.

The mean residual life function is defined as

$$m_{PSBTPAD}(x;\theta,\alpha,\beta) = E\left(X - x \mid X > x\right)$$
$$= \frac{1}{1 - F_{PSBTPAD}(x;\theta,\alpha,\beta)} \int_x^{\infty} \left(1 - F_{PSBTPAD}(t;\theta,\alpha,\beta)\right) dt$$
$$= \frac{1}{\alpha\theta^2\Gamma\left(2, x^{\beta}\theta\right) + \Gamma\left(4, x^{\beta}\theta\right)} \int_x^{\infty} \left(\alpha\theta^2\Gamma\left(2, x^{\beta}\theta\right) + \Gamma\left(4, x^{\beta}\theta\right)\right) dt.$$

The Mills ratio of the PSBTAD is defined as

$$MR_{PSBTPAD}(x;\alpha,\theta,\beta) = \frac{1}{RH_{PSBTPAD}(x;\alpha,\theta,\beta)}$$
$$= \frac{F_{PSBTPAD}(x;\alpha,\theta,\beta)}{f_{PSBTPAD}(x;\alpha,\theta,\beta)}$$
$$= \frac{\alpha\theta^2 + 6 - \alpha\theta^2\Gamma\left(2, x^{\beta}\theta\right) + \Gamma\left(4, x^{\beta}\theta\right)}{\beta\theta^4 x^{2\beta-1}\left(\alpha + x^{2\beta}\right) Exp\left(-\theta x^{\beta}\right)}$$

Plots of the Mills ratio of the PSBTAD are given in Figure (4) for various parameters.

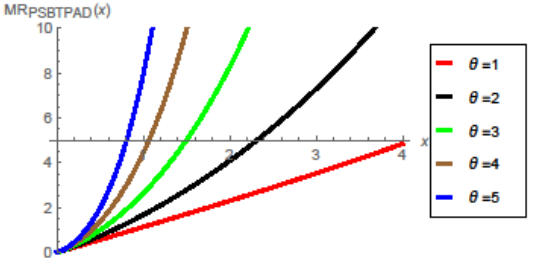


Figure 4. The Mills ratio of the PSBTPAD for $\theta = 1,2,3,4,5$, $\alpha = 1.7$ and $\beta = 0.5$

6. Rényi Entropy

The Rényi entropy (RE) of a random variable X is a measure of variation of the uncertainty. The RE is defined as $RE(\omega) = \frac{1}{1-\omega}\log\left(\int_0^{\infty} f(x)^{\omega} dx\right)$, $\omega > 0$ and $\omega \neq 1$. The entropy can be used for performing a goodness fit test. For more about

entropy see, for example, Al-Omari and Zamanzade (2017, 2018) for goodness of fit for Laplace and logistic distributions, respectively; Zamanzade and Mahdizadeh (2017) for entropy estimation using ranked set sampling; Zamanzade (2014) for testing uniformity using new entropy estimators, and Zamanzade and Arghami (2011) for goodness-of-fit test with correcting moments of modified entropy estimator; Al-Omari and Haq (2019) for novel entropy estimators of a continuous random variables.

Theorem 3: If $X \sim f_{PSBTPAD}(x;\theta,\alpha,\beta)$, the Rényi entropy of X is defined as

$$RE_{\text{PSBTPAD}}(\omega)=\frac{1}{1-\omega}\log\left[\begin{array}{l}\left(\frac{\omega^{-4}}{\alpha\theta^2+6}\right)^{\omega}\left(\beta(\omega\theta)^{\frac{1}{\beta}}\right)^{\omega\text{-}1}\times\\ \sum_{j=0}^{\omega}\binom{\omega}{j}(\alpha\theta^2\omega^2)^j\,\Gamma\left(-2\text{j}+4\omega-\frac{\omega-1}{\beta}\right)\end{array}\right]. \tag{25}$$

Proof: The Rényi entropy of the PSBTPAD can be obtained as

$$\begin{aligned}RE_{PSBTPAD}(\omega)&=\frac{1}{1-\omega}\log\left[\int_0^{\infty}\left(f_{PSBTPAD}(x;\theta,\alpha,\beta)\right)^{\omega}dx\right]\\&=\frac{1}{1-\omega}\log\left[\int_0^{\infty}\left(\frac{\beta\theta^4}{\alpha\theta^2+6}x^{2\beta-1}\left(\alpha+x^{2\beta}\right)Exp\left(-\theta x^{\beta}\right)\right)^{\omega}dx\right]\\&=\frac{1}{1-\omega}\log\left[\left(\frac{\beta\theta^4}{\alpha\theta^2+6}\right)^{\omega}\int_0^{\infty}\left(\alpha x^{2\beta-1}+x^{4\beta-1}\right)^{\omega}e^{-\theta\omega x^{\beta}}dx\right]\\&=\frac{1}{1-\omega}\log\left[\left(\frac{\beta\theta^4}{\alpha\theta^2+6}\right)^{\omega}\int_0^{\infty}\alpha^j\sum_{j=0}^{\omega}\binom{\omega}{j}x^{(2\beta-1)j}\left(x^{4\beta-1}\right)^{\omega-j}e^{-\theta\omega x^{\beta}}dx\right]\\&=\frac{1}{1-\omega}\log\left[\begin{array}{l}\left(\frac{\omega^{-4}}{\alpha\theta^2+6}\right)^{\omega}\left(\beta(\omega\theta)^{\frac{1}{\beta}}\right)^{\omega\text{-}1}\sum_{j=0}^{\omega}\binom{\omega}{j}(\alpha\theta^2\omega^2)^j\\ \Gamma\left(-2\text{j}+4\omega-\frac{\omega-1}{\beta}\right)\end{array}\right].\end{aligned}$$

To investigate the behaviour of the PSBTPAD Rényi entropy, Tables 3 and 4 involve some Rényi entropy values of the PSBTPAD for some values of the distribution parameters.

Table 3. Rényi entropy values for the PSBTPAD with $\beta=\theta=2$, $\omega=9$ and $\alpha=2,3,...,46$

α	$RE_{PSBTPAD}(\omega)$	α	$RE_{PSBTPAD}(\omega)$	α	$RE_{PSBTPAD}(\omega)$
1	0.191233	17	0.037523	32	0.014837
2	0.167737	18	0.034947	33	0.014018
3	0.143430	19	0.032617	34	0.013246
4	0.123293	20	0.030499	35	0.012515
5	0.107162	21	0.028566	36	0.011823
6	0.094186	22	0.026795	37	0.011167
7	0.083608	23	0.025166	38	0.010544
8	0.074857	24	0.023663	39	0.009951
9	0.067515	25	0.022272	40	0.009387
10	0.061276	26	0.020981	41	0.008849
11	0.055914	27	0.019779	42	0.008336
12	0.051260	28	0.018658	43	0.007845
13	0.047184	29	0.017610	44	0.007376
14	0.043585	30	0.016627	45	0.006928
15	0.040386	31	0.015705	46	0.006497

Table 4. Rényi entropy values for the PSBTPAD with $\beta=3$, $\alpha=4$, $\omega=1.1$ and $\theta=1,2,...,45$

θ	$RE_{PSBTPAD}(\omega)$	θ	$RE_{PSBTPAD}(\omega)$	θ	$RE_{PSBTPAD}(\omega)$
1	0.56641	16	4.26337	31	5.36132
2	1.11634	17	4.36373	32	5.41414
3	1.62485	18	4.45843	33	5.46534
4	2.03848	19	4.54806	34	5.51501
5	2.37837	20	4.63313	35	5.56325
6	2.66442	21	4.71409	36	5.61013
7	2.91043	22	4.79132	37	5.65573
8	3.12584	23	4.86514	38	5.70012
9	3.31722	24	4.93583	39	5.74336
10	3.48928	25	5.00366	40	5.78551
11	3.64551	26	5.06884	41	5.82662
12	3.78853	27	5.13158	42	5.86674
13	3.92037	28	5.19205	43	5.90591
14	4.04264	29	5.25040	44	5.94419
15	4.15663	30	5.30678	45	5.98161

Based on Table 3, we can say that the RE values approach zero for $\beta=\theta=2$ and $\omega=9$ as α starts increasing from 2 up to 46. But from Table 4, the RE values are increasing as the values of θ are increasing for fixed values of $\beta=3$, $\alpha=4$ and $\omega=1.1$.

7. Application and goodness of fit

In this section, the proposed PSBTPAD is applied to model data. We compare the fits of the PSBTPAD model with

1) Sushila distribution (SD) suggested Shanker et al. (2013):

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

2) Akash distribution (AD) Shanker (2015):

$$f\left(x,a\right)=\frac{a^3}{a^2+2}\left(1+x^2\right)e^{-ax},x>0,a>0.$$

3) Size biased Akash distribution (SBAD):

$$f\left(x,a\right)=\frac{a^3}{a^2+2}\left(1+x^2\right)e^{-ax},x>0,a>0.$$

4) Two-parameters Akash distribution (TPAD) Shanker and Shukla (2017):

$$f_{TPAD}(x;\theta,\alpha)=\frac{\theta^3}{\alpha\theta^2+2}\left(\alpha+x^2\right)e^{-\theta x};\ x>0,\ \theta>0.$$

5) Two-parameter quasi Akash distribution (TPQAD):

$$f_{TPQAD}(x;\theta,\alpha)=\frac{\theta^2}{\alpha\,\theta+2}\left(\alpha+\theta\,x^2\right)e^{-\theta x};\ x>0,\ \theta>0.$$

6) Marshall-Olkin Esscher Transformed Laplace distribution (MOETL), Georgea and Georgea (2013):

$$f(x)=\frac{\lambda k}{1+k^2}\begin{cases}Exp\left(\frac{\lambda}{k}x\right), & x<0\\ Exp\left(-k\lambda x\right), & x\geq 0.\end{cases}$$

We considered the negative maximized log-likelihood values (-MLL), Hannan-Quinn Information Criterion (HQIC), Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), Consistent Akaike Information Criterion (CAIC) and Kolmogorov-Smirnov (K-S) test statistic. These measures are defined as

$$AIC=-2MLL+2i, CAIC=-2MLL+\frac{2in}{n-i-1},$$

$$BIC=-2MLL+iLog(n) \text{ and } HQIC=2ln\left[ln(n)(i-2MLL)\right],$$

where i is the number of parameters and n is the sample size. Also, the Kolmogorov-Smirnov (KS) test is defined as $KS = Sup_n \left|F_n(x) - F(x)\right|$, where $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{x_i \le x}$ is the empirical distribution function and $F(x)$ is the cumulative distribution function. In general, lesser values of the above measures indicate a better fit of the model to the data set. The data set represent the strength data of glass of the aircraft window reported by Fuller et al. (1994). The data are as follows:

18.83, 20.80, 21.657, 23.03, 23.23, 24.05, 24.321, 25.5, 25.52, 25.80, 26.69, 26.77, 26.78, 27.05, 27.67, 29.90, 31.11, 33.2, 33.73, 33.76, 33.89, 34.76, 35.75, 35.91, 36.98, 37.08, 37.09, 39.58, 44.045, 45.29, 45.381.

Table 5. The -2LL, KS, P-value, AIC, CAIC, BIC, HQIC and the MLE based on the real data

Model	AIC	CAIC	BIC	HQIC	KS	P-Value	-2LL	MLE
AD	242.68	242.82	244.12	243.15	0.2987	0.0060	120.34	$\hat{\alpha} = 0.0971$
SD	256.48	256.91	259.35	257.42	0.3616	0.0004	126.24	$\hat{\alpha} = 0.1327$
SBAD	545.82	546.00	547.25	546.29	0.6472	3.4 e-13	271.91	$\hat{\alpha} = 0.1298$
								$\hat{\delta} = 0.0086$
MOETL	278.57	279.00	281.44	279.51	0.4585	1.8 e-06	137.29	$\hat{k} = -0.0262$
								$\hat{\lambda} = -1.2363$
TPAD	244.56	244.99	247.43	245.50	0.2902	0.0083	120.28	$\hat{\theta} = 0.0959$
								$\hat{\alpha} = 0.3316$
TPQAD	238.77	239.20	241.64	239.70	0.4520	2.7 e-06	117.38	$\hat{\theta} = 0.0904$
								$\hat{\alpha} = 11.7621$
PSBTPAD	**215.84**	**216.72**	**220.14**	**217.24**	**0.1074**	**0.8295**	**104.92**	$\hat{\theta} = 0.0052$
								$\hat{\alpha} = 0.5914$
								$\hat{\beta} = 1.9242$

Accordingly, the PSBTPAD is the appropriate model for fitting the data since it has the smallest values of AIC, CAIC, BIC, HQIC and KS with larger P-value as compared to the competitive models considered in this study.

7. Conclusions

In this paper, we proposed a new continuous distribution which generalizes the size biased two-parameter Akash distribution. The distribution is named power size biased two-parameter Akash distribution. Various statistical properties of the PSBTPAD are derived and discussed such as the moments, coefficient of variation, coefficient of

skewness, coefficient of kurtosis and the distribution of order statistics. The model parameters are estimated using the maximum likelihood estimation procedure. Finally, the distribution is fitted to real data. The new distribution is found to provide a better fit than its competitors used in this study.

Acknowledgements

The authors would like to thank the editor and anonymous referees for their several valuable comments and suggestions, which significantly improved this paper

REFERENCES

ABEBE, B., SHANKER , R., (2018). A discrete Akash distribution with applications. *Turkiye Klinikleri J Biostat*, 10(1), p. 1012.

AL-OMARI, A. I., ALHYASAT, K. M., IBRAHIM, K., ABU BAKER, M. A., (2019). Power length-biased Suja distribution: Properties and application. *Electronic Journal of Applied Statistical Analysis*, Vol. 12(2), pp. 429–452.

AL-OMARI, A. I. AL-NASSER, A. D., CIAVOLINO, E., (2019). A size-biased Ishita distribution application to real data. *Quality and Quantity,* Vol. 53(1), pp. 493–512.

AL-OMARI, A. I., ALSMAIRAN, I. K., (2019). Length-biased Suja distribution: Properties and application. *Journal of Applied Probability and Statistics*, Vol. 14(3), pp. 95–116.

AL-OMARI, A.I., ZAMANZADE, E., (2017). Goodness-of-fit tests for Laplace distribution in ranked set sampling. *Revista Investigacin Operacional*, Vol. 38(4), pp. 366–276.

Al-OMARI, A. I., ZAMANZADE, E., (2018). Goodness of fit tests for logistic distribution based on Phi-divergence. *Electronic Journal of Applied Statistical Analysis*, Vol. 11(1), pp. 185–195.

AL-OMARI, A.I., HAQ, A., (2019). Novel entropy estimators of a continuous random variable. *International Journal of Modeling, Simulation, and Scientific Computing*, Vol. 10(2), 1950004.

FISHER, R. A., (1934). The effects of methods of ascertainment upon the estimation of frequencies. *Ann. Eugen*, Vol. 6, pp. 13–25.

FULLER, J.R., E.R., FRIEMAN, S.W., QUINN, J. B. QUINN, G. D. and CARTER, W. C., (1994). Fracture mechanics approach to the design of glass aircraft windows: A case study. *SPIE Proc*., 2286, pp. 419–430.

GEORGEA, D., GEORGEA, S., (2013). Marshall–Olkin Esscher transformed Laplace distribution and processes. *Brazilian Journal of Probability and Statistics*, 27(2), pp. 162–184.

GHITANY, M. E., AL-MMUTAIRI, D. K., BALAKRISHANAN, N., AL-ENEZI, L. J., (2013). Power Lindley distribution and Associated Inference. *Comput. Stat. Data Anal.*, Vol. 64, pp. 20–33.

GUPTA, R. C., KEATING, J. P., (1986). Relations for reliability measures under length biased sampling. *Scand. J. Stat*, Vol. 13, pp. 49–56.

GUPTA, R. C., KIRMANI, S. N., (1990). The role of weighted distributions in stochastic modeling. *Commun. Stat. - Theory Methods*, Vol. 19, pp. 3147–3162.

HAQ, M. A., USMAN, R. M., HASHMI, S. and AL-OMARI, A. I., (2017). The Marshall-Olkin length-biased exponential distribution and its applications. *JKSUS*, Vol. 31(2), pp. 246–251.

OLUYEDE, B. O., (1999). On inequalities and selection of experiments for length-biased distributions. *Probab. Eng. Inf. Sci*. Vol. 13(2), pp. 169–185.

PATIL, G. P., RAO, G. R., (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics*, Vol. 34, pp. 179–189.

RAO, R., (1965). On discrete distributions arising out of methods of ascertainment. *Sankhya. Series A*. Vol. 27(2), pp. 311–324.

SAGHIR, A., HAMEDANI, G.G. TAZEEM, S. and KHADIM, A., (2017). Weighted distributions: A brief review, perspective and characterizations, *IJSP*, Vol. 6(3), pp. 109–131.

SHANKER, R., (2015). Akash distribution and its applications. *International Journal of Probability and Statistics*, 4 (3), pp. 65–75.

SHANKER, R., (2016). A Quasi Akash distribution and its applications. *ASR,* 30(1), pp. 135–160.

SHANKER, R., (2017). Size-biased Poisson-Akash distribution and its applications. *International Journal of Statistics and Applications*, 7(6), pp. 289–297.

SHANKER, R., (2017). Suja distribution and its application. Int J Probab Stat, Vol. 6(2), pp. 11–19.

SHANKER, R., FESSHAYE, H. and TESFAZGHI, T., (2016). On Poisson-Akash distribution and its applications. *Biometrics & Biostatistics International Journal*, 3(6), pp. 1–9.

SHANKER, R., SHUKLA, K. K., (2017). On two-parameter Akash distribution. *BBIJ,* Vol. 6(5), pp. 1–11.

SHANKER, R., SHUKLA, K. K., and LEONIDA, T. A., (2017). A two-parameter Poisson-Akash distribution with properties and applications. International Journal of Probability and Statistics, 7(4), pp. 114–123.

SHANKER, R., SHUKLA, K. K., SHANKR, R. and PRATAB, A., (2018). A generalized Akash distribution. *Biometrics & Biostatistics International Journal*, 7(1), pp. 18–26.

SHANKER, R., SHUKLA, K.K., (2017). Power Akash Distribution and Its Application. *Journal of Applied Quantitative Methods*, 12(3), pp. 1–10.

SHUKLA, K. K., SHANKER, R., (2018). Power Ishita distribution and its application to model lifetime data. *Statistics in Transition New Series*, Vol. 19(1), pp. 135–148.

ZAMANZADE, E., MAHDIZADEH, M., (2017). Entropy estimation from ranked set samples with application to test of fit. *Revista Colombiana de Estadstica*, Vol. 40(2), pp. 223–241.

ZAMANZADE, E., (2014). Testing uniformity based on new entropy estimators, *Journal of Statistical Computation and Simulation*, Vol. 85(16), pp. 3191–3205.

ZAMANZADE, E., ARGHAMI, N. R., (2011). Goodness-of-fit test based on correcting moments of modified entropy estimator. *Journal of Statistical Computation and Simulation*, Vol. 81(12), pp. 2077–2093.

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 3 pp. 93–107, DOI 10.21307/stattrans-2020-046
Received – 21.11.2019; accepted – 14.06.2020

Statistical Properties and Estimation of Power-Transmuted Inverse Rayleigh Distribution

Amal S. Hassan[1], Salwa M. Assar[2], Ahmed M. Abdelghaffar[3]

ABSTRACT

A three-parameter continuous distribution is constructed, using a power transformation related to the transmuted inverse Rayleigh (TIR) distribution. A comprehensive account of the statistical properties is provided, including the following: the quantile function, moments, incomplete moments, mean residual life function and Rényi entropy. Three classical procedures for estimating population parameters are analysed. A simulation study is provided to compare the performance of different estimates. Finally, a real data application is used to illustrate the usefulness of the recommended distribution in modelling real data.

Key words: transmuted inverse Rayleigh, mean residual life function, maximum likelihood, percentiles.

1. Introduction

Trayer (1964) introduced an important model for lifetime analysis, known as the inverse Rayleigh **(IR)** distribution. The probability density function **(pdf)** and the cumulative distribution function **(cdf)** of a random variable Y have the IR distribution with scale parameter θ and are defined by:

$$f_{IR}(y;\theta) = 2\theta y^{-3} e^{-\theta y^{-2}}, \; ; y > 0, \; \theta > 0.$$

and

$$F_{IR}(y;\theta) = e^{-\theta y^{-2}}; \quad y, \theta > 0.$$

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

[2] Department of Mathematical Statistics, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt. E-mail: salwaassar@yahoo.com. ORCID: https://orcid.org/0000-0001-7450-7486.

[3] Central Bank of Egypt, Egypt. E-mail: ahmad.m.abdelghaffar@gmail.com. ORCID: https://orcid.org/0000-0002-7163-927X.

Voda (1972) studied some properties of the maximum likelihood (ML) of its scale parameter. Gharraph (1993) provided closed-form expressions for the mean, harmonic mean, geometric mean, mode and the median of the IR distribution. A lot of works have been done in the literature upon estimation of the IR distribution; the reader can refer to Mohsin and Shahbaz (2005), Soliman et al. (2010), Dey (2012), Sindhu et al. (2013), Fan (2015), Rasheed et al. (2015), Panwar et al. (2015), Rasheed and Aref (2016).

In recent years, a number of extensions for the IR distribution have been developed using different methods of generalization by several authors, see, for example, beta IR distribution (Leao et al.; 2013), transmuted IR **(TIR)** distribution (Ahmed et al. 2014), modified IR **(MIR)** distribution (Khan; 2014), transmuted modified IR **(TMIR)** distribution (Khan and King; 2015) transmuted exponentiated IR **(TEIR)** distribution (Haq; 2015), Kumaraswamy exponentiated IR **(KEIR)** distribution (Haq; 2016), weighted IR distribution (Fatima and Ahmad; 2017) and odd Fréchet IR distribution (Elgarhy and Alrajhi; 2018).

The power transformation **(PT)** methodology has been used in many statistical aspects, although PT has been first proposed by Box and Cox (1964). One of the most important uses of the PT methodology is developing new distributions out of well-known distributions by adding an additional parameter, which gives several desirable properties and more flexibility in the form of the hazard rate and density functions. Also, it offers a more flexible model that can describe different types of real data. So, our objective in this study is developing a power transmuted inverse Rayleigh **(PTIR)** distribution out of the TIR distribution via the PT technique. Several statistical properties and different methods of estimation are discussed to obtain the point estimators regarding the proposed distribution.

This paper is organized as follows. Section 2 introduces the formation of the PTIR model. The structural characteristics of the PTIR distribution are studied in Section 3. Section 4 discusses parameter estimators for the PTIR distribution based on ML, least squares and percentile methods. Simulation schemes are performed in Section 5. A real life data application illustrates the potential of the PTIR distribution compared with some other distributions in Section 6. The article ends with some concluding remarks.

2. Model Formulation

The TIR distribution is a generalization of the IR distribution using the quadratic rank transmutation map (see Ahmed et al. 2014). The cdf of the TIR distribution is given by:

$$F_{TIR}(y;\theta,\lambda) = e^{-\theta y^{-2}}(1+\lambda-\lambda e^{-\theta y^{-2}}),\ ;\theta>0, |\lambda| \le 1,\ \ y>0.$$

Here, we propose a new extension of the TIR distribution by considering $X = Y^{1/\beta}$, where the random variable Y follows the TIR distribution with parameters θ and λ. The distribution function of a random variable X has the PTIR distribution and is defined as follows:

$$F_{PTIR}(x;\theta,\lambda,\beta) = e^{-\theta x^{-2\beta}}(1+\lambda-\lambda e^{-\theta x^{-2\beta}});\ \theta,\beta>0, |\lambda| \le 1,\ x>0. \tag{1}$$

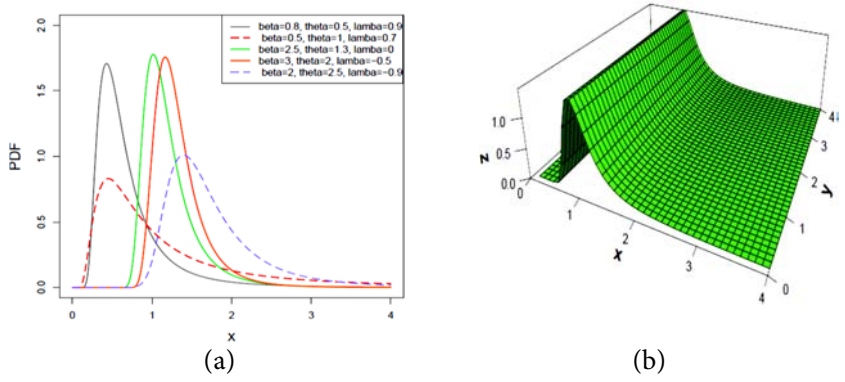
The pdf of the PTIR distribution corresponding to (1) is given by

$$f_{PTIR}(x;\theta,\lambda,\beta) = \frac{2\theta\beta}{x^{2\beta+1}} e^{-\theta x^{-2\beta}}(1+\lambda-2\lambda e^{-\theta x^{-2\beta}});\ \theta,\beta>0, |\lambda| \le 1,\ x>0. \tag{2}$$

A random variable X that follows the distribution (2) is denoted by $X \sim (\theta,\lambda,\beta)$. Two special sub models can be obtained from (2) as follows.

- For $\lambda = 0$, the pdf (2) reduces to a power IR **(PIR)** distribution as a new model.
- For $\lambda = 0$ and $\beta = 1$, the pdf (2) reduces to the IR distribution.

Some descriptive pdf plots of X have the PTIR distribution, which is illustrated in Figure 1 for some specific values of parameters.



(a) (b)

Figure 1. The pdf plots of the PTIR distribution (a) for some choices of parameters (b) for β=1.5, θ=1.0, λ=0.5

From Figure 1, it can be shown that the shape of the PTIR distribution is unimodal. It can also be said that the distribution is positively skewed.

Furthermore, the survival function and the hazard rate function **(hrf)** are given, respectively, by

$$S_{PTIR}(x;\theta,\lambda,\beta) = 1 - e^{-\theta x^{-2\beta}}(1+\lambda-\lambda e^{-\theta x^{-2\beta}}),$$

and

$$h_{PTIR}(x;\theta,\lambda,\beta) = 2\theta\beta e^{-\theta x^{-2\beta}} x^{-(2\beta+1)}(1+\lambda-2\lambda e^{-\theta x^{-2\beta}})\left(1-e^{-\theta x^{-2\beta}}(1+\lambda-\lambda e^{-\theta x^{-2\beta}})\right)^{-1}.$$

Some descriptive hrf plots of X are illustrated in Figure 2 for some specific values of parameters.

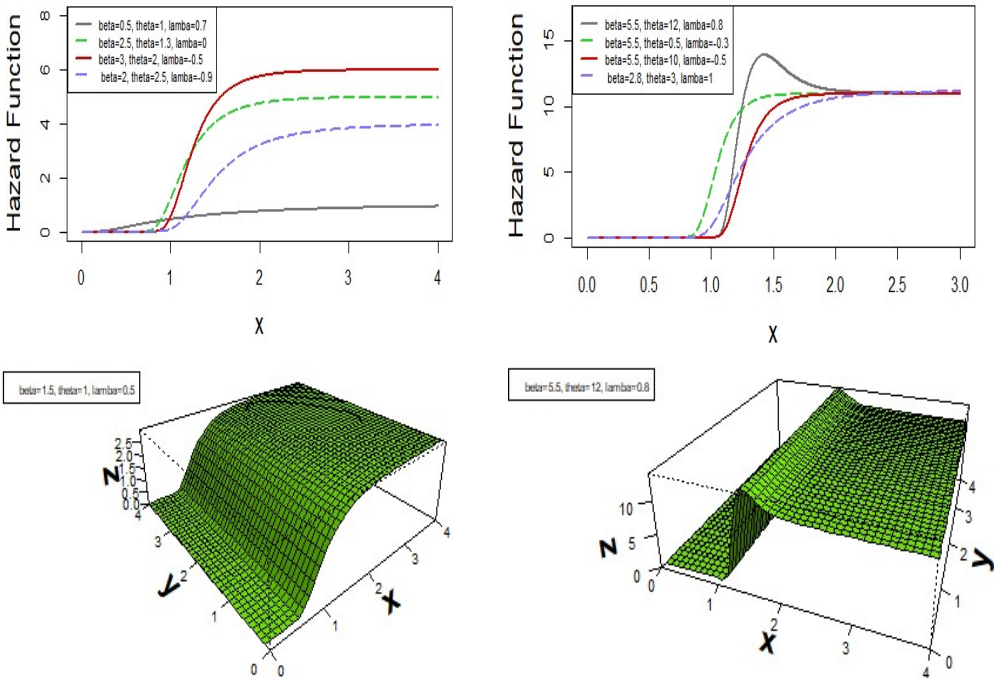


Figure 2. The hrf plots of the PTIR distribution for some choices of parameters

From Figure 2, it can be shown that the plots at several selected values of the parameters of hrfs have an increasing tendency.

The reversed hrf and cumulative hrf are given, respectively, by:

$$r_{PTIR}(x;\theta,\lambda,\beta) = 2\theta\beta x^{-2\beta-1},$$

and

$$H_{PTIR}(x;\theta,\lambda,\beta) = -\ln\left(1-e^{-\theta x^{-2\beta}}(1+\lambda-\lambda e^{-\theta x^{-2\beta}})\right).$$

3. Some Structural Properties

In this section some structural properties are provided.

3.1. Quantile Function

The quantile function of the PTIR distribution, say $Q(u) = F^{-1}(u)$ of X can be obtained by inverting (1) as follows:

$$\lambda\left(e^{-\theta/(Q(u))^{2\beta}}\right)^2 - (1+\lambda)e^{-\theta/(Q(u))^{2\beta}} + u = 0, \tag{3}$$

Factorizing (3) leads to

$$Q(u) = \left[-\theta \Big/ \ln\left[\frac{(1+\lambda) - \sqrt{(1+\lambda)^2 - 4\lambda u}}{2\lambda}\right]\right]^{\frac{1}{2\beta}}, \tag{4}$$

where u has a uniform random variable on (0, 1). Also, (4) can be used in simulating PTIR random variables when the parameters θ, λ and β are known. Median (***m***) of the distribution is obtained by setting $u = 0.5$ in (4). Also, the first and third quantiles can be obtained by setting $u = 0.25$ and $u = 0.75$ in (4).

3.2. Moments of the PTIR Distribution

Moments are used to understand various characteristics of a frequency distribution. They have been applied in order to obtain mean, variance, in addition to some measures, such as skewness and kurtosis.

The r^{th} moment of X has the PTIR distribution and is derived by using (2) as follows:

$$E(X^r) = \int_0^\infty x^r \left[\frac{2\theta\beta}{x^{2\beta+1}} e^{\frac{-\theta}{x^{2\beta}}} (1 + \lambda - 2\lambda e^{\frac{-\theta}{x^{2\beta}}})\right] dx. \tag{5}$$

Let $z = \theta x^{-2\beta}$, then the r^{th} moment of the PTIR distribution is given by

$$E(X^r) = \theta^{\frac{r}{2\beta}} \left[\left(\int_0^\infty z^{\frac{-r}{2\beta}} e^{-z} + \lambda \int_0^\infty z^{\frac{-r}{2\beta}} e^{-z} - 2\lambda \int_0^\infty z^{\frac{-r}{2\beta}} e^{-2z}\right) dz\right],$$

which is the gamma function, so the r^{th} moment can be formed as follows:

$$E(X^r) = \theta^{\frac{r}{2\beta}} \Gamma\left(1 - \frac{r}{2\beta}\right)\left[1 + \lambda - 2^{\frac{r}{2\beta}}\lambda\right], \; r < 2\beta, \; r = 1, 2, 3, \ldots$$

Hence, the mean and variance of the PTIR distribution are given, respectively, by

$$\mu = \theta^{\frac{1}{2\beta}} \Gamma\left(1 - \frac{1}{2\beta}\right)\left[1 + \lambda - 2^{\frac{1}{2\beta}}\lambda\right], \; \beta > 0.5,$$

and

$$Var(X) = \theta^{\frac{1}{\beta}}\Gamma\left(1-\frac{1}{\beta}\right)\left[1+\lambda-2^{\frac{1}{\beta}}\lambda\right]-\left(\theta^{\frac{1}{2\beta}}\Gamma\left(1-\frac{1}{2\beta}\right)\left[1+\lambda-2^{\frac{1}{2\beta}}\lambda\right]\right)^2,\ \beta>1.$$

Measurement of skewness and kurtosis of the distribution is obtained from complete moments using the well-known relationship. Plots of the PTIR skewness and kurtosis for some selected values are displayed in Figure 3.

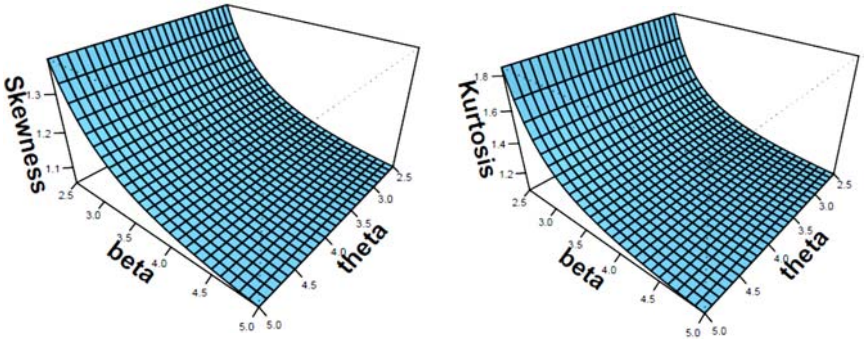


Figure 3. The skewness and kurtosis of the PTIR for $\lambda = 0.5$ and different values of θ and β

From Figure 3, it can be seen that both the skewness and the kurtosis are decreasing functions of θ, λ and β.

3.3. Incomplete Moments

The answer to many important questions in economics requires more than just knowing the mean of a distribution, but its shape as well. This is obvious not only in the study of econometrics and income distribution, but in other areas as well (see Butler and McDonald; 1989).

The s^{th} incomplete moment of a random variable X has the PTIR distribution and is obtained as follows:

$$£_{(s)}(t) = \int_0^t x^s f_{PTIR}(x;\theta,\lambda,\beta)dx = \int_0^t x^s\left[2\theta\beta x^{-(2\beta+1)}e^{\frac{-\theta}{x^{2\beta}}}(1+\lambda-2\lambda e^{\frac{-\theta}{x^{2\beta}}})\right]dx.$$

Let $z = \theta x^{-2\beta}$, then the s^{th} incomplete moment of the PTIR distribution is given by:

$$£_{(s)}(t) = \theta^{\frac{s}{2\beta}} \int_{\frac{\theta}{t^{2\beta}}}^{\infty} \left(z^{\frac{-s}{2\beta}} e^{-z} + \lambda z^{\frac{-s}{2\beta}} e^{-z} - 2\lambda z^{-\frac{s}{2\beta}} e^{-2z} \right) dz ,$$

which is the upper incomplete moments, so

$$£_{(s)}(t) = \theta^{\frac{s}{2\beta}} \left[\Gamma\left(1 - \frac{s}{2\beta}, \frac{\theta}{t^{2\beta}}\right) + \lambda\Gamma\left(1 - \frac{s}{2\beta}, \frac{\theta}{t^{2\beta}}\right) - 2^{\frac{s}{2\beta}} \lambda\Gamma\left(1 - \frac{s}{2\beta}, \frac{2\theta}{t^{2\beta}}\right) \right], \tag{6}$$

where $\Gamma(.,x)$ is the upper incomplete moments. The first incomplete moment can be obtained by setting $s = 1$ in (6). The mean deviation about the mean (μ), denoted by δ_1 , and the mean deviation about the median, denoted by δ_2, can be obtained, respectively, as follows:

$$\delta_1 = 2\mu F_{PTIR}(\mu) - 2£_{(1)}(\mu)$$

$$= 2\left(\theta^{\frac{1}{2\beta}} \Gamma\left(1 - \frac{1}{2\beta}\right) \left[1 + \lambda - 2^{\frac{1}{2\beta}} \lambda \right] \right) \left(e^{\frac{-\theta}{\mu^{2\beta}}} (1 + \lambda - \lambda e^{\frac{-\theta}{\mu^{2\beta}}}) \right)$$

$$- 2\theta^{\frac{1}{2\beta}} \left[\Gamma\left(1 - \frac{1}{2\beta}, \frac{\theta}{\mu^{2\beta}}\right) + \lambda\Gamma\left(1 - \frac{1}{2\beta}, \frac{\theta}{\mu^{2\beta}}\right) - 2^{\frac{1}{2\beta}} \lambda\Gamma\left(1 - \frac{1}{2\beta}, \frac{2\theta}{\mu^{2\beta}}\right) \right].$$

$$\delta_2 = \mu - 2£_{(1)}(m).$$

$$= \left(\theta^{\frac{1}{2\beta}} \Gamma\left(1 - \frac{1}{2\beta}\right) \left[1 + \lambda - 2^{\frac{1}{2\beta}} \lambda \right] \right) - 2\theta^{\frac{1}{2\beta}} \left[\Gamma\left(1 - \frac{1}{2\beta}, \frac{\theta}{m^{2\beta}}\right) + \lambda\Gamma\left(1 - \frac{1}{2\beta}, \frac{\theta}{m^{2\beta}}\right) - 2^{\frac{1}{2\beta}} \lambda\Gamma\left(1 - \frac{1}{2\beta}, \frac{2\theta}{m^{2\beta}}\right) \right].$$

Lorenz curve of the PTIR distribution is obtained as follows:

$$L_F(t) = \frac{£_{(1)}(t)}{E(T)} = \frac{\left[\Gamma\left(1 - \frac{1}{2\beta}, \frac{\theta}{t^{2\beta}}\right) + \lambda\Gamma\left(1 - \frac{1}{2\beta}, \frac{\theta}{t^{2\beta}}\right) - 2^{\frac{1}{2\beta}} \lambda\Gamma\left(1 - \frac{1}{2\beta}, \frac{2\theta}{t^{2\beta}}\right) \right]}{\Gamma\left(1 - \frac{1}{2\beta}\right) \left[1 + \lambda - 2^{\frac{1}{2\beta}} \lambda \right]}.$$

Bonferroni curve is obtained as follows:

$$B_F(t) = \frac{L_F(t)}{F(t)} = \frac{\left[\Gamma\left(1 - \frac{1}{2\beta}, \frac{\theta}{t^{2\beta}}\right) + \lambda\Gamma\left(1 - \frac{1}{2\beta}, \frac{\theta}{t^{2\beta}}\right) - 2^{\frac{1}{2\beta}} \lambda\Gamma\left(1 - \frac{1}{2\beta}, \frac{2\theta}{t^{2\beta}}\right) \right]}{\Gamma\left(1 - \frac{1}{2\beta}\right) \left[1 + \lambda - 2^{\frac{1}{2\beta}} \lambda \right] \left(e^{\frac{-\theta}{t^{2\beta}}} (1 + \lambda - \lambda e^{\frac{-\theta}{t^{2\beta}}}) \right)}.$$

3.4. Mean Residual Life Function

Mean residual life **(MRL)** function has been used in estimating time to failure for one or more existing and future failure modes. For an example nowadays MRL or remaining useful life is recognized as a key feature in maintenance strategies, while the

real prognostic systems are rare in industry, even in mining industry. The n^{th} moment of the residual life of X is given by

$$m_n(t) = E\left((X-t)^n \,|X>t\right) = \frac{1}{S(t)}\int_t^{\infty}(X-t)^n f(x)dx$$

Using the binomial expansion, for the term $(X\text{-}t)^n$, then $m_n(t)$ will be

$$m_n(t) = E\left((X-t)^n \,|X>t\right) = \frac{\sum_{j=0}^{n}\binom{n}{j}(-t)^{n-j}}{S(t)}\int_t^{\infty} x^j f(x)\,\mathrm{d}x. \tag{7}$$

The n^{th} moment of the residual life is obtained by substituting (2) in (7) and using $z = \theta x^{-2\beta}$, which leads to

$$m_n(t) = \frac{1}{S_{PTIR}(t)}\sum_{j=0}^{n}\binom{n}{j}(-t)^{n-j}(\theta)^{\frac{j}{2\beta}}\int_0^{\frac{\theta}{t^{2\beta}}} z^{\frac{-j}{2\beta}}\left[e^{-z}(1+\lambda-2\lambda e^{-z})\right]dz,$$

which is the lower incomplete gamma function, so the n^{th} moment of the PTIR distribution takes the following form:

$$m_n(t) = \frac{1}{S_{PTIR}(t)}\sum_{j=0}^{n}\binom{n}{j}(-t)^{n-j}(\theta)^{\frac{j}{2\beta}}\left[\gamma\left(1-\frac{j}{2\beta},\frac{\theta}{t^{2\beta}}\right)\lambda\gamma\left(1-\frac{j}{2\beta},\frac{\theta}{t^{2\beta}}\right)-2^{\frac{j}{2\beta}}\lambda\gamma\left(1-\frac{j}{2\beta},\frac{2\theta}{t^{2\beta}}\right)\right],$$

where $\gamma(.,x)$ is the lower incomplete moments.

3.5. Rényi Entropy

The Rényi entropy is used to quantify the diversity, uncertainty or randomness of a system; it has various fields of application such as ecology, statistics. Also, it is important in quantum information, where it can be used as a measure of entanglement.

$$I_R(X) = \frac{1}{1-\rho}\ln\left\{\int_R f^{\rho}(x)dx\right\},$$

where for some real values $\rho>0$ and $\rho\neq 1$, the entropy of the PTIR random variable X has the pdf (2) and is given by

$$I_R(X) = \frac{1}{1-\rho}\ln\left\{\int_0^{\infty}\frac{(2\theta\beta)^{\rho}}{x^{\rho(2\beta+1)}}e^{\frac{-\rho\theta}{x^{2\beta}}}\left(1+\lambda-2\lambda e^{\frac{-\theta}{x^{2\beta}}}\right)^{\rho}\right\}.$$

So,

$$I_R(X) = \frac{1}{1-\rho}\ln\left\{\int_0^{\infty}\frac{(2\theta\beta)^{\rho}}{x^{\rho(2\beta+1)}}e^{\frac{-\rho\theta}{x^{2\beta}}}(1+\lambda)^{\rho}\left(1-\frac{2\lambda}{1+\lambda}e^{\frac{-\theta}{x^{2\beta}}}\right)^{\rho}\right\}.$$

By using binomial expansion and after simplification, the Rényi entropy is

$$I_R(X)=\frac{1}{1-\rho}\ln\left\{\theta^{\rho}(2\beta)^{\rho-1}(1+\lambda)^{\rho}\sum_{j=0}^{\infty}\binom{\rho}{j}(-1)^j\left(\frac{2\lambda}{1+\lambda}\right)^j\Gamma\left(\frac{\frac{\rho(2\beta+1)-1}{2\beta}}{[\theta(\rho+j)]^{\frac{\rho(2\beta+1)-1}{2\beta}}}\right)\right\}.$$

4. Parameter Estimation

In this section, parameter estimators are obtained for the PTIR distribution based on **ML**, least squares **(LS)** and percentiles **(PR)** methods.

4.1. Maximum Likelihood Estimators

The ML estimator procedure is considered to estimate the population parameters of the PTIR distribution. The likelihood function is given by

$$L=(2\theta\beta)^n\prod_{i=1}^{n}x_i^{-(2\beta+1)}e^{\frac{-\theta}{x_i^{2\beta}}}(1+\lambda-2\lambda e^{\frac{-\theta}{x_i^{2\beta}}}).$$

The log likelihood function is given by

$$\ln L=n\ln 2\beta+n\ln\theta-(2\beta+1)\sum_{i=1}^{n}\ln x_i-\sum_{i=1}^{n}\frac{\theta}{x_i^{2\beta}}+\sum_{i=1}^{n}\ln(1+\lambda-2\lambda e^{\frac{-\theta}{x_i^{2\beta}}}). \tag{8}$$

Therefore, the ML estimators of θ,λ and β, which maximizes (8), satisfy the following normal equations.

$$\frac{\partial\ln L}{\partial\theta}=\frac{n}{\theta}-\sum_{i=1}^{n}x_i^{-2\beta}+2\lambda\sum_{i=1}^{n}\frac{e^{-\theta x_i^{-2\beta}}}{x_i^{2\beta}(1+\lambda-2\lambda e^{-\theta x_i^{-2\beta}})}, \tag{9}$$

$$\frac{\partial\ln L}{\partial\lambda}=\sum_{i=1}^{n}\frac{1-2e^{-\theta x_i^{-2\beta}}}{(1+\lambda-2\lambda e^{-\theta x_i^{-2\beta}})}, \tag{10}$$

and

$$\frac{\partial\ln L}{\partial\beta}=\frac{n}{\beta}-2\sum_{i=1}^{n}\ln x_i+2\theta\sum_{i=1}^{n}x_i^{-2\beta}\ln x_i+4\lambda\theta\sum_{i=1}^{n}\frac{x_i^{-2\beta}e^{\frac{-\theta}{x_i^{2\beta}}}\ln x_i}{(1+\lambda-2\lambda e^{\frac{-\theta}{x^{2\beta}}})}. \tag{11}$$

Then ML estimators of the parameters θ,λ and β denoted by $\hat{\theta},\hat{\lambda}$ and $\hat{\beta}$ are determined by solving numerically the non-linear Equations (9), (10) and (11) after setting them equal to zeros simultaneously.

4.2. Least Squares Estimators

Let $X_1,...,X_n$ be a random sample of size n from the PTIR distribution. Suppose that $X_{(1)},...,X_{(n)}$ denotes the corresponding ordered sample. Therefore, the LS estimators of θ,λ and β say, $\tilde{\theta},\tilde{\lambda}$ and $\tilde{\beta}$ respectively, can be obtained by minimizing the following function with respect to θ,λ and β.

$$LS = \sum_{i=1}^{n}\left[\left(e^{\frac{-\theta}{x_{(i)}^{2\beta}}} + \lambda e^{\frac{-\theta}{x_{(i)}^{2\beta}}} - \lambda e^{\frac{-2\theta}{x_{(i)}^{2\beta}}}\right) - \frac{i}{n+1}\right]^2. \tag{12}$$

Differentiating (12) with respect to θ, λ and β respectively, and equating with zeros, allows the LS estimators $\tilde{\theta}$, $\tilde{\lambda}$ and $\tilde{\beta}$ to be obtained.

4.3. Percentiles Estimators

Let $X_1,\ldots,X_n$ be a random sample of size n from the PTIR distribution. Suppose that $X_{(1)},\ldots,X_{(n)}$ denotes some estimates of $F\left(x_{(i)};\theta,\lambda,\beta\right)$ then the estimates of θ,λ and β can be obtained by minimizing the following equation:

$$PR = \sum_{i=1}^{n}\left[x_{(i)} - \left[\frac{-1}{\theta}\ln\left[\frac{(1+\lambda)-\sqrt{(1+\lambda)^2 - 4\lambda p_i}}{2\lambda}\right]\right]^{\frac{-1}{2\beta}}\right]^2, \tag{13}$$

with respect to θ,λ and β. In percentiles method, we estimate the unknown parameters θ,λ and β by equating the sample percentile points with the corresponding population percentile points, where $p_i = i/n+1$ is the estimates for $F\left(x_{(i)};\theta,\lambda,\beta\right)$. Then the PR estimators of θ,λ and β say, $\overline{\theta},\overline{\lambda}$ and $\overline{\beta}$ respectively, can be obtained by minimizing (13) with respect to θ,λ and β.

5. Simulation Studies

A numerical study is performed to evaluate and compare the performance of the estimates with respect to their absolute biases **(ABs)**, and mean square errors **(MSEs)** for different sample sizes and for different parameter values. The numerical procedures are described as follows:

Step (1): A random sample $X_1,\ldots,X_n$ of sizes n=10,20,30,100 is selected. These random samples are generated from the PTIR distribution by using the transformation (4).

Step (2): Four different set values of the parameters are selected as:

Set 1 = ($\theta=1.0,\ \lambda=0.5,\ \beta=0.5$), **Set 2** =($\theta=1.0,\ \lambda=0.5,\ \beta=1.5$), **Set 3** = ($\theta=1.0,\ \lambda=0.5,\ \beta=2$) and **Set 4** = ($\theta=0.5,\ \lambda=-0.7,\ \beta=1$).

Step (3): The ML, LS and PR estimates of θ,λ and β are computed for each set of parameters and for each sample size.

Step (4): Steps from 1 to 3 are repeated 5000 times for each sample size and for selected sets of parameters. Then, the ABs and MSEs of the ML, LS, PR estimates are computed.

Table 1. ABs and MSEs of the PTIR distribution for Set 1, Set 2, Set 3 and Set 4

n	Method	Properties	Set 1			Set 2		
			θ	λ	β	θ	λ	β
10	ML	MSE	0.0022	0.0055	0.0002	0.0017	0.0041	0.0013
		AB	0.0425	0.0667	0.0132	0.0365	0.0540	0.0319
	LS	MSE	0.0048	0.0122	0.0006	0.0005	0.0027	0.0007
		AB	0.0695	0.1106	0.0241	0.0226	0.0516	0.0259
	PR	MSE	0.0121	0.0397	0.0008	0.0060	0.0175	0.0038
		AB	0.1101	0.1993	0.0282	0.0774	0.1325	0.0619
20	ML	MSE	0.0003	0.0006	0.0000	0.0010	0.0025	0.0007
		AB	0.0043	0.0016	0.0013	0.0282	0.0442	0.0240
	LS	MSE	0.0015	0.0035	0.0002	0.0005	0.0016	0.0007
		AB	0.0386	0.0595	0.0128	0.0219	0.0396	0.0258
	PR	MSE	0.0013	0.0039	0.0001	0.0031	0.0088	0.0018
		AB	0.0367	0.0622	0.0095	0.0554	0.0939	0.0428
30	ML	MSE	0.0002	0.0004	0.0000	0.0005	0.0013	0.0005
		AB	0.0027	0.0010	0.0009	0.0177	0.0292	0.0181
	LS	MSE	0.0011	0.0022	0.0001	0.0003	0.0011	0.0002
		AB	0.0333	0.0470	0.0116	0.0160	0.0326	0.0155
	PR	MSE	0.0006	0.0014	0.0000	0.0013	0.0038	0.0008
		AB	0.0244	0.0375	0.0060	0.0360	0.0618	0.0275
100	ML	MSE	0.0001	0.0002	0.0000	0.0001	0.0002	0.0001
		AB	0.0020	0.0058	0.0005	0.0035	0.0058	0.0048
	LS	MSE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		AB	0.0068	0.0070	0.0029	0.0009	0.0050	0.0006
	PR	MSE	0.0001	0.0003	0.0000	0.0005	0.0014	0.0003
		AB	0.0098	0.0170	0.0040	0.0218	0.0373	0.0187

(cont.)

N	Method	Properties	Set 3			Set 4		
			θ	λ	β	θ	λ	β
10	ML	MSE	0.0006	0.0013	0.0002	0.0002	0.0018	0.0000
		AB	0.0065	0.0176	0.0091	0.0053	0.0226	0.0023
	LS	MSE	0.0156	0.0287	0.0044	0.0008	0.0071	0.0000
		AB	0.1248	0.1695	0.0663	0.0277	0.0844	0.0068
	PR	MSE	0.0129	0.0222	0.0025	0.0004	0.0051	0.0006
		AB	0.1135	0.1489	0.0500	0.0189	0.0712	0.0240
20	ML	MSE	0.0003	0.0006	0.0001	0.0001	0.0011	0.0000
		AB	0.0043	0.0017	0.0015	0.0040	0.0166	0.0007
	LS	MSE	0.0054	0.0112	0.0016	0.0007	0.0061	0.0000
		AB	0.0734	0.1060	0.0401	0.0266	0.0784	0.0060
	PR	MSE	0.0045	0.0093	0.0010	0.0001	0.0016	0.0001
		AB	0.0668	0.0967	0.0309	0.0115	0.0400	0.0108
30	ML	MSE	0.0002	0.0005	0.0001	0.0001	0.0007	0.0000
		AB	0.0057	0.0076	0.0005	0.0027	0.0105	0.0009
	LS	MSE	0.0040	0.0079	0.0012	0.0005	0.0041	0.0000
		AB	0.0632	0.0887	0.0340	0.0219	0.0637	0.0028
	PR	MSE	0.0030	0.0059	0.0005	0.0001	0.0008	0.0000
		AB	0.0550	0.0766	0.0232	0.0086	0.0291	0.0061
100	ML	MSE	0.0001	0.0002	0.0000	0.0000	0.0002	0.0000
		AB	0.0077	0.0101	0.0038	0.0009	0.0049	0.0015
	LS	MSE	0.0004	0.0007	0.0001	0.0000	0.0002	0.0000
		AB	0.0206	0.0270	0.0106	0.0045	0.0124	0.0007
	PR	MSE	0.0002	0.0003	0.0000	0.0000	0.0001	0.0000
		AB	0.0132	0.0170	0.0052	0.0046	0.0106	0.0002

The following conclusions can be observed on the properties of estimated parameters (see Table 1).

- The MSEs of the ML, LS and PR estimates decrease as the sample sizes increase for selected sets of parameters.
- The MSEs for the ML estimates of θ, λ and β take the smallest values compared to the MSEs of the LS and PR estimates in almost all of the cases.
- The ABs of the ML estimates are smaller than the ABs of the PR and LS estimates in almost all of the cases especially at small and moderate sample sizes.
- The ABs and MSEs of the ML, PR and LS estimates of β are smaller than the corresponding estimates of θ and λ in almost all of the cases.

6. Applications to Real Data

In this section, a real data analysis is provided in order to assess the goodness-of-fit of the PTIR model comparing with some known distributions such as IR, TIR, PIR, MIR, TMIR, KEIR.

In order to compare the models, criteria like maximized likelihood ($-2\hat{\ell}$), Akaike information criterion (**AIC**), consistent AIC (**CAIC**), Bayesian information criterion (**BIC**) and Hannan-Quinn information criterion (**HQIC**) are applied. The model with the minimum values of AIC, BIC, CAIC and HQIC is considered to be the best model to fit the proposed data.

The data set represents the survival times (in days) of 72 guinea pigs infected with virulent tubercle bacilli, observed and reported by Bjerkedal (1960). Plots of the estimated PTIR density and cumulative functions in addition to that of the compared models (TIR – PIR – IR – KEIR – MIR - TMIR) for the data set are displayed in Figure 4.

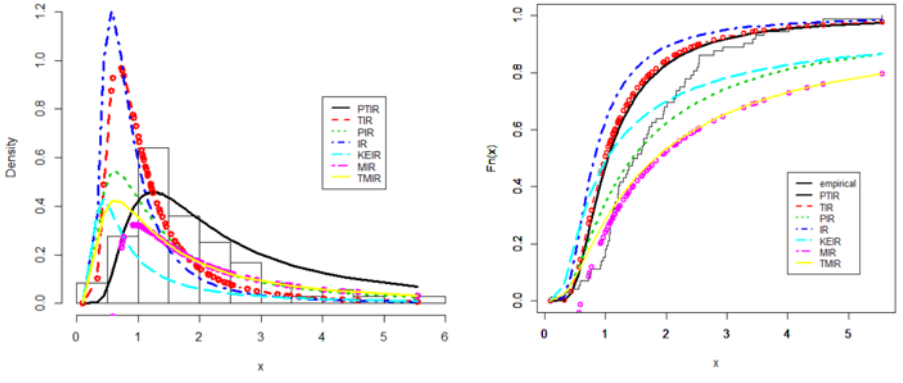


Figure 4. Estimated pdfs and cdfs of models for the data set

It can be observed from Figure 4 that the PTIR distribution is the most fitted distribution compared with the other models mentioned above concerning the Bjerkedal data.

The ML estimates and their standard errors (**SEs**) of the PTIR model compared with some known distributions such as IR,TIR, PIR,MIR, TMIR, KEIR are computed (see Table 2). Also, the corresponding measures of fit statistic using $-2\hat{\ell}$, AIC, BIC, CAIC, and HQIC, are provided in Table 3.

Table 2. ML estimates of the model parameters and the corresponding SEs

Model	θ	β	λ	α	a	b
PTIR	0.6056	0.6577	-0.9108			
	(0.0808)	(0.0463)	(0.0873)			
TIR	0.3525		-0.9416			
	(0.0434)		(0.0539)			
PIR	1.0691	0.5865				
	(0.1325)	(0.0421)				
IR	0.4629					
	(0.0546)					
KEIR	0.4001			0.3657	1.4444	0.4045
	(4.7575)			(4.3316)	(17.4921)	(0.0581)
MIR	0.0465			1.2500		
	(0.0187)			(0.1537)		
TMIR	0.0105		-0.9166	0.6575		
	(0.0278)		(0.0989)	(0.0960)		

Table 3. The statistics $-2\hat{\ell}$, AIC, CAIC, BIC and HQIC

Distribution	PTIR	IR	TIR	PIR	MIR	TMIR	KEIR
$-2\hat{\ell}$	225.273	327.518	280.538	236.332	237.825	236.819	280.492
AIC	231.273	329.518	284.538	240.332	241.825	243.825	288.492
CAIC	231.625	329.575	284.712	240.506	241.999	244.178	289.089
BIC	238.103	331.795	289.092	244.885	246.378	250.655	297.599
HQIC	233.992	330.424	286.351	242.145	243.638	246.544	292.118

Also, it can be confirmed from Table 3 that the PTIR distribution is the most fitted distribution among other models for the data set as the PTIR distribution has the minimum values of AIC, BIC, CAIC and HQIC.

7. Concluding Remarks

In this article, a new model, called a power transmuted inverse Rayleigh distribution is introduced. Some statistical properties of the proposed distribution are derived and discussed. The estimation of the model parameters is discussed through the maximum likelihood, least squares and percentiles methods. A simulation study is carried out to compare the performance of different estimates. The simulation study revealed that the ML performs better than the LS and PR estimates, in approximately most of the situations. An application to a real data set indicates that the new model is superior to the fits than the other suggested distributions.

Acknowledgements

The authors would like to thank the editor and the anonymous referees for their valuable and very constructive comments, which have greatly improved the contents of the paper.

REFERENCES

AHMAD, A., AHMAD, S.P., AHMED, A., (2014). Transmuted inverse Rayleigh distribution: A generalization of the inverse Rayleigh distribution. Mathematical Theory and Modeling, 4(7), pp. 90–98.

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

BOX, G. E. P., COX, D. R., (1964). An analysis of transformations. Journal of the Royal Statistical Society, Series B, 26, pp. 211–252.

BUTLER, R.J., MCDONALD, J. B., (1989). Using of incomplete moments to measure inequality. Journal of Econometrics, 42(1), pp. 109–119.

DEY, S., (2012). Bayesian estimation of the parameter and reliability function of an inverse Rayleigh distribution. Malaysian Journal of Mathematical Sciences, 6(1), pp. 113–124.

ELGARHY, M., ALRAJHI, S., (2018). The odd Fréchet inverse Rayleigh distribution: Statistical properties and applications. Journal of Nonlinear Sciences and Applications, 12, pp. 291–299.

FAN, G., (2015). Bayes estimation for inverse Rayleigh model under different loss functions. Research Journal of Applied Sciences, Engineering and Technology, 9(12), pp. 1115–1118.

FATIMA, K., AHMAD, S. P., (2017). Weighted inverse Rayleigh distribution. International Journal of Statistics and Systems, 12(1), pp. 119–137.

GHARRAPH, M.K., (1993). Comparison of estimators of location measures of an inverse Rayleigh distribution. The Egyptian Statistical Journal, 37, pp. 295–309.

HAQ, M. A., (2015). Transmuted exponentiated inverse Rayleigh distribution. Journal of Statistics Applications and Probability, 5(2), pp. 337–343.

HAQ, M. A., (2016). Kumaraswamy exponentiated inverse Rayleigh distribution. Mathematical Theory and Modeling, 6(3), pp. 93–104.

KHAN, M. S., (2014). Modified inverse Rayleigh distribution. International Journal of Computer Applications, 87(13), pp. 28–33.

KHAN, M. S., KING, R., (2015). Transmuted modified inverse Rayleigh distribution. Austrian Journal of Statistics, 44, pp. 17–29.

LEAO, J., SAULO, H., BOURGUIGNON, M., CINTRA, J., REGO, L., CORDEIRO, G., (2013). On some properties of the beta Inverse Rayleigh distribution. Chilean Journal of Statistics, 4(2), pp. 111–131.

MOHSIN, M., SHAHBAZ, M. Q., (2005). Comparison of negative moment estimator with maximum likelihood estimator of inverse Rayleigh distribution. Pakistan Journal of Statistics Operation Research, 1, pp. 45–48.

PANWAR, M. S., SUDHIR, B. A., BUNDEL, R., TOMER, S. K., (2015). Parameter estimation of Inverse Rayleigh distribution under competing risk model for masked data. Journal of Institute of Science and Technology, 20(2), pp. 122–127.

RASHEED, H. A., ISMAIL, S. Z., JABIR, A. G., (2015). A comparison of the classical estimators with the Bayes estimators of one parameter inverse Rayleigh distribution. International Journal of Advanced Research, 3(8), pp. 738–749.

RASHEED, H. A., AREF, R. K. H., (2016). Reliability estimation in inverse Rayleigh distribution using precautionary loss function. Mathematics and Statistics Journal, 2(3), pp. 9–15.

SINDHU, T. N., ASLAM, M., FEROZE, N., (2013). Bayes estimation of the parameters of the inverse Rayleigh distribution for left censored data. ProbStat Forum, 6, pp. 42–59.

SOLIMAN, A., AMIN, E. A., ABD-EI AZIZ, A. A., (2010). Estimation and prediction from inverse Rayleigh distribution based on lower record values. Applied Mathematical Sciences, 4, pp. 3057–3066.

TRAYER, V. N., (1964). Proceedings of the Academy of Science Belarus, USSR.

VODA, V. G. H., (1972). On the inverse Rayleigh distributed random variable. Rep. Statistics Application and Research, JUSE., 19(4), pp. 13–21.

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 3, pp. 109–128, DOI 10.21307/stattrans-2020-047
Received – 16.05.2018; accepted – 22.05.2020

Generalised Odd Frechet Family of Distributions: Properties and Applications

Shahdie Marganpoor[1], **Vahid Ranjbar**[2] **Morad Alizadeh**[3]
Kamel Abdollahnezhad[4]

ABSTRACT

A new distribution called Generalized Odd Fréchet (GOF) distribution is presented and its properties explored. Some structural properties of the proposed distribution, including the shapes of the hazard rate function, moments, conditional moments, moment generating function, skewness, and kurtosis are presented. Mean deviations, Lorenz and Bonferroni curves, Rényi entropy, and the distribution of order statistics are given. The maximum likelihood estimation technique is used to estimate the model parameters, and finally applications of the model to a real data set are presented to illustrate the usefulness of the proposed distribution.

Key words: Fréchet distribution, Wiebull distribution, structural properties, failure-time, maximum likelihood estimation.

1. Introduction

Recently, some attempts have been made to define new families of distributions to extend well-known models and at the same time provide great flexibility in modelling data in practice. Several techniques could be employed to form a larger family from an existing distribution by incorporating extra parameters. These generalized distributions give more flexibility by adding one "or more" parameters to the baseline model. For example, Gupta et al. (1998) proposed the exponentiated-G class, which consists of raising the cumulative distribution function (cdf) to a positive power parameter. Many other classes can be cited such as the Marshall-Olkin-G family by Marshall and Olkin (1997), beta generalized-G family by Eugene et al. (2002), the gamma-generated family by Zografos and Balakrishnan (2009), Kumaraswamy G family by Cordeiro and de Castro (2011), Generalized beta generated distributions by Alexanderet et al. (2015a), exponentiated generalized-G family by Cordeiro et al. (2013), a new method for generating families of continuous distributions by Alzaatreh et al. (2013), exponentiated T-X family of distributions by Alzaghal et al. (2013), the Lomax generator of distributions by Cordeiro et al. (2014), the WeibullG family of probability distributions by Bourguignon et al. (2014), beta Marshall-Olkin by Alizadeh et al. (2015a), Kumaraswamy odd log-logistic by Alizadeh et al. (2015b), beta odd log-logistic by Cordeiro et al. (2015), Kumaraswamy Marshall-Olkin by Alizadeh et

[1]Golestan University, Gorgan, 49138-15739, Iran. sh.marganpoor@gmail.com.
[2]Golestan University, Gorgan, 49138-15739, Iran, v.ranjbar@gu.ac.ir, vahidranjbar@gmail.com.
ORCID: https://orcid.org/0000-0003-3743-0330.
[3]Persian Gulf University, Bushehr, 751691-3798, Iran. moradalizadeh78@gmail.com.
ORCID: https://orcid.org/0000-0001-6638-2185.
[4]Golestan University, Gorgan, 49138-15739, Iran. k.abdollahnezhad@gu.ac.ir.

al. (2015c), transmuted exponentiated generalized-G family by Yousof et al. (2015), generalized transmuted-G by Nofal et al. (2015), generalized transmuted family by Alizadeh et al. (Alizadeh2015a), another generalized transmuted family by Merovci et al. (2015), Kumaraswamy transmuted-G by Afify et al. (2016a), transmuted geometric-G by Affify et al. (2016b), beta transmuted-H by Afify et al. (2016c), Burr X-G by Yousof et al. (2016), the odd Lindley-G family of distributions by Silva et al. (2016), exponentiated transmuted-G family by Merovci et al. (2016), odd-Burr generalized family by Alizadeh et al. (2016a) the complementary generalized transmuted Poisson family by Alizadeh et al. (2016b), logistic-X by Tahir et al. (2016a), a new Weibull-G by Tahir et al. (2016b), the two-sided power-G class by Korkmaz and Genc (2016), the type I half-logistic family by Cordeiro et al. (2016a), the Zografos-Balakrishnan odd log-logistic family of distributions by Cordeiro et al. (2016b), the generalized odd log-logistic family by Cordeiro et al.(2016c), the beta odd log-logistic generalized family of distributions by Cordeiro et al. (2016d), the Kumaraswamy odd log-logistic family of distributions by Alizadeh et al. (2016d) and a new generalized odd log-logistic family of distributions by Haghbin et al. (2017), the generalized odd log-logistic family of distributions: properties, regression models and applications by Cordeiro et al. (2017), the odd power Cauchy family of distributions by Alizadeh et al. (2018), a new family of the continuous distributions: the extended Weibull-family by Korkmaz (2018a), the Marshll-Olkin generalized G Piosson of distributions by Korkmaz et al. (2018b) and a new family of distributions with properties, regression models and applications by Yousof et al. (2018), among others.

The article is outlined as follows: in Section 2, we introduce the GOF distribution and provide plots of the density and hazard rate functions. Shapes, quantile function, moments, and moment generating function are also obtained. Moreover, mean deviation, order statistics, Lorenz and Bonferroni curves and finally asymptotic properties are presented in this section. Estimation by the method of maximum likelihood and an explicit expression for the observed information matrix are presented in Section 3. The simulation study is presented in Section 4. The applications to real data sets are considered in Section 5. Finally, Section 6 offers some concluding remarks.

2. Generalized Odd Frechet Family of distribution

The cdf of the Generalized Odd Frechet (GOF) Family of distributions is given by

$$F(x;a,b,\xi) = \exp\left\{-(G(x,\xi)^{-a}-1)^{b}\right\} \tag{1}$$

where $\xi = (\xi_1;\xi_2;...)$ is a parameter vector, and a and b are positive parameters. The corresponding probability density function (pdf) is

$$f(x;a,b,\xi) = a\,b\,g(x,\xi)G(x,\xi)^{-a-1}[G(x,\xi)^{-a}-1]^{b-1}\exp\left\{-(G(x,\xi)^{-a}-1)^{b}\right\} \tag{2}$$

For $a=1$ we obtain Odd Frechet family. Some of the possible shapes of the density function (2) of generalized odd Frechet Wiebull distribution (GOFW), for the selected parameter

values are illustrated in Figure 1. As seen in Figure 1, the density function can take various forms depending on the parameter values.

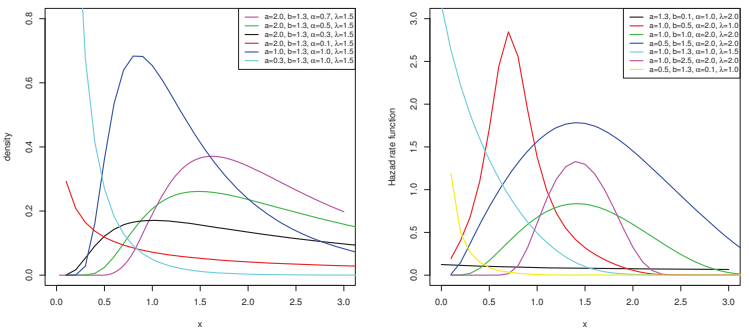


Figure 1: Different shapes of GOFW pdf (left) and Hazard function (right)

2.1. Survival and Hazard Rate Functions

A central role is played in the reliability theory by the quotient of the pdf and survival function. We obtain the survival function corresponding to (1) as

$$R(x) = 1 - \exp\left\{-(G(x,\xi)^{-a} - 1)^{b}\right\}$$

In reliability studies, The hazard rate [h(x)], reversed-hazard rate function [r(x)] and cumulative hazard rate function [H(X)] are important characteristics and fundamental to the design of safe systems in a wide variety of applications. Therefore, we discuss these properties of the GOF distribution. The $h(x), r(x)$ and $H(x)$ of X take the form

$$h(x) = \frac{a\,b\,g(x)\,G(x)^{-a-1}\,[G(x)^{-a} - 1]^{b-1}\,\mathrm{e}^{-[\mathrm{G(x)}^{-\mathrm{a}}-1]^{\mathrm{b}}}}{1 - \mathrm{e}^{-[\mathrm{G(x)}^{-\mathrm{a}}-1]^{\mathrm{b}}}}$$

$$r(x) = a\,b\,g(x,\xi)G(x,\xi)^{-a-1}[G(x,\xi)^{-a} - 1]^{b-1}$$

and

$$H(x) = -\log\left(1 - \exp\left\{-(G(x,\xi)^{-a} - 1)^{b}\right\}\right)$$

Plots of the hrf of the GOFW distribution for several parameter values are displayed in Figure 1.

2.2. Mixture representations for the pdf and cdf

Several structural properties of the extended distributions may be easily explored using mixture forms of Exp-G models. Therefore, we obtain mixture forms of exponentiated-G ("Exp-G") for F (x) and f (x). In this subsection, we provide alternative mixture representations for the pdf and cdf of X. Some useful expansions for (1) can be derived by using the concept of power series and generalized binomial expansion. We have

$$\begin{aligned} F(x) &= \exp\left(-(G(x)^{-a}-1)^{b}\right) = \exp\left(-\left(\frac{1-G(x)^{a}}{G(x)^{a}}\right)^{b}\right) \\ &= \sum_{i=0}^{\infty}\frac{(-1)^{i}}{i!}\left(\frac{1-G(x)^{a}}{G(x)^{a}}\right)^{bi} = \sum_{i,j=0}^{\infty}\frac{(-1)^{i+j}}{i!}\binom{bi}{j}G(x)^{aj}\,G(x)^{-abi} \qquad (3) \\ &= \sum_{j,k=0}^{\infty}\sum_{l=0}^{k} w_{j,k,l}\,G(x)^{aj+l} \qquad (4) \end{aligned}$$

where

$$w_{j,k,l} = \sum_{i=0}^{\infty}\frac{(-1)^{i+j+k+l}}{i!}\binom{bi}{j}\binom{-abi}{k}\binom{k}{l}$$

Furthermore, the corresponding GOF density function is obtained by differentiating (4)

$$f(x) = \sum_{j,k=0}^{\infty}\sum_{l=0}^{k} w_{j,k,l}\,(aj+l)\,g(x)\,G(x)^{aj+l-1} \qquad (5)$$

Using relation (3) we obtain another form of expansions for (1) as bellow, which is used in rest of the paper,

$$F(x) = \sum_{i,j=0}^{\infty}\frac{(-1)^{i+j}}{i!}\binom{bi}{j}G(x)^{a(j-bi)} = \sum_{k=0}^{\infty} e_k\,H_k(x) \qquad (6)$$

where $\bar{G}(x) = 1-G(x)$,

$$e_k = \sum_{i,j=0}^{\infty}\frac{(-1)^{i+j+k}}{i!}\binom{bi}{j}\binom{a(j-bi)}{k} \qquad (7)$$

and $H_\delta(x) = (1-G(x))^{\delta}$ is the survival function of the Exp-G distribution with power parameter δ. Then the corresponding GOF density function is obtained by differentiating (6)

$$f(x) = \sum_{k=0}^{\infty} e_k\,h_k(x) \qquad (8)$$

where $h_\delta(x) = \delta\,g(x)\overline{G}(x)^{\delta-1}$.

2.3. Moments and Moment Generating Function

Some of the most important features and characteristics of a distribution can be studied through moments (e.g. tendency, dispersion, skewness and kurtosis). Now we obtain ordinary moments and the moment generating function (mgf) of the GOF distribution. The rth ordinary moment of X is given by

$$\mu'_r = E(X^r) = \int x^r f(x)dx = \sum_{k=0}^{\infty} e_k E(Y_k^r) \tag{9}$$

where $E(Y_k^r) = \int x^r k\, g(x)\overline{G}(x)^{k-1}dx$; which can be computed numerically for most parent distributions. The skewness and kurtosis measures can be calculated from the ordinary moments using well-known relationships. One can also find the kth central moment of the GOF distribution through the following well-known equation

$$\mu_k = E(X-\mu)^k = \sum_{r=0}^{k} \binom{k}{r} \mu'_r(-\mu)^{k-r}. \tag{10}$$

Using (10), the variance, skewness and kurtosis measures can be obtained. Skewness measures the degree of the long tail and kurtosis is a measure of the degree of tail heaviness. The skewness can be computed as

$$S = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1}{(\mu'_2 - \mu'^2_1)^{3/2}}$$

and the kurtosis is based on octiles as

$$K = \frac{\mu_4}{\mu_2^2} = \frac{\mu'_4 - 4\mu'_1\mu'_3 + 6\mu'^2_1\mu'_2 - 3\mu'^4_1}{\mu'_2 - \mu'^2_1}.$$

When the distribution is symmetric $S = 0$, and when the distribution is right (or left) skewed $S > 0 (or\ S < 0)$. As K increases, the tail of the distribution becomes heavier. These measures are less sensitive to outliers and they exist even for distributions without moments.

The rth moment of generalized odd Frechet Weibull (GOFW) distribution using relation (8) is given by

$$\begin{aligned} \mu'_r &= \int_0^{\infty} x^r f(x)dx = \sum_{k=0}^{\infty} k e_k \int_0^{\infty} x^r \frac{\alpha}{\lambda}\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-(\frac{x}{\lambda})^{\alpha}} \left(e^{-(\frac{x}{\lambda})^{\alpha}}\right)^{k-1} dx \\ &= \sum_{k=0}^{\infty} k e_k \int_0^{\infty} x^r \frac{\alpha x^{\alpha-1}}{\lambda^{\alpha}} e^{-k(\frac{x}{\lambda})^{\alpha}} dx = \lambda^r A(\lambda,\alpha,r) \end{aligned} \tag{11}$$

where $\Gamma(a) = \int_0^{\infty} x^{a-1}e^{-x}dx$ is gamma function and

$$A(\lambda,\alpha,r) = \sum_{k=0}^{\infty} \left(\frac{e_k}{k^{r/\alpha}}\right)\Gamma\left(1 + \frac{k^{1/\alpha}r}{\lambda}\right).$$

Using power series, the moment generating function of GOFW is as bellow

$$M_X(t) = E(e^{tX}) = \sum_{n=0}^{\infty} \frac{t^n}{n!} E(X^n) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \lambda^n A(\lambda, \alpha, n)$$

It is to be highlighted that the equation (11) can be easily computed numerically using mathematical or statistical software. For this purpose, one can compute this equation for a large natural number, say N, instead of infinity in the sums. Therefore, several quantities of X such as moments, skewness and kurtosis can be computed numerically using (11). Plots for skewness and kurtosis are presented in Figure 2.

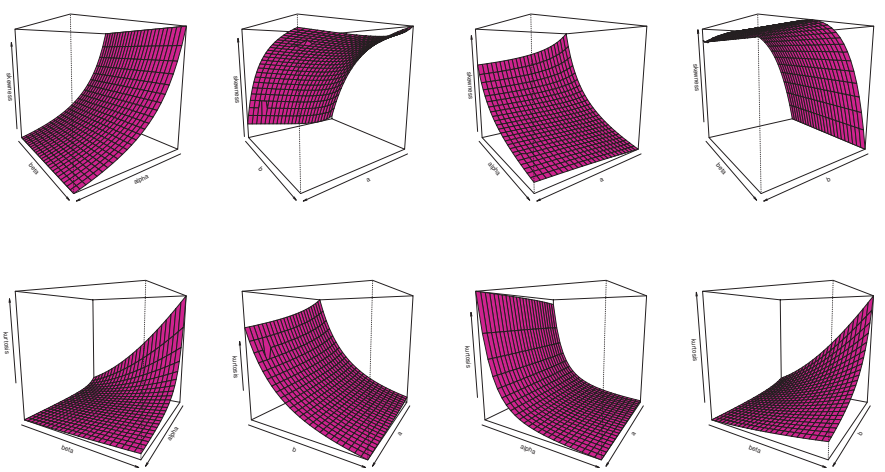


Figure 2: The skewness and kurtosis plots of GOF distribution for selected a, b, α, β.

2.4. Order statistics

Order statistics make their appearance in many areas of statistical theory and practice. Suppose $X_1, \ldots, X_n$ is a random sample from any GOF distribution. Let $X_{i:n}$ denote the ith order statistic. The pdf of $X_{i:n}$ can be expressed as

$$f_{i:n}(x) = K f(x) F^{i-1}(x) \{1 - F(x)\}^{n-i} = K \sum_{j=0}^{n-i} (-1)^j \binom{n-i}{j} f(x) F(x)^{j+i-1},$$

where $K = 1/B(i, n-i+1)$. We use the result of Gradshteyn and Ryzhik (2000) for a power series raised to a positive integer n (for $n \geq 1$)

$$\left(\sum_{i=0}^{\infty} a_i u^i \right)^n = \sum_{i=0}^{\infty} d_{n,i} u^i, \tag{12}$$

where the coefficients $d_{n,i}$ (for $i = 1,2,\ldots$) are determined from the recurrence equation (with $d_{n,0} = a_0^n$)

$$d_{n,i} = (i a_0)^{-1} \sum_{m=1}^{i} [m(n+1) - i] a_m d_{n,i-m}. \tag{13}$$

We can show that the density function of the ith order statistic of any GOF distribution can be expressed as

$$f_{i:n}(x) = \sum_{r,k=0}^{\infty} m_{r,k} H_{r+k+1}(x), \tag{14}$$

where $H_{r+k+1}(x)$ stands for the the survival function of the Exp-G distribution with power parameter $r+k+1$.

$$m_{r,k} = \frac{n!\,(r+1)\,(i-1)!\,e_{r+1}}{(r+k+1)} \sum_{j=0}^{n-i} \frac{(-1)^j f_{j+i-1,k}}{(n-i-j)!\,j!}.$$

Here, e_r is given by (7) and the quantities $f_{j+i-1,k}$ can be determined given that $f_{j+i-1,0} = e_0^{j+i-1}$ and recursively we have:

$$f_{j+i-1,k} = (k e_0)^{-1} \sum_{m=1}^{k} [m\,(j+i) - k]\, e_m f_{j+i-1,k-m}, k \geq 1.$$

Equation (14) is the main result of this section. Therefore, several mathematical quantities of these order statistics like ordinary and incomplete moments, factorial moments, mgf, mean deviations and others can be derived using this result.

2.5. Mean Deviations, Lorenz and Bonferroni Curves

Mean deviation about the mean and mean deviation about the median as well as Lorenz and Bonferroni curves for the GOF distribution are presented in this section. Bonferroni and Lorenz curves are a widely used tool for analysing and visualizing income inequality. Lorenz curve, L(p) can be regarded as the proportion of total income volume accumulated by those units with income lower than or equal to the volume y, and Bonferroni curve, B(p) is the scaled conditional mean curve, that is, ratio of group mean income of the population.

2.5.1 Mean deviations

The amount of scatter in a population may be measured to some extent by deviations from the mean and median. These are known as the mean deviation about the mean and the mean deviation about the median, defined by

$$\delta_1(X) = \int_0^{\infty} |x-\mu|\, f(x)\,\mathrm{d}x, \quad \textit{and} \qquad \delta_2(X) = \int_0^{\infty} |x-M|\, f(x)\,\mathrm{d}x.$$

respectively, where $\mu = E(X)$ and $M = \text{Median}(X) = Q(0.5)$ denotes the median and $Q(p)$ is the quantile function. The measures $\delta_1(X)$ and $\delta_2(X)$ can be calculated using the relationships

$$\delta_1(X) = 2\mu F(\mu) - 2\int_0^{\mu} x f(x)\,\mathrm{d}x, \quad \textit{and} \quad \delta_2(X) = \mu - 2\int_0^{M} x f(x)\,\mathrm{d}x$$

Finally for GOFW distribution we have

$$\begin{aligned}\delta_1(X) &= 2\mu F(\mu) - 2\sum_{k=0}^{\infty} k e_k \int_0^{\mu} x \frac{\alpha x^{\alpha-1}}{\lambda^{\alpha}} e^{-k(\frac{x}{\lambda})^{\alpha}} dx \\ &= 2\mu F(\mu) - 2\lambda B(\lambda,\alpha,\mu)\end{aligned}$$

where $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt$ is lower incomplete gamma function and

$$B(\lambda,\alpha,\mu) = \sum_{k=0}^{\infty} \frac{e_k}{k^{1/\alpha}} \gamma(2, \frac{\mu\lambda^{\alpha}}{k})$$

And

$$\delta_2(X) = \mu - 2\lambda B(\lambda,\alpha,M).$$

2.5.2 Bonferroni and Lorenz curves

The Bonferroni and Lorenz curves have applications in economics as well as other fields like reliability, medicine and insurance. Let $X \sim GOFW(a,b,\alpha,\lambda)$ and $F(x)$ be the cdf of X, then the Bonferroni curve of the GOFW distribution is given by

$$B(F(x)) = \frac{1}{\mu F(x)} \int_0^x t f(t)\mathrm{d}t,$$

where $\mu = E(X)$. Therefore, from (15), we have

$$B(F(x)) = \frac{1}{\mu F(x)} \times \lambda B(\lambda,\alpha,x).$$

The Lorenz curve of the GOFW distribution can be obtained using the relation

$$L(F(x)) = F(x)B(F(x)) = \frac{\lambda}{\mu} B(\lambda,\alpha,x).$$

2.6. Asymptotic Properties

One of the main usage of the idea of an asymptotic distribution is in providing approximations to the cumulative distribution functions of the statistical estimators.

The asymptotic of cdf, pdf and hrf of the GOF distribution as $x \to 0$ are, respectively, given by

$$\begin{aligned} F(x) &\sim \exp(-G(x)^{-ab}) \quad as \quad x \to 0, \\ f(x) &\sim a\,b\,g(x)\,G(x)^{-ab-1}\exp(-G(x)^{-ab}) \quad as \quad x \to 0, \\ h(x) &\sim a\,b\,g(x)\,G(x)^{-ab-1} \quad as \quad x \to 0. \end{aligned}$$

The asymptotic of cdf, pdf and hrf of the GOF distribution as $x \to \infty$ are, respectively, given by

$$\begin{aligned} 1-F(x) &\sim (a\bar{G}(x))^{b} \quad as \quad x \to \infty, \\ f(x) &\sim b\,a^{b}\,g(x)\bar{G}(x)^{b-1} \quad as \quad x \to \infty, \\ h(x) &\sim \frac{b\,g(x)}{\bar{G}(x)} \quad as \quad x \to \infty. \end{aligned}$$

These equations show the effect of parameters on the tails of the GOF distribution.

3. Estimation

Several approaches for parameter estimation have been proposed in the literature but the maximum likelihood method is the most commonly employed. Here, we consider estimation of the unknown parameters of the GOF distribution by the method of maximum likelihood. Let $x_1, x_2, ..., x_n$ be observed values from the GOF distribution with parameters a, b and ξ, where ξ is the parameter of based distribution function. The log-likelihood function for $(a; b; \xi)$ is given by

$$\begin{aligned} \ell_n &= n\log(a) + n\log(b) + \sum_{i=1}^{n}\log(g(x_i,\xi)) - (a+1)\sum_{i=1}^{n}\log(G(x_i,\xi)) \\ &\quad + (b-1)\sum_{i=1}^{n}\log(G(x_i,\xi)^{-a}-1) - \sum_{i=1}^{n}(G(x_i,\xi)^{-a}-1)^{b}. \end{aligned}$$

The derivatives of the log-likelihood function with respect to the parameters $(a; b; \xi)$ are given respectively, by

$$\begin{aligned} \frac{\partial \ell_n}{\partial a} &= \frac{n}{a} - \sum_{i=1}^{n}\log(G(x_i,\xi)) + (b-1)\sum_{i=1}^{n}\frac{-\log(G(x_i,\xi))G(x_i)^{-a}}{G(x_i,\xi)^{-a}-1} \\ &\quad + \sum_{i=1}^{n} b(G(x_i,\xi)^{-a}-1)^{b-1}G(x_i,\xi)^{-a}\log(G(x_i,\xi)) \end{aligned}$$

$$\frac{\partial \ell_n}{\partial b} = \frac{n}{b} + \sum_{i=1}^{n}\log(G(x_i,\xi)^{-a}-1) - \sum_{i=1}^{n}\log(-(G(x_i,\xi)^{-a}-1))(G(x_i,\xi)^{-a}-1)^{b}$$

and

$$\frac{\partial \ell_n}{\partial \xi} = \sum_{i=1}^{n} \frac{g'(x_i,\xi)}{g(x_i,\xi)} - (a+1)\sum_{i=1}^{n} \frac{G'(x_i,\xi)}{G(x_i,\xi)} - (b-1)\sum_{i=1}^{n} \frac{aG'(x_i,\xi)G(x_i,\xi)}{G(x_i,\xi)^{-a}-1} + \sum_{i=1}^{n} abG'(x_i,\xi)(G^{-a}-1)^{b-1}$$

where

$$g'(x_i,\xi) = \frac{\partial g(x_i,\xi)}{\partial \xi}, \quad G'(x_i,\xi) = \frac{\partial G(x_i,\xi)}{\partial \xi}$$

The maximum likelihood estimates (MLEs) of $(a;b;\xi)$, say $(\widehat{a};\widehat{b};\widehat{\xi})$, are the simultaneous solution of the equations $\frac{\partial \ell_n}{\partial a} = 0; \frac{\partial \ell_n}{\partial b} = 0; \frac{\partial \ell_n}{\partial \xi} = 0$.

For estimating the model parameters, numerical iterative techniques should be used to solve these equations. We can investigate the global maxima of the log-likelihood by setting different starting values for the parameters. The information matrix will be required for interval estimation. Let $\theta = (\alpha;\beta,\gamma,\lambda)^T$, then the asymptotic distribution of $\sqrt{n}(\theta - \hat{\theta})$ is $N_4(0, K(\theta)^{-1})$, under standard regularity conditions (see Lehmann and Casella, 1998, pp. 461-463), where $K(\theta)$ is the expected information matrix. The asymptotic behaviour remains valid if $K(\theta)$) is superseded by the observed information matrix multiplied by $1/n$, say $I(\theta)/n$, approximated by $\hat{\theta}$, i.e. $I(\hat{\theta})/n$. We have

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

where

$$I_{\alpha\alpha} = \frac{\partial^2 \ell_n}{\partial \alpha^2}; \quad I_{\alpha\beta} = I_{\alpha\beta} = \frac{\partial^2 \ell_n}{\partial \alpha \partial \beta}; \quad I_{\alpha\gamma} = I_{\alpha\gamma} = \frac{\partial^2 \ell_n}{\partial \alpha \partial \gamma}; \quad I_{\alpha\lambda} = I_{\alpha\gamma} = \frac{\partial^2 \ell_n}{\partial \alpha \partial \lambda}$$

$$I_{\beta\gamma} = I_{\gamma\beta} = \frac{\partial^2 \ell_n}{\partial \beta \partial \gamma}; \quad I_{\beta\lambda} = I_{\lambda\beta} = \frac{\partial^2 \ell_n}{\partial \beta \partial \lambda}; \quad I_{\gamma\lambda} = I_{\lambda\gamma} = \frac{\partial^2 \ell_n}{\partial \gamma \partial \lambda}.$$

4. Simulation study

In this section, we propose the inverse cdf method for generating random data from the GOF distribution. If $U \sim U(0,1)$ and if G has an inverse function, then

$$x = G^{-1}\left(\left[1 + (-ln(u))^{\frac{1}{b}}\right]^{\frac{-1}{a}}\right)$$

has cdf (1). Particularly,

$$x = \lambda \left[-\ln\left(\left([1 + (-\ln(u))^{\frac{1}{b}}]^{\frac{-1}{a}} \right) \right) \right]^{\frac{-1}{\alpha}}$$

is a random data with GOFW distribution.

Moreover, the performance of the maximum likelihood method is evaluated for estimating the GOFW parameters using a Monte Carlo simulation study. The mean square error (MSES) and the bias of the parameter estimates are calculated. We generate $N = 10,000$ samples of sizes $n = 50, 55, ..., 300$ from the GOFW distribution with $a = 2$, $b = 1.5$, $\alpha = 1.5, \lambda = 1$. Let $(\widehat{\alpha}, \widehat{\lambda}, \widehat{a}, \widehat{b})$ be the MLEs of the new model parameters and $(s_{\widehat{\alpha}}, s_{\widehat{\lambda}}, s_{\widehat{a}}, s_{\widehat{b}})$ be the standard errors of the MLEs. The estimated biases and MSEs are given by

$$\widehat{Bias}_{\varepsilon}(n) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\varepsilon}_i - \varepsilon)$$

and

$$\widehat{MSE}_{\varepsilon}(n) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\varepsilon}_i - \varepsilon)^2,$$

for $\varepsilon = \alpha, \lambda, a, b$. Figure 3 displays the numerical results for the above measures. We conclude below results from these plots:

- ✓ The estimated biases decrease when the sample size n increases,
- ✓ The estimated MSEs decay toward zero as n increases,

These results reveal the consistency property of the MLEs.

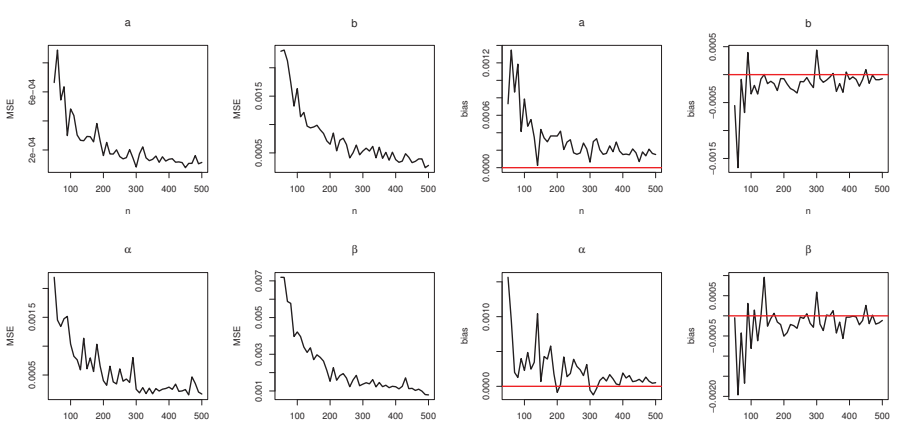


Figure 3: Estimated biases and MSEs for the selected parameter values.

5. Application

In this section, we illustrate the fitting performance of the GOFW distribution using two real data sets. For the purpose of comparison, we fitted the following models to show the fitting performance of GOFW distribution by means of real data set:
i) Weibull Distribution, $W(\alpha,\lambda)$.
ii) Exponentiated Weibull distribution, $EW(\alpha,\lambda,a)$, with distribution function given by

$$F_{ew}(x) = \left(1-e^{-(\frac{x}{\lambda})^{\alpha}}\right)^{a}.$$

iii) Kumaraswamy Weibull, $KwW(a,b,\alpha,\lambda)$

$$F_{kww}(x) = 1-[1-W(x,\alpha,\lambda)^{a}]^{b}.$$

iv) Beta Weibull, $BW(a,b,\alpha,\lambda)$, with distribution function given by

$$F_{bw}(x) = \int_0^{W(x,\alpha,\lambda)} t^{a-1}(1-t)^{b-1}dt.$$

v) Mc Weibull distribution $McW(a,b,\alpha\lambda,c)$, with distribution function given by

$$F_{mcw}(x) = \int_0^{(W(x,\alpha,\lambda))^{c}} t^{a-1}(1-t)^{b-1}dt.$$

vi) Generalized Odd Log-Logistic Weibull distribution $GOLLW(a,b,\alpha,\lambda)$, with distribution function given by

$$F_{gollw}(x) = \frac{W(x,\alpha,\lambda)^{ab}}{W(x,\alpha,\lambda)^{ab}+(1-W(x,\alpha,\lambda)^{a})^{b}}.$$

vii) Type I General Exponential Weibull distribution $TIGEW(a,b,\alpha,\lambda)$, with distribution function given by

$$F_{tigew}(x) = e^{b\left\{1-W(x,\alpha,\lambda)^{-a}\right\}}.$$

viii) Odd Frechet Weibull distribution $OFW(b,\alpha,\lambda)$, with distribution function given by

$$F(x;a,b,\xi) = \exp\left\{-(G(x,\xi)-1)^{b}\right\}$$

Estimates of the parameters of GOF distribution, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Cramer Von Mises and Anderson-Darling statistics (W^{*} and A^{*}) are presented for each data set. We have also considered the Kolmogorov-

Smirnov (K-S) statistic and its corresponding p-value and the minimum value of the minus log-likelihood function (-Log(L)) for the sake of comparison. Generally speaking, the smaller values of AIC, BIC, W^* and A^*, the better fit to a data set. All the computations were carried out using the software R.

In the rest of the paper, we model the lower discharge of at least seven consecutive days and return period (time) of ten years ($Q_{7,10}$) of the Cuiab*á* River, Cuiab*á*, Mato Grosso, Brazil. We consider the data set presented by Andrade et al. (2007). The calculation of the lower discharge for seven consecutive days and return period (time) of ten years ($Q_{7,10}$) is an important hydrological parameter with applications in the study planning and management of the use of water resources. This study aims to model the lower flood (discharge) of at least seven consecutive days and return period (time) of 10 years ($Q_{7,10}$) in Cuiab*á* River, part of the Brazilian Pantanal (Swamp), since the ecosystem is strongly influenced by the hydrological system. The calculations of $Q_{7,10}$ use a data series from 38 years (January 1962 to October 1999) relating to lower flows of n^o66260001 hydrological station, installed in the Cuiab*á* River in the city of Cuiab*á*, Mato Grosso, Brazil. The data, which have also been analysed by Cordeiro et al. (2012), are listed in Table 1.

Table 1: Data set.

43.86	44.97	46.27	51.29	61.19	61.20	67.80	69.00	71.84
77.31	85.39	86.59	86.66	88.16	96.03	102.00	108.29	113.00
115.14	116.71	126.86	127.00	127.14	127.29	128.00	134.14	136.14
140.43	146.43	146.43	148.00	148.43	150.86	151.29	151.43	156.14
163.00	186.43							

The ML estimates of the parameters and the goodness-of-fit test statistics for the real data set are presented in Table 3 and 4 respectively. As we can see, the smallest values of AIC, BIC, A^*, W^* and $-l$ statistics and the largest p-values belong to the GOFW distribution. Therefore, the GOFW distribution outperforms the other competitive considered distribution in the sense of this criteria.

Here, we also applied likelihood ratio (LR) tests. The LR tests can be used for comparing the GOFW distribution with its sub-models. For example, the test of $H_0 : \alpha = 1$ against $H_1 : \alpha \neq 1$ is equivalent to comparing the GOFW and OFW distributions with each other. For this test, the LR statistic can be calculated by the following relation:

$$LR = 2\left[l(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}, \widehat{\lambda}) - l(\widehat{\alpha^*}, 1, \widehat{\gamma^*}, \widehat{\lambda^*})\right],$$

where $\widehat{\alpha^*}, \widehat{\gamma^*}$ and $\widehat{\lambda^*}$ are the ML estimators of α, γ and λ, respectively, obtained under H_0. Under the regularity conditions and if H_0 is assumed to be true, the LR test statistic converges in distribution to a chi square with r degrees of freedom, where r equals the difference between the number of parameters estimated under H_0 and the number of parameters estimated in general, (for $H_0 : \beta = 1$, we have $r = 1$). Table 4 gives the LR statistics and the corresponding p-value. From Table 4, we observe that the computed p-value is too small so

Table 2: Parameter ML estimates (standard errors in the parentheses).

Model	$\widehat{a}$	$\widehat{b}$	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{c}$
$Weibull(\alpha,\lambda)$	–	–	3.3298	123.2008	–
			(0.4430)	(6.3067)	–
$EW(a,\alpha,\lambda)$	0.3522	–	6.8679	150.7316	–
	(0.2271)		(3.3323)	(14.4243)	–
$KwW(a,b,\alpha,\lambda)$	42.6066	1964.352	0.2112	7.2576	–
	(33.5498)	(8567.8074)	(0.1208)	(18.8249)	–
$BL(a,b,\alpha,\lambda)$	0.4034	0.3105	5.7524	114.9745	–
	(0.2071)	(0.3381)	(2.1554)	(33.4861)	–
$McW(a,b,\alpha,\lambda,c)$	0.1293	868.3850	0.5352	24.3734	112.9874
	(0.1093)	(4921.221)	(0.8856)	(122.134)	(401.6450)
$GOLLW(a,b,\alpha,\lambda)$	0.1734	4.7498	5.2297	94.0411	–
	(0.0234)	(0.0093)	(0.0039)	(0.0039)	–
$TIGEW(a,b,\alpha,\lambda)$	1.9133	0.0787	9.8806	164.239	–
	(2.2559)	(0.0555)	(5.6555)	(15.3034)	–
$OFW(b,\alpha,\lambda)$	–	3.3892	0.8968	49.1821	–
	–	(6.624)	(0.8048)	(54.0428)	–
$GOFW(a,b,\alpha,\lambda)$	2.2737	0.1542	5.0860	92.4172	–
	(0.5557)	(0.0274)	(0.0034)	(0.0034)	–

Table 3: Goodness-of-fit test statistics.

Model	W^*	A^*	$p-value$	*AIC*	*BIC*	$-l$
$Weibull(\alpha,\lambda)$	0.1019	0.6238	0.4312	386.6742	389.9494	191.3371
$EW(a,\alpha,\lambda)$	0.0585	0.4091	0.8515	386.8977	391.8104	190.4488
$KwW(a,b,\alpha,\lambda)$	0.1210	0.7323	0.3251	391.7345	398.2848	191.8672
$BL(a,b,\alpha,\lambda)$	0.0540	0.3879	0.8466	388.6756	395.2260	190.3378
$McW(a,b,\alpha,\lambda,c)$	0.0616	0.4093	0.5995	389.6777	397.8656	189.8388
$GOLLW(a,b,\alpha,\lambda)$	0.0358	0.2887	0.7564	385.1893	391.7396	188.5946
$TIGEW(a,b,\alpha,\lambda)$	0.0615	0.4140	0.6144	387.4896	394.0399	189.7448
$OFW(b,\alpha,\lambda)$	0.2655	1.6203	0.1651	400.1903	405.1031	197.0951
$GOFW(a,b,\alpha,\lambda)$	0.0285	0.2391	0.9775	382.8198	389.3701	187.4099

we reject the null hypotheses and conclude that the GOFW fits the first data better than the considered sub-model according to the LR criterion.

Table 4: The LR test results.

	Hypotheses	LR	p-value
GOFW versus OFW	$H_0 : a = 1$	18.8816	0.00001

In addition, PP plot of the GOFW distribution are plotted in Figure 4. We also plotted the fitted pdfs and cdfs of the considered models for the sake of visual comparison, in Figure 5. Figure 4 and 5 suggest that the GOFW fits the skewed data very well.

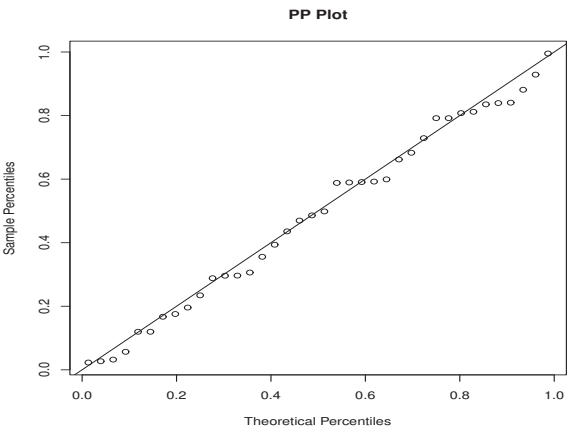


Figure 4: The PP plot.

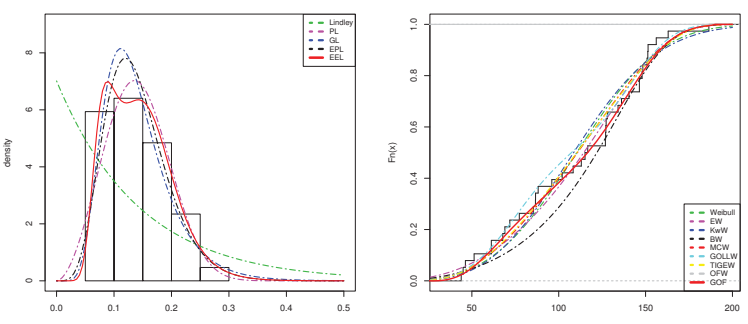


Figure 5: Fitted densities of distributions.

6. Conclusion

In this paper, we present a new class of distributions called the Generalized Odd Frechet (GOF) family of distributions. The statistical properties of the GOF distribution including the hazard and reverse hazard functions, quantile function, moments, incomplete moments, generating functions, mean deviations, Bonferroni and Lorenz curves, order statistics and maximum likelihood estimation for the model parameters are given. Simulation studies were conducted to examine the performance of the new GOF distribution. We also present applications of this new model to a real life data set in order to illustrate the usefulness of the distribution.

REFERENCES

AFIFY A.Z., ALIZADEH, M., YOUSOF, H. M., ARYAL, G. and AHMAD, M. (2016a). The transmuted geometric-G family of distributions: theory and applications. *Pak. J. Statist.*, 32(2), pp. 139–160.

AFIFY A.Z., CORDEIRO, G. M., YOUSOF, H. M., ALZAATREH, A. and NOFAL, Z. M. (2016b). The Kumaraswamy transmuted-G family of distributions: properties and applications. *J. Data Sci.*, 14(2), pp. 245–270.

AFIFY, A.Z., YOUSOF, H.M. and NADARAJAH, S. (2016c). The beta transmuted-H family of distributions: properties and applications. *Statistics and its Inference*, 10(3), pp. 505–520.

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

ALIZADEH, M., CORDEIRO, G. M., DE BRITO, E. and DEMÉTRIO C.G.B. (2015a). The beta Marshall- Olkin family of distributions. *Journal of Statistical Distributions and Applications*, 23(3), pp. 546–557.

ALIZADEH, M., CORDEIRO, G.M., MANSOOR, M., ZUBAIR, M. and HAMEDANI, G.G. (2015b). The Kumaraswamy Marshal-Olkin family of distributions. *Journal of the Egyptian Mathematical Society,* 23, pp. 546–557.

ALIZADEH, M., MEROVCI, F. and HAMEDANI, G.G. (2015c). Generalized transmuted family of distributions: properties and applications. *Hacettepa Journal of Mathematics and Statistics*, 46(4), pp. 645–667.

ALIZADEH M., EMADI M., DOOSTPARAST M., CORDEIRO G.M., ORTEGA E.M.M. and PESCIM R.R., (2016) A new family of distributions: the Kumaraswamy odd log-logistic, properties and applications. *Hacettepe Journal of Mathematics and Statistics*, 44(6), pp. 1491–1512.

ALIZADEH, M., CORDEIRO, G.M., NASCIMENTO, A.D.C. LIMA M.D.S. AND ORTEGA, E.M.M. (2016a). Odd-Burr generalized family of distributions with some applications. *Journal of Statistical Computation and Simulation*, 83, pp. 326–339.

ALIZADEH, M., YOUSOF, H.M., AFIFY A.Z., CORDEIRO, G.M., and MANSOOR, M. (2016b). The complementary generalized transmuted Poisson-G family. *Austrian Journal of Statistics*, 47(4), pp. 60–80.

ALIZADEH, M., ALTUN, E., CORDEIRO, G. M., and RASEKHI, M. (2018). The odd power cauchy family of distributions: properties, regression models and applications. *Journal of Statistical Computation and Simulation*, 88(4), pp. 785–807.

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

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

ANDRADE NLR, MOURA RMP, SILVEIRA A (2007) Determinação da $Q_{7,10}$ para o Rio Cuiab*á*, Mato Grosso, Brasil e comparação com a vazão regularizada após a implantação do reservatório de aproveitamento múltiplo de manso. 24^o Congresso Brasileiro de Engenharia Sanitária e Ambiental. *Belo Horizonte, Minas Gerais Brasil*.

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

CORDEIRO, G. M., DE CASTRO, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, 81, pp. 883–898.

CORDEIRO, GAUSS M., SARALEES NADARAJAH, and EDWIN MM ORTEGA, (2012). The Kumaraswamy Gumbel distribution. *Statistical Methods and Applications*, 21.2, pp. 139–168.

CORDEIRO, G. M., GOMES, A. E., DA-SILVA, C. Q. and ORTEGA, E. M., (2013). The beta exponentiated Weibull distribution. *Journal of Statistical Computation and Simulation*, 83, pp. 114–138.

CORDEIRO, G. M., HASHIMOTO, E. M. and ORTEGA, E. M., (2014). McDonald Weibull model. *Statistics: A Journal of Theoretical and Applied Statistics*, 48, pp. 256–278.

CORDEIRO, G. M.,ALIZADEH, M., TAHIR, M. H., MANSOOR, M., BOURGUIGNON, M., & HAMEDANI, G. G., (2015). The beta odd log-logistic generalized family of distributions, *Hacettepe Journal of Mathematics and Statistics*, 45(73), pp. 126–139.

CORDEIRO, G. M., ALIZADEH, M. and DINIZ MARINHO, P. R., (2016a). The type I half-logistic family of distributions. *Journal of Statistical Computation and Simulation*, 86, pp. 707–728.

CORDEIRO, G. M., ALIZADEH, M., ORTEGA, E. M. and SERRANO, L. H., V.(2016b). The Zografos- Balakrishnan odd log-logistic family of distributions: Properties and Applications. *Hacet. J. Math. Stat.*, 7(1), pp. 211–234.

CORDEIRO, G. M., ALIZADEH, M., OZEL, G., HOSSEINI, B., ORTEGA, E. M. M. and ALTUN, E., (2016c). The generalized odd log-logistic family of distributions: properties, regression models and applications, *Journal of Statistical Computation and Simulation*, 87, pp. 908–932.

CORDEIRO, G. M., ALIZADEH, M., TAHIR, M. H., MANSOOR, M., BOURGUIGNON, M. and HAMEDANI G. G., (2016d). The beta odd log-logistic generalized family of distributions, *Hacettepe Journal of Mathematics and Statistics*, 45(6), pp. 1175–1202.

CORDEIRO, G. M., ALIZADEH, M., OZEL, G., HOSSEINI, B., ORTEGA, E. M. M. and ALTUN, E., (2017). The generalized odd log-logistic family of distributions: properties, regression models and applications. *Journal of Statistical Computation and Simulation*, 87(5), pp. 908–932.

EMLET, R. B., MCEDWARD, L. R. and STRATHMANN, R. R., (1987) Echinoderm larval ecology viewed from the egg, in: M. Jangoux and J.M. Lawrence (Eds.) *Echinoderm Studies*,2, pp. 55–136.

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

GRADSHTEYN, I. S., RYZHIK, I. M., (2000). Table of integrals, series, and products. *Academic Press*, San Diego.

GUPTA, R. C., GUPTA, P. L. and GUPTA, R. D., (1998). Modeling failure time data by Lehmann alternatives. *Commun. Stat. Theory Methods*, 27, pp. 887–904.

HAGHBIN H., OZEL G., ALIZADEH, M. and HAMEDANI, G. G., (2017) A new generalized odd log-logistic family of distributions. *Communications in Statistics - Theory and Methods*, 46(20), pp. 9897–9920.

KORKMAZ, M. C., GENC, A. I., (2016). A new generalized two-sided class of distributions with an emphasis on two-sided generalized normal distribution. *Communications in Statistics - Simulation and Computation*, 46, pp. 1441–1460.

KORKMAZ, M. C., (2018). A new family of the continuous distributions: the extended Weibull-G family. *Communications Faculty of Sciences University of Ankara Series A1 Mathematics and Statistics*, 68(1), pp. 248–270.

KORKMAZ, M. C. , YOUSOF H. M., HAMEDANI, G. G. and ALI, M. M., (2018). The Marshall-Olkin Generalized G Poisson Family of Distributions, *Pakistan Journal of Statistics*, 34(3), pp. 251–267.

LEHMANN, E. L., CASELLA, G., (1998) Theory of Point Estimation, *Springer*.

MARSHALL, A. W., OLKIN, I., (1997). A new methods for adding a parameter to a family of distributions with application to the Exponential and Weibull families. *Biometrika*, 84, pp. 641–652.

MEROVCI, F., ALIZADEH, M. and HAMEDANI, G. G., (2016). Another generalized transmuted family of distributions: properties and applications, *Austrian Journal of Statistics*, 45, pp. 71–93.

MEROVCI, F., ALIZADEH, M., YOUSOF, H. M. and HAMEDANI, G. G., (2016). The exponentiated transmuted-G family of distributions: theory and applications. *Commun. Stat. Theory Meth- ods*, 46(21), pp. 10800–10822.

NOFAL, Z. M., AFIFY, A. Z., YOUSOF, H. M. and CORDEIRO, G. M., (2017). The generalized transmuted-G family of distributions. *Commun. Stat. Theory Methods*,46(8), pp. 4119–4136.

SILVA, F. G., PERCONTINI, A., DE BRITO, E., RAMOS, M. W., VENANCIO, R. and CORDEIRO, G. M., (2016). The odd Lindley-G family of distributions, *Austrian Journal of Statistics*, VV, 1–xx.

TAHIR, M. H., CORDEIRO, G. M., ALZAATREH, A., MANSOOR, M. and ZUBAIR, M., (2016a). The logistic- X family of distributions and its applications. *Commun. Stat. Theory Methods*,45(24), pp. 7326–7349.

TAHIR, M. H., ZUBAIR, M., MANSOOR, M., CORDEIRO, G. M., ALIZADEH, M. and HAMEDANI, G. G., (2016b). A new Weibull-G family of distributions. *Hacet. J. Math. Stat.*,45 (2), pp. 629–647.

YOUSOF, H. M., AFIFY, A. Z., ALIZADEH, M., BUTT, N. S., HAMEDANI, G. G. and ALI, M. M., (2015). The transmuted exponentiated generalized-G family of distributions, *Pak. J. Stat. Oper. Res.*, 11, pp. 441–464.

YOUSOF, H. M., AFIFY, A. Z., HAMEDANI, G. G. and ARYAL, G., (2016). the Burr X generator of distributions for lifetime data. *Journal of Statistical Theory and Applications*,16(3), pp. 288–305.

YOUSOF, H. M., ALTUN, E., RAMIRES, T. G., ALIZADEH, M., and RASEKHI, M., (2018). A new family of distributions with properties, regression models and applications. *Journal of Statistics and Management Systems*, 21(1), pp. 163–188.

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, September 2020
Vol. 21, No. 3 pp. 129–148, DOI 10.21307/stattrans-2020-048
Received – 03.09.2018; accepted – 31.05.2020

Ukraine's State Regulation of the Economic Development of Territories in the Context of Budgetary Decentralisation

Oleksandr H. Osaulenko[1], Taisiia Bondaruk[2], Liudmyla Momotiuk[3]

ABSTRACT

The paper presents the theoretical and methodological foundations of Ukraine's state legislation regulating the economic development of territories, in the context of budget decentralization. The study also describes the transformation of the public administration system necessitated by the above-mentioned phenomenon. The authors discuss the basic methods by which the state can regulate the activity of local self-government bodies: the legislative regulation, where the intervention of public authorities is minimized, and the administrative regulation, which provides rules and instructions which determine the relations between central and local authorities. The authors conduct and describe a methodologically consistent, systematic analysis of state regulations which support the local self-governments' activity. The paper also discusses the recent economic changes in Ukraine which demonstrate that the reform of the local self-government system and the decentralization of authority entail both prospects and problems for the country's development. As might be expected, the authors focus particularly on those problems that have not been solved yet. Additionally, statistical estimations of the phenomena relating to the process of producing state legislation regulating the economic development of territories in the context of budgetary decentralization have been provided. The authors conclude that a successful territorial development strategy requires a joint transformation of the way of the society's thinking and the modernisation of both the Ukrainian business and the state.

Key words: state regulation, budgetary decentralization, local self-government, local budget, economic development of territories.

[1] Doctor of Public Administration, Professor, Corresponding Member of National Academy of Sciences of Ukraine, Rector of National Academy of Statistics, Accounting and Audit, Kyiv, Ukraine. E-mail: o.osaulenko@nasoa.edu.ua. ORCID: https://orcid.org/0000-0002-7100-7176.

[2] Doctor of Economic Sciences, Professor, Head of the Department of Finance, Banking and Insurance of National Academy of Statistics, Accounting and Audit, Kyiv, Ukraine. E-mail: bondaruk23@ukr.net. ORCID: https://orcid.org/0000-0001-9410-6428.

[3] Doctor of Economic Sciences, Professor, Vice-Rector of National Academy of Statistics, Accounting and Audit, Kyiv, Ukraine. E-mail: momotuyk_le@ukr.net. ORCID: https://orcid.org/0000-0002-0445-5948.

1. Introduction

State regulation plays a significant role in the efficient functioning of the economy and is an effective tool for the economic development of the country and its territories. The problem of transformation of the state role, its goals in regulating the development of local self-government under conditions of budgetary decentralization requires special consideration. The urgency of studying the problems of state regulation of economic development of territories in terms of budget decentralization for Ukraine is conditioned by the need to develop an effective system of macroeconomic regulation of socio-economic processes in the context of decentralization. The process of socio-economic development of Ukraine on a democratic and legal basis is impossible without strengthening the role of local self-government. Budget decentralization is one of the main drivers of the much needed reform of self-government today.

There are many problems concerning the management of territories which are not only solved by financial and budgetary methods. This is the first and foremost problem of economic development of territories. Lack of investment can lead to systematic degradation and "extinction" of particular settlements. However, it is not possible to focus on the ongoing support of apparently unpromising territories.

Having gained independence from the state in terms of economic and financial activity as well as the right to its own regional policy, the local government authorities faced the problems of forming the local budget, distribution of state property at the regional and local levels, implementation of administrative reform, etc.

With the same analysis of the territorial entities within the regions there is even more disharmony in the issues of conformity with their economic development. Practice shows that quite often miscalculations of local management in the financial policy are explained by low skilled risk management, poor management training and so on. At the same time, there are many bureaucratic obstacles in management regulation procedures which hinder the use of local reserves.

The purpose of the study is to deepen the theoretical and methodological foundations of state regulation of economic development of territories in conditions of budgetary decentralization.

2. Transformation of public administration in conditions of decentralization

One of the main problems facing local authorities of any country is the problem of relations with the state government bodies, especially central and regional ones. Taking into account this fact, a key problem is about the autonomy of local self-government. Among many beliefs regarding the solution of the problems of public administration,

the issue of decentralization, as one of the means of improving the efficiency of the functioning of public power, has been in the field of view of scholars and practitioners for a long period of time.

In particular, this is due to the successful implementation of the principles of decentralization in the practice of most European Union countries. Thus, changes in the distribution of competences during 2000-2016 show that there has been a considerable reform in the allocation of competences between levels of government in the field of governance and spatial planning in the EU since 2000 (COMPASS (2016-2018). In particular, there are dominant trends that show decentralization and centralization tendencies in different parts of Europe. In many countries there were processes of management decentralization and planning of competences from national and sub-national levels to local level and strengthening local authorities' autonomy (COMPASS (2016-2018)), (Lidström, 2007). However, an increase in the processes of planning at the sub-national level (regionalization) has been observed. The third group of countries shows strengthening of national or sub-national government authority.

Local government autonomy is always relative since it is characterized by the presence of two types of restrictions (Grybanova G., 1998). The first type is the economic and social restrictions that are from different sources. Firstly, the conditions of local economy functioning limit the tax base. In the absence of subsidies from the central government, it is much more difficult for the "poor" areas than it is for the "rich" one to finance an adequate level of public services. In order to avoid a financial crisis local authority should take care of the productive use of land, capital and labour, with a well-developed strategy for developing a particular district. They also have to do it transparently so that the residents could see the feasibility of the ratio of costs (in the form of taxes paid to them) and the benefits received (through the use of services provided locally) are not worse than in neighbouring areas. Secondly, local-dominant interests can put pressure on political decision-making. First and foremost it deals with the interests of business and relevant elites (coalitions). The opinions are also expressed that the primary point in organizing state regulation is the issue of distribution of expenditures between the levels of government and consequently the budgetary system (Brosio G., 1985). As a result, the consequences of misallocation of expenditures lead to inefficient allocation of resources (Musgrave R.A., 1985).

The second type of restriction which is crucial for a political system and a society, peculiar to a particular country is imposed on local autonomy by senior levels of government. At the same time the following factors influence the autonomy of local self-government bodies: the sphere of competence, basic functions of local self-government bodies, forms and methods of their implementation; forms and methods of control over the activity of local self-government bodies by public authorities.

Management and spatial planning competencies are generally shared at different levels in most countries (Hooghe and Marks, 2003). The study of the evolution of the concepts of territorial management and spatial planning in terms of research methods and comparative planning trends has been addressed by Nadin V., Stead D. (Nadin V., Stead D., 2013), and Reimer M. (Reimer M., Getimis P., Blotevogel H., (eds.) 2014).

The studies of Central European countries that have successfully undergone economic and institutional transformation in the area of territorial governance and finance (for example, regarding the foreign capital involvement policy and its impact on Poland's economic development) have been successfully conducted by such authors as W. Dziemianowicz, B. Jałowiecki (Wojciech Dziemianowicz, Bohdan Jałowiecki, 2004).

3. Methodology of state regulation in the sphere of activity of local self-government authorities

When characterizing the methodological basis of cardinal decision-making in the system of state regulation it is important to evaluate the effectiveness of implementation of the regulatory policy of the state in the sphere of ensuring the activity of the local self-government authorities. In order to develop the effective forms and methods of such a policy of the state in a market economy it is necessary to determine its effectiveness without fail. Consequently, it is necessary to develop a comprehensive system for monitoring and assessing the impact of state regulation on the activities of local governments.

In terms of complexity and consistency, the state regulation is determined by the mode of influence, and this is a radical or liberal intervention; or non-interference by ignoring the adverse market situation due to the lack of effective tools of influence. The nature of state regulation is also manifested in the need for solidarity or individual balancing of the complex interests of local governments: economic, social and financial in particular.

There are two main methods with the help of which the state can regulate the activity of self-government authorities.

The first is legislative regulation, where the intervention of public authorities is minimized; it is a kind of "remote" control. Following the adoption of the relevant laws local governments may act at their discretion as long as they remain within the law. The main instrument of legislative regulation is the constitution. The respective constitutional position of local self-government bodies in a particular country is determined primarily by the legislative consolidation of the right to local self-government within the constitutional system. The legal basis of local self-government is not only constitutional provisions but also the rules of current legislation. Typically,

local self-government issues are regulated in detail in specific local self-government laws as well as in some sectoral legislative acts. However, in some countries there is a single law on local self-government and in others – laws on certain types of local self-government. As a rule, the federal states do not have a special federal law on local self-government granting the subjects of the federation (lands, cantons) the right to fully regulate the issues of local self-government (Grybanova G., 1998). Most often political discussions between the state and local self-government unfold precisely around issues of current legislation.

Thus, the factors of state regulation policy are characterized by a set of objective and subjective, structured by directions, means of predominantly regulatory content with the help of which this policy is formed, implemented and evaluated. Such factors include macroeconomic, structural and dynamic, administrative and organizational, pricing, financial, credit, technical and technological, infrastructural and transport, foreign economic, social and demographic, ecological and recreational, and historical and cultural ones (Kvasha G. 2013). Conditions for the implementation of state regulation policy are shaped by the influence of external and internal political, economic and social environments.

Another method is administrative regulation, which provides such an order of relations between the central and local authorities where by creating rules and instructions the state in the person of the central authorities gives a detailed account to local self-government bodies on one or another course of action. When implementing this kind of regulation the legitimacy of the actions of local self-government bodies is determined by the individual decisions of state officials (Grybanova G., 1998). In addition, administrative regulation often requires the prior approval of certain actions, which is legally stipulated. At the same time, in the legislative regulation only a judicial evaluation of the action is taken into account. Administrative control is by no means "remote".

It is important to evaluate the consequences of state regulation of socio-economic development of territories in terms of achieving results, although the effectiveness of regulatory policy will undoubtedly depend on the decision-making procedure that is predetermined by the political process and on the tasks that ensure the implementation of such decisions.

When analysing the regulatory framework on the state regulation and relevant practical measures, a special attention is drawn to the fact that at the present stage of implementation of state regulation policy in Ukraine the main emphasis has been shifted from the stage of development of regulatory acts to the stage of gathering data on the effects of the adopted regulatory acts, monitoring their effectiveness as well as the efficiency of making decisions about changing or repealing them.

Preliminary evaluation of the results of state regulation makes it possible to determine more precisely the amount of financing of socio-economic development of territories, because it is at this stage that the idea of priority goals of state regulation, the tasks needed to be fulfilled to achieve the goals and the volume of necessary resources (material, human, time, information, etc.) are formed.

The methodological orientation of the systematic analysis of state regulation in support of the activity of local self-government bodies in the general form will be presented by the chain of operations shown in Figure 1.

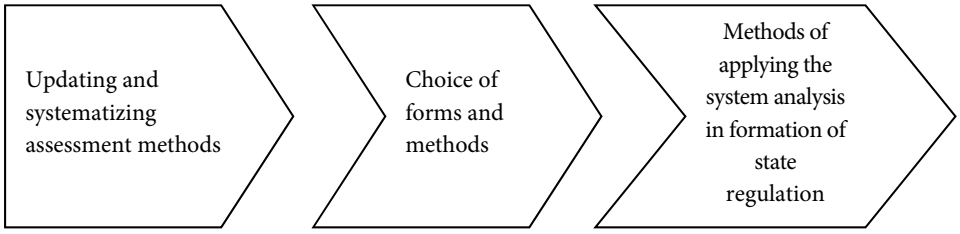


Figure 1. Methodological consistency in conducting a systematic analysis of state regulation on ensuring the activities of local self-government bodies (Compiled by the authors)

The methodology of systematic analysis of the process of state regulation does not only open the space for qualitative analysis but also allows to make an analytical description of the mechanism of interaction in the model of state regulation, the development of methods, methodological recommendations and provisions on the formation of the objectives and their solutions on the basis of the developed models, etc.

In the study of mechanisms and instruments of state regulation, the synthesis of the action of streamlining the system of public administration and self-organization of the economic system is used (Borysenko O., 2017), which can also be used in the assessment of state regulation for ensuring the activities of local self-government authorities.

With their rational ratio, the synergy effect (that is the excess of the final effect compared to a simple summation of the effects of the action of certain instruments of state regulation) will give a much better result than the applied resources of public administration. This effect depends on the quality of the identified priorities in the public administration system, the establishment of proper internal interdependence and the interplay of tasks that are solved in the process of state regulation. Therefore, the purpose and objectives of state regulation are at the heart of the synergy effect. The use of synergies in a systematic approach ensures a qualitative transition from simple planning technology to public administration programming.

Taking into account the aforementioned information, it is necessary, firstly, to maintain the system of regulators of public administration in the sphere of ensuring the activity of local self-government bodies in a controlled state (i.e. within the parameters that must ensure and maintain the stability of management); and secondly, to apply methodological approaches that will facilitate the choice of the right management decision on state regulation as well as the identification of issues that determine further starting parameters of state regulation on the support of activities of local governments.

Thus, the system-synergistic approach to the analysis of state regulation for support of the activities of local governments is the most constructive of the applied areas of systemic research. It directs researchers not only to establish certain regularities in the functioning and development of mechanisms of public administration, but also to develop a methodology for organizing the decision-making process in the context of interconnection and interaction of factors that are in constant motion. This methodology requires the involvement of experts from different fields of knowledge and the application of different research methods, as well as systematic analysis of the public administration system itself and the assessment of the synergistic effect of applying elements of state regulation to support the activities of local governments.

4. Assessment of regulation of the economic development of territories

4.1. Economic changes that have taken place in Ukraine

In 2014, the local self-government reform was launched and the course on decentralization of power was developed in Ukraine. The course on decentralization outlines both prospects and problems of Ukraine's development. Despite receiving positive results of fiscal decentralization reform, the issues of forming and implementing local budgets still remain relevant.

Five years have passed since the introduction of the new model of intergovernmental budgetary relations, but the bulk of local budget revenues is still being generated by deductions from the state budget. In recent years, the volume of transfers in the structure of local budget revenues has increased, so in 2017 intergovernmental transfers from the state budget to local budgets were 1.5 times higher than in 2014. It is appropriate to note that the budget autonomy is largely determined by the level of own revenues. At the same time, the possibilities of local taxation were rather limited.

In addition to the problems caused by management risks and the implementation of the targets, other problems that have a direct or indirect impact on certain budget revenues, such as occurrence of adverse events in the national economy, deterioration of internal macroeconomic conditions of economy functioning (instability of the level

of industrial production and consumption, inflation and other factors causing increase in production costs), significant volumes of shadow economy, the dangers of the budget system and the management of the budget process are related to the inefficient redistribution of revenue sources and liabilities between the state and local budgets, low payment discipline, etc.

The issues of the expenditure part of the budget cause the following risks:

- making decisions that affect the increase in budget expenditures in excess of the approved amounts (increase in social payments, subsidies, benefits);
- increasing the share of budget expenditures on defence and security sector financing due to the military conflict and conducting an anti-terrorist operation in eastern Ukraine;
- increase in budget expenditures due to the influence of foreign economic factor (increase in energy prices, unfavourable change in prices for imported products, changes in exchange rates, etc.);
- debt component of budgetary risks as growth of expenses for servicing and repayment of public debt (as a result of currency, interest, price and credit risks).

In the analysis of the impact of the economic changes that have taken place in Ukraine regarding budgetary decentralization, it is established that they are characterized by such trends. The high level of GDP redistribution through the budgetary system remains. In 2016, the share of consolidated budget revenues in GDP was 32.9% and the share of consolidated budget expenditures in GDP was 35.1% (the highest figure in the last six years). An increase in the total amount of public and government guaranteed debt of Ukraine is observed, as well as a significant increase in budget expenditures to finance its servicing and repayment. The high level of the state budget deficit remains, the growth of which from 1.6% of GDP in 2015 to 2.9% of GDP in 2016 was conditioned particularly by the need to make debt payments, secure defence spending, social protection and security.

Failure to comply with the plan of revenues and expenditures of the consolidated, state and local budgets is caused by management risks and risks of failure to meet the targets. In the structure of state budget expenditures, in particular for 2016, the largest share is spent on financing intergovernmental transfers (28.5%), on social protection and social security (22.1%), on national functions (17.2%), public order, security and the judiciary (10.5%) and defence (8.7). The high level of centralization of budgetary funds remains, which results in the increase in the volume of intergovernmental transfers in the structure of budget revenues. The identified trends make it necessary to assess the unresolved issues for Ukraine.

4.2. Assessment of unresolved issues of regulation of economic development of territories in Ukraine

According to the economic policy pursued by local governments, we are increasingly observing the belief that the natural condition for the formation of the usual reproductive process in the region is the material and financial balance of the development of the region (Dolishniy M., 2001, Varnaliy Z., 2005).

Presently, no appropriate coordination mechanisms have been developed: on the one hand, on the long-term policies of the central executive bodies among themselves over a specific territory; on the other hand, between central and local governments in accordance with the development of goals and priorities at the state and local levels, which leads to a slow reform of the local government in the process of implementing administrative reform at the local level and insufficient rates of economic and social transformation.

Effective implementation of the regulation of economic development of the territories is also hindered by the insufficient provision of local self-government bodies with financial resources. Local governments should have adequate financial capacity to implement development policies. The lack of such opportunities will lead to territorial dispersion of state and local financial resources, inefficient use of them.

The issues of providing local budget revenues for the fulfilment of their own powers are not fully resolved. In rural areas, the list of incomes for fulfilling one's own powers is not enough: for one resident the level of own incomes of rural budgets is 4–6 times lower than the corresponding level of incomes of city budgets (Varnaliy Z., 2005). The relationship of the regional budget with the budgets of the local self-government body can be determined by the indicator of average budgetary provision per capita, calculated for two types of local government: urban and rural. In modern conditions, the cost per inhabitant in rural areas is much higher than in cities through transportation costs, the use of unstable sources of electricity and energy, etc.

Reforming intergovernmental budgetary relations at the basic level should be done in conjunction with the reform of the administrative-territorial structure and the formation of territorial communities in rural areas, which are capable of providing quality services to the population at the level of socially guaranteed standards. Another reason that hinders the effective implementation of territorial economic development regulation is the lack of a mechanism for forming local budgets based on socially guaranteed standards for providing services to the population, regardless of their place of residence. Local budgets are planned depending on the available capabilities of the state budget, which provides only the fair distribution of state resources but does not take into account the objective needs of territorial communities. Budget expenditures per inhabitant of a village, town or city fluctuate 10–15 times (Varnaliy Z, 2005).

This demonstrates significant territorial differences both in the economic development of rural, urban territorial communities and in the living standards of the population.

Finally, the mechanism for regulating the relationship between regional budgets and local government budgets needs to be improved particularly with regard to the setting of rates for local budgets. The state of financial support in most territorial entities depends as before on deductions from the state budget. At the same time the possibilities of local taxation are very limited. It is appropriate to note that the independence of budgets is largely determined by the size and level of their own incomes. Today the vast majority of local budgets are subsidized.

The fulfilment of the tasks facing local self-government requires the solution of a number of problems and the search for priorities. The right choice of priorities is the most important condition for the success of economic transformations, especially the structural ones. In this context, the stabilization of the economy which should be comprehensive in nature, that is be carried out simultaneously in all areas, namely production, finance, budget, taxes, property relations, politics and management and recognized as the paramount task. Macroeconomic stabilization should be based and complemented by specific approaches to local problems.

4.3. Statistical evaluation of the processes of state regulation of economic development of territories in the process of budgetary decentralization

In the process of regulating the economic development of territories, it is important to analyse the factors that determine the need for such regulation. First of all, the efficiency of regulating the economic development of territories is characterized by the process of generating local budget revenues. Therefore, we will analyse the formation of local budget revenues, which on the one hand allow to study the dynamics and structure of these revenues and on the other hand characterize the process of forming the local budget revenues in the current conditions of decentralization of the budget system in order to implement regulatory processes that would meet the real needs of citizens, society and the state.

The main tasks of the analysis of budget revenues are to determine their volume and dynamics. The level of income redistribution through the consolidated budget in the years of Ukraine's independence is characterized by the data presented in Figure 2.

According to Figure 2 data, a considerable part of the budget resources is concentrated in the local budgets of Ukraine. However, in recent years there has been a steady downward trend in the share of local budget revenues in the consolidated budget revenue structure – from 47.6% in 1992 to 18.5% in 2015, which is more than double. In recent years, about 80% of budget resources have been accumulated in the State Budget of Ukraine, which indicates a high degree of centralization of the budget system.

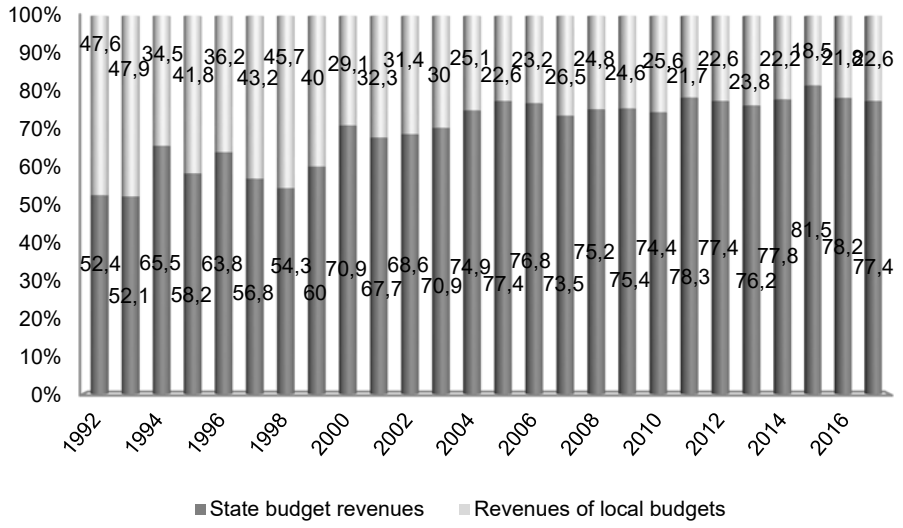


Figure 2. Dynamics of revenue distribution in the consolidated budget of Ukraine, %
Source: Calculated by the authors according to the State Treasury Service of Ukraine.

Expenditures play a leading role in regulating the economic development of the territories, i.e. the need for resources shapes the need for their accumulation. Expenditures of local budgets are considered to reflect the degree of decentralization of power, as they characterize the volume of meeting the needs of the population of a certain administrative-territorial formation, the priorities of its socio-economic development.

As can be seen from Figure 3 in 1992, 1993 and 1997, the share of local budget revenues exceeded their share of expenditures; since 1998 the situation has changed dramatically – each year the share of expenditures exceeds the share of local budget revenues. In the period of 1992–2014, the growth of expenditures of local budgets of Ukraine has significantly outstripped the dynamics of their revenues.

Moreover, each year the lag of such excess increases; if in 1998 the share of revenues relative to the share of expenditures of local budgets in the consolidated budget of Ukraine was 0.95, in 2012 it was already 0.5, and in 2014 and 2017 – 0.52, i.e. it has almost halved (Figure 3), which indicates a significant increase at the level of centralization of budgetary funds.

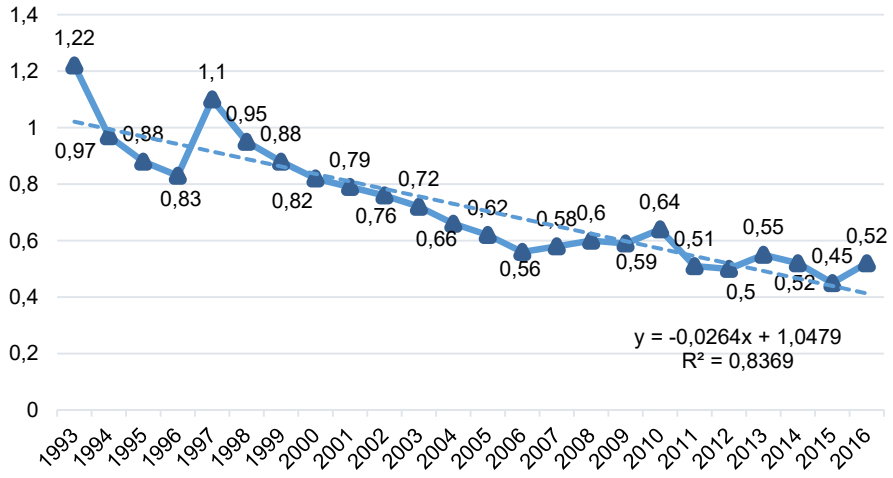


Figure 3. Revenue share relative to the share of local budget expenditures in the consolidated budget of Ukraine

Source: Calculated by the authors according to the State Treasury Service of Ukraine.

Analysis of the local budgets of Ukraine allows to draw a conclusion that existing approaches to their formation do not create any economic incentives for local authorities as for development of their regions (administrative-territorial units), expanding their own tax base as well as to efficiently use the budgetary funds (Lunina I. O., 2014).

The need for state regulation of the economic development of the territories also predetermines that the formation of the revenue part of local budgets takes place in quite difficult conditions. The lack of financial autonomy of local self-government bodies, namely the lack of financial resources and the instability of their revenue sources have become an urgent and acute problem.

Foreign experience is not sufficiently used in the construction of the budget system in the country, and effective economic and mathematical models are hardly applied when determining the revenue bases of the local budgets (Stavnichna M., 2014).

At present, there is no clear justification for the level of decentralization that should be in Ukraine. The level of decentralization is proposed to be measured by such an index as the share of Gross Domestic Product, which that is redistributed through local budgets (Bondaruk T., 2009). We believe that this concerns particularly the share of local budget revenues in the total GDP.

As can be seen from Figure 4, from 1997 to 2000 the share of local budget revenues in GDP decreased; if in 1997 it was 15.7%, in 2000 – 11%. During 2001–2007 there was a steady positive tendency to increase the share of local budget revenues in the GDP of

Ukraine. In 2001 it was 12.2%, in 2007 already 14.9%. In 2012 and 2017 we note the maximum value of this indicator – 16% and 16.8% respectively.

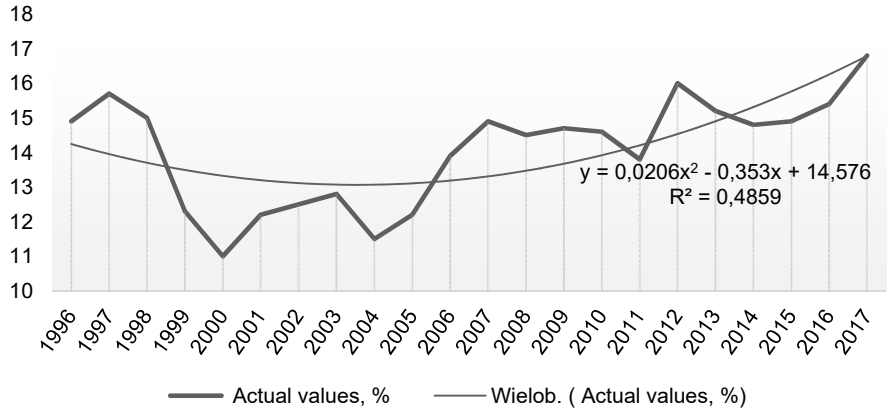


Figure 4. The share of local budget revenues in the GDP of Ukraine

Source: Calculated by the authors according to the State Treasury Service of Ukraine.

Significant fluctuations in the share of local budget revenues in GDP do not allow us to make a conclusion on its clear tendency (decrease or increase). It is possible to identify the trend by building a trend line (Figure 4). The constructed trend line shows that despite the sharp fluctuations in the share of local budget revenues in the GDP of Ukraine during the analysed period its overall trend during this period is upward.

It is advisable to use the indicator of the share of local budget revenues and the share of transfers in the total GDP in order to carry out an in-depth analysis of the trends in regulating the economic development of the territories. The share of local budget revenue in GDP is a fairly significant financial and economic indicator. The place and role of local budgets in the GDP redistribution system are reflected in the data presented in Table 1.

Table 1. Dynamics of local budget revenues and transfers from the state budget to GDP in 1996–2017

Years	GDP (in actual prices)	Revenues		Transfers from the state budget		% of the share of transfers in GDP to the share of revenue in GDP
		Mil. UAH	% of GDP	Mil. UAH	% of GDP	
Before the processes of budgetary decentralization						
1996	81519	12138.6	14.9	1186.2	1.5	10.07
1997	93365	14615.0	15.7	2476.8	2.7	17.20
1998	102593	15413.6	15.0	2202.8	2.1	14.00
1999	130442	16094.8	12.3	2942.4	2.3	18.70
2000	170070	18689.8	11.0	4378.0	2.6	23.64
2001	204190	24972.7	12.2	7237.1	3.5	28.69
2002	225810	28247.4	12.5	8818.1	3.9	31.20
2003	267344	34306.5	12.8	11729.1	4.4	34.38
2004	345113	39593.1	11.5	16819.4	4.9	42.61
2005	441452	53677.3	12.2	23361.1	5.3	43.44
2006	544153	75895.2	13.9	34150.3	6.3	45.32
2007	720731	107050.5	14.9	48701.5	6.8	45.64
2008	948056	137455.3	14.5	63583.2	6.7	46.21
2009	914720	134552.4	14.7	63523.7	6.9	46.94
2010	1094607	159397.1	14.6	78881.3	7.2	49.32
2011	1314000	181600.0	13.8	94900.0	7.2	52.17
2012	1411238	225273.4	16.0	124459.6	8.8	55.00
2013	1454931	221019.4	15.2	115848.3	8.0	52.63
2014	1566728	231702.0	14.8	130160.0	8.3	56.08
During the process of budgetary decentralization						
2015	1979500	294500.0	14.9	173980.0	8.8	59.06
2016	2383182	366143	15.4	195935	8.2	53.51
2017	2982920	502098	16.8	272603	9.1	54.29

Source: Calculated by the authors according to the State Treasury Service of Ukraine, Ministry of Finance of Ukraine.

The analysis of local budget revenues in Ukraine shows that their level increased from 14.9% of GDP in 1996 to 15.7% of GDP in 1997 but in the following years there was a decrease to 11.0% in 2000. From 2005 to 2013, not only the stable dynamics of

nominal earnings growth was observed but also the highest annual growth rates since 2001. Only in 2011 the share of local budget revenues did not exceed 14%, in 2013 this indicator reached 15.2%, but in 2014 it decreased to 14.8%, and in 2016 and 2017 it increased to 15.4 and 16.8%. This trend indicates an increase in the role of local budgets in the distribution of GDP during 2016–2017, i.e. during the budget decentralization process.

An indicator of increasing dependence of local budgets on the state budget is the growth in the volume of transfers. Thus, in the structure of local budget revenues the share of transfers from the state budget grew from 1.5% of GDP in 1996 to 8.8% in 2012, i.e. five times, in 2013–2014 it decreased slightly compared to 2012. In 2013 this indicator decreased to 8.0% compared to 2012. However, in 2014 the share of transfers from the state budget of Ukraine increased relatively to 8.3% and in 2017 it amounted to 9.1%, which is the highest indicator for the analysed period. Having considered the ratio of transfers to GDP to the share of revenues in GDP we note a steady upward trend from 10% in 1996 to 59.06% in 2015, which also indicates an increase in the financial dependence of local budgets on transfers.

The steady growth of transfers from the State Budget of Ukraine to the local budgets leads to important risk factors for financial decentralization in Ukraine. Thus, the priority of the budget policy of Ukraine is to ensure favourable conditions for state regulation of the activities of local self-government bodies, particularly in the formation of local budgets.

5. Analysis and discussion of the results

As a rule, most European countries use the approach of combining an active and passive policy of state regulation, which involves increasing the role and responsibility of local governments for the economic development of the territory, the need to find new tools to stimulate economic development.

The purpose of the first approach is a qualitative change in the structure of the economy. Such a policy is first and foremost applied when it is considered that the market conditions are not sufficient to resolve disparities in territorial development. It envisages raising the level of labour productivity in regions with low levels of development through public investment in local infrastructure, stimulating local development by providing the right conditions for creating and functioning of small and medium-sized businesses.

The aim of the second trend of state regulation policy is to improve regional development by implementing measures that promote the effective functioning of market mechanisms by removing obstacles to labour and capital mobility and ensuring better exchange of information and technology between regions.

Local governments play a key role in organizing territorial development. The state delegates the powers them to maximize administrative and social services to the population, enhances the capacity of local communities to solve local problems.

As a rule, the state is entrusted with the task of forming the concept of a state regulation strategy, the key goal of which should be to maximize the criteria for the sustainability of territorial development, the coherence of interests of territories.

In this case, the state's task is to forecast and plan further regional development. Forecasting and planning at the level of local self-government must be complex, systematic, scientifically based and legally binding. Regional planning is a form of state regulation of the economy and social sphere at the local level in order to resolve acute regional imbalances and social contradictions.

The important tasks of the state are also to ensure the regional unity of reproductive macroeconomic processes, to promote active socio-economic activity of the regions, to form and ensure stable links vertically – between the centre and the regions, and horizontally – between the regions in order to achieve the goal of providing sustainable development of the regions.

In the process of analysing numerous approaches and trends regarding the specific participation of the state and local self-government bodies in regulating the development of territories, a general understanding of the three most important ways of their activity has been formed.

Firstly, the creation of legal and organizational conditions necessary for the functioning of market institutions.

Secondly, state restructuring of the principles of democracy in accordance with the requirements of the market economy. This means a profound transformation, including mastering new methods of managing the economy.

Thirdly, the transition to new forms of regulation, economic and social policies, the goal of which is to find the best way of solving the most important three-fold problem: 1) to maintain stability in a society where social stratification is increasing, the subsistence level is not provided for a large part of the population, the unemployment rate is rising (L. Grygoriev, 2008); 2) stabilize the economy; 3) ensure economic growth.

If the first two areas of state involvement in the economy, namely the creation of a legislative framework for a market economy and the reform of the state itself, are explained by researchers as the need for leading state participation in these processes, the same cannot be stated about all the three areas in general. Starting from this block of questions the differences become particularly noticeable and get a specific history.

At the initial stage of market transformations some opinions were expressed about the possibility of combining the processes of stabilization and structural adjustment of the economy. Some scientists found it unrealistic to carry out a structural transformation in the face of high inflation and a deep decline in production.

Supporters of the possibility of combining these processes argued their own position with the scientific substantiation and positive practical experience of Japan and Germany, in which the processes of economic stabilization and its structural adjustment were combined.

6. Summary and conclusion

Generalization of theoretical developments and positive experience makes it possible to say that in order to carry out structural adjustment in the economy, at least several goals must be achieved:

- to ensure the readiness of the state to carry out structural adjustment of the economy considering such a goal as a long-term strategic one;
- to develop a scientifically sound program of socio-economic development and financial stabilization of the territories in conditions of limited budget resources at a proper professional level, to carry out a thorough review of goals and priorities of a structural policy.

There are both objective and subjective problems, the solution of which will contribute to the economic stabilization of the country and the economic and social development of the territories. Objective problems of the country (differences in the level of development of its regions, difficulties in the coexistence and interaction of public institutions) should not remove responsibility from the political and intellectual elite for the fate of citizens and the state. The subjective reasons that mostly relate to Ukraine include the following:

1. Inability of politicians to take into account the interests of leading social groups that change dynamically in the course of economic and social development, the absence of a long-term strategy based not on the faith but on the conscious participation of citizens in its implementation.
2. Constant preferences for certain oligarchic clans, who try to keep their own income at the expense of other layers of society.
3. Depriving citizens of liberty for protecting themselves from external or other threats and subsequent restriction of their activities, which means stagnation for society, and for politicians – the loss of support from population, government and finally, a good name in history.
4. Negligent attitude to the scientific and social creativity of the individual, the emphasis on simple diligence instead of activity. Attractive for many, the high American standard of living is based not only on the rich resources and vast expanses of the country but also on democratic values promoting the idea of the personal success and vertical mobility within public institutions. The problem of the country is the lack or weakness of encouraging (from above) of protection of one's

own dignity as well as the responsibility of local authorities to citizens instead of paternalism, etc.

The major disadvantage of most of Ukraine's state programs of the past years was that one or the other path of development was offered either as an ideological dogma, whether it was the foundations of the former state plan or liberal programs, or as a set of projects and expenditures. In the transformation period, there was a traditionally high activity of theories, schemes which have not been confirmed by the world science.

Their biggest drawback is the inadequate understanding of the interests of participants of the modernization process: big and small business, different layers of the population, etc. In some cases the efficiency of the market is exaggerated and the importance of forming market institutions is ignored; in others, the efficiency of state regulation is praised and no attention is paid to the objectives of creating the quality market institutions.

The development of a successful territorial development strategy requires the modernization of civil society, business and the state at the same time.

There is definitely dependence of future modernization of the development of territories on modernization in the country and the sustainability of civil society for modernization based on the implementation of the inevitability of compromises and compensations considering that it is impossible to solve simultaneously all the problems in the regions with a significant differentiation of economic development and in a socially heterogeneous society.

REFERENCES

BONDARUK, T. H., (2009). Mistseve samovriaduvannia ta yoho finansove zabezpechennia v Ukraini [Local self-governance and its financing in Ukraine], Kyiv: Express [in Ukrainian].

BORYSENKO, O. P., (2012). Synehriia u konteksti systemnoho pidkhodu do problem derzhavnoi rehuliatornoi polityky u sferi zovnishnoekonomichnoi diialnosti [Synergy in the context of the system approach to regulatory policy problems in the foreign economic activities], Retrieved from, http://www.dy.nayka.com.ua/?op=1&z=459 [in Ukrainian].

BROSIO, G., (1985). Fiscal Autonomy of Non-Central Government and the Problem of Public-Spending Growth. In: Public Expenditure and Government Growth, F. Forte & A. Peacock (Eds.). (pp. 110–135), Oxford: Blackwell.

Comparative Analysis of Territorial Governance and Spatial Planning Systems in Europe/Final Report, Retrieved from, https://www.nordregio.org/research/espon-compass-2016-2018/).

DOLISHNII, M., (ed.), (2001). Rehionalna polityka: metodolohiia, metody, praktyka [The regional economy: methodology, methods, practice]. (pp. 11–15). NAS of Ukraine, Institute for Regional Studies of the NAS of Ukraine, Lviv [in Ukrainian].

DZIEMIANOWICZ, W., JAŁOWIECKI, B.; współpr. KRAJEWSKA M., (2004). Polityka miejska a inwestycje zagraniczne w polskich metropoliach, [Uniwersytet Warszawski]. Centrum Europejskich Studiów Regionalnych i Lokalnych. Warszawa: "Scholar", pp. 130–134.

GRIBANOVA, G. M., (1998). Mestnoe samoupravlenie v Zapadnoy Evrope: Sravnitelnyiy analiz politiko-sotsiologicheskih aspektov [Local self-governance in Western Europe: a comparative analysis of political and sociological aspects], Saint-Petersburg: Herzen State Pedagogical University of Russia [in Russian].

GRIGOREV, L., TAMBOVTSEV, V., (2008). Modernizatsiya cherez koalitsii [Modernization through coalitions]. Voprosyi ekonomiki – Problems of economics, 1, pp. 59–70 [in Russian].

HEIETS, V. M., (1999). Transformatsiia. Modeli ekonomiky Ukrainy [Transformation. Models of the Ukrainian economy]. Kyiv [in Ukrainian].

HOOGHE, E., MARKS, G., (2003). Unraveling the Central State, but How? Types of Multi-level Governance. American Political Science Review, 97(2), pp. 233–243.

KVASHA, S. M., (2013). Metodolohichnyi bazys pryiniattia suspilnykh rishen v ahrarnii politytsi [The methodological framework for taking social decisions in the agrarian policy]. Ekonomika APK – Economics of agroindustrial complex, 8, pp. 12–21 [in Ukrainian].

LIDSTRÖM, A., (2007). Territorial Governance in Transition. Regional and Federal Studies, 17(4), pp. 499–508.

LUNINA, I., (2014). Biudzhetna detsentralizatsiia: tsili ta napriamy reform [Budget decentralization: goals and directions of reforms]. Ekonomika Ukrainy – Economy of Ukraine, 11, pp. 61–75 [in Ukrainian].

MUSGRAVE, R. A., (1959). The Theory of Public Finance: A Study in Public Economy. New York: McGraw-Hill. P. 84.

MYKHASIUK, I. R., MELNYK, A. F., KRUPKA, M. I., ZALOHA, Z. M., (1999). Derzhavne rehuliuvannia ekonomiky [Regulation of the economy]. Lviv: Ukrainski tekhnolohii [in Ukrainian].

NADIN, V., STEAD, D., (2013). Opening up the compendium: An evaluation of international comparative planning research methodologies. European Planning Studies, 21(10), pp. 1542–1561.

REIMER, M., GETIMIS, P., BLOTEVOGEL, H., (eds.) (2014). Spatial planning systems and practices in Europe: A comparative perspective on continuity and changes. London and New York: Routledge.

STAVNYCHA, M. M., SOBETSKA, YU. S., (2014). Tendentsii formuvannia dokhidnoi chastyny mistsevykh biudzhetiv Ukrainy v umovakh podolannia kryzovykh yavyshch [Forming the income part of local budgets in Ukraine in the conditions of crises]. Molodizhnyi ekonomichnyi daidzhest – Youth economic digest, 1(1), pp. 81–87 [in Ukrainian].

VARNALII, Z. S., (ed.), (2005). Rehiony Ukrainy: problemy ta priorytety sotsialno-ekonomichnoho rozvytku [The regions of Ukraine: problems of priorities of socio-economic development], Kyiv: Znannia Ukrainy [in Ukrainian].

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 3 pp. 149–169, DOI 10.21307/stattrans-2020-049
Received – 10.01.2019; accepted – 03.12.2019

The Unobserved Economy – Invisible Production in Households. The Household Production Satellite Account and the National Time Transfer Account

Marta Marszałek[1]

ABSTRACT

Standard measures of economic activity relate to goods and services offered by the market. Stiglitz's report, however, suggests that not only monetary value or economic products create welfare, but non-monetary components should also be included in the System of National Accounts. Although household production is registered in official statistics, the main part of it, i.e. nearly 75-80% of the total home production remains outside of the GDP. The Household Production Satellite Account (HHSA) is a macroeconomic analysis covering both market and non-market home production. The National Time Transfer Accounts (NTTA) is, next to HHSA, an analysis aimed to register and observe the directions of transfers and to present the recipients and givers of home production. Regular estimations provided by the HHSA and NTTA may prove a valuable supporting tool to national accounts, pension systems, or social policy as they provide a great deal of macroeconomic information regarding households, their economic and living conditions, social changes, and welfare.

Key words: generational economy, household production, unpaid work, GDP, Household Production Account, National Time Transfer Accounts.

1. Historical view of the valuation of unpaid work and household production in Poland

The estimations of the unpaid work done by household members for their own use and to satisfy their needs have its relatively long tradition in Poland. The latest one was a foundation to provide the first full sequence of accounts titled the Household Production Satellite Account (Marszałek, 2015).

The first attempts to estimate monetary value of housework were made in the 1970s. In 1976 L. Szczerbińska estimated the unpaid work in Poland as PLN 448 007 million, which was 25.6% in relation to GDP. The monthly value of the unpaid work

[1] Warsaw School of Economics, Collegium of Economic Analysis, Institute of Statistics and Demography, Poland. E-mail: mmars1@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-6810-7977.

per person accounted PLN 1585, which constituted 39.9% of the average monthly net remuneration in the economy. According to L. Szczerbińska's estimations, the monetary value of women's unpaid work was 79.2%, for men's 20.8% of the total housework done in 1976. In that analysis to estimate the monetary value of housework, the replacement cost approach and the market cost method were used (Szczerbińska, 1987; Błaszczak-Przybycińska, 2008).

An expanded valuation of unpaid housework was carried out in the 1980s by the Central Statistical Office in Poland (GUS) and the Polish Academy of Science (PAN). L. Szczerbińska complied the next valuation of unpaid work made in households in Poland in 1984. The analysis was a continuation of the researches of the estimates of extended final consumption expenditure. Childcare, adult care and disabled persons care, were excluded from the calculation then. Why care as a group of non-market household activities is outside the official estimations of GDP? The main reason is that each of caring activities generate only costs and they are not significant for total consumption of all housework in households (Marszałek, 2015; Błaszczak-Przybycińska, 2008).

The next analyses of housework were provided in 1990s in Warsaw School of Life Science (SGGW). K. Niewierowska observed households of farmers in Drohiczyn commune. Based on empirical analysis she counted the unpaid work in that type of households in one of region in north-eastern Poland. The monthly value of housework was estimated using two methods: replacement cost method and simplified method. The average monthly monetary value was different for each method. First of them was counted as PLN 658, second PLN 546. The average remuneration in the region was PLN 578. The author of that calculation noticed that the highest wage in the analysis was assigned to food management (Niewierowska, 1997).

In 1995, the estimation of the monetary value of housework focused on women's work and their participation in creating non-market household production. B. Mikuta applied and implemented two different approaches: *simplified method* and *replacement cost method* (Mikuta, 1998). The *simplified method* was calculated as the amount of time of performing activities multiplied by unified gross remuneration rate. The *replacement cost method* was estimated as a sum of an average duration of some groups of household activities multiplied by an average gross remuneration rate for each group of housework (Błaszczak-Przybycińska, 2008: pp. 111-112).

The monthly monetary value of home activities was counted as PLN 808 (*simplified method*) and PLN 722 (*replacement cost method*). The average monthly gross remuneration in Płock region in 1996, where the survey was carried out, amounted to PLN 929 (the average monthly remuneration in Poland was PLN 874). The dimension and the monetary value of unpaid work confirm that economic impact of households'

activities could be more significant in the economy than it is observed in official statistics, which registers only a small part of household production.

The most comprehensive estimations of the monetary value of unpaid work for women and men in households in Poland were based on time distribution of households in Time Use Survey for Poland 2003/2004 and 2013. The Time Use Survey 2003/2004 and 2013 was harmonized with the European Time Use Survey, and it guarantees the comparability with results of other European countries, where the survey was carried out.

I. Błaszczak-Przybycińska developed the basis and methodological guidelines to further analyses of the non-market household production (Marszałek, 2015). The author applied the *input method* to calculate the value of unpaid work in 5 groups of home activities: household upkeep, food management, making and care for textiles, child and adult care, help for other households (Błaszczak-Przybycińska, 2007). Groups of home activities were corresponding with households' functions which are fulfilled to meet own needs or other household member's needs. Only the *household and family care* group was taken into account from TUS 2003/2004 and 2013 because other activities, e.g. personal services, hobby, interests, sport were excluded from the estimation and it was in accordance with the *productivity criterion* also known as *a third part criterion*, *a third person criterion* or *M. Reid criterion*. That criterion assumes that only activities that could be done by a hired person without losing any utility for that household can be valuated (Eurostat, 1999, p.7). Thereby each household's productive activity can be valuated using the market cost of similar services offered on the market.

In order to estimate the monetary value of housework, also *Survey of Wages According to Professions 2002* (GUS, 2004) was used. Hence, average hourly wages of professions were adapted to the housework monetary valuation.

Monthly average gross value of housework in 2004 was assumed at 1000 PLN per person. The relation of women's household work to men's household work was 1:0.574 in 2004. In 2013, the monetary value of housework amounted to PLN 1672. The proportion of women's housework and men's housework was 1:0.576 in 2013.

In comparison with 2004, the highest changes in the value of housework in 2013 were noticed in the case of help for other households and childcare. Probably it is convergent with the tradition. Polish society is more traditional than the societies of Western European countries. The care, mainly childcare, adult care and informal help for other households, e.g. supporting elder parents or grandparents is closed to family model and social expectations. Households in Poland take care of their family members more often than they outsource the care, although they are burdened with other liabilities.

2. Data source of HHSA and NTTA

The time use survey 2003/2004 (TUS 2003/2004) and 2013 (TUS 2013) was the fundamental source of information about the time spent on unpaid household work. Those surveys were harmonized with the European Time Use Survey methodology, which ensures comparability with other countries.

In both time use surveys the respondents were 15 and above, in TUS 2013 also the sample of 10 and above was observed. All activities were registered in diaries in 10-minute intervals. In TUS 2003/2004 and 2013 the lists of more than 200 activities within ten groups were arranged. In both surveys the six household types were distinguished by the main source of income: employees, employees-farmers, farmers, self-employed, retirees and invalid pensioners and those living on unearned sources other than invalid-pension and retirement. Every respondent registered all activities done in 2 days: one day from Monday-Friday and the other day: festive day during Monday-Friday or Saturday-Sunday. In the Household Production Satellite Account 2011 the system of wages was applied in accordance with the wage for the day from the time use survey.

As far as the valuation of household work and production is taken into account, the main group of activities was the *household and family care* group. In accordance with Margaret Reid's third party criterion only productive activities can be valued in the estimation of unpaid work and non-market household production. Some activities such as personal services must be excluded from the estimation. Finally, 47 household activities within 5 groups were taken into account in the calculations. The groups of activities in the estimation of household work were compatible with the household's functions. In the analysis, the following were distinguished: household upkeep, food management, making and care for textiles, care (childcare and adult care), help for other households (voluntary work for other households). Also, transport and household management were estimated in the analysis in proportional part for the each group of activities.

3. Household production satellite account for Poland

The household production satellite account (HHSA or HPSA) is a full sequence of accounts with information about the value of domestic work, intermediate consumption and capital which are collected and used in households for own needs or other households member's needs. The HHSA could be a comprehensive compilation and a supporting tool for the national accounts. It presents the monetary value of unobserved household production generated for themselves and outside their household, e.g. grandparental help, adult care for elder parents, neighbourly help.

International discussions on that topic have been continuing for at least 4-5 decades. The Eurostat (European Union Statistical Office) and other global institutions recommend that to better understand the social and economic conditions of households and their contribution to the national economy is to estimate the monetary value of domestic work and home production as the household production satellite account. That compilation of full sequence of accounts provides information about unpaid products and services which were produced in households but are not offered on the market. The households' goods and services made for own use do not have a price, but they are valuable. No price is not equal with no value. Household members used the products made in home, e.g. home-made dinner, clean and tidy home, washed clothes, childcare, help for other households without any market price and cost. Also, any market transaction exists in home production for own use. If someone acquires the same goods or services, they will buy it on the market and that transaction will be noticed in GDP.

In international calculations of the HHSA, the value of homemade products and services is estimated at nearly 10-20% of total household production (market and non-market), and it was called *market production* and registered in the national accounts. The major part of the home production is generated for own consumption (Marszałek, 2015). It is called *the non-market household production*, so it is non-observed in the national economy and it is made outside GDP. Non-market home production does not generate any monetary transactions so it is excluded from the market. Although the non-market home production is outside the national statistics, it has a crucial role in well-being research of the households' economic and living conditions.

The estimation of non-market household production is based on calculations of the amount and value of housework, intermediate consumption and capital. The domestic work has the basic and crucial share in total home production made for their own final consumption. The next critic point of the home production estimation is calculating consumption. In order to provide a comprehensive view of the production's process in households, consumption is divided into three types: final consumption, intermediate consumption and capital consumption (depreciation). The final consumption, which means the proper using up of a product: eating food, wearing clothes, feeding baby. Secondly, there is intermediate consumption, which covers the products as a part of the production process, e.g. vegetables, meat, fruits used when cooking dinner. Thirdly, consumption refers to capital services produced by the machines, appliances required in the production process. Capital services consist of: consumption of fixed capital, i.e. depreciation of equipment, machinery, appliances used at home, and interest referring to the acquisition of capital. In the Household Production Satellite Account only the consumption of fixed capital (depreciation) is included (Varjonen & Aalto, 2006, p. 22).

4. Valuing of housework and household production in HPSA – methods

The full sequence of accounts for household production (market and non-market) in Poland and the methodology is based on the framework of national accounts consisting of the whole sequence of accounts (Eurostat 1999; Eurostat 2003).

In literature and economic practice two different methods are recognized and used – *input method* and *output method*. *Input method* was implemented more frequent than *output approach*. When considering the choice of a specific production valuation method, some important assumptions should be included.

Input method is better known and permanently developing. The pioneer Eurostat's framework recommends to apply the *input method* for estimations of unpaid work in households (Eurostat, 1999).

Input method is based on the structure of time distributing during the 24-hours by all household's members. Data of time budget is using from the *time use survey*. Time is a main component to estimate the total monetary value of housework in each group of activity, i.e. house maintenance, food preparation, making and caring of clothes, childcare and adult care, volunteer work. Afterwards, when the structure of daily distribution of time between all housekeepers is recognized, the selection of a specific approach should be implemented (Figure 1).

Input method	Output method
The total value of housework (hours x professional rates of similar market work or service) + other taxes on production subsidies on production + consumption of capital = gross value added + intermediate consumption = **total household production**	The value of output (quantity x market price) = **total household production** intermediate consumption = gross value added consumption of the capital other taxes on production + other subsidies on production = mixed income (with compensation of employees and capital)

Figure 1. Input – output method of valuing the non-market household production

Source: Based on Eurostat 2003, pp. 12.

Finland (Varjonen, Hamunen & Soinne, 2014; Varjonen & Aalto, 2006) and Germany (Varjonen & Rüger, 2008) also applied the input method and compared the results per capita between their countries. Hungary adjusted the input approach to valuing home production by size of household and type of the family. France provided

the estimation based on input method with three definitions of housework (unpaid work). The results of the widest perspective showed that over 80% in relation to French GDP were generated in households (Poissonnier and Roy 2013).

Poland reflects continuing works of the valuing housework and home production with the input method (Błaszczak-Przybycińska & Marszałek, 2019; Błaszczak-Przybycińska, 2008, 2007; Marszałek, 2015).

Individual estimation was proposed by the United Kingdom (Holloway, Short, Tamplin, 2002; Ironmonger & Soupourmas, 2009). The UK used the output method, which was more proper to make comparisons with GDP and production counted in the national accounts, but it did not cover a lot of controversial issues in home production estimation, e.g. caregiving activities, volunteering work, etc. The most crucial and debatable point is that the output method does not ensure the total overview of productive results of home activities. The output of washing clothes is possible to recognize if we have, for example, the total amount of clean clothing. The result of some housework in the output method is impossible to indicate if the effect is hard to identify, e.g. the output of caring children.

The input method is based on time spent used on home activities, so it can be countable and it is more useful to compare between the regions or the countries. This method is divided into two different approaches: replacement cost and opportunity cost. The replacement cost approach uses the rates of professions' salaries, which is calculated in the sum of the value of housework (1.).

Replacement cost method:

$$y = t_{m/w} * r, \tag{1.}$$

where:

$t_{w,}\ t_m$ – time of all housework for women or men (in hours and minutes)
r – average rate per hour of professions (in market price)

The opportunity cost method provides the information about the hypothetical value of housework in relation to type of profession which is realized by individuals. This approach is less applicable than the replacement cost approach because the specification and differences between rates of the salaries determined not the volume but the monetary value of housework. If more paid specialists live in households, e.g. doctors of medicine, lawyers or others, their value of housework will be more expensive than housework of lower paid jobs, e.g. builders, nurses, teachers, home-cleaners (2.).

Opportunity cost method:

$$value\ of\ housework\ (opportunity\ cost\ method) = t_{m/w} * r_s\ , \tag{2.}$$

where:

$t_{m/w}$ – time of all housework for women or men (in hours and minutes)

r_s – rate per hour for different professions according with the specialization of someone's job (in market price)

The whole sequence of valuing processes (equations, scheme of the uses-resources tables of accounts) were presented in the previous analyses of the valuation of housework (Błaszczak-Przybycińska, 2008 & 2007; Błaszczak-Przybycińska & Marszałek, 2019 & 2015) and in the Household Production Satellite Account for Poland (Marszałek, 2018 & 2015).

5. Full sequence of accounts – Household Production Satellite Account 2011 – results

The first full sequence of accounts in the Household Production Satellite Account for Poland was in 2011. The output of household production (sum of market and non-market production) in Poland reached PLN 1109.8 billion (Table 1). Gross value added of household production was PLN 807.3 billion, of which 15% was included in the national accounts. The major part of household production is outside the market and official statistics. The fact that such a large amount of household production is not registered in the system of national accounts (SNA) might contribute to incomplete information about conditions of households.

Value added of housework is counted as more than 75% of total household production, while intermediate consumption – goods and services used in the production process – constitute 16 per cent of total output. The value of unpaid work made at home is the most important and a major component of non-market household production, because it provides information not only about time distribution in households by functions, but also informs about cost inputs of time spent doing housework. Input of domestic work constitutes the starting point for other social and macroeconomic estimations, e.g. for advanced analyses of childcare or adult care for family and social policy.

Sums of capital and intermediate consumption are lower than one third part of the total non-market production, which confirms that the value of labour is the most significant category of the household production valuation. Therefore, working on regular implementation and providing the household production satellite account should be pointed out at solving problems of estimation and harmonized methodology of domestic labour calculation.

Table 2. Household production in Poland in 2011 (million PLN)

Components of household production	Household production		
	SNA (market production) in GDP	**NonSNA** (non-market production) outside GDP	**Total (SNA + nonSNA)**
Value of labour (number of hours spent on housework x hourly rates)	◦	644 390	644 390
Paid domestic staff	796	◦	796
Housing services produced by owner occupiers (rents of equivalent rented accommodation)	53 160	◦	53 160
Own-account house construction	42 088	◦	42 088
Agricultural production for own use (hunting, fishing, picking berries and mushrooms)	7 598	3 301	10 899
Taxes on production	4 458	893	5 351
Subsidies on production	-5 030	-20 619	-25 650
Net value added	**103 070**	**627 965**	**731 035**
Consumption of fixed capital (depreciation)	21 515	54 708	76 223
Gross value added	**124 585**	**682 673**	**807 258**
Intermediate consumption	145 595	156 973	302 567
Output (household production)	**270 179**	**839 646**	**1 109 825**

Source: Own calculations based on the method proposed by Marszałek (2015), p. 163-167.

The incomplete data in official statistics, skipped in the non-market value of home production, might provide an incorrect view on social-economic analyses of welfare and living conditions, and as a result it may generate false conclusions of the situation of households. Households use all their resources: individual and group, cultural, social, monetary, and others to well-organized life and to fulfil needs. The utility that households strive for is in some sense produced by them. Households, which are both consumers and producers, perform basic functions with using not only monetary goods and services, but also non-monetary units. Therefore, it should considered by official

statistics, not only in the core of national accounts, but as an additional complementary analysis – Satellite Account.

The contribution of households in the formation of GDP was counted as PLN 124.6 billion, and it was near 8.2% of total market production in 2011 in Poland. If the non-market home production sums up with market production the household production in relation to GDP will achieve 52.8%. The home production made outside the market was assumed at 44.7% in comparison with GDP (Figure 2).

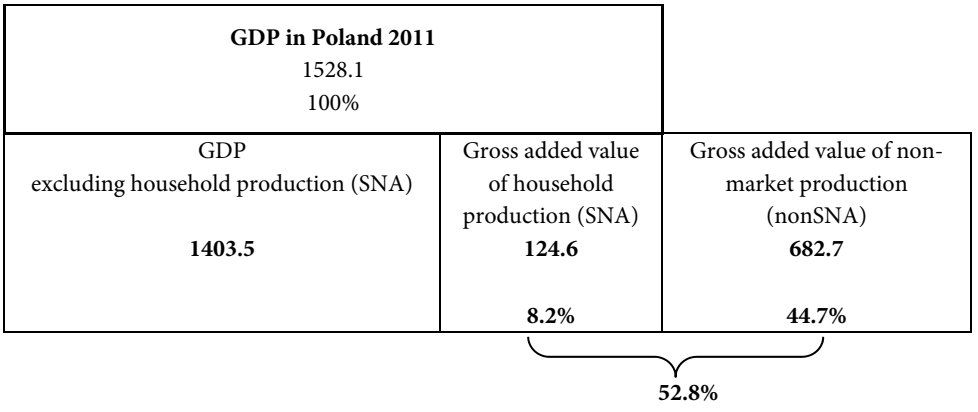


Figure 2. Structure of GDP, market and non-market household production (billion PLN)

Source: Own calculations.

The non-market household production is invisible and it has no full reflection in the European System of National Accounts, which constitutes the gap of that value in the economy. If the non-market household production is included in official statistics, GDP will increase more than 30.9%. The extended GDP concept assumes inclusion of the non-market home production in the national accounts, therefore total production made in households as a goods and services offer to other households members and on the market will achieve 36.5% of the extended GDP measure.

Households carry out a lot of different functions to fulfil individual and group needs inside and for the other family members or neighbours outside home. The results of the monetary valuation of market and home production is presented in Figure 3.

The most diverse of the principal functions is housing. The sum of SNA home production and household upkeep spent for own use without any monetary transactions is the most valuable of all the groups of activities made at home. It assumes more than one third of total household production in 2011.

In the economic practise, housing consists of a wide scope of different domestic works carried out in a space called a dwelling. Housing production is understood in the Satellite Account in a deeper and broader sense than in the official statistics. Only a small part of the total home production made as housing is covered in the core national accounts. SNA-housing production contained: *housing services produced by owner-occupiers*, *own-account house construction*. Also, *paid domestic staff* is included in market boundaries called SNA housing. Non-SNA home production includes all other equipment related to maintain home clean and tidy. Home production covers also furnishing, minor repairs, gardening and yard maintenance. Only goods related to hobbies and interests are excluded from the calculations in the Household production satellite account (Błaszczak-Przybycińska & Marszałek, 2019).

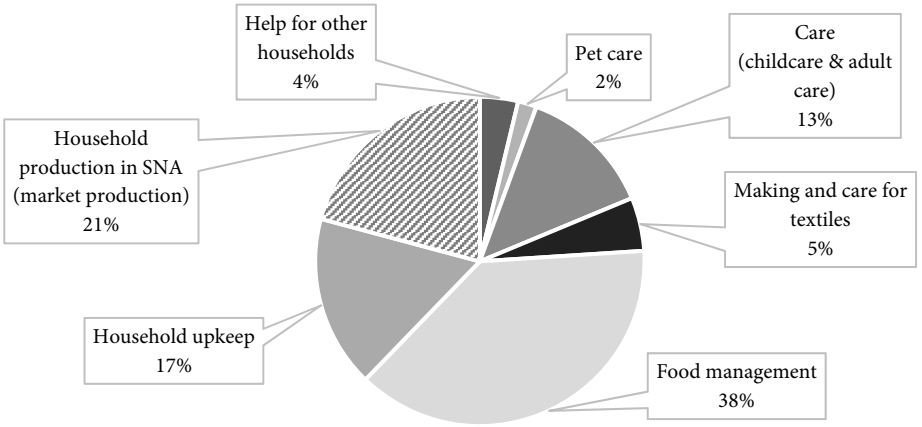


Figure 3. Structure of the household production for Poland in 2011 by functions (in %)
Source: Own calculations.

Considering groups of domestic work separately, the most valuable and crucial to life is the food management. The preparation of meals and snacks consumed within the household, so the output of the services, is fully clearly visible and tangible in opposite to other services offered in the households, e.g. childcare or help for other adults from the same household or outside home. The SNA food management covers *the agricultural production for own use (hunting, fishing, picking berries and mushrooms)*.

Non-SNA food management provides the production of meals, snacks, baking, preserving and other related activities, such as buying groceries, utensils and appliances for the food preparation. Also, non-SNA home production of food management included housework relative to washing dishes, setting the table, cleaning after a meal and other related activities. The food management covers 38% of total home production made in households in 2011 (Figure 3).

Making and care textiles do not have a crucial role in the formation of home production in Polish households now. Twenty-thirty years ago, when the lack of many goods and products was noticed in Poland, the households members, mainly women, were engaged in making clothes by themselves for own use or for others. Currently, the most visible services made in households are related to washing and ironing clothes, rarely mangling, repairing clothes or footwear. Near 5% of home production is provided by making and care textiles.

Childcare and adult care is the most complicated function to organize and estimate the monetary value. Caregiving assumes not only services offered to other dependent underage or adult persons but also goods used during the production process. It is hard to distinct, select and integrate them into the Household production satellite accoun. The most troublesome for home production of care is to estimate the value of time input dedicated to children or adults. Some of activities are treated as a second activity done during other housework, e.g. the main activity is cooking dinner, the second is passive taking care of a child. Therefore, it is important to count the proper part of home production in providing care. Production related to the care of a household member is not registered as such in the core national statistics. The home care of one's own children or adult family member who lives in the same household or outside in a separate household is supported by allowances, e.g. parent's allowance, nursing support for elder or disabled person. In the Household Production Satellite Accoun, allowances were taken into account in the form of subsidies on production. Care provided at home was counted more than one fifth of the household production.

Pet care has a similar concept of the estimation to childcare and adult care. Some researchers claim that caring for pet is discussable to calculate it into the Household production satellite account. If it is treated as a hobby, it should not be included into the calculation of home production. But if it is considered as a work which could be done on the market by a third person, it will be productive for households and in accordance with *the third part criterion*.

Help for other households has provided 4% of home production. Some activities are dedicated by elder person to other family member who lives in a separate household or to friends or neighbours. That group of housework is also a minor component of the household production determined by the amount and the value of domestic work but it is important for social and family life. Probably, in near future the role of help for other households or voluntary work will increase, which is related to demographical changes in the Polish society.

6. National Time Transfer Account for Poland

The National Time Transfer Account (NTTA) for Poland in 2013 was based on the Time Use Survey 2013 results of time dimension of households' members. The NTTA is the part of the global concept of understanding the generational economy (Mason & Lee, 2011). Making home production and consumption visible is a crucial assumption of the NTTA conception. Also, private and public transfers in households are important to be registered into the National Transfer Accounts (NTA). Both formations of accounts: the NTTA and NTA, present the receivers and givers of home production and consumption, public and private transfers. They figure interactions between family inside and outside their own households. Those calculations of the NTTA and NTA could be important, valuable and crucial information carriers for the core national accounts. In the case of the NTTA, the number of members living in a household is not relevant, aspects such as age and sex of the receiver or giver of the unpaid home production are more informative. The fundament aim of the estimation of different transfers could provide the existing gap in social statistics of households' role and households' productivity in the economy.

In the National Time Transfer Accounts, each group of housework which defined the household production was counted as a result of time spent on housework multiply by the average rate per hour of professions corresponding to selected domestic productive activity. The monetary value of childcare is based on an aggregate of some types of average rates of different professions, e.g. teachers, nurses, coaches, lecturers. The same concept of estimation might also be applied to other types of home services: household upkeep, food management, making and care for textiles, help for other households. The distinction between average rates of professions for each group of domestic work is significant because the knowledge, skills and abilities are different for them (Table 2).

Table 2. The average net hourly rates for monetary valuation of home production in National Time Transfer Accounts for Poland in October 2013 (in PLN)

Groups of activities of housework	Average net hourly rates Oct. 2013 (PLN)
Household upkeep (cleaning)	8.01
Making and care for textiles (laundry)	8.68
Food management (cooking)	8.31
Household maintenance	10.37
Gardening	10.09
Household management	13.41
Pet care	9.00
Shopping and other services	11.21

Table 2. The average net hourly rates for monetary valuation of home production in National Time Transfer Accounts for Poland in October 2013 (in PLN) (cont.)

Groups of activities of housework	Average net hourly rates Oct. 2013 (PLN)
Travelling	10.98
Childcare (household)	21.61
Childcare (non-household)	22.44
Help to an adult family member (household adults)	11.27
Help to an adult (non-household adults)	14.54
Informal help to other households (volunteering care)	10.41

Source: Own analysis based on POLNTA project realized at SGH.

The transfers of production and consumption in the NTTA are based on detailed estimations and the sum of the time units allocated in different home productive activities to fulfil own needs or other housekeepers' needs or expectations, and also for outside the household, e.g. for the elder parents, grandparents, other family, neighbours, etc., multiplied by average rates of professions for different groups of housework. The value of household production was estimated separately for women and men. Their arrangement of performing home activities is different, which is registered in the final calculation of the transfers between generations and households (Table 2).

In the NTTA, the process of selected information is analogical to the HPSA. Time is an important component to estimate the monetary value of unpaid work (housework) and home production. The differences focus on the average rates of professions. In the HPSA, the average monthly wages of professions are directly from the "*Survey of Wages According to Professions*". Market wages were matched to similar home activities, e.g. cooking dinner with average monthly wage of a sub chef of the cook, not the cook – because in the HPSA only the lowest wages were implemented in the estimation of housework. It indicates that the valuation of housework and home production is not overestimated but it is more possible that some activities which were registered in time use survey as a secondary activity were not counted and observed in HPSA.

The rates (wages) implemented in the NTTA are different than in the HPSA because they are the sum of average of few various rates of professions for the group of activities, e.g. to calculate the value of childcare from own households of the selected wages of teachers, tutors, nurses, babysitter, etc. were used in the estimation. Both methods implemented in the NTTA and in the HPSA are effective, but they ensure various type of information for the analyses of distribution and transfers of time in households.

7. Unobserved production in households – NTTA results

The idea that ageing is only a developed country problem is no longer valid in practise. Rapid fertility decline in many developing countries and results of that have been observed. The response is the concept of National Time Transfer Accounts (NTTA), which provides the statistical tool to observe how societies are dealing with age or generational issues.

Households is the most differential sector in national accounts, which covers the entire economy although the monetary transactions which exist are an insignificant part of the total home production. The major part of goods and services is invisible and unobserved in official statistics and registry. Even though a lot of products that are consumed in households are not available and do not take action on the market, they allow to fulfil fundamental individual or group needs. The non-market production is not a monetary cycle but it occurs outside the market. Observing and registering transfers inside and between the households could be provided by the National Transfer Accounts and the National Time Transfer Accounts, which compare the public and private transfers.

The Time use survey 2013 (TUS 2013) for Poland registered that men spend on average one hour per day more than women doing paid work, while women spend more time at home or making duties related to housework (Marszałek, 2016). The proportion of total time spent on doing tasks at work and home is different. The entire time of paid work and domestic work is higher for women than men (Figure 4).

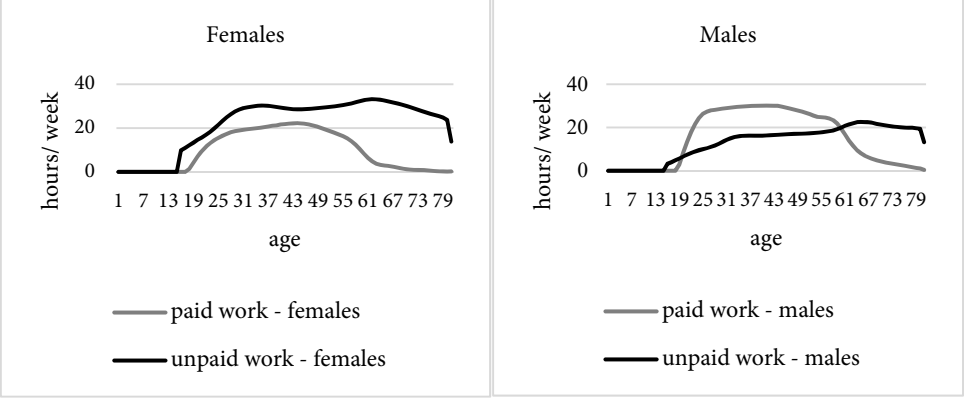


Figure 4. Time transfers of total paid and unpaid work by age and sex in Poland 2013 (in hours/ week)
Source: Own calculations based on POLNTA project carried out at SGH.

Moreover, women aged 25-40 and 50+ are the most important producers of the non-market household production, which means that they are the highest givers of

unpaid domestic work. Men are responsible for gaining income for their households, women for home, even though they are both employed. It confirms the women's double-burdened of domestic work but also stereotypes, social roles and patterns that Polish society is still deep traditional. However, mainly young couples in big towns or cities declare that households in Poland sharing the majority of home tasks between spouses or partners. During the age women are more often unemployed than men in the age. In Poland, a social expectation of women is a deeper commitment in home tasks than in market paid job. Especially help for elder parents or disabled person is dedicated more often for women than men. Therefore, social, cultural and traditional factors are crucial and decisive indicators influence in the population transfers of time, and next in home production and consumption.

The National Transfer Accounts (NTA) and the National Time Transfer Accounts (NTTA) present the economic life cycle as a universal feature of the society. For a long period at the beginning and the end of life people consume more than produce regardless of the type of work: housework (unpaid) or market (paid). In the middle of life there is a period when more is produced than consumed. Many social, behavioural, cultural, educational, political and other factors influence how the labour income, consumption and home production vary with age.

The NTTA profiles of production and consumption present a longitudinal formation of the households' inside and outside transfers (United Nations, 2013). They indicate the toward of transfers not the person to whom the production is offered. It also provides the information about receivers of home production (consumers) and givers of the non-market goods and services.

The current aggregate level of the economic life cycle also reflects the population age structure and the results of activities performed during life. At the beginning, very young and teenager populations, the life cycle deficit equal consumption minus production is dominated. Over the years, when the demographic transition in population age exists, the proportion of the life cycle deficit or surplus is melting down.

During the observations of time distribution in households, the current life cycle stadium reflects. The highest receivers of home production are children, both females and males aged 0-6 (Figure 5). Over the age of 6, they are more decisive and have more skills and abilities to better organize their life and to arrange domestic tasks. In TUS 2013 for Poland it was noticed that children aged 10 and over are not involved in doing housework. Probably, their parents do not expect any or only small portion of help at home as they take the view that children should focus mainly on how better to organize scholar activities and the rest, not on being involved in home tasks.

The most burdened group is women aged 50+ (Figure 5). Women in this cohort are a part of *a sandwich generation*. It means that people in this group are doubly burdened, they help for their elder parents and they take care of their growing children,

who need a lot of attention and support. Sometimes the *sandwich generation* participates in help or care of grandchildren. In Polish households women aged 50+ are pensioners, but they still have the ability to support others, e.g. elder neighbours in daily activities or they work outside the market registration, usually women do care jobs: childcare or adult care.

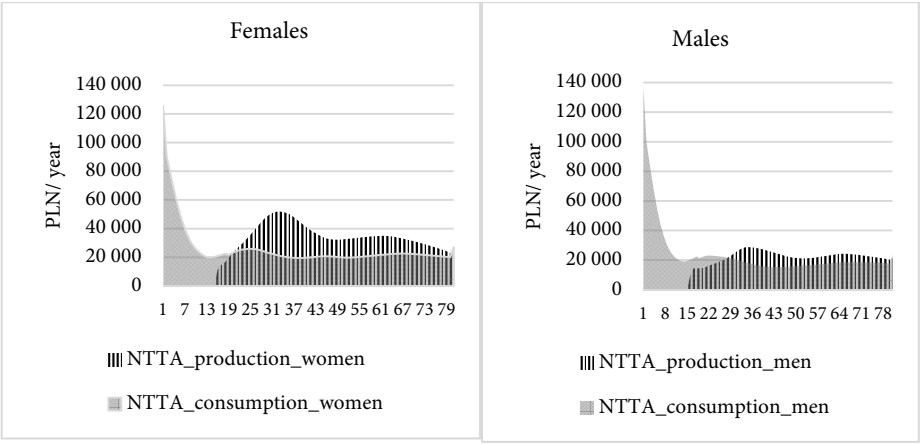


Figure 5. NTTA profiles of the production and consumption transfers by age and sex (PLN/ year)
Source: Own calculations based on POLNTA project implemented at SGH.

In Poland, social expectations focus on providing care and help for the elder family members or children by women. Even if men share housework, the major part of total domestic duties are the women' domain. Tradition and a fundamental view are stronger than the social changes which are observed especially in the big cities and in households with men with higher education.

Men make less non-market production than women across their life. The highest volume of domestic work is for men aged 30 to 45 and 60+. Based on the NTTA data, the value and amount of home production is observed by age and sex. The towards of the transfers is not fully registered. Men aged 30-45 do a lot of their housework made not for themselves but for children. It is a time when men and women have children, and they carry out different domestic work for the youngest generation. The opposite perspective is observed in a cohort of men aged 60+, especially the ones who live alone, make a non-market household production for themselves. In Poland, the most valuable group of home duties is food management. Men aged 60+ spend most of their time on activities related to the preparation of a meal (Figure 5).

8. Conclusions

The social and economic changes which have been observed in the last decades focus not only on the market production made in the economy, but also on the non-market factors that influence the general and current macroeconomic horizon. The macro perspective centres on development and economic growth. The micro perspective, which is households' domain, covers all individual and group needs, social expectations, economic decisions. Home production provides the fulfilment of different needs, which is indirectly reflected in companies, financial and governmental institutions, and finally in the national economy. Moreover, households generate value added of their non-market productive activities, e.g. home repairs, cleaning, preparing food, making textiles and clothes, childcare, etc. Although domestic work does not have any market price, it provides a lot of different needs, so it has a value. Therefore, it should be reflected in the core national system of accounts. GDP, value added, national income are formatted not only based on market decision. The social behaviour, needs and expectations create the final demand even if a lot of housework is made for themselves in own households. The real impact of households for the economy should be regularly estimated as an additional comprehensive sort of households information to the official statistics called *Household Production Satellite Account* (HHSA or HPSA). The HHSA with the *National Time Transfer Accounts* (NTTA) are the multidimensional sequence of accounts with information about the volume and value of home production and consumption across the life cycle. Using the information on the social and the economic situation of households can provide a solution to better organize e.g. social and family system, pension system, the law, entrepreneurs' decisions. The observation of time transfers can be supporting in organizing the families' life or adjusting the working system, especially in more flexible work time or partly-time jobs, which will be reflected in better use of the labour resources. The influence of the changes in some areas is necessary, because a lot of international phenomena are observed, such as the aging of society or low fertility rate. Therefore, new but not costly methods and statistical tools to measure and observe the households situation is required and needful. It is necessary to better understand and register the real conditions of the largest and most dimensional sector of the economy (in a more detailed way).

REFERENCES

BŁASZCZAK-PRZYBYCIŃSKA, I., MARSZAŁEK, M., (2019). Satellite account of household production. Methodological remarks and results for Poland, *Econometrics*, Vol. 23, No. 1, Wrocław, pp. 61–76.

BŁASZCZAK-PRZYBYCIŃSKA, I., MARSZAŁEK, M., (2015). Own work of households value estimation on the basis of time use survey (Wycena wartości pracy własnej gospodarstw domowych na podstawie badania budżetu czasu), in: GUS, Time Use Survey 2013, Part I, ZWS, Warszawa (in Polish), pp. 131–183.

BŁASZCZAK-PRZYBYCIŃSKA, I., (2008). Household Production as a Source of Wealth (Produkcja gospodarstw domowych jako czynnik dochodotwórczy), Oficyna Wydawnicza SGH, Warszawa (in Polish).

BŁASZCZAK-PRZYBYCIŃSKA, I., (2007). Estimation of Unpaid Work in Polish Household, *Statistics in Transition*, Vol. 8, No. 3, pp. 547–561.

BŁASZCZAK-PRZYBYCIŃSKA, I., (2006). Methodology and Empirical Results of Time Use Surveys in Poland, *Statistics in Transition*, Vol. 7, No. 6, pp. 1345–1360.

BŁASZCZAK-PRZYBYCIŃSKA, I., (2005). Estimation of Housework on the Basis of Time Use Survey Data, in: CENTRAL STATISTICAL OFFICE, Time Use Survey 1st July 2003–31st May 2004, Statistical Studies and Analyses, Warsaw (in Polish).

DURAN, M.-A., (2007). The Satellite account for unpaid work in the community of Madrid. La Suma de Todos, Community de Madrid 36.

EUROSTAT, (2003). Household Production and Consumption. Proposal for a Methodology of Household Satellite Accounts, Task force report for Eurostat, Unit E1, Luxembourg.

EUROSTAT (1999), Proposal for a Satellite Account of Household Production, Eurostat Working Papers, 9/1999/A4/11, Luxembourg.

GUS, (2014). Survey of Wages According to Professions 2012 (Struktura wynagrodzeń według zawodów w październiku 2012 r.), ZWS, Warszawa (in Polish).

GUS, (2004). Survey of Wages According to Professions 2002 (Struktura wynagrodzeń według zawodów w październiku 2002 r.), ZWS, Warszawa (in Polish).

GOLDSCHMIDT-CLERMONT, L., (1993). Monetary value of non-market productive time. Methodological Considerations. *The Review of Income and Wealth*, No. 4, pp. 419-433.

HAWRYLYSHYN, O., (1977). Towards a Definition of Non-Market Activities. *The Review of Income and Wealth*, No. 23(1), pp. 79-96.

HAWRYLYSHYN, O., (1976). The value of household services: a survey of empirical estimates. *The Review of Income and Wealth*, No. 22, pp. 101-131.

HOLLOWAY, S. & SHORT, S. & TAMPLIN, S., (2002). Household Satellite Account (experimental) methodology. UK: Office for National Statistics. http://www.ons.gov.uk/ons/guidemethod/method-quality/specific/social-and-welfare-methodology/household-satelliteaccount/index.html

IRONMONGER, D. & SOUPOURMAS, F., (2009). Estimating household production outputs with time use episode data. eIJTUR, No. 6(2), pp. 240-268.

MARSZAŁEK, M., (2018). National Time Transfer Accounts in Poland (Narodowe Rachunki Transferów Czasu (NTTA) w Polsce), in: Międzypokoleniowe relacje z perspektywy ekonomicznej. Narodowe Rachunki Transferów i Narodowe Rachunki Transferów Czasu w Polsce. Raport opracowany w ramach projektu Narodowe Rachunki Transferów i Narodowe Rachunki Transferów Czasu dla Polski, SGH, Warszawa (in Polish).

MARSZAŁEK, M., (2016). The rhythm of activities in the time use survey (Rytm zajęć w budżecie czasu), in: GUS, Time Use Survey 2013, Part II, ZWS, Warszawa, pp. 203–231 (in Polish).

MARSZAŁEK, M., (2015). Household satellite account in Poland in the concept of the system of social statistics (Rachunek produkcji domowej w Polsce w koncepcji system statystyki społecznej), Oficyna Wydawnicza SGH, Warszawa (in Polish).

MASON, A., LEE, R., (2011). Population aging and the generational economy: key findings in Ronald Lee and Andrew Mason, eds., Population Aging and the Generational Economy: A Global Perspective. Cheltenham, Edward Elgar, United Kingdom and Northampton, Massachusetts, pp. 3–31.

MIKUTA, B., (1998). Wycena wartości pracy domowej na wsi, Ekonomika i Organizacja Gospodarki Żywnościowej, *Zeszyty Naukowe SGGW*, No. 32, Warszawa, pp. 151–159 (in Polish).

NIEWIEROWSKA, K., (1997). Próba oszacowania wartości pracy domowej na przykładzie gospodarstw domowych rolników zlokalizowanych na terenie gminy Drohiczyn, woj. białostocki, SGGW, Warszawa (in Polish).

STIGLITZ J., SEN A., FITOUSSI J. P., (2009). Report by the Commission in the Measurement of economic performance and social progress, http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf (accessed: 17.10.2018).

SZCZERBIŃSKA, L., (1987). Analiza zmian w strukturze budżetu czasu oraz wartości pracy gospodarstw domowych w latach 1976 i 1984, Z Prac ZBSE, No. 168, GUS, Warszawa (in Polish).

SZCZERBIŃSKA, L., (1986). Wartość pracy gospodarstw domowych w Polsce w 1983 r., Z Prac ZBSE, No. 156, GUS, Warszawa (in Polish).

UNITED NATIONS, (2013). National Transfer Accounts Manual. Measuring and Analysing the Generational Economy, New York.

VARJONEN, J., HAMUNEN, E., SOINNE K., (2014). Satellite Accounts on Household Production: Eurostat Methodology and Experiences to Apply It, Working Papers 1/2014, http://www.iariw.org (accessed: 15.10.2019)

VARJONEN, J., AALTO, K., (2006). Household Production and Consumption in Finland 2001. Household Satellite Account, Statistics Finland, Helsinki.

VARJONEN, J., RÜGER, Y., (2008). Value of Household Production in Finland and Germany. Analysis and Recalculation of the Household Satellite Account System in both countries. Working Papers 112/2008, National Consumer Research Centre, Helsinki.

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 3 pp. 171–184, DOI 10.21307/stattrans-2020-050
Received – 20.04.2020; accepted – 16.05.2020

Poisson Weighted Ishita Distribution: Model for Analysis of Over-Dispersed Medical Count Data

Bilal Ahmad Para[1], **Tariq Rashid Jan**[2]

ABSTRACT

A new over-dispersed discrete probability model is introduced, by compounding the Poisson distribution with the weighted Ishita distribution. The statistical properties of the newly introduced distribution have been derived and discussed. Parameter estimation has been done with the application of the maximum likelihood method of estimation, followed by the Monte Carlo simulation procedure to examine the suitability of the ML estimators. In order to verify the applicability of the proposed distribution, a real-life set of data from the medical field has been analysed for modeling a count dataset representing epileptic seizure counts.

Key words: compounding model, coverage probability, simulation, count data, epileptic seizure counts.

1. Introduction

Compounding mechanism for generating new count data probability models has received a great attention from researchers to obtain new probability distributions to fit data sets not adequately fit by common parametric distributions. Compound distributions serve well to describe various phenomena in biology, epidemiology and so on. The work has been done in this particular area since 1920. Using compounding mechanism, Greenwood and Yule (1920) established a relationship between Poisson distribution and a negative binomial distribution by treating the rate parameter in Poisson model as gamma variate. Skellam (1948) proposed a probability distribution from the binomial distribution by regarding the probability of success as a beta variable between sets of trials. Lindely (1958) proposed a one parameter probability distribution to illustrate the difference between fiducial distribution and posterior distribution. Gerstenkorn (1993,1996) introduced several compound distributions and obtained compound of gamma distribution with exponential distribution by treating the

[1] Department of Statistics, GDC Anantnag, J&K, India. E-mail: parabilal@gmail.com.
ORCID: http://orcid.org/0000-0002-0077-3391.
[2] Department of Statistics, University of Kashmir. India. E-mail: drtrjan@gmail.com.
ORCID: https://orcid.org/0000-0002-0093-0748

parameter of gamma distribution as an exponential variate and also obtained compound of polya with beta distribution. Mahmoudi et al. (2010) generalized the Poisson-Lindely distribution of Sankaran (1970) and showed that their generalized distribution has more flexibility in analysing count data. Zamani and Ismail (2010) proposed a new compound distribution by compounding negative binomial with one parameter Lindley distribution that provides good fit for count data where the probability at zero has an inflated value. A new generalized negative binomial distribution was proposed by Gupta and Ong (2004). This distribution arises from Poisson distribution if the rate parameter follows generalized gamma distribution; the resulting distribution so obtained was applied to various data sets and can be used as a better alternative to negative binomial distribution. Rashid, Ahmad and Jan (2016) proposed a new competitive count data model, by compounding negative binomial distribution with Kumaraswamy distribution, which finds its application in biological sciences. Para and Jan (2018) introduced two compounding models with applications to handle count data in medical sciences.

In this paper, we propose a new compounding distribution by compounding Poisson distribution with weighted Ishita distribution. Ishita distribution is a flexible probability model introduced by Shanker and Shukla (2017) and its weighted version was introduced by Shukla and Shanker (2019) as a new life time probability model. The new model is introduced as there is a need to find more flexible models for analyzing over-dispersed count data.

2. Definition of Proposed Model (Poisson Weighted Ishita Distribution)

If $X|\lambda \sim$ Poisson(λ), where λ is itself a random variable following weighted Ishita distribution with parameter c and θ, then determining the distribution that results from marginalizing over λ will be known as a compound of Poisson distribution with that of weighted Ishita distribution, which is denoted by $PWID(X;c,\theta)$. It may be noted that the proposed model will be a discrete since the parent distribution is discrete.

Theorem 2.1: The probability mass function of a Poisson weighted Ishita distribution, i.e. $PWID(X;c,\theta)$ is given by

$$P(X=x)=\frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]$$

$$x=0,1,2,3,\ldots \;;\; \theta>0, c>0$$

Proof: Using the definition (2), the pmf of a Poisson weighted Ishita distribution, i.e. $PWID(X;c,\theta)$ can be obtained as

$$g(x\mid\lambda)=\frac{e^{-\lambda}\lambda^x}{x!} \;;\; x=0,1,2,3,\ldots;\lambda>0$$

When its parameter λ follows weighted Ishita distribution (WID) with pdf

$$h(\lambda;c,\theta)=\frac{\lambda^{c}\theta^{(c+3)}(\theta+\lambda^{2})e^{-\theta\lambda}}{c!(\theta^{3}+(c+1)(c+2))},\quad ;\lambda>0,c>0,\theta>0$$

We have

$$P(X=x)=\int_{0}^{\infty}g(x\mid\lambda).h(\lambda;\theta)d\lambda$$

$$P(X=x)=\frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^{3}+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^{2}+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right] \tag{2.1}$$

$$x=0,1,2,3,\ldots\ ;\ \theta>0,c>0$$

which is the pmf of Poisson weighted Ishita distribution.

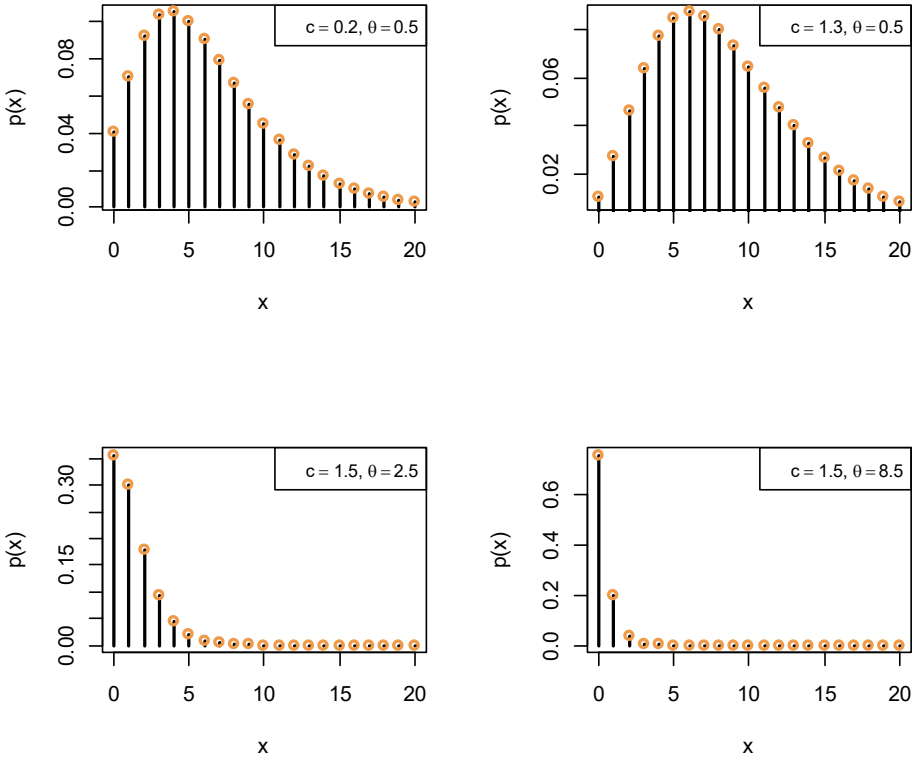


Figure 1. pmf plot of Poisson weighted Ishita distribution for different parameter combinations

The corresponding cdf of Poisson weighted Ishita distribution is obtained as:

$$F_X(x) = \sum_{x=0}^{x} \frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right] \tag{2.2}$$

$$x = 0,1,2,3,\ldots \;;\; \theta > 0, c > 0$$

Cdf is not in the closed form and it can be solved using software like mathematica and MathCAD for getting numerical results.

2.1. Random Data Generation from Poisson weighted Ishita distribution

In order to simulate the data from Poisson weighted Ishita distribution, we employ the discrete version of inverse cdf method. Simulating a sequence of random numbers $y_1, y_2, \ldots, y_n$ from Poisson weighted Ishita random variable K with pmf $p(K = y_i) = p_i, \sum_{i=0}^{x} p_i = 1$ and a cdf $F(K;c,\theta)$, where x may be finite or infinite can be described as in the following steps:

Step1: Generate a random number *u* from uniform distribution *U(0,1).*

Step2: Generate random number y_i based on

$$if\ u \le p_0 = F(y_0;c,\theta)\ then\ K = y_0$$
$$if\ p_0 < u \le p_0 + p_1 = F(y_1;c,\theta)\ then\ K = y_1$$
.
.
$$if \sum_{j=0}^{x-1} p_j < u \le \sum_{j=0}^{x} p_j = F(y_x;c,\theta)\ then\ K = y_x$$

In order to generate n random numbers $y_1, y_2, \ldots, y_n$ from Poisson weighted Ishita distribution, repeat step-1 and step-2 *n* times. We have employed R studio software for running the simulation study of the proposed model.

3. Statistical properties

In this section, structural properties of the Poisson weighted Ishita model have been evaluated. These include the moment, moment generating function and probability generating function.

3.1. Factorial Moments

Using (2.1), the r[th] factorial moment about origin of the Poisson weighted Ishita distribution (2.1) can be obtained as

$$\mu_{(r)}{}' = E[E(X^{(r)} \mid \lambda)], \ \ where\, X^{(r)} = X(X-1)(X-2)...(X-r+1)$$

$$\mu_{(r)}{}' = \int_0^\infty \left[\sum_{x=0}^{\infty} x^{(r)} \frac{e^{-\lambda}\lambda^x}{(x)!} \right] . \frac{\lambda^c \theta^{(c+3)} (\theta + \lambda^2) e^{-\theta\lambda}}{c!(\theta^3 + (c+1)(c+2))} d\lambda$$

$$\mu_{(r)}{}' = \frac{\theta^{(c+3)}}{c!(\theta^3 + (c+1)(c+2))} \int_0^\infty \left[\lambda^r \left(\sum_{x=r}^{\infty} \frac{e^{-\lambda}\lambda^{x-r}}{(x-r)!} \right) \right] \lambda^c \theta^{(c+3)} (\theta + \lambda^2) e^{-\theta\lambda} d\lambda$$

Taking $u = x - r$, we get

$$\mu_{(r)}{}' = \frac{(r+c)!}{c!(\theta^3 + (c+1)(c+2))} \left(\frac{\theta^3 + (r+c+2)(r+c+1)}{\theta^r} \right) \tag{3.1.1}$$

Taking r=1,2,3,4 in (3.1.1), the first four factorial moments about origin of Poisson weighted Ishita distribution can be obtained as

$$\mu_{(1)}{}' = \frac{(c+1)}{(\theta^3 + (c+1)(c+2))} \left(\frac{\theta^3 + (c+3)(c+2)}{\theta} \right)$$

$$\mu_{(2)}{}' = \frac{(c+2)(c+1)}{(\theta^3 + (c+1)(c+2))} \left(\frac{\theta^3 + (c+4)(c+3)}{\theta^2} \right)$$

$$\mu_{(3)}{}' = \frac{(c+3)(c+2)(c+1)}{(\theta^3 + (c+1)(c+2))} \left(\frac{\theta^3 + (c+5)(c+4)}{\theta^3} \right)$$

$$\mu_{(4)}{}' = \frac{(c+4)(c+3)(c+2)(c+1)}{(\theta^3 + (c+1)(c+2))} \left(\frac{\theta^3 + (c+6)(c+5)}{\theta^{4r}} \right)$$

3.1.2. Moments about origin (Raw moments)

Using the relationship between factorial moments about origin and the moments about origin of Poisson weighted Ishita distribution (2.1), we have

$$\mu_1{}' = \frac{(c+1)}{(\theta^3 + (c+1)(c+2))} \left(\frac{\theta^3 + (c+3)(c+2)}{\theta} \right)$$

$$\mu_2' = \mu_{(2)}' + \mu_1' = \frac{(c+2)(c+1)}{(\theta^3+(c+1)(c+2))}\left(\frac{\theta^3+(c+4)(c+3)}{\theta^2}\right)+\frac{(c+1)}{(\theta^3+(c+1)(c+2))}\left(\frac{\theta^3+(c+3)(c+2)}{\theta}\right)$$

$$= \frac{(c+1)}{\theta(\theta^3+(c+1)(c+2))}\left[(c+2)\left(\theta^3+(c+4)(c+3)\right)+\theta\left(\theta^3+(c+3)(c+2)\right)\right]$$

4. Reliability Analysis

In this section, we have obtained the reliability and hazard rate function of the proposed Poisson weighted Ishita distribution.

4.1. Reliability Function R(x)

The reliability function is defined as the probability that a system survives beyond a specified time. It is also referred to as survival function of the distribution. The reliability function or the survival function of Poisson weighted Ishita distribution is given by

$$R(x,\theta) = 1-\sum_{x=0}^{x}\frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]$$

4.2. Hazard Function

The hazard function is also known as the hazard rate, instantaneous failure rate or force of mortality, and is given as:

$$\text{H.R}=\text{h}(x,\theta) = \frac{f(x,\theta)}{R(x,\theta)}$$

$$= \frac{(x+c)!\theta^{c+3}}{\left(1-\sum_{x=0}^{x}\frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]\right)x!c!\,(\theta^3+(c+1)(c+2))\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]}$$

5. Order statistics

Let $X_{(1)}, X_{(2)}, X_{(3)}\ldots., X_{(n)}$ be the ordered statistics of the random sample $X_1, X_2, X_3, \ldots. X_n$ drawn from the discrete distribution with cumulative distribution

function $F_X(x)$ and probability mass function $P_X(x)$, then the probability mass function of r[th] order statistics $X_{(r)}$ is given by:

$$f_{x(r)}(x,c,\theta)=\frac{n!}{(r-1)!(n-r)!}P(x)[F(x)]^{r-1}[1-F(x)]^{n-r}.\ r=1,\ 2,\ 3,\ldots,n$$

Using the equations (2.1) and (2.2), the probability mass function of rth order statistics of Poisson weighted Ishita distribution is given by:

$$f_{(r)}(x,,\theta)=\frac{n!}{(r-1)!(n-r)!}\frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]$$
$$\left[\sum_{x=0}^{x}\frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]\right]^{r-1}$$
$$\left[1-\sum_{x=0}^{x}\frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]\right]^{n-r}$$

Then, the pmf of first order $X_{(1)}$ Poisson weighted Ishita distribution is given by:

$$f_1(x,\theta)=n\frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]$$
$$\left[1-\sum_{x=0}^{x}\frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]\right]^{n-1}.$$

and the pmf of nth order $X_{(n)}$ Poisson Ishita model is given as:

$$f_{w(n)}(x,\theta)=n\frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]$$
$$\left[\sum_{x=0}^{x}\frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]\right]^{n-1}$$

6. Estimation of Parameters

In this section, we estimate the parameters of the Poisson weighted Ishita distribution using methods of maximum likelihood estimation.

6.1. Method of Maximum Likelihood Estimation

This is one of the most useful method for estimating the different parameters of the distribution. Let $X_1, X_2, X_3,\ldots,X_n$ be the random size of sample n draw from Poisson

weighted Ishita distribution. Then, the likelihood function of Poisson weighted Ishita distribution is given as:

$$L(x \mid \theta) = \prod_{i=1}^{n} \left(\frac{(x+c)!\theta^{c+3}}{x!c!\ (\theta^3 + (c+1)(c+2))} \left[\frac{\theta(1+\theta)^2 + (x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}} \right] \right)$$

$$\log L = \sum_{i=1}^{n} \log(x_i + c)! + n(c+3)\log(\theta) - \sum_{i=1}^{n} \log(x_i!) - n\log(c!) - n\log(\theta^3 + (c+1)(c+2)) + \sum_{i=1}^{n} \log(\theta(1+\theta)^2 + (x_i + c + 1)(x_i + c + 2)) - \sum_{i=1}^{n} (x_i + c + 3)\log(1+\theta)$$

By differentiating log-likelihood function with respect to c and θ, and equating them to zero we get normal equations for estimating the parameters of the Poisson weighted Ishita distribution.

$$\frac{\partial}{\partial c}\log L = \frac{\partial}{\partial c}\left(\sum_{i=1}^{n} \log(x_i + c)! + n(c+3)\log(\theta) - \sum_{i=1}^{n} \log(x_i!) - n\log(c!) - n\log(\theta^3 + (c+1)(c+2)) + \sum_{i=1}^{n} \log(\theta(1+\theta)^2 + (x_i + c + 1)(x_i + c + 2)) - \sum_{i=1}^{n} (x_i + c + 3)\log(1+\theta) \right) = 0$$

$$\frac{\partial}{\partial \theta}\log L = \frac{\partial}{\partial \theta}\left(\sum_{i=1}^{n} \log(x_i + c)! + n(c+3)\log(\theta) - \sum_{i=1}^{n} \log(x_i!) - n\log(c!) - n\log(\theta^3 + (c+1)(c+2)) + \sum_{i=1}^{n} \log(\theta(1+\theta)^2 + (x_i + c + 1)(x_i + c + 2)) - \sum_{i=1}^{n} (x_i + c + 3)\log(1+\theta) \right) = 0$$

These two derivative equations cannot be solved analytically, therefore $\hat{c}$ *and* $\hat{\theta}$ will be obtained by maximizing the log likelihood function numerically using the Newton-Raphson method, which is a powerful technique for solving equations iteratively and numerically.

6.2. Monte Carlo Simulation

In order to investigate the performance of the maximum likelihood estimators for a finite sample size n using Monte Carlo simulation procedure. Using the inverse cdf method discussed in sub-section 2.1, random data is generated from Poisson weighted Ishita distribution. We took four random parameter combinations as $c = 0.4, \theta = 0.4$, $c = 0.8, \theta = 0.9, c = 1.5, \theta = 1.7$ and $c = 2.5, \theta = 3.2$, to carry out the simulation study and the process was repeated 1000 times by going from small to large sample sizes n = (10,25,75,200,300,600). From Table 1, it is clear that the estimated variances and MSEs

decrease when the sample size n increases. The coverage probabilities (CP) are near to 0.95 when the sample size increases. Thus, the agreement between theory and practice improves as the sample size n increases. Hence, the maximum likelihood method performs quite well in estimating the model parameters of the Poisson weighted Ishita distribution.

Table 1. Simulation study of ML estimators of Poisson weighted Ishita distribution

Sample size (n)	Parameters	$c = 0.4, \theta = 0.4$				$c = 0.8, \theta = 0.9$			
		Bias	Variance	MSE	Coverage Probability	Bias	Variance	MSE	Coverage Probability
10	c	-0.0948	0.008298	0.017285	0.899	0.054353	0.016773	0.0197272	0.919
	θ	0.214834	0.091871	0.1380246	0.922	0.033567	0.039843	0.0409697	0.911
25	c	-0.07802	0.005691	0.0117781	0.939	-0.00876	0.005471	0.0055477	0.929
	θ	0.078574	0.043124	0.0492979	0.943	0.037584	0.010208	0.0116206	0.922
75	c	-0.06463	0.003127	0.007304	0.949	0.016155	0.000861	0.001122	0.942
	θ	-0.05445	0.031839	0.0348038	0.953	0.011167	0.000993	0.0011177	0.949
200	c	-0.04363	0.003372	0.0052756	0.959	-0.00825	0.001847	0.0019151	0.956
	θ	-0.01411	0.007967	0.0081661	0.957	0.007271	0.001054	0.0011069	0.955
300	c	-0.02236	0.001644	0.002144	0.958	0.001713	0.000411	0.0004139	0.959
	θ	-0.00354	0.004377	0.0043895	0.961	0.006739	0.000308	0.0003534	0.966
600	c	-0.01746	0.000182	0.0004869	0.959	0.006854	0.000149	0.000196	0.963
	θ	-0.00046	0.002975	0.0029752	0.971	0.002234	0.000175	0.000180	0.969
Sample size (n)	Parameters	$c = 1.5, \theta = 1.7$				$c = 2.5, \theta = 3.2$			
		Bias	Variance	MSE	Coverage Probability	Bias	Variance	MSE	Coverage Probability
10	c	-0.050778	0.000497	0.003075	0.939	0.054253	0.016673	0.054253	0.829
	θ	0.040750	0.003891	0.005552	0.889	0.033467	0.039743	0.033467	0.915
25	c	-0.044688	0.000719	0.002716	0.938	-0.00886	0.005371	-0.00886	0.927
	θ	-0.016717	0.002150	0.002429	0.952	0.037484	0.010108	0.037484	0.943
75	c	-0.032848	0.000382	0.001461	0.956	0.016055	0.000761	0.016055	0.949
	θ	0.000015	0.000310	0.000310	0.953	0.011067	0.000893	0.011067	0.943
200	c	0.003141	0.000628	0.000638	0.962	-0.00835	0.001747	-0.00835	0.959
	θ	0.003232	0.000001	0.000011	0.958	0.007171	0.000954	0.007171	0.959
300	c	-0.005717	0.000003	0.000036	0.961	0.001613	0.000311	0.001613	0.962
	θ	-0.001419	0.000012	0.000014	0.959	0.006639	0.000208	0.006639	0.961
600	c	0.002955	0.000040	0.000049	0.965	0.006754	0.000049	0.006754	0.972
	θ	0.000943	0.000036	0.000037	0.962	0.002134	0.000075	0.002134	0.968

7. Applications of Poisson Weighted Ishita Distribution

In this section, we fit our proposed model and other related models to a vaccine adverse event count data studied by Rose et al. (2006). The data are the frequencies which correspond to 4020 observed systemic adverse events for four injections for each of the 1005 study participants. The data set is given in Table 2.

Table 2. Data set representing vaccine adverse event count data studied by Rose et al. (2006)

Vaccine adverse event	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	1437	1010	660	428	236	122	62	34	14	8	4	4	1

Maximum likelihood estimation method is used in estimating the parameters for all the suggested models using R software. Parameter estimates with standard errors in parenthesis for each fitted model are given in Table 3.

Table 3. Estimated Parameters by ML method for fitted distributions for data set representing epileptic seizure counts

Distribution	**Parameter Estimates (Standard Error)**	**Model function**
Poisson Weighted Ishita	$\hat{c} = 0.33\ (0.08)$ $\theta = 1.51(0.06)$	$P(X=x) = \frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))} \left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]$ $x = 0,1,2,3,\ldots \;;\; \theta > 0, c > 0$
Poisson Ishita	$\theta = 1.29(0.01)$	$P(X=x) = \frac{\theta^3}{\theta^3+2}\left[\frac{\theta(1+\theta)^2+(x+1)(x+2)}{(1+\theta)^{x+3}}\right]$ $x = 0,1,2,3,\ldots, \theta > 0$
Poisson	$\lambda = 1.50(0.01)$	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \lambda > 0,\ x = 0,1,2,\ldots$
Poisson Lindley	$\theta = 0.99\ (0.016)$	$p(x) = \frac{\theta^2(x+\theta+2)}{(\theta+1)^{x+3}}\ x = 0,1,2,\ldots, \theta > 0$
Geometric	$p = 0.398\ (0.004)$	$p(x) = q^x p \quad 0 < q < 1, q = 1\text{-}p, x = 0,1,2,\ldots$
Negative Binomial	$p = 0.50\ (0.013)$ $r = 1.53\ (0.08)$	$p(x) = \binom{x+r-1}{x} p^r q^x,\ x = 0,1,2,\ldots$ $r > 0,\ 0 < p < 1$

Zero Inflated Poisson	$\alpha = 0.26\ (0.009)$ $\lambda = 2.04\ (0.031)$	$p(x) = \begin{cases} \alpha + (1-\alpha)\dfrac{e^{-\lambda}\lambda^x}{x!}, \lambda > 0, x = 0 \\ (1-\alpha)\dfrac{e^{-\lambda}\lambda^x}{x!}, \lambda > 0, x = 0,1,2,... \end{cases}$ $0 < \alpha < 1,\ \lambda > 0$
Discrete Weibull	$q = 0.65\ (0.007)$ $\gamma = 1.15\ (0.017)$	$p(x) = q^{x^\gamma} - q^{(x+1)^\gamma}, \quad x = 0,1,2,...$ $0 < q < 1,\ \gamma > 0$

We compute the expected frequencies for fitting Poisson weighted Ishita, Poisson Ishita, Poisson, Geometric, Negative Binomial, Zero Inflated Poisson, Poisson Lindley and discrete Weibull distributions with the help of R studio statistical software, and Pearson's chi-square test is applied to check the goodness of fit of the models discussed. The calculated expected frequencies for each fitted model are given in Table 4. For Poisson weighted Ishita, negative binomial and discrete Weibull distributions, p-value is >0.05, hence it fits the data statistically good. Poisson Ishita, Poisson, Geometric, zero inflated Poisson and Poisson Lindley does not fit the data at all as p-value in the case of these models is <0.05. Based on the chi-square, we observe that Poisson weighted Ishita distribution has the highest p-value (0.8162), which signifies that Poisson weighted Ishita provides a better fit for the data set representing vaccine adverse event count data studied by Rose et al. (2006) as compared to other fitted models.

Table 4. Fitted proposed distribution and other competing models to a data set representing epileptic seizure counts

Epileptic seizure (X)	Observed	Poisson Weighted Ishita	Poisson Ishita	Poisson	Geometric	Negative Binomial	Zero Inflated Poisson	Poisson Lindley	Discrete Weibull
0	1437	1427.4	1518.0	890.8	1603.5	1409.1	1437.0	1500.1	1410.7
1	1010	1035.2	965.9	1342.3	963.9	1068.7	787.3	1003.5	1065.4
2	660	665.7	620.3	1011.4	579.4	670.7	803.3	629.2	667.7
3	428	401.0	386.5	508.1	348.3	391.6	546.4	378.7	393.1
4	236	229.5	231.9	191.4	209.4	220.2	278.7	221.6	222.6
5	122	126.0	134.4	57.7	125.9	120.9	113.8	127.0	122.6
6	62	66.9	75.5	14.5	75.7	65.3	38.7	71.7	66.0
7	34	34.5	41.4	3.1	45.5	34.9	11.3	39.9	34.9
8	14	17.4	22.2	0.6	27.3	18.5	2.9	22.0	18.2
9	8	8.6	11.7	0.1	16.4	9.7	0.7	12.1	9.3
10	4	4.2	6.1	0.0	9.9	5.1	0.1	6.6	4.7
11	4	2.0	3.1	0.0	5.9	2.6	0.0	3.5	2.4
12	1	1.8	3.1	0.0	9.0	2.8	0.0	4.1	2.3
P-value		0.8162	0.0036	0.0003	0.0001	0.2619	<0.0001	0.0322	0.3564

Furthermore, in order to compare our proposed distribution and other competing models, we consider the criteria like AIC (Akaike information criterion), AICC (corrected Akaike information criterion) and BIC (Bayesian information criterion). The better distribution corresponds to lesser AIC, AICC and BIC values. From Table 5, it is observed that the Poisson weighted Ishita distribution has lesser AIC, AICC and BIC values as compared to other competing models. Hence, we can conclude that the Poisson weighted Ishita distribution leads to a better fit than the other competing models for analysing the data set given in Table 2.

Table 5. Model comparison criterion for fitted models to a data set

Criterion	Poisson Weighted Ishita	Poisson Ishita	Poisson	Geometric	Negative Binomial	Zero Inflated Poisson	Poisson Lindley	Discrete Weibull
-logL	6737.2	6747.5	7231.1	6778.0	6740.6	6868.8	6746.0	6739.7
AIC	13478.4	13496.9	14464.3	13558.1	13485.2	13741.6	13494.0	13483.4
BIC	13491.0	13503.2	14470.6	13564.4	13497.8	13754.2	13500.3	13496.0

We also use Likelihood Ratio (LR) test to check whether the fitted Poisson weighted Ishita distribution for a given data set is statistically "superior" to the fitted Poisson Ishita distribution. In any case, hypothesis tests of the type $H_0 : \Theta = \Theta_0$ versus $H_1 : \Theta \neq \Theta_0$ can be performed using LR statistics. In this case, the LR statistic for testing H_0 versus H_1 is $\omega = 2(L(\hat{\Theta}) - L(\hat{\Theta}_0))$ where $\hat{\Theta}$ and $\hat{\Theta}_0$ are the MLEs under H_1 and H_0. The statistic ω is asymptotically $(as\, n \to \infty)$ distributed as χ_k^2, with k degrees of freedom, which is equal to the difference in dimensionality of $\hat{\Theta}$ and $\hat{\Theta}_0$. H_0 will be rejected if the LR-test p-value is <0.05 at 95% confidence level.

Table 6. Likelihood Ratio test of Poisson weighted Ishita distribution versus Poisson Ishita distribution

Model	Model Function	-logl	Likelihood Ratio Statistic
Poisson weighted Ishita	$P(X=x) = \frac{(x+c)!\theta^{c+3}}{x!c!\,(\theta^3+(c+1)(c+2))}\left[\frac{\theta(1+\theta)^2+(x+c+1)(x+c+2)}{(1+\theta)^{x+c+3}}\right]$ $x = 0,1,2,3,\ldots,\ \theta > 0, c > 0$	6737.2	20.60
Poisson Ishita	$P(X=x) = \frac{\theta^3}{\theta^3+2}\left[\frac{\theta(1+\theta)^2+(x+1)(x+2)}{(1+\theta)^{x+3}}\right]$ $x = 0,1,2,3,\ldots,\ \theta > 0$	6747.5	

We have $\chi_1^2 = 6.35 <$ Likelihood Ratio Statistic (20.60), thus the null hypothesis is rejected and it is concluded that parameter c is playing a significant role in Poisson weighted Ishita distribution for analysing the data set given in Table 2.

8. Conclusion

A new over-dispersed probability distribution is introduced using the compounding technique. Statistical properties of the proposed model are studied and application in handling count data set representing epileptic seizure counts is analyzed.

REFERENCES

ADIL, R., ZAHOOR, A., JAN, T. R., (2016). A new count data model with application in genetics and ecology. *Electronic Journal of Applied Statistical Analysis,* 9(1), pp. 213–226.

GERSTENKORN, T., (1993). A compound of the generalized gamma distribution with the exponential one, *Recherches surles deformations,* 16(1), pp. 5–10.

GERSTENKORN, T., (1996). A compound of the Polya distribution with the beta one, *Random Operators and Stochastic Equations*, 4(2), pp. 103–110.

GREENWOOD, M., YULE, G. U., (1920). An inquiry into the nature of frequency distribution representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of Royal Statistical Society*, 83, pp. 255–279.

GUPTA, R. C., ONG, S. H., (2004). A new generalization of the negative binomial distribution, *Journal of Computational Statistics and Data Analysis,* 45, pp. 287–300.

LINDELY, D. V., (1958). Fiducial Distributions and Bayes theorem. *Journal of the Royal Statistical Society*, 20(1), pp. 120-107.

MAHMOUDI E., ZAKERZADEH H., (2010). Generalized Poisson-Lindely Distribution. *Communications in statistics – Theory and Methods,* 39(10*),* pp. 1785–1798.

PARA, B. A., JAN, T. R., (2018). An Advanced Discrete Model with Applications in Medical Science, *Journal of Multiscale Modelling*, 9(1), DOI: 10.1142/S1756973718500014, ISSN: 1756-9737 (print).

PARA, B. A., JAN, T. R., (2018). Discrete Inverse Weibull Beta Model: Properties And Applications In Health Science. *Pakistan Journal of Statistics,* 34(3), pp. 229–349.

PLACKETT, R. L., (1953). The truncated Poisson distribution. *Biometrics*, 9(4), pp. 485–488.

R CORE TEAM, (2019). R version 3.5.3: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

ROSE, C. E., MARTIN, S. W., WANNEMUEHLER, K. A., PLIKAYTIS, B. D., (2006), On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data, *Journal of Biopharmaceutical Statistics,* 16, pp. 463–481.

SHANKER, R., SHUKLA, K. K., (2017). Ishita distribution and its Applications, *Biometrics & Biostatistics International Journal*, 5(2), pp. 1–9.

SHUKLA, K. K., SHANKER, R., (2019). Weighted Ishita Distribution and Its Application to Survival Data, *International Journal of Mathematics and Statistics*, 20(1), pp. 67–82.

SKELLER, J. G., (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials, *Journal of the Royal Statistical Society, Series B,* 10, pp. 257–261.

ZAMANI, H., ISMAIL, N., (2010). Negative Binomial–Lindley Distribution And Its Application. *Journal of Mathematics and Statistics*, 1, pp. 4–9.

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 1, pp. 185–193, DOI 10.21307/stattrans-2020-051
Received – 15.06.2016; accepted – 06.05.2020

Through a Random Route to the Goal: Theoretical Background and Application of the Method in Tourism Surveying in Poland

Sebastian Wójcik [1]

ABSTRACT

Classic survey methods are ineffective when surveying a small or rare population. Several methods have been developed to address this issue, but often without providing a full mathematical justification. In this paper we propose estimators of parameters relating to Random Route Sampling and explore their basic properties. A formula for the Horvitz-Thompson estimator weights is presented. Finally, a case of a tourism-related survey conducted in Poland is discussed.

Key words: random route, Horvitz-Thompson estimator.

1. Introduction

Nowadays official statistics is looking for cost-effective and time-effective survey methods. It is particularly noticeable when we deal with surveys of small populations such as unemployed, foreigners, homeless, etc. Usually a frame for such a subpopulation is not available. Some methods for solving these problems have been developed, but often without a full theoretical background. The representativeness and unbiasedness of the sample surveyed in that way is a question of concern.

2. Random Route Sampling

We shall present some details on the Random Route Sampling method. Assume that we want to survey a subpopulation S of a population P. The frame of members of P is available, but the frame of members of S is unknown. In this paper we focus on household (or dwelling) population. In the Random Route procedure, interviewers walk from house to house and survey households on a prescribed route that ensures randomness. At the first stage, a group of n households (list of starting points) is sampled. At the second stage, an interviewer sets out from the starting point. If a household is not a member of S, then the interviewer continues walking and surveying. There are two alternative models. In the first one the interviewer is surveying until he/she finds a member of S. In the second model the interviewer continues until he/she finds a member of S or he/she reaches the limit of K steps. Each visited dwelling is called *a step*. The interviewer follows some rules that ensure randomness such as: always on the right, always clockwise or always downstairs.

In this paper we unify two aforementioned models by assuming that:

[1] Institute of Mathematics, University of Rzeszow, Division of Mathematical Statistics, Statistical Office in Rzeszów, Poland. E-mail: s.wojcik@stat.gov.pl. ORCID: https://orcid.org/0000-0003-2425-9626.

- the interviewer makes up to $K = 1, 2, ..., \infty$ steps until he/she finds a member of S. The starting point is the first step. The sequence of up to K steps will be called *a route*;
- the interviewer walks only within his/her district (Primary Statistical Unit);
- there is only one interviewer per district;
- the interviewer does not survey the household that has been already surveyed;
- if in a route interviewer surveys another starting point from the list then he/she counts it as a step. The next household to be surveyed after the route is completed, becomes a new (replaced) starting point.

Clearly, the better survey completeness and the larger size of the subpopulation S in a relation to the size of the population P, the lower number of steps made in a route. Obviously, if $K = 1$ then the Random Route (RR) becomes the Simple Random Sampling (SRS). Thus, SRS can be treated as a special case of RR and all of the results for RR for $K = 1$ should be consistent with results for SRS.

Several papers assess the quality of random route samples. Biasedness and sample representativeness are studied based on case studies (Hoffmeyer-Zlotnik (2003), de Rada, Martin (2014)) or simulations (Bauer (2014), Bauer (2016)).
Hoffmeyer-Zlotnik compared three different models of Random Route Sampling:

1) uncontrolled Random Route Sampling with Kish tables and net number of interviews defined used in German General Social Survey (ALLBUS) in 1992,
2) controlled Random Route Sampling with Kish tables and gross number of addresses defined used in German General Social Survey (ALLBUS) in 1998,
3) Random Route plus quota design with net number of interviews defined used in a national survey of the German Youth Institute.

Hoffmeyer-Zlotnik, based on data analysis, found out that the uncontrolled Random Route Sampling saves 30% of expenses in comparison to the controlled version and to the version with quota. In the extreme, case the version with quota caused the walk to be very long and the interviewer had to contact about 100 households for carrying out 10 interviews. Moreover, with modification of the sampling process in the controlled version or the version with quota, the Random Route Sampling becomes non-probability sampling and the sampling error cannot be calculated.

In this paper we refer to the uncontrolled Random Route Sampling. We propose an estimator of the fraction of a subpopulation S in a population P and prove that this estimator is asymptotically unbiased and consistent. Further, we derive sample weights.

3. Parameter estimation under the Random Route Sampling

Since the Random Route method is focused on surveying members of S the question that arises naturally is how to estimate the size of a population S. We will derive some estimators

and check their consistency and unbiasedness. Let us denote by M the size of population - the only known parameter. We introduce two further parameters r and p - the unknown parameters of the level of completeness and the fraction of a subpopulation S in a population P, respectively. We assume that r and p are such that prM and pM are integers.

Now, we divide our analysis into two cases.

3.1. Unlimited number of steps in a single route

Let X be a number of visited $P\backslash S$ members (members of P but not S) until an S member is surveyed. Clearly X is a random variable. Furthermore, for every $k \in \{0,1,...,(1-pr)M\}$, X takes the value k provided the following two conditions are satisfied:

- in each of the first k steps of the route the interviewer either visited a member of $P\backslash S$ or he/she did not get an answer due to the incompleteness;
- in the $(k+1)^{th}$ he/she surveyed a member of S.

Therefore,

$$P(X=k) = \frac{\binom{M-k-1}{(1-pr)M-k}}{\binom{M}{(1-pr)M}} \text{ for } k = 0,1,...,(1-pr)M. \tag{1}$$

That is X follows the negative hypergeometric distribution $HYP^{-}(M,(1-pr)M,1)$. In order to survey exactly n members of S the interviewer will make $n+\sum_{i=1}^{n} X_i$ steps, where X_i for $i = 1,..,n$ are i.i.d. random variables with probability distribution described by (1). Guenther (1975) proposed the following maximum-likelihood estimator of pr

$$Y^{\infty}_{HYP^{-}}(n) = \frac{n}{n+\sum_{i=1}^{n} X_i}. \tag{2}$$

The estimator $Y^{\infty}_{HYP^{-}}$ being the maximum-likelihood estimator has a number of attractive limiting properties such as consistency and efficiency (Pfanzagl (1994)). Nevertheless, the estimator given by (2) is biased (Zhang, Johnson (2011)). It is still an open question if this estimator is asymptotically unbiased. Therefore, we are going to modify the model in such a way that the estimator defined by (2) becomes asymptotically unbiased.

Assume that M is relatively large compared to n. According to the result of Johnson and Kotz (1969), if $M \to \infty$ with p and r being fixed, then

$$Y^{\infty}_{HYP^{-}}(n) \to_D Y$$

where Y is a random variable following the negative binomial distribution $BIN^{-}(1-pr,1)$. Therefore, for relatively large M, we can treat the Random Route Sampling as a sampling with replacement. Note, however, that

$$BIN^{-}(1-pr,1) = GEO(1-pr)$$

where $GEO(1-pr)$ denotes the geometric distribution with parameter $1-pr$. So, it is reasonable to replace the underlying negative hypergeometric distribution by the geometric

one. Then, we get

$$P\left(\sum_{i}^{n} X_i = k\right) = \binom{k+n-1}{k}(pr)^n(1-pr)^k \text{ for } k = 0,1,2,....$$

Furthermore, $\sum_i^n X_i$ being a sum of i.i.d. random variables having the geometric distribution with parameter $1-pr$ (cf. Bandyopadhyay P.S., Forster M.R. (2011)), follows the negative binomial distribution $BIN^{-}(1-pr,n)$. Hence,

$$E\left(\sum_{i}^{n} X_i\right) = \frac{n(1-pr)}{pr} \quad \text{and} \quad D^2\left(\sum_{i}^{n} X_i\right) = \frac{n(1-pr)}{(pr)^2}. \tag{3}$$

Define an estimator $Y^{\infty}_{BIN^-}$ as follows:

$$Y^{\infty}_{BIN^-}(n) = \frac{n}{n+\sum_{i=1}^{n} X_i}. \tag{4}$$

For every $n \in \mathbb{N}$, $Y^{\infty}_{BIN^-}(n)$ expresses the ratio of the number of surveyed S-members in a relation to the number of surveyed P-members in an n-route survey. Note that $Y^{\infty}_{BIN^-}$ is the maximum-likelihood estimator of pr (cf. Hilbe (2011)). Moreover, taking into account (3) and applying the result by Stuart (Stuart (1998), p. 351) we obtain

$$E\left(Y^{\infty}_{BIN^-}(n)\right) = E\left(\frac{n}{\sum_{i=1}^{n} X_i + n}\right) = \frac{n}{E\left(\sum_{i=1}^{n} X_i\right)+n} + O\left(\frac{1}{n}\right) =$$

$$\frac{n}{\frac{n(1-pr)}{pr}+n} + O\left(\frac{1}{n}\right) = pr + O\left(\frac{1}{n}\right).$$

Thus,

$$\lim_{n\to\infty} E\left(Y^{\infty}_{BIN^-}(n)\right) = pr, \tag{5}$$

which shows that $Y^{\infty}_{BIN^-}$ is asymptotically unbiased.

3.2. Limited number of steps in a single route

In practice, the most important case of the Random Route is when the number of steps is finite. In this approach, on the one hand surveying is less sensitive to clustering of $P\backslash S$-members, but on the other, we need more starting points to survey the same number of S-members. Moreover, in this setting it is possible that the interviewer will not survey a S-member in his/her route. Thus, the number of surveyed S-members is a random variable, taking two values: 0 and 1. Let us denote it by L^K. Then, assuming that M is relatively large comparing to n and treating the Random Route Sampling as the sampling with replacement, we obtain the following probability distribution of L^K

$$P\left(L^K = l\right) = \begin{cases} 1-(1-pr)^K & \text{for } l = 1, \\ (1-pr)^K & \text{for } l = 0. \end{cases}$$

We shall determine the maximum-likelihood estimator of pr. To this end, assume that $l = (l_1, ..., l_n)$ is a vector of observed data in a K-step route. Then, the log-likelihood function $\mathscr{L}(pr, l, K)$ is given by

$$\mathscr{L}(pr, l, K) = \sum_{i=1}^{n} \ln[(1-(1-pr)^K)l_i + (1-pr)^K(1-l_i)].$$

Derivating $\mathscr{L}(pr, l, K)$ with respect to pr, we obtain

$$\frac{\partial \mathscr{L}(pr, l, K)}{\partial pr} = \sum_{i=1}^{n} \frac{2l_i - 1}{(1-pr)^{1-K} l_i + (1-pr)(1-2l_i)}. \tag{6}$$

Hence, the maximum-likelihood estimator of pr is of the form

$$Z^K(n) = 1 - \left(1 - \frac{\sum_{i=1}^{n} L_i}{n}\right)^{\frac{1}{K}}. \tag{7}$$

Obviously, Z^K is consistent and efficient. It is an open question if Z^K is asymptotically unbiased.

Let X^K be the number of visited households until an S member is surveyed in a K-step route. Then X^K is a random variable taking the values $1, ..., K$. Furthermore, assuming as previously that M is relatively large compared to n and treating the Random Route Sampling as the sampling with replacement, we conclude that X^K has the following probability distribution:

$$P\left(X^K = k\right) = \begin{cases} pr(1-pr)^k & \text{for } k < K, \\ (1-pr)^K & \text{for } k = K. \end{cases} \tag{8}$$

Thus, we have

$$E\left(X^K\right) = \frac{(1-pr)(1-(1-pr)^K)}{pr} \tag{9}$$

and

$$D^2\left(X^K\right) = \tfrac{1}{(pr)^2}[(1-pr) - (2K-1)pr(1-pr)^K - (1-pr)^K(1-(1-pr)^K)p^2] +$$

$$\tfrac{1}{(pr)^2}[2pr(1-pr)(1-(1-pr)^K)^2 - (1-pr)^{2K}] -$$

$$\tfrac{1}{(pr)^2}[2pr(1-pr)(1-(1-pr)^K - K(1-pr)^{K-1} + K(1-pr)^K)]. \tag{10}$$

Note that

$$\lim_{K\to\infty} E\left(X^K\right) = \frac{1-pr}{pr} \quad \text{and} \quad \lim_{K\to\infty} E\left(L^K\right) = 1.$$

Hence, (9)-(10) are consistent with the case of $K = \infty$.

Now, we are going to determine the maximum-likelihood estimator of pr. Assume that

$x = (x_1, ..., x_n)$ is a vector of observed data in a K-step route. Let $l = (l_1, ..., l_n)$ where

$$l_i = \begin{cases} 0 & \text{whenever } x_i = K, \\ 1 & \text{whenever } x_i < K. \end{cases}$$

The log-likelihood function $\mathscr{L}(pr, x, K)$ is given by

$$\mathscr{L}(pr, x, K) = \sum_{i=1}^{n} \ln[pr(1-pr)^{k_i-1} l_i + (1-pr)^{K-1}(1-l_i)]. \tag{11}$$

Derivating $\mathscr{L}(pr, x, K)$ with respect to pr, we get

$$\frac{\partial \mathscr{L}(pr, x, K)}{\partial pr} =$$

$$\sum_{i=1}^{n} \frac{(l_i - 1)(K-1)(1-pr)^{K-2} + l_i((1-pr)^{x_i-1} - pr(x_i-1)(1-pr)^{k_i-2})}{l_i pr(1-pr)^{x_i-1} + (1-l_i)(1-pr)^{K-1}}.$$

Hence, the maximum-likelihood estimator of pr is of the form

$$Y^K(n) = \frac{\sum_{i=1}^{n} L_i^K}{\sum_{i=1}^{n} L_i^K X_i^K + K(n - \sum_{i=1}^{n} L_i^K) + \sum_{i=1}^{n} L_i^K}. \tag{12}$$

Note that for $i = 1, .., n$ we have

$$L_i^K \to_D 1 \text{ with } K \to \infty.$$

Moreover, if X_i follows $GEO(1-pr)$ for $i = 1, .., n$, then taking into account (8), we get

$$X_i^K \to_D X_i \text{ with } K \to \infty.$$

Thus, in view of (12), for every n, we obtain

$$Y^K(n) \to_D Y^{\infty}_{BIN^-}(n) \text{ with } K \to \infty.$$

The estimator Y^K is asymptotically unbiased. In fact, applying the result of Stuart (Stuart (1998), p. 351) and making use of (9)-(10), we obtain

$$E\left(Y^K(n)\right) = E\left(\frac{\sum_{i=1}^{n} L_i^K}{\sum_{i=1}^{n} X_i^K}\right) = \frac{E\left(\sum_{i=1}^{n} L_i^K\right)}{E\left(\sum_{i=1}^{n} X_i^K\right)} + O\left(\frac{1}{n}\right) =$$

$$\frac{n(1-(1-pr)^K)}{n\frac{1-(1-pr)^K}{pr}} + O\left(\frac{1}{n}\right) = pr + O\left(\frac{1}{n}\right).$$

Hence,

$$\lim_{n\to\infty} E\left(Y^K(n)\right) = pr.$$

4. Horvitz-Thompson estimators

In order to use direct estimators we need to know probabilities of inclusion of a household in the sample. In the Simple Random Sampling, these probabilities can be calculated before survey is conducted. In fact, in the Random Route we can derive them after we data from the survey are collected.

By $\pi_{i|S}$ and $\pi_{i|P\setminus S}$ we denote probability of inclusion of i^{th} household into the sample for S-member and $P\setminus S$-member, respectively. Assume that M is the size of population, p is the share of S-members in population, r is the level of completeness and $q = 1 - pr$.

If $K = \infty$ then

$$1 - \pi_{i|S} = \frac{prM-1}{M} + \frac{qM}{M}\frac{prM-1}{M-1} + \frac{qM}{M}\frac{qM-1}{M-1}\frac{prM-1}{M-2} + ... = \frac{prM-1}{prM},$$

$$1 - \pi_{i|P\setminus S} = \frac{prM}{M} + \frac{qM-1}{M}\frac{prM}{M-1} + \frac{qM-1}{M}\frac{qM-2}{M-1}\frac{prM}{M-2} + ... = \frac{prM}{prM+1}.$$

Hence,

$$\pi_{i|P\setminus S} = \frac{1}{prM+1} < \frac{1}{prM} = \pi_{i|S}$$

so $\pi_{i|P\setminus S}$ and $\pi_{i|S}$ are not equal and depend on size of S only. However, the difference between these probabilities becomes negligible for relevant size of S, e.g. if $prM \geq 1000$ then we have $\pi_{i|S} - \pi_{i|P\setminus S} \leq 10^{-6}$. In the case $K = \infty$, the parameter pr can be estimated from (4).

Consider the case where K is finite. Then

$$1 - \pi_{i|S} = \frac{prM-1}{M} + \frac{qM}{M}\frac{prM-1}{M-1} + \frac{qM}{M}\frac{qM-1}{M-1}\frac{prM-1}{M-2} + ...+$$

$$\frac{qM}{M}\frac{qM-1}{M-1} \times ... \times \frac{qM-K+2}{M-K+2}\frac{prM-1}{M-K+1},$$

$$1 - \pi_{i|P\setminus S} = \frac{prM}{M} + \frac{qM-1}{M}\frac{prM}{M-1} + \frac{qM-1}{M}\frac{qM-2}{M-1}\frac{prM}{M-2} + ...+$$

$$\frac{qM-1}{M}\frac{qM-2}{M-1} \times ... \times \frac{qM-K+1}{M-K+2}\frac{prM}{M-K+1}.$$

In the case of a finite K, the parameter pr can be estimated from (12).

5. The Random Route in practice. Case of tourism survey in Poland.

The survey "Participation of Polish residents in tourism" has been carried out by the Statistical Office in Rzeszów since the first quarter of 2014. The target population are Polish people who travelled abroad (for one day or more) and people who made a domestic trip for at least one night. Taking into account possibly low completeness rate and the fact that the population of travellers is not very big, the Random Route was applied with 8 steps and 18,750 starting points in a population over 13 million of households.

The data base from the third quarter of 2017 was analysed to assess the aforementioned methods of estimating *pr*. We collected information about the step on which in a route the household was surveyed. All of the households on the first step of the route could be treated as data collected in the Simple Random Sampling. Thus, it allowed us to estimate *pr* like a simple fraction. We expected that the level of completeness *r* should be higher in the Simple Random Sampling than in the Random Route Sampling because all of the households in the starting points are informed about the survey from a letter of President of Statistics Poland. Therefore, an estimate of *pr* should be also higher.

The table below presents the precision of formulas described above.

	Fraction formula	Formula (12)	Formula (7)	Formula (4)
Estimate of *pr*	0.104	0.091	0.083	0.081

Clearly, (12) estimate is the closest to *pr* obtained from the fraction formula. It may stem from the observation that (4) is derived under the assumption of infinite number of steps while (7) is not utilizing information on the total number of steps.
Further investigations based on simulations may create a better picture of properties of the estimators given by (4), (7) and (12).

6. Conclusions

As a cost-effective and time-effective survey method, the Random Route may be preferred to the Simple Random Sampling, especially when we deal with small populations. Under some natural assumptions the weights for the Horvitz-Thompson estimator are easy to compute. The Random Route proved its usefulness also in practice.

REFERENCES

HOFFMEYER-ZLOTNIK, J. H. P., (2003). New Sampling Designs and the Quality of Data. *Methodoloski zvezki - Advances in Methodology and Statistics*, 19. Ljubljana: FDV, pp. 205–217.

DE RADA, V. D., MARTIN, V. M., (2014). Random Route and Quota Sampling: Do They Offer Any Advantage over Probably Sampling Methods?, *Open Journal of Statistics*, 4 (5). DOI: 10.4236/ojs.2014.45038.

BAUER, J. J., (2014). Selection Errors of Random Route Samples, *Sociological Methods & Research*, 43 (3), pp. 519–544. DOI: 10.1177/0049124114521150.

BAUER, J. J., (2016). Biases in Random Route Surveys, *Journal of Survey Statistics and Methodology*, 4 (2), pp. 263–287. DOI: 10.1093/jssam/smw012.

PFANZAGL, J., (1994). *Parametric Statistical Theory*, Berlin: Walter de Gruyter.

GUENTHER, W. C., (1975). *The Inverse Hypergeometric - A Useful Model.* Statistica Neerlandica, 29, pp. 129–144.

ZHANG, L., JOHNSON, W. D., (2011). *Approximate Confidence Intervals for a Parameter of the Negative Hypergeometric Distribution* Proceedings of the Survey Research Methods Section, American Statistical Association.

JOHNSON, N. L., KOTZ, S., (1969). *Distributions in statistics, discrete distributions*, Wiley.

BANDYOPADHYAY, P. S., FORSTER, M. R., (2011). *Philosophy of Statistics*, North Holland.

HILBE, J. M., (2011). *Negative binomial regression*, Cambridge University Press.

STUART, A., (1998). *Kendall's Advanced Theory of Statistics*, Wiley.

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 3 pp. 195–198,

About the Authors

Abdollahnezhad Kamel is an Assistant Professor of statistics at the Department of Statistics in Golestan University, Goran, Iran. His research interests are statistical inference, distribution theory and lifetime distributions. He has published over 20 research articles.

Abu Bakar Mohd Aftar is a Senior Lecturer at the Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan, Malaysia. His main areas of interest include time series analysis, machine learning, data science and statistical modelling.

Abuzaid Ali is an Associate Professor in the Department of Mathematics with a concentration in statistics at Al-Azhar University – Gaza, Palestine. He holds a PhD and MSc in statistics from University of Malaya, Malaysia. His main areas of interest include circular data, survival analysis, time series and development of outlier detection procedures in different types of data. He has been the dean of planning and quality at Al Azhar University – Gaza.

Alizadeh Morad is an Assistant Professor at the Department of Statistics, Faculty of Sciences, Persian Gulf University, Bushehr, Iran. His main areas of interest include univariate continuous distributions, lifetime distributions and related inference. He has published over 100 papers.

Assar Salwa Mahmoud is an Associate Professor at the Department of Mathematical Statistics, Faculty of Graduate Studies for Statistical Research, Cairo University. Her main areas of interest include life testing, stochastic processes, survival analysis and the area of estimation.

Abdelghaffar Ahmed M. is an Assistant Manager at Central Bank of Egypt (Banking Supervision Department). He holds a master's degree in statistics from the Faculty of Graduate Studies for Statistical Research. His main areas of interest include probability distributions, record values, goodness of fit tests, high-dimensional statistics, Bayesian statistics and machine learning.

Bondaruk Taisiia is the Head of the Department of Finance, Banking and Insurance of the National Academy of Statistics, Accounting and Audit, Doctor of Economic Sciences, Professor. Her main areas of interest include budget decentralization, local budget and economic development of territories. She is an author of over 170 scientific works, monographs and handbooks on the problems of public and local

finance, local self-government and budgetary security. She is a member of the Specialized Scientific Council of the Institute of Economics and Forecasting of NAS of Ukraine. She is a member of the editorial board of the Scientific Bulletin of National Academy of Statistics, Accounting and Audit.

Hassan Amal Soliman is a Professor of Statistics at the Department of Mathematical Statistics, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt. She received her PhD degree in statistics from the Institute of Statistical Studies & Research, Cairo University, Egypt, in 1999. Currently, she holds the position of Vice-Dean of Community Service & Environmental Development in Faculty of Graduate Studies for Statistical Research, Cairo University. Her main research interests are: probability distributions, record values, ranked set sampling, stress-strength models, accelerated life tests and goodness of fit tests.

Jan Tariq Rashid is presently acting as a Professor in the Department of Statistics, University of Kashmir. He obtained his doctorate from the University of Kashmir in 2004. He has visited some foreign universities in the USA, Malaysia and his research interests are in the areas of bio-statistics, reliability theory and generalised models in biological sciences. He has published several research articles in reputed international journals of mathematical and biosciences.

Krzyśko Mirosław is a Full Professor of Mathematics and Statistics. His research interests are multivariate statistical analysis, analysis of multivariate functional data, statistical inference and data analysis in particular. Professor Krzyśko has published over 150 research papers in international/national journals and conferences. He has also published five books/monographs. Professor Krzyśko is an active member of many scientific professional bodies.

Marganpoor Shahdie has a master's degree from the Golestan University, Goran, Iran. Her research interests are asymptotic properties, distribution theory and lifetime distributions.

Marszałek Marta is an Assistant Professor at the Institute of Statistics and Demography, SGH – Warsaw School of Economics. Simultaneously she holds the position of an expert of "Third economy satellite accounts" at the Social Surveys Department in Statistics Poland. Her main areas of interest include household production, care and housework valuation, national accounts and satellite accounts. She is a member of the global team of generational economy "National Transfer Accounts and National Time Transfer Account".

Momotiuk Liudmyla is the Vice-Rector of the National Academy of Statistics, Accounting and Audit, Doctor of Economic Sciences, Professor. Her main areas of interest include public finance, financial statistics. She is an author of over 100 scientific works, monographs and handbooks on the problems financial stability,

public finance statistics, balance of payment, financial account. She is a member of the Specialized Scientific Council of the National Academy of Statistics, Accounting and Audit. She is a member of two editorial boards: Statistics of Ukraine, Scientific Bulletin of National Academy of Statistics, Accounting and Audit.

Nandram Balgobin has been a Full Professor of Statistics at Worcester Polytechnic Institute since 2003. He does research in Bayesian statistics and small area estimation interfacing. He is an Adjunct Professor at three other major universities. He has published numerous research articles in statistical journals such as the Journal of the American Statistical Association, where he was a two-term Associate Editor. He has supervised many MS and PhD dissertations both nationally and internationally. He is a fellow of the ASA and an elected member of the ISI and Sigma Xi. In 2003-2005, he was Sinclair Professor of Mathematical Sciences at WPI and in 2006 he received the prestigious SPAIG award for WPI from the ASA. Since 2015, he has been a Senior Research Scientist at the National Agricultural Statistics Service, USDA.

Osaulenko Oleksandr H. is the Rector of the National Academy of Statistics, Accounting and Audit, Doctor of Public Administration, Professor, Corresponding Member of the National Academy of Sciences of Ukraine, Honored Economist of Ukraine. During 1996-2014, he headed the national statistical office of Ukraine. He is an author of over 200 scientific works, including 20 monographs and handbooks on the problems of statistics, public administration and information security. He is the Head of the Specialized Scientific Council of the National Academy of Statistics, Accounting and Audit. He is the editor-in-chief of two editorial boards: Statistics of Ukraine, Scientific Bulletin of National Academy of Statistics, Accounting and Audit. His research results became the theoretical and methodological basis for the preparation of long-term programs for development of national statistics and improvement of statistical support management at national and regional levels, and received practical implementation in a number of legislative and other normative legal acts.

Para Bilal Ahmed is a researcher in the field of biostatistics and is presently acting as a faculty member in the Department of Statistics, Government Degree College Anantnag (J&K), India. He has completed his PhD from the Department of Statistics, University of Kashmir. He has published several research publications in Sci and Scopus indexed journals and has attended and presented in over twenty national and international conferences.

Ranjbar Vahid is an Assistant Professor of statistics at the Department of Statistics in Golestan University, Goran, Iran. His research interests are asymptotic properties, distribution theory and lifetime distributions. He has published (accepted) over 25 research articles.

Shanker Rama is working as an Associate Professor and the Head at the Department of Statistics, Assam University, Silchar, India. His main areas of research include distribution theory, statistical inference, biostatistics, reliability and mathematical programming. He has proposed several new lifetime distributions for modelling lifetime data from biomedical sciences and engineering. He has over 150 research papers in reputed journals of statistics. He is a member of editorial boards of several reputed journals of statistics and biostatistics and is presently working as an associate editor of Biometrics and Biostatics International Journal. He was the founding editor-in-chief of Eritrean Journal of Science and Engineering.

Shukla Kamlesh Kumar has been working as an Associate Professor at Department of Statistics, Mainefhi College of Science, Asmara, Eritrea since February 2016. He has been awarded PhD in statistics from Banaras Hindu University, and have more than 15 years of teaching experience at colleges/universities including international experience. He has worked on six projects in international and national organization, viz.: IIPS, WHO, DST, New Delhi, India, and presented many papers in international and national conferences. He has published over 100 research papers in international and national journal. His research fields of interest are distribution theory, mathematical modelling and migration (demography).

Smaga Łukasz received his MSc, PhD and postdoc in mathematics from the Faculty of Mathematics and Computer Science of Adam Mickiewicz University in Poznań, Poland, in 2009, 2013 and 2020 respectively. Currently, he is an Assistant Professor at this University. His research interests include mathematical statistics, machine learning and their applications. He is an author or co-author of over 40 research papers.

Yin Jiani is a Principal Statistician at Takeda Pharmaceuticals. Prior to Takeda, Jiani worked at Alkermes and Veristat. Her main areas of interest include Bayesian small area estimation and survey sampling.

Wójcik Sebastian is an Assistant at the Institute of Mathematics, the University of Rzeszów and a Head of the Mathematical Statistics Division in the Statistical Office in Rzeszów. His main areas of interest include the probability calculus, functional equations and the utility theory as well as machine learning, data analysis and visualization in R.

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).