STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

CONTENTS

Volume 20, Number 3, September 2019

EDITORIAL OFFICE **ISSN 1234-7655**

Scientific Secretary
Marek Cierpiał-Wolan, e-mail: m.cierpial-wolan@stat.gov.pl
Secretary
Patryk Barszcz, e-mail: p.barszcz@stat.gov.pl, phone number + 48 22 — 608 33 66
Technical Assistant
Rajmund Litkowiec, e-mail: r.litkowiec@stat.gov.pl

Address for correspondence

Statistics Poland, al. Niepodległości 208, 00-925 Warsaw, Poland, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. I−IV

FROM THE EDITOR

I am pleased to announce that, with this issue, our journal's Editorial Board is being extended by another distinguished scholar as Partha Lahiri has accepted our invitation to join the group of scientific advisors and supporters of the Statistics in Transition new series (SiTns). Known for his extraordinary scientific achievements and organizational activity worldwide, Partha has also agreed to take on a leading role, as a Guest Editor-in-Chief, in organizing a special issue of SiTns devoted to statistical data integration – see call for papers (page 187).

The third issue of Statistics in Transition new series 2019 is composed of a set of seven research articles and of two papers based on presentations at the conferences held in Łodz (Multivariate Statistical Analysis, 2018) and in Ciechocinek (Classification and Data Analysis. Theory and Applications. 2018).

The issue is opened by **Vivek Verma's** and **Dilip C. Nath's** paper ***Characterization of the sum of binomial random variables under ranked set sampling*** in which authors examine the characteristics of the sum of independent and nonidentical set of binomial ranked set samples, where each set has a different order depending success probability. The characterization is done by establishing the general recurrence relations for two different situations based on the number of cycle, which is initially pre-assumed as a constant integer and when it is a random variable. To extend the knowledge about the characteristics of the sum in terms of their behaviour and pattern, first four moments, i.e. mean, variance, skewness and kurtosis are derived and compared with the sum of binomial simple random samples with the same success probability. The proposed procedure is illustrated with a real-life data on survivorship of children aged under one in Empowered Action Groups (EAG) states of India. Results show that the sum based on ranked set samples provides more reliable and accurate estimates than that of alternative one, for all selected states taken into account.

The next article, by **Mohammed Abduljaleel, Habshah Midi** and ***Mostafa Karimi, Outlier detection in the analysis of nested Gage R&R, random effect model***, starts with an observation that measurement system analysis is a comprehensive valuation of a measurement process and characteristically includes a specially designed experiment that strives to isolate the components of variation in that measurement process. Gage repeatability and reproducibility is the adequate technique to evaluate variations within the measurement system. Repeatability refers to the measurement variation obtained when one person repeatedly measures the same item with the same Gage, while reproducibility refers to the variation due to different operators using the same Gage. The two factors factorial design, either crossed or nested factor, is usually used for a Gage R&R study. In this study, the focus is only on the nested factor, random effect model. Presently, the classical method (the method of analysing data without taking into consideration the existence of outliers) is used to analyse the nested Gage R&R data. However, this method is easily affected by outliers and, consequently,

the measurement system's capability is also affected. Therefore, the aims of this study are to develop an identification method to detect outliers and to formulate a robust method of the measurement analysis of the nested Gage R&R, random effect model. The proposed methods of outlier detection are based on a robust mm location and scale estimators of the residuals. The results of the simulation study and real numerical example show that the proposed outlier identification method and the robust estimation method are the most successful methods for the detection of outliers. However, the other two methods are not performing well and suffer from masking effect.

Mansoor Rashid Malik and **Devendra Kumar** discuss ***Generalized Pareto distribution based on generalized order statistics and associated inference*** taking into account various structural properties of the distribution that are derived, including (quantile function, explicit expressions for moments, mean deviation, Bonferroni and Lorenz curves and Renyi entropy). Authors provided simple explicit expressions and recurrence relations for single and product moments of generalized order statistics from generalized Pareto distribution. The method of maximum likelihood is adopted for estimating the model parameters. Authors are considering the Bayes estimators of the unknown parameters under the assumption of gamma priors with respect to the shape and the scale parameters. The Bayes estimators are inaccessible in explicit forms, therefore authors analyse the above with reference to both symmetric and asymmetric loss functions. The Bayes interval of this distribution is also derived and - for different parameter settings and sample sizes - various simulation studies are performed and compared to the performance of the generalized Pareto distribution.

In the paper ***estimation of product of two population means by multi-auxiliary characters under double sampling the non-respondents*** **Brij Behari Khare, Raghaw Raman Sinha** consider the problem of estimating the product of two population means using the information on multi-auxiliary characters with double sampling the non-respondents. Classes of estimators are proposed for estimating P under two different situations in the literature using known population mean of multi auxiliary characters. Further, this problem is extended to the case when population means of the auxiliary characters are unknown and they are estimated on the basis of a larger first phase sample. In this situation, a class of two phase sampling estimators for estimating P is suggested using multi-auxiliary characters with unknown population means in the presence of non-response. The expressions of bias and mean square error of all the proposed estimators are derived and their properties are studied. An empirical study using real data sets is given to justify the theoretical considerations.

Olga Komorowska, Arkadiusz Kozłowski and **Teresa Słaby** present the results of ***Comparative analysis of poverty in families with a disabled child and families with non-disabled children in Poland in the years 2014 and 2016***. The presence of a child with disabilities in a family presents more challenging conditions than the presence of a non-disabled child. One of the difficulties is of financial nature. One of the parents often has to give up their job to care for the child, which shrinks the household income. At the same time, the family has higher expenses resulting from, e.g. costs of treatment. All this increases the risk of falling into poverty. The goal of this paper is to analyse the financial situation of households with a disabled child, mainly in the context of poverty, and compare it to the financial

situation of households with non-disabled children. The study is based on data from Polish Household Budget Survey, covering two years, 2014 and 2016. The study revealed that families with a disabled child are generally poorer than families with non-disabled children. The financial situation improved over the studied period in both types of families, but the improvement in the families with a disabled child was much greater. The main factor in reducing the risk of poverty in both types of families is the education attainment level of the reference person (the household head), which should be at least upper secondary.

OlaOluwa Simon Yaya, Olalekan J. Akintande, Ahamuefula E. Ogbonna, Hammed Mumimi Adegoke in the paper ***CPI inflation in Africa: fractional persistence, mean reversion and nonlinearity*** discuss the price stability as the key mandates that apex monetary authorities strive to achieve globally. While most developed economies have achieved single digit inflation rates, most developing economies, especially African countries, still experience alarming double-digit inflation rates. Therefore, this paper examine the dynamics of inflation in sixteen African countries. Authors employed the fractional persistence framework with linear trend and non-linear specifications based on Chebyshev's polynomial in time. The results indicated nonlinear time trend in inflation for most of the countries. With the exception of Burkina Faso, which exhibited plausibility of naturally reverting to its mean level, the majority of the selected African countries would require stronger interventions to revert their observed inflationary levels to their mean levels.

Authors conclude that mean reversion is likely to occur in CPI inflation of Burkina Faso. In the choice of methodology for analysing inflation in Africa, this work recommends a careful selection of the estimation approach, particularly in countries where nonlinearities are detected.

Wojciech Roszka's paper ***spatial microsimulation of personal income in Poland at the level of subregions*** presents application of spatial microsimulation methods for generating synthetic population to estimate personal income in Poland in 2011 using census tables and EU-SILC 2011 microdata set. In the first section a research problem is presented along with a brief overview of modern estimation methods in application to small domains with particular emphasis on spatial microsimulation. The second section contains an overview of selected synthetic population generation methods. In the last section personal income estimation on NUTS 3 (sub-region) level is presented with special emphasis placed on the quality of estimates. Solving the problem of the sample size, correction of random and non-random errors, the possibility of performing different simulations are undoubted advantages of the discussed (SMM) methods, which encourage deepening the work and analysis of the effectiveness and reliability of the estimates.

The section Other articles containing post-conference papers starts with **Katarzyna Budny's** article ***Power generalization of Chebyshev's inequality – multivariate case***. Some qualities of the multivariate power generalizations of Chebyshev's inequality are discussed and some improvements with extension to a random vector with singular covariance matrix are suggested. For these generalizations, the cases of the multivariate normal and the multivariate *t* distributions are considered along with presenting some financial application.

In the paper ***Decomposition of gender wage gap in poland using counterfactual distribution with sample selection*****, Joanna Małgorzata**

Landmesser compares income distributions in Poland taking into account gender differences. The gender pay gap can only be partially explained by differences in men's and women's characteristics. The unexplained part of the gap is usually attributed to the wage discrimination. The objective of this study is to extend the Oaxaca-Blinder decomposition procedure to different quantile points along the income distribution. The RIF-regression method is used to describe differences between the incomes of men and women along the two distributions and to evaluate the strength of the influence of personal characteristics on the various parts of the income distributions using data from the EU-SILC for Poland in 2014. As the sample selection is a serious issue for the study, the applied decomposition is adjusted for sample selection problems. The results suggest existence of not only differences in income gap along the income distribution (in particular sticky floor and glass ceiling), but also differences in the contribution of selection effects to the pay gap at different quantiles.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. V

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/

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:

- **The New Scimago Journal & Country Rank**
- BASE – Bielefeld Academic Search Engine
- CEEOL
- CEJSH
- CNKI Scholar
- CIS
- Dimensions
- EconPapers
- Elsevier – Scopus
- ERIH Plus
- Google Scholar
- Index Copernicus
- J-Gate
- JournalGuide
- JournalTOCs
- Keepers Registry
- MIAR
- OpenAIRE
- ProQuest – Summon
- Publons
- RePec
- Wanfang Data
- WorldCat
- Zenodo

STATISTICS IN TRANSITION new series, September 2019
Vol. 20 , No. 3, pp. 1-29, DOI 10.21307/stattrans-2019-022
Submitted – 09.06.2018; Paper ready for publication – 19.03.2019

CHARACTERIZATION OF THE SUM OF BINOMIAL RANDOM VARIABLES UNDER RANKED SET SAMPLING

Vivek Verma[1], **Dilip C. Nath**[2]

ABSTRACT

In this paper, we examined the characteristics of the sum of independent and non-identical set of binomial ranked set samples, where each set has different order depending success probability. The characterization is done by establishing the general recurrence relations for two different situations based on the number of cycle, which is initially pre-assumed as a constant integer and when it is a random variable. To extend the knowledge about the characteristics of sum in terms of their behaviour and pattern, first four moments $i.e.$, mean, variance, skewness and kurtosis are derive and compared with the sum of binomial simple random samples with same success probability. The proposed procedure has been illustrated through a real-life data on survivorship of children below one year in Empowered Action Groups (EAG) states of India.

Key words: Factorial moment generating function, Skewness; Kurtosis, Poisson distribution.

1. Introduction

The role of Ranked Set Sampling (RSS) as an alternative method of Simple Random Sampling (SRS) have been investigated since the time McIntyre(1952), who first introduced this sampling procedure. Since, then many authors have discussed about the efficacy of RSS either theoretically or analytically. RSS is found to be very effective in contexts where exact measurement of sampling units is expansive in time or toil; but the sample unit can be readily ranked either through subjective or *via* the use of relevant concomitant variables.

[1]**Corresponding Author**. Department of Statistics, Gauhati University, Guwahati, Assam, 781014, India, **Current**: Department of Neurology, All India Institute of Medical Sciences, New Delhi, 110029, India. E-mail: viv_verma456@yahoo.com
[2]Assam University, Silchar, 788011, Assam, India. E-mail: dilipc.nath@gmail.com

Comparing with earlier studies where the variable of interest is continuous, infrequent research discuss about the effectiveness of RSS where the variable is binary. In cases, where the variable of interest is binary, there are two possible outcomes, success (denoted as 1) and failure (denoted as 0) and is supposed to follow Bernoulli with success probability p, say. Here, the probability p can be viewed as a proportion of individuals with certain characteristic in the population. The foregoing studies of RSS, where the response is a binary variable, Terpstra (2004), Chen et al. (2005, 2006, 2007, 2008), Chen (2008), Verma et.al(2017), Das et.al (2017) Lacayo et al. (2002), Terpstra and Miller (2006) and Chen et al. (2009), are mainly concerned about estimation of population proportion and variance, and the comparison with SRS is done using these estimates.

Obtaining the behaviour of a sum of Bernoulli random variables based on simple random sample has found of greater importance in various applications like formalization random walk process (Takacs, 1991), the Stein-Chen method for approximation of Poisson (Barbour and Holst, 1989), obtaining bounds for entropy (Sason, 2013), characterization of flows in internet traffic (Chabchoub et al. 2010), and approximation of rare events (Chen and Rollin, 2013). The problem of estimating the characteristics of sum of independent binary variable in terms of their moments based on simple random samples has already been emphasized by many researchers like Malik (1969), Ahuja (1970), Percus and Percus (1985), Ling (1988), Horvath (1989), Yu and Zelterman (2002), and Kadane (2016). As an alternative procedure of SRS, RSS has found to be more efficient and reliable, but the characterization of a sum of independent and non-identical Bernoulli random variables based on ranked set sample has not been considered in the literature. In this connection the present article has mainly concerned to establish a recurrence relations between the factorial moments of sum of independent and non-identical sets of binary variables, which is never procured in case of RSS. These relations also assists to reduce the number of independent calculations required for evaluation of moments under RSS. And, helps in characterizing the sum of binary variables under RSS by using recurrence relations and compare with SRS.

In this paper, the recurrence relationship of sum of binary variables under SRS and balanced RSS for fixed set size, s, and probabilities of success, p, are obtained under two different situations. In the first case, the relationship

is obtained when number of cycles, m is assumed to be known fixed values. In the second case, an attempt is being made to extend the previously mentioned results of recurrence relations among the sum of independent binary variable where the number of cycle is a random variable. To understand the characteristics of sum of binary variables under SRS and RSS, the first four moments are derived using factorial moments that by establishing the recurrence relationships. The technique of asymptotic approximations has been found very helpful in various aspects like in Monte Carlo simulation (Hastings, 1970) and bootstrap techniques (Freedman and Peters, 1984, Brown and Newey, 2002), for obtaining numerical estimates and their asymptotic variance and asymptotic confidence intervals. In characterization of any distributions, asymptotical aspects adds additional information of the distribution and it is an important feature to describe the how large the sample size is required to achieve the asymptotic approximation. A simulation based comparison among SRS and RSS has been discussed to numerically illustrate the requirement of sample size to achieve the asymptotic normality. A practical illustration of the proposed procedure with a real-life data on child survivorship for all eight selected Empowered Action Groups (EAG) Indian states viz., Bihar, Uttaranchal, Chhatisgarh, Jharkhand, Orissa, Rajasthan, Madhya Pradesh and Uttar Pradesh, has also been presented.

2. Sampling Design

Suppose the variable of interest is dichotomous variable, say X, and $n(=ms)$, denotes size of the sample drawn from the population by adopting the procedures of SRS and RSS, respectively, for prefixed set size,s and number of cycles, m. Let $\{X_{[r]i}; r = 1(1)s, i = 1(1)m\}$ symbolizes a ranked set sample of size ms, where $X_{[r]i}$ denotes the i^{th} observation in the r^{th} ranking class. Because of RSS procedure, $X_{[r]i}$'s are independently distributed and corresponding to each r^{th} set, $(X_{[r]1}, X_{[r]2}, \cdots, X_{[r]m})$ are independently and identically (i.i.d.) distributed and $X_{[r]1}$ is the r^{th} order statistic from a simple random sample of s observations on X. Let $\mathbf{X_{SRS}} = (X_1, X_2, \cdots, X_n)$ is an i.i.d. simple random sample from Bernoulli(p) and $W(= \sum_{i=1}^{n} X_i)$, denotes their sum and its density is given by,

$$f_W(w) \quad = \quad \binom{n}{w} p^w \, (1-p)^{n-w} \quad ; w = 0(1)n; 0 \leq p \leq 1 \qquad (2.1)$$

Let $\mathbf{X}_{[\mathbf{r}]} = (X_{[r]1}, X_{[r]2}, \cdots, X_{[r]m})$ is an vector of i.i.d. ranked set samples of r^{th} set from Bernoulli $(p_{[r]})$, for all $r = 1(1)s$ and $Y_r = \sum_{j=1}^{m} X_{[r]j}$, is the number of times the event occurred in r^{th} class, follows Binomial $(m, p_{[r]})$ and is given by,

$$P(Y_r = y_r) = \binom{m}{y_r} p_{[r]}^{y_r} (1-p_{[r]})^{m-y_r} \quad ; y_r = 0, 1, \cdots, m. \tag{2.2}$$

Here, $p_{[r]} = I_p(s-r+1, r)$, denotes the standard incomplete beta integral and is given by,

$$I_x(a,b) = \frac{1}{\mathscr{B}(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} dt, \;\; 0 < x < 1.$$

where $\mathscr{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. And, $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_r, \cdots, Y_s)$ is vector of independent Binomial variate with parameters $m > 0$ and $p_{[r]}$ for all $r = 1(1)s$ and $Z = \sum_{r=1}^{s} Y_r$, denotes their sums.

3. Characterization using the Recurrence Relations

Let $G(t) = \sum_{x=0}^{n} t^x P(X = x)$, denotes the probability generating function (*pgf*) of a random variable X having distribution $P(X = x)$, with support $0, 1, \cdots$, $n(= ms) \in \mathbf{Z}^+$.

3.1. Case-I: The number of cycles, m, is a known fixed value

Theorem-1: For fixed m, the recursive relationship among factorial moment of sum W and Z, under SRS and RSS, respectively, is given by

$$\mu'_{[k]}(W) = ms \sum_{j=0}^{k-1} (-1)^j \frac{(k-1)!}{(k-1-j)!} p^{(j+1)} \mu'_{[k-1-j]}(W) \tag{3.1}$$

and

$$\mu'_{[k]}(Z) = m \sum_{j=0}^{k-1} (-1)^j \frac{(k-1)!}{(k-1-j)!} \left(\sum_{i=1}^{s} p_{[i]}^{j+1} \right) \mu'_{[k-1-j]}(Z). \tag{3.2}$$

where $\mu'_{[0]} = 1$.

Proof: Suppose that W, denotes the sum of $n(= ms)$ $i.e., = \sum_{i=1}^{n} X_i$ and Y_r;

$\forall r = 1(1)s$, is the sum of m *i.e.*, $= \sum_{j=1}^{m} X_{[r]j}$, Bernoulli variables with parameters p and $p_{[r]}$ respectively. The factorial moment generating function (*fmgf*) of W and Y_r using equations (2.1)-(2.2) respectively, are given by

$$\begin{aligned} G_W(t+1) &= \prod_{i=1}^{n} G_{X_i}(t+1) = (1+pt)^n \qquad (3.3) \\ G_{Y_r}(t+1) &= E((t+1)^{Y_r}) = \prod_{j=1}^{m} E((t+1)^{X_{[r]j}}) \\ &= (1+p_{[r]}t)^m. \qquad (3.4) \end{aligned}$$

Since, $\{Y_r\}$'s is a set of mutually independent Binomial variate, therefore, the *fmgf* of $Z(= \sum_{r=1}^{s} Y_r)$ using equation-(3.4) is given by

$$G_Z(t+1) = \prod_{r=1}^{s} G_{Y_r}(t+1) = \prod_{r=1}^{s} (1+p_{[r]}t)^m. \qquad (3.5)$$

If D denotes the differential operator *i.e.*, $\frac{d}{dt}$, then the recursive relationship between the factorial moments of W, based on simple random samples, can be obtained by successive differentiation of equation-(3.3) and are as follows

$$\begin{aligned} D(G_W(1+t)) &= \frac{np}{1+tp} G_W(1+t), \\ D^2(G_W(1+t)) &= \frac{np}{1+tp} D(G_W(1+t)) - \frac{np^2}{(1+tp)^2} G_W(1+t), \\ D^3(G_W(1+t)) &= \frac{np}{1+tp} D^2(G_W(1+t)) - \frac{2np^2}{(1+tp)^2} D(G_W(1+t)) + \\ &\quad \frac{2np^3}{(1+tp)^3} G_W(1+t), \\ &\vdots = \vdots, \end{aligned}$$

$$D^k(G_W(1+t)) = n \sum_{j=0}^{k-1} (-1)^j \frac{(k-1)!}{(k-1-j)!} \left(\frac{p}{1+tp}\right)^{j+1} D^{k-1-j}(G_Z(1+t)), \qquad (3.6)$$

and setting $t = 0$ in equation-(3.6) gives

$$\mu'_{[k]}(W) = ms \sum_{j=0}^{k-1} (-1)^j \frac{(k-1)!}{(k-1-j)!} p^{(j+1)} \mu'_{[k-1-j]}(W),$$

where $\mu'_{[0]}(W) = 1$.

The recursive relationship between *fmgf* of Z, which is based on ranked set samples, using equation-(3.5), is given by

$$\begin{aligned}
D(G_Z(1+t)) &= \sum_{i=1}^{s} \left(\frac{mp_{[i]}}{1+tp_{[i]}} \right) \prod_{r=1}^{s} (1+tp_{[r]})^m = \sum_{i=1}^{s} \left(\frac{mp_{[i]}}{1+tp_{[i]}} \right) G_Z(1+t), \\
D^2(G_Z(1+t)) &= D(G_Z(1+t)) \sum_{i=1}^{s} \left(\frac{mp_{[i]}}{1+tp_{[i]}} \right) - G_Z(1+t) \sum_{i=1}^{s} m \left(\frac{p_{[i]}}{1+tp_{[i]}} \right)^2 \\
D^3(G_Z(1+t)) &= D^2(G_Z(1+t)) \sum_{i=1}^{s} \left(\frac{mp_{[i]}}{1+tp_{[i]}} \right) - \\
&\quad D(G_Z(1+t)) \sum_{i=1}^{s} 2m \left(\frac{p_{[i]}}{1+tp_{[i]}} \right)^2 * \\
&\quad G_Z(1+t) \sum_{i=1}^{s} 2m \left(\frac{p_{[i]}}{1+tp_{[i]}} \right)^3, \\
\vdots &= \vdots, \\
D^k(G_Z(1+t)) &= m \sum_{j=0}^{k-1} (-1)^j h(k)_j \left(\sum_{i=1}^{s} p_{[i]}^{j+1} \right) D^{k-1-j}(G_Z(1+t)), \qquad (3.7)
\end{aligned}$$

where $h(k)_j = \frac{(k-1)!}{(k-1-j)!}$, setting $t = 0$ in equation-(3.7) provides

$$\mu'_{[k]}(Z) = m \sum_{j=0}^{k-1} (-1)^j \frac{(k-1)!}{(k-1-j)!} \left(\sum_{i=1}^{s} p_{[i]}^{j+1} \right) \mu'_{[k-1-j]}(Z),$$

where $\mu'_{[0]}(Z) = 1$.

Corollary 1: The first four factorial moments *i.e.,* for $k = 1, 2, 3$ and 4, using equation-(3.1) of W and (3.1) of Z, based on simple random samples and ranked set sample, are given in Appendix-(7.1).

3.2. Case-II: The number of cycles, m, is a random variable

Suppose that the number of cycles, N, is a random variable, where $m \in N^+ = \{1,2,,\cdots\}$. The probability mass function of $N = m$ is given by

$$f_N(m) = \frac{e^{-\lambda}\ \lambda^{m-1}}{(m-1)!}\ ;m = 1,2,\cdots. \tag{3.8}$$

i.e., $N-1 \sim \mathscr{P}(\lambda)$ (Poisson with mean λ), $\lambda > 0$.

Theorem-2: The recursive relationship between factorial moments of marginal sums of W and Z respectively, where N is a random variable and follows the Poisson distribution of equation-(3.8), is given by

$$\mu'_{[k]}(W) = s\sum_{j=0}^{k-1}\frac{(k-1)!}{(k-1-j)!}p^{(j+1)}\left((-1)^j + \frac{\lambda(s-1)!}{(s-1-j)!j!}\right)\mu'_{[k-1-j]}(W), \tag{3.9}$$

$$\mu'_{[k]}(Z) = \lambda\left(\sum_{i=1}^{s}p_{[i]}\right)\mu'_{[k-1]}(Z) + \sum_{j=0}^{k-1}(-1)^j\frac{(k-1)!}{(k-1-j)!}\left(\sum_{i=1}^{s}p_{[i]}^{j+1}\right)\mu'_{[k-1-j]}(Z) \tag{3.10}$$

Proof: Suppose W and Y_r, are the mixtures of binomial distributions with a fixed probability of success p and $p_{[r]}$, respectively, but a variable number of cycles, m, modelled with Poisson distribution discussed in equation-(3.8). The conditional distribution of $W|N = m$ is given by,

$$f_W(w|N=m) \quad = \quad \binom{n}{w}\ p^w\ (1-p)^{n-w} = \binom{ms}{w}\ p^w\ (1-p)^{ms-w}$$

where $w = 0,1,\cdots,ms$ and $0 \le p \le 1$. The *fmgf* of mixture of W using the equations-(2.1) and (3.8) can be derived as,

$$\begin{aligned} G_W(1+t) &= e^{-\lambda}\sum_{m=1}^{\infty}\frac{\lambda^{m-1}}{(m-1)!}\sum_{w=0}^{ms}\binom{ms}{w}\ ((1+t)p)^w\ (1-p)^{ms-w}, \\ &= e^{-\lambda}\sum_{m=1}^{\infty}\frac{\lambda^{m-1}}{(m-1)!}(1+tp)^{ms}, \\ &= e^{-\lambda}(1+tp)^s e^{\lambda(1+tp)^s}. \end{aligned} \tag{3.11}$$

If D denotes the differential operator *i.e.*, $\frac{d}{dt}$, then the recursive relationship between the factorial moments of the mixture of W can be obtained by suc-

cessive differentiation of equation-(3.11) and are as follows,

$$\begin{aligned} D(G_W(1+t)) &= e^{-\lambda}\{ps(1+tp)^{s-1}e^{\lambda(1+tp)^s}+(1+tp)^s e^{\lambda(1+tp)^s}\lambda sp(1+tp)^{s-1}\}, \\ &= G_W(1+t)\left\{\frac{sp}{1+tp}+\lambda sp(1+tp)^{s-1}\right\}, \\ &= s\,G_W(1+t)\left(\frac{p}{1+tp}\right)+s\lambda p\,G_W(1+t)(1+tp)^{s-1}, \end{aligned}$$

$$\begin{aligned} D^2(G_W(1+t)) &= \frac{sp}{1+tp}D(G_W(1+t))-\frac{sp^2}{(1+tp)^2}G_W(1+t)+ \\ &\quad \lambda sp(1+tp)^{s-1}D(G_W(1+t))+ \\ &\quad \lambda s(s-1)p^2(1+tp)^{s-2}G_W(1+t), \end{aligned}$$

$$\begin{aligned} D^3(G_W(1+t)) &= \frac{sp}{1+tp}D^2(G_W(1+t))-\frac{2sp^2}{(1+tp)^2}D(G_W(1+t))+ \\ &\quad \frac{2sp^3}{(1+tp)^3}G_W(1+t)+\lambda sp(1+tp)^{s-1}D^2(G_W(1+t)) \\ &\quad 2\,\lambda s(s-1)p^2(1+tp)^{s-2}G_W(1+t) \\ &\quad \lambda s(s-1)(s-2)p^3(1+tp)^{s-3}G_W(1+t), \\ \vdots &= \vdots, \end{aligned}$$

$$D^k(G_W(1+t)) =$$

$$s\sum_{j=0}^{k-1}\frac{(k-1)!}{(k-1-j)!}\left(\frac{p}{1+tp}\right)^{j+1}\left((-1)^j+\frac{\lambda(s-1)!}{(s-1-j)!j!}(1+tp)^s\right)D^{k-1-j}, \tag{3.12}$$

setting $t=0$ in equation-(3.12) gives the recursive relationship,

$$\mu'_{[k]}(W)=s\sum_{j=0}^{k-1}\frac{(k-1)!}{(k-1-j)!}p^{(j+1)}\left((-1)^j+\frac{\lambda(s-1)!}{(s-1-j)!j!}\right)\mu'_{[k-1-j]}(W) \tag{3.13}$$

where $\mu'_{[0]}(Z)=1$.

The *fmgf* of mixture of Y_r of the r^{th} set of $\mathbf{Y}$ by using the equations-(2.2)

and (3.8) can be derived as,

$$\begin{aligned} G_{Y_r}(t+1) &= e^{-\lambda}\sum_{m=1}^{\infty}\frac{\lambda^{m-1}}{(m-1)!}\sum_{y_r=0}^{m}(1+t)^{y_r}\binom{m}{y_r}\,p_{[r]}^{y_r}\,(1-p_{[r]})^{m-y_r} \\ &= e^{-\lambda}\sum_{m=1}^{\infty}\frac{\lambda^{m-1}}{(m-1)!}(1+tp_{[r]})^{m} \\ &= (1+tp_{[r]})e^{-\lambda}\sum_{m=1}^{\infty}\frac{((1+tp_{[r]})\lambda)^{(m-1)}}{(m-1)!} \\ &= (1+tp_{[r]})\,e^{-\lambda(1-(1+tp_{[r]}))} = (1+tp_{[r]})\,e^{\lambda tp_{[r]}}. \end{aligned} \tag{3.14}$$

The *fmgf* of Z using the equation-(3.14) is given by,

$$\begin{aligned} G_Z(1+t) &= \prod_{r=1}^{s}G_{Y_r}(1+t) = e^{t\lambda\sum_{r=1}^{s}p_{[r]}}\prod_{r=1}^{s}(1+tp_{[r]}) \\ &= e^{ta}\prod_{r=1}^{s}(1+tp_{[r]}), \end{aligned} \tag{3.15}$$

where $z=0,1,\cdots,ms$ and $0\le p_{[r]}\le 1$, $a=\lambda sp$ and $p=\sum_{r=1}^{s}p_{[r]}/s$, for all $r=1(1)s$. The recursive relationship between the factorial moments of the mixture of Z can be obtained by successive differentiation of equation-(3.15) and are as follows,

$$\begin{aligned} D(G_Z(1+t)) &= ae^{ta}\prod_{r=1}^{s}(1+tp_{[r]})+e^{ta}\sum_{i=1}^{s}\left(\frac{p_{[i]}}{1+tp_{[i]}}\right)\prod_{r=1}^{s}(1+tp_{[r]}), \\ &= aG_Z(1+t)+G_Z(1+t)\sum_{i=1}^{s}\left(\frac{p_{[i]}}{1+tp_{[i]}}\right), \\ D^2(G_Z(1+t)) &= aD(G_Z(1+t))+D(G_Z(1+t))\sum_{i=1}^{s}\left(\frac{p_{[i]}}{1+tp_{[i]}}\right)- \\ &\quad G_Z(1+t)\sum_{i=1}^{s}\left(\frac{p_{[i]}}{1+tp_{[i]}}\right)^2, \end{aligned}$$

$$
\begin{aligned}
D^3(G_Z(1+t)) &= aD^2(G_Z(1+t)) + D^2(G_Z(1+t)) \sum_{i=1}^{s} \left(\frac{p_{[i]}}{1+tp_{[i]}} \right) - \\
&\quad D(G_Z(1+t)) \sum_{i=1}^{s} 2 \left(\frac{p_{[i]}}{1+tp_{[i]}} \right)^2 + \\
&\quad G_Z(1+t) \sum_{i=1}^{s} 2 \left(\frac{p_{[i]}}{1+tp_{[i]}} \right)^3, \\
\vdots &= \vdots,
\end{aligned}
$$

$$
D^k(G_Z(1+t)) =
$$

$$
aD^{k-1}(G_Z(1+t)) + \sum_{j=0}^{k-1} (-1)^j \frac{(k-1)!}{(k-1-j)!} \left(\sum_{i=1}^{s} p_{[i]}^{j+1} \right) D^{k-1-j}(G_Z(1+t)), \quad (3.16)
$$

setting $t = 0$ in equation-(3.16) gives recursive relationship as,

$$
\mu'_{[k]}(Z) = a\mu'_{[k-1]}(Z) + \sum_{j=0}^{k-1} (-1)^j \frac{(k-1)!}{(k-1-j)!} \left(\sum_{i=1}^{s} p_{[i]}^{j+1} \right) \mu'_{[k-1-j]}(Z)
$$

where $\mu'_{[0]}(Z) = 1$.

Corollary 2: The first four factorial moments, *i.e.,* $k = 1, 2, 3$ and 4, of the mixture of W and Z using equation-(3.9) and equation-(3.10) based on simple random sample and ranked set sample, respectively, where the number of cycles, m, is a random variable and follows the Poisson distribution, is given in Appendix-(7.2).

4. Comparison of Moments

In this section, a comparison is being made among the factorial moments of W and Z. Let $\mathscr{D}_{[k]} = \mu'_k(W) - \mu'_k(Z)$, denotes the difference among factorial moments of W and Z, of order k. For the situation, where m, is a known constant, the difference, $\mathscr{D}_{[k]}$, by using equations-(3.1) and (3.2), is given by

$$
\mathscr{D}_{1[k]} = m \sum_{j=0}^{k-1} (-1)^j \frac{(k-1)!}{(k-1-j)!} \left(sp^{(j+1)} \mu'_{[k-1-j]}(W) - \mu'_{[k-1-j]}(Z) \sum_{i=1}^{s} p_{[i]}^{j+1} \right). \quad (4.1)
$$

Under the assumption that m, is a random variable, $\mathscr{D}_{[k]}$ can be obtained by using the equations-(3.9) and (3.10) and is given by,

$$\mathscr{D}_{2[k]} =$$

$$\sum_{j=0}^{k-1}\left(sp^{(j+1)}\ \gamma_j\ \mu'_{[k-1-j]}(W)-(-1)^j\left(\sum_{i=1}^{s}p_{[i]}^{j+1}\right)\mu'_{[k-1-j]}(Z)\right) \times\frac{(k-1)!}{(k-1-j)!}-\lambda\left(\sum_{i=1}^{s}p_{[i]}\right)\mu'_{[k-1]}(Z), \quad (4.2)$$

where $\gamma_j=\left((-1)^j+\frac{\lambda(s-1)!}{(s-1-j)!j!}\right)$.

Note-1: Since, $p_{[i]}=I_p(s-i+1,i),\ \forall\ i=1,2,\cdots,s$, therefore, a sum of $p_{[i]}$'s is given by,

$$\begin{aligned}\sum_{i=1}^{s}p_{[i]} &= \sum_{i=1}^{s}I_p(s-i+1,i)=\sum_{i=1}^{s}\int_0^p\frac{t^{s-i}(1-t)^{i-1}}{\mathscr{B}(s-i+1,i)}dt\\ &= \int_0^p\sum_{i=1}^{s}\binom{s-1}{i-1}t^{s-i}(1-t)^{i-1}dt=s\int_0^p dt\ =\ sp.\end{aligned}$$

Note-2: Let $\gamma(v)=\frac{1}{p^v}\sum_{i=1}^{s}p_{[i]}^v$, is a constant depends on the order v and for $v=1$, $\gamma(1)=sp/p=s$, *i.e.*, the minimum value of $\gamma(v)$ is s, that implies $\gamma(v)\geq s\ ;\forall v$. Suppose that,

$$\sum_{i=1}^{s}p_{[i]}^v=Cp^v,$$

where $C>0$, is a proportionality constant such that,

$$\sum_{i=1}^{s}p_{[i]}^v-Cp^v = 0 \text{ if } C=\gamma(v) \quad (4.3)$$

$$\sum_{i=1}^{s}p_{[i]}^v-Cp^v > 0 \text{ if } C\in(0,\gamma(v)] \quad (4.4)$$

$$\sum_{i=1}^{s}p_{[i]}^v-Cp^v < 0 \text{ if } C>\gamma(v). \quad (4.5)$$

Note-3: The difference equations of $\mathscr{D}_{1[k]}$ and $\mathscr{D}_{2[k]}$, of equations-(4.1)-(4.2),

respectively, for $k = 1,2$ are such that

$$\mathscr{D}_{1[1]} = m\left(sp - \sum_{i=1}^{s} p_{[i]}\right) = 0 \tag{4.6}$$

$$\mathscr{D}_{1[2]} = m\left(\sum_{i=1}^{s} p_{[i]}^2 - sp^2\right) > 0; \text{ from-equations (4.6) and (4.4)} \tag{4.7}$$

$$\mathscr{D}_{2[1]} = sp(1+\lambda) - \sum_{i=1}^{s} p_{[i]} - \lambda \sum_{i=1}^{s} p_{[i]} = 0 \tag{4.8}$$

$$\begin{aligned}\mathscr{D}_{2[2]} &= sp(1+\lambda)\mu'_{[1]}(W) + sp^2(-1+\lambda(s-1)) - \mu'_{[1]}(Z)\left[\lambda sp + \sum_{i=1}^{s} p_{[i]}\right] + \sum_{i=1}^{s} p_{[i]}^2 \\ &= \left[\sum_{i=1}^{s} p_{[i]}^2 - sp^2\right] + sp^2\lambda(s-1) > 0; \text{from-equations (4.8) and (4.4)}\end{aligned} \tag{4.9}$$

Since, $\sum_{i=1}^{s} p_{[i]} = sp$ and $s \in (0, \gamma(v)]; \forall v$, therefore, from equation-(4.4), we find that the difference between, $\left(\sum_{i=1}^{s} p_{[i]}^v - sp^v\right) > 0$.

5. Simulations

To assess the performance and changes in the moments of sums, when set size, m, is known and unknown, a simulation study is done for different combination of $p \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, $s = 2,4$ and 6, and $m = \lambda = 10, 50, 100, 200$ and 500, under SRS and RSS, are presented in Table 1-2 of Appendix. To compare the accuracy of the estimator under RSS with respect to SRS, the relative efficiency (RE)

$$RE = \frac{\mu_2(SRS)}{\mu_2(RSS)}$$

is also obtained. In addition of that pattern of the skewness and kurtosis of sums based on both SRS and RSS regarding their asymptotic behaviour are also depicted in Figure 1-4 of Appendix.

Discussion: From Table 1, it has observed that the RE of the estimator under RSS with respect to SRS, always greater than 1 for all combination of s, m and p. For fixed number of cycles, m and proportion, p, the RE of the estima-

tor under RSS with respect to SRS, is also increasing in most of the cases with increase in set size, s, such as for $m = 10$ and $p = 0.1$, with increase the value of $s = 2,4,6$, the RE is 1.21, 1.32 and 1.39, respectively. When the set size, s, is fixed and number of cycles are significantly high values, the RE of the estimator under RSS with respect to SRS has followed an increasing trend with increase in proportion p, for e.g., when $m = 500$ and $s = 6$, the RE has obtained as 1.38, 1.81, 2.02, 2.26 and 2.31 for $p = 0.1, 0.2, \cdots 0.5$, respectively. Table 2 is based on the assumption that the number of cycles, is a random quantity and follows zero truncated poisson(λ). Under this truncated poisson assumption, similar pattern of the RE of the estimator under RSS with respect to SRS as previous has obtained. Results has shown that with increase in s and p higher will be the values of RE. It has also found that for fixed s and p, changing in poisson parameter λ does not affect the efficiency of estimators under RSS as comparing to SRS, and remains almost same. It is found from the simulated results that even though the mean under both SRS and RSS are same (Verma et.al (2017)) but the variances of sums based on SRS are often higher than that of RSS, for all m and p, which shows that the number of success obtained using RSS is more reliable and efficient as compare to SRS.

The interaction of sample size, $n = ms$, and skewness and kurtosis, respectively, of sums under SRS and RSS has depicted in Figure 1-2, where the number of cycles m is a fixed quantity. Figures 1 and 2 represents the pattern of five skewness and kurtosis curves, respectively, obtained for fixed set size, s, at different choices of $p \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. When m is a random quantity and follows zero truncated poisson(λ), the pattern of five skewness and kurtosis curves for different choices of p have also presented in Figure 3-4. Through these figures, one can compare and calculate the required sample size, n, to meet that asymptotic normality (Small, 1980, Bai and NG,2005, Sunklodas, 2014, and Butler and Stephens, 2017), *i.e.,* $skewness = 0$ and $kurtosis = 3$, under RSS as compare to SRS, for given p. For fixed set size, s, and proportion p, it has found that with a minimum number of cycles, m or parameter λ, one can achieve asymptotic normality, under RSS as compare to the required number of sample based on SRS.

6. Illustration with Real-life Data

To illustrate a practical significance of the discussed methodology, a real-life data on children aged 0-1 years to mothers aged 15 to 39 years, who

are residing in eight Indian states ((a) Bihar (b) Uttaranchal (c) Chhatisgarh (d) Jharkhand (d) Orissa (e) Rajasthan (f) Madhya Pradesh and (g) Uttar Pradesh,) has considered. These selected eight states are socio-economically backward and reports highest infant mortality rates ($<$ 50 per 1000), and are also known as Empowered Action Groups (EAG) states of India. The data has been obtained from the National Family Health Survey-3 (2005-06), preceding five years of the survey. Here, our objective is to characterize the number of babies that remains alive in EAG states of India, under both SRS and RSS.

The event of survivorship of a child is positively correlated with mother's age (Finlay et.al (2011) and Selemani et.al (2014)). One can say that chance of survivorship of a child is low in mother of lower ages than that of those of higher ages. So mother's age (in months) is used as an auxiliary variable for ranking purpose in ranked set sampling. The procedure adopted for sampling through RSS has discussed below (Das et al. 2017):

1. A simple random sample of s^2 units, say $X_i; i = 1, 2, \cdots, s$, is drawn from the target population and are randomly partitioned into s sets each having s units, , say X_{rj} for all $r = 1, 2, \cdots, s;\ j = 1, 2, \cdots, s$.

2. In each of s sets the units are ranked according to the mother's age, denoted as $X_{[r]j}$. In situation of ties the observations are ordered systematically in the sequence, as discussed by Terpstra and Nelson (2005)

3. From the first set, the unit corresponding to the mother with lowest age is selected ($X_{[1]1}$). From the second set, the unit corresponding to mother with second lowest age is selected ($X_{[2]2}$) and so on. Finally, from the s^{th} set, the unit corresponding to the mother with highest age ($X_{[s]s}$) is selected. The remaining $s(s-1)$ sampled units are discarded from the data set.

4. The Steps 1 - 3, called a cycle, are repeated m times to obtain a ranked set sample of size $n = ms$.

Using the sample we have computed various moments discussed in previous sections under both SRS and RSS. The results have reported in Tables 3 and 4. When the number of cycles m has assumed as a fixed quantity, the obtained result have shown that characterization of the sums based on ranked set sampling for all states are much reliable and its efficiency lies to 10-34%. The kurtosis based on both simple random sample and ranked

set sample have found closer to 3, but significant deviation from 0 have observed in the skewness (negatively skewed). Under the assumption that m is a random quantity, the efficiency increases 3 to 4 times that of earlier case. It has also observed that the variance under RSS is converging towards the mean, which shows the asymptotic convergence to possion distribution. The kurtosis based on both simple random sample and ranked set sample have found far away from 3 and a significant deviation from 0 have observed in the skewness (positively skewed). Statistical Analysis System (SAS) package, University edition has used for sampling units and all other computation is carried out by using R package (version-3.0.3).

7. Conclusion

The goal of present article is to characterize a sum of independent and non-identical set of binomial ranked set samples and compare it with a sum of independent and identical binomial simple random samples for two different situations based on the number of cycles, which is first pre-assumed as a constant integer and when it is a random variable. Our comparison depends only on establishing the variability and their behaviour using some moments. Results show that the sum based on ranked set samples, which is same as that of simple random sample, are more precise and achieve asymptotic normality using comparatively with smaller sample than that of simple random sample. In the context of real-life data study related to child's survivorship in selected eight EAG Indian states, it is found that RSS provides much reliable and accurate estimates than that of SRS for all selected states taken into account.

Acknowledgement

The first author would like to express his deepest gratitude and sincere thanks to the Department of Science & Technology, India (Grant No - IF130365) for funding, while working at Gauhati University. The authors would like to thank the anonymous referees and editorial board for their constructive comments and suggestions to improve the quality of this manuscript.

REFERENCES

AHUJA, J. C., (1970). On the distribution of sum of independent positive binomial variables. Canad. Math. Bull, 13(1), pp. 151–152.

BAI, J., NG, S., (2005)., Tests for skewness, kurtosis, and normality for time series data. Journal of Business & Economic Statistics, 23(1), pp. 49–60.

BARBOUR, A. D., HOLST, L., (1989). Some applications of the Stein-Chen method for proving Poisson convergence. Advances in Applied Probability, 21(1), pp. 74–90.

BROWN, B. W., NEWEY, W. K., (2002). Generalized method of moments, efficient bootstrapping, and improved inference. Journal of Business & Economic Statistics, 20(4), pp. 507–517.

BUTLER, K., STEPHENS, M. A., (2017). The distribution of a sum of independent binomial random variables. Methodology and Computing in Applied Probability, 19(2), pp. 557–571.

CHABCHOUB, Y., FRICKER, C., GUILLEMIN, F., ROBERT, P., (2010). On the statistical characterization of flows in Internet traffic with application to sampling. Computer Communications, 33(1), pp. 103–112.

CHEN, H., (2008). Alternative Ranked set Sample Estimators for the Variance of a Sample Proportion. Applied Statistics Research Progress. Nova Publishers, pp. 35–38.

CHEN, L. H., ROLLIN, A., (2013). Approximating dependent rare events. Bernoulli, 19(4), pp. 1243-1267.

CHEN, H. STASNY E. A., WOLFE, D. A., (2005). Ranked Set Sampling for Efficient Estimation of a Population Proportion. Statistics in Medicine, 24, pp. 3319–3329.

CHEN, H., STASNY, E. A., WOLFE, D. A., (2008). Ranked set sampling for ordered categorical variables. Canadian Journal of Statistics, 36(2), pp. 179–191.

CHEN, H., STASNY, E. A., WOLFE, D. A., (2006). Unbalanced Ranked Set Sampling for estimating a Population Proportion. Biometrics, 62,, pp. 150–158.

CHEN, H., STASNY, E. A., WOLFE, D. A., (2007). Improved Procedures for Estimation of Disease Prevalence Using Ranked Set Sampling. Biometrical Journal, 49(4),, pp. 530–538.

CHEN, H., STASNY, E. A., WOLFE, D. A., MACEACHERN, S. N., (2009). Unbalanced Ranked Set Sampling for Estimating a Population Proportion Under Imperfect Rankings. Communications in Statistics: Theory and Methods, 38(12), pp. 2116–2125.

DAS, R. VERMA, V., NATH, C. N., (2017). Bayesian Estimation of Measles Vaccination Coverage under Ranked Sets Sampling. Statistics in Transition new series, 18(4), pp. 589–608.

FINLAY, J. E., OZALTIN, E., CANNING, D., (2011). The association of maternal age with infant mortality, child anthropometric failure, diarrhoea and anaemia for first births: evidence from 55 low-and middle-income countries. BMJ open, 1(2), e000226.

FREEDMAN, D. A., PETERS, S. C., (1984). Bootstrapping a regression equation: Some empirical results. Journal of the American Statistical Association, 79(385), pp. 97–106.

HASTINGS, W. K., (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1), pp. 97–109.

HORVATH, L., (1989). The limit distributions of likelihood ratio and cumulative sum tests for a change in a binomial probability. Journal of Multivariate Analysis, 31(1), pp. 148–159.

KADANE, J. B., (2016). Sums of possibly associated Bernoulli variables: The Conway-Maxwell-binomial distribution. Bayesian Analysis, 11(2), pp. 403–420.

LING, K. D., (1988). On binomial distributions of order k. Statistics & Probability Letters, 6(4), pp. 247–250.

MALIK, H. J., (1969). Distribution of the sum of truncated binomial variates. Canadian Mathematical Bulletin, 12(3), pp. 334–336.

PERCUS, O. E., PERCUS, J. K., (1985). Probability bounds on the sum of independent nonidentically distributed binomial random variables. SIAM Journal on Applied Mathematics, 45(4), pp. 621–640.

SASON, I., (2013). Entropy bounds for discrete random variables via coupling. In 2013 IEEE International Symposium on Information Theory (pp. 414–418). IEEE.

SELEMANI, M., MWANYANGALA, M. A., MREMA, S., SHAMTE, A., KAJUNGU, D., MKOPI, A., NATHAN, R., (2014). The effect of mother?s age and other related factors on neonatal survival associated with first and second birth in rural, Tanzania: evidence from Ifakara health and demographic surveillance system in rural Tanzania. BMC pregnancy and childbirth, 14(1), p. 240.

SMALL, N. J. H., (1980). Marginal skewness and kurtosis in testing multivariate normality. Applied Statistics, pp. 85–87.

SUNKLODAS, J. K., (2014). On the normal approximation of a binomial random sum. Lithuanian Mathematical Journal, 54(3), pp. 356–365.

TAKACS, L., (1991). A Bernoulli excursion and its various applications. Advances in Applied Probability, 23(3), pp. 557–585.

TERPSTRA, J. T., (2004). On estimating a population proportion via ranked set sampling. Biometrical Journal, 46,, pp. 264–272.

TERPSTRA, J. T., MILLER, Z. A., (2006). Exact Inference for a Population

Proportion Based on a Ranked Set Sample, Communications in Statistics: Simulation and Computation, 35(1), pp. 19–27.

TERPSTRA, J. T., NELSON, E. J., (2005). Optimal Rank Set Sampling Estimates for a Population Proportion, Journal of Statistical Planning and Inference, 127, pp. 309–321.

VERMA, V., NATH, D. C., DAS, R., (2017). Bayesian bounds for population proportion under ranked set sampling. Communications in Statistics-Simulation and Computation, pp. 1–16. DOI: 10.1080/03610918.2017.1387659.

YU, C., ZELTERMAN, D., (2002). Sums of dependent Bernoulli random variables and disease clustering. Statistics & probability letters, 57(4), pp. 363–373.

APPENDIX

INTER-RELATIONSHIP BETWEEN FACTORIAL, RAW AND CENTRAL MOMENTS

$$
\begin{aligned}
\mu'_{[1]} &= \mu'_1 = \mu_1 \\
\mu'_{[2]} &= \mu'_2 - \mu'_1 \; ; \mu'_2 = \mu'_{[2]} + \mu'_1 \; ; \; \mu_2 = \mu'_2 - \mu_1'^2 \\
\mu'_{[3]} &= \mu'_3 - 3\mu'_2 + 2\mu'_1 \; ; \; \mu'_3 = \mu'_{[3]} + 3\mu'_2 - 2\mu'_1 \; ; \; \mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu_1'^3 \\
\mu'_{[4]} &= \mu'_4 - 6\mu'_3 + 11\mu'_2 - 6\mu'_1 \; ; \; \mu'_4 = \mu'_{[4]} + 6\mu'_3 - 11\mu'_2 + 6\mu'_1; \\
&\quad \mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu_1'^2\mu'_2 - 3\mu_1'^4
\end{aligned}
$$

7.1. Case-I: The number of cycles, m, is a known fixed value

Using equation-(3.1) the factorial moments based on simple random sample are given by

$$
\begin{aligned}
\mu'_{[1]}(W) &= \frac{n!}{(n-1)!}p \\
\mu'_{[2]}(W) &= \frac{n!}{(n-2)!}p^2 \\
\mu'_{[3]}(W) &= \frac{n!}{(n-3)!}p^3 \\
\mu'_{[4]}(W) &= \frac{n!}{(n-4)!}p^4
\end{aligned}
$$

Using equation-(3.2) the factorial moments based on ranked set sample are given by

$$
\begin{aligned}
\mu'_{[1]}(Z) &= m\left(\sum_{i=1}^{s} p_{[i]}\right) \\
\mu'_{[2]}(Z) &= m\left(\sum_{i=1}^{s} p_{[i]}\right)\mu'_{[1]} - m\left(\sum_{i=1}^{s} p_{[i]}^2\right) \\
\mu'_{[3]}(Z) &= m\left(\sum_{i=1}^{s} p_{[i]}\right)\mu'_{[2]} - 2m\left(\sum_{i=1}^{s} p_{[i]}^2\right)\mu'_{[1]} + 2m\left(\sum_{i=1}^{s} p_{[i]}^3\right) \\
\mu'_{[4]}(Z) &= m\left(\sum_{i=1}^{s} p_{[i]}\right)\mu'_{[3]} - 3m\left(\sum_{i=1}^{s} p_{[i]}^2\right)\mu'_{[2]} + 6m\left(\sum_{i=1}^{s} p_{[i]}^3\right)\mu'_{[1]} - 6m\left(\sum_{i=1}^{s} p_{[i]}^4\right)
\end{aligned}
$$

$$\mu'_{[k]}(W) = s\sum_{j=0}^{k-1}\frac{(k-1)!}{(k-1-j)!}p^{(j+1)}\left((-1)^j + \frac{\lambda(s-1)!}{(s-1-j)!j!}\right)\mu'_{[k-1-j]}(W). \quad (7.1)$$

7.2. Case-II: The number of cycles, m, is a random variable

Using equation-(3.9) the first four factorial moments based on simple random samples, where $N \sim$ Poisson (λ) ;$N = 1,2,\cdots$, are given by

$$\begin{aligned}
\mu'_{[1]}(W) &= sp(1+\lambda)\\
\mu'_{[2]}(W) &= s\{p(1+\lambda)\mu'_{[1]} + p^2(-1+\lambda(s-1))\}\\
\mu'_{[3]}(W) &= s\{p(1+\lambda)\mu'_{[2]} + 2p^2(-1+\lambda(s-1))\mu'_{[1]} + 2p^3(1+0.5\lambda(s-1)(s-2))\}\\
\mu'_{[4]}(W) &= s\{p(1+\lambda)\mu'_{[3]} + 3p^2(-1+\lambda(s-1))\mu'_{[2]}\\
&\quad +6p^3(1+\frac{\lambda(s-1)(s-2)}{2})\mu'_{[1]} + 6p^3(-1+\frac{\lambda(s-1)(s-2)(s-3)}{3!})\ \}
\end{aligned}$$

$$\mu'_{[k]}(Z) = \lambda\left(\sum_{i=1}^{s}p_{[i]}\right)\mu'_{[k-1]}(Z) + \sum_{j=0}^{k-1}(-1)^j\frac{(k-1)!}{(k-1-j)!}\left(\sum_{i=1}^{s}p_{[i]}^{j+1}\right)\mu'_{[k-1-j]}(Z). \quad (7.2)$$

Using equation-(3.10) the first four factorial moments based on ranked set samples, where $N \sim$ Poisson (λ) ;$N = 1,2,\cdots$, are given by

$$\begin{aligned}
\mu'_{[1]}(Z) &= a + \left(\sum_{i=1}^{s}p_{[i]}\right)\\
\mu'_{[2]}(Z) &= a\mu'_{[1]} + \left(\sum_{i=1}^{s}p_{[i]}\right)\mu'_{[1]} - \left(\sum_{i=1}^{s}p_{[i]}^2\right)\\
\mu'_{[3]}(Z) &= a\mu'_{[2]} + \left(\sum_{i=1}^{s}p_{[i]}\right)\mu'_{[2]} - 2\left(\sum_{i=1}^{s}p_{[i]}^2\right)\mu'_{[1]} + 2\left(\sum_{i=1}^{s}p_{[i]}^3\right)\\
\mu'_{[4]}(Z) &= a\mu'_{[3]} + \left(\sum_{i=1}^{s}p_{[i]}\right)\mu'_{[3]} - 3\left(\sum_{i=1}^{s}p_{[i]}^2\right)\mu'_{[2]} + 6\left(\sum_{i=1}^{s}p_{[i]}^3\right)\mu'_{[1]} - 6\left(\sum_{i=1}^{s}p_{[i]}^4\right)
\end{aligned}$$

where $a = \lambda sp$ and $p = \frac{1}{s}\sum_{r=1}^{s}p_{[r]}$.

Table 1: Mean and relative precision of sum of Binomial variate under SRS and RSS, for the given set size s, m and p.

p	m	s=2			s=4			s=6		
		Mean		RE	Mean		RE	Mean		RE
		SRS	RSS		SRS	RSS		SRS	RSS	
0.1	10	3	3	1.21	3	3	1.32	5	5	1.39
	50	14	14	1.19	18	18	1.23	29	29	1.53
	100	16	16	1.10	42	42	1.23	46	46	1.60
	200	38	38	1.12	83	83	1.36	103	103	1.42
	500	101	101	1.11	196	196	1.29	273	273	1.38
0.2	10	5	5	1.50	7	7	1.41	12	12	1.66
	50	20	20	1.25	42	42	1.98	58	58	1.51
	100	44	44	1.20	89	89	1.66	139	139	2.17
	200	91	91	1.22	157	157	1.48	234	234	1.82
	500	197	197	1.20	375	375	1.48	593	593	1.81
0.3	10	5	5	1.50	12	12	1.83	18	18	1.50
	50	31	31	1.33	60	60	1.99	92	92	2.33
	100	67	67	1.46	120	120	1.63	175	175	1.98
	200	117	117	1.19	247	247	1.54	356	356	2.08
	500	298	298	1.27	597	597	1.70	899	899	2.02
0.4	10	3	3	1.21	23	23	2.51	29	29	2.11
	50	39	39	1.29	75	75	1.60	122	122	2.07
	100	79	79	1.10	168	168	1.79	240	240	2.18
	200	164	164	1.33	297	297	1.77	481	481	2.25
	500	422	422	1.30	792	792	1.91	1205	1205	2.26
0.5	10	10	10	1.56	17	17	2.51	28	28	4.39
	50	49	49	1.33	107	107	2.38	154	154	1.99
	100	108	108	1.24	212	212	1.78	297	297	2.53
	200	204	204	1.43	400	400	1.83	596	596	2.05
	500	508	508	1.38	1000	1000	1.89	1509	1509	2.31

Table 2: Marginal mean and relative precision of sum of Binomial variate under SRS and RSS, for the given set size s, λ and p.

p	λ	s=2			s=4			s=6		
		Mean		RE	Mean		RE	Mean		RE
		SRS	RSS		SRS	RSS		SRS	RSS	
0.1	10	2.2	2.2	1.10	4.4	4.4	1.30	6.6	6.6	1.50
	50	10.2	10.2	1.10	20.4	20.4	1.30	30.6	30.6	1.50
	100	20.2	20.2	1.10	40.4	40.4	1.30	60.6	60.6	1.50
	200	40.2	40.2	1.10	80.4	80.4	1.30	120.6	120.6	1.50
	500	100.2	100.2	1.10	200.4	200.4	1.30	300.6	300.6	1.50
0.2	10	4.4	4.4	1.20	8.8	8.8	1.60	13.2	13.2	1.99
	50	20.4	20.4	1.20	40.8	40.8	1.60	61.2	61.2	2.00
	100	40.4	40.4	1.20	80.8	80.8	1.60	121.2	121.2	2.00
	200	80.4	80.4	1.20	160.8	160.8	1.60	241.2	241.2	2.00
	500	200.4	200.4	1.20	400.8	400.8	1.60	601.2	601.2	2.00
0.3	10	6.6	6.6	1.30	13.2	13.2	1.89	19.8	19.8	2.49
	50	30.6	30.6	1.30	61.2	61.2	1.90	91.8	91.8	2.50
	100	60.6	60.6	1.30	121.2	121.2	1.90	181.8	181.8	2.50
	200	120.6	120.6	1.30	241.2	241.2	1.90	361.8	361.8	2.50
	500	300.6	300.6	1.30	601.2	601.2	1.90	901.8	901.8	2.50
0.4	10	8.8	8.8	1.40	17.6	17.6	2.19	26.4	26.4	2.98
	50	40.8	40.8	1.40	81.6	81.6	2.20	122.4	122.4	3.00
	100	80.8	80.8	1.40	161.6	161.6	2.20	242.4	242.4	3.00
	200	160.8	160.8	1.40	321.6	321.6	2.20	482.4	482.4	3.00
	500	400.8	400.8	1.40	801.6	801.6	2.20	1202.4	1202.4	3.00
0.5	10	11.0	11.0	1.49	22.0	22.0	2.48	33.0	33.0	3.47
	50	51.0	51.0	1.50	102.0	102.0	2.50	153.0	153.0	3.49
	100	101.0	101.0	1.50	202.0	202.0	2.50	303.0	303.0	3.50
	200	201.0	201.0	1.50	402.0	402.0	2.50	603.0	603.0	3.50
	500	501.0	501.0	1.50	1002.0	1002.0	2.50	1503.0	1503.0	3.50

Table 3: State-wise mean, variance, skewness, kurtosis and relative precision of sum of Binomial variate under SRS and RSS, for the given set size s, m and p.

State	p	s	m	Mean	SRS			RSS			RE
					Vari	Skew	Kurt	Vari	Skew	Kurt	
Bihar	0.94	4	140	523	34.56	-0.148	3.018	29.11	-0.108	3.001	1.19
Uttaranchal	0.95	4	75	285	14.25	-0.238	3.050	12.96	-0.204	3.025	1.10
Chhatisgarh	0.92	4	90	328	29.16	-0.152	3.018	22.56	-0.091	2.993	1.29
Jharkhand	0.93	4	100	373	25.18	-0.172	3.025	21.15	-0.126	3.002	1.19
Orissa	0.93	5	60	278	20.39	-0.189	3.029	14.63	-0.090	2.981	1.39
Rajasthan	0.93	5	70	324	24.07	-0.174	3.024	19.34	-0.116	2.997	1.24
Madhya Pr.	0.93	5	100	465	32.55	-0.151	3.019	26.23	-0.101	2.998	1.24
Uttar Pr.	0.91	5	80	367	30.28	-0.152	3.018	23.44	-0.093	2.995	1.29

Table 4: State-wise marginal mean, variance, skewness, kurtosis and relative precision of sum of Binomial variate under SRS and RSS, for the given set size s, λ and p.

State	p	s	λ	Mean	SRS			RSS			RE
					Vari	Skew	Kurt	Vari	Skew	Kurt	
Bihar	0.94	4	140	527.74	1995.18	0.086	3.008	524.20	0.043	3.021	3.81
Uttaranchal	0.95	4	75	288.80	1097.44	0.117	3.014	285.17	0.058	3.041	3.85
Chhatisgarh	0.92	4	90	335.69	1250.82	0.108	3.012	332.23	0.054	3.033	3.76
Jharkhand	0.93	4	100	374.71	1403.58	0.102	3.011	371.22	0.051	3.030	3.78
Orissa	0.93	5	60	284.67	1325.64	0.131	3.017	280.25	0.059	3.049	4.73
Rajasthan	0.93	5	70	328.63	1524.07	0.122	3.015	324.27	0.055	3.042	4.70
Madhya Pr.	0.93	5	100	469.65	2195.13	0.102	3.010	465.26	0.046	3.029	4.72
Uttar Pr.	0.91	5	80	369.56	1697.65	0.114	3.013	365.30	0.052	3.036	4.65

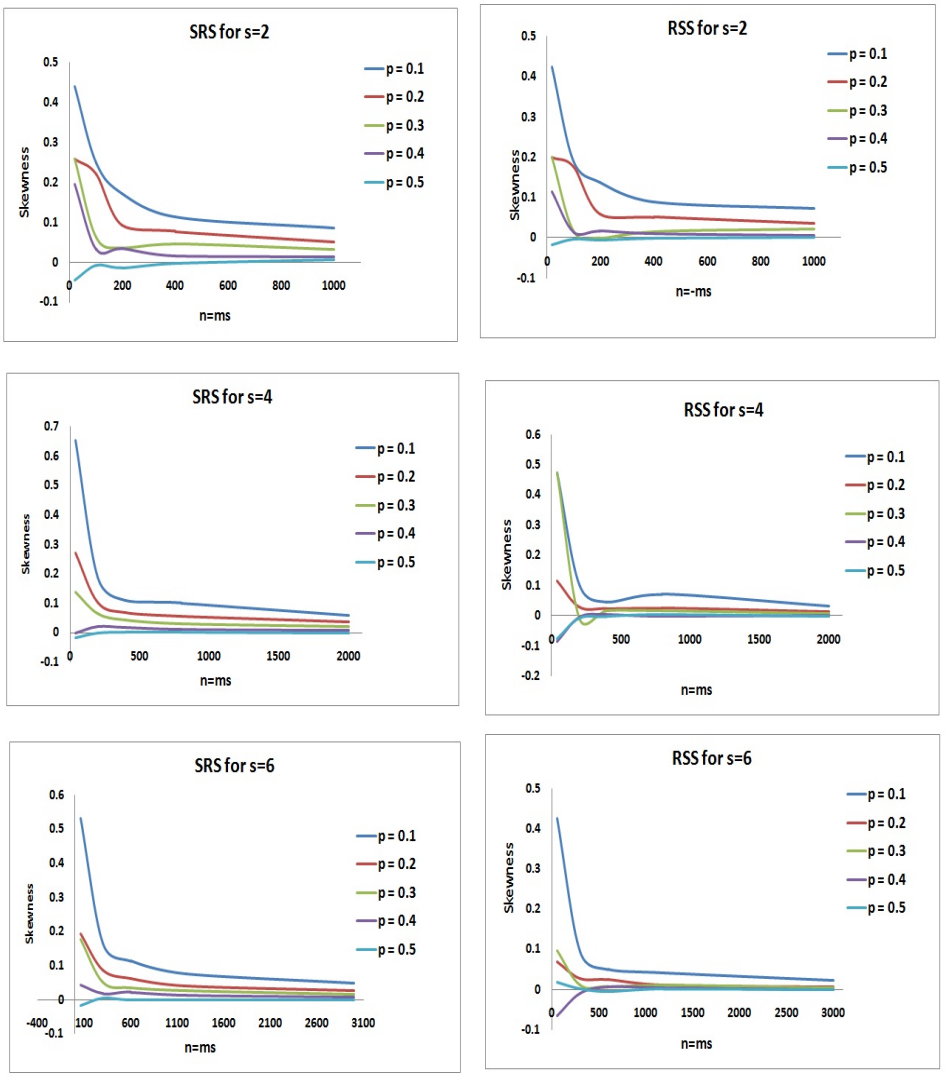


Figure 1: Skewness pattern of sum of Binomial variable under SRS and RSS for set size $s = 2,4,6$ and sample size $n = ms$, where $m = \{10,50,100,200,500\}$, and for fixed p.

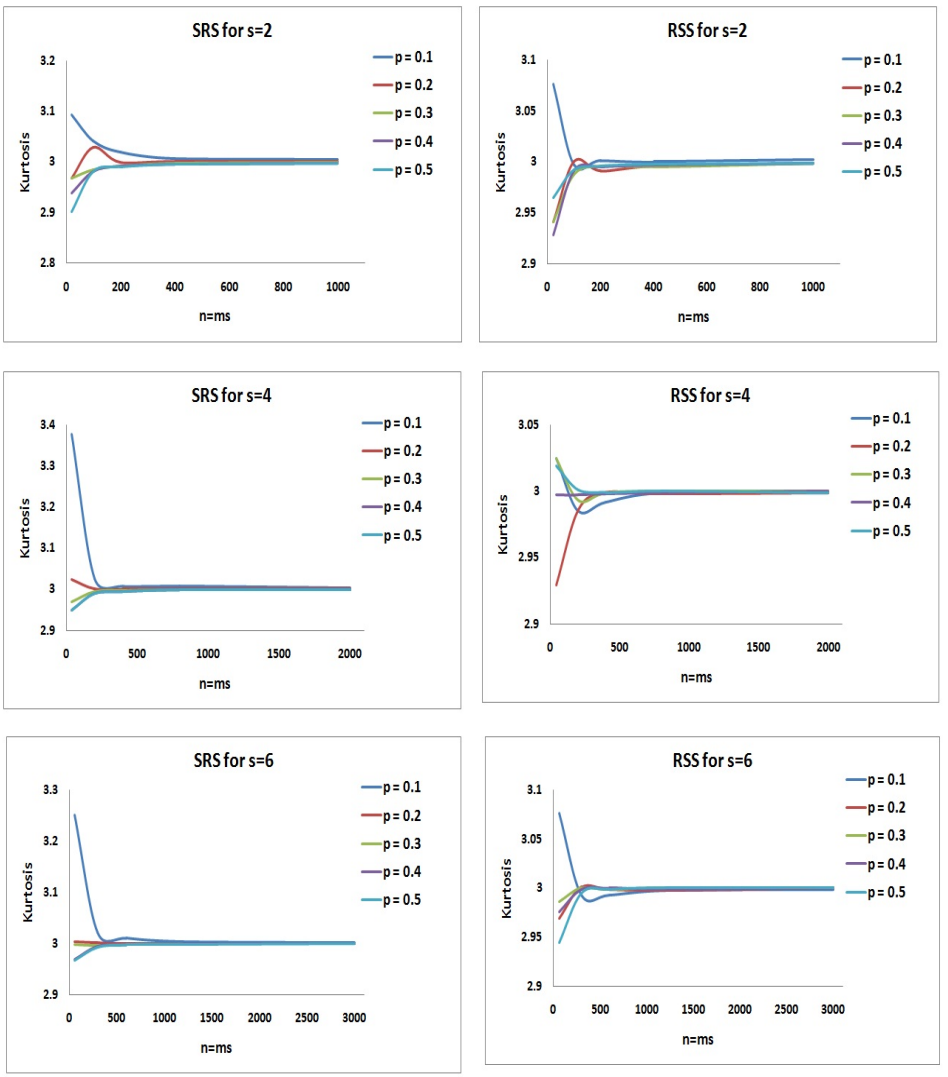


Figure 2: Kurtosis pattern of sum of Binomial variable under SRS and RSS for set size $s = 2,4,6$ and sample size $n = ms$, where $m = \{10,50,100,200,500\}$, and for fixed p.

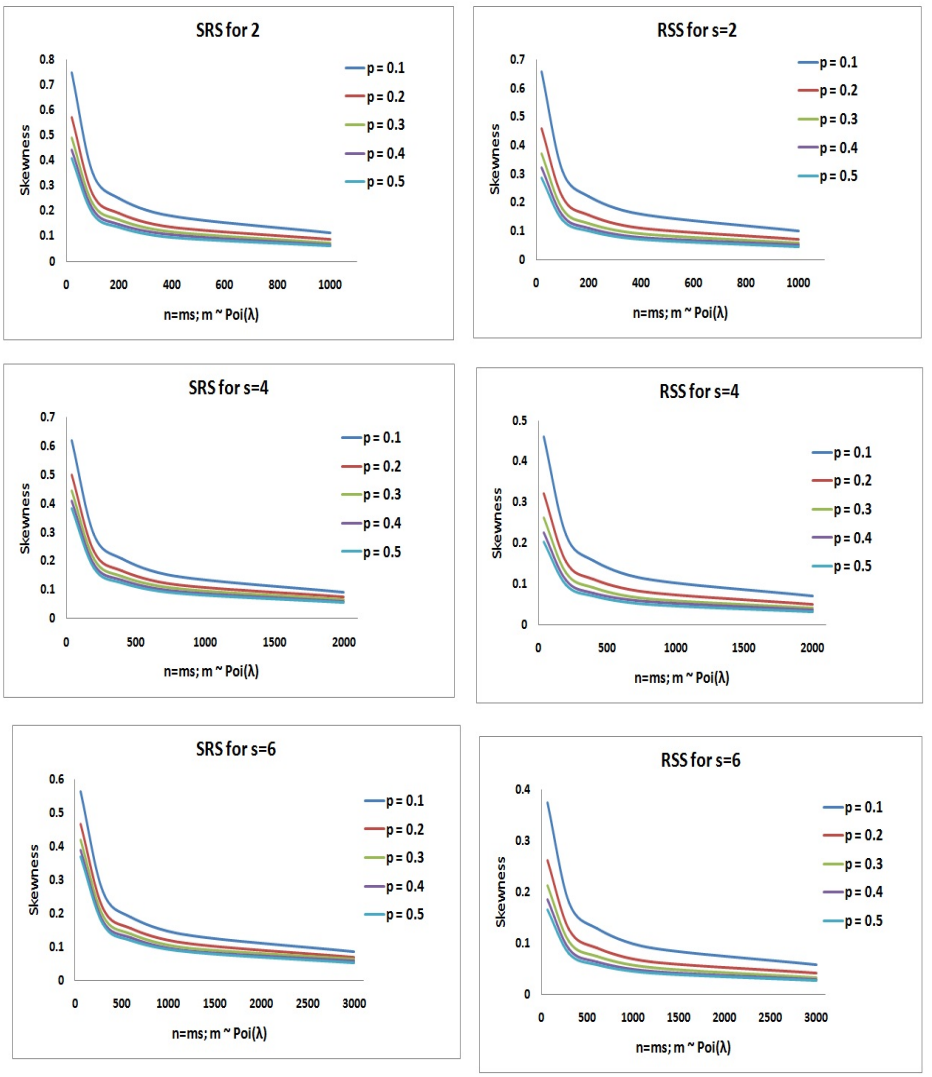


Figure 3: Marginal skewness pattern of sum of Binomial variable under SRS and RSS for set size $s = 2,4,6$ and sample size $n = ms$, where $m \sim Poisson(\lambda)$, $\lambda = \{10,50,100,200,500\}$, and for fixed p.

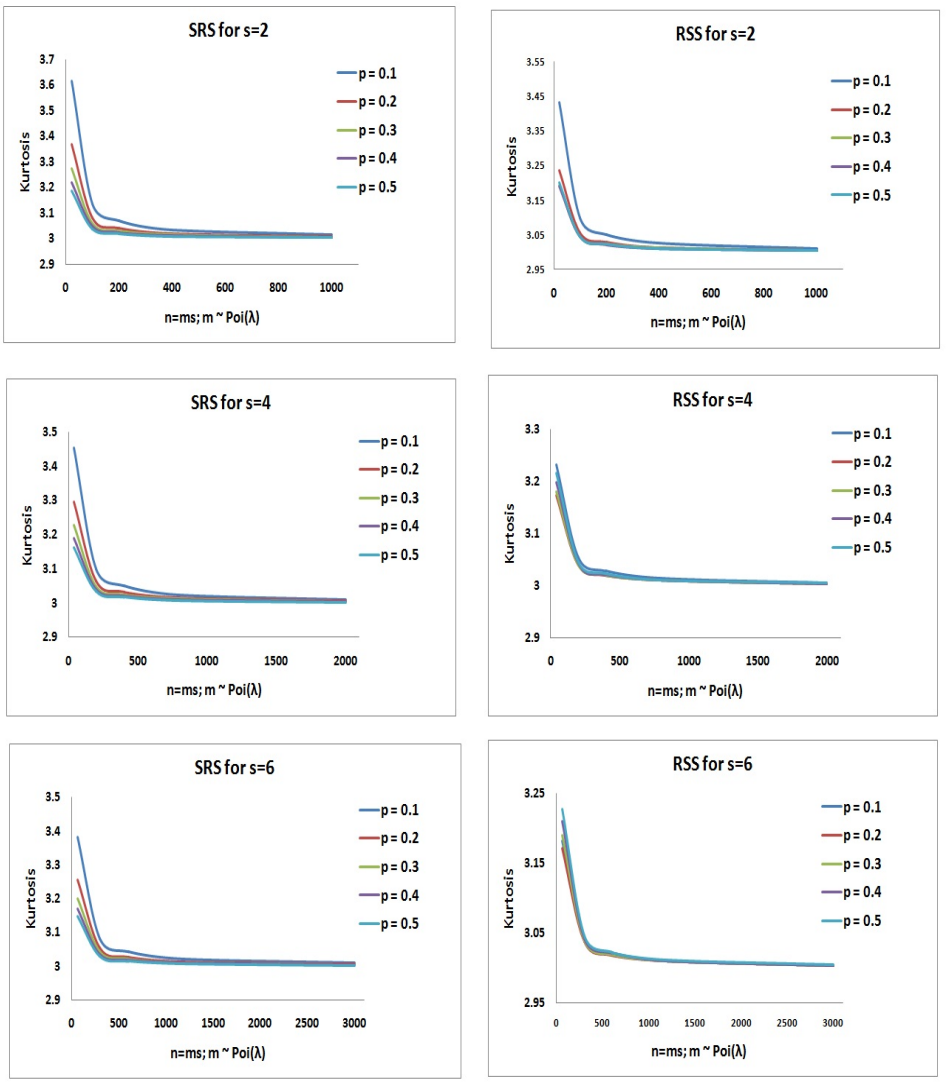


Figure 4: Marginal kurtosis pattern of sum of Binomial variable under SRS and RSS for set size $s = 2,4,6$ and sample size $n = ms$, where $m \sim Poisson(\lambda)$, $\lambda = \{10,50,100,200,500\}$, and for fixed p.

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. 31–56, DOI 10.21307/stattrans-2019-023
Submitted – 20.07.2019; Paper ready for publication – 11.08.2019

OUTLIER DETECTION IN THE ANALYSIS OF NESTED GAGE R&R, RANDOM EFFECT MODEL

Mohammed Abduljaleel[1], **Habshah Midi**[2], **Mostafa Karimi**[3]

ABSTRACT

Measurement system analysis is a comprehensive valuation of a measurement process and characteristically includes a specially designed experiment that strives to isolate the components of variation in that measurement process. Gage repeatability and reproducibility is the adequate technique to evaluate variations within the measurement system. Repeatability refers to the measurement variation obtained when one person repeatedly measures the same item with the same Gage, while reproducibility refers to the variation due to different operators using the same Gage. The two factors factorial design, either crossed or nested factor, is usually used for a Gage R&R study. In this study, the focus is only on the nested factor, random effect model. Presently, the classical method (the method of analysing data without taking into consideration the existence of outliers) is used to analyse the nested Gage R&R data. However, this method is easily affected by outliers and, consequently, the measurement system's capability is also affected. Therefore, the aims of this study are to develop an identification method to detect outliers and to formulate a robust method of measurement analysis of nested Gage R&R, random effect model. The proposed methods of outlier detection are based on a robust mm location and scale estimators of the residuals. The results of the simulation study and real numerical example show that the proposed outlier identification method and the robust estimation method are the most successful methods for the detection of outliers.

Key words: measurement system analysis, mm location, nested Gage R&R, outlier, residuals.

1. Introduction and background

Control is a contentious word that on occasions can be identified with having power (Macintosh and Quattrone, 2010) or training oppression, but in a structural background it has been defined as the ability to create and monitor rules and regulations which should be followed (Ouchi and Maguire, 1975) or on the

[1] Ministry of Electricity, Baghdad, Iraq.
[2] Department of Applied and Computational Statistics, Institute for Mathematical Research, University Putra Malaysia, Serdang, Selangor 43400, Malaysia. E-mail: habshahmidi@gmail.com
[3] Department of Applied and Computational Statistics, Institute for Mathematical Research, University Putra Malaysia, Serdang, Selangor 43400, Malaysia. E-mail: mostafa.karimi.ir@gmail.com

opposing, has been seen as a routine, uninteresting task of observing, supervising, measuring and providing feedback (Reeves and Woodward, 1970). Whatever the definition, the concept is viewed by many as the central nervous system of the processes in every organization.

Montgomery (2007) clarified that the quality control system always has been an integral part of virtually all products and services. However, wakefulness of its importance and the introduction of formal methods for quality control and improvement have been an evolutionary development. An important part of the statistical quality control is the six sigma system (Smith, 1993). Six Sigma is a severe, focused and highly effective application of proven quality principles and techniques. Companies operating at six sigma typically spend less than 5 per cent of their profits fixing problems. In contrast with non-six sigma companies, these costs are often extremely high. Companies operating at three or four sigmas typically spend between 25 and 40 per cent of their profits fixing problems (Pyzdek and Keller, 2014). Based on (Kwak and Anbari, 2006), the authors showed that understanding the key features, impediments, and confines of the six sigma method allow organizations to better support their strategic directions and increasing needs for monitoring and training. Although Six Sigma provides assistance over prior approaches to quality management, it also creates new challenges for researchers and experts (Schroeder et al., 2008).

The important part of the six sigma quality is the measurement system analysis (MSA) used to isolate the variation among devices being measured from the error in the measurement system. The measurement system analysis has been the focus of substantial attention because of its ability to determine the level and range of variation in data. In a process that is important to a measurement system, some variation is likely to occur. The measurement system analysis is an important part of a study that is able to determine the amount of variation (Bourne et al., 2007).

To ensure that the measurement system variability is not adversely large, it is necessary to conduct the measurement system analysis (MSA). Such a study can be conducted in virtually any type of manufacturing industry. According to (He et al., 2011), MSA helps to measure the ability of a Gage or measuring device to produce data that support the analyst's decision-making requirements. Also, MSA is an important section of Six Sigma as well as of the ISO/TS 16949 standards. Burdick et al., (2003) showed that Gage repeatability and reproducibility (Gage R&R) is the most common study in MSA to assess the precision of measurement systems.

Awad et al., (2009) and Peruchi et al., (2013) showed that the repeatability represents the variability from the Gage or measurement tool when it is used to measure the same part (with the same operator or setup or in the same time period), whereas reproducibility reveals the variability arising from different operators, setups, or time periods. As stated by (Grejda et al., 2005; Parker et al., 2005; Piratelli-Filho et al., 2012), some works have used repeatability and/or reproducibility perceptions and ignored Gage R&R statistical analysis in matching measurement system variation to process variation. These studies comprise only Gage variability which are lacking to determine whether the measurement system is able to monitor a particular manufacturing process or not. In case the variation of the measurement system is small relative to the variation of the process, the

measurement system is reflected as acceptable measurements. Furthermore, Gage R&R studies must be performed any time a process is adopted. This is because as the process variation decreases, a once-capable measurement system may now be unqualified. As identified by (Burdick et al., 2003; Wang and Chien, 2010), there are two methods commonly used in the analysis of a Gage R&R study, namely the analysis of variance ANOVA, X bar and R chart. Furthermore, analysts prefer the ANOVA method because it measures the operator-to-part interaction Gage error.

Larsen (2003) extended the univariate Gage R&R study to a common industrial test scenario where multiple features were tested on each device. Providing examples from an industrial application, the author showed that the total yield, false failures and missed false estimates could lead to improvements in the production test process and hence to lower production costs, and finally to customers receiving higher quality products. (Flynn et al., 2009) used regression analysis to analyse the qualified performance capability between two functionally equal but technically different automatic measurement systems. For such accurate measurements as repeatability and reproducibility, the authors found the "pass/fail" criteria for the unit being tested incorrect. Hence, they proposed a methodology based on principal components analysis (PCA) and MANOVA to examine whether there was a statistically significant difference among the measurement systems. He et al., (2011) proposed a PCA-based approach in MSA for the in-process monitoring of all instruments in multisite testing. The approach considers a defective instrument to be one whose statistical distribution of measurements differs significantly from the overall distribution across multiple test tools. Their approach can be implemented as an online monitoring procedure for test instruments so that, until a faulty instrument is identified, production goes continuously. Whereas, Parente et al., (2012) applied univariate and multivariate methods to evaluate the repeatability and reproducibility of the measurement of opposite phase chromatography (RP-HPLC) peptide profiles of excerpts from cheddar cheese. The ability to discriminate different samples was assessed according to the sources of variability in their measurement and analysis procedure. The authors showed that their study had an important impact on the design and analysis of experiments for summarizing of cheese proteolysis. Inferential statistical procedures helped them to analyse the relationships between design variables and proteolysis. In evaluating a measurement system's variation, the most an adequate technique, once an instrument is calibrated, is Gage repeatability and reproducibility Gage R&R (Hoffa and Laux, 2007). The primary purpose of a Gage study is to determine how much variation in the data is due to the measurement system, and whether the measurement system is accomplished by assessing the process performance. The first type of Gage R&R is crossed Gage R&R, which is developed to analyse data from typical measurement system studies. It adopts the most common approach to the appropriate measurement of data with an ANOVA model and evaluates different sources of variation in the measurement system using the variance components in the model. The second type of Gage R&R is the nested Gage R&R, which is developed to measure the system analysis when all operators in the system measure different parts.

2. Mathematical model and measurement guality of nested Gage R&R, random effect model

The nested design is the first option for destructive testing since each operator measures unique parts. If a part can be measured multiple times by different operators, then it is necessary to use the crossed design. In this study, we focus only on the nested design. For the nested experiment, as well as the part is nested within each operator, it is impossible to assess the operator and part interaction. The data of nested Gage R&R is represented as shown in Table 1.

Table 1. The experimental format for nested Gage R&R data

Operator i	Parts $j_{(i)}$	Replication			$\bar{y}$	S^2
1	1	y_{111}	y_{112}	$\ldots\ldots y_{11n}$	$\bar{y}_{11}$	$s^2{}_{11}$
	. .				.	.
	. .				.	.
	. .				.	.
	p	y_{1p1}	y_{1p2}	$\ldots\ldots y_{1pn}$	$\bar{y}_{1p}$	$S^2{}_{1p}$
2	1	y_{211}	y_{212}	$\ldots\ldots y_{21n}$	$\bar{y}_{21}$	$s^2{}_{21}$
	2 .				.	.
	. .				.	.
	. .				.	.
	p	y_{2p1}	y_{2p2}	$\ldots\ldots y_{2pn}$	$\bar{y}_{2p}$	$S^2{}_{2p}$
.						
.						
.						
.						
.						
o	1	y_{o11}	y_{o12}	$\ldots\ldots y_{o1n}$	$\bar{y}_{o1}$	$s^2{}_{o1}$
	2 .				.	.
	. .				.	.
	. .				.	.
	p	y_{op1}	y_{op2}	$\ldots\ldots y_{opn}$	$\bar{y}_{op}$	$S^2{}_{op}$

The analysis of variance, random effect model of nested Gage R&R, is represented in Equation 1.

$$y_{ijk} = \mu + \tau_i + \beta_{j(i)} + \varepsilon_{ijk} \qquad \begin{cases} i = 1,2,3 \ldots\ldots\ldots o \\ j = 1,2,3, \ldots\ldots\ldots p \\ k = 1,2,3, \ldots\ldots\ldots n \end{cases} \tag{1}$$

where

μ is the overall mean

τ_i is the effect for the i_{th}operator, $\tau_i \sim^{iid} N(0, \sigma^2{}_\tau)$

$\beta_{j(i)}$ is the effect of the j_{th} part nested within the i_{th}operator, $\beta_{j(i)} \sim^{iid} N(0, \sigma^2{}_{\beta})$
ε_{ijk} is random error where $\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2)$

The total variation and the total degree of freedom of the nested design random effect model can be partitioned into three components as follows:

$$SS_{Total} = SS_{operator} + SS_{part(operator)} + SS_{error}$$

$$\sum\sum\sum (y_{ijk} - \bar{y}_{...})^2 = \sum\sum\sum (\bar{y}_{i..} - \bar{y}_{...})^2 + \sum\sum\sum (\bar{y}_{ij.} - \bar{y}_{i..})^2 + \sum\sum\sum (y_{ijk} - \bar{y}_{ij.})^2$$

$$\sum\sum\sum (y_{ijk} - \bar{y}_{...})^2 = pn \sum_{i} (\bar{y}_{i..} - \bar{y}_{...})^2 + n \sum_{I} \sum_{j} (\bar{y}_{ij.} - \bar{y}_{i..})^2 + \sum\sum\sum (y_{ijk} - \bar{y}_{ij.})^2$$

Partitioning of Degree of Freedom:

$$opn - 1 = (o - 1) + o(p - 1) + op(n - 1)$$

$$opn - 1 = o - 1 + op - p + opn - op$$

where:

$$y_{i..} = \sum_j \sum_k y_{ijk}; \quad \bar{y}_{i..} = \frac{y_{i..}}{np}$$

$$y_{ij.} = \sum_k y_{ijk}; \quad \bar{y}_{ij.} = \frac{y_{ij.}}{n}$$

$$y_{...} = \sum\sum\sum y_{ijk}; \quad \bar{y}_{...} = \frac{y_{...}}{opn}$$

The Expected Mean Squares of the nested Gage R&R random effect model are presented in Table 2.

Table 2. Expected Mean Squares

Mean Squares	Degree Of Freedom	Expected Mean Squares
MS_o	$o - 1$	$\sigma^2 + np\sigma^2{}_{\tau} + n\sigma^2{}_{\beta(i)}$
$MS_{P(O)}$	$o(p - 1)$	$\sigma^2 + n\sigma^2{}_{\beta(i)}$
MSE	$op(n - 1)$	σ^2

The nested experiment calculations for a total sum of squares (SS_{Total}), the sum of squares of the operator (SS_O), the sum of squares of the part nested within operator ($SS_{part(operator)}$) and the sum of a square of error (SSE) are shown in Table 3.

Table 3. Analysis of variance table for the random effect model for Gage R&R study nested design

Source	S.S	D.F	MS	F
$operator$	$ss_{operator} = pn\sum_{i}(\bar{y}_{i..} - \bar{y}_{...})^2$	$o-1$	$MSo = \frac{ss_{operator}}{o-1}$	$F_O = \frac{MSo}{MS_{P(o)}}$
$Parts_{(Operator)}$	$ss_{part(o)} = n\sum_{i}\sum_{j}(\bar{y}_{ij.} - \bar{y}_{i..})^2$	$o(p-1)$	$MS_{p(o)} = \frac{ss_{part}}{o(p-1)}$	$F_{P(O)} = \frac{MS_{p(o)}}{MSE}$
$Error$	$SSE = \Sigma\Sigma\Sigma(y_{ijk} - \bar{y}_{ij.})^2$	$op(n-1)$	$MSE = \frac{ss_{error}}{op(n-1)}$	
$\boldsymbol{Total}$	$ss_{Total} = \Sigma\Sigma\Sigma(y_{ijk} - \bar{y}_{...})^2$	$opn-1$		

As we mentioned previously, nested Gage R&R is a measurement system analysis whereby the variation in the system is due to repeatability and reproducibility. Repeatability is a variation from a measurement instrument and, on the other hand, reproducibility is a variation from the operators using the instrument (Erdmann et al., 2009).

In this design, the interest is to test for the operator effect and the part (operator) effect. The test on the operator effect is expected to be non-significant that implies operators have no difficulty in making consistent measurements.

The part (operator) is anticipated to be significant, which indicates the ability of the Gage/instrument to distinguish between units of measurement.

The following hypothesis and test statistics in Equation 2 are used to test the operator's effect:

$H_0: {\sigma^2}_O = 0$ (No significant difference between the operator's effects)
$H_1: {\sigma^2}_O > 0$ (Significant difference between the operator's effects)

$$\text{Test statistic: } F = \frac{MS_O}{MS_{P(o)}} \tag{2}$$

The estimated variance of the operator's effect can be formulated as follows:

$E(MS_O) - E\big(MSP(o)\big) = (\sigma^2 + np{\sigma^2}_\tau + n{\sigma^2}_{\beta(i)})$- $(\sigma^2 + n{\sigma^2}_{\beta(i)})$

$MS_O - MS_{P(O)} = \hat{\sigma}^2 + np{\hat{\sigma}^2}_\tau + n{\hat{\sigma}^2}_{\beta(i)} - \hat{\sigma}^2 - n{\hat{\sigma}^2}_{\beta(i)}$

Therefore:

$$\hat{\sigma}^2{}_\tau = \frac{MS_O - MS_{P(O)}}{pn} \tag{3}$$

${\hat{\sigma}^2}_\tau$ is the estimated variance for the operator denoted as ${\hat{\sigma}^2}_{operator}$.

The following Equation (4) is used to test the part's effect

$H_0: {\sigma^2}_{\beta(i)} = 0$ (No significant difference between the parts)

$H_1: {\sigma^2}_{\beta(i)} > 0$ (Significant difference between the parts)

$$\text{Test statistic: } F = \frac{MS_{p(O)}}{MSE} \tag{4}$$

With a similar approach as in Equation (3) and Equation (4), the estimated variance of the parts' effect can be formulated as follows:

$$E\left(MS_{P(O)}\right) - E(MSE) = \sigma^2 + n\sigma^2{}_{\beta(i)} - \sigma^2$$

$$MS_{P(O)} - MSE = \hat{\sigma}^2 + n\hat{\sigma}^2{}_{\beta(i)} - \hat{\sigma}^2 = n\hat{\sigma}^2{}_{\beta(i)}$$

$$\hat{\sigma}^2{}_{\beta(i)} = \frac{MS_{P(O)} - MSE}{n} \tag{5}$$

$\hat{\sigma}^2{}_{\beta(i)}$ is the estimated variance for the part (operator) denoted as $\hat{\sigma}^2{}_{\text{parts(operator)}}$.

$$\hat{\sigma}_{part(operator)} = \sqrt{\hat{\sigma}^2{}_{parts(operator)}} \tag{6}$$

$$\hat{\sigma}^2{}_e = MSE \tag{7}$$

$$\hat{\sigma}^2{}_{repeatability} = MSE = \hat{\sigma}^2;\ \hat{\sigma} = \sqrt{MSE}$$
$$\hat{\sigma}^2{}_{reproducibility} = \hat{\sigma}^2{}_O$$

If the value of the variance components is less than 0, treat them as equal to 0 because variance cannot be negative.

The estimated variance for Gage R&R is given by:

$$\hat{\sigma}^2{}_{Gage\ R\&R} = \hat{\sigma}^2{}_{repeatability} + \hat{\sigma}^2{}_{reproducibility} = \hat{\sigma}^2 + \hat{\sigma}^2{}_o \tag{8}$$

$$\text{The standard deviation of Gage R\&R} = \sqrt{\hat{\sigma}^2{}_{Gage\ R\&R}} \tag{9}$$

Estimated variance of Total variation= $\hat{\sigma}^2{}_{Gage\ R\&R} + \hat{\sigma}^2{}_{part(operator)}$

$$\text{The estimated standard deviation of Total variation} = \sqrt{\hat{\sigma}^2{}_{Gage\ R\&R} + \hat{\sigma}^2{}_{parts(operat}} \tag{10}$$

In the study of Gage R&R design, the Gage capability can be measured by using the precision-to-tolerance ratio (or P/T ratio), as follows:
$\frac{P}{T} = \frac{6\hat{\sigma}_{Gage\ R\&R}}{USL-LSL}$ where $\hat{\sigma}_{GageR\&R}$ is the standard deviation of Gage R&R as stated in Equation (9).

USL and LSL are the upper and lower specification limits of the product under study (given in each nested Gage R&R data). If the P/T ratio is 0.1 or less, this indicates acceptable Gage capability (Headquarters, 2015). But there are clear dangers in relying too much on the P/T ratio, in some nested Gage R&R data. For example, the ratio may be made randomly small by increasing the width of the specification tolerance (Stevens, 2013). As such other measures are employed such as using the percentage contribution of the variance component

of the total variation of Gage R&R and also the percentage contribution of the variation component of part (operator).

$$\text{Percentage contribution of Total variation of Gage R\&R} = \frac{\text{variation of Gage R\&R}}{Total\ variation} \quad (11)$$

The percentage contribution of variation of Gage R&R measures the contribution of the nested Gage R&R in the total variation. The small value of per cent contribution of Gage R&R means adequate Gage. % contribution $< 30\%$ indicates the measurement system is capable.

Another important measure is by using percentage contribution of variance component of part (operator) as follows:

$$\text{Percentage contribution of part (operator)} = \frac{\text{the variance of part (operator)}}{Total\ variation} \quad (12)$$

A high percentage of contribution indicates good measurement, which implies that the measurement system can distinguish between parts. Most of the total variation in the measurement is due to differences between parts, which is desirable.

An equivalent measure of Gage capability is by using the percentage contribution of standard deviation namely:

$$\%\text{contribution of total sd of Gage R\&R} = \frac{standard\ deviation\ of\ Gage\ R\&R}{standard\ deviation\ of\ Total} \quad (13)$$

$$\%\text{contribution of total 6 sd of Gage R\&R} = \frac{6\ standard\ deviation\ of\ Gage\ R\&R}{6\ standarad\ deviation\ of\ Total} \quad (14)$$

$$\%\text{contribution of sd of part(operator)} = \frac{standard\ deviation\ of\ part(operator)}{\text{standard deviation of Total}} \quad (15)$$

3. Methodology

This section presents the methodology of this study. The study proposes a method to identify outliers in nested Gage R&R, namely, a method based on a highly efficient estimator which has a high breakdown point. Moreover, to reduce the negative effect of the outliers, a robust estimation method has been presented to obtain a reliable measurement with low variation after detecting the outliers. Then, two numerical examples and the analysis of the data are presented. Moreover, the simulation study is illustrated.

3.1. Data analysis: how outliers affect the analysis of Gage R&R

In this section, a numerical nested example is presented to show the effect of outliers on the nested Gage R&R measurements.

This data is taken from (Excel, 2013), which described an industrial application whereby heat treating of parts is inducted to perform a Gage R&R analysis on the hardness tester. For the reason of measuring the hardness, a piece of the product is cut, prepared and tested. That piece was altered, so it cannot be retested. It is assured that the parts within operators are homogeneous. For this process, three operators are included in the Gage R&R study, that is operator A, B and C. Each operator is needed to test two parts. But there are not always enough measurements for each operator to test parts from each operator. Based on that, a nested design has been used. Three operators are used and five parts from each operator and two measurements from each part have been taken. The total number of measurements is 30. Table 4 shows the collected data of industrial section.

To see the effect of outliers on the variability's measurements, we purposely contaminate the data with a certain number of outliers. The outliers are created by replacing one observation of each operator by $(maximum\ value\ of\ observation + 10\ sd\ in\ each\ operator)$. The outlier is represented in bold, in Table 4.

Table 4. Nested Gage R&R numerical example of three different operators from the industrial section

Operators	**Parts**									
A	1		2		3		4		5	
	1	2	1	2	1	2	1	2	1	2
	33.4	33.2	32.4	31.7	34.4	34.5	33.9	34.5	34.5	**35.7 (47.4)**
B	6		7		8		9		10	
	1	2	1	2	1	2	1	2	1	2
	32.5	32.1	32.1	32.3	**35.1 (48.1)**	34.7	32.4	33.1	34.8	34.9
C	11		12		13		14		15	
	1	2	1	2	1	2	1	2	1	2
	32.6	32.7	32.3	32.1	**34.9 (47.3)**	34.7	33.0	33.2	31.6	30.9

Nested ANOVA or nested Gage R&R table is represented to show the significance of the parts and the operators and the table of variance components to measure the Gage variation, part-to-part variation and the total variation as shown in Table 5-A. The components of variance and standard deviations contribution are shown in Table 5-B and Table 5-C, respectively. Five useful graphs for the interpretation of the experimental results are displayed in Figure 3 and Figure 4.

Table 5. Gage R&R (nested) for results without and with outliers

A. Nested ANOVA

Source	df	SS		MS		F		p	
		Without outlier	With outlier	Without outlier	With outlier	Without outlier	With outlier	Without outlier	With outlier
Operator	2	5.256	4.741	2.628	2.371	0.793	0.083	0.475	0.921
Part (operator)	12	37.766	342.178	3.133	28.514	26.024	1.687	$10e^{-5}$	0.168
Repeatability	15	1.911	253.455	0.127	16.897				
Total	29	46.932	600.374						

B. Components of variance analysis

Gage R&R Source	Var Comp		%Contribution (of Var Comp)	
	Without outlier	With outlier	Without outlier	With outlier
Total variation of Gage R&R	0.127	16.897	7.41	74.42
Repeatability	0.127	16.897	7.41	74.42
Reproducibility	$10e^{-5}$	$10e^{-5}$	$10e^{-5}$	$10e^{-5}$
$part\ to\ part\ or\ part_{(operator)}$	1.593	5.808	92.61	25.58
Total variation	1.721	22.705	100.00	100.00

C. Components of standard deviation and 6 × standard deviation analysis

Gage R&R Source	StdDev (SD)		Study Var (6 × SD)		%Study Var (%SV)	
	Without outlier	With outlier	Without outlier	With outlier	Without outlier	With outlier
Total s.d of Gage R&R	0.356	4.111	2.141	24.663	27.21	86.27
Repeatability	0.356	4.111	2.141	24.663	27.21	86.27
Reproducibility	$10e^{-5}$	$10e^{-5}$	$10e^{-5}$	$10e^{-5}$	$10e^{-5}$	$10e^{-5}$
$part\ to\ part\ or\ part_{(operator)}$	1.262	2.411	7.573	14.461	96.23	50.58
Total variation	1.311	4.765	7.871	28.591	100.00	100.00

*Specification tolerance (upper specification limit-lower specification limit=8). From the example and process standard deviation is 2.5.

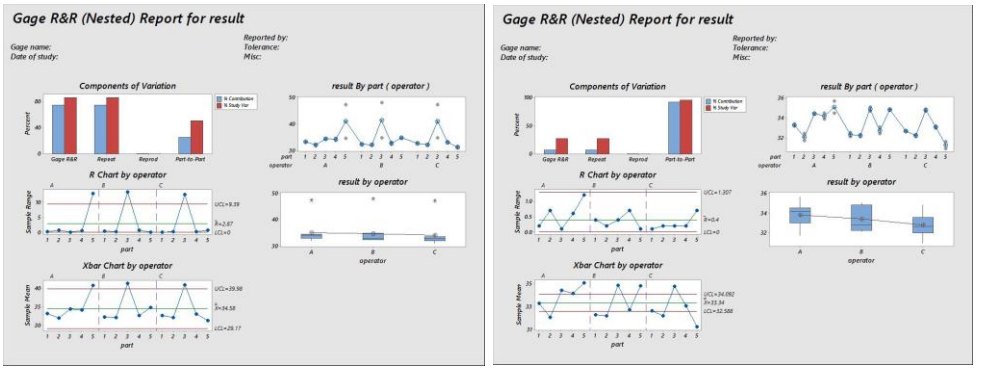


Figure 1. Nested Gage R&R: components of variation, result by part (operator), R chart by the operator, X bar chart by the operator, result by part (operator), and result by the operator without outlier (left) and with outliers (right).

It can be observed from Table 5-A, when there are no outliers there is a significant difference between parts when nesting within operators ($p - value < 0.05$). This results indicate that the Gage is capable of distinguishing between different units. The test on operator suggests that the operator has no difficulty of making consistent measurements ($p - value > 0.05$). These two conditions are desired. Gage R&R studies quantify this by determining the % Gage R&R value. Based on these results, the hardness tester (Gage R&R) is responsible for about 7% of the total variation. This test method appears to be very reliable because the % contribution of Gage R&R variation is less than 30%.

The percentage of contribution for the difference between parts, when nesting within operators ($part - to - part = 96.23$) as shown in Table 5-B, is high, which is close to 100%. The higher percentage contribution for the parts indicates good performance of system Gage R&R. These results can be seen from graphics. The components of the variation graph are placed in the upper left corner in Figure 3 and that means reliable data. Also in Table 5-B, the reproducibility is 0 because all the variations are due to the Gage variation and part-to-part (the part when nesting within operator variation) not due to the interaction between operators and parts.

Most of the variations are due to part-to-part (parts nested within operators) variation, with a low percentage of variation due to errors in the measurement system of Gage R&R at the $\bar{x}$ chart-located in the lower left corner in Figure 1. Most of the points in the $\bar{x}$ chart are outside the control limits when the variation is mostly due to part-to-part variation, 27 points of 30 outside the control limits, i.e. about 90%, the Gage is capable (should be more than 75% is outside the control limits) (Headquarters, 2015).

$$\text{The percentage of contribution for total sd of Gage R\&R } = \frac{(6 \times \text{SD})}{Total\ variation}$$

$= 2.141/7.871 = 27.2\%$. This means that the "spread" of the Gage R&R takes up to 27.2% of the total spread. This result implies that the Gage is acceptable (the Gage variation and spread should be less than 30%) (Headquarters, 2015).

Now, let us focus on the results with outliers of Table 5-A (with outliers). It can be seen that the part (operator) effect is not significant ($p - value > 0.05$). This indicates that the Gage cannot distinguish between different units, which is not desirable.

From Table 5-B (with outlier's columns), in particular, the percentage of contribution for the difference between parts ($part - to - part = 25.58$) is much smaller than the percentage of contribution to the variation of the measurement system 92.61 when there is no outlier. It is noticed that the % contribution of variation of hardness tester (Gage R&R) has increased to 74.42% of the total variation. This indicates that the Gage is not capable in the presence of outliers.

Figure 1 shows that the components of the variation graph are placed in the upper left corner. Most of the variations are due to errors in the measurement variation Gage R&R, with a low percentage of variation due to the part-to-part (parts nested within operators) variation.

We have seen the effect of outliers on the measurement variation of the nested Gage R&R data in the numerical example just presented.

3.2. Outlier identification method

We have seen in the previous section that an outlier has a negative effect on the nested Gage R&R analysis. In this situation, it is very crucial to detect an outlier in the nested Gage R&R model. To the best our knowledge no such work has been devoted to identifying an outlier in the nested Gage R&R, random effect model.

Fearn (2001) and Walsh (2016) developed Cochran's C test to decide if a single estimate of variance (or a standard deviation) is significantly larger than a group of variances. Rousseeuw and Mia Hubert (2011) developed a modified Rousseeuw and Mia Hubert method to identify outliers in univariate data. Tukey (2011) also discussed the method of identifying outliers in such type of data. Bagheri and Midi (2011) noted that the traditional approach of identifying outliers in univariate data is by using T statistics, $T = \frac{x-\bar{x}}{s}$.

3.2.1. Rousseeuw and Mia Hubert method

Rousseeuw and Mia Hubert (2011) proposed a method to detect outliers in a univariate data as presented in the following steps:

- Compute the median of all observations in a data set
- Calculate $MAD = 1.483\ median\ \left|observation - median\right|$
- Calculate $z_i = \frac{observation-median}{MAD}$
- Any value of $|z_i| > 2.5$ is considered as an outlier

This method is denoted as Z_{RM}

3.2.2. Tukey method

Hubert and Vandervieren (2008) defined the Tukey method in the following steps:

- Compute Interquartile Range, $\mathrm{IQR} = Q_3 - Q_1$

where:

$Q_3 = X[3n/4]$
$Q_1 = X[n/4]$
$x = variable\ of\ observation$
$n = sample\ size$

An observation is detected as an outlier if it lies outside the following interval

$$[Q_1 - 1.5\ \mathrm{IQR},\ Q_3 + 1.5\ \mathrm{IQR}]$$

This method is denoted as Int_T.

3.3. Proposed method of outlier detection in nested Gage R&R data

As already mentioned, no specific test is developed to identify outliers in nested Gage R&R. The Rousseeuw and Hubert method and the Tukey method is designed for univariate data. Hence, it can be adopted in the formulation of the detection of outliers in nested Gage R&R, with slight modification.

Instead of using the observed value of x as noted in the Rousseeuw and Hubert method and the Tukey method, the residuals can be computed in this regard. It can be observed that MAD and IQR are used as the scale estimator in the Rousseeuw and Hubert and the Tukey method, respectively.

Even though this estimator is resistant to outliers, its weakness is that it is not reliable under normality assumption (Lee et al., 2007). Another shortcoming of this method is that the use of the median is not very reliable because it has low efficiency under normal errors (Mazlina and Habshah, 2015).

As such, we propose to formulate a new test measure, which is based on highly efficient mm estimator, which has a high breakdown point.

The proposed method is summarized as follows:

Step 1: Perform Analysis of the variance method to nested Gage R&R, random effect model.

Step 2: Compute the fitted value as follows:

Referring to Equation 3.1; $E(y_{ijk}) = \mu$ because:

$\tau_i \sim^{iid} N(0, {\sigma^2}_\tau)$, $\beta_{j(i)} \sim^{iid} N(0, {\sigma^2}_{\beta(\tau)})$ and $\varepsilon_{ijk} \sim^{iid} N(0, \sigma^2)$

Hence, the fitted value is written as $\hat{y}_{ijk} = \hat{\mu} = \bar{y}_{...}$ where $\bar{y}_{...} = \frac{\sum\sum\sum y_{ijk}}{opn}$

Step 3: Compute the residual (e_{ijk}) of each observation as follows:

$$e_{ijk} = y_{ijk} - \bar{y}_{...}$$

Let e_{11}, e_{21},……, e_{opn} be opn residuals represented by $r_1, r_2, \ldots\ldots, r_{opn}$

The location-scale model can be written as follows:

$$r_i = \mu + \sigma\varepsilon_i$$

The mm location and scale of r_i is computed in three steps:

Step i: Use robust S estimator to obtain the initial consistent estimator μ_0 and scale σ_0.

Step ii: Compute m estimate of the scale of the residuals from the initial estimates of the location.

Step iii: Using m estimation method, compute the location and scale of the mm estimates

denoted as $\hat{\mu}_{mm}$ and $\hat{S}_{mm}$.

Step 4: Compute the mm location and scale estimates of the residuals. The mm location and
the scale estimates are chosen because according to [18] they has high breakdown points and high efficiency under normal errors.

Step 5: Compute $T_{mm} = \frac{e_{ijk} - \hat{\mu}_{mm}}{\hat{s}_{mm}}$.

Step 6: Any value of $|T_{mm}| > 2.5$ is declared as an outlier.

3.3.1. Simulation study

In order to assess the performance of our proposed method a simulation study is firstly carried out by considering three operators, five parts and two replicates. For each operator $i = 1,2,3; j = 1,2,3,4,5; k = 1,2$.

$$y_{i1j} \text{ is generated from } N(30,0.1)$$

$$y_{i2j} \text{ is generated from } N(32,0.1)$$

$$y_{i3j} \text{ is generated from } N(34,0.1)$$

$$y_{i4j} \text{ is generated from } N(36,0.1)$$

$$y_{i5j} \text{ is generated from } N(38,0.1)$$

The above process is repeated for various number of operators, parts and samples; such as (3 operators, 5 parts and 2 samples), (3 operators, 10 parts and 2 samples), (4 operators, 5 parts and 2 samples), (5 operators, 6 parts and 2 samples), (5 operators, 8 parts and 2 samples), (6 operators, 10 parts and 2 samples), (10 operators, 12 parts and 2 samples). For each design layout, the data is then contaminated by replacing good observation with a certain number of outliers. The outliers are created by taking the maximum value of each data set +3, 5 and10 standard deviation. The same process is repeated for samples equal to 4. Since in practice it is expensive to collect data, it is not recommended to have more than 5 sample sizes. The proposed method is evaluated based on the number of correct detection of outliers. The number of iterations for each design layout is equal to 1000. The results are presented in Tables (6, 7, 8, 9, 10 and 11).

It can be observed from all tables that the T_{mm} is very successful in detecting outliers in the data set compared to the other two methods. The Rousseeuw and Tukey methods become very poor as the number of outliers increases. Both methods suffer from the masking effect. It is very interesting to see that our proposed method is capable of identifying the correct outliers with no masking effect.

Table 6. Percentage of correct detection of an outlier for 3 standard deviations, $\boldsymbol{n = 2}$, the Rousseeuw method and the Tukey method

Operator (part) and two samples	Number of outliers	Proposed Method. Number of correct detection	Rousseeuw method in percentage	Tukey method in percentage
3(5)	0	100	55.6	73.8
	1	100	53.4	43.8
	2	100	44.6	41.6
	3	100	36.3	19.4
3(10)	0	100	37.7	72.2
	1	100	33.5	52.8
	2	100	32.1	38.3
	3	100	30.9	24.3
	4	100	29.8	12.4
	5	100	19.3	7.8
	6	100	0	0
4(5)	0	100	53.5	72.7
	1	100	45.8	47.4
	2	100	51.2	29.3
	3	100	35.4	19.7
	4	100	23.1	11.9
5(6)	0	100	39.3	74.9
	1	100	36.4	56.6
	2	100	32.2	36.7
	3	100	30.9	23.3
	4	100	30.1	10.3
	5	100	18.5	7.2
	6	100	0	0
5(8)	0	100	48.8	71.4
	1	100	42.6	59.6
	2	100	41.1	45.3
	3	100	40.3	30.7
	4	100	39	15.5
	5	100	26.8	8
	6	100	1.7	0.7
	7	100	0	0
	8	100	0	0
6(10)	0	100	35.4	65.3
	1	100	31.8	61.7
	2	100	29	50
	3	100	28.5	37.1
	4	100	21.9	19.2
	5	100	24.1	13
	6	100	9.5	2.4
	12	100	0	0
10(12)	0	100	43.7	64
	1	100	18.5	45.5
	2	100	10.2	42.9
	3	100	8.3	40.6
	4	100	7.9	38.5
	5	100	7.4	33.5
	6	100	6.6	14.8
	7	100	1.1	0.4
	8	100	0	0
	9	100	0	0
	10	100	0	0
	11	100	0	0
	12	100	0	0
	24	100	0	0

Table 7. Percentage of correct detection of an outlier for 3 standard deviations, $\boldsymbol{n = 4}$, the Rousseeuw method and the Tukey method

Operator (part) and two samples	Number of outliers	Proposed method Number of correct detection	Rousseeuw method in percentage	Tukey method in percentage
3(5)	0	100	49.3	74.9
	1	100	36.3	29.9
	2	100	27.8	16.6
	3	100	13.6	11.5
	4	100	5.5	6.7
	5	100	3.7	3.5
	6	100	0	0
3(10)	0	100	49	73.1
	1	100	34.7	37.2
	2	100	35.1	19.2
	3	100	24.3	9.8
	4	100	11.2	5.6
	5	100	5.1	3.3
	6	100	0.2	0.7
	12	100	0	0
4(5)	0	100	48.8	71.4
	1	100	39.3	29.4
	2	100	31.4	15.9
	3	100	15.6	9.9
	4	100	7.3	7.6
	5	100	2.7	4.9
	6	100	0.3	0.4
	7	100	0	0
	8	100	0	0
5(8)	0	100	47.7	70.7
	1	100	42.6	39.7
	2	100	40.2	21.6
	3	100	26.1	10.1
	4	100	12.6	5
	5	100	6.7	2.3
	6	100	3.2	1.9
	12	100	0	0
6(10)	0	100	43.7	64
	1	100	33.1	42.6
	2	100	29.1	28.8
	3	100	25.9	12.9
	4	100	21.1	6.5
	5	100	12.6	2.7
	6	100	11.8	1.9
	12	100	0	0

Table 8. Percentage of correct detection of an outlier for 5 standard deviations, $\boldsymbol{n = 2}$, the Rousseeuw method and the Tukey method

Operator (part) and two samples	Number of outliers	Proposed method. Number of correct detection	Rousseeuw method in percentage	Tukey method in percentage
3(5)	0	100	58.6	73.8
	1	100	56.9	65.8
	2	100	55.8	53.3
	3	100	46.4	38.4
3(10)	0	100	58.4	68.5
	1	100	58.1	66.5
	2	100	57.9	66.5
	3	100	54.7	64.2
	4	100	50.8	61.9
	5	100	27.8	14.9
	6	100	0	0
4(5)	0	100	53.5	72.7
	1	100	49.8	69.3
	2	100	52	59.1
	3	100	43.5	41.4
	4	100	34.3	22.7
5(6)	0	100	39.3	74.9
	1	100	35.3	74.2
	2	100	30.6	65.6
	3	100	28.9	50.6
	4	100	27	29.1
	5	100	26.3	12.4
	6	100	0	0
5(8)	0	100	48.8	71.4
	1	100	33.1	70.8
	2	100	31.9	69.6
	3	100	30.6	58.8
	4	100	29.7	38.3
	5	100	26.9	32.8
	6	100	23.1	29.6
	7	100	19.5	21.7
	8	100	16.3	18.5
6(10)	0	100	35.4	65.3
	1	100	30.6	61.3
	2	100	20.1	59.6
	3	100	17.7	58.7
	4	100	15.2	40.2
	5	100	13.7	33.2
	6	100	9.8	26.9
	12	100	6.4	15.6
10(12)	0	100	43.7	64
	1	100	28.6	55.4
	2	100	24.9	52.7
	3	100	22.1	51
	4	100	19.3	48.5
	5	100	18.1	43.9
	6	100	15.4	40.3
	7	100	13.9	29.5
	8	100	10.6	21.4
	9	100	6.3	18.2
	10	100	4.5	13.9
	11	100	2.8	5.4
	12	100	1.3	4.6
	24	100	0	0

Table 9. Percentage of correct detection of an outlier for 5 standard deviations, $\boldsymbol{n = 4}$, the Rousseeuw method and the Tukey method

Operator(part) and two samples	Number of outliers	Proposed method. Number of correct detection	Rousseeuw method in percentage	Tukey method in percentage
3(5)	0	100	64.1	72.3
	1	100	32.7	33.5
	2	100	18.7	16.6
3(10)	0	100	63.5	72
	1	100	43.7	46.2
	2	100	30.6	29.3
	3	100	22.1	20.3
	4	100	19.8	15.4
	5	100	16.1	12.9
	6	100	10.5	8.7
	12	100	0	0
4(5)	0	100	62.8	72.3
	1	100	44.3	46.1
	2	100	26.6	25.8
	3	100	22.9	21.3
	4	100	18.9	20.8
	5	100	16.2	18.9
	6	100	10.7	13.2
	7	100	0	0
	8	100	0	0
5(8)	0	100	62.2	70.6
	1	100	40.4	47.9
	2	100	33.4	27.3
	3	100	22.7	19.8
	4	100	14.5	12.3
	5	100	11.8	9.6
	6	100	8.7	5.4
	12	100	0	0
6(10)	0	100	42.7	59.5
	1	100	35.8	45.5
	2	100	31.4	30.9
	3	100	24.7	21.7
	4	100	20.7	14.1
	5	100	16.2	9.8
	6	100	12.2	6.4
	7	100	6.3	4.4
	8	100	3.4	4.1
	9	100	1.7	2.9
	10	100	1.2	0.9
	11	100	0.4	0.7
	12	100	0.3	0.6
	24	100	0	0

Table 10. Percentage of correct detection of an outlier for 10 standard deviations, $\boldsymbol{n = 2}$, the Rousseeuw method and the Tukey method

Operator (part) and two samples	Number of outliers	Proposed method Number of correct detection	Rousseeuw method in percentage	Tukey method in percentage
3(5)	0	100	58.6	73.8
	1	100	58	71.2
	2	100	57.4	62.1
	3	100	54.5	52.4
3(10)	0	100	55.4	72.6
	1	100	54.5	71
	2	100	54.3	64.9
	3	100	52.4	60.6
	4	100	44.9	58.5
	5	100	36.3	51.2
	6	100	28.5	49.8
4(5)	0	100	53.5	72.7
	1	100	52.8	71.8
	2	100	51.9	68.4
	3	100	50.6	55.4
	4	100	42.5	35.7
5(6)	0	100	39.3	74.9
	1	100	37.3	74.7
	2	100	34.3	72.2
	3	100	32.3	61.9
	4	100	30.4	40.9
	5	100	29.5	33.1
	6	100	0	0
5(8)	0	100	48.8	71.4
	1	100	24.9	76.2
	2	100	22.5	74.4
	3	100	20.4	68.1
	4	100	19.3	49.5
	5	100	0	0
	6	100	0	0
	7	100	0	0
	8	100	0	0
6(10)	0	100	35.4	65.3
	1	100	34.6	64.2
	2	100	32.8	61.9
	3	100	29.5	59.7
	4	100	24.6	56.7
	5	100	22.1	52.1
	6	100	19.8	44.9
	12	100	0	0
10(12)	0	100	43.7	64
	0	100	45.3	66.9
	1	100	42.9	63.7
	2	100	37.3	59.8
	3	100	28.5	52.4
	4	100	21.8	49.6
	5	100	19.6	38.5
	6	100	16.5	32.4
	7	100	12.5	30.9
	8	100	9.8	28.4
	9	100	7.2	25.1
	10	100	5.4	22.9
	11	100	3.7	21.4
	12	100	2.1	17.6
	24	100	0	0

Table 11. Percentage of correct detection of an outlier for 10 standard deviations, $\boldsymbol{n = 4}$, the Rousseeuw method and the Tukey method

Operator (part) and two samples	Number of outliers	Proposed method Number of correct detection	Rousseeuw method in percentage	Tukey method in percentage
3(5)	0	100	39.3	74.9
	1	100	38.9	66.6
	2	100	36.8	54.5
	3	100	33.7	40.2
	4	100	29.5	38.4
	5	100	14.6	30.6
	6	100	10.9	25.4
3(10)	0	100	40.5	72.8
	1	100	56.1	54.5
	2	100	50.9	40.2
	3	100	44.5	38.6
	4	100	25.1	35.4
	5	100	23.9	32.4
	6	100	21.2	29.6
	12	100	0	0
4(5)	0	100	48.8	71.4
	1	100	52.2	50.5
	2	100	47.4	33.6
	3	100	32.1	33.9
	4	100	20.6	25.3
	5	100	2.5	6.8
	6	100	1.3	4.1
	7	100	0	0
	8	100	0	0
5(8)	0	100	47.7	70.7
	1	100	46.1	54.3
	2	100	34.3	36.9
	3	100	40.2	43.9
	4	100	28.5	33.7
	5	100	21.9	30.4
	6	100	19.1	28.5
	7	100	12.9	21.4
	8	100	0	0
6(10)	0	100	43.7	64
	1	100	41.6	54.6
	2	100	38.4	45.3
	3	100	36.8	42.1
	4	100	35.2	38.8
	5	100	32.1	36.9
	6	100	29.4	34.6
	7	100	27.5	31.2
	8	100	24.3	29.5
	9	100	21.6	25.9
	10	100	18.7	23.1
	11	100	13.6	21.8
	12	100	10.2	17.3
	24	100	0	0

3.3.2. Proposed method on numerical example

To show the superiority of the proposed method in outlier detection in nested Gage R&R study, a numerical example is presented. This data has been taken from (Erdmann et al., 2009), described in the health section. This features of the quality improvement were to take measurements for a body temperature of patients. The measurement of the temperature has been taken using an ear thermometer. The normal body temperature for any individual range from 35 °C, which is a lower specification limit (LSL), to 40 °C, which is an upper specification limit (USL). The quality of the temperature measurement is assessed through a Gage R&R study. The nurses handling the ear thermometer may cause some variation. The other group of variation for the experiment involved different healthy persons. A single ear thermometer is used by all the nurses. Each patient is measured in the right and left ear. The experiment has been assumed to involve 3 nurses (operators) and each nurse measures 10 different healthy persons, four times. Table 12 shows the collected data of the health section.

To see the effect of outliers on the variability's measurements, we purposely contaminate the data with a certain number of outliers. The outliers are created by replacing one observation of each operator with $(maximum\ value\ of\ observation + 10\ sd\ in\ each\ operator)$. The outlier is represented in bold, in Table 12.

Table 12. Data of the nested Gage R&R experiment showing 3 different operators, 10 parts, and 4 samples in each part

Patients	Operator 1 Jolan				Operator 2 Mariska				Operator 3 Paula			
	1		2		1		2		1		2	
	r	1	r	1	r	1	r	1	r	1	r	1
1	37.3	37.5	37.3	37.5	37.5	37.7	37.3	37.6	37.5	37.6	37.4	37.5
2	37	37.3	36.7	36.8	37.5	37.3	37.4	37.2	37.4	37.4	37.3	37.1
3	36.4	37	37.3	37	37.5	37.3	37.4	37.1	37.6	37.4	37.2	37
4	37.6	37.5	37.6	37.4	37.5	37.5	37.5	37.7	37.7	37.6	37.6	37.5
5	36.7	37.6	**37.8 (41.8**)	37.5	37.9	37.5	37.6	37.6	37.9	37.6	**37.9 (41.2)**	37.8
6	37.5	37.7	37.6	37.3	**38.4 (41.6)**	38	37.8	37.8	37.6	37.9	37.8	37.8
7	37	36.9	37.1	37.3	37.1	37.3	37.4	37.5	37.2	37.4	37.1	37.2
8	37.7	37.4	37.6	37.4	37.6	37.5	37.5	37.1	37.5	37.4	37.2	36.9
9	36.4	36.5	37.6	36.1	37.1	36.9	36.7	36.8	37	36.4	36.9	36.8
10	37.2	37.4	37	37.3	37.1	37.2	37.2	37.2	37.1	37.2	37	37.3

The residuals r_i, the z_i of the Rousseeuw and Mia Hubert method, the interval of the Tukey method and our proposed T_{mm} method are presented in [Table 13 A and Table 13 B].

Table 13. A. Residuals r_i, z_i, the interval of Tukey and T_{mm}

No	Residuals r_i	z_i	Tukey interval	T_{mm}
1	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
2	-0.3157705	-1.3486	(-0.46317,4.3683)	-0.56743098
3	-0.7634017	-3.3715	(-0.46317,4.3683)	-1.37166387
4	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
5	-0.5395861	-2.3601	(-0.46317,4.3683)	-0.96954743
6	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
7	-0.3157705	-1.3486	(-0.46317,4.3683)	-0.56743098
8	0.20646592	1.01146	(-0.46317,4.3683)	0.370840721
9	-0.7634017	-3.3715	(-0.46317,4.3683)	-1.37166387
10	-0.1665601	-0.6743	(-0.46317,4.3683)	-0.29935335
11	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
12	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
13	-0.3157705	-1.3486	(-0.46317,4.3683)	-0.56743098
14	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
15	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
16	0.20646592	1.01146	(-0.46317,4.3683)	0.370840721
17	-0.3903757	-1.6858	(-0.46317,4.3683)	-0.7014698
18	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
19	-0.6887965	-3.0344	(-0.46317,4.3683)	-1.23762506
20	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
21	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
22	-0.5395861	-2.3601	(-0.46317,4.3683)	-0.96954743
23	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
24	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
25	3.26527912	14.8348	(-0.46317,4.3683)	5.866432121
26	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
27	-0.2411653	-1.0115	(-0.46317,4.3683)	-0.43339217
28	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
29	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
30	-0.3157705	-1.3486	(-0.46317,4.3683)	-0.56743098
31	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
32	-0.4649809	-2.0229	(-0.46317,4.3683)	-0.83550861
33	-0.3157705	-1.3486	(-0.46317,4.3683)	-0.56743098
34	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
35	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
36	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
37	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
38	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
39	-0.9872173	-4.383	(-0.46317,4.3683)	-1.77378031
40	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
41	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
42	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
43	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
44	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
45	0.35567632	1.68577	(-0.46317,4.3683)	0.63891835
46	3.11606872	14.1605	(-0.46317,4.3683)	5.598354492
47	-0.2411653	-1.0115	(-0.46317,4.3683)	-0.43339217
48	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
49	-0.2411653	-1.0115	(-0.46317,4.3683)	-0.43339217
50	-0.2411653	-1.0115	(-0.46317,4.3683)	-0.43339217
51	0.20646592	1.01146	(-0.46317,4.3683)	0.370840721
52	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
53	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
54	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
55	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
56	0.43028152	2.02293	(-0.46317,4.3683)	0.772957164
57	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
58	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
59	-0.3903757	-1.6858	(-0.46317,4.3683)	-0.7014698
60	-0.1665601	-0.6743	(-0.46317,4.3683)	-0.29935335

Table 13. B. Residuals r_i, z_i, the interval of Tukey and T_{mm}

No	Residuals r_i	z_i	Tukey interval	T_{mm}
61	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
62	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
63	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
64	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
65	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
66	0.28107112	1.34862	(-0.46317,4.3683)	0.504879535
67	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
68	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
69	-0.5395861	-2.3601	(-0.46317,4.3683)	-0.96954743
70	-0.1665601	-0.6743	(-0.46317,4.3683)	-0.29935335
71	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
72	-0.1665601	-0.6743	(-0.46317,4.3683)	-0.29935335
73	-0.2411653	-1.0115	(-0.46317,4.3683)	-0.43339217
74	0.20646592	1.01146	(-0.46317,4.3683)	0.370840721
75	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
76	0.28107112	1.34862	(-0.46317,4.3683)	0.504879535
77	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
78	-0.2411653	-1.0115	(-0.46317,4.3683)	-0.43339217
79	-0.4649809	-2.0229	(-0.46317,4.3683)	-0.83550861
80	-0.1665601	-0.6743	(-0.46317,4.3683)	-0.29935335
81	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
82	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
83	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
84	0.20646592	1.01146	(-0.46317,4.3683)	0.370840721
85	0.35567632	1.68577	(-0.46317,4.3683)	0.63891835
86	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
87	-0.1665601	-0.6743	(-0.46317,4.3683)	-0.29935335
88	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
89	-0.3157705	-1.3486	(-0.46317,4.3683)	-0.56743098
90	-0.2411653	-1.0115	(-0.46317,4.3683)	-0.43339217
91	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
92	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
93	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
94	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
95	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
96	0.35567632	1.68577	(-0.46317,4.3683)	0.63891835
97	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
98	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
99	-0.7634017	-3.3715	(-0.46317,4.3683)	-1.37166387
100	-0.1665601	-0.6743	(-0.46317,4.3683)	-0.29935335
101	-0.0173497	3.4E-15	(-0.46317,4.3683)	-0.03127572
102	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454
103	-0.1665601	-0.6743	(-0.46317,4.3683)	-0.29935335
104	0.13186072	0.67431	(-0.46317,4.3683)	0.236801906
105	2.81764792	12.8119	(-0.46317,4.3683)	5.062199233
106	0.28107112	1.34862	(-0.46317,4.3683)	0.504879535
107	-0.2411653	-1.0115	(-0.46317,4.3683)	-0.43339217
108	-0.1665601	-0.6743	(-0.46317,4.3683)	-0.29935335
109	-0.3903757	-1.6858	(-0.46317,4.3683)	-0.7014698
110	-0.3157705	-1.3486	(-0.46317,4.3683)	-0.56743098
111	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
112	-0.2411653	-1.0115	(-0.46317,4.3683)	-0.43339217
113	-0.3157705	-1.3486	(-0.46317,4.3683)	-0.56743098
114	0.05725552	0.33715	(-0.46317,4.3683)	0.102763091
115	0.28107112	1.34862	(-0.46317,4.3683)	0.504879535
116	0.28107112	1.34862	(-0.46317,4.3683)	0.504879535
117	-0.1665601	-0.6743	(-0.46317,4.3683)	-0.29935335
118	-0.3903757	-1.6858	(-0.46317,4.3683)	-0.7014698
119	-0.4649809	-2.0229	(-0.46317,4.3683)	-0.83550861
120	-0.0919549	-0.3372	(-0.46317,4.3683)	-0.16531454

It can be observed from Tables [13 A and Table 13 B] that our proposed method can detect the 3 outliers that are purposely placed in the data set. However, the Rousseeuw method detects 8 outliers and the Tukey method detects 15 outliers.

5 Conclusion

Outliers have an adverse effect on the analysis of the nested Gage R&R measurements and give a misleading conclusion. Therefore, they should be detected at the outset before further analysis is carried out. Once the outlier is detected, the management should find out whether this outlier was caused by the parts or operators handling the equipment or it is a true error from random variation. Proper action should be taken if those outliers are due to operators or parts. As such, it is very crucial to have an efficient method of identifying outliers. We propose a new method, T_{mm} in this regard. The simulation study and the numerical example clearly show that our proposed method is able to successfully identify an outlier with no masking effect. Nonetheless, the other two methods are not performing well and suffer from masking effect.

REFERENCES

MACINTOSH, N. B., QUATTRONE, P., (2010). Management accounting and control systems: An organizational and sociological approach: John Wiley Sons.

BOURNE, M., MELNYK, S., FAULL, N., FRANCO-SANTOS, M., KENNERLEY, M., MICHELI, P., GRAY, D., (2007). Towards a definition of a business performance measurement system. International Journal of Operations and Production Management, 27(8), pp. 784–801.

HE, S.-G., WANG, G. A., COOK, D. F., (2011). Multivariate measurement system analysis in multisite testing: An online technique using principal component analysis. Expert Systems with Applications, 38(12), pp. 14602–14608.

BURDICK, R. K., BORROR, C. M., MONTGOMERY, D. C., (2003). A review of methods for measurement systems capability analysis. Journal of Quality Technology, 35(4), p. 342.

AWAD, M., ERDMANN, T. P., SHANSHAL, Y., BARTH, B., (2009). A measurement system analysis approach for hard-to-repeat events. Quality Engineering, 21(3), pp. 300–305.

PERUCHI, R. S., BALESTRASSI, P. P., DE PAIVA, A. P., FERREIRA, J. R., DE SANTANA CARMELOSSI, M., (2013). A new multivariate gage R&R method for correlated characteristics. International Journal of Production Economics, 144(1), pp. 301–315.

GREJDA, R., MARSH, E., VALLANCE, R., (2005). Techniques for calibrating spindles with nanometer error motion. Precision engineering, 29(1), pp. 113–123.

PARKER, D. H., ANDERSON, R., EGAN, D., FAKES, T., RADCLIFF, B., SHELTON, J. W., (2005). Weighing the world's heaviest telescope at eight points with corrections for lifting perturbations. Precision engineering, 29(3), pp. 354–360.

PIRATELLI-FILHO, A., FERNANDES, F. H. T., ARENCIBIA, R. V., (2012). Application of virtual spheres plate for AACMMs evaluation. Precision engineering, 36(2), pp. 349–355.

BURDICK, R. K., BORROR, C. M., MONTGOMERY, D. C., (2003). A review of methods for measurement systems capability analysis. Journal of Quality Technology, 35(4), p. 342.

WANG, F.-K., CHIEN, T.-W., (2010). Process-oriented basis representation for a multivariate gauge study. Computers Industrial Engineering, 58(1), pp. 143–150.

FLYNN, M. J., SARKANI, S., MAZZUCHI, T. A., (2009). Regression analysis of automatic measurement systems. IEEE Transactions on Instrumentation and Measurement, 58(10), pp. 3373–3379.

PARENTE, E., PATEL, H., CALDEO, V., PIRAINO, P., MCSWEENEY, P. L., (2012). RP-HPLC peptide profiling of cheese extracts: a study of sources of variation, repeatability and reproducibility. Food Chemistry, 131(4), pp. 1552–1560.

HOFFA, D. W., LAUX, C. M., (2007). Gauge R&R: an effective methodology for determining the adequacy of a new measurement system for micron-level metrology.

ERDMANN, T. P., DOES, R. J., BISGAARD, S., (2009). Quality quandaries* a gage R&R study in a hospital. Quality Engineering, 22(1), pp. 46–53.

NO, A., COMMITTEE, A. M., (2015). Using the Grubbs and Cochran tests to identify outliers. Analytical Methods, 7(19), pp. 7948–7950.

STEVENS, N. T., STEINER, S. H., BROWNE, R. P., & MACKAY, R. J., (2013). Gauge R&R studies that incorporate baseline information. IIE Transactions, 45(11), pp. 1166–1175.

EXCEL SPC FOR EXCEL, (2013). Access date: 1 January 2017 "Destructive Gage R&R Analysis" https://www.spcforexcel.com/knowledge/measurement-systems-analysis/destructive-gage-rr-analysis.

RELIASOFT CORPORATION., EXPERIMENT DESIGN AND ANALYSIS REFERENCE, (2015), Worldwide Headquarters, p. 379.

LEE, E. F., CZABOTAR, P. E., SMITH, B. J., DESHAYES, K., ZOBEL, K., COLMAN, P. M., FAIRLIE, W. D., (2007). Crystal structure of ABT-737 complexed with Bcl-x L: implications for selectivity of antagonists of the Bcl-2 family. Cell death and differentiation, 14(9), p. 1711.

MIDI, H., ABU BAKAR, N. M., (2015). The Performance of Robust-Diagnostic F in the Identification of Multiple High Leverage Points. Pakistan Journal of Statistics, 31(5).

FEARN, T., THOMPSON, M., (2001). A new test for 'sufficient homogeneity'. Analyst, 126(8), pp. 1414–1417.

WALSH, S. J., MACSIK, Z., WEGRZYNEK, D., KRIEGER, T., BOULYGA, S., (2016). Model diagnostics for detecting and identifying method repeatability outliers in precision studies: application to a homogeneity study under a two-stage nested ANOVA. Journal of Analytical Atomic Spectrometry, 31(3), pp. 686–699.

ROUSSEEUW, P. J., HUBERT, M., (2011). Robust statistics for outlier detection. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), pp. 73–79.

STATISTICS IN TRANSITION new series, September 2019
Vol.20, No. 3, pp. 57–79, DOI 10.21307/stattrans-2019-024
Submitted – 21.02.2018; Paper ready for publication – 28.05.2019

GENERALIZED PARETO DISTRIBUTION BASED ON GENERALIZED ORDER STATISTICS AND ASSOCIATED INFERENCE

Mansoor Rashid Malik, **Devendra Kumar**[1]

ABSTRACT

In this paper, we have considered the generalized Pareto distribution. Various structural properties of the distribution are derived including (quantile function, explicit expressions for moments, mean deviation, Bonferroni and Lorenz curves and Renyi entropy). We have provided simple explicit expressions and recurrence relations for single and product moments of generalized order statistics from the generalized Pareto distribution. The method of maximum likelihood is adopted for estimating the model parameters. For different parameter settings and sample sizes, the simulation studies are performed and compared to the performance of the generalized Pareto distribution.

Key words: generalized order statistics, generalized Pareto distribution, single and product moment, recurrence relations, characterization and maximum likelihood estimation.

1. Introduction

The Pareto distribution has been introduced as a model for the distribution of incomes. It is also used as a model for losses in property and casualty insurance. The Pareto distribution has a heavy right tail behaviour, making it appropriate for including large events in applications such as excess-of-loss pricing[see Arnold (2008) and Verma and Betti (2006)]. The Pareto distribution has probability density function

$$f(x;\alpha,\beta)=\frac{\alpha\beta^{\alpha}}{(x+\beta)^{\alpha+1}};\ \ x>0,\ \alpha,\beta>0,$$

and the corresponding cumulative distribution function is

$$F(x;\alpha,\beta)=1-\left(\frac{\beta}{x+\beta}\right)^{\alpha};\ \ x>0,\ \alpha,\beta>0,$$

where β is a scale parameter and α is the shape parameter. Consider the transformation $Y=X+\beta$ to get another form of the Pareto distribution

$$f(y;\alpha,\beta)=\frac{\alpha\beta^{\alpha}}{y^{\alpha+1}};\ \ \beta\leq y<\infty,\ \alpha,\beta>0.$$

[1]Corresponding Author adress: Department of Statistics, Central University of Haryana, Mahendergarh, India. E-mail: devendrastats@gmail.com. ORCID ID: https://orcid.org/0000-0001-5831-3315.

This study uses the concept of generalized order statistics (GOS), introduced by Kamps (1995), that enables a common approach to several models of ordered random variables, such as ordinary order statistics, record values, progressively Type II censoring order statistics, Pfeifer records and sequential order statistics. The use of such a concept has been steadily growing over the years. Well-known properties of order statistics, progressively censored order statistics and record values can be subsumed, generalized and integrated within the concept of GOS. This concept can be effectively applied, e.g., in reliability theory. The statistical properties and the estimation problems based on generalized order statistics for some lifetime distributions has been studied by several researchers. For instance, Aboeleneen (2010) discussed Bayesian and non-Bayesian estimation methods based on GOS for Weibull distribution. Estimates of the unknown parameters and confidence intervals from progressively type II censoring and record values are obtained. Burkschat (2010) derived the best linear unbiased and best equivariant estimators in location and scale families of GOS from generalized Pareto distribution. Safi and Ahmed (2013) obtained the estimates of the unknown parameters of the Kumaraswamy distribution based on GOS using maximum likelihood method. Recently, Wu et al. (2014) obtained maximum likelihood estimator (MLE) of lifetime performance index for the Burr XII distribution with progressively type II right censored sample and Kim and Han (2014) obtained Bayesian estimators and highest posterior density credible intervals for the scale parameter of Rayleigh distribution based GOS. Also, they derived the Bayesian predictive estimator and the highest posterior density predictive interval for independent future observations. Recently, Kumar and Goyal (2019a, 2019b) obtained the relations for single and product moments of order statistics from power Lindley distribution and generalized lindley distribution respectively. Kumar (2015a, 2015b) and Kumar and Dey (2017a) Kumar and Jain (2018) obtained the relations for moments and moment generating function of type-II exponentiated log-logistic, extended generalized half logistic, extended exponential and power generalized Weibull distribution based on GOS respectively. Recently, Kumar et al. (2017) and Kumar and Dey (2017b) established the relations for order statistics from extended exponential and power generalized Weibull distribution and the reference therein.

The motivation of the paper is twofold: first, to derive the mathematical and statistical properties of this distribution as well as explicit expressions for single and product moments based on GOS of generalized Pareto distribution, and second, to estimate the parameters of the model using maximum likelihood method for different sample sizes and different parameter values for the generalized Pareto distribution, which we think would be of deep interest to applied statisticians.

The remaining of the article is organized as follows. In Section 2, we derive the expressions for survival function, hazard rate function, complete moments, conditional moments, mean deviation, Bonferroni and Lorenz curves, Renyi entropy and quantile function. In Section 3 we derive relations for single and product moments of GOS from generalized Pareto distribution. The obtained relations were used to compute first for moments, variances, skewness and kurtosis of order statistics and

record values. We have also derived the characterization of this distribution by using conditional moments of GOS in Section 4. In Section 5, we derive maximum likelihood estimation of the generalized Pareto distribution. In Section 6, simulations are performed for different sample sizes. Section 7 ends with concluding remarks.

2. Generalized Pareto distribution

The generalized Pareto (GP)distribution was proposed by Pickands (1975). Now it is widely used in analysis of extreme events in the modelling of large insurance claims, and to describe the annual maximum flood at river gauging station.
A random variable X has the GP Distribution with two parameters α and β if it has probability density function (pdf) given by

$$f(x;\alpha,\beta) = \frac{\alpha}{(\beta x+\alpha)^2}\left(\frac{\alpha}{\beta x+\alpha}\right)^{\frac{1}{\beta}-1}, \;\; x>0, \;\alpha,\beta>0 \tag{1}$$

and the corresponding cumulative distribution function (cdf) is

$$F(x;\alpha,\beta) = 1-\left(\frac{\alpha}{\beta x+\alpha}\right)^{\frac{1}{\beta}}, \;\; x>0, \;\alpha,\beta>0 \tag{2}$$

The hazard rate function

$$h(x;\alpha,\beta) = (\beta x+\alpha)^{-1}, \;\; x>0, \;\alpha,\beta>0$$

and the survival function

$$S(x;\alpha,\beta) = \left(\frac{\alpha}{\beta x+\alpha}\right)^{\frac{1}{\beta}}, \;\; x>0, \;\alpha,\beta>0.$$

Note that for GP Distribution defined in (1)

$$\bar{F}(x) = (\beta x+\alpha)f(x). \tag{3}$$

For $\beta>0$, the GP Distribution is known as Pareto type II or Lomax distribution. For $\beta=-1$, GP Distribution reduces uniform distribution on $(0,\alpha)$. As $\beta\to 0$, GP Distribution tends to exponential distribution with scale parameter α. It is well known that the GP Distribution for $\beta>0$, provides reasonably good fit to distributions of income and property values. For more details and some applications of this distribution one may refer to Pickands (1975) and Arnold (1983). Plots of the pdf (Figure 1), hazard function (Figure 2) and survival function (Figure 3), respectively for GP Distribution when $\alpha=1,2,3$ and $\beta=1,2,3$.

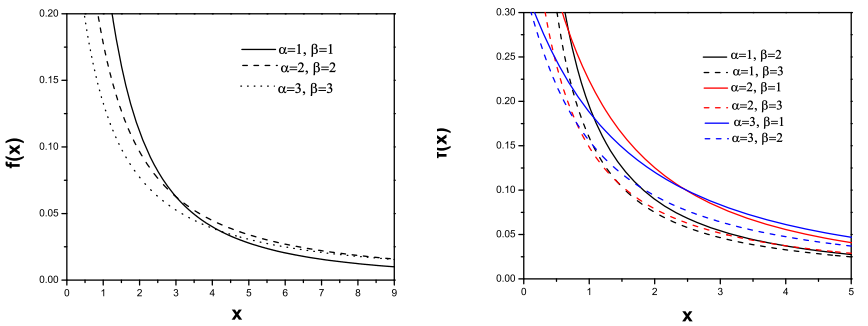


Figure 1: Probability density function of GP Distribution

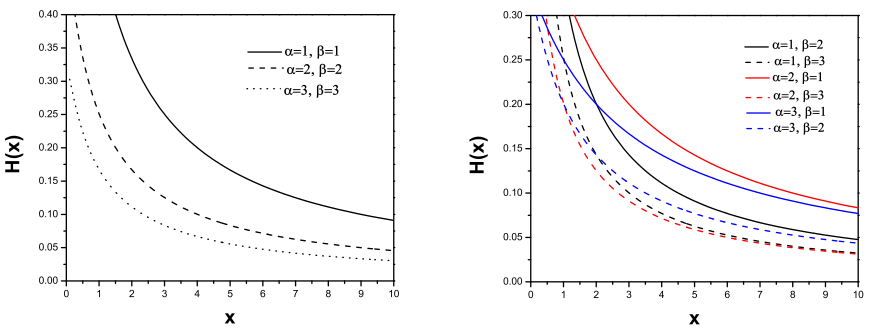


Figure 2: Hazard function of GP Distribution

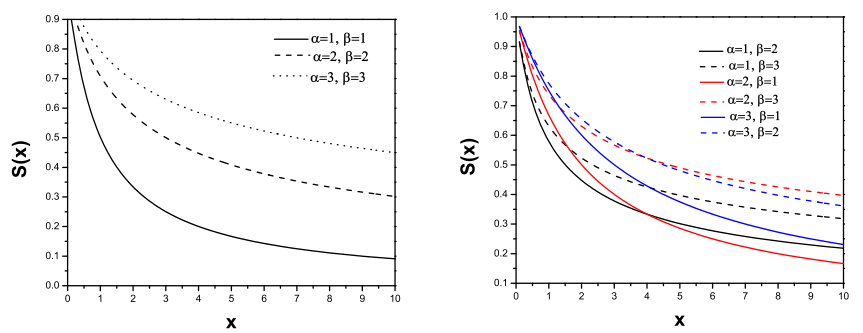


Figure 3: Survival function of GP Distribution

2.1. Quantile function

Let $x_p = Q(p) = F^{-1}(p)$, for $0 < p < 1$ denote the quantile function of the GP Distribution. then

$$x_p = \frac{\alpha[(1-p)^{-\beta} - 1]}{\beta}. \tag{4}$$

In particular, the first three quantiles, Q_1, Q_2 and Q_3, can be obtained by setting $p = 0.25$, $p = 0.5$ and $= 0.75$ in equation (4) respectively.

The effects of the parameters α and β on the skewness and kurtosis can be considered based on quantile measures. The Bowley skewness (Kenney and Keeping 1962) is one of the earliest skewness measures defined by

$$B = \frac{Q(3/4) + Q(1/4) - 2Q(1/2)}{Q(3/4) - Q(1/4)}.$$

Since only the middle two quartiles are considered and the outer two quartiles are ignored, this adds robustness to the measure. The Moors kurtosis (Moors 1988) is defined as

$$M = \frac{Q(3/8) - Q(1/8) + Q(7/8) - Q(5/8)}{Q(6/8) - Q(2/8)}.$$

Clearly, $M > 0$ and there is good concordance with the classical kurtosis measures for some distributions. These measures are less sensitive to outliers and they exist even for distributions without moments. For the standard normal distribution, these measures are 0 (Bowley) and 1.2331 (Moors).

2.2. Moments

Let X be a random variable having the GP Distribution. It is easy to obtain the nth moment of X as the following form

$$\begin{aligned} E(X^k) &= \int_0^\infty x^n f(x)dx = \int_0^\infty x^n \frac{\alpha}{(\beta x+\alpha)^2}\left(\frac{\alpha}{\beta x+\alpha}\right)^{\frac{1}{\beta}-1} dx \\ &= \left(\frac{\alpha}{\beta}\right)^k \sum_{p=0}^{\infty} \frac{(-1)^p \Gamma(k+1)}{p!\Gamma(k+1-p)[\beta(p-k)+1]}. \end{aligned} \tag{5}$$

The variance, skewness and kurtosis of X can be obtained using the relationship

$$Var(X) = E(X^2) - [E(X)]^2$$

$$Skewness(X) = E[X - E(X)]^3/[Var(X)]^{3/2}$$

and

$$Kurtosis(X) = E[X - E(x)]^4/[Var(X)]^2.$$

The variations of $E(X)$, $Var(X)$, $Skewness(X)$ and $Kurtosis(X)$ versus α and β are illustrated in table 1. It appears that $E(X)$, $Var(X)$, $Skewness(X)$ and $Kurtosis(X)$ are increasing function of α for every fixed β. It appears also that the $E(X)$ is greater than its $Var(X)$ for every fixed β.

2.3. Conditional moments

The conditional moments of the GP Distribution, is given by

$$\begin{aligned} E(X^k|X>x) &= \alpha\int_x^\infty \frac{t^k}{(\beta t+\alpha)^2}\left(\frac{\alpha}{\beta t+\alpha}\right)^{\frac{1}{\beta}-1}dt \\ &= \left(\frac{\alpha}{\beta}\right)^k\sum_{p=0}^{\infty}\frac{(-1)^p\Gamma(k+1)}{p!\Gamma(k+1-p)[\beta(p-k)+1]}\left(\frac{\alpha}{\beta x+\alpha}\right)^{p-k+\frac{1}{\beta}}. \end{aligned}$$

The mean residual lifetime function is $E(X|X>x)-x$.

Table 1: Mean, variance, skewness, kurtosis and coefficient of variation for $\beta = 5$ and some values of α

α	Mean	Variance	Skewness	Kurtosis	CV
1	0.025221	0.002242	2.266105	3.254921	4.444709
2	0.050441	0.008969	2.272871	3.269397	8.890585
3	0.075662	0.020179	2.273945	3.272412	13.33496
4	0.100882	0.035875	2.274003	3.272299	17.78067
5	0.126103	0.056055	2.273847	3.272586	22.22588
6	0.151323	0.080719	2.273877	3.272552	26.67109
7	0.176544	0.109867	2.273915	3.272620	31.11604
8	0.201764	0.143500	2.273924	3.272615	35.56135
9	0.226985	0.181618	2.273875	3.272584	40.00661
10	0.252206	0.224219	2.273877	3.272596	44.45156

2.4. Mean deviations

The mean deviations about the mean and the median are used to measure the dispersion and the spread in a population from the centre. The mean deviations about the mean $\mu = E(X)$ and about the median M can be calculated as

$$D(\mu) = E|x-\mu| = \int_0^\infty |x-\mu|f(x)dx$$

and

$$D(m) = E|x-m| = \int_0^\infty |x-m|f(x)dx,$$

respectively. The measures, we obtain $D(\mu)$ and $D(m)$, can be calculated using the following relationships:

$$\begin{aligned} D(\mu) &= \int_0^{\mu}(\mu-x)f(x)dx+\int_{\mu}^{\infty}(x-\mu)f(x)dx \\ &= \mu F(\mu)-\int_0^{\mu}xf(x)dx-\mu[1-F(\mu)]+\int_{\mu}^{\infty}xf(x)dx \\ &= 2\mu F(\mu)-2\mu+2\int_{\mu}^{\infty}xf(x)dx \end{aligned}$$

and

$$\begin{aligned} D(m) &= \int_0^{m}(m-x)f(x)dx+\int_{m}^{\infty}(x-m)f(x)dx \\ &= mF(m)-\int_0^{m}xf(x)dx-m[1-F(m)]+\int_{m}^{\infty}xf(x)dx \\ &= 2mF(m)-m-\mu+2\int_{m}^{\infty}xf(x)dx. \end{aligned}$$

Consider

$$I=\int_{\mu}^{\infty}xf(x)dx. \tag{6}$$

Using the substitution $t=[\bar{F}(x)]^{\beta}$ in (6), we obtain

$$\int_{\mu}^{\infty}xf(x)dx=\frac{\alpha}{\beta(1-\beta)}\left(\frac{\alpha}{\beta\mu+\alpha}\right)^{1/\beta}\left[\left(1+\frac{\beta\mu}{\alpha}\right)+\beta-1\right]$$

and

$$\int_{m}^{\infty}xf(x)dx=\frac{\alpha}{\beta(1-\beta)}\left(\frac{\alpha}{\beta m+\alpha}\right)^{1/\beta}\left[\left(1+\frac{\beta m}{\alpha}\right)+\beta-1\right],$$

so it follows that

$$D(\mu)=2\mu F(\mu)-2\mu+\frac{2\alpha}{\beta(1-\beta)}\left(\frac{\alpha}{\beta\mu+\alpha}\right)^{1/\beta}\left[\left(1+\frac{\beta\mu}{\alpha}\right)+\beta-1\right],$$

and

$$D(m)=2mF(m)-m-\mu+\frac{2\alpha}{\beta(1-\beta)}\left(\frac{\alpha}{\beta m+\alpha}\right)^{1/\beta}\left[\left(1+\frac{\beta m}{\alpha}\right)+\beta-1\right].$$

2.5. Bonferroni and Lorenz curve

Boneferroni and Lorenz curves are proposed by Boneferroni (1930). These curves have applications not only in economics to study income and poverty, but also in

other fields like reliability, demography, insurance and medicine. They are defined as

$$B(p) = \frac{1}{p\mu}\int_0^q xf(x)dx \tag{7}$$

$$L(p) = \frac{1}{\mu}\int_0^q xf(x)dx, \tag{8}$$

and respectively, where $\mu = E(X)$ and $q = F^{-1}(p)$. By using (1), one can reduce (7) and (8) to

$$B(p) = \frac{\alpha}{p\mu(1-\beta)}\left[1-\beta\left(\frac{\alpha}{\beta q+\alpha}\right)^{1/\beta}\left\{\left(1+\frac{\beta q}{\alpha}\right)+\beta-1\right\}\right],$$

and

$$L(p) = \frac{\alpha}{\mu(1-\beta)}\left[1-\beta\left(\frac{\alpha}{\beta q+\alpha}\right)^{1/\beta}\left\{\left(1+\frac{\beta q}{\alpha}\right)+\beta-1\right\}\right],$$

respectively.

2.6. Renyi entropy

The entropy of a random variable X with the density function $f(x)$ is a measure of variation of the uncertainty. Renyi entropy is defined as $I_R(\rho) = (1-\rho)^{-1}log[I(\rho)]$, where $I(\rho) = \int_{\Re} f^{\rho}(x)dx$, $\rho > 0$ and $\rho \neq 1$. If a random variable X has a GP distribution, then, we have

$$\begin{aligned} I(\rho) &= \alpha^{\rho}\int_0^{\infty}\frac{1}{(\beta x+\alpha)^{2\rho}}\left(\frac{\alpha}{\beta x+\alpha}\right)^{\rho\left(\frac{1}{\beta}-1\right)}dx \\ &= \frac{1}{\beta(1+\beta)\alpha^{\rho}}, \end{aligned}$$

[see Gradshteyn and Ryzhik (2014), p-322]. Hence, the Renyi entropy reduces to

$$I_R(\rho) = -\frac{1}{1-\rho}log\beta(\beta+1)+\left(\frac{\rho}{1-\rho}\right)log\alpha.$$

3. Generalized order statistics

The concept of generalized order statistics GOS was introduced by Kamps (1995). Several models of ordered random variables such as order statistics, record values, sequential order statistics, progressive type II censored order statistics and Pfeifer's record values can be discussed as special cases of the GOS. Suppose

$X(1,n,m,k),\ldots,X(n,n,m,k)$, $(k \geq 1, m$ is a real number), are n GOS from an absolutely continuous $cdf\ F(x)$ with $pdf\ f(x)$, if their joint pdf is of the form

$$f_{X(1,n,m,k),\ldots,X(n,n,m,k)}(x_1,x_2,\ldots,x_n)$$
$$= k\left(\prod_{j=1}^{n-1}\gamma_j\right)\left(\prod_{i=1}^{n-1}[1-F(x_i)]^m f(x_i)\right)[1-F(x_n)]^{k-1}f(x_n) \tag{9}$$

on the cone $F^{-1}(0) \leq x_1 \leq x_2 \leq \ldots \leq x_n \leq F^{-1}(1)$,
where $\gamma_j = k+(n-j)(m+1) > 0$ for all $j,\ 1 \leq j \leq n,\ k$ is a positive integer and $m \geq -1$.
If $m = 0$ and $k = 1$, then this model reduces to the ordinary $r-$th order statistic and (9) will be the joint pdf of n order statistics $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ from $cdf\ F(x)$. If $k = 1$ and $m = -1$, then (9) will be the joint pdf of the first n record values of the identically and independently distributed $(i.i.d.)$ random variables with $cdf\ F(x)$ and corresponding $pdf\ f(x)$. In view of (9), the marginal pdf of the $r-$th GOS, $X(r,n,m,k)$, $1 \leq r \leq n$, is

$$f_{X(r,n,m,k)}(x) = \frac{C_{r-1}}{(r-1)!}[\bar{F}(x)]^{\gamma_r-1}f(x)g_m^{r-1}(F(x)), \tag{10}$$

and the joint pdf of $X(r,n,m,k)$ and $X(s,n,m,k)$, $1 \leq r < s \leq n$, $x < y$ is

$$\begin{aligned} f_{X(r,n,m,k),X(s,n,m,k)}(x,y) &= \frac{C_{s-1}}{(r-1)!(s-r-1)!}[\bar{F}(x)]^m f(x)g_m^{r-1}(F(x)) \\ &\times\ [h_m(F(y))-h_m(F(x))]^{s-r-1}[\bar{F}(y)]^{\gamma_s-1}f(y), \end{aligned} \tag{11}$$

where

$$\bar{F}(x) = 1-F(x), \quad C_{r-1} = \prod_{i=1}^{r}\gamma_i\ ,$$

$$h_m(x) = \begin{cases} -\frac{1}{m+1}(1-x)^{m+1}, & m\neq -1 \\ -ln(1-x), & m=-1 \end{cases}$$

and

$$g_m(x) = h_m(x)-h_m(1),\ \ x \in [0,1).$$

3.1. Relations for single moments of generalized order statistics

We shall first establish explicit expressions for jth single moments of the rth generalized order statistics, $E\left(X^j(r,n,m,k)\right)$. For the GP distribution, as given in (1), the $j-$th moments of $X(r,n,m,k)$ is given as,

$$\begin{aligned} E[X^j(r,n,m,k)] &= \int_0^\infty x^j f_{X(r,n,m,k)}(x)dx \\ &= \frac{C_{r-1}}{(r-1)!}\int_0^\infty x^j[\bar{F}(x)]^{\gamma_r-1}f(x)g_m^{r-1}(F(x))dx. \end{aligned} \tag{12}$$

Further, on using the binomial expansion, we can rewrite (12) as

$$E[X^j(r,n,m,k)] = \frac{C_{r-1}}{(r-1)!(m+1)^{r-1}} \sum_{u=0}^{r-1} (-1)^u \begin{pmatrix} r-1 \\ u \end{pmatrix} \times \int_0^\infty x^j [\bar{F}(x)]^{\gamma_{r-u}-1} f(x) dx. \tag{13}$$

Now, letting $t = [\bar{F}(x)]^\beta$ in (13), we get

$$E[X^j(r,n,m,k)] = \frac{C_{r-1}}{(r-1)!(m+1)^r} \left(\frac{\alpha}{\beta}\right)^j \sum_{p=0}^{j} \sum_{u=0}^{r-1} (-1)^{u+p} \begin{pmatrix} r-1 \\ u \end{pmatrix} \begin{pmatrix} j \\ p \end{pmatrix} \times B\left(\frac{k}{m+1} + n - r + u + \frac{\beta(p-j)}{m+1}, 1\right), \tag{14}$$

Since

$$\sum_{a=0}^{b} (-1)^a \begin{pmatrix} b \\ a \end{pmatrix} B(a+k,c) = B(k,c+b), \tag{15}$$

where $B(a,b)$ denotes the complete beta function and defined by $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ Therefore,

$$E[X^j(r,n,m,k)] = \frac{C_{r-1}}{(m+1)^r} \left(\frac{\alpha}{\beta}\right)^j \sum_{p=0}^{j} (-1)^p \begin{pmatrix} j \\ p \end{pmatrix} \times \frac{\Gamma\left(\frac{k+(n-r)(m+1)+\beta(p-j)}{m+1}\right)}{\Gamma\left(\frac{k+n(m+1)+\beta(p-j)}{m+1}\right)} \tag{16}$$

$$= \left(\frac{\alpha}{\beta}\right)^j \sum_{p=0}^{j} (-1)^p \begin{pmatrix} j \\ p \end{pmatrix} \frac{1}{\prod_{a=1}^{r}\left(1 + \frac{\beta(p-j)}{\gamma_a}\right)}, \tag{17}$$

where $\Gamma(.)$ denotes the complete gamma function and defined by $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$.

Special Cases

i) Putting $m = 0$, $k = 1$, in (16), we get moments of order statistics from GP distribution as

$$E[X_{r:n}^j] = \frac{n!}{(n-r)!} \left(\frac{\alpha}{\beta}\right)^j \sum_{p=0}^{j} (-1)^p \begin{pmatrix} j \\ p \end{pmatrix} \frac{\Gamma(n-r+1+\beta(p-j))}{\Gamma(n+1+\beta(p-j))}. \tag{18}$$

ii) Setting $m = -1$ in (17), to get moments of $k-$th record value from GP distribution as;

$$E[X(r,n,-1,k)] = \left(\frac{\alpha}{\beta}\right)^j \sum_{p=0}^{j} (-1)^p \begin{pmatrix} j \\ p \end{pmatrix} \frac{1}{\left(1 + \frac{\beta(p-j)}{k}\right)^r}$$

for upper record values $k = 1$

$$E[X^j_{U(r)}] = \left(\frac{\alpha}{\beta}\right)^j \sum_{p=0}^{j} (-1)^p \begin{pmatrix} j \\ p \end{pmatrix} \frac{1}{[\beta(p-j)]^r}. \tag{19}$$

A recurrence relation for single moment of GOS from cdf (2) can be obtained in the following theorem.

Theorem 1. For the distribution given in (1) and for $2 \leq r \leq n$, $n \geq 2$ $k = 1,2,\ldots,$

$$\left(1 - \frac{j\beta}{\gamma_r}\right) E[X^j(r,n,m,k)] = E[X^j(r-1,n,m,k)] + \frac{j\alpha}{\gamma_r} E[X^{j-1}(r,n,m,k)]. \tag{20}$$

Proof. From (10), we have

$$E[X^j(r,n,m,k)] = \frac{C_{r-1}}{(r-1)!} \int_0^\infty x^j [\bar{F}(x)]^{\gamma_r - 1} f(x) g_m^{r-1}(F(x)) dx.$$

Integrating by parts and using (3) and simplifying the resultant expression we get the result given in (20).

Remark 1: Under the assumption of Theorem 1 with $m = 0$, $k = 1$, we shall deduced the recurrence relations for single moments of ordinary order statistics of the GP distribution

$$\left(1 - \frac{j\beta}{n-r+1}\right) E(X^j_{r:n}) = E(X^j_{r-1:n}) + \frac{j\alpha}{(n-r+1)} E(X^{j-1}_{r:n}).$$

Remark 2: Putting $m = -1$ in Theorem 1 we obtain the recurrence relations for single moments of k record values of the GP distribution.

$$\left(1 - \frac{j\beta}{k}\right) E(X^j_{U(r)}) = E(X^j_{U(r-1)}) + \frac{j\alpha}{k} E(X^{j-1}_{U(r)}).$$

All the tables and figures are made by using R software. The codes of the program are available from the author on request. Table 2-3 lists some numerical values for the first four moments, variances, skewness and kurtosis of order statistics and upper record values from equation (18) and (19) and using numerical integration. The parameter values are taken as $\alpha = 2$ and $\beta = 0.5$. The results in this table show a good agreement between the two methods.

Table 2: First four moments, variances, skewness and kurtosis of some order statistics from equation (18) and via numerical integration

$X_{r:50}\downarrow$	$j-th$ moments $\rightarrow$	$j=1$	$j=2$	$j=3$	$j=4$	Variance	Skewness	Kurtosis
$r=1$	Expression (18)	1.791126	6.076642	20.70228	71.72752	2.868510	0.004433	1.154258
	Numerical	1.787827	6.049550	20.56910	71.71503	2.853225	0.004332	1.226838
$r=5$	Expression (18)	1.171432	2.242526	4.737888	10.80944	0.870273	0.003931	1.879686
	Numerical	1.170642	2.239087	4.726975	10.77690	0.868684	0.003951	1.880719
$r=10$	Expression (18)	0.910562	1.276332	2.024636	3.532368	0.447209	0.012898	2.226165
	Numerical	0.910210	1.275245	2.022027	3.526351	0.446763	0.012913	2.226649
$r=15$	Expression (18)	0.774743	0.896511	1.186052	1.738996	0.296284	0.020178	2.407004
	Numerical	0.774531	0.895981	1.184989	1.736927	0.296083	0.020188	2.407285
$r=20$	Expression (18)	0.687348	0.692336	0.801766	1.033374	0.219889	0.026215	2.521945
	Numerical	0.687203	0.692022	0.801214	1.032429	0.219774	0.026225	2.522204
$r=25$	Expression (18)	0.624858	0.564523	0.588489	0.684613	0.174075	0.031388	2.603020
	Numerical	0.624750	0.564315	0.58816	0.684104	0.174002	0.031396	2.603190
$r=30$	Expression (18)	0.577222	0.476860	0.455715	0.486926	0.143675	0.035917	2.663920
	Numerical	0.577138	0.476712	0.455501	0.486622	0.143624	0.035930	2.664067
$r=35$	Expression (18)	0.539307	0.412941	0.366450	0.364092	0.122089	0.039968	2.711642
	Numerical	0.539238	0.412831	0.366301	0.363895	0.122053	0.039963	2.711900
$r=40$	Expression (18)	0.508172	0.364243	0.303026	0.282557	0.106004	0.043595	2.750593
	Numerical	0.508115	0.364157	0.302918	0.282423	0.105976	0.043604	2.750620
$r=50$	Expression (18)	0.459576	0.294896	0.220099	0.184408	0.083686	0.049953	2.810085
	Numerical	0.459534	0.294840	0.220036	0.184337	0.083669	0.049957	2.810084

Table 3: First four moments, variances, skewness and kurtosis of some upper record values from equation (19) and via numerical integration

$r\downarrow$	$j-th$ moments $\rightarrow$	$j=1$	$j=2$	$j=3$	$j=4$	Variance	Skewness	Kurtosis
1	Expression (19)	1.891126	6.376642	21.70228	73.61752	2.800284	0.004765	1.211188
	Numerical	1.835999	6.226602	21.21964	73.54503	2.85571	0.010477	1.171609
2	Expression (19)	1.328135	3.217226	8.587433	24.52173	1.453283	0.033607	1.712181
	Numerical	1.350241	3.278293	8.764432	25.05541	1.455142	0.027061	1.704100
3	Expression (19)	0.955521	1.682350	3.539371	8.351799	0.76933	0.234001	2.600790
	Numerical	0.964074	1.705065	3.597157	8.504109	0.775626	0.224602	2.575414
4	Expression (19)	0.673323	0.871408	1.451192	2.835484	0.418044	0.622106	3.895456
	Numerical	0.674463	0.878537	1.468775	2.877312	0.423637	0.610888	3.855009
5	Expression (19)	0.467551	0.448047	0.592478	0.960073	0.229443	1.174468	5.628818
	Numerical	0.465080	0.449381	0.597197	0.970923	0.233082	1.159966	5.573523
6	Expression (19)	0.321244	0.229045	0.241045	0.324331	0.125847	1.881817	7.858959
	Numerical	0.317367	0.228555	0.241977	0.326887	0.127833	1.866261	7.795781
7	Expression (19)	0.219019	0.116560	0.097785	0.109353	0.068591	2.760704	10.69797
	Numerical	0.214919	0.115722	0.097766	0.109843	0.069532	2.75115	10.64534
8	Expression (19)	0.148476	0.059104	0.039574	0.036809	0.037059	3.849072	14.31953
	Numerical	0.144723	0.058385	0.039407	0.03685	0.03744	3.856723	14.30993
9	Expression (19)	0.100229	0.029885	0.015984	0.012373	0.019839	5.200595	18.96155
	Numerical	0.097047	0.029374	0.015853	0.012345	0.019956	5.243194	19.04621
10	Expression (19)	0.067447	0.015077	0.006446	0.004154	0.010528	6.886594	24.93712
	Numerical	0.064874	0.014745	0.006367	0.004131	0.010536	6.987957	25.20555

3.2. Relations for product moments of generalized order statistics

We shall first establish explicit expressions for the product moment of ith and jth generalized order statistics, $E\left(X^{(i,j)}_{r,s,n,m,k}\right) = \mu^{(i,j)}_{r,s,n,m,k}$. For GP distribution, the product moment of $X(r,n,m,k)$ and $X(s,n,m,k)$ is given as

$$E[X^i(r,n,m,k),X^j(s,n,m,k)] = \int_0^\infty \int_x^\infty x^i x^j f_{X(r,n,m,k)X(s,n,m,k)}(x,y)dxdy.$$

On using (11) and binomial expansion, we have

$$E[X^i(r,n,m,k),X^j(s,n,m,k)] = \frac{C_{s-1}(m+1)^{2-s}}{(r-1)!(s-r-1)!}\sum_{u=0}^{r-1}\sum_{v=0}^{s-r-1}(-1)^{u+v}$$
$$\times \binom{r-1}{u}\binom{s-r-1}{v}\int_0^\infty x^j[\bar{F}(x)]^{(s-r+u-v)(m+1)-1}f(x)G(x)dx, \qquad (21)$$

where

$$G(x) = \int_x^\infty x^j[\bar{F}(y)]^{\gamma_{s-v}-1}f(y)dy. \qquad (22)$$

By setting $t = [\bar{F}(y)]^\beta$ in (22), we obtain

$$G(x) = \left(\frac{\alpha}{\beta}\right)^j \sum_{p=0}^{j}(-1)^p\binom{j}{p}\frac{[\bar{F}(x)]^{\gamma_{s-v}+\beta(p-j)}}{[\gamma_{s-v}+\beta(p-j)]}.$$

On substituting the above expression of $G(x)$ in (22), and simplifying the resulting equation, we get.

$$\begin{aligned} E[X^i(r,n,m,k),X^j(s,n,m,k)] &= \frac{C_{s-1}}{(r-1)!(s-r-1)!(m+1)^s}\left(\frac{\alpha}{\beta}\right)^{i+j}\sum_{p=0}^{j}\sum_{q=0}^{i}(-1)^{p+q} \\ &\times \binom{j}{p}\binom{i}{q}B\left(\frac{k}{m+1}+n-r+\frac{\beta(p+q-i-j)}{m+1},r\right) \\ &\times B\left(\frac{k}{m+1}+n-s+\frac{\beta(p-j)}{m+1},s-r\right), \qquad (23) \end{aligned}$$

which after simplification yields

$$\begin{aligned} E[X^i(r,n,m,k),X^j(s,n,m,k)] &= \left(\frac{\alpha}{\beta}\right)^{i+j}\sum_{p=0}^{j}\sum_{q=0}^{i}(-1)^{p+q}\binom{j}{p}\binom{i}{q} \\ &\times \frac{1}{\prod_{a=1}^{r}\left(1+\frac{\beta(p+q-i-j)}{\gamma_a}\right)\prod_{b=r+1}^{s}\left(1+\frac{\beta(p-j)}{\gamma_b}\right)}. \qquad (24) \end{aligned}$$

Special cases

i) Putting $m = 0$, $k = 1$ in (23), we shall deduced the explicit formula for product

moments of ordinary order statistics of GP distribution.
ii) Setting $m=-1$ in (24), we obtain the explicit expression for product moments of k record values of GP distribution.

Making use of (3), we can derive recurrence relations for product moments of GOS from (11).

Theorem 2. For the distribution given in (1) and for $1 \le r < s \le n,\ n \ge 2$ and $k = 1,2...$

$$\left(1-\frac{j\beta}{\gamma_s}\right) E[X^i(r,n,m,k)X^j(s,n,m,k)] = E[X^i(r,n,m,k)X^j(s-1,n,m,k)] + \frac{j\alpha}{\gamma_s} E[X^i(r,n,m,k)X^{j-1}(s,n,m,k)]. \quad (25)$$

Proof: Using (11), we have

$$E[X^i(r,n,m,k)X^j(s,n,m,k)] = \frac{C_{s-1}}{(r-1)!(s-r-1)!} \times \int_0^\infty x^i[\bar{F}(x)]^m f(x) g_m^{r-1}(F(x)) I(x), dx \quad (26)$$

where

$$I(x) = \int_x^\infty y^j [h_m(F(y)) - h_m(F(x))]^{s-r-1} [\bar{F}(y)]^{\gamma_s - 1} f(y) dy.$$

Solving the integral in $I(x)$ by parts and using (3) and substituting the resulting expression in (26), we get the result given in (25).
Remark 3 Under the assumption of Theorem 2 with $m=0$, $k=1$ we shall deduced the recurrence relations for product moments of order statistics of the GP distribution.
Remark 4 Putting $m=-1$ in Theorem 2 we obtain the recurrence relations for product moments of $k-$th record values from GP distribution.
Remark 5 At $j=0$ in (25), we have

$$E[X^i(r,n,m,k)] = \left(\frac{\alpha}{\beta}\right)^i \sum_{q=0}^{i} (-1)^q \binom{i}{q} \frac{1}{\prod_{a=1}^{r}\left(1+\frac{\beta(q-i)}{\gamma_a}\right)}.$$

Remark 6 At $i=0$, Theorem 2 reduces to Theorem 1.

4. Characterization

Let $X(r,n,m,k),\ r=1,2,\ldots,n$ be GOS, then from a continuous population with cdf $F(x)$ and pdf $f(x)$, then the conditional pdf of $X(s,n,m,k)$ given $X(r,n,m,k)=x$,

$1 \leq r < s \leq n$, in view of (10) and (11), is

$$f_{X(s,n,m,k)|X(r,n,m,k)}(y|x) = \frac{C_{s-1}}{(s-r-1)!C_{r-1}} \times \frac{[h_m(F(y)) - h_m(F(x))]^{s-r-1}[F(y)]^{\gamma_s-1}}{[\bar{F}(x)]^{\gamma_{r+1}}} f(y). \ x < y \quad (27)$$

Theorem 3: Let X be a non-negative random variable having an absolutely continuous distribution function $F(x)$ with $F(0) = 0$ and $0 < F(x) < 1$ for all $x > 0$, then

$$E[X(s,n,m,k)|X(l,n,m,k) = x] = \frac{(\beta x + \alpha)}{\beta} \left\{ \prod_{j=1}^{s-l} \left(\frac{\gamma_{l+j}}{\gamma_{+j} - \beta} \right) - \alpha \right\}, \ l = r, \ r+1 \quad (28)$$

if and only if

$$F(x;\alpha,\beta) = 1 - \left(\frac{\alpha}{\beta x + \alpha} \right)^{\frac{1}{\beta}} \quad x > 0, \ \alpha, \beta > 0.$$

Proof. From (27), we have

$$E[X(s,n,m,k)|X(r,n,m,k) = x] = \frac{C_{s-1}}{(s-r-1)!C_{r-1}(m+1)^{s-r-1}} \int_x^\infty y \left(\frac{\bar{F}(y)}{\bar{F}(x)} \right)^{\gamma_s-1} \times \left[1 - \left(\frac{\bar{F}(y)}{\bar{F}(x)} \right)^{m+1} \right]^{s-r-1} \frac{f(y)}{\bar{F}(x)} dy. \quad (29)$$

By setting $u = \frac{\bar{F}(y)}{\bar{F}(x)}$ from (2) in (29), we obtain

$$E[X(s,n,m,k)|X(r,n,m,k) = x] = \frac{C_{s-1}}{\beta(s-r-1)!C_{r-1}(m+1)^{s-r-1}} [(\beta x + \alpha)A_1 - \alpha A_2], \quad (30)$$

where

$$A_1 = \int_0^1 u^{\gamma_s - \beta - 1}(1 - u^{m+1})^{s-r-1} du \quad (31)$$

and

$$A_2 = \int_0^1 u^{\gamma_s - 1}(1 - u^{m+1})^{s-r-1} du. \quad (32)$$

Again by setting $t = u^{m+1}$ in (31) and (32) and substituting the values of A_1 and A_2 in (30) and simplifying the resultant expression we get the result given in (28).

To prove sufficient part, we have from (27) and (28)

$$\frac{C_{s-1}}{(s-r-1)!C_{r-1}(m+1)^{s-r-1}}\int_x^\infty y[(\bar{F}(x))^{m+1}-(\bar{F}(y))^{m+1}]^{s-r-1} \times[\bar{F}(y)]^{\gamma_s-1}f(y)dy=[\bar{F}(x)]^{\gamma_{r+1}}H_r(x), \tag{33}$$

where

$$H_r(x)=\frac{(\beta x+\alpha)}{\beta}\left\{\prod_{j=1}^{s-r}\left(\frac{\gamma_{r+j}}{\gamma_{r+j}-\beta}\right)-\alpha\right\}.$$

Differentiating (33) both sides with respect to x and rearranging the terms, we get

$$-\frac{C_{s-1}[\bar{F}(x)]^m f(x)}{(s-r-2)!C_{r-1}(m+1)^{s-r-2}}\int_x^\infty y[(\bar{F}(x))^{m+1}-(\bar{F}(y))^{m+1}]^{s-r-2} \times[\bar{F}(y)]^{\gamma_s-1}f(y)dy=H_r'(x)[\bar{F}(x)]^{\gamma_{r+1}}-\gamma_{r+1}H_r(x)[\bar{F}(x)]^{\gamma_{r+1}-1}f(x)$$

Therefore,

$$\frac{f(x)}{\bar{F}(x)}=-\frac{H_r'(x)}{\gamma_{r+1}[H_{r+1}(x)-H_r(x)]}=\frac{1}{(\beta x+\alpha)},$$

which proves that

$$F(x;\alpha,\beta)=1-\left(\frac{\alpha}{\beta x+\alpha}\right)^{\frac{1}{\beta}} \quad x>0,\ \alpha,\beta>0.$$

5. Estimation of model parameters

In this section we discuss the process of obtaining the maximum likelihood estimators of the parameters α and β. Let $X_1,X_2,\ldots,X_n$ be random sample with observed values $x_1,x_2,\ldots,x_n$ from GP distribution. Let $\Theta=(\alpha,\beta)$ be the parameter vector. The likelihood function based on the random sample of size n is obtained from

$$L(\alpha,\beta|x)=\alpha^{n/\beta}\prod_{i=1}^{n}(\beta x_i+\alpha), \tag{34}$$

The maximum likelihood estimates are the values of α and β that maximize this likelihood function.

5.1. Maximum likelihood estimation

The log likelihood function $l(\alpha,\beta|x) = logL(\alpha,\beta|x)$, dropping terms that do not involve α and β, is

$$l(\alpha,\beta|x) = \frac{n}{\beta} log\alpha - \left(1+\frac{1}{\beta}\right)\sum_{i=1}^{n} log(\beta x_i+\alpha). \tag{35}$$

We assume that the parameters α and β are unknown. To obtain the normal equations for the unknown parameters, we differentiate (35) partially with respect to α and β and equate to zero. The resulting equations are

$$0 = \frac{\partial l(\alpha,\beta|x)}{\partial\alpha} = \frac{n}{\alpha\beta} - \left(1+\frac{1}{\beta}\right)\sum_{i=1}^{n}\frac{1}{(\beta x_i+\alpha)}, \tag{36}$$

and

$$0 = \frac{\partial l(\alpha,\beta|x)}{\partial\beta} = -\frac{n}{\beta^2} log\alpha + \frac{1}{\beta^2}\sum_{i=1}^{n} log(\beta x_i+\alpha) - \left(1+\frac{1}{\beta}\right)\sum_{i=1}^{n}\frac{x_i}{(\beta x_i+\alpha)}. \tag{37}$$

The solutions of the above equations are the maximum likelihood estimators of the GP distribution parameters α and β, denoted as $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$, respectively. As the equations expressed in (36) and (37) cannot be solved analytically, one must use a numerical procedure to solve them.

5.2. Approximate confidence intervals

In this section, we present the asymptotic confidence intervals for the parameters of the GP distribution. Since the $MLEs$ of the unknown parameters α and β cannot be derived in closed form, it is not easy to derive the exact distributions of the $MLEs$. Hence, we cannot obtain exact confidence intervals for the parameters. We must use the large sample approximation. It is known that the asymptotic distribution of the $MLEs$ is $[\sqrt{n}(\hat{\alpha}_{MLE}-\alpha), \sqrt{n}(\hat{\beta}_{MLE}-\beta)] \to N_2(0, I^{-1}(\Theta))$, we can refer to Lawless (1982), where $I^{-1}(\Theta)$, the inverse of the observed information matrix of the unknown parameters $\Theta = (\alpha,\beta)$, is

$$I^{-1}(\Theta) = \begin{pmatrix} -\frac{\partial^2 l(\alpha,\beta)}{\partial^2\alpha} & -\frac{\partial^2 l(\alpha,\beta)}{\partial\alpha\partial\beta} \\ -\frac{\partial^2 l(\alpha,\beta)}{\partial\alpha\partial\beta} & -\frac{\partial^2 l(\alpha,\beta)}{\partial^2\beta} \end{pmatrix}^{-1}_{(\alpha,\beta)=(\hat{\alpha},\hat{\beta})}$$

$$= \begin{pmatrix} Var(\hat{\alpha}) & Cov(\hat{\alpha},\hat{\beta}) \\ Cov(\hat{\alpha},\hat{\beta}) & Var(\hat{\alpha}) \end{pmatrix}.$$

The derivatives in $I(\Theta)$ are given in

$$\frac{\partial^2 l(\alpha,\beta|x)}{\partial\alpha^2} = -\frac{n}{\alpha^2\beta} + \left(1+\frac{1}{\beta}\right)\sum_{i=1}^{n}\frac{1}{(\beta x_i+\alpha)^2}$$

$$\frac{\partial^2 l(\alpha,\beta|x)}{\partial\alpha\,\partial\beta} = -\frac{n}{\alpha\beta^2} + \frac{1}{\beta^2}\sum_{i=1}^{n}\frac{1}{(\beta x_i+\alpha)} - \left(1+\frac{1}{\beta}\right)\sum_{i=1}^{n}\frac{x_i}{(\beta x_i+\alpha)^2} = \frac{\partial^2 l(\alpha,\beta|x)}{\partial\beta\,\partial\alpha}$$

$$\begin{aligned}\frac{\partial^2 l(\alpha,\beta|x)}{\partial\beta^2} &= \frac{2n}{\beta^3}ln\alpha - \frac{2}{\beta^3}\sum_{i=1}^{n}ln(\beta x_i+\alpha) + \frac{2}{\beta^2}\sum_{i=1}^{n}\frac{x_i}{(\beta x_i+\alpha)} \\ &+ \quad + \left(1+\frac{1}{\beta}\right)\sum_{i=1}^{n}\left(\frac{x_i}{\beta x_i+\alpha}\right)^2.\end{aligned}$$

The above approach is used to derive approximate $100(1-\tau)\%$ confidence intervals of the parameters α and β of the forms

$$\hat{\alpha} \pm z_{\tau/2}\sqrt{var(\hat{\alpha})}$$

and

$$\hat{\beta} \pm z_{\tau/2}\sqrt{var(\hat{\beta})},$$

where $z_{\tau/2}$ is the upper $(\tau/2)$th percentile of the standard normal distribution.

6. Numerical Experiments and Discussion

In this section, we examine the performance of maximum likelihood estimates for the two parameter GP distribution by conducting simulation study for different sample sizes $n = 20, 30, 50, 100, 150$. We simulate 1000 samples with four different sets of parameters. The results are presented in Table 4, which shows the averages of $MLEs[Av(\hat{\alpha},\hat{\beta})]$ together with the 95% confidence intervalsfor parameters of GP distribution $[C(\alpha,\beta)]$ and their variances, $[Var(\hat{\alpha}), Var(\hat{\beta})]$. These results suggest that ML estimates performed adequately. The variances of $MLEs$ decrease when the sample size n increases.

The following observations can be drawn from the Tables 4

1. All the estimators show the property of consistency i.e., the $MLEs$ decreases as sample size increases.

2. The variances of $MLEs$ decrease when n increases.

Table 4: Mean of the $MLEs$, their variances and confidence interval

n	(α,β)	$Av(\hat{\alpha},\hat{\beta})$	$C(\alpha,\beta)$	$Var(\hat{\alpha})$	$Var(\hat{\beta})$
20	(2.0, 1.5)	(2.3172, 1.2835)	(0.7983, 0.8972)	1.4585	0.8074
	(0.5, 4.0)	(0.5132, 7.0972)	(0.9732, 0.9636)	0.0820	34.8920
	(4.0, 0.5)	(4.8230, 0.4973)	(0.9201, 0.9930)	7.8013	0.0723
	(2.0, 2.0)	(2.3672, 2.0874)	(0.8920, 0.9874)	1.3210	1.5672
30	(2.0, 1.5)	(2.3124, 1.2213)	(0.9230, 0.9972)	0.6707	0.4872
	(0.5, 4.0)	(0.5217, 6.0132)	(0.9731, 0.9835)	0.0631	23.7234
	(4.0, 0.5)	(4.6133, 0.4017)	(0.8937, 0.9083)	4.5983	0.0692
	(2.0, 2.0)	(1.9967, 2.0313)	(0.9538, 0.9876)	0.9012	1.5078
50	(2.0, 1.5)	(2.0891,1.1942)	(0.9074, 0.9920)	0.3948	0.2672
	(0.5, 4.0)	(0.4838, 5.0120)	(0.9235, 0.9574)	0.0572	13.0789
	(4.0, 0.5)	(4.6103, 0.3031)	(0.9318, 0.9927)	1.9071	0.0563
	(2.0, 2.0)	(1.9956, 1.9897)	(0.9536, 0.9897)	0.6752	0.9032
100	(2.0, 1.5)	(1.9762, 1.0701)	(0.9975, 0.9432)	0.3572	0.2014
	(0.5, 4.0)	(0.4702, 4.9673)	(0.9432, 0.9784)	0.0132	7.3210
	(4.0, 0.5)	(4.0259, 0.2123)	(0.9810, 0.9374)	1.8270	0.0513
	(2.0, 2.0)	(1.8130, 1.2704)	(0.9714, 0.9905)	0.3412	0.2715
150	(2.0, 1.5)	(1.8352, 1.0250)	(0.9512, 0.9930)	0.1327	0.0978
	(0.5, 4.0)	(0.3976, 3.9989)	(0.9618, 0.9568)	0.0115	0.2560
	(4.0, 0.5)	(3.9859, 0.2262)	(0.9805, 0.9907)	0.5098	0.0099
	(2.0, 2.0)	(1.7894, 1.1359)	(0.9758, 0.9853)	0.2081	0.2315

7. Concluding Remarks

In this paper, the various structural properties of the distribution are derived including explicit expressions for moments, mean deviation, Bonferroni and Lorenz curves, Renyi entropy and quantile function. The explicit expressions and recurrence relations for single and product moments of GOS are obtained from the GP distribution. The characterizing result of the GP distribution has been studied using conditional moments of generalized order statistics. The method of maximum likelihood is adopted for estimating the model parameters. For different parameter settings and sample sizes, the simulation studies are performed and compared to the performance of the GP distribution.

Acknowledgments

The authors are grateful for the comments and suggestions by the referees and the editor. Their comments and suggestions have greatly improved the article. Also first author grateful to Dr Neeraj Kumar, Amity University, Noida for his help and suggestions throughout the preparation of this paper.

REFERENCES

ABOELENEEN, Z. A., (2010). Inference for Weibull distribution under generalized order statistics, Mathematics and Computers in Simulation, 8, pp. 26–36.

ARNOLD, B. C., (1983). The Pareto distribution, International Co-operative Publishing Houshing, Fairland, MD.

ARNOLD, B. C. (2008). Pareto and Generalized Pareto Distributions. In: Chotikapanich D. (eds.) Modeling Income Distributions and Lorenz Curves. Economic Studies in Equality, Social Exclusion and Well-Being, vol. 5 , Springer, New York, NY.

BONFERRONI C. E., (1930). Elmenti di statistica generale, Libreria Seber, Firenze.

BURKSHAT, M., (2010). Linear estimators and predictors based on generalized order statistics from generalized Pareto distributions, Comm. Statist. Theory Methords, 39 , pp. 311–326.

Gradshteyn, I. S., Ryzhik, I.M., (2014). Table of Integrals, Series, and Products. Sixth edition, San Diego: Academic Press.

KAMPS, U., (1995). A concept of generalized order statistics, B.G. Teubner Stuttgart, Germany .

KENNEY, J. F., KEEPIN, E., (1962). Mathematics of Statistics, D. Van Nostrand Company.

KIM, C., HAN, K., (2014). Bayesian estimation of Rayleigh distribution based on generalized order statistics, Applied Mathematics Sciences, 8, pp. 7475–7485.

KUMAR, D., (2015a). The extended generalized half logistic distribution based on ordered random variables, Tamkang Journal of Mathematics, 46, pp. 245–256.

KUMAR, D., (2015b). Exact moments of generalized order statistics from type II exponentiated log-logistic distribution, Hacettepe Journal of Mathematics and Statistics, 44, pp. 715–733.

KUMAR, D., DEY, S., (2017a). Relations for moments of generalized order statistics from extended exponential distribution, American Journal of Mathematical and Management Sciences, 17, pp. 378–400.

KUMAR, D., DEY, S., (2017b). Power generalized Weibull distribution based on order statistics, Journal of Statistical Research, 51, pp. 61–78.

KUMAR, D., DEY, S., NADARAJAH, S., (2017). Extended exponential distribution based on order statistics, Communication in Statistics-Theory and Methods, 46, pp. 9166–9184.

KUMAR, D., JAIN N. (2018). Power generalized Weibull distribution based on generalised order statistics, Journal of data Science, 16, pp. 621–646.

KUMAR, D., GOYAL, A., (2019a). Order Statistics from the Power Lindley Distribution and Associated Inference with Application, Annals of Data Sciences, 6, pp. 153–177.

KUMAR, D., GOYAL, A., (2019b). Generalized Lindley Distribution Based on Order Statistics and Associated Inference with Application, Annals of Data Sciences, https://doi.org/10.1007/s40745-019-00196-6.

LAWLESS, J. F., (1982). Statistical models and methods for lifetime data, 2nd Edition, Wiley, New York.

MOORS, J. J. A., (1988). A quantile alternative for kurtosis, Journal of the Royal Statistical Society. Series D (The Statistician), 37, pp. 25–32.

PICKANDS, J., (1975). Statistical inference using extreme order statistics, Ann. Statist. 3, pp. 119–131.

SAFI, S. K., AHMED, R. H., (2013). Statistical estimation based on generalized order statistics from Kumaraswamy distribution, Proceeding of the 14st Applied Stochastic Models and Data Analysis (ASMDA) International Conference, Mataro (Barcelona), Spain, pp. 25–28.

Verma, V., Betti, G., (2006). EU Statistics on Income and Living Conditions (EU-SILC): Choosing the Survey Structure and Sample Design, Statistics in Transition, 7, pp. 935–970.

WU, S. J., CHEN, Y. J., CHANG, C. T., (2014). Statistical inference based on progressively censored samples with random removals from the Burr type XII distribution, Journal of Statistical Computation and Simulations, 77, pp. 19–27.

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. 81–95, DOI 10.21307/stattrans-2019-025
Submitted – 07.04.2018; Paper ready for publication – 28.09.2018

ESTIMATION OF PRODUCT OF TWO POPULATION MEANS BY MULTI-AUXILIARY CHARACTERS UNDER DOUBLE SAMPLING THE NON-RESPONDENTS

B. B. Khare[1], **R. R. Sinha**[2]

ABSTRACT

This paper considers the problem of estimating the product of two population means using the information on multi-auxiliary characters with double sampling the non-respondents. Classes of estimators are proposed for estimating P under two different situations [discussed by Rao (1986, 90)] using known population mean of multi-auxiliary characters. Further, this problem has been extended to the case when population means of the auxiliary characters are unknown and they are estimated on the basis of a larger first phase sample. In this situation, a class of two phase sampling estimators for estimating P is suggested using multi-auxiliary characters with unknown population means in the presence of non-response. The expressions of bias and mean square error of all the proposed estimators are derived and their properties are studied. An empirical study using real data sets is given to justify the theoretical considerations.

Key words: product, bias, mean square error, auxiliary characters, non-response.

1. Introduction

While conducting sample surveys in the field of agriculture, socio-economic and forest research, one may be interested in the estimation of the product of two population means. For example- if we want to estimate the total population of persons in a District using villages as the sampling units, then we will estimate the product of average number of occupied houses in a village and average number of persons in a house in that village and the population of the District can be obtained by multiplying the estimate of product to the number of villages. The auxiliary characters used in this circumstance may be the area, the number of cultivators, the number of agricultural labours, etc., of the village. Similarly one may use the amount of manure, water and seeds supplied to each plot as the auxiliary characters in the estimation of total yield of a crop within an agricultural field, which can be obtained by estimating the product of average area per plot

[1] Department of Statistics, Banaras Hindu University, Varanasi, India. E-mail: bbkhare56@yahoo.com.
[2] Department of Mathematics, Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, India. E-mail: raghawraman@gmail.com (Corresponding author).

and the average yield per plot and then multiplying this estimate of product by the number of plots in the agricultural field.

Many authors like Singh (1965, 67, 69), Shah & Shah (1978), Singh (1982a,b), Ray and Singh (1985), Khare (1987), Srivastativa *et al.* (1988), Khare (1990, 91) considered the problem of estimating the ratio and the product of two population means and suggested estimators/classes of estimators using the information of auxiliary character(s).

When the population means of the auxiliary characters are not known then in this case Singh (1982c) proposed generalized double sampling estimators for the ratio and product of population parameters using double sampling scheme. Further, Khare (1992) proposed the class of estimators for the product of two population means using multi-auxiliary characters with known and unknown population means.

In some cases, it happens that the information on the main characters and auxiliary characters may not be available in practice due to the occurrence of non-response for the selected units in the sample viz. while conducting the yield of a crop in an agricultural field, it may be possible that the data on yield of a crop as well as the area of the plot may be made available for some selected plots for which the information on the auxiliary characters may or may not be known due to lack of reporting the owner of the field. To reduce the effect of non-response in the estimation of parameter for a variable, Hansen and Hurwitz (1946) suggested a method of sub-sampling on the non-responding units and proposed an unbiased estimator for population mean. Further, following Hansen and Hurwitz (1946) strategies of sub-sampling the non-responding units, Khare and Sinha (2002 a, b, 2004 a, b, 2007, 2012 a, b), Singh *et al.* (2007), Singh and Kumar (2008) proposed some classes of estimators for the ratio/product of two population means using auxiliary character(s) in the presence of non-response. The objective of this paper is to suggest classes of estimators for estimating the product of two population means using information available on multi-auxiliary characters in the presence of non-response under different situations and study their theoretical and empirical properties.

2. Proposed classes of estimators

Let $y_i (i = 1, 2)$ and $x_j (j = 1,2, \dots, p)$ be the main and auxiliary characters under study having non-negative k^{th} value $Y_{ik}, X_{jk}; (k = 1,2, \dots, N)$ with population means $\bar{Y}_i \ (i = 1, 2)$ of study characters and $\bar{X}_j \ (j = 1, 2, \dots \dots, p)$ of auxiliary characters respectively. The whole population is supposed to be divided into two non-overlapping unknown strata of N_1 responding and N_2 non-responding units such that $N_1 + N_2 = N$. Let n be the size of the sample drawn from the population of size N using simple random sampling without replacement (SRSWOR) scheme of sampling and it has been observed that n_1 units respond and n_2 units do not respond in the sample of size n. We have considered that the responding and non-responding units are same for both the study and auxiliary characters. The stratum weights of responding and non-responding groups are given by $W_1 = N_1/N$ and $W_2 = N_2/N$ and their estimates are respectively given by $\widehat{W}_1 = w_1 = n_1/n$ and $\widehat{W}_2 = w_2 = n_2/n$. Further, from the non-responding units n_2, we

draw a subsample of size $m(= n_2\,\delta^{-1}, \delta > 1)$ using SRSWOR technique of sampling and collect the information by the direct interview for $y_i\ (i = 1,2)$. Using the methodology of Hansen and Hurwitz (1946), the unbiased estimator for $\bar{Y}_i\ (i = 1,2)$ based on the information of $(n_1 + m)$ units is given by

$$\bar{y}_{i(HH)} = w_1\bar{y}_{i1} + w_2\bar{y}_{i2}^{*}, \quad i = 1,2 \tag{1}$$

where $\bar{y}_{i1}$ and $\bar{y}_{i2}^{*}$ are the sample means of y_i based on n_1 and m units respectively.

The estimator $\bar{y}_{i(HH)}$ is unbiased and has variance given by

$$V\left(\bar{y}_{i(HH)}\right) = V^{(i)} = \lambda\, S_{y_i}^2 + \lambda_\delta\, S_{y_{i(2)}}^2, \tag{2}$$

where $S_{y_i}^2$ and $S_{y_{i(2)}}^2$ are the population mean square of y_i for the entire and non-responding group of the population and $\lambda = \frac{N-n}{Nn}$, $\lambda_\delta = \frac{N_2}{N\,n}(\delta - 1)$.

Similarly, the estimator $\bar{x}_{j(HH)}\ (j = 1,2,\ldots,p)$ for the population mean $\bar{X}_j\ (j = 1,2,\ldots,p)$ is given by

$$\bar{x}_{j(HH)} = w_1\bar{x}_{j1} + w_2\bar{x}_{j2}^{*}, \quad j = 1,2,\ldots,p \tag{3}$$

where $\bar{x}_{j1}$ and $\bar{x}_{j2}^{*}$; $(j = 1,2\ldots,p)$ are the sample means of the character $x_j\ (j = 1,2,\ldots,p)$ based on n_1 and m units respectively.

Let $\hat{P}(= \prod_{i=1}^{2}\bar{y}_{i(HH)})$ denote conventional estimator for the product of two population means $(P = \bar{Y}_1.\bar{Y}_2)$ in the presence of non-response on study characters. Utilizing the information of auxiliary characters with known population means, the two different proposed classes of estimators for P under two different cases discussed by {Rao (1986), page 220} are as follows:

- **For the first case**, it is assumed that the population mean $\bar{X}_j\ (j = 1,2,\ldots,p)$ is known, and incomplete information occurred on $y_i\ (i = 1,2)$ and $x_j\ (j = 1,2,\ldots,p)$ for the selected units in the sample of size n. In such situation, we propose a class of estimators t_p for the product of two population means (P) using multi-auxiliary characters $x_j (j = 1,2,\ldots,p)$ with known population means as:

$$t_p = g\left(\prod_{i=1}^{2}\bar{y}_{i(HH)},\ (u_1,u_2,\ldots,u_p)'\right) = g(\varphi, \boldsymbol{u}'), \tag{4}$$

such that

$$g(P, \boldsymbol{e}') = P, \qquad g_1(P, \boldsymbol{e}') = \left(\frac{\partial}{\partial\varphi}g(\varphi,\ \boldsymbol{u}')\right)_{(P,\ \boldsymbol{e}')} = 1 \tag{5}$$

where $\boldsymbol{u}$ and $\mathbf{e}$ denote the column vectors $(u_1,u_2,\ldots,u_p)'$ and $(1,1,\ldots,1)'$ respectively. We also denote $\varphi = \prod_{i=1}^{2}\bar{y}_{i(HH)}$ and $u_j = \bar{x}_{j(HH)}/\bar{X}_j\,;\ j = 1,2,\ldots,p.$

- **For the second case**, it is assumed that the population mean $\bar{X}_j\ (j = 1,2,\ldots,p)$ is known, and incomplete information occurred on $y_i\ (i = 1,2)$ only but complete information on $x_j\ (j = 1,2,\ldots,p)$ for the selected units in the sample of size n. In this situation, we propose a class of estimators t_p^{*} for the product of two population means (P) using multi-auxiliary characters $x_j (j = 1,2,\ldots,p)$ with known population means as:

$$t_p^* = h\left(\prod_{i=1}^{2} \bar{y}_{i(HH)},\ \ (\omega_1, \omega_2, \dots, \omega_p)'\right) = h(\varphi, \boldsymbol{\omega}'), \tag{6}$$

such that

$$h(P, \boldsymbol{e}') = P, \qquad h_1(P, e') = \left(\frac{\partial}{\partial \varphi} h(\varphi,\ \boldsymbol{\omega}')\right)_{(P,\ \boldsymbol{e}')} = 1 \tag{7}$$

where $\boldsymbol{\omega}$ denotes the column vectors $(\omega_1, \omega_2, \dots, \omega_p)'$ and $\omega_j = \bar{x}_j / \bar{X}_j$; $j = 1,2, \dots, p$.

For the expansion of the functions $g(\varphi, \boldsymbol{u}')$ and $h(\varphi, \boldsymbol{\omega}')$, it is supposed that whatever be the sample chosen for any sampling design, $(\varphi, \boldsymbol{u}')$ [or $(\varphi, \boldsymbol{\omega}')$] assumes a value in a bounded closed convex subset D_p [or D_p^*] of the $(p + 1)$ dimensional real space containing the point $(P,\ \boldsymbol{e}')$. In D_p [or D_p^*], the function $g(\varphi, \boldsymbol{u}')$ [or $h(\varphi, \boldsymbol{\omega}')$] is continuous and bounded. The first and second partial derivatives of $g(\varphi, \boldsymbol{u}')$ [or $h(\varphi, \boldsymbol{\omega}')$] exist and are continuous and bounded in D_p [or D_p^*].

Here, $g_1(\varphi, \boldsymbol{u}')$ [or $h_1(\varphi, \boldsymbol{\omega}')$] and $g_2(\varphi, \boldsymbol{u}')$ [or $h_2(\varphi, \boldsymbol{\omega}')$] denote the first partial derivatives of $g(\varphi, \boldsymbol{u}')$ [or $h(\varphi, \boldsymbol{\omega}')$] with respect to φ and $\boldsymbol{u}'$ [or $\boldsymbol{\omega}'$] respectively. The second partial derivatives of $g(\varphi, \boldsymbol{u}')$, $h(\varphi, \boldsymbol{\omega}')$ with respect to $\boldsymbol{u}'$ and $\boldsymbol{\omega}'$ are respectively denoted by $g_{22}(\varphi, \boldsymbol{u}')$, $h_{22}(\varphi, \boldsymbol{\omega}')$] and first partial derivative of $g_2(\varphi, \boldsymbol{u}')$ [or $h_2(\varphi, \boldsymbol{\omega}')$] with respect to φ is denoted by $g_{12}(\varphi, \boldsymbol{u}')$ [or $h_{12}(\varphi, \boldsymbol{\omega}')$].

Under the regularity conditions imposed on $g(\varphi, \boldsymbol{u}')$ and $h(\varphi, \boldsymbol{\omega}')$, it may be seen that the bias and mean square error of the estimators t_p and t_p^* will always exist.

In order to derive the expressions for bias and mean square error of the estimators under large sample approximation, let us assume

$$\epsilon_{0i} = \frac{\bar{y}_i^* - \bar{Y}_i}{\bar{Y}_i},\ \ \epsilon_j = \frac{\bar{x}_j^* - \bar{X}_i}{\bar{X}_i},\ \epsilon_j' = \frac{\overline{x}_j - \overline{X}_j}{\overline{X}_j},\ \text{with}\ \ E(\epsilon_{0i}) = E(\epsilon_j) = E(\epsilon_j') = 0$$

and $|\epsilon_{0i}| < 1$, $|\epsilon_j| < 1$, $|\epsilon_j'| < 1\ \ \forall\ i = 1,2; j = 1,2, \dots \dots, p$.

Now, using SRSWOR method of sampling, we have

$$E(\epsilon_{0i}^2) = \frac{V(\bar{y}_{i(HH)})}{\bar{Y}_i^2} = \frac{1}{\bar{Y}_i^2}\left[\lambda\, S_{y_i}^2 + \lambda_\delta\, S_{y_{i(2)}}^2\right], \quad E(\epsilon_j^2) = \frac{V(\bar{x}_{j(HH)})}{\bar{X}_j^2} = \frac{1}{\bar{X}_j^2}\left[\lambda\, S_{x_j}^2 + \lambda_\delta\, S_{x_{j(2)}}^2\right],$$

$$E(\epsilon_j'^2) = \frac{V(\bar{x}_j)}{\bar{X}_j^2} = \lambda \frac{S_{x_j}^2}{\bar{X}_j^2},\ \ E(\epsilon_{01}, \epsilon_{02}) = \frac{Cov(\bar{y}_{1(HH)}, \bar{y}_{2(HH)})}{\bar{Y}_1.\bar{Y}_2} = \frac{1}{\bar{Y}_1.\bar{Y}_2}\left[\lambda\, S_{y_1 y_2} + \lambda_\delta S_{y_1 y_2(2)}\right],$$

$$E(\epsilon_{0i}, \epsilon_j) = \frac{Cov(\bar{y}_{i(HH)}, \bar{x}_{j(HH)})}{\bar{Y}_i.\bar{X}_j} = \frac{1}{\bar{Y}_i.\bar{X}_j}\left[\lambda\, S_{y_i x_j} + \lambda_\delta S_{y_i x_j(2)}\right],$$

$$E(\epsilon_{0i}, \epsilon_j') = \frac{Cov(\bar{y}_{i(HH)}, \bar{x}_j)}{\bar{Y}_i.\bar{X}_j} = \lambda \frac{S_{y_i x_j}}{\bar{Y}_i.\bar{X}_j}, \qquad E\left(\epsilon_j', \epsilon_{j'}'\right) = \frac{Cov(\bar{x}_j, \bar{x}_{j'})}{\bar{X}_j.\bar{X}_{j'}} = \lambda \frac{S_{x_j x_{j'}}}{\bar{X}_j.\bar{X}_{j'}},$$

$$E(\epsilon_j, \epsilon_{j'}) = \frac{Cov(\bar{x}_{j(HH)}, \bar{x}_{j'(HH)})}{\bar{X}_j.\bar{X}_{j'}} = \frac{1}{\bar{X}_j.\bar{X}_{j'}}\left[\lambda\, S_{x_j x_{j'}} + \lambda_\delta S_{x_j x_{j'}(2)}\right]. \tag{8}$$

The contribution of the terms involving the powers in ϵ_{0i}, ϵ_j and ϵ_j' of order higher than two in the bias and mean square error is assumed to be negligible.

3. Bias and mean square error (MSE) of t_p and t_p^*

Expanding the functions $g(\varphi, \boldsymbol{u}')$ about the point $(P, \boldsymbol{e}')$ using Taylor's series up to the second order partial derivative, we get

$$t_p = g(P, \boldsymbol{e}') + (\varphi - P)g_1(P, \boldsymbol{e}') + (\boldsymbol{u} - \boldsymbol{e})' g_2(P, \boldsymbol{e}') + \frac{1}{2}[(\varphi - P)^2 g_{11}(\varphi^*, \boldsymbol{u}^{*\prime}) + 2(\varphi - P)(\boldsymbol{u} - \boldsymbol{e})' g_{12}(\varphi^*, \boldsymbol{u}^{*\prime}) + (\boldsymbol{u} - \boldsymbol{e})' g_{22}(\varphi^*, \boldsymbol{u}^{*\prime})(\varphi - P)]$$

Using condition given in equation (5), we get

$$t_p = P + (\varphi - P) + (\boldsymbol{u} - \boldsymbol{e})' g_2(P, \boldsymbol{e}') + \frac{1}{2}[(\varphi - P)^2 g_{11}(\varphi^*, \boldsymbol{u}^{*\prime}) + 2(\varphi - P)(\boldsymbol{u} - \boldsymbol{e})' g_{12}(\varphi^*, \boldsymbol{u}^{*\prime}) + (\boldsymbol{u} - \boldsymbol{e})' g_{22}(\varphi^*, \boldsymbol{u}^{*\prime})(\boldsymbol{u} - \boldsymbol{e})].$$

Similarly, expanding the $h(\varphi, \boldsymbol{\omega}')$ about the point $(P, \boldsymbol{e}')$ and using equation (7), we get

$$t_p^* = P + (\varphi - P) + (\boldsymbol{\omega} - \boldsymbol{e})' h_2(P, \boldsymbol{e}') + \frac{1}{2}\begin{bmatrix}(\varphi - P)^2 h_{11}(\varphi, \boldsymbol{\omega}') + 2(\varphi - P)(\boldsymbol{\omega} - \boldsymbol{e})' h_{12}(\varphi^*, \boldsymbol{\omega}^{*\prime}) \\ +(\boldsymbol{\omega} - \boldsymbol{e})' h_{22}(\varphi^*, \boldsymbol{\omega}^{*\prime})(\boldsymbol{\omega} - \boldsymbol{e})\end{bmatrix} \quad (9)$$

The expressions for bias and mean square error of t_p and t_p^* for any sampling design up to the terms of order n^{-1} are given by

$$Bias(t_p) = Bias\,(\varphi) + E(\varphi - P)(\boldsymbol{u} - \boldsymbol{e})' g_{12}(\varphi^*, \boldsymbol{u}^{*\prime}) + \frac{1}{2}E(\boldsymbol{u} - \boldsymbol{e})' g_{22}(\varphi^*, \boldsymbol{u}^{*\prime})(\boldsymbol{u} - \boldsymbol{e}), \quad (10)$$

$$MSE(t_p) = MSE\,(\varphi) + 2E(\varphi - P)(\boldsymbol{u} - \boldsymbol{e})' g_2(P, \boldsymbol{e}') + E\big(g_2(P, \boldsymbol{e}')\big)'(\boldsymbol{u} - \boldsymbol{e})(\boldsymbol{u} - \boldsymbol{e})'\, g_2(P, \boldsymbol{e}') \quad (11)$$

$$Bias(t_p^*) = Bias\,(\varphi) + E(\varphi - P)(\boldsymbol{\omega} - \boldsymbol{e})' h_{12}(\varphi^*, \boldsymbol{\omega}^{*\prime}) + \frac{1}{2}E(\boldsymbol{\omega} - \boldsymbol{e})' h_{22}(\varphi^*, \boldsymbol{\omega}^{*\prime})(\boldsymbol{\omega} - \boldsymbol{e}), \quad (12)$$

$$MSE(t_p^*) = MSE\,(\varphi) + 2E(\varphi - P)(\boldsymbol{\omega} - \boldsymbol{e})' h_2(P, \boldsymbol{e}') + E\big(h_2(P, \boldsymbol{e}')\big)'(\boldsymbol{\omega} - \boldsymbol{e})(\boldsymbol{\omega} - \boldsymbol{e})'\, h_2(P, \boldsymbol{e}'), \quad (13)$$

where $\varphi^* = P + \theta_P(\varphi - P)$, $\boldsymbol{u}^* = \boldsymbol{e} + \emptyset_{\boldsymbol{1}}(\boldsymbol{u} - \boldsymbol{e})$ and $\boldsymbol{\omega}^* = \boldsymbol{e} + \emptyset_{\boldsymbol{2}}(\boldsymbol{\omega} - \boldsymbol{e})$, such that $0 < \theta_P, \emptyset_{1j}, \emptyset_{2j} < 1$ and $\emptyset_{\boldsymbol{1}}$ and $\emptyset_{\boldsymbol{2}}$ are $(p \times p)$ diagonal matrix with j^{th} diagonal elements $\emptyset_{1j}$ and $\emptyset_{2j}$ respectively.

Differentiating equations (11) and (13) partially with respect to $g_2(P, \boldsymbol{e}')$ and $h_2(P, \boldsymbol{e}')$ respectively and equating them to zero, we get the conditions for the minimum value of the mean square error of t_p and t_p^*

$$g_2(P, \boldsymbol{e}') = -[E(\varphi - P)(\boldsymbol{u} - \boldsymbol{e})' / E(\boldsymbol{u} - \boldsymbol{e})(\boldsymbol{u} - \boldsymbol{e})'] \quad (14)$$

and $$h_2(P, \boldsymbol{e}') = -[E(\varphi - P)(\boldsymbol{\omega} - \boldsymbol{e})' / E(\boldsymbol{\omega} - \boldsymbol{e})(\boldsymbol{\omega} - \boldsymbol{e})'] \quad (15)$$

respectively.

Now, putting the value of $g_2(P,\boldsymbol{e}')$ from equation (14) to (11) and $h_2(P,\boldsymbol{e}')$ from equation (15) to (13), the minimum value of the mean square error t_p and t_p^* will be given by

$$MSE\big(t_p\big)\big|_{min} = MSE\ (\varphi) - E(\varphi - P)(\boldsymbol{u}-\boldsymbol{e})'\ [E(\boldsymbol{u}-\boldsymbol{e})(\boldsymbol{u}-\boldsymbol{e})']^{-1}E(\varphi - P)(\boldsymbol{u}-\boldsymbol{e}) \tag{16}$$

$$MSE\big(t_p^*\big)\big|_{min} = MSE\ (\varphi) - E(\varphi - P)(\boldsymbol{\omega}-\boldsymbol{e})'\ [E(\boldsymbol{\omega}-\boldsymbol{e})(\boldsymbol{\omega}-\boldsymbol{e})']^{-1}E(\varphi - P)(\boldsymbol{\omega}-\boldsymbol{e}) \tag{17}$$

Considering SRSWOR method of sampling, let us define two $p \times p$ positive definite matrices $\boldsymbol{\mathcal{A}} = \left[\mathfrak{a}_{jj'}\right]$ and $\boldsymbol{\mathcal{A}_0} = \left[\mathfrak{a}_{0jj'}\right]$such that

$$\mathfrak{a}_{jj'} = \lambda\ \mathfrak{a}_{0jj'} + \lambda_\delta\ \mathfrak{a}_{0jj'(2)} \qquad \forall\ j \neq j' = 1,2,\dots,p$$

where $\mathfrak{a}_{0jj'} = \rho_{x_jx_{j'}}C_{x_j}C_{x_{j'}}$, $\mathfrak{a}_{0jj'(2)} = \rho_{x_jx_{j'}(2)}C_{x_{j(2)}}C_{x_{j'(2)}}$, $C_{x_j}^2 = S_{x_j}^2/\bar{X}_i^2$, $C_{x_{j(2)}} = S_{x_{j(2)}}^2/\bar{X}_i^2$

$\rho_{x_jx_{j'}}$ - correlation coefficient between x_j and $x_{j'}$ for entire population,

$\rho_{x_jx_{j'}(2)}$ - correlation coefficient between x_j and $x_{j'}$ for non-responding group of population.

Then the expressions for bias and mean square error of t_p and t_p^* up to the terms of order (n^{-1}) in the case of SRSWOR method of sampling are given by

$$Bias\big(t_p\big) = Bias\ (\varphi) + P\big(\lambda\ \boldsymbol{\mathcal{b}} + \lambda_\delta\ \boldsymbol{\mathcal{b}}_{(2)}\big)'\, g_{12}(\varphi^*, \boldsymbol{u}^{*\prime}) + \frac{1}{2} trace\ \boldsymbol{\mathcal{A}}\ g_{22}(\varphi^*, \boldsymbol{u}^{*\prime}), \tag{18}$$

$$MSE\big(t_p\big) = MSE\ (\varphi) + \big(g_2(P,\boldsymbol{e}')\big)'\ \boldsymbol{\mathcal{A}}\ g_2(P,\boldsymbol{e}') + 2\ P\big(\lambda\ \boldsymbol{\mathcal{b}} + \lambda_\delta\ \boldsymbol{\mathcal{b}}_{(2)}\big)'\, g_2(P,\boldsymbol{e}'), \tag{19}$$

$$Bias\big(t_p^*\big) = Bias\ (\varphi) + \lambda\left\{P\boldsymbol{\mathcal{b}}' h_{12}(\varphi^*, \boldsymbol{\omega}^{*\prime}) + \frac{1}{2} trace\ \boldsymbol{\mathcal{A}_0}\ h_{22}(\varphi^*, \boldsymbol{\omega}^{*\prime})\right\}, \tag{20}$$

and $$MSE\big(t_p^*\big) = MSE\ (\varphi) + \lambda\left\{\big(h_2(P,\boldsymbol{e}')\big)'\boldsymbol{\mathcal{A}_0}\ h_2(P,\boldsymbol{e}') + 2\ P\boldsymbol{\mathcal{b}}'\ h_2(P,\boldsymbol{e}')\right\} \tag{21}$$

where $\boldsymbol{\mathcal{b}} = \big(\mathcal{b}_1, \mathcal{b}_2, \dots, \mathcal{b}_p\big)'$ and $\boldsymbol{\mathcal{b}}_{(2)} = \big(\mathcal{b}_{1(2)}, \mathcal{b}_{2(2)}, \dots, \mathcal{b}_{p(2)}\big)'$ are two column vectors such that

$$\mathcal{b}_j = \frac{S_{x_j}}{\bar{X}_j}\left\{\rho_{y_1x_j}\frac{S_{y_1}}{\bar{Y}_1} + \rho_{y_2x_j}\frac{S_{y_2}}{\bar{Y}_2}\right\} \text{ and } \mathcal{b}_{j(2)} = \frac{S_{x_{j(2)}}}{\bar{X}_j}\left\{\rho_{y_1x_{j(2)}}\frac{S_{y_1(2)}}{\bar{Y}_1} + \rho_{y_2x_j(2)}\frac{S_{y_2(2)}}{\bar{Y}_2}\right\},$$

$$Bias\ (\varphi) = P\left\{\lambda\ \rho_{y_1y_2}\frac{S_{y_1}}{\bar{Y}_1}\frac{S_{y_2}}{\bar{Y}_2} + \lambda_\delta\ \rho_{y_1y_2(2)}\frac{S_{y_1(2)}}{\bar{Y}_1}\frac{S_{y_2(2)}}{\bar{Y}_2}\right\}, \tag{22}$$

$$MSE(\varphi) = P^2\left\{\lambda\left(\frac{S_{y_1}^2}{\bar{Y}_1^2} + \frac{S_{y_2}^2}{\bar{Y}_2^2} + 2\rho_{y_1y_2}\frac{S_{y_1}}{\bar{Y}_1}\frac{S_{y_2}}{\bar{Y}_2}\right) + \lambda_\delta\left(\frac{S_{y_1(2)}^2}{\bar{Y}_1^2} + \frac{S_{y_2(2)}^2}{\bar{Y}_2^2} + 2\rho_{y_1y_2(2)}\frac{S_{y_1(2)}}{\bar{Y}_1}\frac{S_{y_2(2)}}{\bar{Y}_2}\right)\right\}, \tag{23}$$

$\rho_{y_1y_2}$- correlation coefficient between y_1 and y_2 for the entire population,
$\rho_{y_1x_j}$- correlation coefficient between y_1 and x_j for the entire population,
$\rho_{y_2x_j}$- correlation coefficient between y_2 and x_j for the entire population,
$\rho_{y_1y_{2(2)}}$- correlation coefficient between y_1 and y_2 for the non-responding group of population,
$\rho_{y_1x_{j(2)}}$- correlation coefficient between y_1 and x_j for the non-responding group of population,
$\rho_{y_2x_{j(2)}}$- correlation coefficient between y_2 and x_j for the non-responding group of population.

The conditions for which $MSE(t_p)$ and $MSE(t_p^*)$ will attain the minimum values are given by

$$g_2(P, \boldsymbol{e}') = -P(\lambda\,\boldsymbol{\mathscr{b}} + \lambda_\delta\,\boldsymbol{\mathscr{b}}_{(2)})\boldsymbol{\mathcal{A}}^{-1} \tag{24}$$

and $$h_2(P, \boldsymbol{e}') = -P\boldsymbol{\mathscr{b}}\boldsymbol{\mathcal{A}}^{-1} \tag{25}$$

respectively. And, the values of the minimum mean square error for t_p and t_p^* are given by

$$MSE(t_p)\big|_{min} = MSE\ (\varphi) - P^2\left\{(\lambda\,\boldsymbol{\mathscr{b}} + \lambda_\delta\,\boldsymbol{\mathscr{b}}_{(2)})'.\boldsymbol{\mathcal{A}}^{-1}.(\lambda\,\boldsymbol{\mathscr{b}} + \lambda_\delta\,\boldsymbol{\mathscr{b}}_{(2)})\right\} \tag{26}$$

and $$MSE(t_p^*) = MSE\ (\varphi) - P^2\lambda\,\boldsymbol{\mathscr{b}}'\boldsymbol{\mathcal{A}_0}^{-1}\,\boldsymbol{\mathscr{b}}\ . \tag{27}$$

4. Extension of the proposed class of estimator to the case when population means of the auxiliary characters are unknown

In the case when the population means of the auxiliary characters $\bar{X}_j\ (j = 1,2,\dots,p)$ are unknown but sampling frame is available, we use two phase sampling technique to estimate the unknown population means of the auxiliary characters $x_1, x_2, \dots, x_p$. In two phase sampling scheme, we first select a larger sample of size n' from N using SRSWOR method of sampling and collect information regarding the auxiliary characters and estimate $\bar{X}_j\ (j = 1,2,\dots,p)$ based on n'units by $\bar{x}_j'\ (j = 1,2,\dots,p)$. Again, a second phase sample of size $n\ (< n')$ is drawn from n' units by SRSWOR method of sampling and observe the study characters $y_i (i = 1,2)$. For the study characters $y_i (i = 1,2)$, we observe that only n_1 units are responding and n_2 units are not responding in the sample of size n. Now, to reduce the effect of non-response, the information is collected by the direct interview on the sub-sampled units of size $m\ (= n_2\,\delta^{-1},\ \delta > 1)$ for $y_i\ (i = 1,2)$ and following Hansen and Hurwitz (1946), the unbiased estimator

$\bar{y}_{i(HH)}$ [given in section 2 equation (1)] is considered for $\bar{Y}_i$ $(i = 1, 2)$ based on the information of $(n_1 + m)$ units.

When the population means $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p$ are unknown but estimated by $\bar{x}_1', \bar{x}_2', \dots, \bar{x}_p'$,which is based on larger first phase sample of size n' and we have incomplete information on y_i $(i = 1, 2)$ but complete information on x_j $(j = 1,2, \dots, p)$ for the sample of size n $(< n')$, then we propose a class of estimators t_p^{**} for P which is given by

$$t_p^{**} = f\left(\prod_{i=1}^{2} \bar{y}_{i(HH)},\ \ (z_1, z_2, \dots, z_p)'\right) = f(\varphi,\ \boldsymbol{z}'), \tag{28}$$

such that

$$f(P, \boldsymbol{e}') = P, \qquad f_1(P, e') = \left(\frac{\partial}{\partial \varphi} f(\varphi,\ \boldsymbol{z}')\right)_{(P,\ \boldsymbol{e}')} = 1 \tag{29}$$

where $\boldsymbol{z}$ denotes the column vectors $(z_1, z_2, \dots, z_p)'$ and $z_j = \bar{x}_j / \bar{x}_j'$; $j = 1,2, \dots, p$.

The function $f(\varphi,\ \boldsymbol{z}')$ satisfies all the necessary regularity conditions similar to those given for the functions $g(\varphi, \boldsymbol{u}')$ and $h(\varphi, \boldsymbol{\omega}')$.

Now, expanding the function $f(\varphi,\ \boldsymbol{z}')$ about the point $(P, \boldsymbol{e}')$ by using Taylor's series up to the second order derivatives, the expressions for bias and mean square error of the estimator t_p^{**} for any sampling design up to the terms of order (n^{-1}) are given by

$$Bias\left(t_p^{**}\right) = Bias\ (\varphi) + E(\varphi - P)(\boldsymbol{z} - \boldsymbol{e})' f_{12}\left(\varphi^*, \boldsymbol{z}^{*'}\right) + \frac{1}{2} E(\boldsymbol{z} - \boldsymbol{e})' h f_{22}(\varphi^*, \boldsymbol{\omega}^{*'})(\boldsymbol{\omega} - \boldsymbol{e}), \tag{30}$$

and

$$MSE\left(t_p^{**}\right) = MSE\ (\varphi) + 2E(\varphi - P)(\boldsymbol{z} - \boldsymbol{e})' f_2(P, \boldsymbol{e}') + E\left(f_2(P, \boldsymbol{e}')\right)'(\boldsymbol{z} - \boldsymbol{e})(\boldsymbol{z} - \boldsymbol{e})'\ f_2(P, \boldsymbol{e}'), \tag{31}$$

where $\boldsymbol{z}^* = \boldsymbol{e} + \emptyset_{\boldsymbol{3}}(\boldsymbol{u} - \boldsymbol{e})$; $0 < \emptyset_{3j} < 1$ and $\emptyset_{\boldsymbol{3}}$ is $(p \times p)$ diagonal matrix having diagonal elements $\emptyset_{3j}$ $(j = 1,2, \dots, p)$.

Here, $f_1(\varphi, \boldsymbol{z}')$ and $f_2(\varphi, \boldsymbol{z}')$ denote the first partial derivatives of $f(\varphi, \boldsymbol{z}')$ with respect to φ and $\boldsymbol{z}'$ respectively. The second partial derivative of $f(\varphi, \boldsymbol{z}')$ with respect to $\boldsymbol{z}'$ is denoted by $f_{22}(\varphi, \boldsymbol{z}')$ and the first partial derivative of $f_2(\varphi, \boldsymbol{z}')$ with respect to φ is denoted by $f_{12}(\varphi, \boldsymbol{z}')$.

The estimator t_p^{**} will attain the minimum value of the mean square error for

$$f_2(P, \boldsymbol{e}') = -[E(\varphi - P)(\boldsymbol{z} - \boldsymbol{e})' / E(\boldsymbol{z} - \boldsymbol{e})(\boldsymbol{z} - \boldsymbol{e})'] \tag{32}$$

and, the minimum value of mean square error of t_p^{**} is given by

$$MSE\left(t_p^{**}\right)\Big|_{min} = MSE\ (\varphi) - E(\varphi - P)(\boldsymbol{z} - \boldsymbol{e})'\ [E(\boldsymbol{z} - \boldsymbol{e})(\boldsymbol{z} - \boldsymbol{e})']^{-1} E(\varphi - P)(\boldsymbol{z} - \boldsymbol{e}). \tag{33}$$

To obtain the expressions of bias and the mean square error of t_p^{**} under SRSWOR method of sampling, we assume $\in_j''=\frac{\bar{x}_j'-\bar{X}_j}{\bar{X}_j}$, such that $E(\in_j'')=0, |\in_j''|<1$ and, therefore, we have

$$E(\in''^2_j)=\frac{V(\bar{x}_j')}{\bar{X}_j^2}=\lambda'\frac{S_{\bar{x}_j}^2}{\bar{X}_j^2},\quad E\left(\in_j'',\in_{j'}''\right)=\frac{Cov(\bar{x}_j',\bar{x}_{j'}')}{\bar{X}_j.\bar{X}_j}=\frac{1}{\bar{X}_j.\bar{X}_{j'}}\left[\lambda' S_{x_j x_{j'}}\right];\ j\neq j'=1,2,\dots,p$$

$$E(\in_{0i},\in_j'')=\frac{Cov(\bar{y}_{i(HH)},\bar{x}_j')}{\bar{Y}_i.\bar{X}_j}=\lambda'\frac{S_{y_i x_j}}{\bar{Y}_i.\bar{X}_j},\qquad E(\in_j',\in_j'')=\frac{Cov(\bar{x}_j,\bar{x}_j')}{\bar{X}_j^2}=\lambda'\frac{S_{\bar{x}_j}^2}{\bar{X}_j^2},$$

where $\lambda'=\frac{N-n'}{N\,n'}$.

Now, the expressions for bias and the mean square error of t_p^{**} up to the order n^{-1} under SRSWOR method of sampling are given by

$$Bias(t_p^{**})=Bias\ (\varphi)+\lambda'\left\{P\boldsymbol{\mathscr{b}}'f_{12}(\varphi^*,\boldsymbol{z}^{*\prime})+\frac{1}{2}trace\ \boldsymbol{\mathcal{A}_0}\ f_{22}(\varphi^*,\boldsymbol{z}^{*\prime})\right\} \tag{34}$$

and $$MSE(t_p^{**})=MSE\ (\varphi)+\lambda'\left\{\left(f_2(P,\boldsymbol{e}')\right)'\boldsymbol{\mathcal{A}_0}\ f_2(P,\boldsymbol{e}')+2\ P\boldsymbol{\mathscr{b}}'\ f_2(P,\boldsymbol{e}')\right\} \tag{35}$$

The conditions for which $MSE(t_p^{**})$ will attain the minimum value are given by

$$f_2(P,\boldsymbol{e}')=P\boldsymbol{\mathcal{A}_0}^{-1}P \tag{36}$$

and the minimum mean square error of t_p^{**} is given by

$$MSE(t_p^{**})=MSE\ (\varphi)-P^2\lambda'\ \boldsymbol{\mathscr{b}}'\boldsymbol{\mathcal{A}_0}^{-1}\ \boldsymbol{\mathscr{b}}. \tag{37}$$

5. Concluding remarks

i) The proposed classes of estimators t_p, t_p^* and t_p^{**} have a wider class of estimators. Following the strategies of Raj (1965), Singh (1967), Abu-Dayyeh *et al.* (2003), Kadilar and Cingi (2005), Perri (2005) and many more, a large number of estimators may be formed, some of them for t_p, t_p^* and t_p^{**} are given in Table 1.

Table 1. Members of classes of estimators t_p, t_p^* and t_p^{**}

t_p	t_p^*	t_p^{**}
$T_{P1}=\ \varphi\prod_{j=1}^{p}u_j^{\theta_{1j}}$	$T_{P1}^*=\ \varphi\prod_{j=1}^{p}\omega_j^{\theta_{1j}^*}$	$T_{P1}^{**}=\ \varphi\prod_{j=1}^{p}z_j^{\theta_{1j}^{**}}$
$T_{P2}=\ \varphi\prod_{j=1}^{p}u_j^{\theta_{2j}}+\sum_{j=1}^{p}c_j\left(\bar{X}_j-\bar{x}_{j(HH)}\right)$	$T_{P2}^*=\ \varphi\prod_{j=1}^{p}\omega_j^{\theta_{2j}^*}+\sum_{j=1}^{p}c_j^*\left(\bar{X}_j-\bar{x}_{j(HH)}\right)$	$T_{P2}^{**}=\ \varphi\prod_{j=1}^{p}z_j^{\theta_{2j}^{**}}+\sum_{j=1}^{p}c_j^{**}\left(\bar{X}_j-\bar{x}_{j(HH)}\right)$
$T_{P3}=\varphi\sum_{j=1}^{p}w_j\,u_j^{\theta_{3j}/w_j},\ \sum_{j=1}^{p}w_j=1$	$T_{P3}^*=\varphi\sum_{j=1}^{p}w_j\,\omega_j^{\theta_{3j}^*/w_j},\ \sum_{j=1}^{p}w_j=1$	$T_{P3}^{**}=\varphi\sum_{j=1}^{p}w_j\,z_j^{\theta_{3j}^{**}/w_j},\ \sum_{j=1}^{p}w_j=1$

Since the estimators (T_{P1}, T_{P2}, T_{P3}), (T_{P1}^*, T_{P2}^*, T_{P3}^*) and (T_{P1}^{**}, T_{P2}^{**}, T_{P3}^{**}) are the members of t_p, t_p^* and t_p^{**} and they satisfy accordingly the conditions (5), (7) and

(29), the values of constants involved in (T_{P1},T_{P2},T_{P3}), $(T_{P1}^{*},T_{P2}^{*},T_{P3}^{*})$ and $(T_{P1}^{**},T_{P2}^{**},T_{P3}^{**})$ can be calculated by the equations (14), (15) and (32) respectively. In the case when the values of parameters in the optimum value of the constants are not known one may estimate it on the basis of the sample values or may use past data. Srivastava and Jhajj (1983) shown that such values do not affect the mean square error of the estimator up to the terms of order n^{-1} while Reddy (1978) shown that such values are stable over time and region. So the proposed class of two phase sampling estimator is preferred in large scale sample survey.

ii) On comparing the estimators t_p, t_p^* and t_p^{**} with φ in terms of precision, we find

a) $MSE(t_p) < MSE\ (\varphi)$ if
$MSE\ (\varphi) < \left(g_2(P,e')\right)' \mathcal{A}\ g_2(P,e') + 2\ P\left(\lambda\ \mathscr{b} + \lambda_\delta\ \mathscr{b}_{(2)}\right)' g_2(P,e') < 0$

b) $MSE(t_p^*) < MSE\ (\varphi)$ if
$MSE\ (\varphi) < \lambda\left\{\left(h_2(P,e')\right)' \mathcal{A}_0\ h_2(P,e') + 2\ P\mathscr{b}'\ h_2(P,e')\right\} < 0$

c) $MSE(t_p^{**}) < MSE\ (\varphi)$ if
$MSE\ (\varphi) < \lambda'\left\{\left(f_2(P,\boldsymbol{e}')\right)' \boldsymbol{\mathcal{A}_0}\ f_2(P,\boldsymbol{e}') + 2\ P\boldsymbol{\mathscr{b}}'\ f_2(P,\boldsymbol{e}')\right\} < 0.$

iii) When there is complete information on the study characters $y_i (i = 1,2)$ and the auxiliary character and $x_j (j = 1,2,\dots,p)$, i.e. $W_2 = 0$, then we find that the estimators t_p and t_p^* are equally efficient for the class of estimators proposed by Khare (1992) for P using known population means of auxiliary characters. Similarly, the estimator t_p^{**} is also equally efficient to for class of two phase sampling estimators proposed by Khare (1992) for unknown population means of auxiliary characters.

iv) Due to the involvement of various parameters, in the mean square error of t_p and t_p^*, it is very difficult to find the condition for superiority of t_p over t_p^*. However, in the case of one auxiliary character, it has been obtained that relative efficiency of t_p with respect to t_p^* increases by increasing the values of $\frac{\rho_{y_1x_{1(2)}}}{\rho_{y_1x_1}}$ and $\frac{\rho_{y_2x_{1(2)}}}{\rho_{y_2x_1}}$ and decreasing the value of $\frac{\rho_{y_2x_1}}{\rho_{y_1x_1}}$. Hence, one may use the estimators t_p and t_p^* depending upon the values of $\frac{\rho_{y_1x_{j(2)}}}{\rho_{y_1x_j}}$, $\frac{\rho_{y_2x_{j(2)}}}{\rho_{y_2x_j}}$ and $\frac{\rho_{y_2x_j}}{\rho_{y_1x_j}}$ for all $j = 1,2,\dots p$.

6. An empirical study

Source: Police-station – Baria, Tahasil – Champua, District Census Handbook-1981, Orissa, Govt. of India. Total number of villages 109, 25% villages (i.e. 27 villages) from the bottom are considered as the non-responding group of the population. In this data set, study characters and auxiliary characters are as follows:

y_1-Number of occupied residential houses in the village,

y_2-Average number of persons in the house in the village,

x_1-Number of cultivators in the village,
x_2-Area (in hectares) of the village,
x_3-Number of main workers in the village.

The values of the parameters of the population under study are as follows :

$\bar{Y}_1$= 88.3670 $\bar{Y}_2$= 5.5832 $\bar{X}_1$= 100.5505 $\bar{X}_2$= 256.3331 $\bar{X}_3$= 165.2661

S_{y_1}= 59.3208 S_{y_2}= 0.6024 C_{x_1}= 0.7314 C_{x_2}= 0.6105 C_{x_3}= 0.6828

$S_{y_{1(2)}}$= 45.2704 $S_{y_{2(2)}}$= 0.5025 $C_{x_{1(2)}}$= 0.5678 $C_{x_{2(2)}}$= 0.4944 $C_{x_{3(2)}}$= 0.5769

$\rho_{y_1x_1}$= 0.795 $\rho_{y_1x_2}$= 0.854 $\rho_{y_1x_3}$= 0.907 $\rho_{y_2x_1}$= -0.084 $\rho_{y_2x_2}$= -0.117 $\rho_{y_2x_3}$= -0.136

$\rho_{y_1x_{1(2)}}$= 0.658 $\rho_{y_1x_{2(2)}}$= 0.759 $\rho_{y_1x_{3(2)}}$= 0.891 $\rho_{y_2x_{1(2)}}$= 0.092 $\rho_{y_2x_{2(2)}}$= 0.199 $\rho_{y_2x_{3(2)}}$= 0.109

$\rho_{x_1x_2}$= 0.715 $\rho_{x_1x_3}$= 0.841 $\rho_{x_2x_3}$= 0.796 $\rho_{x_1x_{2(2)}}$= 0.541 $\rho_{x_1x_{3(2)}}$= 0.785 $\rho_{x_2x_{3(2)}}$= 0.657

$\rho_{y_1y_2}$= -0.194 $\rho_{y_1y_{2(2)}}$= 0.023

To compare the efficiency of the proposed classes of estimators t_p, t_p^* and t_p^{**} with respect to the conventional estimator $\varphi[=\prod_{i=1}^{2}\bar{y}_{i(HH)}]$ through an empirical study based on real data set, their respective members $T_{P1}=\varphi\prod_{j=1}^{p}u_j^{\theta_{1j}}$, $T_{P1}^*=\varphi\prod_{j=1}^{p}\omega_j^{\theta_{1j}^*}$ and $T_{P1}^{**}=\varphi\prod_{j=1}^{p}z_j^{\theta_{1j}^{**}}$ are considered.

The mean square error (MSE) of T_{P1} and T_{P1}^* along with their optimum value of constants $(\boldsymbol{OVC})$ and their percentage relative efficiency (PRE) with respect to φ for different values of sub-sampling fraction $(1/\delta)$ are shown in Table 2, while the $MSE(T_{P1}^{**})$ and $PRE(T_{P1}^{**})$ with respect to φ in the case of fixed sample sizes, i.e. $n'=70$ and $n=40$ for different values of sub-sampling fraction $(1/\delta)$ are given in Table 3.

7. Discussion and conclusion

From Table 2, it has been observed that the estimators T_{P1} and T_{P1}^* are more efficient than φ [i.e. $\hat{P}$] for the different values of the sub-sampling fraction. We also observe that the mean square error of T_{P1} and T_{P1}^*decreases while the relative efficiency of T_{P1} and $T_{P1}^*$with respect to φ increases as the number of auxiliary characters and sub-sampling fraction increase. From the Table 2, it has also been observed that the estimator T_{P1} is more efficient than T_{P1}^* and the efficiency is increasing with the increase in the number of auxiliary characters and sub-sampling fraction.

From Table 3, we observe that the estimator T_{P1}^{**} is more efficient than φ for the different values of the sub-sampling fraction $(1/\delta)$. The relative efficiency of T_{P1}^{**} with respect to φ increases while $MSE(T_{P1}^{**})$ decreases as the number of auxiliary characters and sub-sampling fractions increase. Hence, we conclude that the efficiency of the estimators T_{P1}, T_{P1}^* and T_{P1}^{**} with respect to φ can be increased by increasing the number of the auxiliary characters as well by increasing the values of the sub-sampling fractions.

Table 2. Percentage relative efficiency [PRE (·)] with respect to φ at different values of δ

Estimators	Auxiliary character(s)	$N = 109, n = 40$		
		$1/\delta$		
		$1/4$	$1/3$	$1/2$
φ	–	100.00 (2905.7121)*****	100.00 (2494.6591)	100.00 (2083.6062)
T_{P1}	x_1	220.05 (1320.4901)	229.52 (1086.8873)	244.43 (852.4396)
	OVC	θ_{11} = –0.6702	θ_{11} = –0.6802	θ_{11} = –0.6941
	x_1, x_2	411.70 (705.7874)	428.12 (582.7021)	453.38 (459.5749)
	OVC	θ_{11} = –0.3343, θ_{12} = –0.6326	θ_{11} = –0.3369, θ_{12} = –0.6311	θ_{11} = –0.3408, θ_{12} = –0.6286
	x_1, x_2, x_3	691.16 (420.4097)	701.38 (355.6792)	717.60 (290.3568)
	OVC	θ_{11} = –0.015, θ_{12} = –0.382, θ_{13} = –0.575	θ_{11} = –0.028, θ_{12} = –0.385, θ_{13} = –0.565	θ_{11} = –0.048, θ_{12} = –0.388, θ_{13} = –0.549
T_{P1}^{*}	x_1	157.09 (1849.7092)	173.40 (1438.6562)	202.76 (1027.6033)
	OVC	θ_{11}^{*} = 0.714	θ_{11}^{*} = 0.714	θ_{11}^{*} = 0.714
	x_1, x_2	185.13 (1569.5338)	215.34 (1158.4808)	278.77 (747.4279)
	OVC	θ_{11}^{*} = –0.348, θ_{12}^{*} = –0.347	θ_{11}^{*} = –0.348, θ_{12}^{*} = –0.347	θ_{11}^{*} = –0.348, θ_{12}^{*} = –0.347
	x_1, x_2, x_3	199.42 (1457.1192)	238.48 (1046.0662)	328.12 (635.0133)
	OVC	θ_{11}^{*} = –0.079, θ_{12}^{*} = –0.391, θ_{13}^{*} = –0.523	θ_{11}^{*} = –0.079, θ_{12}^{*} = –0.391, θ_{13}^{*} = –0.523	θ_{11}^{*} = –0.079, θ_{12}^{*} = –0.391, θ_{13}^{*} = –0.523

*****Mean square error of the estimators (·) is shown in the parenthesis.

Table 3. Percentage relative efficiency [PRE (T_{P1}^{**})] with respect to φ at different values of δ for fixed n' and n

Estimators	Auxiliary character(s)	$n' = 70,\ n = 40$		
		$1/\delta$		
		$1/4$	$1/3$	$1/2$
φ	–	100.00 (2905.7121)*	100.00 (2494.6591)	100.00 (2083.6062)
T_{P1}^{**}	x_1	132.74 (2188.9752)	140.31 (1777.9222)	152.44 (1366.8693)
	OVC	θ_{11}^{**} = –0.717	θ_{11}^{**} = –0.717	θ_{11}^{**} = –0.717
	x_1, x_2	144.89 (2005.4099)	156.47 (1594.3569)	176.08 (1183.3040)
	OVC	θ_{11}^{**} = –0.346, θ_{12}^{**} = –0.622	θ_{11}^{**} = –0.346, θ_{12}^{**} = –0.622	θ_{11}^{**} = –0.346, θ_{12}^{**} = –0.622
	x_1, x_2, x_3	150.58 (1929.7012)	164.27 (1518.6482)	188.12 (1107.5953)
	OVC	θ_{11}^{**} = –0.067, θ_{12}^{**} = –0.382, θ_{13}^{**} = –0.538	θ_{11}^{**} = –0.067, θ_{12}^{**} = –0.382, θ_{13}^{**} = –0.538	θ_{11}^{**} = –0.067, θ_{12}^{**} = –0.382, θ_{13}^{**} = –0.538

*Mean square error of the estimators is shown in the parenthesis.

Acknowledgements

Authors are grateful to the referee and the editor for their invaluable suggestions, which helped in further improvisation of the paper.

REFERENCES

ABU-DAYYEH, W. A., AHMED, M. S., AHMED, R. A., MUTTLAK, H. A., (2003). Some estimators of finite population mean using auxiliary information, App. Math. Comp., 139, pp. 287–298.

HANSEN, M. H., HURWITZ, W. N., (1946). The problem of non-response in sample surveys, Jour. Amer. Stat. Assoc., 41, pp. 517–529.

KADILAR, C., CINGI, H., (2005). A new estimator using two auxiliary variables, App. Math. Comp., 162, pp. 901–908.

KHARE, B. B., (1987). On modified class of estimators of ratio and product of two population means using auxiliary character, Proc. Math. Soc., B.H.U., 3, pp. 131–137.

KHARE, B. B., (1990). A generalized class of estimators for a combination of products and ratio of some population means using multi-auxiliary characters, Jour. Stat. Res., 24(1-2), pp. 1–8.

KHARE, B. B., (1991). On generalized class of estimators for ratio of two population means using multi-auxiliary characters, Ali. Jour. Stat., 11, pp. 81–90.

KHARE, B. B., (1992). On class of estimators for product of two population means using multi-auxiliary characters with known and unknown means, Ind. Jour. Appl. Stat., 1, pp. 56–67.

KHARE, B. B., SINHA, R. R., (2002 a). On the class of two phase sampling estimators for the product of two population means using the auxiliary character in presence of non-response, Proceedings of the 5th International Symposium on Optimization and Statistics, pp. 221–232.

KHARE, B. B., SINHA, R. R., (2002 b). Estimation of the ratio of two population means using auxiliary character with unknown population mean in presence of non-response, Prog. Math., 36, pp. 337–348.

KHARE, B. B., SINHA, R. R., (2004 a). Estimation of finite population ratio using two phase sampling scheme in the presence of non-response, Ali. Jour. Stat., 24, pp. 43–56.

KHARE, B. B., SINHA, R. R., (2004 b). On the general class of two phase sampling estimators for the product of two population means using the auxiliary character in the presence of non-response, Ind. Jour. Appl. Stat., 8, pp. 1–14.

KHARE, B. B., SINHA, R. R., (2007). Estimation of the ratio of the two population means using multi-auxiliary characters in the presence of non-response, Proceedings of Statistical Technique in Life Testing, Reliability, Sampling Theory and Quality Control, pp. 205–213.

KHARE, B. B., SINHA, R. R., (2012 a). Improved classes of estimators for ratio of two means with double sampling the non-respondents, Statistika, 49(3), pp. 75–83.

KHARE, B. B., SINHA, R. R., (2012 b). Combined class of estimators for ratio and product of two population means in presence of non-response, Inter. Jour. Stat. Eco., 8, pp. 86–95.

PERRI, P. F., (2005). Combining two auxiliary variables in ratio-cum-product type estimators, Proceedings of Italian Statistical Society, Intermediate Meeting on Statistics and Environment, Messina, 21-23 Sept. 2005, pp. 193–196.

RAJ, D., (1965). On a method of using multi-auxiliary information in sample surveys, Jour. Amer. Stat. Assoc., 60, pp. 154–165.

RAO, P. S. R. S., (1986). Ratio estimation with subsampling the nonrespondents, Survey Methodology, 12(2), pp. 217–230.

RAO, P. S. R. S., (1990). Regression estimators with subsampling of nonrespondents, In-Data Quality Control, Theory and Pragmatics, (Eds.) Gunar E. Liepins and V.R.R. Uppuluri, Marcel Dekker, New York, pp. 191–208.

RAY, S. K., SINGH, R. K., (1985). Some estimators for the ratio and product of population parameters, Jour. Ind. Soc. Ag. Stat., 37, pp. 1–10.

REDDY, V. N., (1978). A study of use of prior knowledge on certain population parameters in estimation, Sankhaya, C, 40, pp. 29–37.

SHAH, S. M., SHAH, D. N., (1978) Ratio cum product estimator for estimating ratio (product) of two population parameters, Sankhya, C, 40, pp. 156–166.

SINGH, H. P., KUMAR, S., (2008). A general family of estimators of finite Population ratio, product and mean using two phase sampling scheme in the presence of non-response, Jour. Stat. Theory and Practice, 2(4), pp. 677–692.

SINGH, H. P., CHANDRA, P., JOARDER, A. H., SINGH, S., (2007). Family of estimators of mean, ratio and product of a finite population using random non-response, Test, 16(3), pp. 565–597.

SINGH, M. P., (1965). On the estimation of ratio and product of population parameters, Sankhya, C, 27, pp. 321–328.

SINGH, M. P., (1967). Ratio cum product method of estimation, Metrika, 12, pp. 34–43.

SINGH, M. P., (1969). Comparison of some ratio cum product estimators, Sankhya, B, 31, pp. 375–378.

SINGH, R. K., (1982 a). Generalized estimators for the estimation of ratio and product of population parameters, Jour. Stat. Res., 16, pp. 15–23.

SINGH, R. K., (1982 b). On estimating ratio and product of population parameters, Cal. Stat. Asso. Bull., 20, pp. 39–49.

SINGH, R. K., (1982 c) Generalized double sampling estimators for the ratio and product of population parameters, Jour. Ind. Stat. Asso., 20, pp. 39–49.

SRIVASTAVA, S. K., JHAJJ, H. S., (1983). A class of estimators of the population mean using multi-auxiliary information, Cal. Stat. Asso. Bull., 32, pp. 47–56.

SRIVASTAVA, S. RANI, KHARE, B. B., SRIVASTAVA, S. R., (1988). On generalised chain estimator for ratio and product of two population means using auxiliary characters, Assam Stat. Rev., 1, pp. 21–29.

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. 97–117, DOI 10.21307/stattrans-2019-026
Submitted – 14.11.2018; Paper ready for publication – 12.08.2019

COMPARATIVE ANALYSIS OF POVERTY IN FAMILIES WITH A DISABLED CHILD AND FAMILIES WITH NON-DISABLED CHILDREN IN POLAND IN THE YEARS 2014 AND 2016

Olga Komorowska[1], **Arkadiusz Kozłowski**[2], **Teresa Słaby**[3]

ABSTRACT

The presence of a child with disabilities in a family presents more challenging conditions than the presence of a non-disabled child. One of the difficulties is of financial nature. One of the parents often has to give up their job to care for the child, which shrinks the household income. At the same time, the family has higher expenses resulting from, e.g. costs of treatment. All this increases the risk of falling into poverty. The goal of this paper is to analyse the financial situation of households with a disabled child, mainly in the context of poverty, and compare it to the financial situation of households with non-disabled children. The study is based on data from Polish Household Budget Survey, covering two years, 2014 and 2016. The study revealed that families with a disabled child are generally poorer than families with non-disabled children. The financial situation improved over the studied period in both types of families, but the improvement in the families with a disabled child was much greater. The main factor in reducing the risk of poverty in both types of families is the education attainment level of the reference person (the household head), which should be at least upper secondary.

Key words: households with a disabled child, factors related to poverty, Household Budget Survey, logistic regression.

1. Introduction

The issue of poverty is frequently addressed in economic, social and political discourses. In “Europe 2020: A strategy for smart, sustainable and inclusive growth”, one of the five headline targets determined for the European Union is combating poverty (European Commission, 2010, p. 3). Poverty is a dangerous phenomenon – both for the entire society, and for people categorised as the poor. Low-income households limit their consumption, both current and related to

[1] Department of Statistics, Faculty of Management, University of Gdansk. E-mail: olga.komorowska@ug.edu.pl, ORCID ID: http://orcid.org/0000-0002-0305-8748.
[2] Department of Statistics, Faculty of Management, University of Gdansk. E-mail: arkadiusz.kozlowski@ug.edu.pl. ORCID ID: http://orcid.org/0000-0001-6282-1494.
[3] Warsaw Management University. E-mail: teresa.slaby@wsm.warszawa.pl. ORCID ID: http://orcid.org/0000-0002-9354-8571.

development and prevention. These limitations usually bring about changes in the behaviour and mentality, which may result in passivity, loss of self-esteem, alcohol abuse and other addictions, pathologies and aggression. All of this may, in turn, lead to reduced participation in various aspects of life, namely – to social exclusion.

A highly dangerous phenomenon is the intergenerational transmission of poverty (Bird, 2013; Harper, Marcus and Moore, 2003; Kruszyński and Warzywoda-Kruszyńska, 2011). From the perspective of the entire society poverty is associated with wastage of human capital, financial outlays for support and the growth of poverty enclaves (Golinowska, et al., 2008, pp. 60-61).

Poverty is related to failure to meet one's needs at the expected level due to too low an income (Panek, 2014, p. 196). This situation develops for a number of reasons. Among the predictors of poverty, the following are mentioned: the source of income of household head from unearned sources other than retirement; number of children in household; education attainment level of household head; voivodship; unemployed persons in household; persons with disabilities in household – especially when they are children with disabilities (Szarfenberg and Szewczyk, 2010, pp. 29-30; GUS, 2015a, pp. 10-11). The impact of some factors on emergence and persistence of poverty is ambiguous: at times it is hard to say whether a given factor is the cause or the effect of poverty (for example alcohol addiction – sometimes it can be the result of living below the poverty line, and sometimes it can be a reason for finding oneself in a group of the impoverished).

One of the factors increasing the risk of poverty is the presence of a disabled person in a household. In 2016, the incidence of extreme poverty (percentage of persons in households with expenditures below extreme poverty threshold set by the Institute of Labour and Social Studies) in households with at least one disabled person was 7.5%, whereas in a household without such members, the corresponding value was 4.2%. With regard to households where a child was the disabled person, the incidence of extreme poverty went further up, reaching 8.3% (GUS, 2017a, p. 4). Two years earlier, all three indicators were higher, amounting to 10.8%, 6.5%, and 14.6% respectively (GUS, 2015b, p. 4). A situation in which the incidence of extreme poverty is higher in households with at least one disabled child than in households with disabled adults had been the case in point for several years. But in 2017 the situation changed; the incidence of extreme poverty in households with at least one disabled person (regardless of age) was 6.7%, while in households with at least one disabled child (under 16) it was 4.9% (GUS, 2018, p. 4).

The analysis in this paper covers households where at least one person is 18 or under. The work aims to describe poverty from various perspectives since this issue is complex, ambiguous and diverse, both in the territorial and social sense. The analysis concerns the research conducted in 2016 and 2014. It is important to know that in 2016 the support programme "Family 500+" was introduced, which is likely to have reduced poverty in households with members under 18 years of age (see: GUS, 2017b, p. 12).

The statistical analysis was carried out using unit data from the Household Budget Survey (HBS) of 2014 and 2016. The sample covered in HBS in 2014 included 12,809 households with non-disabled children and 622 households with

at least one disabled child (in the vast majority of households it was only one such child). The sample used in HBS 2016 covered 12,172 and 635 households respectively. The households were studied as an entirety or taking into account their size and composition with the use of an equivalence scale. In the latter case, the so-called modified OECD equivalence scale was employed, as proposed by Haagenars, de Vos and Zaidi (1994, p. 18), which is currently used by Eurostat. See also Anyaegbu, (2010), and Łukasiewicz, Koszela and Orłowski (2006, pp. 207-217). The scale assigns the weight of 1 to the first person aged 14 or more, 0.5 to every subsequent person of the same age group, and 0.3 to children under 14.

The main contribution of this paper is the exploration of data from HBS, showing the potential of this survey, which allows the analysis of household finances broken down by the characteristics of individual members of the household such as age or having a disability. HBSs are conducted in all European Union Member States, and, although they are not harmonised, similar analyses can be performed in other countries and compared with the following results.

The second section of the article presents the general financial situation of the two groups of households under analysis. First of all, some objective metrics of the situation are considered, namely: source of income, levels of income and expenditure; then a subjective analysis of these households is performed, also in comparison with their actual financial situation. The following section includes typical elements of poverty analysis, namely the poverty thresholds, headcount rates, and depth of poverty, from the objective and subjective perspective. The final part addresses the differentiating factors for poor and non-poor households. In this analysis, classification trees and the logistic regression model were used.

2. General assessment of the income situation of households with children

The first aspect used in the comparison of the two groups of households is the main source of household income. Figure 1 presents frequency distributions in the form of a scatter plot. Thanks to this, any potential differences in distributions are more pronounced. The sources close to the diagonal of the square represent a similar share in both household groups under analysis. The sources above the diagonal are more frequent in households with disabled children, whereas the sources below the diagonal apply more frequently to households with non-disabled children.

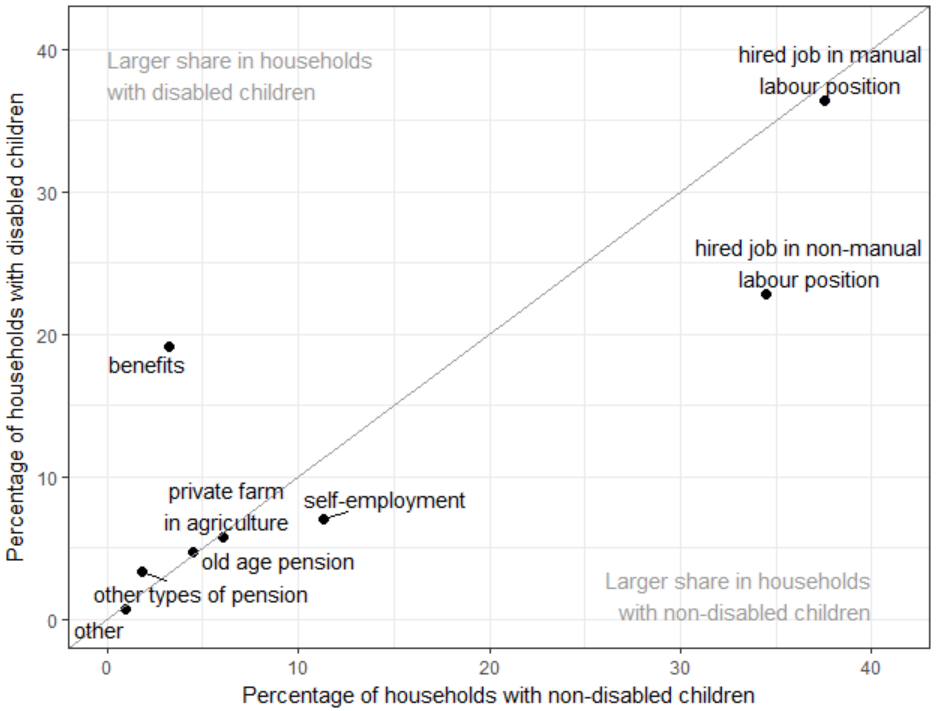


Figure 1. Distributions of main sources of income in 2016
Source: Own study based on unit data from HBS 2016.

The greatest difference in the distribution of income sources can be seen under the position *benefits*, which comprise *unemployment benefit* and *other social benefits*. The share of this source of income is 15.8 percentage point higher in households with disabled children. An even greater disproportion occurs when the main and additional sources of income are taken into account: 67.4% of households with disabled children indicated benefits as the main or additional source of income, with 25.0% of the same households with non-disabled children. In 2014, these percentages were 61.3% and 9.8%, respectively. Next, a smaller share of households with disabled children – as compared to households with non-disabled children – is supported by performing *hired job in non-manual labour position* (difference: 11.6 percentage points) and *self-employment* (difference: 4.3 percentage points).

Figure 2 presents the empirical distribution functions of disposable income and total expenditure calculated per equivalent unit. Only in the case of low levels of income (up to about PLN 1,800) the difference in distribution functions is not big, within the range of 1-2 percentage points. For the remaining values of income and for virtually the entire scope of expenditure variability, the values of distribution functions for households with disabled children are higher. This suggests a worse financial status of such households when compared to

households with non-disabled children. However, it is a noteworthy fact that two years earlier, in 2014, the distribution functions were even more divergent.

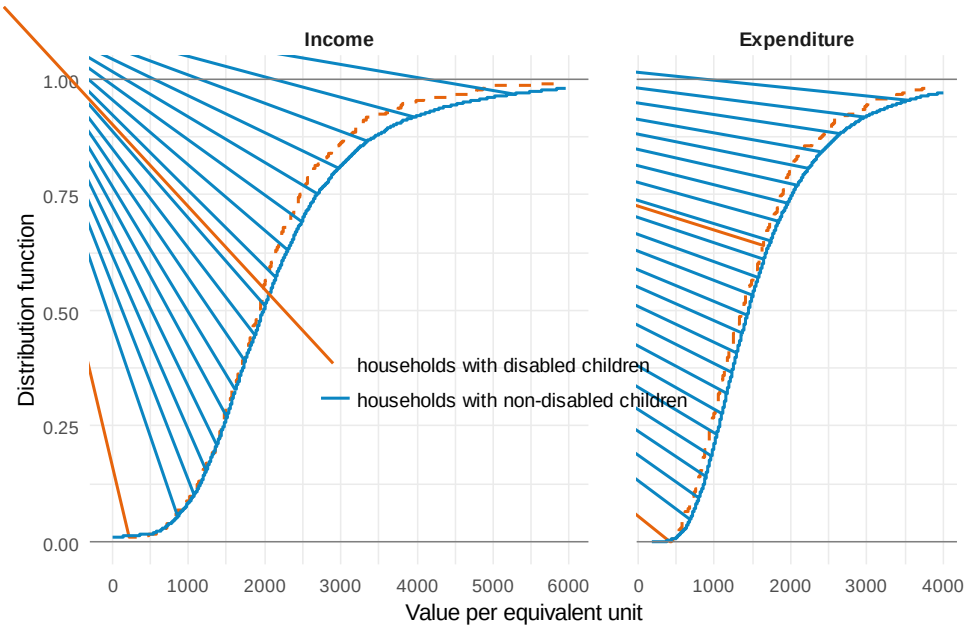


Figure 2. Empirical distribution functions of income and expenditure per equivalent unit in households in 2016

Source: Own study based on unit data from HBS 2016

An interesting question in HBS related to a subjective perception of financial status is the question about the income level in a household that the respondents would consider *very bad*, *unsatisfactory*, *barely satisfactory*, *good*, *very good*. Figure 3 illustrates the distribution of answers to this question using boxplots. An important point is that on the Y-axis a logarithmic scale is used due to the strong right-skewed distributions. On average, households with a disabled child had lower income expectations in all categories than households with non-disabled children. All three quartiles are lower in every income category. As compared to 2014, the values of income indicating a specific standard of living were higher in the case of both household groups. As an example, in 2016, the median of income indicated as *very bad* was PLN 750 in households with non-disabled children and PLN 667 in households with disabled children (as per equivalent unit). In 2014, however, the median was PLN 667 and PLN 588 respectively. The median of income indicated as *good* in 2016 in the first group of households was PLN 2174, and in the other group – PLN 2000. The 2014 results were PLN 1957 and PLN 1786 respectively.

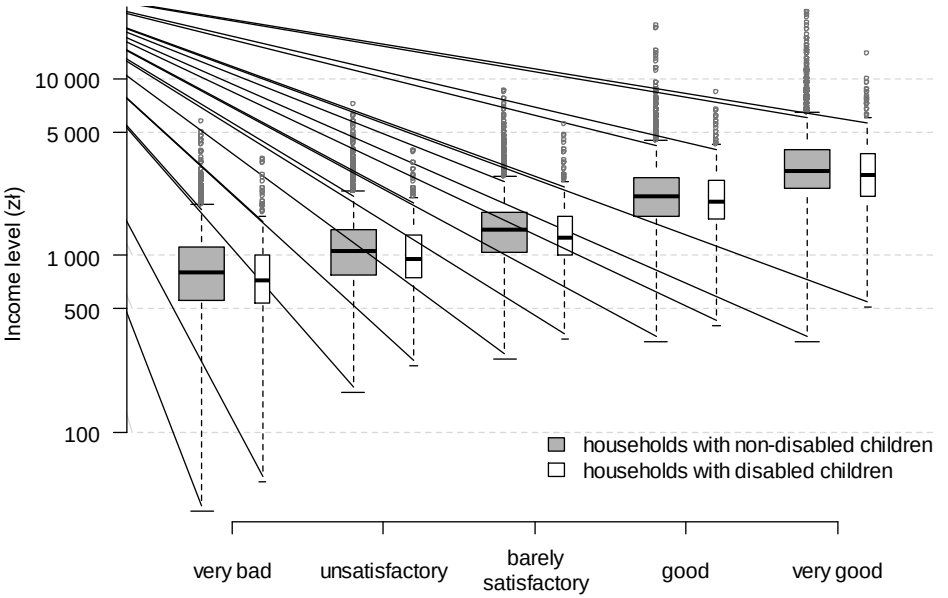


Figure 3. Distributions of answers to the questions about the income level that would mean a specific standard of living for a household

Source: Own study based on unit data from HBS 2016

The comparison above could suggest that respondents in households with a disabled child have on average lower expectations regarding the level of income. It is, however, a superficial pattern, since – as indicated above – households with disabled children are, generally speaking, poorer, whereas their expectations regarding the income level that would mean a specific standard of living are positively correlated with the actual financial status of a given household. To demonstrate this dependency, scatterplots are presented in Figure 4 for income levels indicating various standards of living and actual household expenditure, along with regression lines against the logarithms of both variables, separately for both types of households. The scatterplots and the slopes of the regression line confirm the positive correlation between the variables. Moreover, the regression line for households with disabled children virtually overlaps with the regression line of households with non-disabled children. (The differences in estimated regression coefficients and intercepts are not statistically significant. To show that this is the case, for each *j*-th standard of living, separate regression models were estimated: $\ln y_i^{(j)} = \beta_0 + \beta_1 \ln x_i + \beta_2 gosp_i + \beta_3 \ln x_i \cdot gosp_i + \varepsilon_i$, where $y_i^{(j)}$ is the value of income denoting the *j*-th standard of living, x_i is the actual household expenditure, $gosp_i$ is the household type (0 – with non-disabled children, 1 – with disabled children). These models were estimated for both groups of households together, but thanks to the variable $gosp_i$ and the interaction $\ln x_i \cdot gosp_i$, they can

be used to test the significance of differences in intercepts and regression coefficients between the models of the form: $\ln y_i^{(j)} = \alpha_0 + \alpha_1 \ln x_i + \varepsilon_i$, estimated separately for household groups, which are presented in Figure 4. The test for the significance of the difference in intercepts α_0 is the same as the significance test for coefficient β_2, while the test for the significance of difference in the regression coefficients α_1 is the same as the significance test of the β_3 coefficient. The p-values for $\beta_2$and β_3 coefficients, for each *j* level, are as follows: "very bad" (0.363, 0.365), "unsatisfactory" (0.197, 0.193), "barely satisfactory" (0.098, 0.091), "good" (0.357; 0.338), "very good" (0.372, 0.338). All p-values are greater than 0.05, so the regression line pairs in Figure 4 do not differ significantly from each other). It can be stated that if households with disabled children had higher income levels, their expectations regarding financial resources would also be higher.

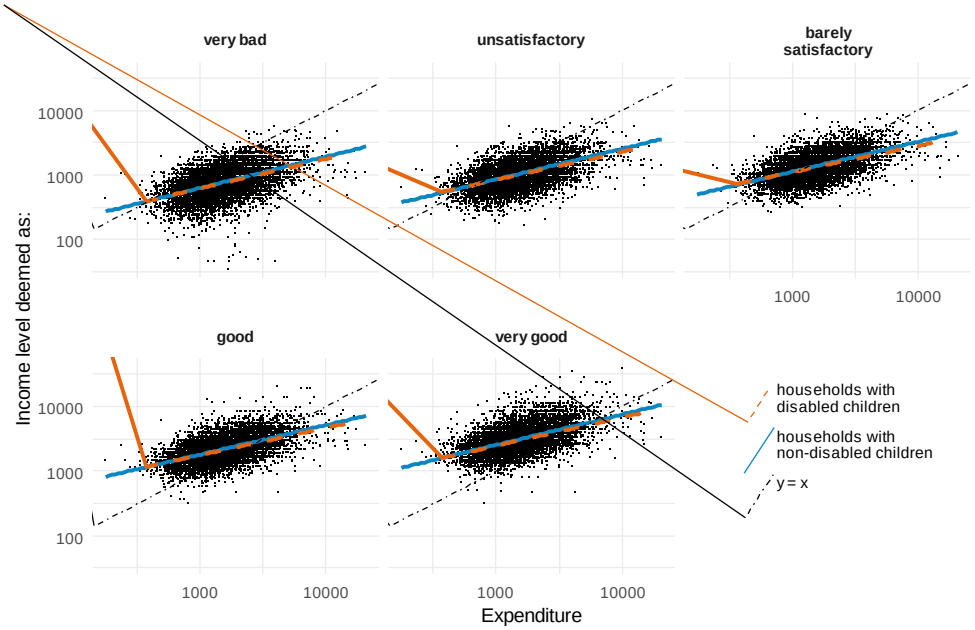


Figure 4. The relationships between income levels that would mean a specific standard of living for households and the actual financial situation of households measured with expenditure in 2016

Source: Own study based on unit data from HBS 2016

Another statement, validating the aforementioned conclusion, is the comparison of income levels indicating a certain standard of living for a household with an actual disposable income. Figure 5 shows the percentage of households the actual income of which was classified between the one considered by the respondents as indicating a certain standard of living in 2016 and 2014. For

example, in 2016 the percentage of households whose actual disposable income was between the level described as *good* and *very good* (by each household individually) was 28.5% for households with non-disabled children and 31.8% for households with disabled children. The highest percentage of households in the two groups and in the two periods had the disposable income which, according to their subjective criteria, was between the level of income described as *barely sufficient* and *good*.

On the basis of this figure, improvement in the subjectively viewed financial situation can be observed in both examined groups of households in 2016 as compared to 2014. Proportions of households with actual income higher than the income subjectively viewed as *good* and *very good* increased, while portions of households with income lower than *good* decreased (an exception to this rule is the group of households with disabled children with the income lower than the income subjectively described as *very bad,* but the difference is small). It must be emphasised that such subjective improvement was greater among households with disabled children.

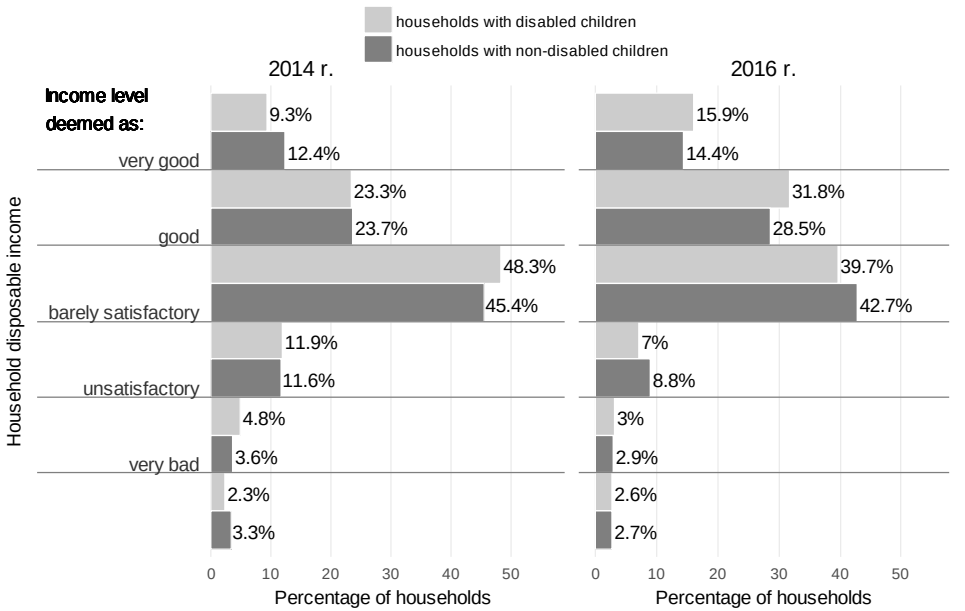


Figure 5. Distribution of disposable income classified by subjective income level for different standards of living in 2014 and 2016

Source: Own study based on unit data from HBS 2014 and 2016.

Figure 5 can also help determine the incidence of subjective poverty by showing a percentage of households with a disposable income below the individually set threshold. The threshold can be the *very bad*, *insufficient* or *barely sufficient* level. In the first case, it can be viewed as subjective extreme poverty.

For both household groups, this rate fluctuated between 2-3%. If the *barely sufficient* level is taken as the poverty threshold, then in 2014, 18.5% of households with non-disabled children and 19% of households with disabled children were subjectively impoverished, while in 2016 – it was 14.4% and 12.6% respectively (the incidence of poverty, i.e. headcount rates, is discussed in more detail in the next section).

Another element of the subjective evaluation of a household's income situation is requesting a respondent to provide an expression which best characterises the way of managing money in his/her household. Table 1 presents the distribution of answers to that question. In both groups, the income situation in 2016 was better when compared to 2014 – the percentage of answers *we have to live economically everyday* (which may be understood as living in privation) and *we have not enough even for basic needs* (which may be understood as living in poverty) decreased. It must be emphasised that similarly to the distributions shown in Figure 5 in the group of households with disabled children, the improvement was greater. The percentage of responses indicating poverty dropped by 3.9 percentage points (in the group with non-disabled children by 1.1 percentage points), while the percentage of responses indicating privation dropped by 15.7 percentage points (in the group with non-disabled children by 7.3 percentage points). In both years the disproportions between the groups were noticeable. More households with disabled children are in a worse financial situation. In 2016, 26.9% of households with disabled children (8.7 percentage points more than in the case of households with non-disabled children) had to live economically every day.

Table 1. Subjective evaluation of money management in households

	Households with children			
	non-disabled		disabled	
	2014	2016	2014	2016
Statement	*In %*			
we can afford some luxury	1.5	2.0	0.6	1.4
we have enough without special saving	11.7	15.2	5.0	9.1
we have enough for everyday living, but we have to save for greater purchases	59.1	63.7	45.2	59.9
we have to live economically everyday	25.5	18.2	42.6	26.9
we have not enough even for basic needs	2.1	1.0	6.6	2.7

Source: Own calculation based on unit data from HBS 2014 and 2016.

When analysing the phenomenon of poverty, it is worth to have a closer look at income inequality. In the case of households with disabled child and households with non-disabled children the Gini coefficient (based on income per person) was on a similar level, i.e. in 2016 it was 0.28 and 0.30 respectively

(the Gini coefficient based on income per equivalent unit was 0.27 and 0.30 respectively), which stands for a relatively low dispersion of income in the two types of households under study. It should be emphasised that in the case of the two types of households in question, income inequality decreased in 2016 in comparison with 2014 (when it was 0.34 and 0.33 respectively).

Another measure also related to poverty is the ratio of two extreme deciles (also quintiles). In 2016, the decile ratio for households with a disabled child was 3.31 (in 2014, 4.24). In the case of households with non-disabled children, the decile ratio was higher, namely 3.63 in 2016, and 4.38 in 2014. It is clear that income dispersion measured with a ratio of extreme deciles is significantly lower between the measured periods in the case of the two household groups.

3. Measures of poverty

In the research on poverty, no general definition of poverty has been established. Consequently, determining who is poor in the examined population is not that easy (Cowell, 2011; Thon, 1979). Therefore, the analysis of poverty must by multifaceted. Generally, a household can be classified as poor when its income or expenditure level is lower than the established threshold (Lisicka, 2013; Panek, 2014, p. 204; Szarfenberg and Szewczyk, 2010, pp. 29-30). In this study five different types of threshold were used, which could by divided into two groups, objective and subjective. In the objective approach, legal and two relative lines were used, while in the subjective approach – the Leyden method and the subjective poverty line were used (Panek, 2011, pp. 35-38).

In general, household expenditure is a better measure of wealth than income (Klugman, 2002, p. 30), therefore expenditure was used in the case of objective thresholds. But in the case of the subjective approach income was used since the question concerning the subjective evaluations refers directly to income.

The legal line is set in order to apply for a benefit from the social service system. It is determined separately for households with a different number of people (irrespective of their age). The relative line most often equals 60% of the median (used by Eurostat) or 50% of the mean (used by Central Statistical Office in Poland). It allows one to identify the poor who are far from the average level of expenditure realised in a given society.

The Leyden method uses answers to the question about the level of household income which the respondents would consider *very bad, unsatisfactory, barely satisfactory, good, very good* (see the previous section). The obtained answers are used to estimate the so-called individual income wealth (utility) functions, which have a form of a distribution function (here log-normal distribution). The poverty line (individual for each household) is set to such level of income for which the utility function takes a certain low, arbitrary chosen value δ (value of the distribution function). In the conducted analysis three values were adopted: 0.3, 0.4 and 0.5.

In order to determine the subjective poverty line, an answer to the question about the income essential to “make ends meet” is used. In the HBS there is no such question, but the same question as in Leyden method can be used, taking into account only the *barely satisfactory* variant, since it has the closest meaning to “making ends meet”.

Table 2. Poverty lines in 2016 (annual average values)

Type of poverty line	2-person household (1 adult + 1 child up to 14 years of age)	4-person household (2 adults + 2 children up to 14 years of age)
	In PLN	
legal	1 028.00	2 056.00
	Based on expenditure	
60% of median	1 157.86	1 870.39
50% of mean	1 131.82	1 828.33
	Based on income	
Leyden (δ = 0.3)	1 261.27	1 482.36
Leyden (δ = 0.4)	1 640.96	1 928.61
Leyden (δ = 0.5)	2 098.56	2 466.43
Subjective poverty line	1 755.76	2 082.19

Source: GUS, 2017, p. 11; Own calculation based on unit data from HBS 2016.

All the line values in Table 2, except for the legal line which remained the same throughout the year, are averaged for the whole year, while subjective limits are additionally averaged for all households. The lines were stated individually for households comprising one adult and one child (e.g. a single parent who raises the child on his or her own) and households comprising two adults and two children (e.g. a married couple with two children). The values were provided for information purposes only as they were not directly used, except for the legal limit, to calculate the headcount rates. The headcount rate (often referred to as "at risk of poverty rate"), i.e. the percentage of persons in households considered to be impoverished, for the relative values was calculated with respect to the lines calculated individually for each quarter (and in comparison with expenditure), while for subjective values – with respect to the lines set individually for each household (and in comparison with income). Naturally, objective lines are the same for households with the same composition, regardless of the presence of a child with a disability.

As for the legal and relative lines, the headcount rate in households with disabled children is much higher than in households with non-disabled children (Table 3). Bigger differences can be observed in 2014: for instance, the difference in the case of relative lines was approx. 10 pp., while in 2016 – 4.6 pp. for the 50%-mean line and 5.7 pp for the 60%-median line. As far as subjective lines are concerned, the situation is different. Here, the headcount index was nearly the same for the two groups of households under consideration. One exception is the values for 2016 calculated with the use of the Leyden line of δ = 0.5, and the subjective poverty line where the percentage of poor households among households with disabled children is lower than the same percentage in the households with non-disabled children. One should emphasise that over the span of the two years in question, the range of poverty decreased, regardless of the definition of impoverished households.

Table 3. Headcount rates in households with disabled children and non-disabled children in 2014 and 2016 (in %)

Households with children	Type of poverty line						
	legal	50% of mean	60% of median	Leyden (δ = 0.3)	Leyden (δ = 0.4)	Leyden (δ = 0.5)	subj. poverty line
				2014			
non-disabled	17.5	16.5	17.3	8.3	14.5	24.2	18.5
disabled	30.0	26.5	27.3	8.1	14.8	23.5	19.0
				2016			
non-disabled	17.4	12.8	14.0	6.4	10.9	18.3	14.4
disabled	25.7	17.4	19.7	6.2	10.3	16.3	12.6

Source: own calculation based on unit data from HBS 2014 and 2016.

Another measure of poverty is its depth, i.e. the poverty gap index. The depth calculated with respect to the relative poverty line (50% of mean expenditure) was at the similar level in the case of the two household groups in 2014 and amounted to 20.6 for households with non-disabled children, and 20.2 for households with disabled children. This means that the average expenditure of impoverished households was by approximately 20% lower than the poverty level calculated as 50% of mean expenditure for all households. In 2016, the depth of poverty in the case of households with a disabled child remained at the same level and amounted to 20.2, while in the case of households with non-disabled children it decreased to the level of 18.9.

4. Factors related to poverty

In the next step of the analysis, the aim was to check if and in what terms impoverished households differ from non-impoverished households and whether such differences are the same in households with non-disabled children and in households with disabled children. To this end, classification trees and the model of logistic regression were employed. The dependent variable in both cases was a dummy variable defined as follows:

$$Y = \begin{cases} 1 & \text{for poor households} \\ 0 & \text{for non} - \text{poor households} \end{cases}$$

A household was deemed impoverished if its expenditure per equivalent unit was lower than the relative poverty line understood as 50% of mean expenditure.

The following nine features of a household which were deemed most important in the context of the phenomenon under analysis and which could be obtained from HBS were selected as explanatory variables:

- number of children (18 or under),
- number of unemployed,
- number of adults,

- age of youngest child,
- education of household head (reference person),
- main source of household income,
- urbanisation degree (class of place of residence),
- voivodship,
- disabled parent (is one of the parents disabled?).

First, the exploratory technique of data analysis was used, i.e. classification trees. Binary trees were used; information gain was used as the criterion for split (see Gatnar, 2001, pp. 33-34); the division was stopped either at the maximum depth of the tree (which was set to 4) or the minimum leaf size (which was set to 1% of the number of units). The classification trees obtained in this way are presented in Figure 6 and Figure 7, separately for the set of households with non-disabled and disabled children. In the figures, the branch widths are proportional to the number of units in sub-sets. The nodes contain information about the variable used in the division and its variants, or a split point; below one can find information about the portion of impoverished households in a node (in the rectangle the shadowing intensity of which depends on the level of the fraction), while at the bottom, information about the number of units in a node (as a percentage of the whole set). Units which satisfy the condition of a node division are sent to the left side, the remaining ones – to the right side. The division is arranged in a way that a group with a smaller fraction of poor households goes to the left.

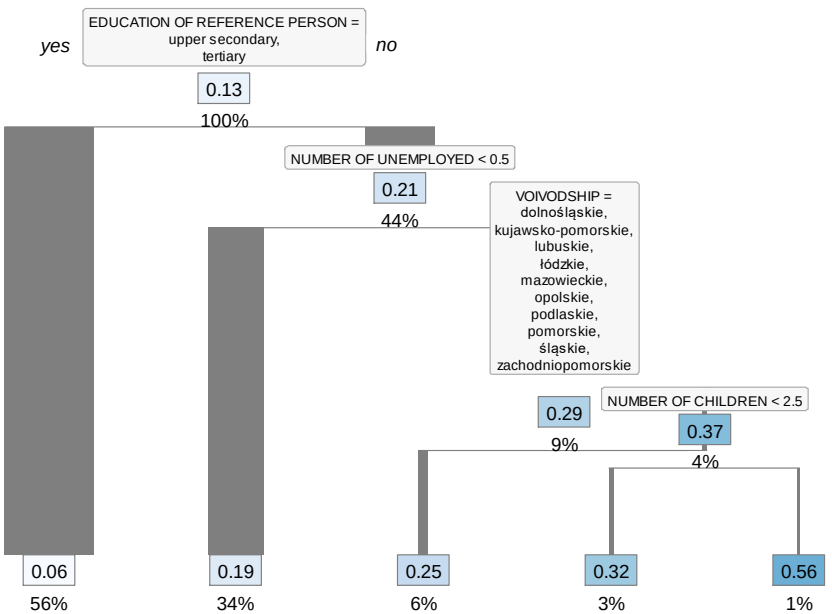


Figure 6. Classification tree of belonging to impoverished households, for households with non-disabled children in 2016

Source: Own study based on unit data from HBS 2016.

In the case of both trees, the first variable used for the division, i.e. the most discriminating variable, turned out to be the *reference person's education*. For households with non-disabled children, upper secondary or tertiary education determined whether the household belonged to the group with the smallest portion of poor households. For households with disabled children, only tertiary education ensured such a division. What is important is that the households in which the reference person had tertiary (or upper secondary in the case of households with non-disabled children) education are not divided further on, so they are relatively homogenous groups, with a low headcount rate.

Households in which the reference person's education is lower, were further divided according to the variable *number of unemployed* in a household. In households with at least one unemployed person, the place of residence became important (*voivodship*), and subsequently – *number of children* (more than 2 children in this node significantly increased chances for poverty). In the case of households with disabled children, where the reference person had no tertiary education, what mattered first was the place of residence (*voivodship*), while secondly – the presence of unemployed (more than one). In the node on the lowest level, chances for finding oneself in a group of poor households increase significantly if the *main source of income* are types of pension other than old-age pension.

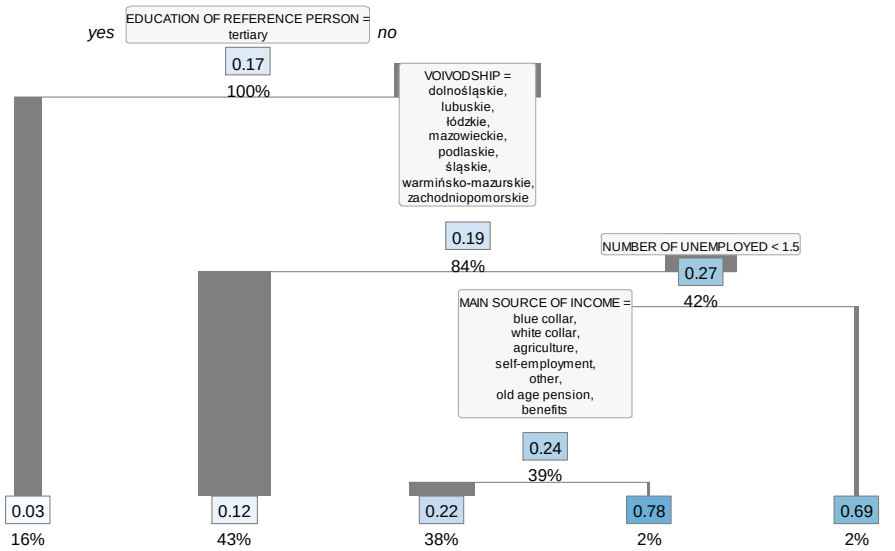


Figure 7. Classification tree of belonging to impoverished households, for households with disabled children in 2016

Source: Own study based on unit data from HBS 2016.

In general, classification trees for households with non-disabled and disabled children are similar. In both cases, only four variables were used for the division, of which three variables were the same, although the way they split the data and the level at which they were used differed slightly. Nevertheless, the general rules are the same – a larger percentage of poor households is associated with a lower level of education, unemployment of at least one member of the household, and a place of residence in the south-eastern voivodships.

A different method of verifying which factors affect the probability of finding oneself in a group of impoverished households is a logistic regression (Fahrmeir, et al., 2013). Just as in the case of classification trees, models were estimated separately for households with non-disabled and disabled children, with the same set of variables. Once the full model was estimated, a stepwise elimination of insignificant variables was applied according to AIC criterion. The basic results of the models are presented in Table 4.

Table 4. Odds ratios for changes in the value of explanatory variables in the logistic regression models for the probability of belonging to the group of poor households in 2016 (values in bold indicate statistically significant variables at the level of 0.05)

	Odds ratio	
Variable: d= *value of change* or Variable: *option under study – reference option*	households with non-disabled children	households with disabled children
Number of children: d=*1*	**1.17**	x
Number of unemployed: d=*1*	**1.55**	**1.82**
Number of adults: d=*1*	**1.29**	**1.27**
Age of youngest child: d=*10*	**1.39**	x
Reference person's education: *lower secondary and lower – upper secondary*	**2.44**	**2.28**
Reference person's education: *basic vocational - upper secondary*	**1.80**	1.05
Reference person's education: *tertiary - upper secondary*	**0.38**	**0.21**
Main source of income: *white-collar wage work - blue-collar wage work*	**0.72**	x
Main source of income: *use of private farm in agricultural - blue-collar wage work*	1.19	x
Main source of income: *self-employment – blue-collar wage work*	**0.64**	x
Main source of income: *other - blue-collar wage work*	**2.33**	x
Main source of income: *old age pension - blue-collar wage work*	0.89	x
Main source of income: *other types of pension - blue-collar wage work*	**1.57**	x
Main source of income: *benefits - blue-collar wage work*	**2.11**	x
Urbanisation degree: *densely populated area – medium populated area*	0.93	1.04

Urbanisation degree: *sparsely populated area - medium populated area*	**1.30**	1.77
Voivodship: *dolnośląskie - pomorskie*	0.97	x
Voivodship: *kujawsko-pomorskie - pomorskie*	1.12	x
Voivodship: *lubelskie - pomorskie*	1.39	x
Voivodship: *lubuskie - pomorskie*	0.76	x
Voivodship: *łódzkie – pomorskie*	0.77	x
Voivodship: *małopolskie - pomorskie*	**1.66**	x
Voivodship: *mazowieckie - pomorskie*	1.02	x
Voivodship: *opolskie - pomorskie*	0.68	x
Voivodship: *podkarpackie - pomorskie*	**1.52**	x
Voivodship: *podlaskie - pomorskie*	1.25	x
Voivodship: *śląskie – pomorskie*	1.14	x
Voivodship: *świętokrzyskie - pomorskie*	**1.74**	x
Voivodship: *warmińsko-mazurskie - pomorskie*	**1.86**	x
Voivodship: *wielkopolskie - pomorskie*	**1.60**	x
Voivodship: *zachodniopomorskie - pomorskie*	0.95	x
Disabled parent: *yes – no*	x	1.88

Source: Own calculation based on unit data from HBS 2016.

In the case of the model for households with non-disabled children, all variables, except for *disabled parent*, were preserved, which to a large extent results from the large sample size. On the other hand, in the model for households with disabled children, only five variables were preserved, which partially results from the small sample size. Quality measures (Table 5) show that both models are moderately fitted to the data. It should be emphasised, however, that the objective of models under assessment, both logistic regression and classification trees, was not developing a predictive tool, but finding out if any relationships exist between the variables under analysis.

Table 5. Quality measures of logistic regression models

	Households with non-disabled children	Households with disabled children
Area under the ROC curve	0.77	0.73
Sommers' D_{xy}	0.54	0.47
Nagelkerke R^2	0.18	0.16
Likelihood ratio test	$\chi^2_{v=31} = 1154.7$ $p < 0.0001$	$\chi^2_{v=8} = 59.1$ $p < 0.0001$

Source: Own calculation based on unit data from HBS 2016.

The interpretation of outcomes in Table 4 is as follows: in the first column there is the name of a variable, followed by a colon, then, for numerical variables, the value of change denoted by the letter *d*, or, for categorical variables, the value for a given odds ratio followed by the reference value for this variable. And the odds ratio is the ratio of the odds of being poor when explanatory variable is greater by *d* (in the case of numerical variables) or equals specific value (in the case of categorical variable), and the odds of being poor when explanatory variable is not changed (in the case of numerical variables) or equals the reference value (in the case of categorical variable). For example: the odds ratio for *number of unemployed* for households with non-disabled children is 1.55 (for *d=1*), which means that increasing the number of unemployed persons by 1 lengthens the odds of being poor by 55%. Another example: the odds ratio for *reference person's education* for households with disabled children (for *lower secondary and lower – upper secondary*) is 2.28, which means that the odds of being poor when the education attainment level is *lower secondary and lower* is by 128% greater compared to *upper secondary* level.

When one compares the odds ratios for statistically significant variables in both models, it can be seen that the direction of impact for specific variants is always the same, but its strength is somewhat different. In the case of the two household groups, tertiary education markedly reduces the odds of becoming poor; however, in households with disabled children, this effect is more pronounced. Lower secondary and lower education, as well as basic vocational education, markedly increase the chances of falling into poverty, but this effect is weaker in households with disabled children. The presence of an unemployed person in a household has a stronger negative impact in families with disabled children.

As an additional element of the assessment of impact of specific explanatory variables on the response variable, the Wald statistics was calculated (Harrell, 2015, p. 191, 194) to test the significance of variables (the statistics have asymptotic chi-squared distribution) and a ranking of predictor importance was created, which is presented in Figure 7. In both models, as in the case of classification trees, the variable having the strongest impact on the chance of belonging to the poor is the *reference person's education*. In the case of households with non-disabled children, the *main source of income* and *voivodship* ranked second and third, whereas in the model for households with disabled children they did not occur at all, as they were removed at the stepwise elimination stage. The second most important variable in the group of households with disabled children turned out to be the *number of unemployed*. Variables that do not affect the chances of finding oneself in a group of impoverished or whose effect is relatively small in both groups of households, are: *age of youngest child*, *urbanisation degree*, *number of children* and *disabled parent*.

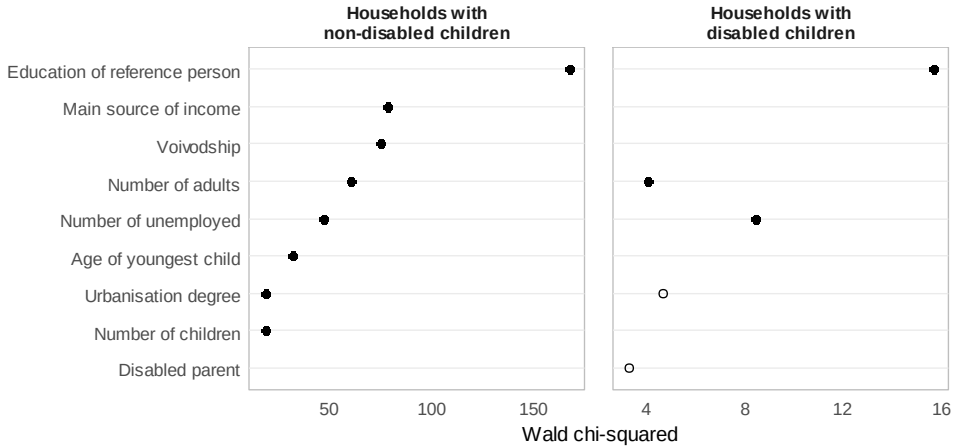


Figure 7. Ranking of predictors in logistic regression models – based on the Wald χ^2 (filled point means a statistically significant variable at the significance level of 0.05, unfilled – significant at the level of 0.1)

Source: Own study based on unit data from HBS 2016.

The above analysis was also conducted for the data from 2014. As far as the classification trees are concerned, for households with non-disabled children the division was very similar to the one presented above for the 2016 data. In the case of households with disabled children, however, the division was very different. The strongest discriminating variable was the *number of unemployed,* followed by the *reference person's education* for the subset of households without the unemployed. Apart from that, the tree was more extensive with 11 leaf nodes compared to five leaf nodes in 2016. Additionally, eight variables were used, and the final subsets were more homogenous.

Both models of logistic regression from the period of two years earlier were similar in general. The differences that could be observed in both types of households included: a stronger negative effect (i.e. greater chance for poverty) of *number of children, number of unemployed* and living in a sparsely populated area. The decrease of the negative impact of those variables in 2016 can be a result of better economic prosperity (lower unemployment rate) and the introduction of the "Family 500+" programme in mid-2016 (thus, the smaller impact of a large number of children on poverty). In a sense, the consequence of these changes is the fact that both regression models in 2016 were fitted to the data worse than in 2014, so it is now more difficult to determine typical characteristics of the poor households on the basis of available data.

5. Conclusions

Based on the analysis, it can be stated that households with a disabled child were in a worse financial situation when compared to households with non-disabled children – both in 2014 and in 2016. Households with a disabled child tend to rely to a greater extent on all sorts of benefits, and so they are more vulnerable to changes in state policies in this area. The 2016 introduction of the "Family 500+" support programme is likely to have been one of the factors that contributed to poverty reduction in both groups under analysis; however a more pronounced improvement can be noticed in households with a disabled child, which led to the reduction of the disproportion in financial situations reported by the two household types.

The factor that discriminates the most between poor and non-poor households, especially in households with a disabled child, was the education attainment level of the household head. In households where the person with the highest income was a university graduate, the percentage share of the poor was the lowest. Furthermore, worse financial condition are linked to unemployment of at least one family member and the fact of residing in the south-eastern voivodships of Poland.

Households with non-disabled children located in less densely populated areas were exposed to a greater risk of poverty as compared to households from more populous areas. However, the impact of this variable is moderate. In the case of households with a disabled child, the impact of the size of their place of residence is even weaker.

An interesting observation is the impact of the number of children on the risk of falling into poverty. It is quite common to associate multi-child families with financial hardship. Although in 2014 the number of children was a fairly important factor contributing to the risk of poverty, in 2016 the impact of this variable was far weaker. What might have brought about this change is the aforementioned "Family 500+" programme, but a different explanation could be the fairly strong impact of the variable *number of adults*. The analysis is based on the assumption that a child is a person aged 18 or under, while all the other household members are treated as adults. In the next stage of the analysis it would be of interest to check whether the "adult children", persons over 18 still living in the household with their parents, are a factor increasing the risk of poverty.

REFERENCES

ANYAEGBU, G., (2010). Using the OECD equivalence scale in taxes and benefits analysis. Economic & Labour Market Review, 4(1), pp. 49–54.

BIRD, K., (2013). The Intergenerational Transmission of Poverty: An Overview. In: A. Shepherd and J. Brunt, eds. Chronic Poverty: Concepts, Causes and Policy, Rethinking International Development Series. London: Palgrave Macmillan UK. pp. 60–84.

COWELL, F., (2011). Measuring Inequality, Oxford University Press.

EUROPEAN COMMISSION, (2010). Communication from the Commission, Europe 2020, A strategy for smart, sustainable and inclusive growth, Brussels.

FAHRMEIR, L., KNEIB, T., LANG, S., MARX, B., (2013). Regression, Models, methods and applications, Heidelberg: Springer.

GATNAR, E., (2001). Nieparametryczna metoda dyskryminacji i regresji, Warszawa: Wydawnictwo Naukowe PWN.

GŁÓWNY URZĄD STATYSTYCZNY, (2015a). Ubóstwo w Polsce w latach 2013-2014, Warszawa: GUS.

GŁÓWNY URZĄD STATYSTYCZNY, (2015b). Aneks tabelaryczny do opracowania sygnalnego Ubóstwo ekonomiczne w Polsce w 2014 r. Warszawa: GUS.

GŁÓWNY URZĄD STATYSTYCZNY, (2017a). Zasięg ubóstwa ekonomicznego w Polsce w 2016 roku, Warszawa: GUS.

GŁÓWNY URZĄD STATYSTYCZNY, (2017b). Ubóstwo w Polsce w latach 2015 i 2016, WARSZAWA: GUS.

GŁÓWNY URZĄD STATYSTYCZNY, (2018). Zasięg ubóstwa ekonomicznego w Polsce w 2017 roku, Warszawa: GUS.

GOLINOWSKA, S., MARECKA, Z., STYRC, M., CUKROWSKA, E., CUKROWSKI, J., (2008), Od ubóstwa do wykluczenia społecznego. Warszawa: IPiSS.

HAGENAARS, A., DE VOS, K., ZAIDI, M. A., (1994). Poverty statistics in the late 1980s: research based on micro-data, Luxembourg: Office for Official Publications of the European Communities.

HARPER, C., MARCUS, R., MOORE, K., (2003). Enduring Poverty and the Conditions of Childhood: Lifecourse and Intergenerational Poverty Transmissions. World Development, 31(3), pp. 535–554.

HARRELL, F. E. J., (2015). Regression modeling strategies, Heidelberg: Springer.

KLUGMAN, J., (2002). A sourcebook for poverty reduction strategies, Vol. 1: Core Techniques and Cross-Cutting Issues, Washington: The World Bank.

KRUSZYŃSKI, K., WARZYWODA-KRUSZYŃSKA, W., (2011). Dziedziczenie biedy i wykluczenia społecznego – w perspektywie lokalnej polityki społecznej. In: R. Szarfenberg, ed. 2011, Ubóstwo i wykluczenie społeczne w Polsce, Warszawa: Kampania Przeciw Homofobii. pp. 49–55.

LISICKA I., (2013). Pomiar ubóstwa gospodarstw domowych w Polsce – ujęcie klasyczne, Ekonomika i Organizacja Gospodarki Żywnościowej, (102), pp. 37–48.

ŁUKASIEWICZ, P., KOSZELA, G., ORŁOWSKI, A., (2006). Wpływ wyboru skali ekwiwalentności na wyniki w zakresie pomiaru ubóstwa i koncentracji

dochodów. Ekonomika i Organizacja Gospodarki Żywnościowej, 60, pp. 207–217.

PANEK, T., (2011). Ubóstwo, wykluczenie społeczne i nierówności. Teoria I Praktyka Pomiaru, Warszawa: Oficyna Wydawnicza SGH.

PANEK, T., (2014). Ubóstwo i wykluczenie społeczne. In: T. Panek, ed. 2014. Statystyka Społeczna, Warszawa: PWE.

SZARFENBERG, R., SZEWCZYK, Ł., (2010). Badania ubóstwa – perspektywa ilościowa i jakościowa. In: R. Szarfenberg, C. Żołędowski, M. Theiss, eds. 2010, Ubóstwo i wykluczenie społeczne – perspektywa poznawcza. Warszawa: ELIPSA.

THON, D., (1979). On Measuring Poverty. Review of Income and Wealth, 25(4), pp. 429–439.

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. 119–132, DOI 10.21307/stattrans-2019-027
Submitted – 02.03.2018; Paper ready for publication – 07.02.2019

CPI INFLATION IN AFRICA: FRACTIONAL PERSISTENCE, MEAN REVERSION AND NONLINEARITY

O. S. Yaya[1], **O. J. Akintande**[2], **A. E. Ogbonna**[3], **H. M. Adegoke**[4]

ABSTRACT

Price stability has been one of the key mandates that apex monetary authorities strive to achieve globally. While most developed economies have achieved single digit inflation rates, most developing economies, especially African countries still experience alarming double-digit inflation rates. This paper therefore examined the dynamics of inflation in sixteen African countries. We employed the fractional persistence framework with linear trend and non-linear specifications based on Chebyshev's polynomial in time. The results indicated nonlinear time trend in inflation for most of the countries. With the exception of Burkina Faso, which exhibited plausibility of naturally reverting to its mean level, the majority of the selected African countries would require stronger interventions to revert their observed inflationary levels to their mean levels.

Key words: Africa, Fractional Integration, Inflation Rate, Mean Reversion, Nonlinear Trend, Structural Break.

JEL: 22.

1. Introduction

The persistent increase in the general price level of goods and services in an economy over a period of time is a feat that cannot be ruled out during the process of policy formation with regards to economic activities. This being the result of the attendant consequence of its possible impacts and effects, either positively or negatively, on the purchasing power of the economy's medium of exchange and unit of account within the economy (Paul *et al.,* 1973). On the negative impact, the general price increase could lead to commodity scarcity, increased opportunity cost of holding money, investment drought as a result of

[1] Economic and Financial Statistics Unit, Department of Statistics, University of Ibadan, Nigeria. E-mail: os.yaya@ui.edu.ng; o.s.olaoluwa@gmail.com. ORCID ID: http://orcid.org/0000-0002-7554-3507
[2] Laboratory for Interdisciplinary Statistical Analysis, Department of Statistics, University of Ibadan, Nigeria. E-mail: aojsoft@gmail.com. ORCID ID: https://orcid.org/0000-0001-5458-9795.
[3] Centre for Econometric and Allied Research, University of Ibadan & Economic and Financial Statistics Unit, Department of Statistics, University of Ibadan, Nigeria. E-mail: ae.ogbonna@cear.org.ng. ORCID ID: http://orcid.org/0000-0002-9127-5231.
[4] Economic and Financial Statistics Unit, Department of Statistics, University of Ibadan, Nigeria. E-mail: hammedmuminiadegoke@gmail.com.

uncertainty of future prices. Positively, it reduces the real burden of public and private debt, reduces unemployment due to nominal wage rigidity and provides monetary authorities with a tool to stabilize the economy, since interest rates are nominally kept above zero (Mankiw, 2001). The severity of this general price level could be low or moderate fluctuations in the real demand for goods and services, or changes in available supplies, especially, during scarcity. Consensually, a long sustained period of inflation is perceived to be the outcome of a faster growth rate of money supply in comparison with economic growth rate. Rather than a zero or negative inflation, a low inflation is preferred as it reduces the severity of economic recession by enabling the labour market to adjust more quickly in a downturn and, consequently, reducing the risk of liquidity trap, which may prevent monetary authorities from performing its stabilizing role for the economy (Svensson, 2003).

The literature is replete with various country specific studies that tried to investigate the dynamics of inflation. These include studies on Nigeria (Adenekan and Nwanna, 2004; Odusanya and Atanda, 2010; Imimole and Enoma, 2011; Bawa et al., 2016); Ethiopia (Wolde-Rufael, 2008); Ghana (Adu and Marbuah, 2011); South Africa (Nell, 2000, 2006; Hodge, 2002, 2006, 2009; Fedderke and Schaling, 2005; Burger and Du Plessis, 2006; Burger and Marinkov, 2005; Vermeulen, 2015); Egypt (Ali, 2011; Osama, 2014; Osama, 2014); Kenya (Kaushik, 2011; Kimani and Mutuku, 2013; Kirimi, 2014); Cameroun (Tabi and Ondoa, 2011). These studies basically focused on the impact of inflation on economic growth, local currency value, money supply and stock prices. In some other studies, inflation thresholds of 6% and 9%, respectively, were obtained for Nigeria (Fabayo and Ajilore, 2006; Ajide and Olukemi, 2010), while Phiri (2013) obtained inflation threshold of 22.5% for Zambia. Methodologically, Fielding *et al.*, 2004, and Mikkelsen and Peiris, 2005, employed VAR in their study of inflation. Barnichon and Peiris, 2008, employed the heterogeneous panel cointegration methodology and established the significant role of the output gap and the real money gap on the evaluation of inflation, with the money gap playing a larger role. Caporale, Carcel and Gil-Alana (2015) investigated inflation persistence and nonlinearity using fractional integration approach in five African countries such as Angola, Botswana, Lesotho, Namibia and South Africa and found nonlinear persistence in the case of Angola and Lesotho, while linear persistence was found in the remaining three countries. Boateng et al. (2017) investigated inflation persistence in Ghana and South Africa by using CPI inflation. The authors applied the fractional autoregressive moving average model with heteroscedasticity innovations. The results obtained showed evidence of mean reverting persistence with asymmetric effects of shocks on the conditional mean of CPI inflation of the two countries. In further enhancing inflation forecast precision, the incorporation of the mixed data sampling methodology was suggested (*see* Salisu and Ogbonna, 2017). Although this is yet to be applied in the African context, it has proven to significantly improve an economy's inflation prediction, especially with regards to OECD member countries.

Careful study on the dynamics of inflationary process in Africa will therefore help in the choice of policy models and estimation methods. There is still an ongoing debate on whether the inflation rate in Africa is a stationary I(0) or nonstationary I(1) process. As our contribution in this paper, we carried out

extensive time series analysis using fractional integration framework to investigate African inflationary persistence and nonlinearity.

The methodological approach in this paper is hardly applied in investigating inflation dynamics and other economic time series. Following Granger and Hyung (2004) and Ohanissian et al. (2008), a process follows long range dependency if over a long time span, far apart observations are still strongly correlated with the current observations. The time processes in mean reverting series are not integrated of order 1 (non-stationary) but are integrated of fractional order less than 1 and the test of fractional order confirms that the I(1) hypothesis should be rejected. In economic theory, mean reversion means that the series can still revert itself to its mean level after the initial shock on the economy, as propelled by a high inflation rate. Nonstationarity in inflation rates means that shocks to inflation have a permanent effect and strong policies would be required by monetary/economic agencies to revert the inflation rate back to its mean level. Stationary or mean reverting inflation means that inflation incurs a lower cost for the monetary/economic agency in the pursuit of monetary policies (Cecchetti and Debelle, 2006). Stationarity/nonstationarity of the inflation rate is controversial, while many authors believe that the series follow I(0) stationary process based on the fact that the generating time series is log-price I(1). Other authors are of the opinion that the series is nonstationary I(1), and it should be included in the system of cointegrating variables (Gil-Alana, Shittu and Yaya, 2012). Using fractional persistence, inflation is neither I(0) nor I(1) but I(d), where d is a value between 0 and 1. Noting that long memory models overestimate the degree of persistence of the series in the presence of structural breaks (Ben Nasr et al., 2014; Gil-Alana, Cunado and Gupta, 2015), and also, with the availability of long time series for many countries, these are very likely. Thus, we supplement our long memory model to accommodate for nonlinear deterministic trends as in Cuestas and Gil-Alana (2016)[5]. The approach employed Chebyshev polynomials in cosine function of time up to third orders in fractional persistence framework to determine nonlinearity in time series, in a smooth fashion, rather than an abrupt fashion as in Gil-Alana (2008).

Gil-Alana, Shittu and Yaya (2012) analyzed Nigerian inflation rates using long range dependence in fractional integration incorporating structural breaks. In their results, they observed long memory behaviour in inflation rates with different periods of breaks. Gil-Alana, Yaya and Solademi (2016) examined unit roots, structural breaks and nonlinearity in inflation rates in G7 countries. Based on classical unit root decisions, the authors first observed inclusive results in the stationarity/nonstationarity of inflation rates in these countries. A test based on fractional unit root analysis showed nonstationarity I(1) process for inflation rates in the case of UK, Canada, France, Japan and the US, while in the case of Italy, evidence of I(d > 1) was observed and in the case of Germany, mean reversion was observed.

Specifically, in this paper, we investigate long range dependency, mean reversion and nonlinearity in inflation dynamics of African countries using fractional persistence approach. This paper is the first, among many, investigating

[5] It is a known fact that fractional persistence, nonlinearities and structural breaks are closely related properties in time series (see Diebold and Inoue, 2001; Kapetanios, Shin and Snell, 2003; Granger and Hyung, 2004; Gil-Alana, Cunado and Gupta, 2015).

unit root in African CPI inflation and inflation rates. The findings clearly expose readers to nonlinear dynamics of CPI, which has effect on the construct of cointegrating econometric models involving CPI inflation. Importance of time dynamics of inflation to monetary policy agents in Africa also gingered this write up.

Following the introductory section, the rest of the paper is presented as follows: Section 2 presents fractional persistence framework for nonlinear deterministic trend. Section 3 presents the results and discussion while Section 4 renders concluding remarks and policy implications.

2. Methodology

A time series process $\{y_t,\ t=0,\pm1,\pm2,...\}$ is integrated of order zero, $I(0)$, if it is a covariance stationary process with a spectral density function that is positive and finite at zero frequency. Thus, a process is integrated of order d if it can be represented as,

$$(1-B)^d y_t = u_t, \quad t=0,\pm1,\pm2,... \tag{1}$$

with $y_t = 0$ for $t \le 0$, and $d > 0$ where B is the backward shift operator such that $By_t = y_{t-1}$ and u_t is $I(0)$ process. The parameter d therefore determines the size of differences needed to render a series stationary $I(0)$. Recall, in the case of classical unit integration, d is restricted as integer, while in fractional persistence, a much richer degree of flexibility in the values of *d* is allowed.

A very appealing case of fractional persistence is $I(0 < d < 0.5)$ time series process, known as long memory. In the sense that the spectral density function of the process is unbounded at the lowest frequency. By a way of time domain definition, let $\gamma(h) = (y_t, y_{t+h})$ be the autocovariances at lag h of the stationary process $\{y_t;\ t \in \square\}$, then the autocovariance of such long memory process is unbounded and infinite, that is,

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| = \infty \tag{2}$$

Thus, in terms of hyperbolic decay of autocovariances,

$$\gamma(h) \square\ h^{2d-1}\ell_1(h) \tag{3}$$

As $h \to \infty$ and $\ell_1(h)$ is a slowly varying function.[6]

The values of fractional d have many implications from both economic and statistical viewpoints. For example, if $d = 0$ in equation (1), $y_t = u_t$, the process y_t is then said to be $I(0)$, stationary with autocovariances decaying exponentially. For $0 < d < 0.5$, as in long memory process, the autocovariances as well as autocorrelations decay at much slower and hyperbolic rates compared to when $d = 0$. For $0.5 < d < 1$, y_t becomes nonstationary as the variance of the partial sums increases in magnitude. In economic terms, $d < 1$ implies that the series is "mean reverting", in the sense that shocks to the series disappears in the long run, and the series reverts back to its mean level. For $d \geq 1$, this is a nonstationary stance, where the effect of any shocks to the series persists forever.

Actually, the mean reversion case is relevant in the context of the inflation rate, since shocks imparts differently in the short and long run, depending on the value of the fractional differencing parameter d.

Robinson (1994) incorporates equation (1) into the conventional regression model of the form,

$$x_t = \alpha + \beta t + y_t, \quad (1-B)^d y_t = u_t, \quad t = 0, \pm 1, \pm 2, \ldots \tag{4}$$

with equation (1), where x_t is now the observed time series, α and β are the coefficients corresponding to the intercept and a linear time trend. Since u_t is $I(0)$, this allows the usage of a Whittle function in the frequency domain to compute the estimates of α, β and the fractional differencing parameter *d* as well as the confidence intervals of the estimates. The approach tests the null hypothesis,

$$H_o : d = d_0 \tag{5}$$

in equations (1) and (4) for a grid of real values d_0. Thus, the null model tested is,

$$x_t = \alpha + \beta t + y_t, \quad (1-B)^{d_o} y_t = u_t, \quad t = \pm 1, \pm 2, \ldots \tag{6}$$

with $I(0)$ disturbances.

By considering the effects of structural breaks on the time series under investigation, we considered the smooth change rather than the abrupt change

[6] A positive measurable function defined on some neighbourhood $[a, \infty)$ of infinity is said to be slowly varying in Karamata's sense if and only if for any $c > 0$, $\ell_1(cx)/\ell_1(x)$ converges to 1 as x tends to infinity (Palma, 2007).

implied by structural breaks.[7] The Chebyshev polynomials in time were first used in the context of unit root in Bierens (1997) since the function is bounded and orthogonal, in cosine function of time. In the context of fractional persistence, Cuestas and Gil-Alana (2016) made the proposition. The testing regression framework is of the form,

$$x_t = \sum_{i=0}^{m} \theta_i P_{iN}(t) + y_t; \quad t = \pm 1, \pm 2, \ldots \tag{7}$$

where *m* is the order of the Chebyshev polynomials and $P_{iN}(t)$ is the Chebyshev polynomial, given as,

$$P_{iN}(t) = \sqrt{2}\cos\left[i\pi\left(t-0.5\right)/N\right], \quad t = 1, 2, \ldots, N; \quad i = 1, 2, \ldots \tag{8}$$

with $P_{0N}(t) = 1$ (see Gil-Alana, Cunado and Gupta, 2015; Yaya, Gil-Alana and Carcel, 2015; Gil-Alana, Yaya and Solademi, 2016 and Caporale, Carcel and Gil-Alana, 2017 for some applications). Now, incorporating equation (8) in equation (7) with equation (1), we obtain simultaneously the fractional persistence estimate *d* along with nonlinear parameters $\theta_0, \theta_1, \theta_2, \ldots, \theta_m$. For $m = 0$, the entire model system contains only an intercept and a linear trend, while $m \geq 1$ indicates a nonlinear model.

By restricting ourselves to a case where $m = 3$, we have $\theta_0, \theta_1, \theta_2$ and θ_3. Nonlinearity is observed when at least one of θ_1, θ_2, and θ_3 is significant. The estimates of the regression here are tested using a Lagrange Multiplier (LM) test of the same form as in Robinson (1994).

Results and discussion

The data considered in this paper are the monthly consumer price index (CPI) of some selected African countries, obtained from the International Monetary Fund (IMF) website. The countries include: Burkina Faso (BKF), Cameroon (CAM), Cote D'Ivoire (COTE'D), Egypt (EGY), Ethiopia (ETH), Gambia (GMB), Ghana (GHN), Kenya (KNY), Madagascar (MADA), Mauritania (MAU), Morocco (MOR), Niger (NIGR), Nigeria (NGR), Senegal (SEN), South Africa (SA) and Swaziland (SWA). The series under consideration spans a 48-year period between January 1969 and December 2016, amounting to a sample size of 576.

Plots of CPI inflation are given in Figure 1. Clearly, we observe an increasing trend in all the 16 plots with price stability between 1970 and 1975. There is a general upward movement of CPI in almost all the countries, with some noticeable upward shift around 1995 for some countries (see Burkina Faso (BKF), Cameroun (CAM), Cote D'Ivoire (COTE'D), Kenya (KNY), Niger (NIGR), Nigeria (NGR) and Senegal (SEN)). Between 1993 and 1994, CAM, COTE'D, NIGER, SEN and SA experienced a sharp increase in CPI, while late 2010 marked the

[7] Ouliaris et al. (1989) proposed regular polynomials to approximate GDP data generating process.

index base year for all the countries, since this allows for easier comparison among the selected countries in Africa

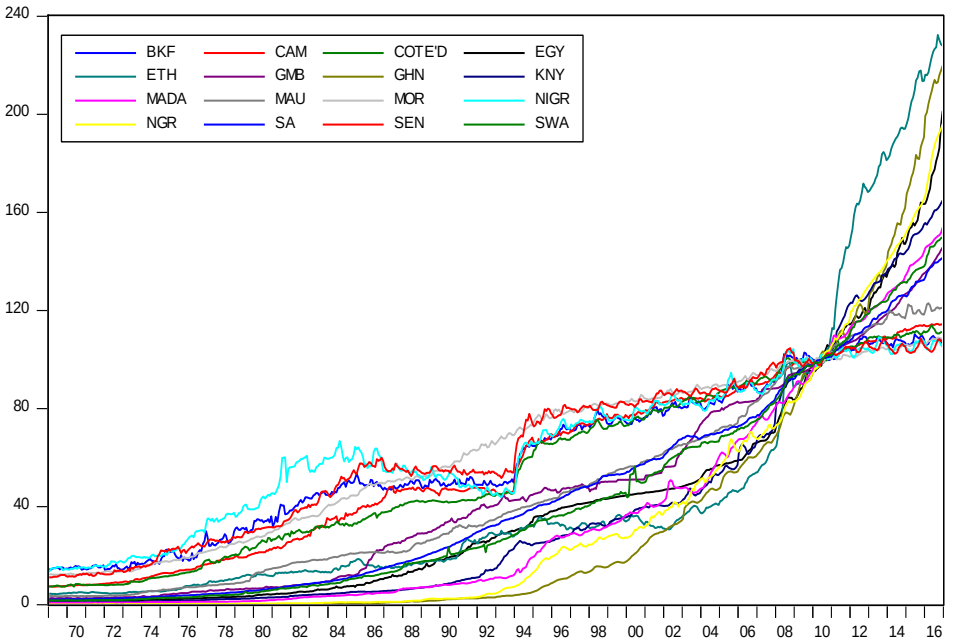


Figure 1. Plots of African CPI inflation

We consider the case of fractional persistence with linear trend, as proposed in Robinson (1994), using white noise, $AR(1)$ and seasonal $AR(1)$ disturbances. The results are given in Table 1. In the three cases, the estimates of persistence d are significant throughout, except in the cases of MADA, MAU and SA, where there were no convergence under $AR(1)$ disturbances. In the case of the white noise disturbance assumption, only BKF indicated tendency of mean reversion (i.e. $d<1$), while explosive behaviours of $d>1$ were observed in the cases of CAM, COTE'D, EGY, ETH, GMB, GHN, KNY, MADA, MAU, NIGR, NGR, SEN and SA. Evidence of unit root was observed for MOR and SWA. By considering $AR(1)$ disturbances, mean reversion is found in BKF, COTE'D, MOR and SEN, and evidence of $d>1$ is found in CAM, EGY, ETH, GMB, GHN, KNY, NIGR, NGR and SWA. With seasonal $AR(1)$ disturbance, both BKF and MOR indicated mean reversion, while NIGR and SWA exhibited unit roots. The remaining countries experienced explosive behaviour of $d>1$. Judging by the fractional persistence with linear trend result alone, one would conclude that inflation rates in the selected African countries might continue to drift farther from their mean level, without any possibility of naturally reverting to their mean level.

Table 1. Fractional persistence with Linear Trend

COUNTRIES	WND	AR(1)	Seasonal AR(1)
BKF	0.9214***a [0.0351]	0.8893***a [0.0651]	0.8993***a [0.0354]
CAM	1.1146***b [0.0339]	1.1199***b [0.0585]	1.1098***b [0.0341]
COTE'D	1.1135***b [0.0373]	0.9720***a [0.0703]	1.1031***b [0.0374]
EGY	1.3911***b [0.0341]	1.3668***b [0.0508]	1.3753***b [0.0354]
ETH	1.3926***b [0.0375]	1.2481***b [0.0574]	1.3783***b [0.0377]
GMB	1.2570***b [0.0291]	1.3362***b [0.0457]	1.2361***b [0.0303]
GHN	1.3372***b [0.0307]	1.3140***b [0.0419]	1.1935***b [0.0322]
KNY	1.3180***b [0.0348]	1.2304***b [0.0351]	1.3050***b [0.0361]
MADA	1.5990***b [0.0509]	-NA-	1.5647***b [0.0507]
MAU	1.1757***b [0.0360]	-NA-	1.1581***b [0.0386]
MOR	1.0310*** [0.0408]	0.8454***a [0.0447]	0.9969***a [0.0440]
NIGR	1.1153***b [0.0413]	1.0911***b [0.0411]	1.0668*** [0.0409]
NGR	1.3659***b [0.0310]	1.3609***b [0.0454]	1.3199***b [0.0317]
SEN	1.1572***b [0.0414]	0.8200***a [0.0866]	1.1221***b [0.0404]
SA	1.3280***b [0.0317]	-NA-	1.3392***b [0.0365]
SWA	1.0711*** [0.0250]	1.1538***b [0.0357]	1.0245*** [0.0092]

Note: Each cell contains the estimated value for d with corresponding standard errors given in squared brackets. The 'a' indicates evidence of I(d) with d < 1, while 'b' indicates evidence of I(d) with d > 1. WND means White Noise Disturbance.

*** denotes statistical significance at 1% level. NA means no convergence in the estimation.

By considering nonlinear deterministic trend in the fractional persistence framework, thereby testing fractional persistence simultaneously with nonlinearity in CPI inflation, we found only BKF to experience mean reversion (see Tables 2 and 3 for $m = 3$ and $m = 2$, respectively), while unit roots were found in MOR and SWA, and the behaviours of the remaining 13 countries were found to be explosive. These results are summarized in Table 4. In terms of nonlinearity for $m = 3$, we found significant evidence of nonlinearities in BKF, CAM, COTE'D, EGY, GMBIA, MAU, MOR, SEN, SA and SWA, while evidence of linearity was found in ETH, GHN, KNY, MADA, NIGR and NGR. Using $m = 2$, we observed

reduction in nonlinearity detection, as we observed BKF, CAM, COTE'D, MAU, MOR, SEN and SWA to be nonlinear. These results are summarized in Table 5.

Table 2. Nonlinear Fractional persistence based on Chebyshev Inequality for $m = 3$

COUNTRY	$\hat{d}$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
BKF	0.9082***a [0.0358]	-0.0118 [11.43]	-29.8581*** [5.104]	0.8418 [2.826]	-3.6135* [1.962]
CAM	1.1124***b [0.0340]	0.0830 [15.36]	-33.8404*** [8.870]	0.6447 [3.900]	-1.9529 [2.484]
COTE'D	1.1017***b [0.0382]	-1.3753 [15.31]	-34.0102*** [8.643]	2.2116 [3.838]	-1.8622 [2.456]
EGY	1.4128***b [0.0318]	-123.333** [54.67]	61.1138* [36.93]	-7.8413 [18.24]	5.8565 [9.877]
ETH	1.3879***b [0.0382]	3.8651 [105.5]	-30.2781 [68.86]	14.5224 [24.18]	-16.8124 [13.70]
GMB	1.2735***b [0.0293]	-12.0803 [23.29]	-25.1045* [14.91]	5.6864 [5.473]	-2.7410 [3.268]
GHN	1.3447***b [0.0309]	-61.4146 [64.46]	12.9233 [41.87]	13.3147 [14.67]	-6.3658 [8.490]
KNY	1.3105***b [0.0364]	-17.5846 [41.48]	-21.4425 [26.76]	14.8487 [9.644]	-6.8724 [9.644]
MADA	1.6010***b [0.0507]	-81.8681 [130.6]	28.5989 [87.63]	3.6018 [24.45]	0.6758 [12.49]
MAU	1.1128***b [0.0415]	1.8902 [11.13]	-35.7005*** [6.431]	8.9859 [2.821]	-4.7569*** [1.796]
MOR	0.9809*** [0.0455]	-4.6256 [25.09]	-31.9419*** [3.445]	-2.9066* [1.758]	-1.1612 [1.182]
NIGR	1.1098***b [0.0417]	-12.5872 [33.59]	-26.5733 [19.23]	0.3892 [8.478]	-5.5325 [5.405]
NGR	1.3767***b [0.0311]	-63.9076 [62.16]	12.6564 [12.59]	10.4945 [12.59]	-2.7215 [7.235]
SEN	1.1546***b [0.0418]	-0.2727 [28.50]	-31.0652* [17.29]	-2.5864 [7.237]	-2.4577 [4.532]
SA	1.3398***b [0.0319]	-16.3606 [21.52]	-23.5638* [13.97]	7.4402 [4.892]	-1.7994 [2.834]
SWA	1.0396*** [0.0281]	3.1819 [18.04]	-38.0410*** [6.456]	13.7196*** [3.077]	-5.9384*** [3.077]

Note: Each cell contains the estimated coefficient with corresponding standard errors given in squared brackets. 'a' indicates evidence of I(d) with d < 1, while 'b' indicates evidence of I(d) with d > 1.

***, ** and * denote statistical significance at 1%, 5% and 10% levels, respectively.

Table 3. Nonlinear Fractional persistence based on Chebyshev Inequality for $m = 2$

COUNTRY	$\hat{d}$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$
BKF	0.9243***a [0.0344]	-3.0395 [13.37]	-30.1741*** [5.620]	0.8907 [3.058]
CAM	1.1157***b [0.0337]	-3.9965 [14.88]	-33.0979*** [9.046]	0.6723 [3.965]
COTE'D	1.1053***b [0.0378]	-4.8150 [14.81]	-33.5735*** [8.817]	2.2078 [3.907]
EGY	1.4102***b [0.0320]	-106.358** [48.53]	54.3631 [35.92]	-7.2192 [17.33]
ETH	1.3975***b [0.0377]	-4.5106 [118.1]	-37.0295 [79.04]	10.3473 [25.60]
GMB	1.2786***b [0.0286]	-18.8667 [22.63]	-22.8764 [15.20]	5.4880 [5.604]
GHN	1.3496***b [0.0301]	-78.8250 [56.70]	19.4258 [38.79]	12.7245 [14.73]
KNY	1.3216***b [0.0349]	-35.5398 [43.63]	-15.1124 [29.44]	14.2944 [10.18]
MADA	1.6008***b [0.0506]	-79.8863 [120.1]	27.8198 [83.53]	3.6559 [23.97]
MAU	1.1528***b [0.0377]	-7.5287 [12.80]	-33.9124*** [8.134]	8.8778** [3.421]
MOR	0.9897*** [0.0444]	-8.5827 [52.72]	-31.9241*** [3.632]	-2.9077 [1.834]
NIGR	1.1170***b [0.0408]	-22.4837 [32.83]	-25.1671 [19.98]	0.3935 [8.769]
NGR	1.3788***b [0.0304]	-71.5437 [54.34]	15.5853 [37.25]	10.2329 [12.66]
SEN	1.1569***b [0.0415]	-6.9309 [27.40]	-29.1434* [17.45]	-2.4043 [7.320]
SA	1.3436***b [0.0313]	-21.1572 [21.12]	-21.8362 [14.35]	7.2951 [4.970]
SWA	1.0683*** [0.0265]	-8.7848 [14.91]	-36.4636*** [7.714]	13.4972*** [3.545]

Note: Each cell contains the estimated coefficient with corresponding standard errors given in squared brackets. 'a' indicates evidence of I(d) with d < 1, while 'b' indicates evidence of I(d) with d > 1.

***, ** and * denote statistical significance at 1%, 5% and 10% levels, respectively.

Table 4. Summary of the Results in terms of value of *d*

	Mean reversion $(d<1)$	**Unit roots** $(d=1)$	**Explosive behaviour** $(d>1)$
White noise disturbances	BKF	MOR, SWA	CAM, COTE'D, EGY, ETH, GMB, GHN, KNY, MADA, MAU, NIGER, NGR, SEN, SA
AR (1) Disturbances	BKF, COTE'D, MOR, SEN	MADA, MAU, SA	CAM, EGY, ETH, GMB, GHN, KNY, NIGR, NGR, SWA
Seasonal AR (1) Disturbances	BKF, MOR	NIGR, SWA	CAM, COTE'D, EGY, ETH, GMB, GHN, KNY, MADA, MAU, NGR, SEN, SA
Nonlinear trend with $m=3$	BKF	MOR, SWA	CAM, COTE'D, EGY, ETH, GMB, GHN, KNY, MADA, MAU, NIGR, NGR, SEN, SA
Nonlinear trend with $m=2$	BKF	MOR, SWA	CAM, COTE'D, EGY, ETH, GMB, GHN, KNY, MADA, MAU, NIGR, NGR, SEN, SA

Table 5. Summary of the Results in terms of Nonlinearities

	Evidence of Nonlinearities	No Evidence of Nonlinearities
$m=3$	BKF, CAM, COTE'D, EGY, GMB, MAU, MOR, SEN, SA, SWA	ETH, GHN, KNY, MADA, NIGR, NGR
$m=2$	BKF, CAM, COTE'D, MAU, MOR, SEN, SWA	EGY, ETH, GMB, GHN, KNY, MADA, NIGR, NGR, SA

4. Concluding Remarks and Policy

In this paper, we have examined time behaviour of African inflation using CPI as an inflation proxy variable. We considered 16 African countries, with data in monthly frequency, spanning between January 1969 and December 2016. We considered memory property, fractional persistence and nonlinearity using a newly developed approach of Cuestas and Gil-Alana (2016). The main results indicated that African inflationary dynamics are mostly explosive and nonlinear, and strong policy intervention is required to bring inflation back to original trend levels in each of these countries. Thus, mean reversion is likely to occur in CPI inflation of Burkina Faso. In the choice of methodology for analyzing inflation in Africa, this work recommends careful selection of the estimation approach, particularly, in countries where nonlinearities are detected.

REFERENCES

ADENEKAN, A. T., NWANNA, G. A., (2004). Inflation Dynamics in a Developing Economy: An Error Correction Approach. African Review of Money Finance and Banking, pp. 77–99.

ADU, G. MARBUAH, (2011). Determinants of Inflation in Ghana: An Empirical, Investigation. South African Journal of Economics, 79 (3), pp. 251–269.

AJIDE, K. B, LAWANSON, O., (2012). Inflation Threshold and Economic Growth: Evidence from Nigeria. Asian Economic and Financial Review, 2 (7), pp. 876–901.

ALI, H., (2011). Inflation Dynamics: The Case of Egypt. MPRA Paper No. 36331. Available at:http://mpra.ub.uni-muenchen.de/36331/.

BAWA, S., ABDULLAHİ, I. S., IBRAHIM, A., (2016). Analysis of Inflation Dynamics in Nigeria (1981 – 2015). CBN Journal of Applied Statistics, 7, pp. 255–276.

BEN NASR, A., AJMİ, A. N., GUPTA, R., (2014). Modelling the volatility of the Dow Jones Islamic market world index using a fractionally integrated time varying GARCH (FITVGARCH) model. Applied Financial Economics, 24, pp. 993–1004.

BIERENS, H. J., (1997). Testing the unit root with drift hypothesis against nonlinear trend stationarity with an application to the US price level and interest rate. Journal of Econometrics, 81, pp. 29–64.

BOATENG, A., LESOANA, M., SIWEYA, H., BELETE, A., GIL-ALANA, L. A., (2017). Modelling persistence in the conditional mean of inflation using the ARFIMA process with GARCH and GJRGARCH innovations: the case of Ghana and South Africa. African Review of Economics and Finance, 9 (2), pp. 96–130.

BURGER, P., MARINKOV, M., (2005). The South African Phillips Curve: A triangular puzzle? Paper presented at the "Development Perspectives – Is Africa Different?" Biennial conference of the Economic Society of South Africa (ESSA), Durban, South Africa.

BURGER, R., DU PLESSIS, S., (2006). A New Keynesian Phillips Curve for South Africa. SARB Conference 2006.

CAPORALE, G. M., CARCEL, H., GIL-ALANA, L. A., (2015). Modelling African inflation rates: nonlinear deterministic terms and long-range dependence. Applied Economics Letters, 22 (5), pp. 421–424.

CAPORALE, G. M., CARCEL, H., GIL-ALANA, L. A., (2017). Central bank policy rates: Are they cointegrated? International Economics, 152, pp. 116–123.

CECCHETTI, S., DEBELLE, G. (2006). Has the inflation process changed? Economic Policy, 21 (46), pp. 311–352.

CUESTAS, J. C., GIL-ALANA, L. A., (2016). A non-linear approach with long range dependence based on Chebyshev polynomials. Studies in Nonlinear Dynamics and Econometrics, 23, pp. 445–468.

DIEBOLD, F.X. AND INOUE, A., (2001). Long memory and regime switching. Journal of Econometrics, 105, pp. 131–159.

FABAYO, J. A., AJILORE, O. T., (2006). Inflation: How much is too much for Economic growth in Nigeria. Indian Economic Review, 41 (2), pp. 129–147.

FEDDERKE, J. W., SCHALING, E. (2005). Modelling Inflation in South Africa: A Multivariate Cointegration Analysis. South African Journal of Economics, 73, pp. 79–92.

FIELDING, D., K. LEE, SHIELD, K., (2004). The Characteristics of Macroeconomic Shocks in the CFA Franc Zone. Research Paper No. 2004/21, March, World Institute for Development.

GIL-ALANA, L. A., (2008). Fractional integration and structural breaks at unknown periods of time. Journal of Time Series Analysis, 29, pp. 163-185.

GIL-ALANA, L. A., CUNADO, J., GUPTA, R., (2015). Persistence, mean reversion and nonlinearities in infant mortality rates. Department of Economics, University of Pretoria Working paper series, No. 2015–74.

GIL-ALANA, L. A., SHITTU, O. I., YAYA, O. S., (2012). Long memory, Structural breaks and Mean shifts in the Inflation rates in Nigeria. African Journal of Business Management, 6 (3), pp. 888–897.

GIL-ALANA, L. A., YAYA, O. S., SOLADEMI, E. A., (2016). Testing unit roots, structural breaks and linearity in the inflation rates of the G7 countries with fractional dependence techniques. Applied Stochastic Models in Business and Industry, 32, pp. 711–724.

GRANGER, C. W. J., HYUNG, N., (2004). Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. Journal of Empirical Finance, 11, pp. 399–421.

HODGE, D., (2002). Inflation versus unemployment in South Africa: is there a trade-off? South African Journal of Economics, 70 (3), pp. 417–443.

HODGE, D., (2006). Inflation and growth in South Africa. Cambridge Journal of Economics (30), pp. 163–180.

HODGE, D., (2009). Growth, Employment and Unemployment in South Africa. South African Journal of Economics, 70 (4), pp. 488–504.

HOSNY, A. S., (2014). What is the Central Bank of Egypt's Implicit Inflation Target? International Journal of Applied Economics, 13 (1), pp. 43-56.

IMIMOLE, B., ENOMA, A., (2011). Exchange Rate Depreciation and Inflation in Nigeria (1986 – 2008). Business and Economics Journal, BEJ28, pp. 1–12.

KAPETANIOS, G., SHIN, Y., SNELL, A., (2003). Testing for a unit root in the nonlinear STAR framework. Journal of Econometrics, 112, pp. 359–379.

KAUSHIK, B., (2011). Understanding inflation and controlling it. Paper. Available at: http://finmin.nic.in/WorkingPaper/understanding_inflation_controlling.pdf

KİMANİ, D. K., MUTUKU, C. M., (2013). Inflation Dynamics on the Overall Stock Market Performance: The Case of Nairobi Securities Exchange in Kenya. Economics and Finance Review, 2 (11), pp. 1–11.

KİRİMİ, W. N., (2014). The Determinants of Inflation in Kenya (1970 – 2013). A research project presented to the School of Economics of the University of Nairobi, Kenya.

KWİATKOWSKİ, D., PHİLLİPS, P. C. B., SCHMİDT, P, SHİN, Y., (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root. Journal of Econometrics, 54, pp. 159–178.

MANKIW, N. G., (2001). The Inexorable and Mysterious tradeoff between Inflation and unemployment. The Economic Journal Conference Papers, 111 (471), C45–C61.

MIKKELSEN, J., PEIRIS, J. S., (2005). Uganda: Selected Issues and Statistical Appendix, IMF Country Report 05/172.

NELL, K. S., (2000). Is Low Inflation a Precondition for Faster Growth, The Case of South Africa, Department of Economics, University of Kent, United Kingdom.

NELL, K. S., (2006). Structural change and nonlinearities in a Phillips Curve model for South Africa. Contemporary Economic Policy, 24 (4), pp. 600–617.

ODUSANYA, I. A., ATANDA, A. A., (2010). Analysis of Inflation and its Determinants in Nigeria, MPRA Paper No. 35837.

OHANISSIAN, A., RUSSELL, J. R., TSAY, R. S., (2008). True or spurious long memory? A new test, Journal of Business Economics and Statistics, 26, pp. 161–175.

OSAMA, EL BAZ, (2014). The Determinants of Inflation in Egypt: An Empirical Study (1991-2012). The Egyptian Center for Economic Studies, MPRA Paper No. 56978. http://mpra.ub.uni-muenchen.de/56978/.

OULIARIS, S., PARK, J. Y., PHILLIPS, P. C. B., (1989). Testing for a unit root in the presence of a maintained trend. In Ray, B. (Ed.). Advances in Econometrics and Modelling. Kluwer, Dordrecht, pp. 6–28.

PAUL, H. W., NORMAN, E. D., ERNEST, I. H., (1973). Financial Accounting, New York: Harcourt Brace Javonovich, Inc. Page 429.

PALMA, W., (2007). Long memory time series: Theory and methods. John Wiley & Sons, New Jersey.

PHIRI, A., (2013). Inflation and Economic Growth in Zambia: A Threshold Autoregressive (TAR) Econometric Approach. Department of Economics and Management Sciences, School of Economics, North West University, Potchefstroom, South Africa, MPRA Paper No. 52093, http://mpra.ub.uni-muenchen.de/52093/.

ROBINSON, P. M., (1994). Efficient tests of nonstationary hypotheses. Journal of the American Statistical Associations, 89, pp. 1420–1437.

SALISU, A. A., OGBONNA, A. E., (2017). Improving the predictive ability of oil for inflation: An ADL-MIDAS Approach - Centre for Econometric and Allied Research, University of Ibadan Working Papers Series, CWPS 0025.

SVENSSON, L. E. O., (2003). Escaping from a Liquidity Trap and Deflation: The Foolproof Way and Others. Journal of Economic Perspectives, 17, pp. 145–166.

TABI, H. N., ONDOA, H. A., (2011). Inflation, Money and Economic Growth in Cameroon. International Journal of Financial Research, 2 (1), pp. 45–56.

VERMEULEN, C., (2015). Inflation, growth and employment in South Africa: Trends and trade-offs, ERSA working paper 547.

WOLDE-RUFAEL, Y., (2008). Budget Deficits, Money and Inflation: The Case of Ethiopia, The Journal of Developing Areas, 42 (1), pp. 183–199.

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. 133–153, DOI 10.21307/stattrans-2019-028
Submitted – 28.08.2018; Paper ready for publication – 16.07.2019

SPATIAL MICROSIMULATION OF PERSONAL INCOME IN POLAND AT THE LEVEL OF SUBREGIONS

Wojciech Roszka[1]

ABSTRACT

The paper presents an application of spatial microsimulation methods for generating a synthetic population to estimate personal income in Poland in 2011 using census tables and EU-SILC 2011 microdata set. The first section presents a research problem and a brief overview of modern estimation methods in application to small domains with particular emphasis on spatial microsimulation. The second section contains an overview of selected synthetic population generation methods. In the last section personal income estimation on NUTS 3 level is presented with special emphasis on the quality of estimates.

Key words: data integration, spatial microsimulation, small area estimation, synthetic data generation.

1. Introduction

Providing reliable, current and multidimensional information for local administrative units is one of the main tasks of official statistics. In particular, it is important to support the state in the struggle against various undesirable social phenomena, such as monetary and non-monetary poverty. Information about its size and spatial differentiation is very desirable. Providing detailed spatial information on life quality indicators may contribute to a better redistribution of income, as well as to indicate places where different types of investments are needed.

To fulfill their obligations, statistical bodies carry out many sample surveys on different socio-economic phenomena. One of the studies in which the indicators of quality of life are measured is the European Union Statistics on Income and Living Conditions (EU-SILC). The sample size in the EU-SILC study, however, allows the aggregation of results at most at the level

[1] Poznań University of Economics and Business. E-mail: wojciech.roszka@ue.poznan.pl. ORCID ID: http://orcid.org/0000-0003-4383-3259.

of NUTS1, because direct[2] estimates at lower levels of spatial aggregation are characterized by an unacceptably large random error.

To increase the usability, in the context of obtaining estimates for small domains, information from sample surveys, small area estimation methods (indirect estimation, SAE) and administrative sources are often used. SAE combines direct estimation with the so-called strength borrowing. Using additional information from a different data source, small domain estimates may characterize in smaller error. The results cannot be aggregated and disaggregated freely though. They are just fixed numbers resulting from a particular model. The estimators used in SAE usually improve the efficiency of estimates for small domains (Rao 2003) and in Poland experimental work has been done on the use of indirect estimation in poverty mapping, i.e. its spatial differentiation (Wawrowski 2014, Szymkowiak *et al.* 2013). Administrative sources contain information on a large amount of individuals for basic socio-economics characteristics. Serving, however, other than statistical purposes, a problem with coverage may appear (Penneck 2007; Walgren, Walgren 2007). Also, their substantive content is less abundant compared to sample surveys. And last but not least, there is a huge problem with data confidentiality, which results in reluctance to disseminate them (*Statistics New Zealand* 2006).

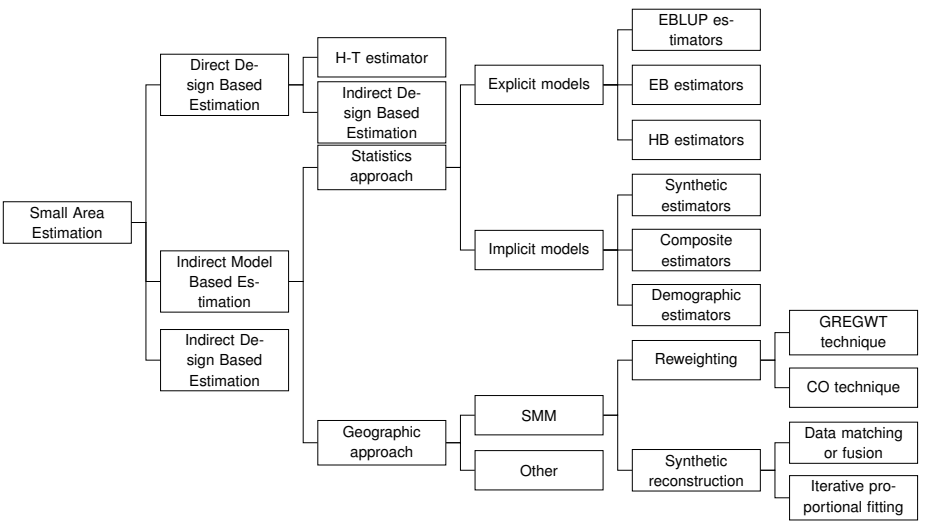
Combining advantages and reducing defects of methods discussed above, spatial microsimulation modelling (SMM) is gaining more and more popularity. The aim of the SMM methods is to create a dataset containing information on all units from a resulting population and a vector of many socio-economic characteristics (Ballas *et al.* 2005, Tanton, Edwards 2013; Rahman, Harding 2017; Rahman 2009; Tanton 2014; O'Donoghue 2014). The creation involves integration of sample survey microdata and small domain census constraints. Using different reconstruction and reweighing algorithms, synthetic units are being created in such a way that the true distribution of a real population small geographical units is reflected. Having a multidimensional, full-coverage dataset not only small area estimation can be performed but flexible aggregation and disaggregation is possible. In the context of poverty, Eurostat has already undertaken the first works on the use of EU-SILC for the construction of this type of pseudo-populations (Alfons *et al.*, 2011).

Microsimulation models are becoming more and more popular in the SAE literature (Rahman, Harding 2017; Tanton, Edwards *et al.* 2013; Templ,

[2]e.g. Horvitz-Thompson (H-T) estimators.

Filzmoser 2014; Tanton *et al.* 2011; Whitworth *edt* 2013; Rahman *et al.* 2010; Rahman 2009). Methods involving creation of pseudo-populations (or *synthetic* populations) are ascribed to "geographic approach" towards small area estimation (Rahman 2008; see Figure 1). The main idea of spatial microsimulation is a creation of anonymised full-coverage synthetic dataset with adequate variables and with marginal and joint distribution, which are at least *quasi*-identical to reality (Templ *et al.* 2017).

Figure 1. Small area estimation methods in spatial microsimulation (after Rahman 2008)



The paper presents an application of methods for generating a synthetic population. The aim of this study is to estimate personal income in Poland in 2011 using census tables and EU-SILC 2011 microdata set. In the first section a research problem and a brief overview of modern estimation methods in application to small domains with particular emphasis on spatial microsimulation is presented . The second section contains an overview of selected synthetic population generation methods. In the last section personal income estimation on NUTS 3 level is presented with special emphasis on the quality of estimates.

The resulting pseudo-population should satisfy the following conditions (Münnich, Schürle 2003):

- the true distribution in terms of small geographical units should be reflected in the synthetic population,
- marginal and joint distribution between variables – the interdependence of true population – should be preserved,

- heterogeneity in subpopulations should be reflected, especially in spatial terms,
- simple units' replication based on integer sample weights leads to a reduction of variability. Hence, it should not be performed,
- data confidentiality must be ensured.

The complex dataset is synthesized by integration of two data types:

1. **Survey sample microdata file** - which contains comprehensive information about many socio-economic phenomena of persons and/or households.
2. **Census benchmarks (tables)** - which deliver (implicitly) true frequencies in small areas (domains).

The starting point of microsimulation is a construction of a microdata file (Rahman 2009). Even if the data file is provided by a particular statistical body, it is most likely burdened by non-random errors. The number of refusals to respond increases every year. Also, item non-response problems are often handled by imputation methods, which result in a model value rather than a real one. To overcome these problems, new weights are calculated based on census constraints and given sample weights. In another step, the Monte-Carlo sampling is performed to create new close-to-reality complex dataset.

Spatial microsimulation has a certain advantage over "traditional" statistical models (where estimates are calculated only for a particular area). First of all, having a complex microdata set allows a dynamic aggregation and dissaggregation of the data. The multidimensionality of resulting file gives the opportunity of flexible estimation in terms of choice of a spatial scale. Data integration approach in microsimulation uses the synergy effect, which links the comprehensiveness of sample survey and the full-coverage of census. And last but not least, with set of attributes stored as lists for each individual it is possible to perform different simulations.

2. SMM methods overview

Spatial microsimulation methods can be divided into two subgroups (Rahman 2010): (1) synthetic reconstruction and (2) reweighting.

2.1. Synthetic reconstruction

Synthetic reconstruction is a method where synthetic populations are reconstructed in such a way that all small area census constraints are met. Two techniques are introduced (Rahman 2008): data matching and iterative proportional fitting.

Data matching is a mass imputation technique where on the basis o p-dimensional vector of common variables units form a sample survey microdata file are matched with units in census microdata[3] (vide Figure 1). When personal identifiers are available in both files n sample units are deterministically matched to its census counterparts (such an approach is called *exact matching*). The rest of census units are matched with sample units using non-parametric, parametric or mixed framework of probabilistic data matching (for a detailed description of statistical matching methods see D'Orazio *et al.* 2006 and Rässler 2002).

The iterative proportional fitting algorithm is an iterative procedure that matches the *n*-dimensional table of sample frequencies to known population benchmarks. Sample weights are calibrated to known sums from the entire population.

A detailed description of IPF method can be found in (Norman 1999).

On the basis of original sample weights and expected frequencies the inclusion probability is computed and units are randomly selected until the expected numbers in census domains are reached. As in the case of data matching, all q-dimensional vectors of attributes are automatically selected (Templ *et al.* 2017).

2.2. Reweighting

The are two reweighting techniques in SMM - GREGWT (**G**eneralized **Reg**ression and **W**eigh**t**ing) and combinatorial optimization (CO). Both are widely used in spatial microsimulation models in small area estimation.

The GREGWT technique is one of the calibration methods. It is an iterative process using the Newton-Raphson method of iteration. The algorithm uses a constrained distance function known as the truncated chi-squared distance function that is minimized subject to the calibration equations for each small area (Rahman 2013). Generally speaking, the method produces new weights according to known small domains counts in such a way that the new weights are characterized by a minimum distance from the origi-

[3] Census microdata is usually obtained by disaggregation of published census tables.

nal weights. The algorithm is described in detail in Tanton *et al.* (2011), Rahman, Harding (2017) and Munoz *et al.* (2015).

The CO re-weighting method is motivated towards selecting *an appropriate combination of units* from survey data to attain the known constraints at **small area levels** using an optimization tool (Voas, Williamson 2000; Rahman *et al.* 2010; Williamson 2013; Rahman, Harding 2017). The CO reweighing involves the following steps:

1. Collection of sample survey microdata and small area benchmark constraints.

2. Selection of a set of units randomly from the survey sample, which will act an initial combination of units from a small area.

3. Tabulation of selected units and calculation of total absolute differences (TAD) from the known small area constraints:

$$TAD = \sum |x_i - x_i^*|, \tag{1}$$

   where x_i is a true value of x in i-th contingency cell and x_i^* is a value resulting from the created combination.

4. Choosing one of the selected units randomly and replacing it with a new unit drawn at random from the survey sample, and then follow step 3 for the new set of combination of units.

5. Repetition of step 4 until no further reduction in TAD is possible.

It is worth noting that with finite populations it is theoretically possible to calculate all the combinations and find the one with the minimal possible TAD. However, in practice, to fit a small area of 10 units out of 1000 in a population one would have to calculate 2.63×10^{23} combinations. This is an approximate number of grains of sand on Earth[4] and the number of stars in the observable universe according to European Space Agency[5]. In order to overcome that obvious computational problem, the simulated annealing (SA) probabilistic technique for approximating the global optimum of a given function has been adapted to combinatorial optimization (Pham, Karaboga 2000). SA is a type of a heuristic algorithm that searches the space of alternative problem solutions to find the best solutions. The mode of operation

[4]Wolfram Alpha provides that this number varies from 10^{20} to 10^{24}.

[5]`https://www.esa.int/Our_Activities/Space_Science/Herschel/How_many_stars_are_there_in_the_Universe` (access from 15.08.2018)

of the simulated annealing is similar to the annealing in the metallurgy (for details see Rahman, Harding 2017).

2.3. Quality assessment

The quality of the obtained synthetic population is assessed mainly by comparison to the real, known values. No standardized variance estimation method has been developed yet (Rahmnan, Harding 2017). In most cases the quality assessment is carried out in two stages (Rahman, Harding 2017; Templ *et al.* 2017; Templ, Filzmoser 2014; Alfons *et al.* 2011). Firstly, the internal validation is performed. Marginal and joint distributions of census variables are compared to those in the synthetic dataset. Also, the distribution of the target variable in the synthetic dataset is compared to the distribution in the sample.

If internal validation is passed, the synthetic population estimates are compared to real values known from other sources. To perform inference about the lack of differences between the synthetic population estimates and real values the use of standard significance tests was proposed (Williamson 2013; Templ *et al.*. 2017). Such an approach, although methodologically correct, has some disadvantages. First of all, the use of population size in test statistics may lead to rejecting null hypothesis even with very low differences due to the "artificial" increase of test statistics' value. Subsequently, having real values of the estimated variables puts into question the meaning of conducting the microsimulation – the goal is to estimate unknown values. And third, using parametric tests the assumptions about the normality of distributions are omitted (not to say ignored). Still, work on estimating standard errors and the properties of SMM estimators is ongoing (Goedemé 2013; Whitworth *et al.* 2016).

3. Empirical study

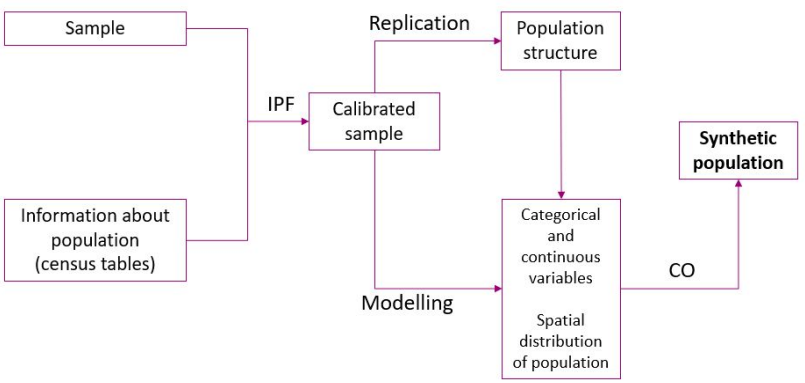
The main aim of the empirical study was to estimate personal net income in terms of 72 NUTS 3 geographical units on the basis of EU-SILC study in Poland. Such estimates are unavailable due to insufficient size of the sample in these areas. The secondary goal is to verify the suitability of the discussed methods in the estimates for small domains for socio-economic issues in Poland. Due to the conduct of census, the year 2011 was selected as the year of the study.

In 2011 in EU-SILC 12871 households were surveyed, in which 36720

inhabitants lived. There were 30421 people for whom income was measured[6] (also economic status and education level). Such a sample size allowed publishing the results at the level of NUTS 1 only. Publications including estimates at lower levels had an experimental character and are not considered official estimates of official statistics (Szymkowiak *et al.* 2017).

The EU-SILC microdata included 19 variables selected for the study (see Table 1). Variables SYMTER and KLM were added by the Polish NSI to facilitate spatial analysis. Variables PY010N – PY140N contained information about the size of different sources of *net* income (in € per year[7]). For the purpose of the study, after summing up all sources of income and creating a *nIncome* variable, the variables were dichotomized in such a way that they took a value of 1 for non-zero values and 0 otherwise. Census tables contained joint distributions estimated by National Census of Population and Housing 2011 of NUTS 3 $\times$ gender $\times$ age.

Figure 2. Structure of the study



Source: Templ *et al.* (2017)

[6] At the age of 16 years and more.
[7] The previous year was the reference period.

Table 1. Variables in the study

Variable	Definition
SYMTER	Symbol of territorial unit
RB090	Gender
RX010	Age at the time of interview
PL031	Self-defined economic status
PE040	Highest ISCED level attained
KLM	Class of place of residence
PY010N	Net employee cash or near cash income
PY020N	Net Non-Cash employee income
PY021N	Company car
PY035N	Contributions to individual pension plans
PY050N	Net cash benefits/losses from self-employment
PY080N	Regular pension from private plans
PY090N	Unemployment benefits
PY100N	Old-age benefits
PY110N	Survivor benefits
PY120N	Sickness benefits
PY130N	Disability benefits
PY140N	Education-related allowances
nIncome	Total net personal income (sum of "PY" vars)

The plan of the study (see Figure 2) starts with the calibration of original EU-SILC sample weights given census constraints using IPF algorithm in the first step. In the second step, on the basis of the calibrated weights, the units are replicated through sampling. The probability of unit being selected is an inverse of the calibrated weight. The units are drawn until census constraints are met. Next, the target variable is modelled. In order to overcome a very likely situations where category appears in the population but not in the sample data, categories are estimated by conditional probabilities using multinomial logistic regression (Alfons *et al.* 2011). One categorical variable is simulated as follows:

1. Simulated variable is selected from sample S. Independent variables

must be present in both sample S and population U,

$$S = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,j} & x_{1,p+1} \\ x_{2,1} & x_{2,2} & \dots & x_{2,j} & x_{2,p+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,j} & x_{n,p+1} \end{bmatrix}$$

where $i = 1,\dots,n$ are sample units and $k = 1,\dots,j$ is the number of variables. X_1 to X_j is an independent variable vector and X_{p+1} is the target (dependent) variable.

2. The model is estimated in every small area using sample S units. As a result β coefficients are obtained.

3. For every $i = 1,\dots,N$ unit of the selected variable, new outcome category is predicted. The conditional probability of selecting r-th category for each i-th $\hat{x}^*_{i,j+1}$ is:
$\hat{p}_{i1} = \frac{1}{1+\sum_{r=2}^{R} exp(\hat{\beta}_{0r}+\hat{\beta}_{1r}\hat{x}_{i,1}+\dots+\hat{\beta}_{jr}\hat{x}_{i,1})}$,
$\hat{p}_{ir} = \frac{exp(\hat{\beta}_{0r}+\hat{\beta}_{1r}\hat{x}_{i,1}+\dots+\hat{\beta}_{jr}\hat{x}_{i,1})}{1+\sum_{r=2}^{R} exp(\hat{\beta}_{0r}+\hat{\beta}_{1r}\hat{x}_{i,1}+\dots+\hat{\beta}_{jr}\hat{x}_{i,1})}$,
where $r = 2,\dots,R$ and $\hat{\beta}_{0r},\dots,\hat{\beta}_{jr}$ are the estimates of multinomial logistic regression model. The new $\hat{x}^*_{i,j+1}$ values are computed.

4. The population U is:

$$U = \begin{bmatrix} \hat{x}_{1,1} & \hat{x}_{1,2} & \dots & \hat{x}_{1,j} & \hat{x}^*_{1,j+1} \\ \hat{x}_{2,1} & \hat{x}_{2,2} & \dots & \hat{x}_{2,j} & \hat{x}^*_{2,j+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{x}_{N,1} & \hat{x}_{N,2} & \dots & \hat{x}_{N,j} & \hat{x}^*_{N,j+1} \end{bmatrix}.$$

Such an approach minimizes the appearance of the so-called *random zeroes* (domains that exist in the population but did not occur in the sample).

For continuous variables one of the suggested approaches (Templ *et al.* 2017) involves the following:

1 Dependent x_{j+1} in discretized is y_{j+1} by creating R cut-off values $c_1 \leq \dots \leq c_R$:

$$y = \begin{cases} 1 & \text{if } c_1 \leq x_{ij} < c_2, \\ 2 & \text{if } c_2 \leq x_{ij} < c_3, \\ \vdots & \vdots \\ R & \text{if } c_{R-1} \leq x_{ij} \leq c_R. \end{cases}$$

2 Multinomial logistic regression model (the same as for categorical variables) is estimated with the dependent variable y_{j+1} and the independent variables vector $x_1, x_2, \ldots, x_j$ for each k-th domain (small area) separately.

3 Within each r-th class estimates $\hat{x}$ are drawn from a uniform distribution with boundaries of classes as parameters. The exception is the last class, where due to outliers values are drawn using generalized Pareto distribution:

$$\hat{x}^*_{i,j+1} \approx \begin{cases} U(c_r, c_{r+1}) & \text{if } \hat{y}_i = r \text{ and } 1 \leq r \leq R-1, \\ GPD(\mu, \sigma, \xi, x) & \text{if } \hat{y}_i = R. \end{cases}$$

With replicated units and modelled values of the target variable(s), the population is once again reweighed to known small domain constraints using CO algorithm. The relocation of units in domains is necessary when a perfect match is required.The replication of units using IPF weights does not meet the constraints exactly due to the random process of replication. After reweighting, the final synthetic population is ready for quality assessment and then for estimation.

As a result a synthetic dataset of 38,113,162 individuals in Poland was created. Every unit was described by a vector of variables listed in Table 1.

The internal validation was largely descriptive and performed in two stages. In the first stage, marginal and joint distributions of matching variables were compared. The comparison was conducted using mosaic plots representing differences in joint distribution of sample and synthetic estimates in the form of a three-dimensional contingency table, which presents the relative differences between them. Due to the lack of official statistics for net personal income, the estimates obtained were compared to the annual average gross salary[8] in terms of subregions for 2010[9].

The distributions of selected matching variables in the sample and synthetic populations are largely consistent (see Figure 3 and 4). Figure 3 shows differences in the joint distribution of the variables: sex, self-defined economic status and highest ISCED level attained[10]. Relatively small differences prevailed (colours derived from green, yellow and pink) – up to 2%. The biggest differences prevailed in the smallest domains - which was to be

[8] Treated as a proxy variable.

[9] The reference year for income in EU-SILC 2011

[10] jsol – junior secondary or lower; sapn – secondary and post-secondary non-tertiary; t – tertiary

expected, as the less frequent domains are characterized by a greater error of estimate. It should also be noted that the highest observed difference (unemployed women with secondary and post-secondary non-tertiary education level) did not exceed 4.6%, which can be considered a good result.

Figure 4 shows differences in the distribution in terms of sex, age and self-defined economic status[11]. The differences in this case were greater due to the much smaller domains determined by the analyzed variables. However, it should be noted that the vast majority of them were characterized by a difference up to 4.6%, which should be regarded as a good result with these relatively small domains. The biggest error (unemployed women aged 16-19) was 16%, but it was characterized by one of the smallest domains.

[11] w – working; u – unemployed; i – inactive

Figure 3. Differences in joint distribution of gender, self-defined economic status and highest ISCED level

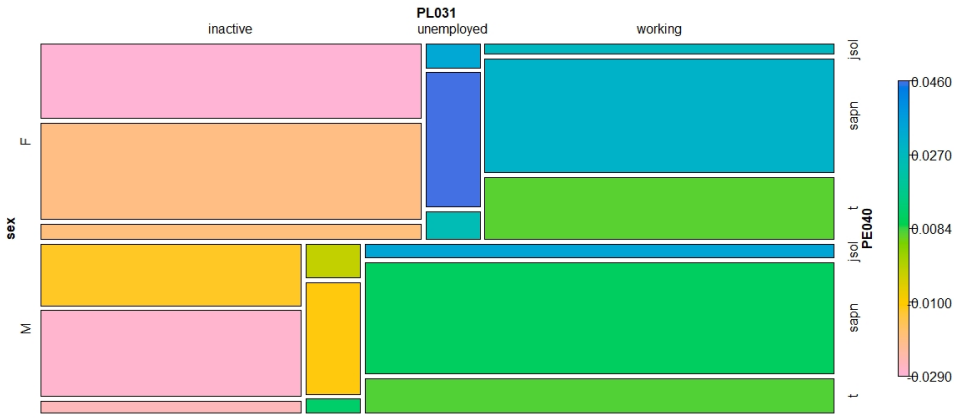
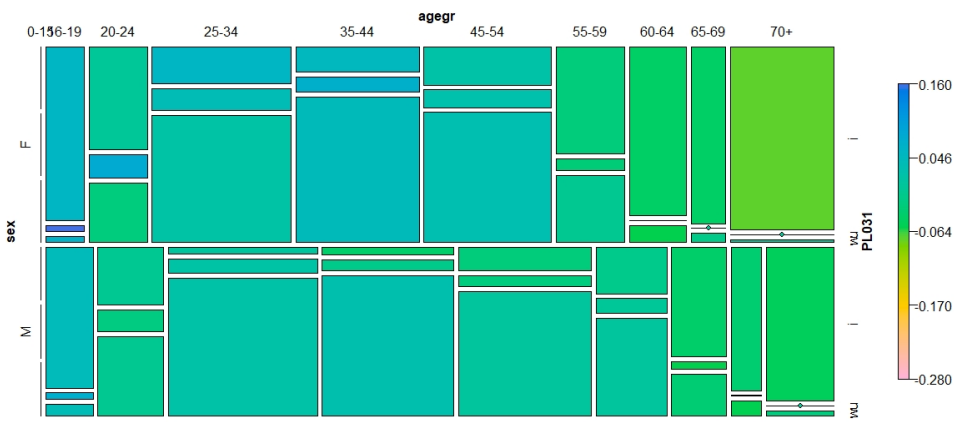
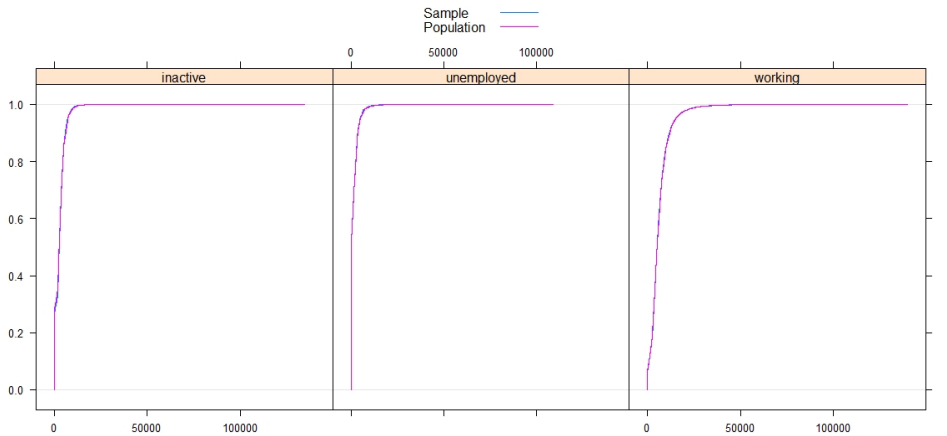


Figure 4. Differences in joint distribution of gender, age groups and self-defined economic status



The distribution of net personal income in terms of the self-defined economic status is largely consistent in the sample (blue line) and synthetic population (pink line; see Figure 5). A similar situation is observed in other domains.

Figure 5. Distribution of net personal income in terms of self-defined economic status



The spatial distribution of mean personal income is consistent with general knowledge (see Figure 7 and Figure 8). One can distinguish 5 areas of relatively high income. The first is Warsaw (with estimated average yearly personal net income equal to 7529.7 €; see Figure 9), which is the centre of services and financiers in Poland, and its surroundings (subregion Warsaw West - 4926.8 € and Warsaw East - 4600.9 €), which are often referred to as the bedroom of the capital and their inhabitants largely work in Warsaw. Secondly, one can mention Poznań (6216.5 €), Tri-City (5841.5 €), Cracow (5756.7 €), Wroclaw (5685.5 €) and Legnica with Głogów in one subregion (5484.8 €; this subregion "crept" between large urban centres due to the location of a huge mining conglomerate). Upper Silesia, where many mines and industrial plants are located, is one of the richest areas in Poland. Although none of its subregions found themselves in the top six, 7 of them are in the top twenty. In the top 25 there were almost all large urban centres and their neighbourhoods. One can also notice the disproportion of income in spatial terms. The east is poorer than the west. Ten subregions with the lowest average income[12] are in the east with the values between 2803.8 € and 3535.6 €.

Figure 7. Means personal income in terms of NUTS 3 geographical units

[12] nowosądecki, puławski, ostrołęcki, przemyski, chełmsko-zamojski, bialski, krośnieński, tarnowski, nowotarski, sandomiersko-jędrzejowski.

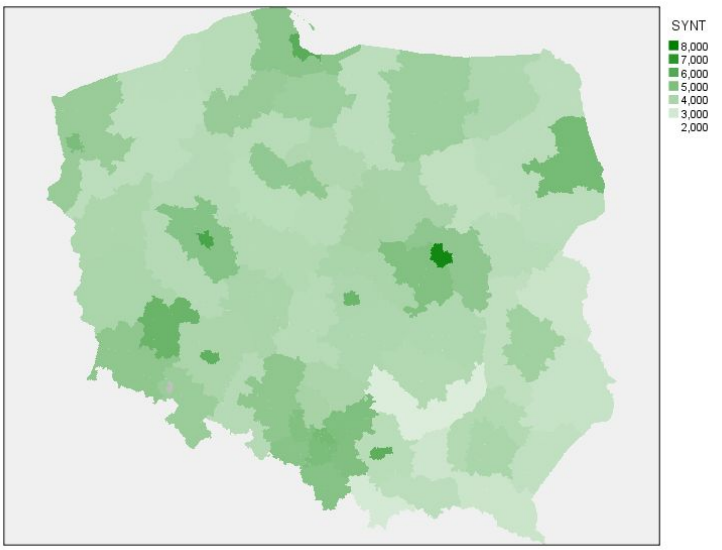
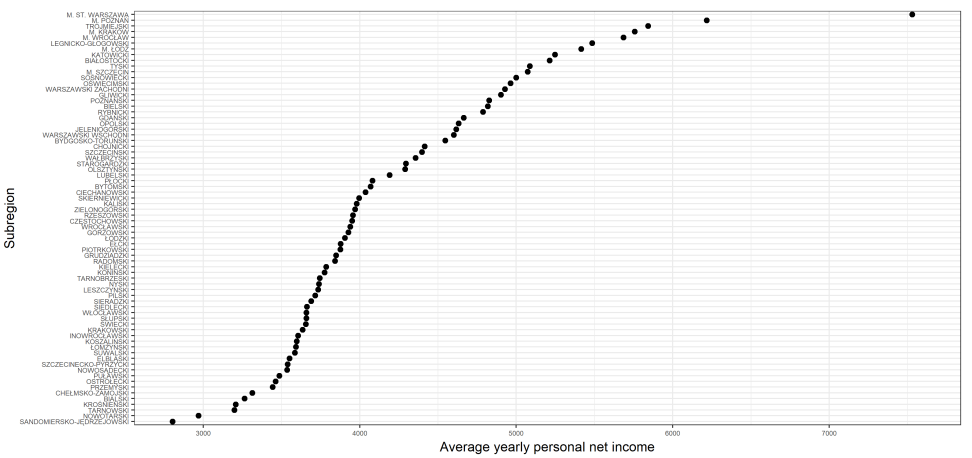


Figure 8. Estimated personal net income in terms of subregions

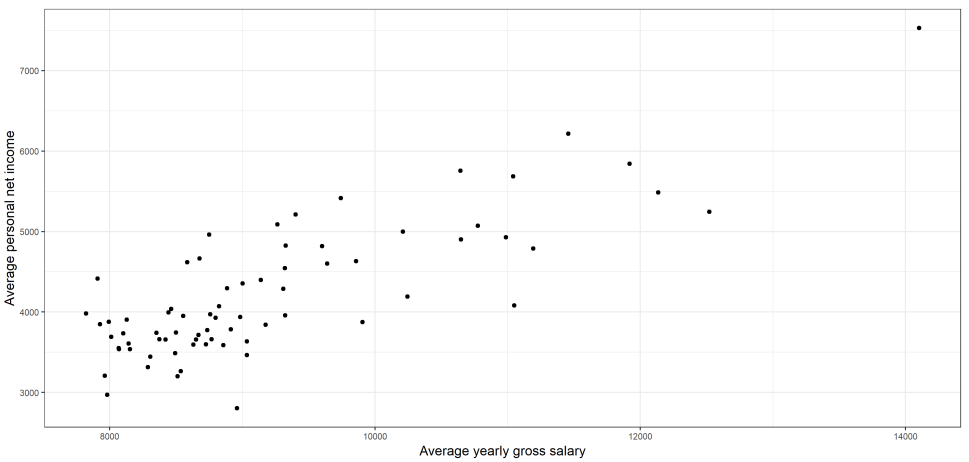


The comparison of the results with actual values is not possible. No data on personal income in terms of NUTS 3 (or any) spatial units is published. For the needs of the study the estimated mean personal net income was compared to the average yearly gross salary (excluding economic entities employing up to 9 persons), for which official statistics on the level of NUTS 3 are published (data are collected from monthly corporate reporting). The average salary was used as a proxy variable, which is correlated with the target variable thus it can serve as a reference point when trying to assess

the quality of the estimates. Salary is also one of the main components of personal income (for working people), so the definitions are similar.

As expected, the variables are strongly correlated with r=0.813 (see Figure 9[13]). This means that the estimates of the average net personal income in terms of subregions are convergent with reality.

Figure 9. Correlation diagram of estimated mean personal net income and average yearly gross salary in 2010 in subregions cross-section



The set of data created by spatial microsimulation techniques satisfactorily reflects the spatial distribution of the average annual net personal income in Poland. It also indicates the need to develop further techniques for assessing the quality of estimates.

4. Conclusions

Spatial microsimulation modeling satisfactorily reflected the spatial distribution of net personal income in Poland. The resulting synthetic population was characterized by consistent distributions, both spatial and joint. The main problem with the described methodology is inability to estimate the variance of estimators. This causes not only doubts about the legitimacy of using this method, but also prevents the comparison of results with other SAE estimators. The use of multiple data sources may cause overlapping

[13]The big difference in the size of personal income and remuneration is due to the fact that income is also calculated for the unemployed and inactive people, for whom the income is low, and zero in many cases.

with errors that accompany them. Random and non-random errors of the sample survey, possible coverage and administrative measurement errors of administrative data sources, discrepancy between census measurement and sample frame used in samples, spatial microsimulation model misspecification - all this (and more) affects the results and many are very difficult to recognize and verify. Reliable description of the properties of estimators is the most important task at the moment.

Nevertheless, the results of this and many other studies show that SMM is a good development direction of SAE methodology for socio-economic phenomena. Getting a full data matrix creates opportunities that have not been offered by any popular methods so far. This is particularly important when studying socio-economic phenomena of vital importance, like poverty, income, housing stress (and Laeken indicators), which are not the subject of any other research. Solving the problem of sample size, correction of random and non-random errors, the possibility of performing different simulations - these are undoubted advantages of the SMM methods that encourage to deepen the work and analysis of the effectiveness and reliability of the estimates.

REFERENCES

ALFONS A., KRAFT S. · TEMPL M., FILZMOSER P., (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. Statistical Methods and Applications, 20, pp. 383–407, Springer-Verlag.

BALLAS D., ROSSITER D., THOMAS B., CLARKE G.P., DORLING, D., (2005). Geography Matters: Simulating the Local Impacts of National Social Policies. York, Joseph Rowntree Foundation, UK.

D'ORAZIO M., DI ZIO M., SCANU M., (2006). Statistical Matching. Theory and Practice. John Wiley & Sons Ltd., England.

GOEDEMÉ T., (2013). Testing the Statistical Significance of Microsimulation Results: A Plea. International Journal Of Microsimulation, 6(3), pp. 50–77, International Microsimulation Association.

O'DONOGHUE C., (2014). Spatial Microsimulation Modeling: a Review of Applications and Methodological Choices. International Journal of Microsimulation, 7(1), pp. 26–75, International Microsimulation Association.

MUNOZ E., TANTON R., VIDYATTAMA Y., (2015). A comparison of the GREGWT and IPF methods for the re-weighting of surveys. 5th World Congress of the International Microsimulation Association (IMA).

MÜNNICH R., SCHÜRLE J., (2013). On the Simulation of Complex Universes in the Case of Applying the German Microcensus, DACSEIS research paper series no. 4.

NORMAN P., (1999). Putting Iterative Proportional Fitting on the Researcher's Desk. School of Geography, University of Leeds, UK.

PENNECK S., (2007). Using administrative data for statistical purposes. Economic & Labour Market Review.

PHAM D.T., and KARABOGA D., (2000). Intelligent optimization techniques: genetic algorithms, taboo search, simulated annealing and neural networks. London, Springer.

RAHMAN A., (2008). A review of small area estimation problems and methodological developments. *Online Discussion Paper - DP66*, NATSEM, University of Canberra.

RAHMAN A., (2009). Small Area Estimation Through Spatial Microsimulation Models: Some Methodological Issues. Paper Presented at the 2^{nd} International Microsimulation Association Conference, Ottawa, Canada, 8-10 June 2009, NATSEM, University of Canberra.

RAHMAN A., HARDING A., (2017). Small Area Estimation and Microsimulation Modeling. CRC Press, A Chapman & Hall Book, Boca Raton, Florida, USA.

RAHMAN A., HARDING A., TANTON R., LIU S., (2010). Methodological Issues in Spatial Microsimulation Modeling for Small Area Estimation. International Journal of Microsimulation, 3(2), pp. 3–22, International Microsimulation Association.

RAO J. N. K., (2003). Small Area Estimation. John Wiley & Sons.

RÄSSLER S., (2002). Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. Springer, New York, USA.

STATISTICS NEW ZEALAND, (2006). Data Integration Manual.

SZYMKOWIAK M., BERĘSEWICZ M., JÓZEFOWSKI T., KLIMANEK T., KOWALEWSKI J., MAŁASIEWICZ A., MŁODAK A., WAWROWSKI Ł., (2013). Mapy ubóstwa na poziomie podregionów w Polsce z wykorzystaniem estymacji pośredniej. Urząd Statystyczny w Poznaniu, Ośrodek Statystyki Małych Obszarów.

SZYMKOWIAK M., MŁODAK A., WAWROWSKI Ł., (2017). Mapping Poverty At The Level Of Subregions In Poland Using Indirect Estimation. STATIS-

TICS IN TRANSITION new series, December 2017, Vol. 18, No. 4, pp. 609–635.

TANTON R., (2014). A Review of Spatial Microsimulation Methods. International Journal of Microsimulation, 7(1), pp. 4-25, International Microsimulation Association.

TANTON R., EDWARDS K. L. *eds.*, (2013). Spatial Microsimulation: A Reference Guide for Users. Springer.

TANTON R., VIDYATTAMA Y., NEPAL B., MCNAMARA J., (2011). Small area estimation using a reweighing algorithm. Journal of the Royal Statistical Society, 174, Part 4, pp. 931–951.

TEMPL M., FILZMOSER P., (2014). Simulation and quality of a synthetic close-to-reality employer-employee population. Journal of Applied Statistics, Vol. 41, No. 5, pp. 1053–1072.

TEMPL M., MEINDL B., KOWARIK A., DUPRIEZ O., (2017). Simulation of Synthetic Complex Data: The R Package simPop. Journal of Statistical Software, August 2017, Vol. 79, Issue 10.

VOAS D., WILLIAMSON P., (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. International Journal of Population Geography, Vol. 6, pp. 349–366.

WALLGREN A., WALLGREN B., (2007). Register-based Statistics. Administrative Data for Statistical Purposes. John Wiley and Sons Ltd.

WAWROWSKI Ł., (2014). Wykorzystanie metod statystyki małych obszarów do tworzenia map ubóstwa w Polsce. Wiadomości Statystyczne, Vol. 9, pp. 46–56.

WILLIAMSON P., (2013). An Evaluation of Two Synthetic Small-Area Microdata Simulation Methodologies: Synthetic Reconstruction and Combinatorial Optimization [in:] Spatial Microsimulation: A Reference Guide for Users. Springer.

WHITWORTH (*edt*), (2013). Evaluation and improvements in small area estimation methodologies. National Centre for Research Methods, Methodological Review paper, University of Sheffield.

WHITWORTH A., CARTER E., BALLAS D., MOON G., (2016). Estimating uncertainty in spatial microsimulation approaches to small area estimation: A new approach to solving an old problem. Computers, Environment and Urban Systems,
http://dx.doi.org/10.1016/j.compenvurbsys.2016.06.004.

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. 155–170, DOI 10.21307/stattrans-2019-029
Submitted – 04.03.2019; Paper ready for publication – 14.05.2019

POWER GENERALIZATION OF CHEBYSHEV'S INEQUALITY – MULTIVARIATE CASE

Katarzyna Budny [1]

ABSTRACT

In the paper some multivariate power generalizations of Chebyshev's inequality and their improvements will be presented with extension to a random vector with singular covariance matrix. Moreover, for these generalizations, the cases of the multivariate normal and the multivariate *t* distributions will be considered. Additionally, some financial application will be presented.

Key words: multivariate Chebyshev's inequality, Mahalanobis distance, multivariate normal distribution, multivariate *t* distribution.

1. Introduction

Chebyshev's inequality yields a bound on the probability of a univariate random variable taking values close to the mean expressed by its variance. Pearson (1919) proposed its univariate power generalization presenting bounds by the central moments of a random variable of even orders.

Theorem 1.1. (Pearson, 1919). If we take a random variable $\xi : \Omega \to R$ with finite central moments of *2s* order $\left(\mu_{2s}\right)$, then for all $\varepsilon > 0$

$$P\left(\left|\xi - E\left(\xi\right)\right| \geq \varepsilon\sigma\right) \leq \frac{\mu_{2s}}{\varepsilon^{2s}\sigma^{2s}} . \tag{1.1}$$

There also exist multivariate generalizations of Chebyshev's inequality (see, e.g. Olkin and Pratt, 1958, Marshall and Olkin, 1960, Osiewalski and Tatar, 1999).

In the paper we present one of those providing upper bounds on the probability that the Mahalanobis distance of a random vector from its mean is greater or equal than the fixed value. These bounds will be given by the power transformations and will constitute the multivariate extension of (1.1).

There are many applications of the Mahalanobis distance in statistical analysis. In particular, this is used in classification methods and in cluster analysis. The multivariate power generalization of Chebyshev's inequality presented below can be exploited to detect outliers.

[1] Department of Mathematics, Cracow University of Economics, Poland.
E-mail: budnyk@uek.krakow.pl. ORCID ID: https://orcid.org/0000-0002-3683-0327.

2. Multivariate power generalization of Chebyshev's inequality

We begin by recalling the inequality which is given by the measure of multivariate kurtosis.

Theorem 2.1. (Mardia, 1970) Let $\mathbf{X}:\Omega\to R^n$ be a random vector with nonsingular covariance matrix $\mathbf{\Sigma}$ and finite fourth–order moments. Then, for any $\varepsilon>0$ the following inequality holds

$$P\Big(\big(\mathbf{X}-E(\mathbf{X})\big)^T\mathbf{\Sigma}^{-1}\big(\mathbf{X}-E(\mathbf{X})\big)\geq\varepsilon\Big)\leq\frac{\beta_{2,n}(\mathbf{X})}{\varepsilon^2}, \tag{2.1}$$

where $\beta_{2,n}(\mathbf{X})=E\left[\Big(\big(\mathbf{X}-E(\mathbf{X})\big)^T\mathbf{\Sigma}^{-1}\big(\mathbf{X}-E(\mathbf{X})\big)\Big)^2\right]$ is Mardia's kurtosis of a random vector (Mardia, 1970).

Chen (2007, 2011) proposed a tight upper bound (see Navarro, 2014) in the case of a random vector for which only mean and covariance matrix are known.

Theorem 2.2. (Chen, 2007, Chen, 2011) Assume that $\mathbf{X}:\Omega\to R^n$ is a random vector with positive covariance matrix $\mathbf{\Sigma}$. Then, for all $\varepsilon>0$ we get

$$P\Big(\big(\mathbf{X}-E(\mathbf{X})\big)^T\mathbf{\Sigma}^{-1}\big(\mathbf{X}-E(\mathbf{X})\big)\geq\varepsilon\Big)\leq\frac{n}{\varepsilon}. \tag{2.2}$$

Budny (2014) obtained the multivariate power generalization of Chebyshev's inequality.

Theorem 2.3. (Budny, 2014) Suppose that $\mathbf{X}:\Omega\to R^n$ is a random vector with nonsingular covariance matrix $\mathbf{\Sigma}$. Let us consider any $s>0$ such that

$$I_{s,n}(\mathbf{X})=E\left[\Big(\big(\mathbf{X}-E(\mathbf{X})\big)^T\mathbf{\Sigma}^{-1}\big(\mathbf{X}-E(\mathbf{X})\big)\Big)^s\right]$$

exists. Then, for all $\varepsilon>0$

$$P\Big(\big(\mathbf{X}-E(\mathbf{X})\big)^T\mathbf{\Sigma}^{-1}\big(\mathbf{X}-E(\mathbf{X})\big)\geq\varepsilon\Big)\leq\frac{I_{s,n}(\mathbf{X})}{\varepsilon^s}. \tag{2.3}$$

Remark 2.1. (Budny, 2014) Observe that theorems 2.1 and 2.2 can be considered as the special cases of theorem 2.3. Taking $s=1$ we get (2.2) and for $s=2$ we obtain (2.1).

Budny (2016), following Navarro (2016), extended (2.3) to the case of a random vector with singular covariance matrix by using the spectral decomposition.

Assume that $\mathbf{X}:\Omega\to R^n$ is a random vector with covariance matrix $\mathbf{\Sigma}$, $\operatorname{rank}\mathbf{\Sigma}=m$, $m\in\{1,...,n\}$. Let $\mathbf{\Sigma}=P\Lambda P^T$ be a spectral decomposition of a covariance matrix, i.e. P is an orthogonal matrix such that $PP^T=P^TP=I_n$ and $\Lambda=\operatorname{diag}(\lambda_1,...,\lambda_m,0,...,0)$ is the diagonal matrix with the ordered eigenvalues

$$\lambda_1\geq...\geq\lambda_m>\lambda_{m+1}=...=\lambda_n=0.$$

Hence, the Moore-Penrose generalized inverse matrix of $\mathbf{\Sigma}$ is of the form $\mathbf{\Sigma}^+=PCP^T$, where $C=\operatorname{diag}(\lambda_1^{-1},...,\lambda_m^{-1},0,...,0)$.

Let us consider any $s>0$ such that

$$I_{s,m}(\mathbf{X})=E\left[\left((\mathbf{X}-E(\mathbf{X}))^T\mathbf{\Sigma}^+(\mathbf{X}-E(\mathbf{X}))\right)^s\right]$$

exists.

Theorem 2.4. (Budny, 2016) Under the above assumptions, for any $\varepsilon>0$, we have

$$P\left((\mathbf{X}-E(\mathbf{X}))^T\mathbf{\Sigma}^+(\mathbf{X}-E(\mathbf{X}))\geq\varepsilon\right)\leq\frac{I_{s,m}(\mathbf{X})}{\varepsilon^s}. \tag{2.4}$$

We will denote by S the set of all $s>0$ such that $I_{s,m}(\mathbf{X})$ exists. Let us define, for fixed $\varepsilon>0$, the function $Bd:S\to R_+$:

$$Bd(s)=\frac{I_{s,m}(\mathbf{X})}{\varepsilon^s},\qquad s\in S.$$

It is easily seen that for $s_1,s_2\in S$ if $s_1<s_2$, then the following conditions are equivalent:

$$Bd(s_1)>Bd(s_2)\quad\Leftrightarrow\quad\varepsilon>\left(\frac{I_{s_2,m}(\mathbf{X})}{I_{s_1,m}(\mathbf{X})}\right)^{\frac{1}{s_2-s_1}}$$

and

$$Bd(s_1)<Bd(s_2)\quad\Leftrightarrow\quad0<\varepsilon<\left(\frac{I_{s_2,m}(\mathbf{X})}{I_{s_1,m}(\mathbf{X})}\right)^{\frac{1}{s_2-s_1}}.$$

Summarizing, we get following remark.

Remark 2.2. For $s_1, s_2 \in S$ if $s_1 < s_2$, then the upper bound $Bd(s_2)$ of $P\left(\left(\mathbf{X} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^+ \left(\mathbf{X} - E(\mathbf{X})\right) \geq \varepsilon\right)$ is better than $Bd(s_1)$ for all $\varepsilon > \left(\frac{I_{s_2,m}(\mathbf{X})}{I_{s_1,m}(\mathbf{X})}\right)^{\frac{1}{s_2 - s_1}}$. On the contrary, the upper bound $Bd(s_1)$ is better than $Bd(s_2)$ for all $\varepsilon \in \left(0, \left(\frac{I_{s_2,m}(\mathbf{X})}{I_{s_1,m}(\mathbf{X})}\right)^{\frac{1}{s_2 - s_1}}\right)$.

In particular, if we consider $s_1 = 1$, $s_2 = 2$ and $\operatorname{rank}\boldsymbol{\Sigma} = n$, then the upper bound $Bd(s_2)$ is better than $Bd(s_1)$ for all $\varepsilon > \frac{\beta_{2,n}(\mathbf{X})}{n}$. Conversely, the upper bound $Bd(s_1)$ is better than $Bd(s_2)$ for all $\varepsilon \in \left(0, \frac{\beta_{2,n}(\mathbf{X})}{n}\right)$.

3. The case of the multivariate normal distribution

Budny (2016) proposed the form of the multivariate power generalization of Chebyshev's inequality for a normally distributed random vector for all $s \in N \setminus \{0\}$. In the next theorem we extend this result to the case of any real $s > 0$.

Theorem 3.1. Let $\mathbf{X}: \Omega \to R^n$ be a normally distributed random vector with mean μ and covariance matrix $\boldsymbol{\Sigma}$, $\mathbf{X} \sim N_n(\mu, \boldsymbol{\Sigma})$. Suppose that $\operatorname{rank}\boldsymbol{\Sigma} = m$, $m \in \{1, ..., n\}$. Then, for all $\varepsilon > 0$ and $s > 0$ we obtain

$$P\left(\left(\mathrm{X} - \mu\right)^T \boldsymbol{\Sigma}^+ \left(\mathrm{X} - \mu\right) \geq \varepsilon\right) \leq \left(\frac{2}{\varepsilon}\right)^s \cdot \frac{\Gamma\left(\frac{m}{2} + s\right)}{\Gamma\left(\frac{m}{2}\right)}. \tag{3.1}$$

Proof: The proof is similar to that presented for theorem 3.1 in Budny (2016). A slight change is that we consider *s*th uncorrected moment (*s*th moment about zero) of a chi-square distribution with *m* degrees of freedom for any real $s > 0$ (not only for $s \in N \setminus \{0\}$). Hence, for $s > 0$:

$$I_{s,m}(\mathbf{X}) = \frac{2^s \Gamma\left(\frac{m}{2} + s\right)}{\Gamma\left(\frac{m}{2}\right)}$$

(Johnson, Kotz and Balakrishnan, 1994, p. 420) and it completes the proof.

Remark 3.1. For $s \in N \setminus \{0\}$ the inequality (3.1) takes the following form

$$P\left(\left(\mathbf{X}-\mu\right)^T \boldsymbol{\Sigma}^+ \left(\mathbf{X}-\mu\right) \geq \varepsilon\right) \leq \frac{m \cdot \left(m+2\right) \cdot \ldots \cdot \left(m+2\left(s-1\right)\right)}{\varepsilon^s}$$ (Budny, 2016).

Remark 3.2. On account of remark 2.2, if $s_1 < s_2$, then the upper bound $Bd\left(s_2\right)$ is better than $Bd\left(s_1\right)$ for all $\varepsilon > 2 \cdot \left(\Gamma\left(\frac{m}{2}+s_2\right) \Big/ \Gamma\left(\frac{m}{2}+s_1\right)\right)^{\frac{1}{s_2-s_1}}$. Conversely, the upper bound $Bd\left(s_1\right)$ is better than $Bd\left(s_2\right)$ for all $\varepsilon \in \left(0,\, 2 \cdot \left(\Gamma\left(\frac{m}{2}+s_2\right) \Big/ \Gamma\left(\frac{m}{2}+s_1\right)\right)^{\frac{1}{s_2-s_1}}\right)$.

Particularly if we take $s_1 = 1$ and $s_2 = 2$, then the upper bound $Bd\left(s_2\right)$ is better than $Bd\left(s_1\right)$ for all $\varepsilon > m+2$. Conversely, the upper bound $Bd\left(s_1\right)$ is better than $Bd\left(s_2\right)$ for all $\varepsilon \in \left(0,\, m+2\right)$.

Example 3.1. Let us consider normally distributed random vector $\mathbf{X}: \Omega \to R^n$ with mean μ and covariance matrix $\boldsymbol{\Sigma}$, $\mathbf{X} \sim N_n\left(\mu, \boldsymbol{\Sigma}\right)$. Assume that $\operatorname{rank}\boldsymbol{\Sigma} = m = 3$.

A random variable $\left(\mathbf{X}-\mu\right)^T \boldsymbol{\Sigma}^+ \left(\mathbf{X}-\mu\right)$ has a chi-square distribution with *m* degrees of freedom (Kotz, Balakrishnan and Johnson, 2000, p. 110, Budny, 2016), hence we know the exact value of P.

From remark 3.2 for $s_1 = 1$ and $s_2 = 2$ we get that the upper bound $Bd\left(s_2\right)$ is better than $Bd\left(s_1\right)$ for all $\varepsilon > m+2 = 5$ and the upper bound $Bd\left(s_1\right)$ is better than $Bd\left(s_2\right)$ for all $\varepsilon \in \left(0, 5\right)$ (see Figure 3.1).

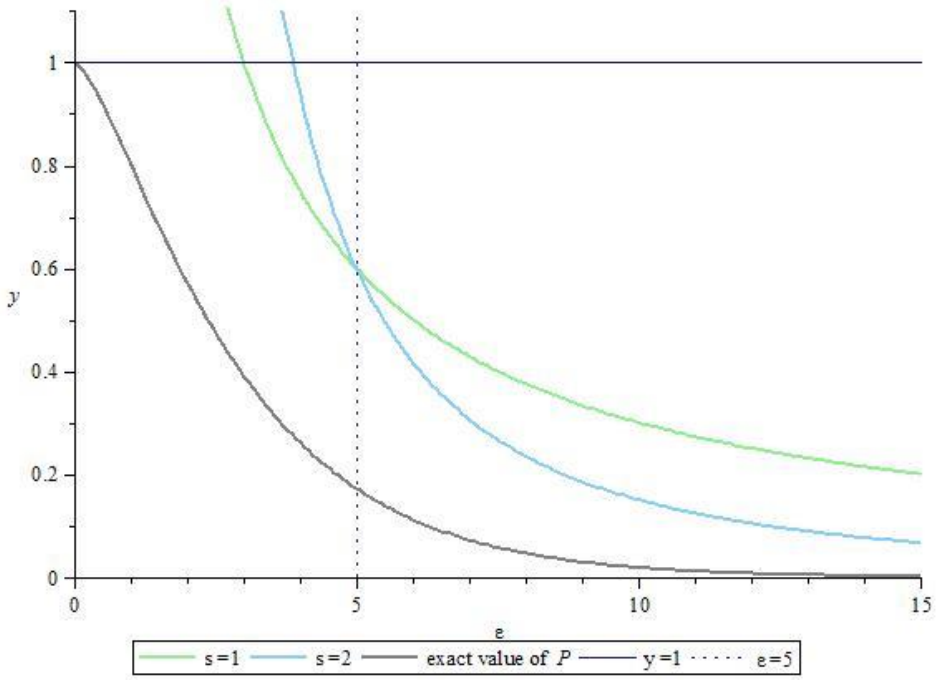


Figure 3.1. The upper bounds ($s=1$, $s=2$) and exact value of P for $\mathbf{X} \sim N_n(\mu, \mathbf{\Sigma})$, $\operatorname{rank} \mathbf{\Sigma} = 3$.

In turn Figure 3.2 shows the upper bounds of $P\left((\mathbf{X}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{X}-\mu) \geq \varepsilon\right)$ for various values of s.

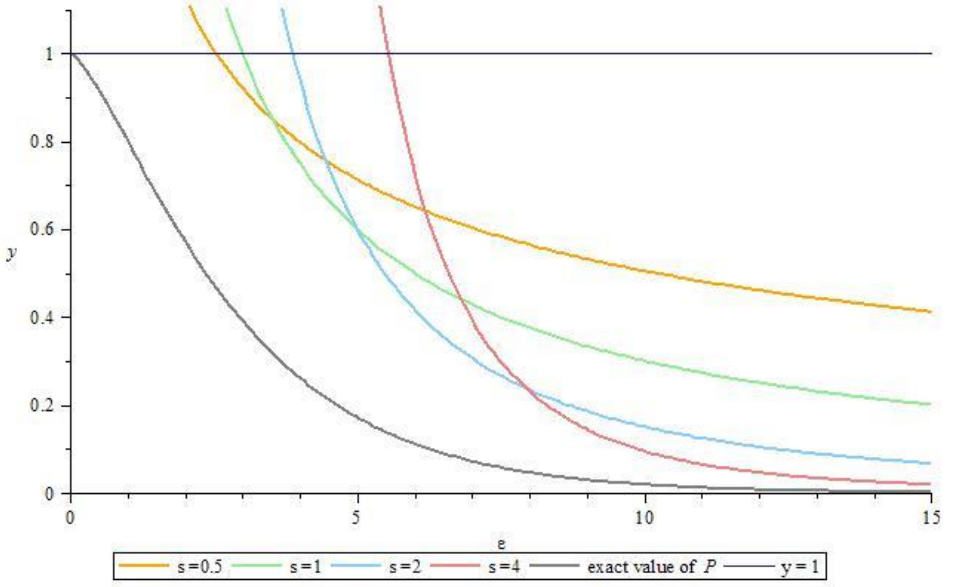


Figure 3.2. The upper bounds ($s=0.5$, $s=1$, $s=2$, $s=4$) and exact value of P for $\mathbf{X} \sim N_n(\mu, \mathbf{\Sigma})$, rank $\mathbf{\Sigma} = 3$.

4. The case of the multivariate *t* distribution

A $n-$variate random vector $\mathbf{X}: \Omega \to R^n$ is said to have multivariate *t* distribution with degrees of freedom ν, mean μ and nonsingular covariance matrix $\frac{\nu}{\nu-2}\mathbf{R}$, $\nu > 2$, denoted by $t_\nu(\mu, \mathbf{R}, n)$, if its joint probability density function (pdf) is given by

$$f(x) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\pi\nu)^{n/2}\Gamma\left(\frac{\nu}{2}\right)|R|^{1/2}}\left[1+\frac{1}{\nu}(x-\mu)^T R^{-1}(x-\mu)\right]^{-(\nu+n)/2} \qquad \left(x \in R^n\right).$$

If $\mathbf{X} \sim t_\nu(\mu, \mathbf{R}, n)$, then the random variable $\xi = \frac{(\mathbf{X}-\mu)^T \mathbf{R}^{-1}(\mathbf{X}-\mu)}{n}$ has a central *F*-distribution with n, ν degrees of freedom, $\xi \sim F(n, \nu)$ (Lin, 1972).

It follows that for any $s < \frac{\nu}{2}$ we get

$$E\left[\xi^s\right] = \left(\frac{\nu}{n}\right)^s \cdot \frac{\Gamma\left(\frac{n}{2}+s\right)\Gamma\left(\frac{\nu}{2}-s\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{\nu}{2}\right)} \tag{4.1}$$

(Johnson, Kotz and Balakrishnan, 1995, p. 349).

The power generalization of Chebyshev's inequality for multivariate *t* distribution is established by our next theorem.

Theorem 4.1. Assume that $\mathbf{X} \sim t_\nu(\mu, \mathbf{R}, n)$, $\operatorname{rank}\mathbf{R} = n$. Then, for any $\varepsilon > 0$ the inequality (2.3) takes the following form

$$P\left((\mathbf{X}-\mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mu) \geq \varepsilon\right) \leq \left(\frac{\nu-2}{\varepsilon}\right)^s \cdot \frac{\boldsymbol{\Gamma}\left(\frac{n}{2}+s\right)\boldsymbol{\Gamma}\left(\frac{\nu}{2}-s\right)}{\boldsymbol{\Gamma}\left(\frac{n}{2}\right)\boldsymbol{\Gamma}\left(\frac{\nu}{2}\right)} \tag{4.2}$$

for any $s > 0$ such that $s < \frac{\nu}{2}$.

Proof: We first observe that $\boldsymbol{\Sigma}^{-1} = \frac{\nu-2}{\nu}\mathbf{R}^{-1}$. From this it is obvious that

$$I_{s,n}(\mathbf{X}) = \left(\frac{\nu-2}{\nu}\right)^s \cdot n^s \cdot E\left[\xi^s\right]. \tag{4.3}$$

Substituting (4.1) into (4.3) yields

$$I_{s,n}(\mathbf{X}) = (\nu-2)^s \cdot \frac{\Gamma\left(\frac{n}{2}+s\right)\Gamma\left(\frac{\nu}{2}-s\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}. \tag{4.4}$$

This establishes the inequality (4.2).

Remark 4.1. For $s \in N \setminus \{0\}$, $s < \frac{\nu}{2}$ from (4.4) we get

$$I_{s,n}(\mathbf{X}) = \frac{(\nu-2)^s \cdot n \cdot (n+2) \cdot \ldots \cdot (n+2(s-1))}{(\nu-2)\cdot(\nu-4)\cdot\ldots\cdot(\nu-2s)}. \tag{4.5}$$

Hence, the inequality (4.2) is of the form

$$P\left((\mathrm{X}-\mu)^T \boldsymbol{\Sigma}^{-1}(\mathrm{X}-\mu) \geq \varepsilon\right) \leq \left(\frac{\nu-2}{\varepsilon}\right)^s \frac{n \cdot (n+2) \cdot \ldots \cdot (n+2(s-1))}{(\nu-2)\cdot(\nu-4)\cdot\ldots\cdot(\nu-2s)}.$$

Remark 4.2. According to remark 2.2, if $s_1 < s_2 < \frac{\nu}{2}$, then the upper bound $Bd(s_2)$ is better than $Bd(s_1)$ for all $\varepsilon > (\nu-2)\left(\frac{\Gamma\left(\frac{n}{2}+s_2\right)\Gamma\left(\frac{\nu}{2}-s_2\right)}{\Gamma\left(\frac{n}{2}+s_1\right)\Gamma\left(\frac{\nu}{2}-s_1\right)}\right)^{\frac{1}{s_2-s_1}}$.

On the contrary, the upper bound $Bd(s_1)$ is better than $Bd(s_2)$ for all

$$\varepsilon \in \left(0, (\nu-2)\left(\frac{\Gamma\left(\frac{n}{2}+s_2\right)\Gamma\left(\frac{\nu}{2}-s_2\right)}{\Gamma\left(\frac{n}{2}+s_1\right)\Gamma\left(\frac{\nu}{2}-s_1\right)}\right)^{\frac{1}{s_2-s_1}}\right).$$

In particular, if we consider $s_1 = 1$, $s_2 = 2$ and $\nu > 4$, then from (4.5) the upper bound $Bd(s_2)$ is better than $Bd(s_1)$ for all $\varepsilon > \frac{(\nu-2)\cdot(n+2)}{(\nu-4)}$. Conversely, the upper bound $Bd(s_1)$ is better than $Bd(s_2)$ for all $\varepsilon \in \left(0, \frac{(\nu-2)\cdot(n+2)}{(\nu-4)}\right)$.

Example 4.1. We will consider a random vector $\mathbf{X}: \Omega \to R^3$ that has multivariate *t* distribution with degrees of freedom $\nu = 9$, $\mathbf{X} \sim t_9(\mu, \mathbf{R}, 3)$, $\operatorname{rank}\mathbf{R} = 3$. Let us observe that

$$P\left((\mathbf{X}-\mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mu) \geq \varepsilon\right) = P\left(\frac{(\mathbf{X}-\mu)^T \mathbf{R}^{-1}(\mathbf{X}-\mu)}{n} \geq \frac{\nu\varepsilon}{(\nu-2)n}\right),$$

thus we know the exact value of P. For $s_1 = 1$ and $s_2 = 2$ we get $\varepsilon_0 = \frac{(\nu-2)\cdot(n+2)}{(\nu-4)} = 7$. From this it follows that the upper bound $Bd(s_2)$ is better than $Bd(s_1)$ for all $\varepsilon > 7$ and the upper bound $Bd(s_1)$ is better than $Bd(s_2)$ for all $\varepsilon \in (0, 7)$ (see Figure 4.1).

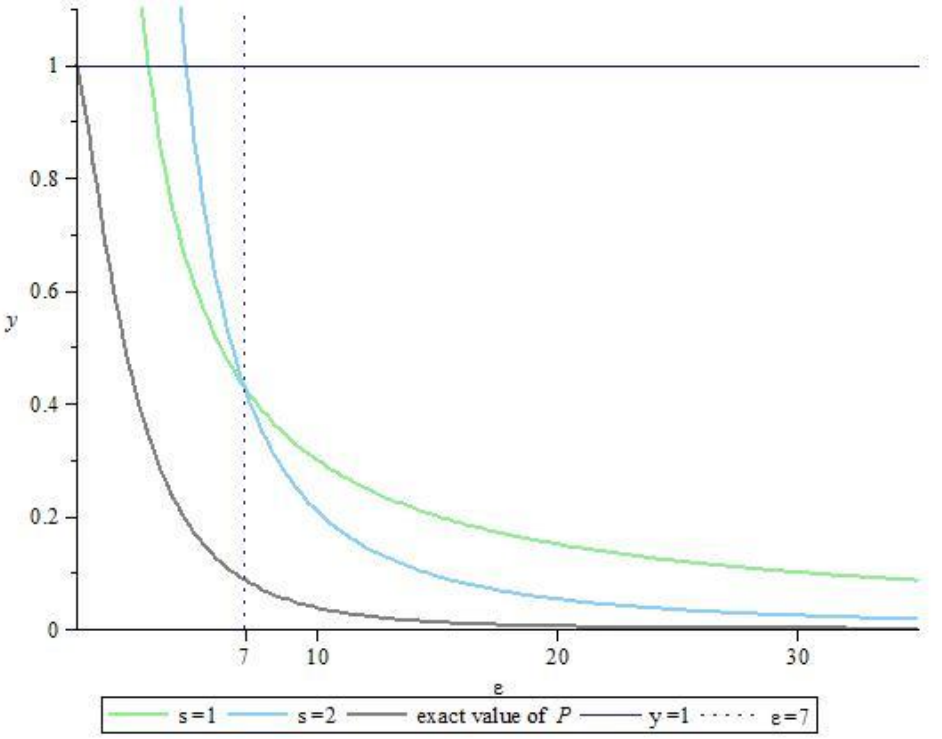


Figure 4.1. The upper bounds ($s=1$, $s=2$) and exact value of P for $\mathbf{X} \sim t_9(\mu, \mathbf{R}, 3)$.

The next figure presents the upper bounds of $P\left((\mathbf{X}-\mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mu) \geq \varepsilon\right)$ for various values of $s < \frac{\nu}{2} = 4.5$.

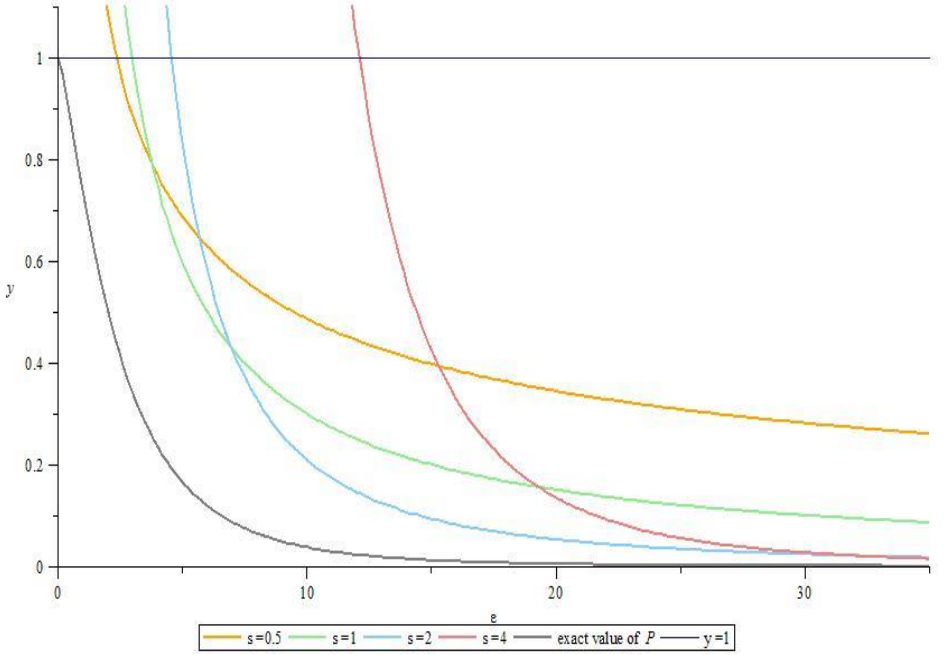


Figure 4.2. The upper bounds ($s=0.5$, $s=1$, $s=2$, $s=4$) and exact value of P for $\mathbf{X} \sim t_9(\mu, \mathbf{R}, 3)$.

5. Improvement of some multivariate power generalization of Chebyshev's inequality – extension to a random vector with a singular covariance matrix

In this section we will consider some improvement of multivariate power generalization of Chebyshev's inequality. We should mention that this improvement will be given by restricting the range of ε, it means for ε sufficiently large.

Loperfido (2014) proposed improvement of the inequality (2.4) for $s=2$ in the case of a random vector with nonsingular covariance matrix.

Theorem 5.1. (Loperfido, 2014) Let $\beta_{2,n}(\mathbf{X})$ be Mardia's kurtosis of a random vector $\mathbf{X}: \Omega \to R^n$ with nonsingular covariance matrix $\boldsymbol{\Sigma}$ and finite fourth–order moments. Then for any $\varepsilon > n$ the following inequalities hold

$$P\left(\left(\mathbf{X}-E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}-E(\mathbf{X})\right) \geq \varepsilon\right) \leq \frac{\beta_{2,n}(\mathbf{X})-n^2}{\varepsilon^2-2n\varepsilon+\beta_{2,n}(\mathbf{X})} \leq \frac{\beta_{2,n}(\mathbf{X})}{\varepsilon^2}. \qquad (5.1)$$

In the next theorem we will extend (5.1) to the case of a random vector with singular covariance matrix.

Theorem 5.2. Assume that $\mathbf{X}:\Omega\to R^n$ is a random vector with covariance matrix $\mathbf{\Sigma}$, $\operatorname{rank}\mathbf{\Sigma}=m$, $m\in\{1,...,n\}$. Let Y denote the random variable $Y=(\mathbf{X}-E(\mathbf{X}))^T\mathbf{\Sigma}^+(\mathbf{X}-E(\mathbf{X}))$. Let us consider Y such that $I_{2,m}(\mathbf{X})=E(Y^2)$ is finite. Then, for all $\varepsilon>m$ we obtain

$$P\left((\mathbf{X}-E(\mathbf{X}))^T\mathbf{\Sigma}^+(\mathbf{X}-E(\mathbf{X}))\geq\varepsilon\right)\leq\frac{I_{2,m}(\mathbf{X})-m^2}{\varepsilon^2-2m\varepsilon+I_{2,m}(\mathbf{X})}\leq\frac{I_{2,m}(\mathbf{X})}{\varepsilon^2}. \qquad (5.2)$$

Proof: At the beginning we should mention that the proof will be similar to that presented in Loperfido (Loperfido 2014, proof of Theorem 1) for $s=2$.

We first consider the random variable $Z=\left\{(Y-m)(\varepsilon-m)+I_{2,m}(\mathbf{X})-m^2\right\}^2$. Let us observe that

$$E(Z)=(\varepsilon-m)^2E\left((Y-m)^2\right)+\left(I_{2,m}(\mathbf{X})-m^2\right)^2=\left(I_{2,m}(\mathbf{X})-m^2\right)\cdot\left(\varepsilon^2-2m\varepsilon+I_{2,m}(\mathbf{X})\right)$$

.

This gives that expected value of a nonnegative random variable

$$W=\frac{Z}{\left\{\varepsilon^2-2m\varepsilon+I_{2,m}(\mathbf{X})\right\}^2}$$

exists for all $\varepsilon>m$ and takes the form $E(W)=\dfrac{I_{2,m}(\mathbf{X})-m^2}{\varepsilon^2-2m\varepsilon+I_{2,m}(\mathbf{X})}$. Markov's inequality implies that

$$P(W\geq1)\leq\frac{I_{2,m}(\mathbf{X})-m^2}{\varepsilon^2-2m\varepsilon+I_{2,m}(\mathbf{X})}.$$

The assumption $\varepsilon>m$, in turn, leads to equality $P(Y\geq\varepsilon)=P(W\geq1)$. Indeed, the set $D=\{Y\geq\varepsilon\}$ takes the form $D=\{Y-m\geq\varepsilon-m\}$. It follows that for $\varepsilon>m$:

$$D=\left\{(Y-m)(\varepsilon-m)+\left(I_{2,m}(\mathbf{X})-m^2\right)\geq\varepsilon^2-2m\varepsilon+I_{2,m}(\mathbf{X})\right\}.$$

Thus, from Jensen's inequality we have:

$$D=\left\{\ Z\geq\left\{\varepsilon^2-2m\varepsilon+I_{2,m}(\mathbf{X})\right\}^2\ \right\}=\{W\geq1\}.$$

Finally, we get

$$P\left((\mathbf{X}-E(\mathbf{X}))^T\mathbf{\Sigma}^+(\mathbf{X}-E(\mathbf{X}))\geq\varepsilon\right)\leq\frac{I_{2,m}(\mathbf{X})-m^2}{\varepsilon^2-2m\varepsilon+I_{2,m}(\mathbf{X})}.$$

To prove the second inequality $\frac{I_{2,m}(\mathbf{X})}{\varepsilon^2} \geq \frac{I_{2,m}(\mathbf{X})-m^2}{\varepsilon^2-2m\varepsilon+I_{2,m}(\mathbf{X})}$, let us observe that simple transformations lead to equivalent form $\frac{\left(I_{2,m}(\mathbf{X})-m\varepsilon\right)^2}{\varepsilon^2\left\{(\varepsilon-m)^2+I_{2,m}(\mathbf{X})-m^2\right\}} \geq 0$, which holds for all $\varepsilon > m$, since $I_{2,m}(\mathbf{X}) \geq m^2$.

Remark 5.1. If $\operatorname{rank}\boldsymbol{\Sigma} = m = n$, then $\boldsymbol{\Sigma}^+ = \boldsymbol{\Sigma}^{-1}$ and $I_{2,m}(\mathbf{X}) = \beta_{2,n}(\mathbf{X})$. Hence, we obtain the inequalities (5.1).

Remark 5.2. Let $\mathbf{X}: \Omega \to R^n$ be normally distributed random vector with mean μ and covariance matrix $\boldsymbol{\Sigma}$, $\mathbf{X} \sim N_n(\mu, \boldsymbol{\Sigma})$. Assume that $\operatorname{rank}\boldsymbol{\Sigma} = m$, $m \in \{1,..., n\}$. Then, for any $\varepsilon > m$ we get:

$$P\left((\mathbf{X}-\mu)^T \boldsymbol{\Sigma}^+ (\mathbf{X}-\mu) \geq \varepsilon\right) \leq \frac{m}{\varepsilon^2 - 2m\varepsilon + m\cdot(m+2)} \leq \frac{m\cdot(m+2)}{\varepsilon^2}. \qquad (5.3)$$

We will illustrate the improvement (5.3) for $m = 3$ with Figure 5.1.

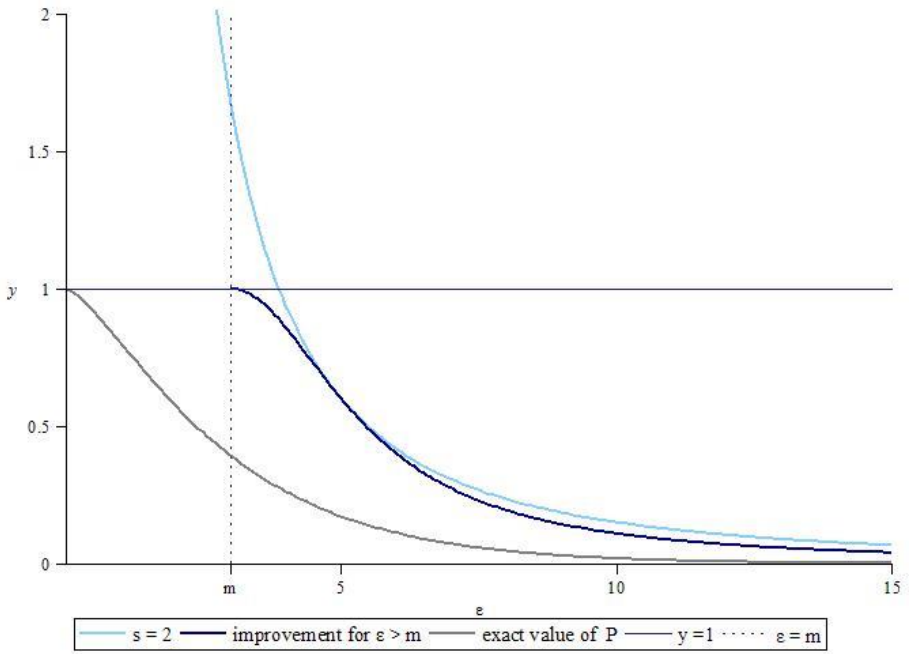


Figure 5.1. The upper bound $s = 2$, its improvement and exact value of P for $\mathbf{X} \sim N_n(\mu, \boldsymbol{\Sigma})$, $\operatorname{rank}\boldsymbol{\Sigma} = 3$.

Remark 5.3. Let us consider $\mathbf{X} \sim t_\nu(\mu, \mathbf{R}, n)$, rank $\mathbf{R} = n$, $\nu > 4$. Then, from (4.5) for any $\varepsilon > n$ the inequalities (5.1) take the following form

$$P\left((\mathbf{X}-\mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mu) \geq \varepsilon\right) \leq \frac{\left(\frac{\nu-2}{\nu-4}\right)\cdot n\cdot(n+2) - n^2}{\varepsilon^2 - 2n\varepsilon + \left(\frac{\nu-2}{\nu-4}\right)\cdot n\cdot(n+2)} \leq \frac{\left(\frac{\nu-2}{\nu-4}\right)\cdot n\cdot(n+2)}{\varepsilon^2} \tag{5.4}$$

As the example, we will consider a 3-variate random vector $\mathbf{X} \sim t_9(\mu, \mathbf{R}, 3)$, rank $\mathbf{R} = 3$ (see example 4.1). The inequalities (5.4) are presented in Figure 5.2.

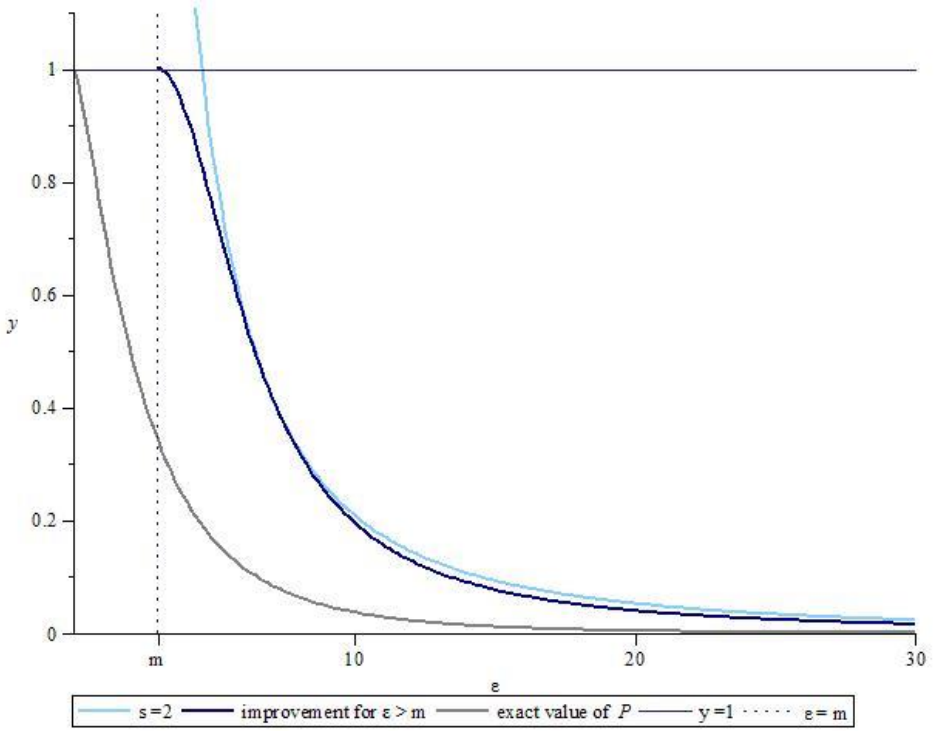


Figure 5.2. The upper bound $s = 2$, its improvement and exact value of P for $\mathbf{X} \sim t_9(\mu, \mathbf{R}, 3)$.

6. Applications in finance

Let us take a random vector r_t of *n* assets returns on a specific day t with mean μ (sample mean vector of historical returns) and covariance matrix $\mathbf{\Sigma}$ (sample covariance matrix of historical returns). Kritzman and Li (2010) propose to use the Mahalanobis distance as a measure of financial turbulence, which is understood as occurrence of unusual multivariate financial data. They defined (the so-called "the turbulence index") turbulence for a particular time t as:

$$d_t = (r_t - \mu)^T \mathbf{\Sigma}^{-1} (r_t - \mu).$$

In the examples presented in section 3 and 4, for any $\varepsilon > 0$, we know the exact value of $P\left(d_t = (r_t - \mu)^T \mathbf{\Sigma}^{-1} (r_t - \mu) \geq \varepsilon\right)$. In the general case, this probability may not be easy to compute and if we are able to calculate the upper bounds (5.2), then we can estimate the exact value of *P*.

Other financial applications of the Mahalanobis distance were presented by Stöckl and Hanke (2014).

Acknowledgement

The publication was financed from the funds granted to the Faculty of Finance and Law at Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

REFERENCES

BUDNY, K., (2014). A generalization of Chebyshev's inequality for Hilbert-space-valued random elements. Statistics and Probability Letters, 88, pp. 62–65.

BUDNY, K., (2016). An extension of the multivariate Chebyshev's inequality to a random vector with a singular covariance matrix, Communications in Statistics – Theory and Methods, 45 (17), pp. 5220–5223.

CHEN, X., (2007). A new generalization of Chebyshev inequality for random vectors. Available at: <https://arxiv.org/abs/0707.0805>[Accessed 5 July 2007].

CHEN, X., (2011). A new generalization of Chebyshev inequality for random vectors. Available at: <https://arxiv.org/abs/0707.0805v2>[Accessed 24 June 2011].

JOHNSON, N.L., KOTZ, S., BALAKRISHNAN, N., (1994). Continuous univariate distribution. Vol. 1, 2nd ed. John Wiley & Sons Inc.

JOHNSON, N.L., KOTZ, S., BALAKRISHNAN, N., (1995). Continuous univariate distribution, Vol. 2, 2nd ed. John Wiley & Sons Inc.

KRITZMAN, M., Li, Y., (2010). Skulls, financial turbulence, and risk management. Financial Analysts Journal, 66 (5), pp. 30–41.

KOTZ, S., BALAKRISHNAN, N., JOHNSON, N.L., (2000). Continuous multivariate distribution, Vol. 1: Models and applications, 2nd ed. John Wiley & Sons Inc.

LIN, P., (1972). Some characterizations of the multivariate *t* distribution, Journal of Multivariate Analysis, 2, pp. 339–344.

LOPERFIDO, N., (2014). A probability inequality related to Mardia's kurtosis. In: C. Perna, M. Sibillo (eds.). Mathematical and statistical methods of actuarial science and finance, Springer: Springer International Publishing Switzerland 201. pp. 129–132.

MARDIA, K.V., (1970). Measures of multivariate skewness and kurtosis with applications, Biometrika, 57 (3), pp. 519–530.

MARSHALL, A., OLKIN, I., (1960). Multivariate Chebyshev inequalities, The Annals of Mathematical Statistics, 31, pp. 1001–1014.

NAVARRO, J., (2014). Can the bounds in the multivariate Chebyshev inequality be attained? Statistics and Probability Letters, 91, pp. 1–5.

NAVARRO, J., (2016). Avery simple proof of the multivariate Chebyshev's inequality. Communications in Statistics – Theory and Methods, 45 (12), pp. 3458–3463.

OLKIN, I., PRATT, J.W., (1958). A multivariate Tchebycheff inequality. The Annals of Mathematical Statistics, 29, pp. 226–234.

OSIEWALSKI, J., TATAR, J., (1999). Multivariate Chebyshev inequality based on a new definition of moments of a random vector, Przegląd Statystyczny (Stat. Rev.), 2, pp. 257–260.

PEARSON, K., (1919). On generalised Tchebycheff theorems in the mathematical theory of statistics, Biometrika,12(3–4), pp. 284–296.

STÖCKL, S., HANKE, M., (2014). Financial applications of the Mahalanobis distance, Applied Economics and Finance, 1 (2), pp. 78–84.

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. 171–186, DOI 10.21307/stattrans-2019-030
Submitted – 10.01.2019; Paper ready for publication – 11.04.2019

DECOMPOSITION OF GENDER WAGE GAP IN POLAND USING COUNTERFACTUAL DISTRIBUTION WITH SAMPLE SELECTION

Joanna Małgorzata Landmesser[1]

ABSTRACT

In the paper, we compare income distributions in Poland taking into account gender differences. The gender pay gap can only be partially explained by differences in men's and women's characteristics. The unexplained part of the gap is usually attributed to the wage discrimination. The objective of this study is to extend the Oaxaca-Blinder decomposition procedure to different quantile points along the income distribution. We use the RIF-regression method to describe differences between the incomes of men and women along the two distributions and to evaluate the strength of the influence of personal characteristics on the various parts of the income distributions. As the sample selection is a serious issue for the study, therefore our decomposition approach will be adjusted for sample selection problems. The results suggest not only differences in the income gap along the income distribution (in particular sticky floor and glass ceiling), but also differences in the contribution of selection effects to the pay gap at different quantiles. The analysis is based on data from the EU-SILC data for Poland in 2014.

Key words: gender wage gap, sample selection, decomposition of income inequalities.

1. Introduction

Gender differences in pay are a well-known phenomenon of labour markets. Also, researchers investigating the wage gap in Poland have found significant gender differences. Goraus, Tyrowicz and van der Velde (2017) report that the raw wage gap is around 10% and the adjusted pay gap estimates oscillate between 14% and 24% depending on the method of calculation. All studies indicate that women are paid only a part of what men with similar demographic characteristic, family situations, working hours, educational levels and work experience are paid. Goraus and Tyrowicz (2014) studied the gender wage gap in Poland using the Labour Force Survey data covering the time span of 1995–2012. The raw gap was highest in the first and last five years of the analyzed

[1] Warsaw University of Life Sciences – SGGW. E-mail: joanna_landmesser@sggw.pl. ORCID ID: http://orcid.org/0000-0001-7286-8536.

period and amounted to around 15% of average females' wages. In the year 1999 the gap started decreasing and reached the level of around 2% in the years 2003 and 2004. Then the gap was increasing until it reached its previous level. The lowest levels of wage gap were observed after economic downturn in Poland. The adjusted gender wage gap was twice as high as the raw gender wage gap. The authors suggested that the adjusted gender wage gap has cyclical properties and it conforms to the bahaviour of unit labour costs - the more lax the labour market conditions, the higher the chances for women to be less unequally compensated.

According to the annual Global Gender Gap Report from the World Economic Forum (2018) women around the world earn about 20-30 percent less on average than their male counterparts. The inequalities between the sexes had closed by only a small amount in the past years. The largest gaps exist in politics, healthcare and education. A meta-analysis by Weichselbaumer and Winter-Ebmer (2005) of more than 260 published pay gap studies for over 60 countries found that, from the 1960s to the 1990s, raw wage differentials worldwide fell substantially from around 65% to 30%. The bulk of this decline was due to better labour market endowments of women (i.e. better education, training, and work attachment). In the period from 2010 to 2015, the gap decreased in only 10 of the 30 European countries (World Bank calculations using EU-SILC surveys for 30 countries in Europe). The most notable decreases were in Estonia, the Slovak Republic, and Switzerland. For others, the gap increased, particularly in Poland, Bulgaria, Lithuania, and France.

Numerous foreign publications show that such a factor as self-selection into the labour force is also crucial for the gender pay gap (e.g. Albrecht, van Vuuren and Vroman, 2009; Töpfer, 2017). If all women participated in the labour force, the observed gap would have a different size. Previous researchers in Poland focused on decomposing the gender pay gap at the means of the wage distribution using a procedure developed by Oaxaca (1973) and Blinder (1973). Using this method the gender wage gap could only be partially explained by differences in men's and women's characteristics (e.g. Słoczyński (2012), Śliwicki and Ryczkowski (2014)). Later, attention shifted to investigating the degree to which gender pay gaps might vary across the wage distribution (Rokicka and Ruzik (2010), Landmesser, Karpio and Łukasiewicz (2015), Magda, Tyrowicz and van der Velde(2015), Landmesser (2016)). Various techniques for the decomposition of differences in income distributions were considered but they lack women self-selection.

The objective of this study is to extend the Oaxaca-Blinder decomposition procedure to different quantile points along the income distribution. The employment rates in Poland differ by gender and the sample selection is a serious issue for the study. Therefore, our decomposition approach will be adjusted for sample selection problems.

We focus our attention on people's decisions regarding full-time or part-time employment, which seem to be non-random. According the Polish Central Statistical Office (2016), the share of part-time employees among the total working population for men is equal to 4.7 and for women to 10.7. Part-time employment in Poland is therefore more concentrated among women. The main reasons given by people for part-time work are: a person prefers this kind of work

(the answer given by 46.2% of men and 42.3% of women who worked part-time), a person is unable to find full-time work (25.1% of men and 27.3% of women), care for children and other people, other personal or family reasons (only 3.3% of men and 13.1% of women), education, illness or disability (Central Statistical Office, 2016). However, our empirical investigation is based on data from the European Union Statistics on Income and Living Conditions project for Poland, which made it impossible to take into account all the factors mentioned above.

To decompose the differences between two distributions one uses the so-called counterfactual distribution, which is a mixture of a conditional distribution of the dependent variable and a distribution of the explanatory variables. Such a counterfactual distribution can be constructed in various ways (DiNardo, Fortin and Lemieux (1996), Donald, Green and Paarsch (2000), Machado and Mata (2005), Fortin, Lemieux and Firpo (2010)). We will examine the differences in the entire range of income values by the use of the *RIF*-regression method (recentered influence function) (Firpo, Fortin and Lemieux (2009)) corrected in a way that allows us to include non-random selection into the sample. It will also be found how the men's and women's characteristics (the explanatory variables in estimated models) influence various ranges of income distributions.

2. Methods of the analysis

Let Y_g denote the outcome variable in group g (e.g. the personal income in men's group, $g=M$, or in women's group, $g=W$) and X_g the vector of individual characteristics of the person in group g (e.g. age, education level, number of years spent in paid work). The expected value of y conditionally on X is a linear function $y_g = X_g\beta_g + v_g, g = M,W$, where β_g are the returns to the characteristics. The Oaxaca-Blinder decomposition for the average income inequality between two groups at the aggregate level can be expressed as

$$\hat{\Delta}^{\mu} = \overline{Y}_M - \overline{Y}_W = \overline{X}_M\hat{\beta}_M - \overline{X}_W\hat{\beta}_W = \underbrace{(\overline{X}_M - \overline{X}_W)\hat{\beta}_M}_{\hat{\Delta}^{\mu}_{\text{explained}}} + \underbrace{\overline{X}_W(\hat{\beta}_M - \hat{\beta}_W)}_{\hat{\Delta}^{\mu}_{\text{unexplained}}}. \quad (1)$$

The first component, on the right side of the equation, gives the effect of characteristics and expresses the difference of the potentials of people in two groups (the so-called explained effect). The second term called the unexplained effect, is the result of differences in the returns to observables. This is the result of differences in the estimated parameters, and so in the "prices" of individual characteristics of group representatives. It can be interpreted as the labour market discrimination. Also, the detailed decomposition may be calculated. A drawback of the approach is that it focuses only on average effects, which may lead to a misleading assessment if the effects of covariates vary across the wage distribution.

In our study people participate in the labour market full-time or part-time. Those who work full-time have higher incomes and those who work part-time have lower incomes. The decision on working time is subjective and non-random. Therefore, our decomposition approach will be adjusted for sample selection.

Most of the literature on gender wage gaps, using decomposition methods such as the Blinder-Oaxaca decomposition, typically ignores selection bias. But some of the studies on this topic have tried to control for self-selection of individuals in the labour market, e.g. by using two-step Heckman procedure (Heckman, 1976; 1979). In this case the correction for sample-selection bias will be included in the wage equation whose initial form is as follows:

$$y_{1i} = X_{1i}\beta + \varepsilon_{1i}\,. \tag{2}$$

The dependent variable y_1, in our case a wage of a full-time worker, is not always observed. The dependent variable is rather observed if a person works full-time ($y_{2i} = 1$).

Therefore, in the first stage of the procedure, we formulate a model for the probability of working full-time (the selection equation). The specification for this relationship is a probit regression of the form

$$y_{2i}^{*} = X_{2i}\gamma + \varepsilon_{2i}\,, \tag{3}$$

where $y_{2i} = 0$ for $y_{2i}^{*} \le 0$ and $y_{2i} = 1$ for $y_{2i}^{*} > 0$. The estimation of the model (3) yields results that can be used to predict the full-time employment probability for each individual $P(Y_{2i} = 1 \mid X_{2i}) = \Phi(X_{2i}\gamma)$.

In the second stage, we correct for self-selection by incorporating a transformation of these predicted individual probabilities as an additional explanatory variable in wage equation (2). The conditional expectation of wages given the person works full-time under the assumption that the error terms in (2) and (3) are jointly normal is then:

$$E(y_{1i} \mid y_{2i} = 1) = X_{1i}\beta + \sigma_{12}\lambda(X_{2i}\gamma)\,, \tag{4}$$

where $\lambda(.)$ is the inverse Mills ratio. In the Heckman two-step procedure, an exclusion restriction is required to generate proper estimates: there must be at least one variable which appears in the selection equation but does not appear in the equation of interest. If no such variable is available, it may be difficult to correct for sampling selectivity.

In our study we consider that the selection into the full-time employment occurs for both groups (for men and for women). Thus, we correct for overall selection (of both groups) and then apply Oaxaca-Blinder decomposition method to the overall estimation.

Let us return to the main goal of our work. The objective of this study is to extend the Oaxaca-Blinder decomposition procedure to different quantile points along the income distribution taking into account the problem of sample selection. Let $F_{Y_g}(y)$ be the distribution function for the variable *Y* in group *g*, which can be expressed using the conditional distribution $F_{Y_g \mid X, D_g}(y \mid X = x)$ of *Y* and the joint distribution $F_{X \mid D_g}(X)$ of all elements of *X* ($D_g = 1$ if *g*=*M*; $D_g = 0$ if *g*=*W*):

$$F_{Y_g \mid D_g}(y) = \int F_{Y_g \mid X, D_g}(y \mid X = x) \cdot F_{X \mid D_g}(X)\, dx, \qquad g = M, W\,. \tag{5}$$

We can extend the mean decomposition analysis to the case of differences between the two distributions using the counterfactual distribution $F_{Y_W^C}(y) = \int F_{Y_W|X_W}(y|X) \cdot dF_{X_M}(X)$ (distribution of incomes that would prevail for people in group *W* if they had the distribution of characteristics of group *M*):

$$F_{Y_M}(y) - F_{Y_W}(y) = \underbrace{[F_{Y_M}(y) - F_{Y_W^C}(y)]}_{\substack{\hat{\Delta}^{\mu}_{\text{unexplained}} \\ \text{(structure effect)}}} + \underbrace{[F_{Y_W^C}(y) - F_{Y_W}(y)]}_{\substack{\hat{\Delta}^{\mu}_{\text{explained}} \\ \text{(composition effect)}}}. \quad (6)$$

The counterfactual distribution can be constructed using the *RIF*-regression method by Firpo, Fortin and Lemieux (2009). This method is similar to a linear regression, except that the variable *Y* is replaced by the recentered influence function of the statistic of interest. The recentered influence function is defined as:

$$RIF(y, Q_\tau) = Q_\tau + IF(y, Q_\tau) = Q_\tau + \frac{\tau - \mathrm{I}\{y \le Q_\tau\}}{f_Y(Q_\tau)}, \quad (7)$$

where $IF(y, Q_\tau)$ is the influence function corresponding to an income *y* for the quantile Q_τ of the distribution F_Y, and $\mathrm{I}\{y \le Q_\tau\}$ is the indicator variable for whether the income *y* is smaller or equal to the quantile Q_τ. Firpo, Fortin and Lemieux (2009) model the conditional expectation of $RIF(y, Q_\tau)$ as a linear function of the explanatory variables $E[RIF(y, Q_\tau|X)] = X\beta_\tau$, where the parameters β_τ can be estimated by OLS ($\hat{\beta}_{g,\tau} = (\sum_{i \in g} X_i^T \cdot X_i)^{-1} \cdot \sum_{i \in g} X_i^T \cdot RIF(y_{g,i}, Q_{g,\tau})$, $g = M, W$). In the approach, we first compute the sample quantile $\hat{Q}_\tau$ and estimate the density $\hat{f}_Y(\hat{Q}_\tau)$ using kernel methods. Then, we estimate the linear probability model for the proportion of people with income less than $\hat{Q}_\tau$, calculate *RIF* of each observation and run regressions of *RIF* on the vector *X*. The aggregated and detailed decomposition for any unconditional quantile is then:

$$\begin{aligned} \hat{\Delta}^{\tau} &= (\overline{X}_M - \overline{X}_W)\hat{\beta}_{W,\tau} + \overline{X}_M(\hat{\beta}_{M,\tau} - \hat{\beta}_{W,\tau}) = \\ &= \sum_{j=1}^{k} ((\overline{X}_{jM} - \overline{X}_{jW})\hat{\beta}_{jW,\tau} + \overline{X}_{jM}(\hat{\beta}_{jM,\tau} - \hat{\beta}_{jW,\tau})) \end{aligned} \quad (8)$$

Now, following Buchinsky (1998) we propose the *RIF*-model corrected for selectivity bias at the τth quantile:

$$RIF(y, Q_\tau) = X_1\beta_\tau + \delta_{1\tau}\lambda(X_2\gamma) + \delta_{2\tau}\lambda(X_2\gamma)^2 + \varepsilon_\tau, \quad (9)$$

where $\lambda(.)$ is the standard inverse Mills ratio (compare with other approaches in Albrecht, van Vuuren and Vroman (2009), Töpfer (2017)). By that means, we will examine the differences in the entire range of income values by the use of the

RIF-regression method corrected in a way that allows us to include non-random selection into the sample. It would be also possible to compare the income distributions of men and women by building a model for people working only full-time, although it seems that this would be an oversimplification of the problem.

3. Data basis

The empirical data used were collected within the European Union Statistics on Income and Living Conditions project for Poland in 2014 (research proposal 234/2016-EU-SILC). The sample consists of 5,181 men and 4,734 women. Each person is described by the following characteristics: *age* (in years), *educlevel* (education level, 1 – primary, . . ., 5 – tertiary), *married* (marital status, 1 – married, 0 – unmarried), *yearswork* (number of years spent in paid work), *permanent* (type of contract, 1 – permanent job/work contract of unlimited duration, 0 – temporary contract of limited duration), *parttime* (1 – person working part-time, 0 – person working full-time), *manager* (managerial position, 1 – supervisory, 0 – non-supervisory). The sample features are presented in Table 1 and Table 2.

Table 1. The selected sample features

Characteristic		Men	Women	Characteristic		Men	Women
No. of obs.		5,181	4,734	Average *age*		42.07	42.36
Average *income*		7,165.94	5,900.21	Average *yearswork*		20.09	18.46
educlevel	= 1	4.91%	3.89%				
	= 2	1.45%	0.55%	*married*	= 1	71.53%	69.60%
	= 3	68.57%	47.32%	*permanent*	= 1	70.60%	71.63%
	= 4	2.55%	7.91%	*parttime*	= 1	4.31%	10.09%
	= 5	22.52%	40.32%	*manager*	= 1	18.68%	15.74%

Source: Own calculations.

Table 2. The number of observations and the average annual net employee incomes of men and women

	Men		Women	
	full-time	part-time	full-time	part-time
Number of observations	4957	224	4257	477
The average annual net employee incomes in thousands of Euros	7,326.65	3,562.56	6,206.51	3,147.30
The average logarithm of the annual income	8.707	7.817	8.564	7.765

Source: Own calculations.

Since there was no information in the EU-SILC database on the hourly income, in our analysis the annual net employee (cash or near cash) incomes of men were compared with those obtained by women. Employee income is defined as the total remuneration payable by an employer to an employee in return for

work done by the latter during one year. The net employee income corresponds to the gross employee income (mainly wages and salaries paid for the time worked or work done in the main and any secondary job(s), remuneration for the time not worked, enhanced rates of pay for overtime, payments for fostering children, supplementary payments (e.g. thirteenth month payment)) but the tax at source, the social insurance contributions are deducted. In our empirical decomposition analysis the logarithm of the annual income (*log_income*) constitutes the outcome variable. Figure 1 contains the kernel density estimates of *income* and *log_income* for men and women.

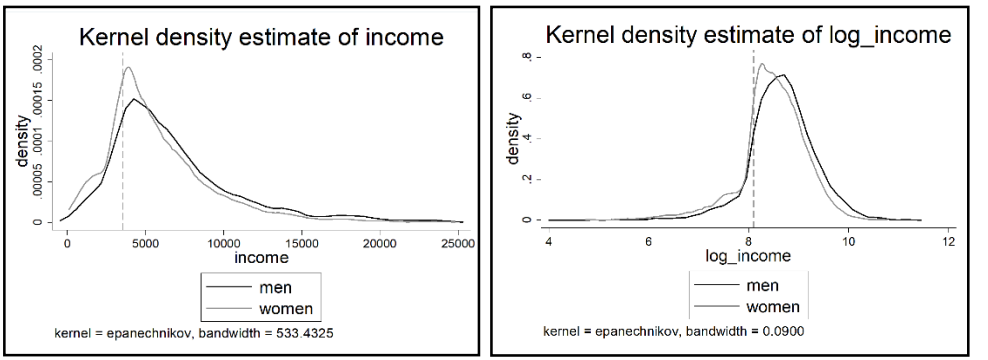


Figure 1. Kernel density estimates of *income* and *log_income* for men and women (the annual income level obtained at the minimum wage is marked with a dotted line)

Source: Own elaboration.

4. Empirical analysis

4.1. Results of Oaxaca-Blinder decomposition for differences in mean log incomes

Table 3 presents the results of the estimation of models for the probability of working full-time for men and women, separately (the selection equations (3) in the form of probit regression).

Table 3. The results of probit models estimation

	Men	Women
age	-0.027 ***	-0.009 ***
married	0.451 ***	0.195 ***
educlevel	-0.002	0.080 ***
permanent	0.878 ***	0.708 ***
cons	2.103 ***	0.815 ***
No. of observations	5181	4734
lnL	-803.91942	-1431.5893

Source: Own calculations.

Then, after estimating the wage equations, we compare the results of aggregate and detailed Oaxaca-Blinder decomposition of inequalities between men's and women's log incomes without and with selection adjustment (Table 4). For this purpose, we have used the following code prepared in Stata:

```
probit fulltime age married educlevel permanent if men==0
predict Xbw if men==0, xb
gen millsw = . if men==0
replace millsw =  normalden(Xbw) / normal(Xbw)  if men==0 & fulltime==1
replace millsw = -normalden(Xbw) / normal(-Xbw) if men==0 & fulltime==0
probit fulltime age married educlevel permanent if men==1
predict Xbm if men==1, xb
gen millsm = . if men==1
replace millsm =  normalden(Xbm) / normal(Xbm)  if men==1 & fulltime==1
replace millsm = -normalden(Xbm) / normal(-Xbm) if men==1 & fulltime==0
gen mills = .
replace mills = millsw if men==0
replace mills = millsm if men==1
oaxaca2 logincome educlevel yearswork manager, by(men) weight(1)
oaxaca2 logincome educlevel yearswork manager mills, by(men) weight(1)
```

The variables which appear in the selection equation but do not appear in the income equation are *age*, *married*, *permanent*. Hence, the identification requirement for the model has been satisfied. Only the variable *kids* would be a better instrument, but it does not appear in the database used.

Table 4. The Oaxaca-Blinder decomposition of the average log income differences without and with selection adjustment

Mean log income men	8.670			
Mean log income women	8.484			
Raw differential	0.186			
	without selection		with selection	
	Aggregate decomposition			
Explained component	-0.071		-0.072	
Unexplained component	0.257		0.258	
% explained	38.13%		38.71%	
% unexplained	-138.13%		-138.71%	
	Detailed decomposition			
	explained	unexplained	explained	unexplained
educlevel	-0.108 ***	-0.166 ***	-0,109 ***	-0,166 ***
yearswork	0.029 ***	-0.143 ***	0,028 ***	-0,136 ***
manager	0.009 ***	0.018 ***	0,008 ***	0,020 ***
mills			0,000	0,000
cons		0.548 ***		0,540 ***
Total	-0.071 ***	0.257 ***	-0,072 ***	0,258 ***

Source: Own elaboration using the Stata command 'oaxaca2'.

The mean predicted log income for men equals 8.670 (annual net income = 5,825.50 Euro), and for women equals 8.484 (annual net income = 4,831.92 Euro) (see Table 4). There is a positive difference between the mean values of log incomes for men or women (the mean log income differential is 0.186). The difference between the mean log income values was decomposed into two components: the first one explaining the contribution of the attributes differences (the explained part) and the second one explaining the contribution of the different values of models coefficients (the unexplained part). The explained is very low and negative, but the unexplained effect is huge and positive, which means that the inequalities examined should be assigned, in the majority, to the coefficients of estimated models (rather than to the differentiation of individual characteristics). Unfortunately, the selection effect is statistically insignificant.

The detailed decomposition, which was also carried out, made it possible to isolate the factors explaining the inequality observed to a different extent. The strong effect of different education levels of men and women can be noticed. The negative value of the adequate component means that the difference of the average log incomes between men and women is mostly reduced by the women's higher education levels. On the other hand, the values of *manager* attribute possessed by men and women increase the inequality in the average log incomes. A different "evaluation" of personal characteristics (the unexplained component) allow the conclusion that women are discriminated against men (but not because of the education levels and years of work).

4.2. Results of the aggregate decomposition along the income distribution using the residual imputation approach

Since the Oaxaca-Blinder technique focuses only on average effects, next we present the decomposition of inequalities along the distribution of log incomes for men and women using the *RIF*-approach without and with selection adjustment. The results of this decomposition are shown in Table 5, where the inequalities are expressed in terms of percentiles. The symbols p10, …, p90 stand for 10th, …, 90th percentile (e.g. the 10th percentile is the log income value below which 10% of the observations may be found). For each of the nine percentiles the total differences between the values of log incomes for men and women were computed. Then these differences are expressed as the sum of the explained and unexplained components.

Table 5. The results of the decomposition using the *RIF*-regression approach

Percentile	total difference	without selection		with selection	
		explained	unexplained	explained	unexplained
p10	0.287	-0.036	0.322	0.004	0.283
p20	0.094	-0.030	0.124	0.014	0.081
p30	0.152	-0.045	0.198	-0.001	0.153
p40	0.147	-0.054	0.201	-0.002	0.149
p50	0.140	-0.060	0.200	-0.010	0.151
p60	0.141	-0.062	0.203	-0.019	0.160
p70	0.135	-0.064	0.199	-0.025	0.160
p80	0.163	-0.078	0.242	-0.043	0.207
p90	0.216	-0.086	0.302	-0.065	0.282

Source: Own elaboration using the Stata command 'rifreg'.

There are positive differences between the values of log incomes for men and women along the whole log income distribution without and with selection adjustment. Going across the rows to compare quantile effects shows that the income differences increase at the bottom and at the top of the log income distribution. Most of the quantile-specific pay gaps are accounted for by how men and women are rewarded, i.e. by the unexplained component. This finding is in conformity with results obtained in other studies on gender differences in pay, for example Blau and Kahn (2016). The unexplained effect (effect of coefficients) is bigger in absolute value and the explained (effect of characteristics) is lower, which indicates the importance of the "labour market value" of men's and women's attributes. The share of the unexplained part is high and the effect of coefficients is positive in the whole range of the income distribution. This is the result of differences in the "market prices" of individual characteristics of men and women (interpreted as the labour market discrimination).

Additionally, Figure 2 contains the differences between the log income distributions for men and women vs. quantile rank. The total effect is U-shaped. The positive values indicate higher log income values for men than for women.

The explained differential is falling as we move toward the top of the income distribution and is higher for the model with selection adjustment. We can see that the effect of characteristics is negative without sample adjustment. For the sample adjustment we note the positive at the bottom and the negative at the top of the distribution effects of characteristics. The positive (negative) values observed mean that the different values of characteristics of men and women increase (decrease) the income inequalities in these income ranges.

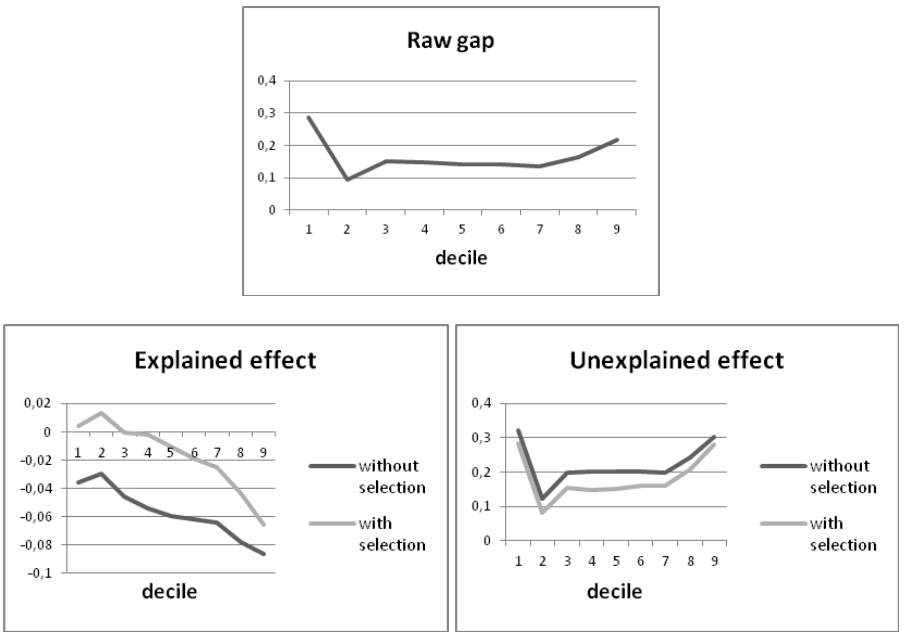


Figure 2. The differences between the log income distributions for men and women calculated using the *RIF*-approach without and with selection adjustment

Source: Own elaboration.

The lower values of the unexplained effect after taking into account the selection imply that without selection correction we overestimate the part attributed to gender-wage discrimination. All in all, the selection component is one of the most important components explaining gender differences in pay along the income distribution.

4.3. Results of the detailed decomposition using the *RIF*-regression approach

The *RIF*-regression method enables us also to extend our analysis to the case of the detailed decomposition. Table 6 shows one of many results obtained of the detailed decomposition of inequalities along log income distributions.

Table 6. The example results of the *RIF*-regression approach – for 70th percentile only

Total gap	0.135 ***		0.135 ***	
	without selection		with selection	
	explained	unexplained	explained	unexplained
educlevel	-0.091 ***	-0.304 ***	-0.089 ***	-0.276 ***
yearswork	0.013 ***	-0.141 ***	0.014 ***	-0.122 ***
manager	0.014 ***	0.021 ***	0.013 ***	0.021 ***
mills			0.000	0.000
mills^2			0.038 ***	-0.015
cons		0.624 ***		0.552 ***
Total	-0.064 ***	0.199 ***	-0.025 **	0.160 ***

Source: Own elaboration using the Stata command 'rifreg'.

These are only the results for 70th percentile of log income distributions. In all, nine detailed decompositions for each decile were carried out (the results for the remaining eight deciles and the bootstrap errors are not presented here due to lack of space).

For better understanding of the results obtained and in order to formulate general conclusions, in Figure 3 we draw the values of explained component for each variable and for each decile group (vs. quantile rank), for the log income inequalities observed between men and women. The ordinate axes present the values $(\overline{X}_{jM} - \overline{X}_{jW})\hat{\beta}_{jM,\tau}$ (detailed explained effects).

The *educlevel* has the greatest reduced influence on the differences between the log income distributions for men and women. It means that on average higher level of education among women decreases the income inequalities, especially as we move toward the top of the income distribution. For the variable *yearswork* we observe the influence which enlarges log income differences but increasingly less as we move toward the top of the income distribution. The variable *manager* also

enlarges log income differences but more and more as we move toward the top of the distribution.

With the increase in incomes the importance of the explained *selection* effect (which is unfavourable for women because of its positive share in the gap) decreases. According to Figure 2, after taking the selection into account, the explained effect is initially positive, which means that the characteristics of the poorest people enlarge the gap. Then, the explained effect is negative, meaning that the characteristics of the richest reduce the gap. Thus, with the increase in incomes, the importance of the people's characteristics increases in the sense that they reduce the gap effect for women.

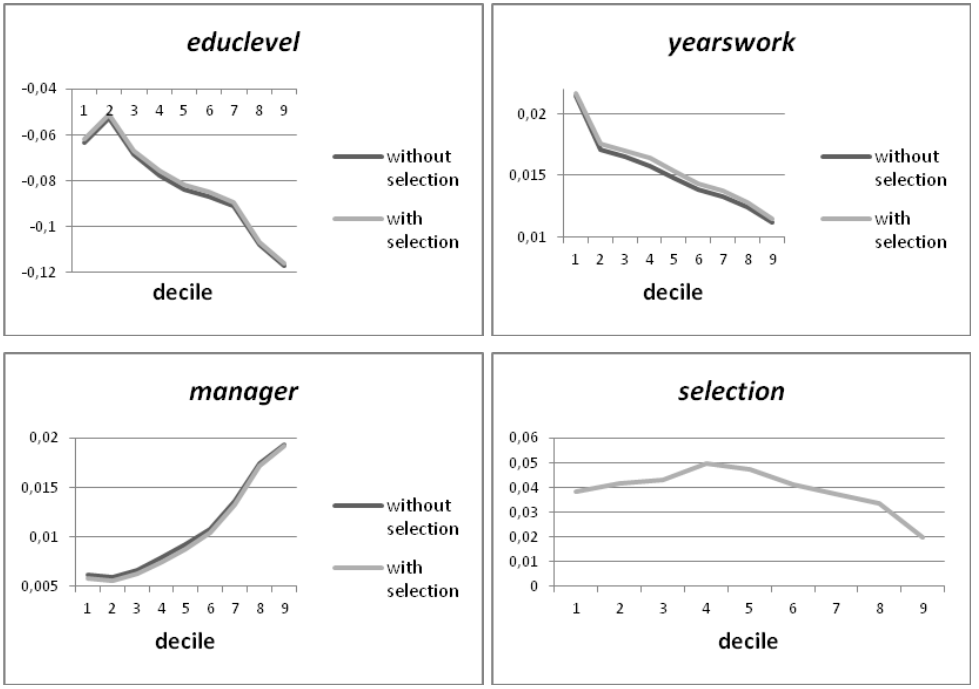


Figure 3. The results of the *RIF*-regression approach for the detailed income inequalities decomposition without and with selection adjustment – the explained effects

Source: Own elaboration.

5. Conclusions

In recent times, particular attention of European politicians has focused on the gender pay gap. There is no clear idea of whether to use the raw or the adjusted gender pay gap for the analysis. Some experts claim that the explained part of the gap may reflect discriminatory social norms or discrimination related to education and occupational choice. Therefore, the use of the unadjusted gap but in tandem

with the employment rates among women should be preferred to analyze the problem.

According to European Union politicians, active policies to close wage gaps are required. The European Commission recommendations point to several possible directions, from pay transparency to improving legal frameworks. The progress on gender pay equality depends on developing specific gender equality policies. However, postulated policies for flexible working can lead to the widening of the gender pay gap if they result, e.g. in an increase in part-time work. A good example is Sweden, which has been at the forefront of gender equity efforts for decades. The universal child care, the protection of mothers' right to work or tough employment laws on pay equity were practiced there. However, the last OECD report reveals the persistence of significant gender earnings gaps in Sweden. The way to close this gap would be if women worked longer hours or if men worked fewer hours. But the majority of Swedish women prefer to work part-time and less than men for a variety of reasons and the gender pay gap is largely a result of the freely-made women's choices. Therefore, maybe we should not consider gaps that arise in these ways to be problems in need of fixing.

Like other researchers, we expected that such a factor as self-selection into the labour force may affect the size of the gender pay gap. Therefore, the goal of this paper was to present the decomposition of inequalities between log incomes for men and women in Poland, taking into account sample selection issues. We started with the decomposition of the average values for log incomes by using the Oaxaca-Blinder method. We found that there is a positive difference between the mean income values for men and women. The unexplained effect was big, but the explained was low. The decomposition showed the influence of the men's and women's attributes on the average log income differences. The selection effect was statistically insignificant.

Then, we decomposed the inequalities between log incomes along the whole distribution using the *RIF*-regression method. The total effect was U-shaped. The explained effect was low again. We claim that Polish women experience both a 'glass ceiling effect' and also a 'sticky floor effect' because gender differences primarily affect women at the top and the bottom of the distribution.

In our research the method of *RIF*-regression provided a way of showing the detailed decomposition of income inequalities and helped to exhibit the influence of the attributes on the whole log income distribution. The variable *educlevel* exerted the greatest reduced influence on the differences between the income distributions for men and women. Higher average levels of education among women decreased the income inequalities. The importance of *educlevel* characteristic increased with income. For the variable *yearswork* (years spent in paid work) we noted the influence which increases income differences but the effect was weaker as the income grew. A woman with the same number of years of work as a man will be more discriminated in the group of low-income people. We also observed strong impact of managerial position in higher quantiles of income distribution, which indicates a shift of big incomes towards men. Being a manager puts men in a more privileged position, especially when it concerns highly paid executive positions.

Also, self-selection into the labour force is crucial for gender gaps: if all women participated full-time in the labour force, the observed gap would be

different at all quantiles. The results showed that selection effects explain a substantial part of the gender pay gap that would otherwise remain unobserved or be attributed to discrimination. Moreover, the contribution of the selection component to the gender pay gap varied across the wage distribution. With the increase in incomes the importance of the selection effect, which is unfavourable for women because of its positive share in the gap, decreased. The total effect of characteristics was negative without sample adjustment. For the sample adjustment we noted the positive at the bottom and the negative at the top of the distribution effects of characteristics.

The article focused on how non-randomness of the sample leads to biased estimates of the wage equation as well as of the components of the wage gap. The method applied was the parametrically extension of the *RIF*-regression method to account for the sample selection problem. In the future, the author intends to estimate the selection correction terms using semiparametric models (as in Töpfer (2017)). In models for sample selection the distributional assumptions may play an important role, therefore semiparametric methods for binary choice (such as the Ichimura or Klein-Spady estimators), although computationally costly, may be more informative.

REFERENCES

ALBRECHT, J., VAN VUUREN, A., VROMAN, S., (2009). Counterfactual distributions with sample selection adjustments: Econometric theory and an application to the Netherlands. Labour Economics, 16 (4), pp. 383–96.

BLAU, F.D., KAHN, L.M., (2016). The Gender Wage Gap: Extent, Trends, and Explanations, Tech. rep. Cambridge: National Bureau of Economic Research.

BLINDER, A., (1973). Wage Discrimination: Reduced Form and Structural Estimates. Journal of Human Resources, 8, pp. 436–55.

BUCHINSKY, M., (1998). The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. Journal of Applied Econometrics, 13 (1), pp. 1–30.

CENTRAL STATISTICAL OFFICE, (2016). Kobiety i mężczyźni na rynku pracy. Warsaw: Central Statistical Office.

DINARDO, J., FORTIN, N. M., LEMIEUX, T., (1996). Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. Econometrica, 64, pp. 1001–44.

DONALD, S. G., GREEN, D. A., PAARSCH, H. J., (2000). Differences in Wage Distributions between Canada and the United States: An Application of a Flexible Estimator of Distribution Functions in the Presence of Covariates. Review of Economic Studies, 67, pp. 609–33.

FIRPO, S., FORTIN, N. M., LEMIEUX, T., (2009). Unconditional Quantile Regressions. Econometrica, 77 (3), pp. 953–73.

FORTIN, N., LEMIEUX, T., FIRPO, S., (2010). Decomposition methods in economics. NBER Working Paper, 16045.

GORAUS, K., TYROWICZ, J., (2014). Gender Wage Gap in Poland – Can It Be Explained by Differences in Observable Characteristics? University of Warsaw, Faculty of Economic Sciences, Working Papers, 11/2014 (128).

GORAUS, K., TYROWICZ, J., VAN DER VELDE, L., (2017). Which Gender Wage Gap Estimates to Trust? A Comparative Analysis. Review of Income and Wealth, 63 (1), pp. 118–46.

HECKMAN, J., (1976). The Common Structure of Statistical Models of Truncation. Annals of Economic and Social Measurement, 5(4), pp. 475-92.

HECKMAN, J., (1979). Sample Selection Bias as a Specification Error. Econometrica, 47(1), pp. 153–62.

JUHN, CH., MURPHY, K. M., PIERCE, B., (1993). Wage Inequality and the Rise in Returns to Skill. Journal of Political Economy, 101, pp. 410–42.

LANDMESSER, J. M., KARPIO, K., ŁUKASIEWICZ, P., (2015). Decomposition of Differences Between Personal Incomes Distributions in Poland. Quantitative Methods in Economics, XVI (2), pp. 43–52.

LANDMESSER, J., (2016). Decomposition of Differences in Income Distributions Using Quantile Regression. Statistics in Transition - new series, 17 (2), pp. 331-48.

MACHADO, J. F., MATA, J., (2005). Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression. Journal of Applied Econometrics, 20, pp. 445–65.

MAGDA, I., TYROWICZ, J., VAN DER VELDE, L., (2015). Nierówności płacowe kobiet i mężczyzn. Pomiar, trendy, wyjaśnienia. Warszawa: Instytut Badań Systemowych.

OAXACA, R., (1973). Male-Female Wage Differentials in Urban Labor Markets. International Economic Review, 14, pp. 693-709.

ROKICKA, M., RUZIK, A., (2010). The Gender Pay Gap in Informal Employment in Poland. CASE Network Studies and Analyses, 406.

SŁOCZYŃSKI, T., (2012). Wokół międzynarodowego zróżnicowania międzypłciowej luki płacowej. International Journal of Management and Economics, 34, pp. 169–85.

ŚLIWICKI, D., RYCZKOWSKI, M., (2014). Gender Pay Gap in the micro level – case of Poland. Quantitative Methods in Economics, XV(1), pp. 159–73.

TÖPFER, M., (2017). Detailed RIF Decomposition with Selection - The Gender Pay Gap in Italy. Hohenheim Discussion Papers in Business, Economics and Social Sciences, 26-2017.

WEICHSELBAUMER, D., WINTER-EBMER, R., (2005). A meta-analysis on the international gender wage gap. Journal of Economic Surveys, 19 (3), pp. 479–511.

WORLD ECONOMIC FORUM, (2018). Global Gender Gap Report. http://reports.weforum.org/global-gender-gap-report-2018/.

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. 187

CALL FOR PAPERS:
STATISITCAL DATA INTEGRATION (SDI)

Statistical data integration has been a subject of research and practical operations for many years. Interest in the subject has grown markedly in recent years, especially in survey statistics due to the increasing accessibility of various unconventional data sources and their potential to supplement the traditional sample survey data to produce a wide range of statistics.

Statistics in Transition new series (SiTns) will bring out a special issue on statistical data integration featuring papers that address theoretical, methodological and practical issues on the subject. The topic of this special issue is timely and is undoubtedly one of the important transitions taking place in statistics.

We are seeking papers on any aspect of statistical data integration, including, but not limited to, the following topics:

Big Data in survey sampling and official statistics
Longitudinal and panel surveys
Multiple Imputation
Nonprobability sampling
Statistical matching
Microsimulation models
Social networks
Multiframe and multi-mode surveys
Small area estimation
Statistical disclosure limitation
Synthetic data
Synthetic population
Record Linkage and Entity Resolution
Total survey errors

Each paper will go through a peer-review process according to the usual standard of the journal. If you are interested in publishing a paper in this special issue, please send your paper to sit@stat.gov.pl by January 31, 2020. It is our goal to notify the authors about our final decisions on acceptance/rejections by June 30, 2020.

For details on editorial requirements and the submission procedure, please consult our journal's link: **http://sit.stat.gov.pl.**

Partha Lahiri, Guest Editor-in-Chief

Włodzimierz Okrasa, Editor-in-Chief

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. 189–192

ABOUT THE AUTHORS

Abduljaleel Mohammed has been a chief statistician in the Ministry of Electricity, Iraq since 2001. He completed his primary education at Dijlah school, Baghdad in 1986. He received his secondary education at AL- Mansour secondary school, Baghdad from 1986 to 1989. The author obtained his bachelor degree in statistics at University of Al-Mustansiriya in 1998. The author got a master degree in applied and computational statistics from the University Putra Malaysia in December 2017**.**

Adegoke Hammed Mumimi is an MSc graduate of Time Series.

Akintande Olalekan J. is a Tutorial Assistant in Department of Statistics, University of Ibadan, Nigeria. He simultaneously works as a Research Assistant and Senior Collaborator at the University of Ibadan Laboratory for Interdisciplinary Statistical Analysis (UI-LISA), Department of Statistics, and a Volunteer - Data Management Specialist at Centre for Petroleum Energy Economic and Law (CPEEL), University of Ibadan. He is a PhD candidate in statistics with research focus in statistical data mining methods, machine learning algorithm and Bayesian theory. He has mentored several undergraduates' students in the Department and has several merit awards to his name. He is an active member of professional bodies: Royal Statistical Society (RSS), International Statistical Institute (ISI), and American Statistical Association (ASA). He enjoys coding (R & Python), teaching, tutoring, researching, writing, trekking/road walking, preaching and meeting new people.

Budny Katarzyna is an Assistant Professor in the Department of Mathematics at the Cracow University of Economics. Her main areas of interest are: mathematical statistics and probability theory. Her research concentrates on the theory of multivariate distributions, estimation methods and multivariate probabilistic inequalities (especially generalizations of Chebyshev's inequality). She is also interested in applications of statistical methods in finance.

Karimi Mostafa is a researcher and data analyst at ENIO Services, located in Kuala Lumpur, Malaysia. He has a PhD in Statistical Inference and Computation from Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia. His main areas of interest include survival analysis, modeling censored data, regression analysis, robust estimators, and outlier detection. He has more than 10 years of experience in data analysis using R programming language. Currently he is working on a data analysis project related to employment rate in Kuala Lumpur, Malaysia.

Khare Brij Behari is a Professor in the Department of Statistics, Banaras Hindu University (BHU), Varanasi, India and obtained his PhD degree from Banaras Hindu University, Varanasi, India in 1984. He is a life member of the National Academy of Sciences, India and Honorable Member of Research Board of

Advisors, ABI (USA). He has established an active research group in BHU making significant contributions in the field of sampling theory and guided seven PhD candidates. Prof. Khare has published over 95 research papers in international/national journals and conferences and presented over 67 research papers in international/national conferences. He has also published two books/monographs and written five chapters in different recognized books. His area of specialization is sampling theory, inference and population studies.

Komorowska Olga is an Assistant Professor at the Department of Statistics, Faculty of Management, University of Gdansk, Poland. Her research focuses on various aspects of social statistics, mainly quality of life, poverty, social exclusion (especially of people with disability and their families). She is an active member of associations that help people with intellectual disabilities and their caregivers. She is the editor-in-chief of the magazine for people with intellectual disability „Moje Ja Też".

Kozłowski Arkadiusz is an Assistant Professor at the Department of Statistics, Faculty of Management, University of Gdansk, Poland. His main areas of interest include survey sampling, statistical inference from survey data, utilizing auxiliary information in surveys, statistical models for classification and regression. He has experience in conducting sample surveys and analysing survey data for both scientific and commercial purposes. He is a member of the Polish Statistical Association.

Kumar Devendra is an Assistant Professor at the Department of Statistics, School of Physical and Mathematical Sciences, Central University of Haryana, India. His research interests are order statistics, distribution theory, Bayesian inference, statistical inference and data analysis in particular. Dr. Kumar has published over 110 research papers in international/national journals and conferences. He has also published two books/monographs. Dr. Kumar is an active member of many scientific professional bodies.

Landmesser Joanna is an Associate Professor at the Department of Econometrics and Statistics, Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences. She received her PhD degree in Economics in 2002 at the Bundeswehr University Munich in Germany and habilitated in 2014 in Economics ("The use of duration analysis methods for the survey of economic activity of people in Poland") at the Nicolaus Copernicus University in Toruń, Poland. Her research interests focus on evaluation of different policies on the labour market, counterfactual scenarios analysis, comparing the income distributions and decomposition of income inequalities.

Malik Mansoor Rashid is a research scholar at the Department of Statistics, Amity Institute of Applied Sciences, Amity University, Noida, India. His main areas of interest include order statistics, progressive censoring and record values. Currently, he is a working as a senior actuarial analyst with Ernst and Young based out of India. Also, he is pursuing his Actuarial qualification from Institute and Faculty of Actuaries, UK.

Midi Habshah is a professor at the Department of Mathematics, Universiti Putra Malaysia. Having published more than 150 papers in citation-indexed journals,

her research interest's focus on regression diagnostics, robust statistics, quality control, experimental design and application of statistical methods to real life problems. She supervises PhD and Master students with more than 30 have already graduated. Prof Habshah is actively involved in the Malaysia Institute of Statistics and has been appointed as the Vice President since 2014. She has written a book published by John Wiley & Sons, entitled "Robust Nonlinear Regression with Application Using R" co-authored with Dr Hosein Riazoshams, Azad University Iran and Prof Gebrenegus Ghilagaber, Stockholm University, Sweden

Nath Dilip C. is the Vice-Chancellor of Assam University (Central University). He has been a Professor of Statistics in Gauhati University, India. He has been a former Head, department of Statistics and Director, Population Research Centre, Gauhati University. Dr. Nath was nominated as an Affiliate Professor for the period September 1999 – September 2002, in the University of Washington, Seattle, USA. His current research interests are in biostatistics/medical statistics/demography and actuarial statistics. He has successfully guided 29 PhD scholars so far to get their degree in statistics. He has completed 14 research projects. He has successfully collaborated with over 50 national and international scholars and contributed over 190 research papers in reputed national/international journals. He is an active member of twenty scientific/professional bodies.

Ogbonna Ahamuefula E. is a Research Assistant & Research Fellow at the Centre for Econometric and Allied Research (CEAR), University of Ibadan, Nigeria and a PhD candidate in the Department of Statistics, University of Ibadan, Nigeria. His research interests revolves around theoretical and applied Econometrics, with recent focus on Bayesian econometrics, financial time series and macroeconomic modelling. He has several research papers, applying econometric methodologies like Bayesian Model Averaging, Autoregressive Distributed Lag – MIDAS, Volatility Modelling, Fourier Unit Root testing framework, among others. He is also competent in programming with relevant analytical software like SPSS, Eviews, STATA, R, RATS and MATLAB, among others.

Roszka Wojciech is an Assistant Professor at the Department of Statistics, Faculty of Informatics and Electronic Economy, Poznań University of Economics and Business. Simultaneously, he is an employee of Center for Public Policy Studies at the Adam Mickiewicz University in Poznan. His main areas of interest include statistical data integration, especially statistical matching and probabilistic record linkage, and small area estimation with special emphasis on spatial microsimulation techniques. As part of the work in the Center, he deals with researching scientific productivity.

Sinha Raghaw Raman is an Assistant Professor in the Department of Mathematics, Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, India and obtained his PhD degree in "Sampling Techniques" from the Department of Statistics, Banaras Hindu University, Varanasi, India in 2001. He has guided one PhD and three MPhil candidates. He is a life member of the Indian Statistical Association and the International Indian Statistical Association.

Dr. Sinha has published over 21 research papers in international/national journals and conferences and presented over 22 research papers in international/national conferences. His area of specialization is sampling theory, data analysis and inference. ORCID identifier number of Dr. R. R. Sinha is 0000-0001-6386-1973.

Słaby Teresa is a researcher of social phenomena using multivariate statistical analysis. For many years (until 2018) she was an academic teacher at the Warsaw School of Economics. Currently she is a Dean of the Faculty of Management and Technical Sciences at the Warsaw Management University. Her main areas of interest include conditions, level, quality and dignity of the people's lives; inequalities and social exclusion; consumer 60+; consumer consumption and behaviour, consumptive behaviour of singles. Notable publications: Konsumpcja. Eseje Statystyczne. Difin, 2006; Reakcje polskich konsumentów na skutki kryzysu gospodarczego. red., SGH, 2009; Jak żyć w kryzysie? Zachowania polskich konsumentów. SGH, 2011; Quality of life of the emerging upper class in Poland. SGH, 2012; Zachowania konsumpcyjne singli w Polsce (co-author). SGH, 2018.

Verma Vivek has received his PhD in Statistics from Department of Statistics, Gauhati University, Guwahati, Assam, India, and joined AIIMS Cohort Study, Department of Neurology, All India Institute of Medical Sciences (AIIMS), New Delhi, India as a Research Officer thereafter. His research interests are Bayesian modelling, sampling techniques, statistical inference and bio-medical, demographical and epidemiological data analysis in particular. Dr. Verma has published nine research papers in international journals and conferences. He is involved in various multidisciplinary medical research activities.

Yaya OlaOluwa Simon is an Assistant Professor at the Department of Statistics, University of Ibadan, Nigeria. His research interests are economic, financial and computational time series with papers written in fractional integration, unit roots, volatility modelling and regime switching. He has over 75 research papers to his credit, each published in reputable ISI and scopus journals. He is an active members of professional bodies such as the Royal Statistical Society, International Statistical Institute, Nigerian Statistical Society.

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 italics) 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).