



# STATISTICS IN TRANSITION

*new series*

*An International Journal of the Polish Statistical Association*

## CONTENTS

Editor's note and acknowledgements .....	445
Submission information for authors .....	449

### Sampling methods and estimation

GALEONE C., POLLASTRI A., Confidence intervals for the ratio of two means using the distribution of the quotient of two normals .....	451
SWAIN A. K. P. C., On classes of modified ratio type and regression-cum-ratio type estimators in sample surveys using two auxiliary variables .....	473
KHARE B. B., SRIVASTAVA U., KUMAR K., Chain ratio estimator for the population mean in the presence of non-response .....	495
CHOUDHURY S., SINGH B. K., A class of chain ratio-cum-dual to ratio type estimator with two auxiliary characters under double sampling in sample surveys ..	519
YADAV R, UPADHYAYA L. N., SINGH H. P., CHATTERJEE S., Almost unbiased ratio and product type exponential estimators .....	537
MEHTA (RANKA) N., MANDOWARA V. L., A better estimator of population mean with power transformation based on ranked set sampling .....	551

### Congress of Polish Statistics:

#### The 100<sup>th</sup> Anniversary of the Polish Statistical Association

GÓRECKI T., KRZYŚKO M., A kernel version of functional principal component analysis .....	559
MIKULEC A., KUPIS-FIJAŁKOWSKA A., An empirical analysis of the effectiveness of Wishart and Mojena criteria in cluster analysis .....	569
WILK J., Symbolic approach in regional analyses .....	581
The Congress of Polish Statistics to mark the 100 <sup>th</sup> anniversary of the Polish Statistical Association, Poznań, 18–20 April 2012 (Gołata E.) .....	601
Report on survey sampling and small area statistics sessions during the Congress of Polish Statistics in Poznań, 18–20 April 2012 (Wywił J. L., Żądło T.) .....	607

### Conference and other reports

The XXXI International Conference on Multivariate Statistical Analysis, 12–14 November 2012, Łódź, Poland (Mikulec A., Kupis-Fijałkowska A.) .....	611
KRZYŚKO M., The History of the Mathematical Statistics Group at the Horticultural Faculty of the Central College of Agriculture in Warsaw, and the Biometric Laboratory at the Marcei Nencki Institute of the Warsaw Scientific Society .....	617

## EDITOR IN CHIEF

Prof. W. Okrasa, *University of Cardinal Stefan Wyszyński, Warsaw, and CSO of Poland*  
w.okrasa@stat.gov.pl; Phone number 00 48 22 — 608 30 66

---

## ASSOCIATE EDITORS

Sir Anthony B. *University of Oxford,*  
Atkinson *UK*

M. Belkindas, *The World Bank,*  
*Washington D.C., USA*

Z. Bochniarz, *University of Minnesota, USA*

A. Ferligoj, *University of Ljubljana,*  
*Ljubljana, Slovenia*

M. Ghosh, *University of Florida, USA*

Y. Ivanov, *Statistical Committee of the*  
*Common-wealth of Independent*  
*States, Moscow, Russia*

K. Jajuga, *Wroclaw University of*  
*Economics,*  
*Wroclaw, Poland*

G. Kalton, *WESTAT, Inc., USA*

M. Kotzeva, *Statistical Institute of Bulgaria*

M. Kozak, *University of Information*  
*Technology and Management in*  
*Rzeszów, Poland*

D.Krapavickaite, *Institute of Mathematics and*  
*Informatics,*  
*Vilnius, Lithuania*

M. Krzyżsko, *Adam Mickiewicz University,*  
*Poznań, Poland*

J. Lapins, *Statistics Department,*  
*Bank of Latvia, Riga, Latvia*

R. Lehtonen, *University of Helsinki, Finland*  
A. Lemmi, *Siena University,*  
*Siena, Italy*

A. Młodak, *Statistical Office Poznań, Poland*  
C.A. O'Muircheartaigh, *University of Chicago,*  
*Chicago, USA*

V. Pacakova, *University of Economics,*  
*Bratislava, Slovak Republic*  
R. Platek, *(Formerly) Statistics Canada,*  
*Ottawa, Canada*

P. Pukli, *Central Statistical Office,*  
*Budapest, Hungary*  
S.J.M. de Ree, *Central Bureau of Statistics,*  
*Voorburg, Netherlands*

I. Traat, *University of Tartu, Estonia*  
V. Verma, *Siena University,*  
*Siena, Italy*

V. Voineagu, *National Commission for Statistics,*  
*Bucharest, Romania*

J. Wesołowski, *Central Statistical Office of Poland,*  
*and Warsaw University of*  
*Technology,*  
*Warsaw, Poland*

G. Wunsch, *Université Catholique de Louvain,*  
*Louvain-la-Neuve, Belgium*

J. Wywił, *University of Economics in*  
*Katowice, Poland*

---

## FOUNDER/FORMER EDITOR

Prof. J. Kordos, *Formerly Central Statistical Office, Poland*

## EDITORIAL BOARD

Prof. Janusz Witkowski (Chairman), *Central Statistical Office, Poland*

Prof. Jan Paradysz (Vice-Chairman), *Poznań University of Economics*

Prof. Czesław Domański, *University of Łódź*

Prof. Walenty Ostasiewicz, *Wroclaw University of Economics*

Prof. Tomasz Panek, *Warsaw School of Economics*

Prof. Mirosław Szreder, *University of Gdańsk*

Władysław Wiesław Łagodziński, *Polish Statistical Association*

## Editorial Office

Marek Cierpiał-Wolan, Ph.D.: Scientific Secretary

m.wolan@stat.gov.pl

Beata Witek: Secretary

b.witek@stat.gov.pl. Phone number 00 48 22 — 608 33 66

Rajmund Litkowiec: Technical Assistant

## Address for correspondence

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

ISSN 1234-7655

## EDITOR'S NOTE AND ACKNOWLEDGEMENTS

The contents of this issue, which brings to a close the previous year's series of the Journal, continues to reflect the fact that due to celebrating the 100th Anniversary of the Polish Statistical Association (PSA) 2012 year was a special time in Polish and international statistics – at least as far as it was marked by also participation of many distinguished guests from abroad in several meetings of statisticians being held for this occasion.

The first and main part embraces, however, articles traditionally dominating in the journal's mainstream, i.e., devoted to **sampling methods and estimation**. It starts with **C. Galeone** and **A. Pollastri's** paper *Confidence Intervals for the Ratio of Two Means Using the Distribution of the Quotient of Two Normals* which concentrates on the estimator of the ratio of two means defined as a ratio of two random variables normally or asymptotically normally distributed. The importance of the problem emerges from the fact that, according to the authors, generally the approximation to Normal is not satisfied. They propose a new method for building confidence intervals for the ratio of two means which, in contrast to other parametric methods, is worthy to be preferred because it takes into account the skewness in the distribution of the ratio estimator, and the confidence intervals are always bounded. In the next paper, *On Classes of Modified Ratio Type and Regression-Cum-Ratio Type Estimators in Sample Surveys Using Two Auxiliary Variables*, **A. K. P. C. Swain** presents a generalized classes of such a type of estimators of the finite population mean in the presence of two auxiliary variables in simple random sampling (given that the population means of the auxiliary variables are known in advance). Certain aspects of the generalized estimators are compared – their biases and efficiencies – both theoretically and in reference to some natural populations.

Another class of estimators that utilize two auxiliary variables is proposed by **B. B. Khare**, **U. Srivastava** and **K. Kumar** in the paper *Chain Ratio Estimator for the Population Mean in the Presence of Non-response*. Authors found the proposed estimators to be more efficient than the relevant estimators for the fixed values of preliminary sample of size  $n'$  and subsample of size  $n(< n')$  taken from the preliminary sample of size  $n'$  under the specified conditions. The proposed estimators are both more efficient than the corresponding estimators in the case of the fixed cost and have less total cost in comparison to the cost incurred by the corresponding relevant estimators for specified variance. These results are supported by empirical study and Monte Carlo simulation. Two auxiliary variates under double (two-phase) sampling procedure, when the information on another additional auxiliary variate is available along with the main auxiliary variate, are

also used by **S. Choudhury** and **B. K. Singh** in procedures presented in the next paper *A Class of Chain Ratio-Cum-Dual to Ratio Type Estimator with Two Auxiliary Characters under Double Sampling in Sample Surveys*. The asymptotically optimum estimators (AOEs) in the class are identified in two different cases with their biases and variances. The optimum values of the first phase and second phase sample sizes have been obtained for the fixed cost of survey. Theoretical and empirical studies have also been done to demonstrate the efficiency of the proposed estimator with respect to strategies which utilized the information on two auxiliary variates. Also **R. Yadav**, **L. N. Upadhyaya**, **H. P. Singh**, and **S. Chatterjee** look for *Almost Unbiased Ratio and Product Type Exponential Estimators* using information on auxiliary variate. Authors suggest a generalized version of Bahl and Tuteja (1991) estimator and examine its properties to find that asymptotic optimum estimator (AOE) in the proposed generalized version of Bahl and Tuteja (1991) estimator is biased. Since, at least in some applications, biasedness of an estimator is disadvantageous, they apply the Singh and Singh's procedure (1993) to derive an almost unbiased version of AOE. A numerical illustration is given in the support of the study's result.

*A Better Estimator of Population Mean with Power Transformation Based on Ranked Set Sampling* is being sought by **N. Mehta (Ranka)** and **V. L. Mandowara** (in the paper entitled this way) to have increased the efficiency of estimate of the population mean. Authors stress that this method is highly beneficial to the estimation based on simple random sampling (SRS), and they present a modified ratio estimator using prior value of coefficient of kurtosis of an auxiliary variable  $x$ , with the intention to improve the efficiency of ratio estimator in ranked set sampling.

Each of the papers included in the following set of five articles is either directly (as a congress paper) or indirectly (as a post-congress report) associated with the above mentioned congress of the PSA. The first group is opened by *A Kernel Version of Functional Principal Component Analysis*, by **Tomasz Górecki** and **Mirosław Krzyśko**, in which a new construction of functional principal components (FPCA) is proposed based on principal components for vector data. A kernel version of FPCA is also presented and the quality of the two described methods is demonstrated for 20 different data sets.

**Artur Mikulec** and **Aleksandra Kupis-Fijalkowska** in *An Empirical Analysis of The Effectiveness of Wishart and Mojena Criteria in Cluster Analysis* discuss the issue of selecting the optimal grouping result in agglomerative cluster analysis, focusing on comparison of the Mojena criteria (the upper tail rule and moving average quality control rule) and the Wishart criteria (tree validation) with other methods, including those proposed by: Baker and Hubert, Calinski and Harabasz, Davies and Bouldin, Hubert and Levine. The results of empirical analysis (that was carried out in *ClustanGraphics 8* Program and selected packages in R environment for the generated data sets) are presented

in the paper. Cluster analysis along with symbolic data are discussed in the article ***Symbolic Approach in Regional Analyses*** by **Justyna Wilk**, who addresses some problems with conducting regional research associated with the need to consider such difficulties as large data sets, insufficient precision of phenomena description, disregarding territorial diversification of a given phenomenon, as well as incomplete description of problems. Author suggests solutions to these problems by presenting phenomena in the form of symbolic data. After discussing specific nature of symbolic data, methods for collecting symbolic data and methods for these data analysis, the second part presents an empirical example referring to the assessment of labour market situation in Polish regions (NTS-2) using symbolic data and cluster analysis.

Two papers refer to the congress itself. The first, by **Elżbieta Golata** (the congress co-organizer), presents a brief report on ***The Congress of Polish Statistics to Mark the 100th Anniversary of the Polish Statistical Association Poznań, 18–20 April 2012***. Other meetings that took place during the year-long celebration – including seminars and conferences organized by regional sections of the Polish Statistical Association, such as those in Wrocław, Lublin, Toruń and Bydgoszcz – are also mentioned. Some illustration of the congress' sessions is given in ***Report on Survey Sampling and Small Area Statistics Sessions during the Congress of Polish Statistics in Poznań*** in which **Janusz Wywiał** and **Tomasz Żądło** provide information on four sessions on both survey sampling and small area estimation that were organized during the Congress. Altogether fourteen papers in English were presented, including five invited lectures given by **Ray Chambers** from Wollongong University, **Jean-Claude Deville**, **Lorenzo Fattorini** from University of Siena, **Malay Ghosh** from University of Florida and **Li-Chun Zhang** from Statistics Norway.

Two different types of information – one by **Artur Mikulec** and **Aleksandra Kupis-Fijałkowska** about the XXXI Conference on Multivariate Statistical Analysis (that took place on November 12–14, 2012 in Łódź, Poland), and another by **Mirosław Krzyśko**, in the form of a historical note on the Mathematical Statistics Group at the Nencki Institute in Warsaw – conclude this issue.

**Włodzimierz Okrasa**

Editor

## ACKNOWLEDGEMENTS TO REVIEWERS

- Andrew Clark**, Paris School of Economics, France  
**Jacek Bialek**, University of Łódź, Poland  
**Czesław Domański**, University of Łódź, Poland  
**Wojciech Gamrot**, University of Economics in Katowice, Poland  
**Elżbieta Gołata**, University of Economics, Poznań, Poland  
**Henryk Gurgul**, AGH University of Science and Technology, Kraków, Poland  
**Cem Kadilar**, Hacettepe University of Ankara, Turkey  
**Jan Kordos**, Warsaw Management Academy, Warsaw School of Economics, Poland  
**Barbara Kowalczyk**, Warsaw School of Economics, Poland  
**Marcin Kozak**, University of Information Technology and Management in Rzeszów, Poland  
**Mirosław Krzyśko**, Adam Mickiewicz University, Poznań, Poland  
**Sunil Kumar**, Indian Statistical Institute, Kolkatta, India  
**Janis Lapins**, Statistics Department, Bank of Latvia, Riga, Latvia  
**Andrzej Młodak**, Statistical Office Poznań, Poland  
**Jerzy Mika**, University of Economics in Katowice, Poland  
**Jerzy Moczko**, Poznań University of Medical Sciences, Poland  
**Natalia Nehrebecka**, University of Warsaw, Poland  
**Uwe Neumann**, Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI), Essen, Germany  
**Włodzimierz Okrasa**, Central Statistical Office of Poland, and Cardinal Stefan Wyszyński University in Warsaw  
**Aloy Onyeka**, Federal University of Technology, Owerri, Nigeria  
**Walenty Ostasiewicz**, Wrocław University of Economics, Poland  
**Agnieszka Rossa**, University of Łódź, Poland  
**Paweł Strzelecki**, Warsaw School of Economics, and National Bank of Poland  
**Mirosław Szreder**, University of Gdańsk, Poland  
**Junmin Wan**, Peking University, China  
**Jacek Wesolowski**, Central Statistical Office of Poland, and Warsaw University of Technology, Poland  
**Janusz Wywiał**, University of Economics in Katowice, Poland  
**Bohdan Wyżnikiewicz**, Central Statistical Office of Poland  
**Rohini Yadav**, Banaras Hindu University, Varanasi, Uttar Pradesh, India  
**Tomasz Żądło**, University of Economics in Katowice, Poland

STATISTICS IN TRANSITION-new series, December 2012  
Vol. 13, No. 3, pp. 449

## SUBMISSION INFORMATION FOR AUTHORS

*Statistics in Transition – new series (SiT)* is an international journal published jointly by the Polish Statistical Association (PTS) and the Central Statistical Office of 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., followed by a hard copy addressed to  
Prof. Włodzimierz Okrasa,  
GUS / Central Statistical Office  
Al. Niepodległości 208, R. 287, 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://www.stat.gov.pl/pts/15\\_ENG\\_HTML.htm](http://www.stat.gov.pl/pts/15_ENG_HTML.htm)





## CONFIDENCE INTERVALS FOR THE RATIO OF TWO MEANS USING THE DISTRIBUTION OF THE QUOTIENT OF TWO NORMALS

Carlotta Galeone<sup>1</sup>, Angiola Pollastri<sup>2</sup>

### ABSTRACT

In various scientific fields such as medicine, biology and bioassay, several ratio quantities assumed to be Normal, are of potential interest. The estimator of the ratio of two means is a ratio of two random variables normally or asymptotically normally distributed. The present paper shows the importance of considering the real distribution of the estimator of the ratio of two means, because generally the approximation to Normal is not satisfied. The estimated asymptotic cumulative and density function of the estimator of the ratio is presented, with several considerations on the skewness. Finally, a new method for building confidence intervals for the ratio of two means was proposed. In contrast to other parametric methods, this new method is worthy to be preferred because it considers the skewness in the distribution of the ratio estimator, and the confidence intervals are always bounded.

**Key words:** estimator of the ratio of two means, distribution of the ratio of two correlated normals, skewness, confidence intervals for the ratio, Fieller's theorem.

### 1. Introduction

In various scientific fields such as medicine, biology, biometry and bioassay, several ratio quantities assumed to be normal are of potential interest. Pearson (1910) reported the first application of the ratio of normal random variables ( $rv$ ) in medicine with the use of the *opsonic index*. It is a measure of the number of bacteria destroyed by blood cells, expressed as the ratio of opsonin, i.e., a substance that marks foreign bodies in the infected patient's blood to the amount

---

<sup>1</sup> Dipartimento di Medicina del Lavoro “Clinica del Lavoro Luigi Devoto”, Università degli Studi di Milano; Dipartimento di Epidemiologia, Istituto di Ricerche Farmacologiche “Mario Negri”, Milano. E-mail: carlotta.galeone@unimi.it.

<sup>2</sup> Dipartimento di Metodi Quantitativi per le Scienze Economico-Aziendali, Università degli Studi di Milano-Bicocca. E-mail: angiola.pollastri@unimib.it.

found in a healthy person's blood. In biology, an example of a ratio of Normal rvs is the so called *digestibility*, i.e., the ratio of the weight of a component of a plant to that of the whole plant. In bioassay and bioequivalence problems, the relative potency of a new drug to that of a standard drug is often expressed in terms of a ratio. The true potencies of the two drugs are the mean values  $(\mu_1, \mu_2)$  and the relative potency is the ratio of the potency of the new drug with respect to the potency of a standard one, i.e.,  $\mu_1 / \mu_2$ .

The estimator of the ratio of two means is a ratio of two rvs normally or asymptotically normally distributed. Descriptive and inferential aspects of the ratio of two Normal rvs are complex. The first author who specifically considered the problem was Geary (1930), in his paper called "*The frequency distribution of the ratio of two Normal variates*". Subsequently, the distribution of the ratio of two normal random variables was studied by Fieller (1932), Marsaglia (1965, 2006), Frosini (1970) and Aroian and Oksoy (1986) and only few others.

In this paper, we present recent results on descriptive aspects on the distribution of the estimator of the ratio of two means. Finally, in the last paragraph we propose a new method to build confidence intervals based on the distribution of the estimator of the ratio of two means. Differently from other parametric methods, this new method is preferable because it considers the skewness in the distribution of the ratio estimator and the confidence intervals are always bounded.

## 2. The distribution of the ratio of two normals random variables

Let us consider a bivariate correlated normal (bcn) rv  $(X_1, X_2)$  having the following parameters:

$$E(X_1) = \mu_1; E(X_2) = \mu_2; Var(X_1) = \sigma_1^2; Var(X_2) = \sigma_2^2; Corr(X_1, X_2) = \rho.$$

Let us define the rv  $W = \frac{X_1}{X_2}$ , i.e., the ratio of two normal rvs, jointly distributes as a bcn.

The distribution of the ratio of two independent standardized normal rvs is the well-known case of a standard Cauchy rv with median value equal to zero and scale parameter equal to one (Mood, 1974 and Kotz, 1994). The Cauchy distribution has no mean, nor variance or higher moments exist.

The cumulative density function (cdf) of  $W$ , indicated as  $F_W(w)$ , with parameters  $(\mu_1, \mu_2; \sigma_1, \sigma_2; \rho)$ , can be expressed involving the bivariate normal integral tabulated by the National Bureau of Standards (1959), as an extension of

Hinkley's results (1969). Aroian (1986) parameterised the cdf in a more convenient way and showed that the distribution of  $W$  depends on two parameters  $a$  and  $b$  and the variable  $t_w$ , as follows:

$$F_W(w) = L\left(\frac{a - b t_w}{\sqrt{1+t_w^2}}, -b, \frac{t_w}{\sqrt{1+t_w^2}}\right) + L\left(\frac{b t_w - a}{\sqrt{1+t_w^2}}, b, \frac{t_w}{\sqrt{1+t_w^2}}\right) \quad w \in R$$

with

$$a = \sqrt{\frac{1}{1-\rho^2}} \left( \frac{\mu_1}{\sigma_1} - \rho \frac{\mu_2}{\sigma_2} \right) \quad (1)$$

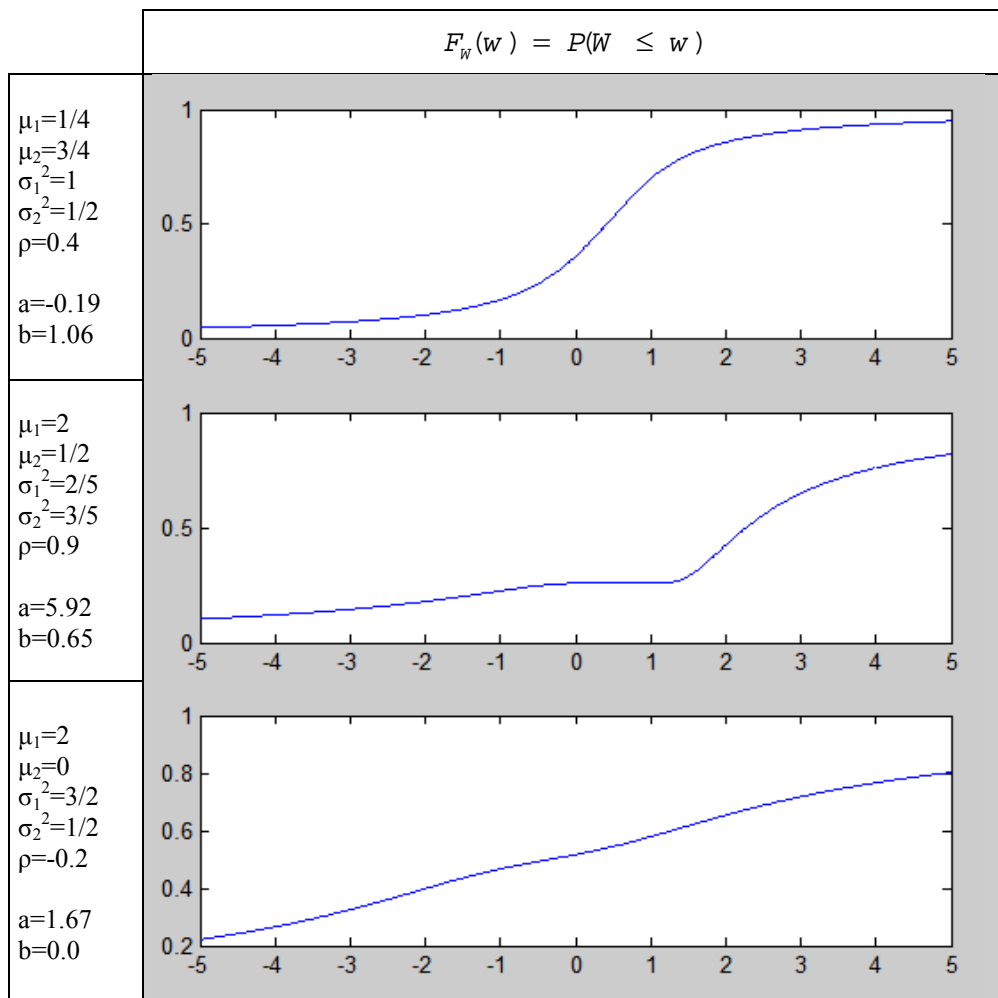
$$b = \left( \frac{\mu_2}{\sigma_2} \right) \quad (2)$$

$$t_w = \sqrt{\frac{1}{1-\rho^2}} \left( \frac{\sigma_2}{\sigma_1} w - \rho \right) \quad (3)$$

and,  $L(h,k,\rho)$  is the bivariate normal integral according to the indication of Kotz et al. (2000), where

$$L(h,k,\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_h^\infty \int_k^\infty \exp\left\{-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)\right\} dx_1 dx_2$$

In Figure 1 the cdf of the ratio of two normal rvs,  $F_W(w)$ , are reported for several selected values of the five parameters  $(\mu_1, \mu_2; \sigma_1, \sigma_2; \rho)$ , and the corresponding  $a$  and  $b$ .

**Figure 1.** Cdf of  $W$  for selected values of the parameters

The probability density function (pdf) of  $W$ , indicated as  $f_W(w)$ , can be expressed as a function of the five parameters  $(\mu_1, \mu_2; \sigma_1, \sigma_2; \rho)$  or, more conveniently, using the previously reported parameterization proposed by Aroian (1986), with  $a$ ,  $b$  and  $t_w$ , as follows:

$$f_W(w) = \frac{\sigma_2}{\sigma_1} \sqrt{\frac{1}{1-\rho^2}} g(t) \quad w \in R$$

where

$$g(t) = \frac{1}{\pi} e^{-\frac{1}{2}(a^2+b^2)} \frac{1}{(1+t^2)} \left\{ 1 + c \int_0^q \varphi(u) du \right\}$$

$$q = \frac{b + at_w}{\sqrt{1+t_w^2}}, \quad c = q \{ \varphi(q) \}^{-1}, \quad \varphi(u) = (2\pi)^{-\frac{1}{2}} e^{-\frac{u^2}{2}}$$

As reported in her PhD thesis (2007), Galeone showed that the pdf of  $W$  can be expressed as a finite non-standard mixture density (Everitt, 1996). In fact, the pdf is decomposable into two component densities ( $c=2$ ) as follows:

$$f_w(w) = p f_1(w) + (1 - p) f_2(w)$$

$$\text{with } 0 \leq p = e^{-\frac{1}{2}(a^2+b^2)} \leq 1$$

where  $a$  and  $b$  were defined in (1) and (2).

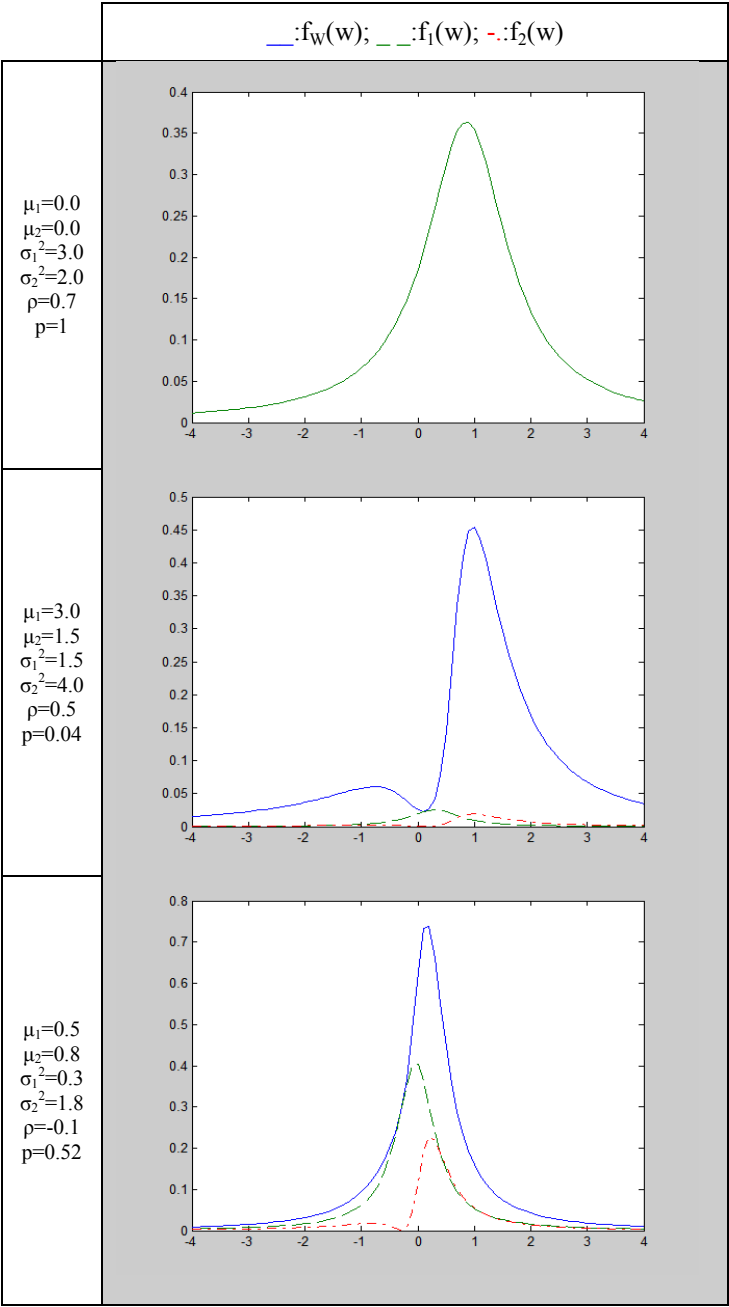
The first pdf,  $f_1$ , is a not central Cauchy rv. The second pdf is a complicate function that depends on  $w$  only through  $q$  and  $t$ , and is not referable to a well-known rv,

$$f_2(w) = \frac{\sigma_2 Q e^{\frac{1}{2}q^2} \int_0^q e^{-\frac{1}{2}s^2} ds}{\pi \sigma_1 \sqrt{1 - \rho^2} (1 + t^2) (e^{\frac{1}{2}(\mu_1^2 + \mu_2^2)} - 1)}$$

The pdf  $f_w$  and the two components of the mixture model  $f_1$  and  $f_2$ , for selected values of the parameters are shown in Figure 2. Generally, the rv  $W$  is not symmetrically distributed, except in the case that the median value is equal to the ratio of the mean values  $\mu_1$  and  $\mu_2$  (Aroian Oksoy, 1986). In some cases, the pdf is bimodal.

The cdf of  $W$  can be computed as a function of  $w$  using Fortran+IMSL library or the scientific software package MATLAB or other libraries, especially those containing routines regarding the cdf of a bcn or other functions which can give the cdf of the bcn. In Appendix, we report the codes of the functions implemented in MATLAB to compute the cdf and the pdf of the ratio of two normal rvs. We have used these functions for all the graphs reported in the article.

**Figure 2.** Pdf  $f_w(w)$  of the ratio of two Normal rvs and of the two components  $f_1(w)$  and  $f_2(w)$  of the mixture model, for selected values of the parameters. The first case is the pdf of central Cauchy random variable, where  $f_2(w)=0$



### 3. The distribution of the estimator of the ratio of two means

Let us suppose we have drawn a simple random sample of  $n$  elements and we have obtained the observations  $(x_{1i}, x_{2i}) (i=1, \dots, n)$ .

If  $(X_1, X_2)$  is a bcn or if  $n$  is large, the rv  $(\bar{X}_1, \bar{X}_2)$  tends to a bcn with the following parameters:

$$E(\bar{X}_1) = \mu_1; E(\bar{X}_2) = \mu_2; \text{Var}(\bar{X}_1) = \frac{\sigma_1^2}{n}; \text{Var}(\bar{X}_2) = \frac{\sigma_2^2}{n}; \text{Corr}(\bar{X}_1, \bar{X}_2) = \rho.$$

Keeping into account the results obtained by Aroian (1986) reported in the previous paragraph, we can obtain the cdf of the rv  $W_n = \frac{\bar{X}_1}{\bar{X}_2}$ , indicated by

$$F_{W_n}(w). \text{ Often } W_n = \frac{\bar{X}_1}{\bar{X}_2} \text{ is used as the estimator of } R = \frac{\mu_1}{\mu_2}.$$

The pdf of  $W_n$  can be expressed as a function of the five parameters  $\left(\mu_1, \mu_2, \frac{\sigma_1^2}{n}, \frac{\sigma_2^2}{n}, \rho\right)$  or, more conveniently, using the previously reported parameterization proposed by Aroian (1986) based on  $a_n$  and  $b_n$  as follows:

$$f_{W_n}(w) = \frac{\sigma_2}{\sigma_1} \sqrt{\frac{1}{1-\rho^2}} g(t_w)$$

where

$$g(t_w) = \frac{1}{\pi} e^{-\frac{1}{2}(a_n^2 + b_n^2)} \frac{1}{(1+t_w^2)} \left\{ 1 + c \int_0^q \varphi(u) du \right\},$$

$$q = \frac{b_n + a_n t_w}{\sqrt{1+t_w^2}}, \quad c = q \{ \varphi(q) \}^{-1},$$

and

$$a_n = \sqrt{\frac{n}{1-\rho^2}} \left( \frac{\mu_1}{\sigma_1} - \rho \frac{\mu_2}{\sigma_2} \right) \quad b_n = \sqrt{n} \left( \frac{\mu_2}{\sigma_2} \right) \quad t_w = \sqrt{\frac{1}{1-\rho^2}} \left( \frac{\sigma_2}{\sigma_1} w - \rho \right)$$

As shown in Figure 2, generally the rv  $W$  is not symmetrically distributed. Mood et al. (1963) defined a rv  $X$  symmetrically distributed about a constant  $C$  if the rv  $(X - C)$  has the same distribution of the rv  $-(X - C)$ . In order to study the distribution of the estimator of  $R$ , we want to verify if the rv  $(W_n - Me)$  is distributed as the rv  $-(W_n - Me)$ , where  $Me$  is the median of the distribution. If this is true, we expect that  $f_{W_n}(Me - w) = f_{W_n}(Me + w)$  (MacGillivray, 1985).

This implies that

$$t_{Me-w} = \sqrt{\frac{1}{1-\rho^2}} \left( \frac{\sigma_2}{\sigma_1} (Me - w) - \rho \right)$$

is equal to

$$t_{Me+w} = \sqrt{\frac{1}{1-\rho^2}} \left( \frac{\sigma_2}{\sigma_1} (Me + w) - \rho \right)$$

Being  $t_{Me-w} \neq t_{Me+w}$ , except in some degenerate case, it is easy to check that  $f_{W_n}(w - Me) \neq f_{W_n}(Me + w)$ .

As a consequence (Frosini, 1971), it is evident that  $F_{W_n}(Me - w) \neq 1 - F_{W_n}(Me + w)$ .

The measures of the asymmetry of a distribution is usually based on the traditional indices based on the moments of the distribution such as the third standardized moment (MacGillivray, 1986). Unfortunately, this approach cannot be used to study the shape of the distribution of the estimator of the ratio because no moment of the rv  $W_n$  exists. For these reasons, we have studied the shape by means of the median and percentiles of the distribution that cannot be obtained analytically but numerical calculations have to be done for each particular case.



First of all, we consider the index proposed by Bowley in 1901 (Brentari, 1990, Groeneveld, 1998)

$$sk(0.25) = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$

where  $Me$  is the median,  $Q_1$  and  $Q_3$  are respectively the first and the third quartile. The index  $sk(0.25)$  varies in the interval  $[-1, 1]$  and so it is easily interpretable.

Then, we use a more analytic index based on the asymmetry of points suggested by David F.N. and Johnson in 1956 (Brentari, 1990); for a continuous rv it is defined as

$$sk(p) = \frac{x(1-p) + x(p) - 2Me}{x(1-p) - x(p)} \quad 0 \leq p < \frac{1}{2}$$

where  $x(p) = F^{-1}(p)$  is the  $p^{th}$  quantile.

The index  $sk(p)$  is the difference of the distance of the  $(1-p)^{th}$  quantile from the median and the distance of the median from the  $p^{th}$  quantile divided by the distance from the  $(1-p)^{th}$  quantile and the  $p^{th}$  quantile. The index  $sk(p)$  is normalized due to  $-1 \leq sk(p) \leq 1$ .

The cdf of  $W_n$ , indicated as  $F_{W_n}(w)$ , may be estimated by substituting in it the maximum likelihood (ML) estimates of  $(\mu_1, \mu_2; \sigma_1, \sigma_2; \rho)$ . The ML estimates of the means  $\mu_1$  and  $\mu_2$  are indicated respectively by  $\bar{x}_1$  and  $\bar{x}_2$ . The ML estimates of  $\sigma_1^2$ ,  $\sigma_2^2$  and  $\rho$  are respectively given by

$$s_1^2 = \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 / n \quad s_2^2 = \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2 / n$$

$$r = \sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) / \sqrt{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2}$$

We can estimate the cdf of  $W_n$  as follows:

$$\hat{F}_{W_n}(w) = L\left(\frac{a_n - b_n t_w}{\sqrt{1 + t_w^2}}, -b_n, \frac{t_w}{\sqrt{1 + t_w^2}}\right) + L\left(\frac{b_n t_w - a_n}{\sqrt{1 + t_w^2}}, b_n, \frac{t_w}{\sqrt{1 + t_w^2}}\right)$$

where

$$a_n = \sqrt{\frac{n}{1 - r^2}} \left( \frac{\bar{x}_1}{s_1} - r \frac{\bar{x}_2}{s_2} \right) \quad b_n = \sqrt{n} \left( \frac{\bar{x}_2}{s_2} \right) \quad t_w = \sqrt{\frac{1}{1 - r^2}} \left( \frac{s_2}{s_1} w - r \right)$$

As reported by Cochran (1977), when  $n$  is not too small, the variance of the estimator  $W_n = \frac{\bar{X}_1}{\bar{X}_2}$  is approximately equal to:

$$\text{Var}(W_n) \cong \frac{R^2}{n} \{ CV_1^2 + CV_2^2 - 2\rho CV_1 CV_2 \}$$

with  $CV_1 = \frac{\sigma_1}{\mu_1}, CV_2 = \frac{\sigma_2}{\mu_2}$

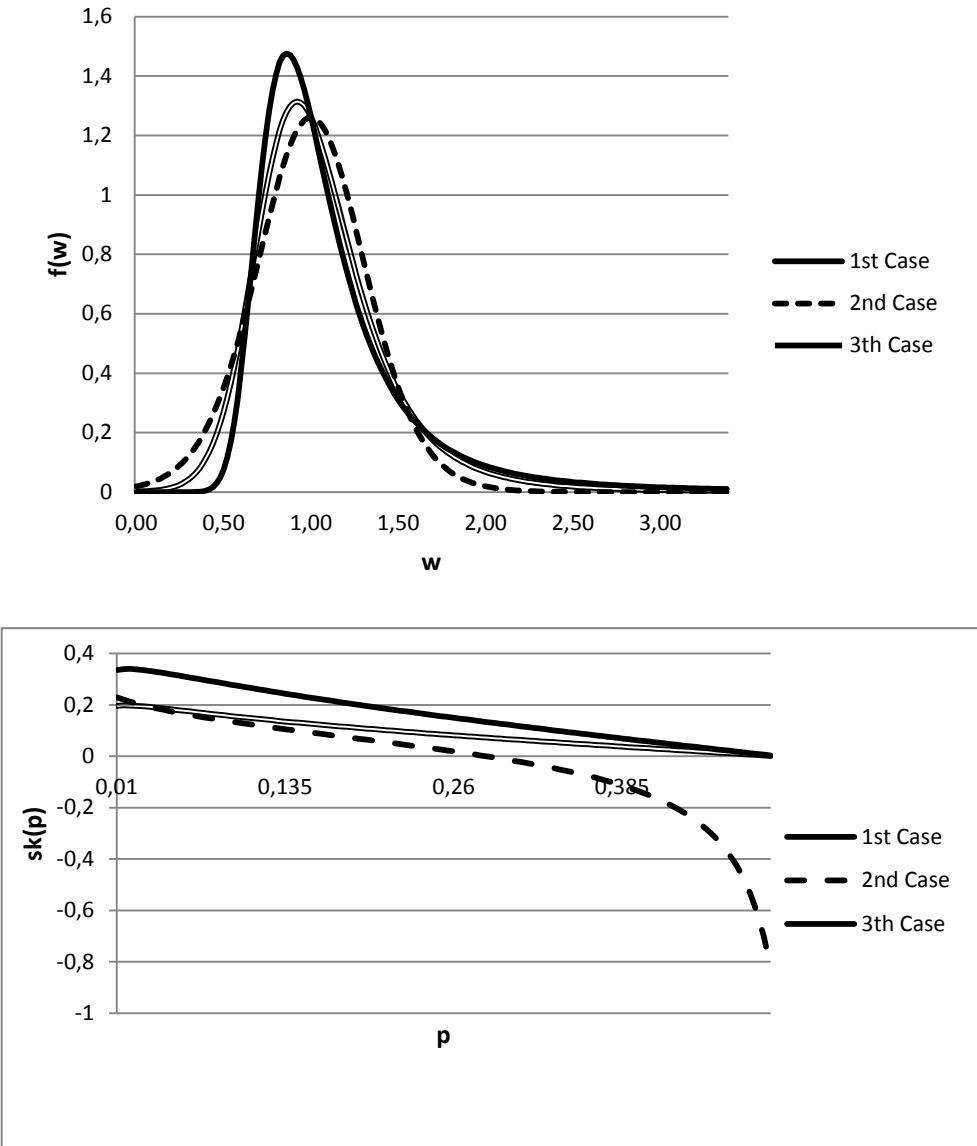
By means of this result, Cochran suggested to use the above approximate variance to build confidence intervals based on the Normal distribution. However, the Normal approximation is not always acceptable, as shown in several situations reported in paragraph 3.2.

### 3.1. Examples regarding the asymmetry of the distribution of the estimator

In order to understand the influences of the parameters of the rv  $W_n$  on the shape of the pdf, we report three situations fixing  $\mu_1 = \mu_2 = 1$  and choosing  $\sigma_1$ ,  $\sigma_2$  and  $\rho$  in order to obtain the same asymptotic variance of  $W_n$ . For each situation, we report the pdf  $f(w)$  and the relative functions  $sk(p)$  of the rv  $W_n$ .

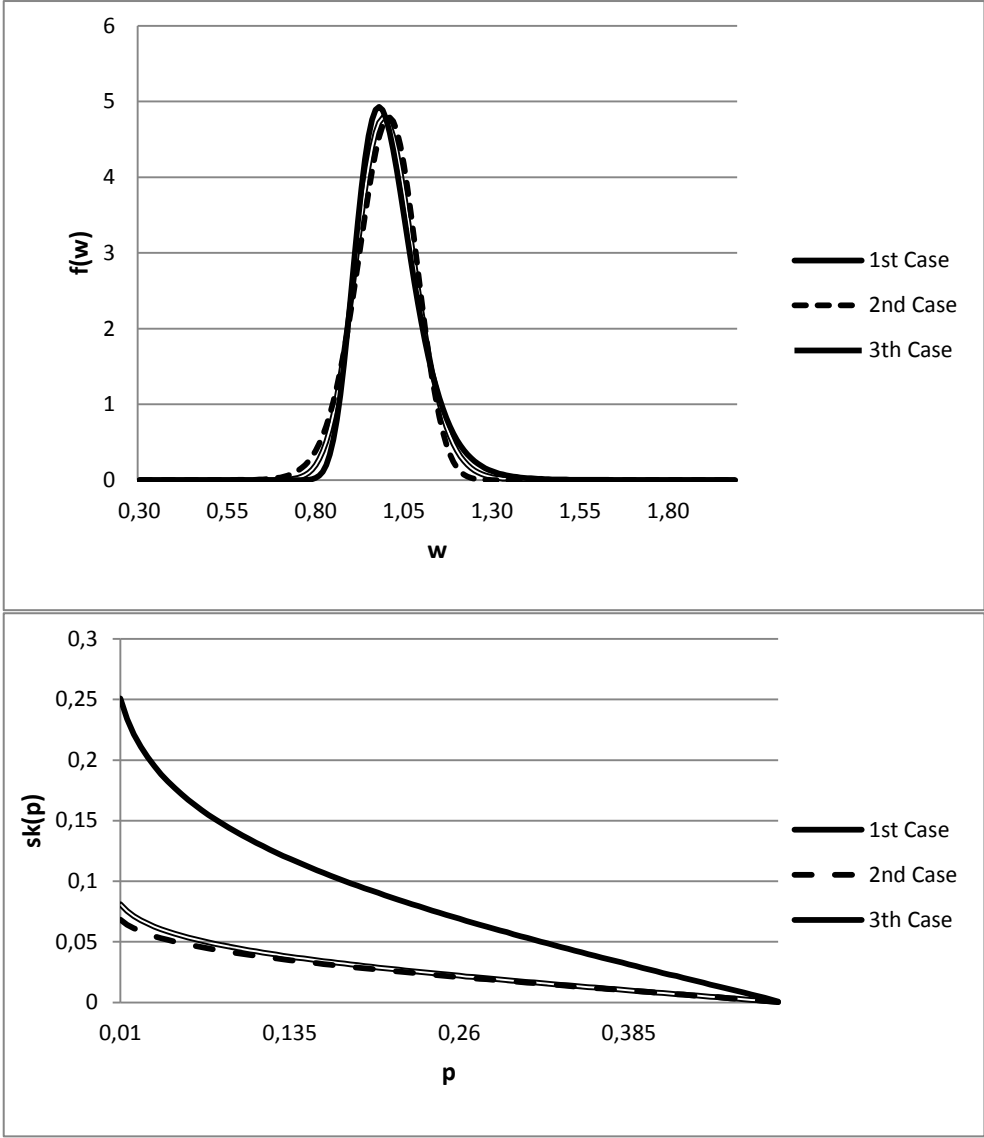
**Figure 3.** Pdf ( $f_w$ ) and asymmetry indexes ( $sk(p)$ ) in three situation with asymptotic variance equal to  $Var(W_n) = 0.01$  (sample size  $n=30$ )

The first case has  $CV_1^2 = 1, CV_2^2 = 4, \rho = 0.5$   
The second case has  $CV_1^2 = 4, CV_2^2 = 1, \rho = 0.5$   
The third case has  $CV_1^2 = 2.5, CV_2^2 = 2.5, \rho = 0.4$



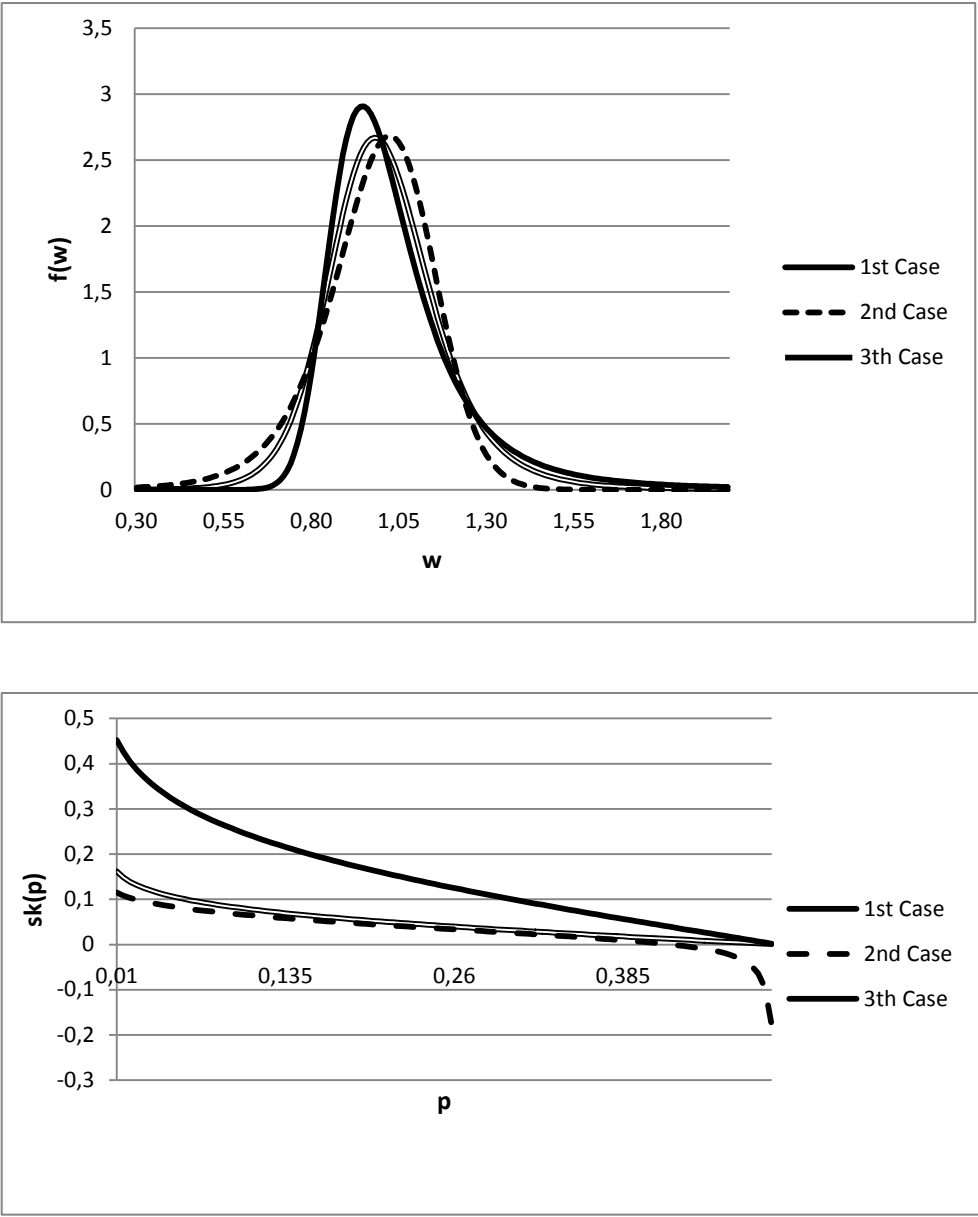
**Figure 4.** Pdf ( $f_w$ ) and asymmetry indexes ( $sk(p)$ ) in three situation with asymptotic variance equal to  $Var(W_n) = 0.0047$  (sample size  $n=30$ )

The first case has  $CV_1^2 = 0.1, CV_2^2 = 0.4, \rho = 0.9$   
The second case has  $CV_1^2 = 0.4, CV_2^2 = 0.1, \rho = 0.9$   
The third case has  $CV_1^2 = 0.25, CV_2^2 = 0.25, \rho = 0.72$



**Figure 5.** Pdf ( $f_w$ ) and asymmetry indexes ( $sk(p)$ ) in three situation with asymptotic variance equal to  $Var(W_n) = 0.01515$  (sample size  $n=30$ )

The first case has  $CV_1^2 = 1, CV_2^2 = 2, \rho = 0.9$   
The second case has  $CV_1^2 = 2, CV_2^2 = 1, \rho = 0.9$   
The third case has  $CV_1^2 = 1.5, CV_2^2 = 1.5, \rho = 0.8485$



In each figure considered above, the asymptotic variance  $Var(W_n)$  was the same obtained with difference values of  $\sigma_1$ ,  $\sigma_2$  and  $\rho$ , and the shapes of the relative pdf were different.

### 3.2. Examples of comparison of the distribution of the estimator with the normal distribution

We fix the mean values equal to one,  $\mu_1 = \mu_2 = 1$ , and we compute the areas under the right and left tails of the exact distribution of  $W_n$ , in different sets of parameters, as follows:

$$Pr(W_n < 0.8) = F_{w_n}(0.8) \text{ and } Pr(W_n > 1.2) = 1 - F_{w_n}(1.2)$$

In Table 1, these areas under the tails were compared to them obtained considering a rv  $Y \sim N(R, Var(W_n))$ , as follows:

$$Pr(Y < 0.8) = Pr(Y > 1.2) = F_Y(0.8)$$

From the values of the differences between the left and the right tails of the true distribution of the rv  $W_n$  and of the normal distribution, it is evident that the distribution of the rv  $W_n$  is skewness and the normal approximation is not appropriate. For this reason, the confidence intervals based on the normal distribution, as suggested by Cochran (1977), are not always convenient.

**Table 1.** Area under the tails of the pdf of  $W_n$  and  $Y$  in several sets of parameters with same asymptotic variance  $Var(W_n)$ .

**Differences between left tails (DLT)**  $= F_Y(0.8) - F_{w_n}(0.8)$

**Differences between right tails (DRT)**  $= F_Y(0.8) - (1 - F_{w_n}(1.2))$

1)  $n=20$   $Var(W_n) \cong 0.025$

$\sigma_1^2$	$\sigma_2^2$	$\rho$	$F_Y(0.8)$	$F_{W_n}(0.8)$	$1 - F_{W_n}(1.2)$	DLT	DRT
0.1	0.4	0.0	0.1030	0.0678	0.1273	0.0352	-0.0243
0.25	0.25	0.0	0.1030	0.0818	0.1203	0.0212	-0.0173
0.4	0.1	0.0	0.1030	0.0943	0.1106	0.0087	-0.0076

2)  $n=20$   $Var(W_n) \cong 0.015$

$\sigma_1^2$	$\sigma_2^2$	$\rho$	$F_Y(0.8)$	$F_{W_n}(0.8)$	$1 - F_{W_n}(1.2)$	DLT	DRT
0.1	0.4	0.5	0.0512	0.0218	0.0843	0.0294	-0.0331
0.25	0.25	0.4	0.0512	0.0369	0.0699	0.0143	-0.0187
0.4	0.1	0.5	0.0512	0.0524	0.0524	-0.0012	-0.0012

**Table 1.** Area under the tails of the pdf of  $W_n$  and  $Y$  in several sets of parameters with same asymptotic variance  $Var(W_n)$  (cont.)3)  $n=20$   $Var(W_n) \cong 0.07$ 

$\sigma^2_1$	$\sigma^2_2$	$\rho$	$F_Y(0.8)$	$F_{W_n}(0.8)$	$1 - F_{W_n}(1.2)$	$DLT$	$DRT$
0.1	0.4	0.9	0.0084	0.0003	0.0346	0.0081	-0.0262
0.25	0.25	0.72	0.0084	0.0052	0.0170	0.0032	-0.0086
0.4	0.1	0.9	0.0084	0.0165	0.0038	-0.0081	0.0046

4)  $n=30$   $Var(W_n) \cong 0.0167$ 

$\sigma^2_1$	$\sigma^2_2$	$\rho$	$F_Y(0.8)$	$F_{W_n}(0.8)$	$1 - F_{W_n}(1.2)$	$DLT$	$DRT$
0.1	0.4	0.0	0.1226	0.0942	0.1475	0.0284	-0.0249
0.25	0.25	0.0	0.1226	0.1054	0.1395	0.0172	0.0865
0.4	0.1	0.0	0.1226	0.1160	0.1298	0.0066	-0.0072

5)  $n=30$   $Var(W_n) \cong 0.01$ 

$\sigma^2_1$	$\sigma^2_2$	$\rho$	$F_Y(0.8)$	$F_{W_n}(0.8)$	$1 - F_{W_n}(1.2)$	$DLT$	$DRT$
0.1	0.4	0.5	0.0668	0.0397	0.0481	0.0271	0.0187
0.25	0.25	0.4	0.0668	0.0538	0.0826	-0.0531	-0.0158
0.4	0.1	0.5	0.0676	0.0668	0.0676	0.0008	0

6)  $n=30$   $Var(W_n) \cong 0.0047$ 

$\sigma^2_1$	$\sigma^2_2$	$\rho$	$F_Y(0.8)$	$F_{W_n}(0.8)$	$1 - F_{W_n}(1.2)$	$DLT$	$DRT$
0.1	0.4	0.9	0.0141	0.0022	0.0382	0.0119	-0.0241
0.25	0.25	0.72	0.0141	0.01	0.0221	-0.0041	-0.008
0.4	0.1	0.9	0.0141	0.0220	0.0085	-0.0079	0.0056

#### 4. The confidence interval for $R$ based on the exact distribution of the estimator

As discussed previously, none of the moments of  $W_n$  exists, and thus it is impossible to infer from the mean value  $E(W_n)$  and variance  $Var(W_n)$ .

Cochran (1977), in order to built the confidence intervals for  $R$ , used a normal distribution having asymptotic expected value  $R$  and  $Var(W_n)$ . Several authors (Fieller, 1932; Hinkley, 1969; Frosini, 1970) showed that the ratio of two normal rvs is approximately normal when the coefficient of variation of the denominator is negligible. Consequently, the rv  $W_n$  is approximately normal also when  $n$  is large. However, this condition is not always satisfied, especially in the practical situations, as discussed in the previous paragraph. An alternative approach to obtain the confidence interval for the ratio of the means in a bivariate normal distribution was proposed by Fieller (1940; 1954), and it is always called as

“Fieller’s theorem”. The calculation of the confidence interval is relatively simple and this approach has been used as a touchstone by several authors (Finney 1964, Rao 1965, Kendall and Stuart 1961), because of its importance in examining the general techniques for constructing confidence intervals using resampling techniques, such as the jackknife or bootstrapping. However, the existence of a bounded  $(1-\alpha)\%$  confidence interval for  $R$  with the Fieller’s theorem is not always guaranteed and in these cases the practical interpretation of the results is impossible. Gardiner et al. (2001) proved that the confidence interval is bounded if and only if the estimated  $\bar{X}_2$  is significantly different from zero at level  $\alpha$ .

In her PhD thesis (2007), Galeone proposed a new approach, called the exact distribution method, to build confidence interval for  $R$ , based on the inverse cdf of  $W_n$ . This approach always guarantees the existence of bounded confidence intervals, since the cdf is a monotonic non-decreasing function that can be inverted with computational methods. The  $(1-\alpha)$  confidence interval of  $R$ , obtained by inverting cdf of  $W_n$ , is given by

$$P\{W_{\alpha/2} \leq R \leq W_{1-\alpha/2}\} = 1 - \alpha$$

where  $W_{\alpha/2}$  and  $W_{1-\alpha/2}$  are the estimators (Galeone and Pollastri, 2008) of  $(\alpha/2)^{th}$  and the  $(1 - \alpha/2)^{th}$  quantile of the rv  $W_n$ .

The implementation of procedures and functions to build confidence intervals with the exact distribution method and with the Fieller’s theorem is available in Matlab, and the codes are reported in Appendix.

### ***Simulation study***

Monte Carlo experiment was used to assess the performances of the Fieller’s theorem and the exact distribution method for computing 90% confidence intervals for  $R$ , by differing levels of correlation between numerator and denominator. We have started using a simulated population with known means (0.25, 1.20) and variances (9, 16) of  $\bar{X}_1$  and  $\bar{X}_2$ , respectively, known correlations between rvs (0, |0.3|, |0.6|, |0.9|) and a known  $R$ . The sample size varied from 25 to 1600 with the rule of the doubling technique. Overall, there were 49 combinations of simulation parameters. For each combination of parameters, we have simulated 5000 independent samples for each treatment group from this population. The criterions used to evaluate the performances of the methods were the probability of coverage of the intervals (denoted as  $(1-\hat{\alpha})$ ), the average width of the intervals (denoted as *Amp*) and the symmetric miscoverage of the intervals (denoted as %*ds*).



### Simulation results

The performances of the two methods for the construction of 90% confidence intervals for  $R$  for  $\rho = 0.3$  were reported in Table 2. For small values of  $n$  ( $n \leq 200$ ), the confidence intervals constructed with the Fieller's theorem were not always bounded. For this reason the corresponding average widths were denoted as “-”, i.e., there was at least one unbounded confidence interval that yielded the average widths not to be expressed as a real number. Consequently, the corresponding coverage probabilities were very low. For elevated values of  $n$ , the performances of the confidence intervals based on the Fieller's theorem and the exact distribution method were very close.

**Table 2.** Simulation study - Performances of the two methods for the construction of 90% confidence interval with  $\rho=0.3$ .

n		Fieller's theorem	Exact distribution method
25	$(1 - \hat{\alpha})$	0.3941	0.9267
	%ds	0.5619	0.6178
	Amp	-	5.0478
50	$(1 - \hat{\alpha})$	0.5947	0.9196
	%ds	0.3476	0.5463
	Amp	-	3.0029
100	$(1 - \hat{\alpha})$	0.8192	0.9101
	%ds	0.2367	0.5222
	Amp	-	1.5692
200	$(1 - \hat{\alpha})$	0.8639	0.8968
	%ds	0.6352	0.5368
	Amp	-	0.7281
400	$(1 - \hat{\alpha})$	0.8974	0.8972
	%ds	0.4815	0.4805
	Amp	0.4338	0.4326
800	$(1 - \hat{\alpha})$	0.9018	0.9010
	%ds	0.5173	0.5192
	Amp	0.2905	0.2901
1600	$(1 - \hat{\alpha})$	0.9008	0.9006
	%ds	0.4980	0.4976
	Amp	0.2002	0.2001

$(1 - \hat{\alpha})$ : probability of coverage of the intervals

%ds : symmetric miscoverage of the intervals

Amp: average width of the intervals

Extending the simulation results to all other values of  $\rho$  considered, the Fieller's theorem always failed for  $n \leq 50$ , with corresponding non-acceptable coverage probabilities. For  $\rho$  equal to -0.6 and -0.9, the Fieller's theorem failed also for  $n$  equal to 100, but in these cases the coverage probabilities were higher as referred to those for  $n < 100$ . For other values of  $\rho$ , i.e., equal to -0.3, 0 and 0.6, the Fieller's theorem failed also for  $n$  equal to 200. The simulation results highlighted that the Fieller's theorem less frequently produces unbounded confidence intervals for  $R$  with increasing values of  $n$ . Finally, the performances of the two methods were satisfactory and very close to each other for high values of  $n$ .

## 5. Conclusions

The present paper shows the importance of considering the real distribution of the estimator of the ratio of two means  $W_n$ , because generally the approximation to normal is not satisfied. For this reason, building the confidence intervals with the Cochran approach is not always appropriate. The Fieller's theorem is a general procedure to construct confidence interval for the ratio of the means in a bivariate normal distribution. The calculation of the confidence interval is relatively simple and this approach has been used as a touchstone by several authors. However, there are several cases, i.e., the estimated  $\bar{X}_2$  is significantly different from zero at level  $\alpha$ , for which the confidence interval is not bounded. In this paper, the authors propose an alternative method to compute the confidence interval, based on the inverse cdf of  $W_n$ , called the exact distribution method. Although the calculus of the confidence intervals by means of the exact distribution method is more complicated, since this involves the calculation of the inverse of a cdf that can be obtained only by a computer support, the novel method proposed always allows one to obtain bounded confidence intervals, also when the Fieller's theorem produces unbounded intervals. Besides theoretical interest, this may be useful in several applications. For example, the problem to treat the cost-effectiveness ratio in the equivalence studies, i.e., when the difference in effectiveness between the new treatment and the control treatment is close to zero, has recently arisen in the cost-effectiveness analysis. In these cases the confidence intervals for the cost-effectiveness ratio cannot be obtained with the Fieller's theorem and in medical literature there are no satisfactory parametric methods to construct confidence intervals. Therefore, the exact distribution method would be valuable. In order to encourage the dissemination of the novel method, in Appendix we have reported the codes of procedures and functions, with relative descriptions, to build confidence intervals with the exact distribution method in the scientific software package MATLAB. For completeness, we have also reported the code for building confidence intervals with the Fieller's theorem.

## REFERENCES

- AROIAN L. A. (1986). The distribution of the quotient of two correlated random variables. *Proceedings of the Am. Stat. Ass. Business and Economic Section*.
- BRENTARI E. (1990). *Asimmetria e misure di Asimmetria*, Giappichelli Ed., Torino.
- COCHRAN W. C. (1997). *Sampling Techniques*, John Wiley and Sons, New York, Wiley; 3<sup>rd</sup> ed.
- EVERITT B. S. (1996). An Introduction to finite mixture distributions, *Stat Methods Med Rev*, 5(2): 107-127.
- FIELLER E. C. (1932). The distribution of the index in a Normal Bivariate population. *Biometrika* 24(3/4): 428-440.
- FIELLER E. C. (1954). Some Problems in Interval Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* 16(2): 175-185.
- FINNEY D. J. (1964). *Statistical Method in Biological Assay*, Hafner, New York.
- FROSINI B. V. (1970). La stima di un quoziente nei grandi campioni. *Giornale degli economisti e annali di economia* : 381-400.
- FROSINI B. V. (1971). Le distribuzioni oblique, *Statistica*, 31: 83-117 (in FROSINI B. V. (2011). *Selected Writings*, Vita e Pensiero, Milano).
- GALEONE C. (2007). On the ratio of two Normal r.v., jointly distributed as a Bivariate Normal, *PhD Thesis, Università di Milano-Bicocca*.
- GALEONE C., POLLASTRI A. (2008). Estimation of the quantiles of the ratio of two correlated Normals, *XLIV Riunione scientifica della Società Italiana di Statistica*.
- GARDINER J. C., HUEBNER M. et al. (2001). On Parametric Confidence Intervals for Cost-Effectiveness Ratio, *Biometrical Journal*, 43 (3): 283-296.
- GEARY R. C. (1930). The frequency distribution of the quotient of two Normal variates. *J Royal Stat Society*, 93(3): 442-446.
- GROENEVELD R.A. (1998). Bowley's measures of skewness, *Encyclopedia of statistical Science*, Update Vol. 2, 619-621, Wiley.
- JOHNSON N. L., KOTZ S., BALAKRISHNAN N. (1994). Continuous Univariate Distributions, Wiley, New York.
- KENDALL M. G., STUART A. (1961). *The advanced Theory of Statistics*, Vol. 2, C. Griffin & Co., London.
- KOTZ S., BALAKRISHNAN N., JOHNSON N. L. (2000). Continuous Multivariate Distributions, Wiley, New York.
- MacGILLIVRAY H. L.(1985). Mean, Median, Mode, Skewness, *Encyclopedia of Statistical Science*, Vol. 5, 364-367, Wiley.

- MacGILLIVRAY H. L. (1986). Skewness and asymmetry: Measures and Orderings, *The Annals Of Mathematical Statistics*, Vol. 14, No. 3 :994-1011.
- MARSAGLIA, G. (1965). Ratios of Normal variables and ratio of sums of Uniform variables. *J Am Statist Ass* 60(309): 193-204.
- MARSAGLIA, G. (2006). Ratios of Normal variables. *Journal of Statistical Software*, 16(4).
- MOOD A. M, GRAYBILL F. A., BOES D. C. (1974). *Introduction to the Theory of Statistics*, McGraw-Hill, New York.
- NATIONAL BUREAU STANDARDS (1959). *Tables of the Bivariate Normal distribution function and related functions*, U.S. Government Printing Office, Washington.
- OKSOY D., AROIAN L. A. (1986). Computational Techniques and examples of the density and the distribution of the quotient of two correlated normal variables, *Proceeding of the American Statistical Association Business and Economic Section*.
- OKSOY D., AROIAN L. A. (1994). The quotient of two correlated normal variables with applications, *Comm Stat Simula*, 23 (1): 223-241.
- RAO C. R. (1965). *Linear Statistical Inference and its applications*, Wiley, New York.

## APPENDIX

Procedures and functions implemented in MATLAB, useful to construct confidence intervals with the exact distribution method and the Fieller's theorem

<i>Function 1</i>	Cumulative density function: $F(w)$
<i>Function 2</i>	Probability density function: $f(w)$
<i>Function 3</i>	Fieller's theorem for the construction of confidence intervals for R
<i>Function 4</i>	Exact distribution method for the construction of confidence intervals for R

### Function 1

% Definition Function: cdf  $F(W)$  of the ratio of two normal random variables ( $w=x1/x2$ )  
function fcum=cdfratio(media1,media2,var1,var2,r,w)

% Insert parameters: means (media1, media2), variances (var1,var2), coefficient of correlation (r) and range of W (w)

```

sqm1=sqrt(var1);
sqm2=sqrt(var2);

raprho=1/(sqrt(1-r.^2));
b=media2/sqm2;
a=((media1/sqm1)-(r*b))*raprho;

t=((((sqm2.*w)/sqm1))-r).*raprho;
rq=sqrt(1+t.^2);
rho=t./rq;
x=(b.*t-a)./rq;
xg=-b;
xs=-x;

pt=zeros(length(w),1);
pp=zeros(length(w),1);
fcum=zeros(length(w),1);

for i=1:length(w)
    pt(i) = bivnormcdf(x(i),b,rho(i));
    pp(i)= bivnormcdf(xs(i),xg,rho(i));

    fcum(i)=pt(i)+pp(i); %cdf
end

```

## Function 2

% Definition Function: pdf  $f(W)$  of the ratio of two normal random variables ( $w=x1/x2$ )  
function fden = pdfratio(media1,media2,var1,var2,r,w)

% Insert parameters: means (media1, media2), variances (var1,var2), coefficient of correlation (r) and range of W (w)

```

sqm1=sqrt(var1);
sqm2=sqrt(var2);

raprho=1/(sqrt(1-r.^2));
b=media2/sqm2;
a=((media1/sqm1)-(r*b))*raprho;

costa=(sqm2/sqm1)*raprho;
gg=(1/pi)*exp(-0.5*(a^2+b^2));
t=((((sqm2/sqm1).*w)-r)*raprho;

q=(b+a.*(t))./(sqrt(1+(t.^2)));
pc=1./(1+t.^2);

funq=(1/sqrt(2*pi))*exp(-0.5*q.^2);
cc=q./funq;
dn=normcdf(q)-0.5;

```

```
og=(1+cc.*dn);
```

```
fden=(costa.*gg.*pc.*og); % pdf
```

### Function 3

% Fieller method for the construction of confidence intervals for the ratio of means of Normal random variables ( $R=\mu_1/\mu_2$ )

```
function [visa]=idcfieller(mcamp1,mcamp2,vcampm1,vcampm2,rhocamp,t,nv)
```

% Insert parameters: sample means (mcamp1, mcamp2), sample variances of means (vcampm1,vcampm2), sample coefficient of correlation (rhocamp), Student's percentile point with n-1 df (t) and number of repetitions (nv)

```
for i=1:nv
```

```
co(i)=rhocamp(i)*(sqrt(vcampm1(i)*vcampm2(i)));
```

```
an(i)=(mcamp2(i)^2)-((t^2)*(vcampm2(i)));
```

```
bn(i)=(mcamp2(i)*mcamp1(i))-((t^2)*(co(i)));
```

```
cn(i)=(mcamp1(i)^2)-((t^2)*vcampm1(i));
```

```
inf(i)=(bn(i)-sqrt((bn(i)^2)-(an(i).*cn(i))))/(an(i));
```

```
sup(i)=(1/an(i))*(bn(i)+sqrt((bn(i)^2)-(an(i).*cn(i))));
```

```
visa=[inf;sup]';
```

```
end
```

### Function 4

% Exact distribution method for the construction of confidence intervals for the ratio of means of Normal random variables ( $R=\mu_1/\mu_2$ )

```
function [visa]=idcinv(mcamp1,mcamp2,vcampm1,vcampm2,rhocamp,x0,alfa,beta,nv)
```

% Insert parameters: sample means (mcamp1, mcamp2), sample variances of means (vcampm1,vcampm2), sample coefficient of correlation (rhocamp),  $x_0=0$ ,  $\alpha/2$  value ( $\alpha$ ),  $1-\alpha/2$  value ( $\beta$ ) and number of repetitions (nv)

```
for i=1:nv
```

```
wmin(i)=fzero(@cdfpr,x0,[],mcamp1(i),mcamp2(i),vcampm1(i),vcampm2(i),rho  
camp(i),alfa);
```

```
wmax(i)=fzero(@cdfpr,x0,[],mcamp1(i),mcamp2(i),vcampm1(i),vcampm2(i),rho  
camp(i),beta);
```

```
visa=[wmin;wmax]';
```

```
end
```

## ON CLASSES OF MODIFIED RATIO TYPE AND REGRESSION-CUM-RATIO TYPE ESTIMATORS IN SAMPLE SURVEYS USING TWO AUXILIARY VARIABLES

A.K.P.C. Swain<sup>1</sup>

### ABSTRACT

In this paper generalized classes of modified ratio type and regression-cum-ratio type estimators of the finite population mean of the study variable are suggested in the presence of two auxiliary variables in simple random sampling without replacement when the population means of the auxiliary variables are known in advance. Some special cases of the generalized estimators are compared with respect to their biases and efficiencies both theoretically and with the help of some natural populations.

**Key words:** ratio type estimators, regression-cum-ratio type estimators, simple random sampling, auxiliary variables, bias, efficiency.

### 1. Introduction

In sample surveys a sampler invariably observes certain auxiliary variables to provide more efficient estimators of the finite population mean of the study variable. The literature on the use of auxiliary information in sample surveys is quite vast and old dating back to early part of the 20th century when the foundation stone of modern sampling theory dealing with stratified random sampling was laid out utilizing the auxiliary information by Bowley (1926) and Neyman (1934, 38). However, the use of auxiliary information in the estimation procedure to improve the precision of estimators was initiated by Watson (1937) and Cochran (1940, 42). Hansen and Hurwitz (1943) were the first to suggest the use of auxiliary information in selecting units with varying probabilities. The customary sources of obtaining auxiliary information on one or more variables having strong correlation with the main variable under study are various census

---

<sup>1</sup> Former Professor of Statistics Utkal University, Bhubaneswar-751004, India. E-mail: swainakpc@yahoo.co.in.

data, previous surveys, pilot surveys, etc., or may be made available while collecting information on the study variable during the survey operations.

Cochran (1940) was the first to introduce a ratio estimator of the population mean of the study variable by using a single auxiliary variable in the form of a ratio of the sample mean of the study variable to the sample mean of the auxiliary variable, multiplied by the population mean of the auxiliary variable. That is, the classical ratio estimator of the population mean  $\bar{Y}$  of the study variable  $y$  in simple random sampling is defined as

$$\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X},$$

where  $\bar{y}$  and  $\bar{x}$  are the sample means of the study variable  $y$  and auxiliary variable  $x$  respectively and  $\bar{X}$  is the population mean of the auxiliary variable  $x$ . Ratio estimator is seen to be most efficient when the regression line of  $y$  on  $x$  passes through the origin. When the regression line does not pass through the origin, Cochran (1940, 42) suggested a linear regression estimator which was generalized by Hansen et al. (1953) in the form of a difference estimator. The linear regression estimator of the population mean  $\bar{Y}$  is defined as

$$\hat{\bar{Y}}_{Reg} = \bar{y} + b_{yx}(\bar{X} - \bar{x}),$$

where  $b_{yx}$  is the sample regression coefficient of  $y$  on  $x$  and the difference estimator is defined as

$$\hat{\bar{Y}}_D = \bar{y} + \lambda(\bar{X} - \bar{x}),$$

where  $\lambda$  is a real constant to be suitably chosen.

To estimate the population mean  $\bar{Y}$  of the study variable  $y$  in the presence of  $p$ -auxiliary variables  $x_1, x_2, \dots, x_p$  with the advance knowledge of the population means  $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p$  respectively, Tripathi (1970, 87) discussed two general classes of estimators for any sampling design, defined by

$$\hat{\bar{Y}}_{MT} = \sum_{i=1}^p W_i \left[ \hat{\bar{Y}} - t_i \left( \hat{\bar{X}}_i - \bar{X}_i \right) \right],$$

and

$$\hat{\bar{Y}}_{MT}^* = \hat{\bar{Y}} - \sum_{i=1}^p t_i^* \left( \hat{\bar{X}}_i - \bar{X}_i \right),$$

where  $\hat{\bar{Y}}$  and  $\hat{\bar{X}}_i$  are unbiased estimators of  $\bar{Y}$  and  $\bar{X}_i$  ( $i = 1, 2, \dots, p$ ) respectively,  $t_i$  and  $t_i^*$  are statistics (or real constants) such that their expected

values exist,  $W_i$ 's are non-negative and  $\sum_{i=1}^p W_i = 1$ .

These classes include Olkin's (1958) multivariate ratio estimator, Raj's (1965) multivariate difference estimator, Ghosh's (1947), Srivastava's (1965) and



Shukla's (1965) multivariate regression estimator. Tripathi (1987) also considered other classes of estimators for any sampling design defined by

$$e_1 = \hat{Y} - \sum_{i=1}^p t_i \left( \hat{X}_i^{\alpha_i} - \bar{X}_i^{\alpha_i} \right), \quad e_2 = \hat{Y} \prod_{i=1}^p \left( \frac{\bar{X}_i}{\hat{X}_i} \right)^{\alpha_i}$$

$$e_3 = \hat{Y} \sum_{i=1}^p W_i \left( \frac{\bar{X}_i}{\hat{X}_i} \right)^{\alpha_i}, \quad e_4 = \hat{Y} \frac{\sum_{i=1}^p W_i \bar{X}_i^{\alpha_i}}{\sum_{i=1}^p W_i \hat{X}_i^{\alpha_i}}$$

$$e_5 = \hat{Y} \frac{\sum_{i=1}^p W_i \hat{X}_i^{\alpha_i}}{\sum_{i=1}^p W_i \bar{X}_i^{\alpha_i}}$$

where  $\alpha_i$ 's are suitably chosen constants and  $W_i$ 's are non-negative weights such that

$$\sum_{i=1}^p W_i = 1.$$

These classes include estimators proposed by Shukla (1966) and John (1969). Singh (1965,67) suggested a ratio-cum-product estimator where some of the auxiliary variables are positively correlated and others are negatively correlated with the study variable.

Srivastava (1971) suggested a general class of estimator in case of simple random sampling without replacement, defined by

$$e_6 = \bar{y} \left( \frac{\bar{x}_1}{\bar{X}_1}, \frac{\bar{x}_2}{\bar{X}_2}, \dots, \frac{\bar{x}_p}{\bar{X}_p} \right) = \bar{y}(u_1, u_2, \dots, u_p),$$

where  $\bar{y}, \bar{x}_1, \bar{x}_2, \dots, \bar{x}_p$  are sample means of  $y, x_1, x_2, \dots, x_p$  respectively and

$u_i = \frac{\bar{x}_i}{\bar{X}_i}$ ,  $i = 1, 2, \dots, p$  and  $h(\cdot)$  is a function of  $u_1, u_2, \dots, u_p$  obeying

regularity conditions, such as

- The point  $(u_1, u_2, \dots, u_p)$  assumes the value in a closed convex subset  $R_p$  of  $p$ -dimensional real space containing the point  $(1, 1, \dots, 1)$ .
- The function  $h(u_1, u_2, \dots, u_p)$  is continuous and bounded in  $R_p$ .
- $h(1, 1, \dots, 1) = 1$ .
- The first and second order partial derivatives of  $h(u_1, u_2, \dots, u_p)$  exist and are continuous and bounded in  $R_p$ .

Subsequently, Srivastava and Jhaji (1983) suggested a wider class of estimators defined by

$$e_7 = g \left( \bar{y}, \frac{\bar{x}_1}{\bar{X}_1}, \dots, \frac{\bar{x}_p}{\bar{X}_p} \right)$$

In what follows we shall consider certain specific classes of modified ratio type, difference-cum-ratio type and regression-cum-ratio type estimators and compare them as regards their biases and efficiencies in the presence of two auxiliary variables having known population means.

## 2. A generalized class of modified ratio type estimators

Let  $U = (U_1, U_2, \dots, U_N)$  be the finite population of size  $N$ . To each unit  $U_i$  ( $i = 1, 2, \dots, N$ ) in the population, the values of the study variable  $y$  and the auxiliary variables  $x$  and  $z$  denoted by the triplet  $(y_i, x_i, z_i)$ , ( $i = 1, 2, \dots, N$ ) are attached.

Now, define the population means of the study variable  $y$  and the auxiliary variables  $x$  and  $z$  respectively as

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i, \quad \bar{X} = \frac{1}{N} \sum_{i=1}^N x_i, \quad \bar{Z} = \frac{1}{N} \sum_{i=1}^N z_i$$

Further, define the finite population variances of  $y$ ,  $x$  and  $z$  and their covariances as

$$\begin{aligned} S_y^2 &= \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2, S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2 \\ S_z^2 &= \frac{1}{N-1} \sum_{i=1}^N (z_i - \bar{Z})^2, S_{yx} = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X}) \\ S_{yz} &= \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})(z_i - \bar{Z}), \text{ and } S_{xz} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})(z_i - \bar{Z}) \end{aligned}$$

Also, the coefficients of variations of  $y$ ,  $x$  and  $z$  and their coefficients of covariation are defined by

$$\begin{aligned} C_y &= \frac{S_y}{\bar{Y}}, \quad C_x = \frac{S_x}{\bar{X}}, \quad C_z = \frac{S_z}{\bar{Z}} \\ C_{yx} &= \frac{S_{yx}}{\bar{Y}\bar{X}} = \rho_{yx} C_y C_x, \quad C_{yz} = \frac{S_{yz}}{\bar{Y}\bar{Z}} = \rho_{yz} C_y C_z, \text{ and } C_{xz} = \frac{S_{xz}}{\bar{X}\bar{Z}} = \rho_{xz} C_x C_z, \end{aligned}$$

where  $\rho_{yx}$ ,  $\rho_{yz}$  and  $\rho_{xz}$  are simple correlations between  $y$  and  $x$ ,  $y$  and  $z$  and  $x$  and  $z$  respectively.

A simple random sample 's' of size  $n$  is selected from  $U$  without replacement and values  $(y_i, x_i, z_i)$ ,  $i = 1, 2, \dots, n$  are observed on the sampled units.

Define the sample means of  $y, x$  and  $z$  as

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } \bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$$

Let us propose a class of generalized modified ratio type estimator defined by

$$\hat{Y}_{gmr} = \bar{y} \left[ \alpha_1 \left( \frac{\bar{X}}{\bar{x}} \right)^{g_1} + (1 - \alpha_1) \left( \frac{\bar{x}}{\bar{X}} \right)^{h_1} \right]^{\delta_1} \left[ \alpha_2 \left( \frac{\bar{Z}}{\bar{z}} \right)^{g_2} + (1 - \alpha_2) \left( \frac{\bar{z}}{\bar{Z}} \right)^{h_2} \right]^{\delta_2}$$

where  $\alpha_1, \alpha_2, g_1, g_2, h_1, h_2, \delta_1$  and  $\delta_2$  are real constants to be determined suitably.  $0 < \alpha_1, \alpha_2 < 1$ . We may fix  $g_1, g_2, h_1, h_2, \delta_1$  and  $\delta_2$  determine the optimum values of  $\alpha_1$  and  $\alpha_2$  by minimizing the mean square of  $\hat{Y}_{gmr}$ . Now, write

$$\bar{y} = \bar{Y}(1 + e_1), \bar{x} = \bar{X}(1 + e_2), \bar{z} = \bar{Z}(1 + e_3),$$

$$\text{where } e_1 = \frac{\bar{y} - \bar{Y}}{\bar{Y}}, e_2 = \frac{\bar{x} - \bar{X}}{\bar{X}} \text{ and } e_3 = \frac{\bar{z} - \bar{Z}}{\bar{Z}}$$

$$\text{We now have } E(e_1) = E(e_2) = E(e_3) = 0,$$

$$V(e_1) = \theta C_y^2, V(e_2) = \theta C_x^2, V(e_3) = \theta C_z^2$$

$$\text{Cov}(e_1, e_2) = \theta C_{yx}, \text{Cov}(e_1, e_3) = \theta C_{yz}, \text{ and } \text{Cov}(e_2, e_3) = \theta C_{xz}$$

$$\text{where } \theta = \left( \frac{1}{n} - \frac{1}{N} \right).$$

Assuming that  $\hat{Y}_{gmr}$  is a continuous function of  $\bar{y}, \bar{x}$  and  $\bar{z}$  and the first and second order partial derivatives of  $\hat{Y}_{gmr}$  exist, we may expand  $\hat{Y}_{gmr}$  in a Taylor's series at the point  $\bar{y} = \bar{Y}, \bar{x} = \bar{X}$  and  $\bar{z} = \bar{Z}$  and write

$$\begin{aligned} \hat{Y}_{gmr} - \bar{Y} = \bar{Y} & \left[ e_1 - \delta_2 \mu_1 + \frac{\delta_2(\delta_2 - 1)}{2} \mu_2 \right. \\ & \left. - \delta_1 \lambda_1 + \delta_1 \delta_2 \mu_1 \lambda_1 + \frac{\delta_1(\delta_1 - 1)}{2} \lambda_2 - \delta_2 \mu_1 e_1 - \delta_1 \lambda_1 e_1 + \dots \right] \end{aligned}$$

where

$$\lambda_1 = \alpha_1 \left[ (g_1 + h_1)e_2 + \left( \frac{h_1(h_1 - 1)}{2} - \frac{g_1(g_1 + 1)}{2} \right) e_2^2 \right] - \left( h_1 e_2 + \frac{h_1(h_1 - 1)}{2} e_2^2 \right)$$

$$\lambda_2 = \{ \alpha_1 (g_1 + h_1) + h_1 \}^2 e_2^2$$

$$\mu_1 = \alpha_2 \left[ (g_2 + h_2)e_3 + \left( \frac{h_2(h_2 - 1)}{2} - \frac{g_2(g_2 + 1)}{2} \right) e_3^2 \right] - \left( h_2 e_3 + \frac{h_2(h_2 - 1)}{2} e_3^2 \right)$$

$$\mu_2 = \{ \alpha_2 (g_2 + h_2) + h_2 \}^2 e_3^2$$

Retaining the first degree terms of  $e_1$ ,  $e_2$  and  $e_3$  we have

$$\begin{aligned} \hat{Y}_{gmr} - \bar{Y} &\cong \bar{Y} \left[ e_1 - \delta_1 \{ \alpha_1 (g_1 + h_1) e_2 - h_1 e_2 \} - \delta_2 \{ \alpha_2 (g_2 + h_2) e_3 - h_2 e_3 \} \right] \\ &= \bar{Y} \left[ e_1 - \delta_1 e_2 \{ \alpha_1 (g_1 + h_1) - h_1 \} - \delta_2 e_3 \{ \alpha_2 (g_2 + h_2) - h_2 \} \right] \end{aligned}$$

Thus, to the first order of approximation, i.e. to  $0\left(\frac{1}{n}\right)$ , the mean square error

(MSE) of  $\hat{Y}_{gmr}$  is given by

$$\begin{aligned} MSE(\hat{Y})_{gmr} &= \theta \bar{Y}^2 \left[ C_y^2 + \delta_1^2 \{ \alpha_1 (g_1 + h_1) - h_1 \}^2 C_x^2 + \delta_2^2 \{ \alpha_2 (g_2 + h_2) - h_2 \}^2 C_z^2 \right. \\ &\quad - 2\delta_1 \{ \alpha_1 (g_1 + h_1) - h_1 \} C_{yx} \\ &\quad - 2\delta_2 \{ \alpha_2 (g_2 + h_2) - h_2 \} C_{yz} \\ &\quad \left. + 2\delta_1 \delta_2 \{ \alpha_1 (g_1 + h_1) - h_1 \} \{ \alpha_2 (g_2 + h_2) - h_2 \} C_{xz} \right] \end{aligned}$$

Now, minimizing  $MSE(\hat{Y}_{gmr})$  with respect to  $\alpha_1$  and  $\alpha_2$ , we have

$$\begin{aligned} \alpha_{1(opt)} &= \frac{h_1}{g_1 + h_1} + \frac{1}{\delta_1 (g_1 + h_1)} \left[ \frac{C_x^2 C_{yx} - C_{xz} C_{yz}}{C_x^2 C_z^2 - C_{xz}^2} \right] \\ \alpha_{2(opt)} &= \frac{h_2}{g_2 + h_2} + \frac{1}{\delta_2 (g_2 + h_2)} \left[ \frac{C_x^2 C_{yz} - C_{xz} C_{yx}}{C_x^2 C_z^2 - C_{xz}^2} \right] \end{aligned}$$

Substituting  $\alpha_{1(opt)}$  and  $\alpha_{2(opt)}$  in the expression for  $MSE(\hat{Y}_{gmr})$  we have to

$$0\left(\frac{1}{n}\right),$$

$$MSE \left( \hat{Y}_{gmr} \right)_{opt} = \theta \bar{Y}^2 C_y^2 (1 - R_{y.xz}^2)$$

where  $R_{y.xz}$  is the multiple correlation coefficient of  $y$  on  $x$  and  $z$ .

Also, to  $O\left(\frac{1}{n}\right)$  the bias of  $\hat{Y}_{gmr}$  with optimum values of  $\alpha_1$  and  $\alpha_2$  is given by

$$\begin{aligned} \text{Bias} \left( \hat{Y}_{gmr} \right)_{opt} &= \theta \bar{Y} \left[ \left\{ \delta_1 \frac{h_1(h_1-1)}{2} + \frac{(\delta_1 h_1 + m_1)(1 + g_1 - h_1)}{2} + \frac{\delta_1 - 1}{2\delta_1} m_1^2 \right\} C_x^2 \right. \\ &\quad \left. + \left\{ \delta_2 \frac{h_2(h_2-1)}{2} + \frac{(\delta_2 h_2 + m_2)(1 + g_2 - h_2)}{2} + \frac{\delta_2 - 1}{2\delta_2} m_2^2 \right\} C_z^2 \right. \\ &\quad \left. - m_1 C_{yx} - m_2 C_{yz} + m_1 m_2 C_{xz} \right] \end{aligned}$$

where

$$m_1 = \frac{C_z^2 C_{yx} - C_{xz} C_{yz}}{C_x^2 C_z^2 - C_{xz}^2} = \frac{C_y}{C_x} \left[ \frac{\rho_{yx} - \rho_{xz} \rho_{yz}}{1 - \rho_{zx}^2} \right]$$

$$m_2 = \frac{C_x^2 C_{yz} - C_{xz} C_{yx}}{C_x^2 C_z^2 - C_{xz}^2} = \frac{C_y}{C_z} \left[ \frac{\rho_{yz} - \rho_{xz} \rho_{yx}}{1 - \rho_{xz}^2} \right]$$

## 2.1. Some special cases of generalized modified ratio type estimators

Consider nine estimators  $t_1, t_2, \dots, t_9$  in Table 1, which are special cases of  $\hat{Y}_{gmr}$  by substituting some simple but arbitrary real constants for  $g_1, g_2, h_1, h_2, \delta_1$  and  $\delta_2$ . The optimal asymptotic mean square errors (optimized with respect to  $\alpha_1$  and  $\alpha_2$ ) of these cases are equal to the optimal asymptotic  $MSE(\hat{Y}_{gmr})$  in the general case, which is independent of free parameters  $g_1, g_2, h_1, h_2, \delta_1$  and  $\delta_2$ .

$$\begin{aligned} \text{Thus, } MSE(t_1) &= MSE(t_2) = \dots = MSE(t_9) = MSE(\hat{Y}_{gmr}) \\ &= \theta \bar{Y}^2 C_y^2 (1 - R_{y.xz}^2). \end{aligned}$$

The biases of estimators to  $O\left(\frac{1}{n}\right)$  in the special cases excepting the constant multiplier  $\theta \bar{Y}$  are presented in Table 1.

**Table 1.** Biases of some modified ratio type estimators

Estimator	$g_1$	$h_1$	$g_2$	$h_2$	$\delta_1$	$\delta_2$	Bias $\left(\hat{\bar{Y}}_{gmr}\right) / \theta \bar{Y}$
$t_1 = \bar{y} \left[ \alpha_1 \left( \frac{\bar{X}}{\bar{x}} \right) + (1 - \alpha_1) \left( \frac{\bar{x}}{\bar{X}} \right) \right]$ $\times \left[ \alpha_2 \left( \frac{\bar{Z}}{\bar{z}} \right) + (1 - \alpha_2) \left( \frac{\bar{z}}{\bar{Z}} \right) \right]$	1	1	1	1	1	1	$\frac{m_1 + 1}{2} C_x^2 + \frac{m_2 + 1}{2} C_z^2$ $-m_1 C_{yx} - m_2 C_{yz} + m_1 m_2 C_{xz}$
$t_2 = \bar{y} \left[ \alpha_1 \left( \frac{\bar{X}}{\bar{x}} \right) + (1 - \alpha_1) \right]$ $\times \left[ \alpha_2 \left( \frac{\bar{Z}}{\bar{z}} \right) + (1 - \alpha_2) \right]$	1	0	1	0	1	1	$m_1 C_x^2 + m_2 C_z^2$ $-m_1 C_{yx} - m_2 C_{yz} + m_1 m_2 C_{xz}$
$t_3 = \bar{y} \left[ \alpha_1 + (1 - \alpha_1) \left( \frac{\bar{x}}{\bar{X}} \right) \right]$ $\times \left[ \alpha_2 + (1 - \alpha_2) \left( \frac{\bar{z}}{\bar{Z}} \right) \right]$	0	1	0	1	1	1	$m_1 m_2 C_{xz} - m_1 C_{yx} - m_2 C_{yz}$
$t_4 = \bar{y} / \left[ \alpha_1 \left( \frac{\bar{X}}{\bar{x}} \right) + (1 - \alpha_1) \left( \frac{\bar{x}}{\bar{X}} \right) \right]$ $\times \left[ \alpha_2 \left( \frac{\bar{Z}}{\bar{z}} \right) + (1 - \alpha_2) \left( \frac{\bar{z}}{\bar{Z}} \right) \right]$	1	1	1	1	-1	-1	$\left( \frac{m_1 - 1}{2} + m_1^2 \right) C_x^2$ $+ \left( \frac{m_2 - 1}{2} + m_2^2 \right) C_z^2$ $-m_1 C_{yx} - m_2 C_{yz} + m_1 m_2 C_{xz}$
$t_5 = \bar{y} / \left[ \alpha_1 \left( \frac{\bar{X}}{\bar{x}} \right) + (1 - \alpha_1) \right]$ $\times \left[ \alpha_2 \left( \frac{\bar{Z}}{\bar{z}} \right) + (1 - \alpha_2) \right]$	1	0	1	0	-1	-1	$m_1 (m_1 + 1) C_x^2 + m_2 (m_2 + 1) C_z^2$ $-m_1 C_{yx} - m_2 C_{yz} + m_1 m_2 C_{xz}$
$t_6 = \bar{y} / \left[ \alpha_1 + (1 - \alpha_1) \left( \frac{\bar{x}}{\bar{X}} \right) \right]$ $\times \left[ \alpha_2 \left( \frac{\bar{Z}}{\bar{z}} \right) + (1 - \alpha_2) \left( \frac{\bar{z}}{\bar{Z}} \right) \right]$	0	1	0	1	-1	-1	$m_1^2 C_x^2 + m_2^2 C_z^2$ $-m_1 C_{yx} - m_2 C_{yz} + m_1 m_2 C_{xz}$
$t_7 = \frac{\alpha_1 \left( \frac{\bar{X}}{\bar{x}} \right) + (1 - \alpha_1) \left( \frac{\bar{x}}{\bar{X}} \right)}{\alpha_2 \left( \frac{\bar{Z}}{\bar{z}} \right) + (1 - \alpha_2) \left( \frac{\bar{z}}{\bar{Z}} \right)}$	1	1	1	1	1	-1	$\frac{m_1 + 1}{2} C_x^2 + \left( \frac{m_2 - 1}{2} + m_2^2 \right) C_z^2$ $-m_1 C_{yx} - m_2 C_{yz} + m_1 m_2 C_{xz}$
$t_8 = \bar{y} \frac{\alpha_1 \left( \frac{\bar{X}}{\bar{x}} \right) + (1 - \alpha_1)}{\alpha_2 \left( \frac{\bar{Z}}{\bar{z}} \right) + (1 - \alpha_2)}$	1	0	1	0	1	-1	$m_1 C_x^2 + m_2 (m_2 + 1) C_z^2$ $-m_1 C_{yx} - m_2 C_{yz} + m_1 m_2 C_{xz}$

**Table 1.** Biases of some modified ratio type estimators (cont.)

Estimator	$g_1$	$h_1$	$g_2$	$h_2$	$\delta_1$	$\delta_2$	Bias $\left(\hat{\bar{Y}}_{gmr}\right)/\theta\bar{Y}$
$t_9 = \bar{y} \frac{\alpha_1 + (1-\alpha_1)\left(\frac{\bar{x}}{\bar{X}}\right)}{\alpha_2 + (1-\alpha_2)\left(\frac{\bar{z}}{\bar{Z}}\right)}$	0	1	0	1	1	-1	$m_2^2 C_z^2 - m_1 C_{yx} - m_2 C_{yz} + m_1 m_2 C_{xz}$

**2.2.Numerical illustrations**

To compare the biases of estimators  $t_1, t_2, \dots$  and  $t_9$  empirically, consider the data on Population-1 and Population-2 referred by Perri (2007) as follows:

Population – 1

The data (Perri, 2007) are taken from the survey of Household Income and Wealth conducted by the Bank of Italy in 2002. The survey covers 8011 Italian households composed of 22148 individuals and 13536 income earners. On the target population comprising of 8011 households three variables –  $y$  (the household net disposable income),  $x$  (household consumption) and  $z$  (the number of household income earners) were observed and the summary statistics are:

$$C_y = 0.787, C_x = 0.668, C_z = 0.4596$$
$$\rho_{yx} = 0.74, \rho_{yz} = 0.458 \text{ and } \rho_{xz} = 0.348$$

Population – 2

The data (Perri, 2007) have been collected by a market research company. The population consists of 2376 points of sale for which three variables are surveyed – the sale area ( $y$ ) in square meters, the number of employees ( $x$ ) and the amount of soft drink sales ( $z$ ) in euro x 1000 in a year. The summary statistics are:

$$C_y = 1.285, C_x = 2.35, C_z = 1.651$$
$$\rho_{yx} = 0.898, \rho_{yz} = 0.861 \text{ and } \rho_{xz} = 0.773$$

The absolute biases of  $t_1, t_2, \dots$  and  $t_9$  without constant multiplier  $\theta\bar{Y}$  are shown in Table 2.

**Table 2.** Absolute biases of estimators without constant multiplier  $\theta\bar{Y}$

Estimator	Absolute bias without constant multiplier	
	Population – 1	Population – 2
$t_1$	0.20862	4.29576
$t_2$	0.09479	1.48338

**Table 2.** Absolute biases of estimators without constant multiplier  $\theta\bar{Y}$  (cont.)

Estimator	Absolute bias without constant multiplier	
	Population – 1	Population – 2
$t_3$	0.33501	1.14016
$t_4$	0.14630	3.11798
$t_5$	0.39732	2.31794
$t_6$	0.3248	0.30560
$t_7$	0.02961	1.85373
$t_8$	0.12702	1.76716
$t_9$	0.30278	0.85639

Comments: The computations show that  $t_7$  for Population -1 and  $t_6$  for Population-2 are least biased.

The asymptotic optimal mean square errors of  $t_1$ ,  $t_2$ , and  $t_9$  to  $0\left(\frac{1}{n}\right)$  are the same.

For Population-1:  $MSE(t_1) = MSE(t_2) = \dots = MSE(t_9) = \theta\bar{Y}^2 [0.367490]$

For Population-2:  $MSE(t_1) = MSE(t_2) = \dots = MSE(t_9) = \theta\bar{Y}^2 [1.445764]$

**3. Difference-cum-ratio estimators**

Consider a difference-cum-ratio estimator defined by

$$T_1 = \left[ \bar{y} + \lambda (\bar{X} - \bar{x}) \right] \left( \frac{\bar{Z}}{\bar{z}} \right)$$

where  $\lambda$  is a real constant or a random variable converging in probability to a constant.  $\lambda$  may be selected in an optimum manner by minimizing the mean square error of  $T_1$  with respect to  $\lambda$ .

Now,  $T_1$  may be expanded in a power series with assumption that  $\left| \frac{\bar{z} - \bar{Z}}{\bar{Z}} \right| < 1$  for all the  ${}^N C_n$  samples. In order to derive the bias and mean square error of  $T_1$  to  $0\left(\frac{1}{n}\right)$ , we retain terms up to and including second degree of the concerned variables and thus

$$T_1 \cong \bar{Y} \left[ 1 - e_3 + e_3^2 + e_1 - e_1 e_3 - \frac{\lambda}{R} e_2 + \frac{\lambda}{R} e_2 e_3 \right],$$



where  $e_1 = \frac{\bar{y} - \bar{Y}}{\bar{Y}}$ ,  $e_2 = \frac{\bar{x} - \bar{X}}{\bar{X}}$ ,  $e_3 = \frac{\bar{z} - \bar{Z}}{\bar{Z}}$  and  $R = \frac{\bar{Y}}{\bar{x}}$

$$E(T_1) = \theta \bar{Y} \left[ C_z^2 + \left( \frac{\lambda}{R} \right)^2 C_x^2 - C_{yz} + \frac{\lambda}{R} C_{xz} \right] + 0 \left( \frac{1}{n^2} \right).$$

Thus,  $T_1$  is a biased estimator of  $\bar{Y}$ , but the bias decreases with increase in sample size.

The mean square error of  $T_1$  to  $0 \left( \frac{1}{n} \right)$  is given by

$$MSE(T_1) = \theta \bar{Y}^2 \left[ C_y^2 + \left( \frac{\lambda}{R} \right)^2 C_x^2 + C_z^2 - 2 \left( \frac{\lambda}{R} \right) C_{yx} - C_{yz} + \frac{2\lambda}{R} C_{xz} \right]$$

Minimizing  $MSE(T_1)$  with respect to  $\lambda$ , the optimum value of  $\lambda$  is given by

$$\begin{aligned} \lambda_{opt} &= R \frac{C_{yx} - C_{xz}}{C_x^2} \\ &= \beta_{yx} - \beta_{xz} \left( \frac{\bar{Y}}{\bar{Z}} \right), \beta_{yx} \text{ and } \beta_{xz} \text{ being the} \end{aligned}$$

population regression coefficients of  $y$  on  $x$  and  $x$  on  $z$  respectively.

Substituting the optimum value of  $\lambda$  in the expression for  $MSE(T_1)$ , the optimum mean square error of  $T_1$  to  $0 \left( \frac{1}{n} \right)$  reduces to

$$MSE(T_1)_{opt} = \theta \bar{Y}^2 \left[ (C_y^2 + C_z^2 - 2C_{yz}) - (\rho_{yx} C_y - \rho_{xz} C_z)^2 \right]$$

Further, the bias of the optimum estimator to  $0 \left( \frac{1}{n} \right)$  is given by

$$\text{Bias}(T_1)_{opt} = \theta \bar{Y} \left[ C_z^2 (1 - \rho_{xz}^2) - C_y C_z (\rho_{yz} - \rho_{yx} + \rho_{xz}) \right]$$

In practice, a consistent estimator of  $\lambda_{opt}$  may be substituted in place of  $\lambda$  in  $T_1$ .

Alternatively, let us consider the regression-cum-ratio estimator.

$$T_1^* = \left[ \bar{y} + b_{yx} (\bar{X} - \bar{x}) \right] \left( \frac{\bar{Z}}{\bar{z}} \right)$$

This estimator was independently proposed by Mohanty (1967) and Swain (1973).

The large sample mean square error of  $T_1^*$  to  $0\left(\frac{1}{n}\right)$  is given by

$$\begin{aligned} MSE(T_1^*) &= \theta \bar{Y}^2 \left[ (C_y^2 + C_z^2 - 2C_{yz}) - \rho_{yx}^2 C_y^2 + 2\rho_{yx}\rho_{xz} C_y C_z \right] \\ &= \theta \bar{Y}^2 \left[ (C_y^2 + C_z^2 - 2C_{yz}) - (\rho_{yx} C_y - \rho_{xz} C_z)^2 + \rho_{xz}^2 C_z^2 \right] \end{aligned}$$

$$\text{Now, } MSE(T_1^*) - MSE(T_1) = \theta \bar{Y}^2 \rho_{xz}^2 C_z^2 \geq 0$$

Therefore,  $T_1^*$  is less efficient than  $T_1$  in large samples.

Consider a class of difference-cum-ratio estimators defined by

$$T_2 = \left[ \bar{y} + \lambda (\bar{X} - \bar{x}) \right] \left( \frac{\bar{Z}}{\bar{z}} \right)^\alpha,$$

where  $\alpha$  and  $\lambda$  are suitably chosen constants and the optimum  $\alpha$  and  $\lambda$  may be obtained by minimizing the approximate mean square error of  $T_2$  with respect to  $\alpha$  and  $\lambda$ .

Following the usual procedure of obtaining the expected values and mean square errors of non-linear estimators to  $0\left(\frac{1}{n}\right)$ , we have

$$\begin{aligned} E(T_2) &= \bar{Y} + \theta \bar{Y} \left[ \frac{\alpha(\alpha+1)}{2} C_z^2 - \alpha C_{yz} + \lambda \alpha \left( \frac{\bar{X}}{\bar{Y}} \right) C_{xz} \right] \\ MSE(T_2) &= \theta \bar{Y}^2 \left[ C_y^2 - \alpha^2 C_z^2 + \left( \frac{\lambda}{R} \right)^2 C_x^2 - 2\alpha C_{yz} - 2\frac{\lambda}{R} C_{yx} + 2\frac{\alpha\lambda}{R} C_{xz} \right] \end{aligned}$$

Minimizing  $MSE(T_2)$  with respect to  $\alpha$  and  $\lambda$ , we have

$$\begin{aligned} \alpha_{opt} &= \frac{C_y}{C_z} \left[ \frac{\rho_{yz} - \rho_{yx}\rho_{xz}}{1 - \rho_{xz}^2} \right] \\ \lambda_{opt} &= \frac{S_y}{S_x} \left[ \frac{\rho_{yx} - \rho_{yz}\rho_{xz}}{1 - \rho_{xz}^2} \right] \end{aligned}$$

$$\text{As such, } MSE(T_2)_{opt} = \theta \bar{Y}^2 C_y^2 (1 - R_{y.xz}^2).$$

Considering an alternative to  $T_2$ , replacing  $\lambda$  by  $b_{yx}$  the sample regression coefficient of  $y$  or  $x$ , we have

$$T_2^* = \left[ \bar{y} + b_{yx} (\bar{X} - \bar{x}) \right] \left( \frac{\bar{Z}}{\bar{z}} \right)^\alpha,$$

suggested by Khare and Srivastava (1981).

Now,

$$MSE(T_2^*) = \theta \bar{Y}^2 \left[ C_y^2 + \alpha^2 C_z^2 + \frac{\beta_{yx}^2}{R^2} C_x^2 - 2\alpha C_{yz} - 2\frac{\beta_{yx}}{R} C_{yx} + 2\alpha \frac{\beta_{yx}}{R} C_{xz} \right] + 0 \left( \frac{1}{n^2} \right)$$

Minimizing  $MSE(T_2^*)$  to  $0 \left( \frac{1}{n} \right)$  with respect to  $\alpha$ , we have

$$\alpha_{opt} = \frac{C_y}{C_z} (\rho_{yz} - \rho_{yx} \rho_{xz})$$

$$\begin{aligned} \text{Thus, } MSE(T_2^*) &= \theta \bar{Y}^2 C_y^2 \left[ (1 - \rho_{yx}^2) - (\rho_{yz} - \rho_{yx} \rho_{xz})^2 \right] \\ MSE(T_2^*) - MSE(T_2) &= \theta \bar{Y}^2 C_y^2 \left[ \frac{\rho_{xz}^2 (\rho_{yz} - \rho_{yx} \rho_{xz})^2}{(1 - \rho_{xz}^2)} \right] \geq 0 \end{aligned}$$

This shows that  $T_2^*$  is less efficient than  $T_2$ .

#### 4. A generalized class of difference-cum-ratio estimator

Define a generalized class of difference-cum-ratio estimator as

$$T_g = \left[ \bar{y} + \lambda (\bar{X} - \bar{x}) \right] \left[ \alpha \left( \frac{\bar{Z}}{\bar{z}} \right)^g + (1 - \alpha) \left( \frac{\bar{z}}{\bar{Z}} \right)^h \right]^\delta$$

where  $\lambda, \alpha, g, h$ , and  $\delta$  are real constants to be determined suitably and  $0 < \alpha < 1$ .

Considering only first degree terms in  $e_1, e_2$  and  $e_3$ ,  $T_g$  may be linearised as

$$T_g \cong e_1 + \delta \{ (1 - \alpha) h - \alpha g \} e_3 - \lambda \frac{\bar{X}}{\bar{Y}} e_2$$

$$\begin{aligned} MSE(T_g) &= \theta \bar{Y}^2 \left[ C_y^2 + \delta^2 \{ (1 - \alpha) h - \alpha g \}^2 C_z^2 + \lambda^2 \frac{\bar{X}^2}{\bar{Y}^2} C_x^2 \right. \\ &\quad \left. + 2\delta \{ (1 - \alpha) h - \alpha g \} C_{yz} - 2\lambda \frac{\bar{X}}{\bar{Y}} C_{yx} - 2\lambda \frac{\bar{X}}{\bar{Y}} \delta \{ (1 - \alpha) h - \alpha g \} C_{xz} \right] \end{aligned}$$

Minimizing  $MSE(T_g)$  with respect to  $\alpha$  and  $\lambda$ , we have

$$\begin{aligned}\alpha_{opt} &= \frac{1}{\delta(g+h)} \left[ \frac{(C_z^2 C_x^2 - C_{xz}^2) \delta h + (C_{yz} C_x^2 - C_{yx} C_{xz})}{C_z^2 C_x^2 - C_{xz}^2} \right] \\ &= \frac{h}{g+h} + \frac{1}{\delta(g+h)} \left[ \frac{C_{yz} C_x^2 - C_{yx} C_{xz}}{C_z^2 C_x^2 - C_{xz}^2} \right] \\ \lambda_{opt} &= \frac{\bar{Y}}{\bar{X}} \left[ \frac{C_z^2 C_{yx} - C_{yz} C_{xz}}{(C_z^2 C_x^2 - C_{xz}^2)} \right]\end{aligned}$$

$$\text{Thus, } MSE(T_g)_{opt} = \theta \bar{Y}^2 C_y^2 (1 - R_{y.xz}^2)$$

$$\begin{aligned}\text{Also, Bias}(T_g)_{opt} &= \theta \bar{Y} \left[ \frac{\delta h(h-1)}{2} C_z^2 + \delta \left\{ \frac{h}{g+h} + \frac{1}{\delta(g+h)} m_2 \right\} \left\{ \frac{g(g+1)}{2} - \frac{h(h-1)}{2} \right\} C_z^2 \right. \\ &\quad \left. + \frac{\delta(\delta-1)}{2} \frac{m_2^2}{\delta^2} C_z^2 - m_2 C_{yz} + m_1 m_2 C_{xz} \right]\end{aligned}$$

$$\text{where } m_1 = \frac{C_{yx} C_z^2 - C_{yz} C_{xz}}{C_z^2 C_x^2 - C_{xz}^2} \text{ and } m_2 = \frac{C_{yz} C_x^2 - C_{yx} C_{xz}}{C_z^2 C_x^2 - C_{xz}^2}$$

Let us consider an alternative class of estimators when  $\lambda$  is substituted by the  $b_{yx}$ , the sample regression estimate, may be defined as

$$T_g^* = [\bar{y} + b_{yx}(\bar{X} - \bar{x})] \left[ \alpha \left( \frac{\bar{Z}}{\bar{z}} \right)^g + (1-\alpha) \left( \frac{\bar{z}}{\bar{Z}} \right)^h \right]^\delta$$

Following usual procedure of finding an approximate mean square error of  $T_g^*$ , it may be seen that to terms of  $O\left(\frac{1}{n}\right)$ ,

$$\begin{aligned}MSE(T_g^*) &= \theta \bar{Y}^2 \left[ C_y^2 + \delta^2 \{(1-\alpha)h - \alpha g\}^2 C_z^2 \right. \\ &\quad \left. + \left( \frac{\beta_{yx}}{R} \right)^2 C_x^2 + 2\delta \{(1-\alpha)h - \alpha g\} C_{yz} \right. \\ &\quad \left. - 2 \left( \frac{\beta_{yx}}{R} \right) C_{yx} - 2 \frac{\beta_{yx}}{R} \delta \{(1-\alpha)h - \alpha g\} C_{xz} \right]\end{aligned}$$

Minimizing the  $MSE(T_g^*)$  with respect to  $\alpha$  , we have

$$\alpha_{opt} = \frac{h}{g+h} + \frac{1}{\delta(g+h)} \cdot \frac{C_{yz}C_x^2 - C_{yx}C_{xz}}{C_z^2C_x^2}$$

$$MSE(T_g^*)_{opt} = \theta \bar{Y}^2 C_y^2 \left[ (1 - \rho_{yx}^2) - (\rho_{yx}\rho_{xz} - \rho_{yz})^2 \right]$$

As  $MSE(T_g^*) > MSE(T_g)$  ,  $T_g$  is more efficient than  $T_g^*$

4.1. Some special cases of  $T_g$

In the following Table 3, we compare the biases of some special cases of  $T_g$ .

**Table 3.** Biases of some special cases of  $T_g$

Estimator	g	h	$\delta$	Bias
$T_1 = \left[ \bar{y} + \lambda (\bar{X} - \bar{x}) \right]$ $\times \left[ \alpha \left( \frac{\bar{Z}}{\bar{z}} \right) + (1-\alpha) \left( \frac{\bar{x}}{\bar{z}} \right) \right]$	1	1	1	$\theta \bar{Y} \left[ \frac{m_2+1}{2} C_z^2 - m_2 C_{yz} + m_1 m_2 C_{xz} \right]$
$T_2 = \left[ \bar{y} + \lambda (\bar{X} - \bar{x}) \right]$ $\times \left[ \alpha \left( \frac{\bar{Z}}{\bar{z}} \right) + (1-\alpha) \right]$	1	0	1	$\theta \bar{Y} \left[ m_2 C_z^2 - m_2 C_{yz} + m_1 m_2 C_{xz} \right]$
$T_3 = \left[ \bar{y} + \lambda (\bar{X} - \bar{x}) \right]$ $\times \left[ \alpha + (1-\alpha) \left( \frac{\bar{z}}{\bar{Z}} \right) \right]$	0	1	1	$\theta \bar{Y} \left[ m_1 m_2 C_{xz} - m_2 C_{yz} \right]$
$T_4 = \frac{\bar{y} + \lambda (\bar{X} - \bar{x})}{\alpha \left( \frac{\bar{Z}}{\bar{z}} \right) + (1-\alpha) \left( \frac{\bar{z}}{\bar{Z}} \right)}$	1	1	-1	$\theta \bar{Y} \left[ m_2^2 C_z^2 + \frac{m_2-1}{2} C_z^2 \right.$ $\left. + m_1 m_2 C_{xz} - m_2 C_{yz} \right]$
$T_5 = \frac{\bar{y} + \lambda (\bar{X} - \bar{x})}{\alpha \left( \frac{\bar{Z}}{\bar{z}} \right) + (1-\alpha)}$	1	0	-1	$\theta \bar{Y} \left[ m_2 (m_2 + 1) C_z^2 - m_2 C_{yz} + m_1 m_2 C_{xz} \right]$
$T_6 = \frac{\bar{y} + \lambda (\bar{X} - \bar{x})}{\alpha + (1-\alpha) \left( \frac{\bar{z}}{\bar{Z}} \right)}$	0	1	-1	$\theta \bar{Y} \left[ m_2^2 C_z^2 + m_1 m_2 C_{xz} - m_2 C_{yz} \right] = 0$

## 4.2. Numerical illustrations

For the Population-1 and Population-2 considered in section 2.2, the absolute biases without constant multiplier are compared in Table 4.

**Table 4.** Comparison of absolute biases of estimators  $T_1 - T_6$ .

Estimator	Absolute bias without constant multiplier	
	Population – 1	Population – 2
$T_1$	0.11464	1.51887
$T_2$	0.39060	0.32265
$T_3$	0.05028	0.59572
$T_4$	0.06436	0.92315
$T_5$	0.08251	0.87949
$T_6$	0	0

## 5. An alternative class of difference-cum-ratio estimator

Consider a generalized class of estimators suggested by Tripathi (1970, 80)

$$T_{g1} = \frac{\bar{y} + \lambda_1 (\bar{X} - \bar{x})}{\bar{z} + \lambda_2 (\bar{X} - \bar{x})} \bar{Z}$$

Following usual procedure of obtaining an approximate mean square error of  $T_{g1}$  to  $O\left(\frac{1}{n}\right)$ , we have

$$\begin{aligned} MSE(T_{g1}) = & \theta \bar{Y}^2 \left[ C_y^2 + \left( \frac{\lambda_2}{R_1} - \frac{\lambda_1}{R} \right)^2 C_x^2 + C_z^2 \right. \\ & \left. + 2 \left\{ \left( \frac{\lambda_2}{R_1} \right) - \left( \frac{\lambda_1}{R} \right) \right\} C_{yx} - 2 C_{yz} - 2 \left\{ \left( \frac{\lambda_2}{R_1} \right) - \left( \frac{\lambda_1}{R} \right) \right\} C_{xz} \right], \end{aligned}$$

where  $R_1 = \frac{\bar{Z}}{\bar{X}}$ .

Minimizing  $MSE(T_{g1})$  with respect to  $\lambda_1$  and  $\lambda_2$  it may be verified that the minimizing equations are not independent and hence  $\lambda_1$  and  $\lambda_2$  cannot be solved uniquely. Therefore, fixing  $\lambda_1$  (or  $\lambda_2$ ) to a suitable real constant, we may solve for  $\lambda_2$  (or  $\lambda_1$ ). Thus, the optimum value for  $\lambda_2$  in terms of  $\lambda_1$  is given by

$$\lambda_2 = \frac{R_1}{C_x^2} (C_{xz} - C_{yx}) + \lambda_1 \left( \frac{R_1}{R} \right)$$

The optimum asymptotic mean square error is given by

$$MSE(T_{g1})_{opt} = \theta \bar{Y}^2 \left[ (C_y^2 + C_z^2 - 2C_{yz}) - (\rho_{xz}C_z - \rho_{yx}C_y)^2 \right]$$

This shows that there is reduction in the mean square error for  $T_{g1}$  by the difference type of adjustment made for  $\bar{y}$  and  $\bar{z}$ , using the second auxiliary variable  $x$ . It may be mentioned here that the optimum value of  $\lambda_2$  in  $T_{g1}$  is not unique because of the presence of free parameter  $\lambda_1$ . A meaningful optimum estimator is obtained by the choice of  $\lambda_1 = \beta_{yx}$ , which results in obtaining  $\lambda_2 = \beta_{zx}$ . Thus, one of the optimum estimators may be obtained as

$$T_{g2} = \frac{\bar{y} + \beta_{yx}(\bar{X} - \bar{x})}{\bar{z} + \beta_{zx}(\bar{X} - \bar{x})} \bar{Z}$$

As  $\beta_{yx}$  and  $\beta_{zx}$  are unknown in practice, we may substitute their consistent sample estimates  $b_{yx}$  and  $b_{zx}$  respectively in  $T_{g2}$ . This reduced estimator was proposed by Sahoo (1984). It may be verified that

$$MSE(T_{g1})_{opt} = MSE(T_{g2})$$

Now, consider the generalized estimator, proposed by Khare and Srivastava (1981) as

$$T_{g3} = \frac{\bar{y} + \lambda_1(\bar{X} - \bar{x})}{[\bar{z} + \lambda_2(\bar{X} - \bar{x})]^\alpha} \bar{Z}^\alpha$$

Minimizing the  $MSE(T_{g3})$  to  $0\left(\frac{1}{n}\right)$  with respect to  $\lambda_1, \lambda_2$  and  $\alpha$  we get

$$\lambda_1 = \beta_{yx}$$

$$\lambda_2 = \beta_{zx}$$

and 
$$\alpha = \frac{C_y}{C_z} \frac{\rho_{yz} - \rho_{yx}\rho_{zx}}{1 - \rho_{zx}^2}$$

As such,

$$MSE(T_{g3})_{opt} = \theta \bar{Y}^2 C_y^2 (1 - R_{y \cdot xz}^2)$$

Now,  $MSE(T_{g1}) - MSE(T_{g3})$

$$= \theta \bar{Y}^2 \left[ C_z \sqrt{1 - \rho_{zx}^2} - C_y \frac{\rho_{yz} - \rho_{yx} \rho_{xz}}{\sqrt{1 - \rho_{xz}^2}} \right] \geq 0,$$

showing thereby that  $T_{g3}$  with optimum values of the parameters  $\lambda_1, \lambda_2$  and  $\alpha$  is more efficient than  $T_{g1}$ .

### 5.1. Comparison of mean square errors

In the following the optimal asymptotic mean square errors of

$\hat{Y}_{Reg}, T_1, T_1^*, T_2, T_2^*, T_g, T_g^*, T_{g1}, T_{g2}$  and  $T_{g3}$  are presented and compared.

$$\text{Now, } MSE(\hat{Y}_{Reg}) = \theta \bar{Y}^2 C_y^2 (1 - \rho_{yx}^2)$$

$$MSE(T_1) = \theta \bar{Y}^2 \left[ (C_y^2 + C_z^2 - 2C_{yz}) - (\rho_{yx} C_y - \rho_{xz} C_z)^2 \right]$$

$$MSE(T_1^*) = \theta \bar{Y}^2 \left[ (C_y^2 + C_z^2 - 2C_{yz}) - (\rho_{yx} C_y - \rho_{xz} C_z)^2 + \rho_{xz}^2 C_z^2 \right]$$

$$MSE(T_2) = \theta \bar{Y}^2 C_y^2 (1 - R_{y.xz}^2)$$

$$MSE(T_2^*) = \theta \bar{Y}^2 C_y^2 \left[ (1 - \rho_{yx}^2) - (\rho_{yz} - \rho_{yx} \rho_{xz})^2 \right]$$

$$MSE(T_g) = \theta \bar{Y}^2 C_y^2 (1 - R_{y.xz}^2)$$

$$MSE(T_g^*) = \theta \bar{Y}^2 C_y^2 \left[ (1 - \rho_{yx}^2) - (\rho_{yx} \rho_{xz} - \rho_{yz})^2 \right]$$

$$MSE(T_{g1}) = MSE(T_{g2})$$

$$= \theta \bar{Y}^2 \left[ (C_y^2 + C_z^2 - 2C_{yz}) - (\rho_{yx} C_y - \rho_{xz} C_z)^2 \right]$$

$$MSE(T_{g3}) = \theta \bar{Y}^2 C_y^2 (1 - R_{y.xz}^2)$$

Thus, we find

$$(i) \quad MSE(T_2) = MSE(T_g) = MSE(T_{g3}) = \theta \bar{Y}^2 C_y^2 (1 - R_{y.xz}^2)$$

$$(ii) \quad MSE(T_2^*) = MSE(T_g^*) = \theta \bar{Y}^2 C_y^2 \left[ (1 - \rho_{yx}^2) - (\rho_{yx} \rho_{xz} - \rho_{yz})^2 \right]$$

$$(iii) \quad MSE(T_1) = MSE(T_{g1}) = MSE(T_{g2})$$



$$= \theta \bar{Y}^2 \left[ \left( C_y^2 + C_z^2 - 2C_{yz} \right) - \left( \rho_{yx} C_y - \rho_{xz} C_z \right)^2 \right]$$

(iv)  $MSE(T_1^*) > MSE(T_1) = MSE(T_{g1}) = MSE(T_{g2})$

5.2.Numerical illustrations

In the following Table 5 we compare the percent relative efficiencies of the difference-cum-ratio estimators/ regression-cum-ratio estimators with respect to the linear regression estimator with  $x$  without adjusting for the second auxiliary variable  $z$  , using data on populations referred to in section 2.2. The efficiency is defined as the inverse of the optimal asymptotic mean square error.

Table 5. Comparison of Percent Relative Efficiencies

Estimator	Population – 1		Population – 2	
	$MSE / \theta \bar{Y}^2$	Relative Efficiency	$MSE / \theta \bar{Y}^2$	Relative Efficiency
$\hat{\bar{Y}}_{reg(x)}$	0.28020	100	0.31967	100
$T_1 = T_{g1} = T_{g2}$	0.32083	87	0.70878	45
$T_1^*$	0.34641	81	2.33752	14
$T_2^* = T_g^*$	0.25531	110	0.27370	117
$T_2 = T_g = T_{g3}$	0.25188	111	0.20546	155

Comments: Comparison of efficiencies of the estimators under consideration shows that  $T_2 = T_g = T_{g3}$  and  $T_2^* = T_g^*$  are more efficient than  $T_1 = T_{g1} = T_{g2}$ ,  $T_1^*$  and  $\hat{\bar{Y}}_{reg(x)}$  .

Further, in case of numerical illustrations under consideration, there has been substantial loss in efficiency in using  $T_1 = T_{g1} = T_{g2}$  and  $T_1^*$  in place of  $\hat{\bar{Y}}_{reg(x)}$  using single auxiliary variable  $x$  in a linear regression set up.

## REFERENCES

- BOWLEY, A. L. (1926). Measurements of precision attained in sampling. *Bull. Inst. Internat. Statist.*, 22, 1-62.
- COCHRAN, W. G. (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. *J. Agricultural Sc.*, 30, 262-275.
- COCHRAN, W. G. (1942). Sampling theory when the sampling units are unequal sizes. *J. Amer. Stat. Assoc.*, 37, 199-212.
- COCHRAN, W. G. (1977). *Sampling Techniques*. John Wiley and Sons, New York.
- HANSEN, M. H. and HURWITZ, W. N. (1943). On the theory of sampling from finite populations. *Ann. Math. Stat.*, 14, 333-362.
- HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953). *Sample survey methods and theory*, Vol. 2, John Wiley and Sons, New York.
- HANIF, M., HAMMD, N. and SHAHBAZ, M. Q. (2010). Some new regression type estimators in two phase sampling. *World Applied Sc. Journal*, 8(7), 799-803.
- JOHN, S. (1969). On multivariate ratio and product estimators, *Biometrika*, 56, 533-536.
- KHARE, B. B. and SRIVASTAVA, S. R. (1980). On an efficient estimator of population mean using two auxiliary variables, *Proc. National Acad. Sc. India*, 50(A), 209-214.
- KHARE, B. B. and SRIVASTAVA, S. R. (1981). A generalized regression ratio estimator for the population mean using two auxiliary variables. *Aligarh J. Statist.*, 1, 43-51.
- MOHANTY, S. (1967). Combination of regression and ratio estimate. *J. Ind. Statist.*, 5, 16-19.
- MURTHY, M. N. (1964). Product method of estimation. *Sankhya*, Series A, 26, 294-307.
- NEYMAN, J. (1934). On the two different aspects of representative method : The method of stratified. Sampling and the method of purposive selection. *J. Roy. Statist. Soc.*, 97, 558-606.
- NEYMAN, J. (1938). Contributions to the theory of sampling human populations. *J. Amer. Statist. Assoc.*, 33, 101-116.

- OLKIN, I. (1958). Multivariate ratio estimation for finite populations, *Biometrika*, 45, 154-165.
- PERRI, P. F. (2007). Improved ratio-cum-product type estimators. *Statistics in Transition (New Series)*, 8, 1, 51-69.
- RAJ, DES (1965). On a method of using multi-auxiliary information in sample surveys. *J. Amer Stat. Assoc.*, 60, 270-277.
- RAO, P. S. R. S. and MUDHOLKAR, G. S. (1967). Generalized multivariate estimator for the mean of finite populations. *J. Amer. Stat. Assoc.*, 62, 1009-12.
- ROBSON, D. S. (1957). Applications of multivariate polykeys to the theory of unbiased ratio type estimators. *J. Amer. Stat. Assoc.*, 52, 511-522.
- SAHOO, L. N. (1984). A note on estimation of the population mean using two auxiliary variables. *Aligarh. J. Statist.*, 3 and 4, 63-66.
- SHUKLA, G. K. (1965). Multivariate regression estimate. *J. Ind. Stat. Assoc.*, 3, 202-211.
- SHUKLA, G. K. (1966). An alternative multivariate ratio estimate for finite population. *Bull Cal. Stat. Assoc.*, 15, 127-134.
- SINGH, M. P. (1965). On the estimation of ratio and product of population parameters, *Sankhya, Series C*, 27, 321-328.
- SINGH, M. P. (1967). Ratio-cum-product method of estimation, *Metrika*, 12, 34-43.
- SINGH, M. P. (1967). Multivariate product method of estimation for finite population, *J. Ind. Soc. Agri. Statist.*, 19, 1-10.
- SRIVASTAVA, S. K. (1965). An estimate of the mean of a finite population using several auxiliary character, *J. Ind. Stat. Assoc.*, 3, 189-194.
- SRIVASTAVA, S. K. (1967). An estimator using auxiliary information in sample surveys. *Cal. Stat. Assoc., Bull.*, 16, 121-132.
- SRIVASTAVA, S. K. (1971). Generalized estimator for the mean of a finite population using multi-auxiliary information. *J. Amer. Stat. Assoc.*, 66, 404-407.
- SRIVASTAVA, S. K. and JHAJJ, H. S. (1983). A class of estimators of the population mean using multi-auxiliary information. *Bull. Cal. Stat. Assoc.*, 32, 47-56.
- SWAIN, A. K. P. C. (1973). Some contributions to the theory of sampling, Ph.D. Thesis submitted to the Utkal University, India.

- TRIPATHI, T. P. (1970). Contributions to the sampling theory using multivariate information. Ph.D. Thesis submitted to Punjabi University, Patiala, India.
- TRIPATHI, T. P. (1980). A general class of estimators of population ratio. *Sankhya Series C*, 42, 63-75.
- TRIPATHI, T. P. (1987). A class of estimators for population means using multivariate auxiliary information under any sampling design. *Aligarh. J. Statist.*, 7, 49-62.
- WATSON, D. J. (1937). The estimation of leaf areas. *J. Agricultural Sc.* 27, 474.

## CHAIN RATIO ESTIMATOR FOR THE POPULATION MEAN IN THE PRESENCE OF NON-RESPONSE

B. B. Khare<sup>1</sup>, U. Srivastava<sup>2</sup>, K. Kumar<sup>1</sup>

### ABSTRACT

In this paper we have proposed two chain ratio type estimators for population mean using two auxiliary variables in the presence of non-response. The proposed estimators have been found to be more efficient than the relevant estimators for the fixed values of preliminary sample of size  $n'$  and subsample of size  $n(< n')$  taken from the preliminary sample of size  $n'$  under the specified conditions. The proposed estimators are more efficient than the corresponding estimators for population mean ( $\bar{Y}$ ) of study variable  $y$  in the case of the fixed cost and have less total cost in comparison to the cost incurred by the corresponding relevant estimators for specified variance. The results have been supported by empirical study as well as Monte Carlo simulation study.

**Key words:** chain ratio type estimator, preliminary sample, bias, mean square error, non-response, additional auxiliary variable.

### 1. Introduction

The research work using auxiliary information in proposing selection procedure and the estimators for the population parameters was initiated by Bowely (1926), Neyman (1934, 1938), Watson (1937), Cochran (1940, 1942), Hansen et al. (1953) and Robson (1957). The population mean of the auxiliary variable ( $x$ ) is not known but the population mean of an additional auxiliary variable ( $z$ ) is known which is cheaper and less correlated to the study variable ( $y$ ) in comparison to the main auxiliary variable (i.e.  $\rho_{yx} > \rho_{yz}$ ,  $\rho_{yx}, \rho_{yz} > 0$ ). In such case Chand (1975), Kiregyera (1980, 1984) and Srivastava et al. (1990) proposed chain ratio type estimators using additional variable with known population mean.

Sometimes, it may not be possible to collect the complete information for all the units selected in the sample due to non-response. The missing observations

<sup>1</sup> Department of Statistics, Banaras Hindu University, Varanasi-221005, India.  
E-mail: bbkhare56@yahoo.com.

<sup>2</sup> Statistics Sections M.M.V., Banaras Hindu University, Varanasi-221005, India.

due to non-response may occur during the investigation, which may be at random, and the ignorance of such missing observations may lead to biased estimator, though the amount of the bias may be very negligible. If the missing observation due to non-response is not at random then the amount of bias in the estimator will be large and may increase the error in the estimation, and the sampling error will also increase. Little and Rubin (2002) suggested to ignore the missing data completely if the percentage of incomplete cases are very low. This practice will reduce the sample size and may increase the bias and the variance of the estimator when the incomplete cases are large. However, some imputation techniques to replace the missing observation are considered by Rao and Toutenburge (1995), and Toutenburge and Srivastava (1998, 2003). Estimation of the population mean in sample surveys when some observations are missing due to non-response not at random was considered by Hansen and Hurwitz (1946), Rao (1986, 1987), Khare and Srivastava (1993, 1995).

In this paper, we have proposed two chain type estimators for the population mean of the study variable in the presence of non-response. The expressions for the bias and mean square error of the proposed estimator are obtained and a comparison of the proposed estimators has been made with the relevant estimators.

The optimum values of the preliminary sample ( $n'$ ), subsample ( $n$ ) and the sub-sampling fraction ( $k^{-1}, k > 1$ ) have been obtained for fixed cost  $C \leq C_0$  and for the specified variance  $V = V'_0$ . A comparative study of the proposed estimators with the relevant estimators has been made with the help of an empirical study as well as Monte Carlo simulation study.

## 2. The estimators

Let  $Y_i$ ,  $X_i$  and  $Z_i$  be the non-negative values for the  $i$ th unit of the population  $U = U_1, U_2, \dots, U_N$  on the study variable  $y$ , the auxiliary variable  $x$  and the additional auxiliary variable  $z$  with their population means  $\bar{Y}$ ,  $\bar{X}$  and  $\bar{Z}$ . Here  $\bar{X}$  is unknown, but  $\bar{Z}$ , the population mean of additional auxiliary variable ( $z$ ) (closely related to  $x$ ) is known, which may be cheaper and less correlated to the study variable ( $y$ ) in comparison to the main auxiliary variable ( $x$ ). Now a preliminary sample of size  $n'$  ( $n' < N$ ) is selected from the population of size  $N$  using simple random sampling without replacement (SRSWOR) scheme, and we estimate the population mean  $\bar{X}$  using additional auxiliary variable with known population mean  $\bar{Z}$  and  $n'$  observations on  $x$  and  $z$ . Again, subsample of size  $n$  ( $n < n'$ ) is selected from the preliminary sample of size  $n'$  by using simple random sampling without replacement (SRSWOR) scheme, and it has been observed that  $n_1$  units respond and  $n_2$  unit do not respond in the sample of size

$n$  for the study variable  $y$ . It is also assumed that the population of size  $N$  is composed of  $N_1$  responding and  $N_2$  non-responding units, though they are unknown. Further, a sub sample of size  $r$  ( $r = n_2 k^{-1}, k > 1$ ) from  $n_2$  non-responding units has been drawn by using SRSWOR method of sampling by making extra effort. Hence, we have  $(n_1 + r)$  observations on the  $y$  variable.

Using Hansen and Hurwitz (1946) technique, the estimator for population mean using  $(n_1 + r)$  observations on  $y$  variable is given by

$$\bar{y}^* = \frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} \bar{y}'_2, \quad (2.1)$$

where  $\bar{y}_1$  and  $\bar{y}'_2$  are the sample means of variable  $y$  based on  $n_1$  and  $r$  units respectively. The estimator  $\bar{y}^*$  is unbiased and has variance

$$V(\bar{y}^*) = \frac{f}{n} S_y^2 + \frac{W_2(k-1)}{n} S_{y(2)}^2, \quad (2.2)$$

where  $f = 1 - \frac{n}{N}$ ,  $W_2 = \frac{N_2}{N}$ ,  $S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$  and  $S_{y(2)}^2 = \frac{1}{N_2-1} \sum_{i=1}^{N_2} (Y_{i(2)} - \bar{Y}_{(2)})^2$  are the population mean square of  $y$  for the entire population and for the non-responding part of the population.

Similarly, the estimator  $\bar{x}^*$  for the population mean  $\bar{X}$  in the presence of non-response based on corresponding  $(n_1 + r)$  observations is given by

$$\bar{x}^* = \frac{n_1}{n} \bar{x}_1 + \frac{n_2}{n} \bar{x}'_2, \quad (2.3)$$

where  $\bar{x}_1$  and  $\bar{x}'_2$  are the sample means of variable  $x$  based on  $n_1$  and  $r$  units respectively, we also have

$$V(\bar{x}^*) = \frac{f}{n} S_x^2 + \frac{W_2(k-1)}{n} S_{x(2)}^2, \quad (2.4)$$

where  $S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$  and  $S_{x(2)}^2 = \frac{1}{N_2-1} \sum_{i=1}^{N_2} (X_{i(2)} - \bar{X}_{(2)})^2$  are population mean square of  $x$  for the entire population and non-responding part of the population.

In case when population mean  $\bar{X}$  is not known, then, it is estimated by taking a preliminary sample of size  $n'(n' < N)$  from the population of size  $N$  by using simple random sampling without replacement (SRSWOR) sampling scheme. In this situation when we have incomplete information both on the study variable  $y$  and incomplete/complete information on auxiliary variable  $x$  from the subsample of size  $n$ , Khare and Srivastava (1995) proposed conventional ( $T_1$ ) / alternative ( $T_2$ ) two phase sampling ratio estimator for population mean in the presence of non-response, which are given as follows:

$$T_1 = \frac{\bar{y}^*}{\bar{x}^*} \bar{x}', \quad T_2 = \frac{\bar{y}^*}{\bar{x}} \bar{x}', \quad (2.5)$$

$$\text{where } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{x}' = \frac{1}{n'} \sum_{i=1}^{n'} x_i$$

Now, we propose two chain ratio type estimators for the population mean in the presence of non-response, in the situation when  $\bar{X}$  is unknown, but  $\bar{Z}$  is known and we have incomplete information on study variable  $y$  and the incomplete /complete information on auxiliary variable  $x$ .

If  $\bar{X}$  is not known, but  $\bar{Z}$ , the population mean of the additional auxiliary variable  $z$  (closely related to  $x$ ) is known which may be cheaper and less correlated to the study variable ( $y$ ) in comparison to main auxiliary variable ( $x$ ), i.e.  $\rho_{yx} > \rho_{yz}$  we take in this situation a preliminary sample of size  $n'$  ( $n' < N$ ) from the population of size  $N$  with SRSWOR scheme and estimate the population mean  $\bar{X}$  by using the sample means  $\bar{x}'$  and  $\bar{z}'$  based on  $n'$  units and the known additional population mean  $\bar{Z}$ . We see that  $\hat{\bar{X}}_r = \frac{\bar{x}'}{\bar{z}'} \bar{Z}$  is more precise

than preliminary sample mean  $\bar{x}'$  if  $\rho_{xz} > \frac{1}{2} \frac{C_z}{C_x}$ , where  $\bar{z}' = \frac{1}{n'} \sum_{i=1}^{n'} z_i$ .

Now, we propose conventional ( $t_1$ ) and alternative ( $t_2$ ) chain ratio type estimators for  $\bar{Y}$  using available information on two auxiliary variables  $x$  and  $z$  in the presence of non-response, which are given as follows:

$$t_1 = \frac{\bar{y}^*}{\bar{x}^*} \frac{\bar{x}'}{\bar{z}'} \bar{Z}, \quad t_2 = \frac{\bar{y}^*}{\bar{x}} \frac{\bar{x}'}{\bar{z}'} \bar{Z} \quad (2.6)$$

### 3. Expressions for the Relative Bias (RB) and Approximated Mean Square Error (AMSE) of the proposed estimators

Using the large sample approximations, the expressions for the relative bias and approximated mean square error of the estimators'  $t_1$  and  $t_2$  up to the terms of order ( $n^{-1}$ ) are given by (see appendix)

$$RB(t_1) = RB(T_1) + \frac{f'}{n'} (C_z^2 - \rho_{yz} C_y C_z), \quad (3.1)$$

$$RB(t_2) = RB(T_2) + \frac{f'}{n'} (C_z^2 - \rho_{yz} C_y C_z), \quad (3.2)$$

$$AMSE(t_1) = AMSE(T_1) + \bar{Y}^2 \frac{f'}{n'} (C_z^2 - 2\rho_{yz} C_y C_z), \quad (3.3)$$



and

$$AMSE(t_2) = AMSE(T_2) + \bar{Y}^2 \frac{f'}{n'} (C_z^2 - 2\rho_{yz} C_y C_z), \quad (3.4)$$

where

$$RB(T_1) = \left( \frac{1}{n} - \frac{1}{n'} \right) (C_x^2 - \rho_{yx} C_y C_x) + \frac{W_2(k-1)}{n} (C_{x(2)}^2 - \rho_{yx(2)} C_{y(2)} C_{x(2)}), \quad (3.5)$$

$$RB(T_2) = \left( \frac{1}{n} - \frac{1}{n'} \right) (C_x^2 - \rho_{yx} C_y C_x), \quad (3.6)$$

$$AMSE(T_1) = \bar{Y}^2 \left[ \left( \frac{1}{n} - \frac{1}{n'} \right) (C_y^2 + C_x^2 - 2\rho_{yx} C_y C_x) + \frac{W_2(k-1)}{n} (C_{y(2)}^2 + C_{x(2)}^2 - 2\rho_{yx(2)} C_{y(2)} C_{x(2)}) + \frac{f'}{n'} C_y^2 \right], \quad (3.7)$$

and

$$AMSE(T_2) = \bar{Y}^2 \left[ \left( \frac{1}{n} - \frac{1}{n'} \right) (C_y^2 + C_x^2 - 2\rho_{yx} C_y C_x) + \frac{W_2(k-1)}{n} C_{y(2)}^2 + \frac{f'}{n'} C_y^2 \right]. \quad (3.8)$$

$$C_y = \frac{S_y}{\bar{Y}}, \quad C_x = \frac{S_x}{\bar{X}}, \quad C_z = \frac{S_z}{\bar{Z}}, \quad C_{y(2)} = \frac{S_{y(2)}}{\bar{Y}}, \quad C_{x(2)} = \frac{S_{x(2)}}{\bar{X}}, \quad f' = \left( 1 - \frac{n'}{N} \right),$$

$$S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2, \quad S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2, \quad S_z^2 = \frac{1}{N-1} \sum_{i=1}^N (Z_i - \bar{Z})^2,$$

$$S_{y(2)}^2 = \frac{1}{N_2-1} \sum_{i=1}^{N_2} (Y_{i(2)} - \bar{Y}_{(2)})^2 \text{ and } S_{x(2)}^2 = \frac{1}{N_2-1} \sum_{i=1}^{N_2} (X_{i(2)} - \bar{X}_{(2)})^2.$$

#### 4. The estimators of $RB(t_1)$ , $RB(t_2)$ , $AMSE(t_1)$ and $AMSE(t_2)$

$$R\hat{B}(t_1) = R\hat{B}(T_1) + \frac{f'}{n'} (c_z^2 - c_{yz}), \quad (4.1)$$

$$R\hat{B}(t_2) = R\hat{B}(T_2) + \frac{f'}{n'} (c_z^2 - c_{yz}), \quad (4.2)$$

$$AM\hat{S}E(t_1) = AM\hat{S}E(T_1) + \bar{y} \frac{f'}{n'} (c_z^2 - 2c_{yz}) \quad (4.3)$$

$$\text{and } AM\hat{S}E(t_2) = AM\hat{S}E(T_2) + \bar{y} \frac{f'}{n'} (c_z^2 - 2c_{yz}), \quad (4.4)$$

where

$$R\hat{B}(T_1) = \left( \frac{1}{n} - \frac{1}{n'} \right) (c_x^2 - c_{yx}) + \frac{W_2(k-1)}{n} (c_{x(2)}^2 - c_{yx(2)}), \quad (4.5)$$

$$RB(\hat{T}_2) = \left( \frac{1}{n} - \frac{1}{n'} \right) (c_x^2 - c_{yx}), \quad (4.6)$$

$$AM\hat{SE}(T_1) = \bar{y}^2 \left[ \left( \frac{1}{n} - \frac{1}{n'} \right) (c_y^2 + c_x^2 - 2c_{yx}) + \frac{W_2(k-1)}{n} (c_{y(2)}^2 + c_{x(2)}^2 - 2c_{yx(2)}) + \frac{f'}{n'} c_y^2 \right] \quad (4.7)$$

$$AM\hat{SE}(T_2) = \bar{y}^2 \left[ \left( \frac{1}{n} - \frac{1}{n'} \right) (c_y^2 + c_x^2 - 2c_{yx}) + \frac{W_2(k-1)}{n} c_{y(2)}^2 + \frac{f'}{n'} c_y^2 \right] \quad (4.8)$$

$$c_y = \frac{s_y}{\bar{y}}, \quad c_x = \frac{s_x}{\bar{x}}, \quad c_z = \frac{s_z}{\bar{z}}, \quad c_{yx} = \frac{s_{yx}}{\bar{y}\bar{x}}, \quad c_{yz} = \frac{s_{yz}}{\bar{y}\bar{z}}, \quad c_{y(2)} = \frac{s_{y(2)}}{\bar{y}}, \quad c_{x(2)} = \frac{s_{x(2)}}{\bar{x}},$$

$$c_{yx(2)} = \frac{s_{yx(2)}}{\bar{y}\bar{x}}, \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2, \quad s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad s_z^2 = \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2,$$

$$s_{yx} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}), \quad s_{yz} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z}), \quad s_{y(2)}^2 = \frac{1}{r-1} \sum_{i=1}^r (y'_{i2} - \bar{y}'_2)^2,$$

$$s_{x(2)}^2 = \frac{1}{r-1} \sum_{i=1}^r (x'_{i2} - \bar{x}'_2)^2 \text{ and } s_{yx(2)} = \frac{1}{r-1} \sum_{i=1}^r (y'_{i2} - \bar{y}'_2)(x'_{i2} - \bar{x}'_2),$$

where  $(y_i, x_i, z_i)$  are the values of the  $i^{th}$  unit of the sample of size  $n$  for variables  $y, x$  and  $z$  respectively.  $(y'_{i2}, x'_{i2})$  are the values of  $i^{th}$  unit in the subsample of size  $r$  drawn from  $n_2$  non-responding units for the variables  $y$  and  $x$  respectively by using SRSWOR scheme of sampling.

The proposed estimators of  $RB(t_1)$ ,  $RB(t_2)$ ,  $AMSE(t_1)$  and  $AMSE(t_2)$  given by (4.1) to (4.4) are almost unbiased up to terms of order  $(n^{-1})$  and the  $AMSE$  of these proposed estimators will be of order  $(n^{-2})$  and will be dependent upon the values of higher order moment ( $> \text{order } 2$ ) involved in it.

## 5. Comparison of the proposed estimators $t_1$ and $t_2$ with respect to

$\bar{y}^*, T_1$  and  $T_2$

Comparing the estimator  $t_1$  and  $t_2$  with  $\bar{y}^*$ ,  $T_1$  and  $T_2$  in terms of precision, we see

$$AMSE(t_1) < AMSE(\bar{y}^*) \text{ if } \rho_{yz} > \frac{1}{2} \frac{C_z}{C_y}, \quad \rho_{yx} > \frac{1}{2} \frac{C_x}{C_y} \text{ and } \rho_{yx(2)} > \frac{1}{2} \frac{C_{x(2)}}{C_{y(2)}} \quad (5.1)$$

$$AMSE(t_2) < AMSE(\bar{y}^*) \text{ if } \rho_{yz} > \frac{1}{2} \frac{C_z}{C_y}, \quad \rho_{yx} > \frac{1}{2} \frac{C_x}{C_y}. \quad (5.2)$$

Under the condition (5.1), we have

$$AMSE(t_1) < AMSE(T_1) \text{ if } \rho_{yz} > \frac{1}{2} \frac{C_z}{C_y}. \quad (5.3)$$

Under the condition (5.2), we have

$$AMSE(t_2) < AMSE(T_2) \text{ if } \rho_{yz} > \frac{1}{2} \frac{C_z}{C_y}. \quad (5.4)$$

## 6. Determination of $n'$ , $n$ and $k$ (for the fixed cost $C \leq C_0$ )

Let  $C_0$  be the total cost (fixed) of the survey apart from overhead cost.

The cost function  $C'$  can be written as:

$$C' = (c'_1 + c'_2)n' + c_1n + c_2n_1 + c_3 \frac{n_2}{k}, \quad (6.1)$$

since  $C'$  vary from sample to sample, so the expected cost can be written as:

$$C = E(C') = (c'_1 + c'_2)n' + n(c_1 + c_2W_1 + c_3 \frac{W_2}{k}), \quad (6.2)$$

where

$c'_1$  - the cost per unit of obtaining information on auxiliary variable  $x$ ,

$c'_2$  - the cost per unit of obtaining information on additional auxiliary variable  $z$ ,

$c_1$  - the cost per unit of mailing questionnaire/visiting the unit at the subsample,

$c_2$  - the cost per unit of collecting, processing data obtained from  $n_1$  responding units,

$c_3$  - the cost per unit of obtaining and processing data (after extra efforts) for the subsampling units and  $W_1 = N_1 / N$ , response rate in the population.

Using the  $AMSE(t_1)$  and  $AMSE(t_2)$  from (3.3) and (3.4), the expression for  $AMSE(t_i)$ ,  $i=1, 2$  can be written in term of notations  $V_{0i}, V_{1i}, V_{2i}$  which is given as:

$$AMSE(t_i) = \frac{V_{0i}}{n} + \frac{V_{1i}}{n'} + \frac{k V_{2i}}{n} - \frac{1}{N} (S_y^2 + R^2 S_z^2 - 2R S_{yz}), \quad (6.3)$$

where  $V_{0i}, V_{1i}$  and  $V_{2i}$  are respectively the coefficients of the terms of  $n^{-1}, n'^{-1}$ , and  $k n^{-1}$  in the expression of  $AMSE(t_i)$  and  $R = \bar{Y} / \bar{Z}$ . Now, for minimizing the  $AMSE(t_i)$  for the fixed cost  $C \leq C_0$  and to obtain the optimum values of  $n'$ ,  $n$  and  $k$ , we define a function  $\phi$  as:

$$\phi = AMSE(t_i) + \lambda_i \left\{ (c'_1 + c'_2)n' + n(c_1 + c_2W_1 + c_3 \frac{W_2}{k}) - C_0 \right\}, i=1, 2, \quad (6.4)$$

where  $\lambda_i$  is the Lagrange's multiplier. Now, differentiating  $\phi$  with respect to  $n'$ ,  $n$  and  $k$  and equating to zero, we have

$$n' = \sqrt{\frac{V_{1i}}{\lambda_i(c'_1 + c'_2)}}, \quad n = \sqrt{\frac{V_{0i} + k V_{2i}}{\lambda_i(c_1 + c_2W_1 + c_3 \frac{W_2}{k})}} \quad (6.5)$$

$$\text{and } k_{opt} = \sqrt{\frac{V_{0i}c_3W_2}{V_{2i}(c_1 + c_2W_1)}}. \quad (6.6)$$

Now, putting the values of  $n'$ ,  $n$  from (6.5) in (6.2) and using  $k_{opt}$  given by (6.6), we have

$$\sqrt{\lambda_i} = \frac{1}{C_0} \left[ \sqrt{V_{1i}(c'_1 + c'_2)} + \sqrt{(V_{0i} + k_{opt}V_{2i})(c_1 + c_2W_1 + c_3 \frac{W_2}{k_{opt}})} \right], i=1, 2. \quad (6.7)$$

It has been observed that the determinant of the matrix of second order derivative of  $\phi$  with respect to  $n'$ ,  $n$  and  $k$  is positive for the optimum values of  $n'$ ,  $n$  and  $k$ , which shows that the solutions for  $n'$ ,  $n$  given by (6.5) using (6.6) and (6.7), and the optimum of  $k$  under the condition  $C \leq C_0$  minimize the mean square error of  $t_i$ . It is also important to note here that the subsampling fraction ( $k_{opt}^{-1}$ ) will decrease as  $\sqrt{c_3 / (c_1 + c_2W_1)}$  increases. The minimum value of  $AMSE(t_i)$  can be obtained by putting the optimum values of  $n'$ ,  $n$  and  $k$  from (6.5), and (6.6) in the expression (6.3), so we have

$$AMSE(t_i)_{\min} = \frac{1}{C_0} \left[ \sqrt{V_{1i}(c'_1 + c'_2)} + \sqrt{(V_{0i} + k_{opt}V_{2i})(c_1 + c_2W_1 + c_3 \frac{W_2}{k_{opt}})} \right]^2 - \frac{1}{N} (S_y^2 + R^2 S_z^2 - 2RS_{yz}). \quad (6.8)$$

Now, neglecting the term of  $O(N^{-1})$ , we have

$$AMSE(t_i)_{\min} = \frac{1}{C_0} \left[ \sqrt{V_{1i}(c'_1 + c'_2)} + \sqrt{(V_{0i} + k_{opt}V_{2i})(c_1 + c_2W_1 + c_3 \frac{W_2}{k_{opt}})} \right]^2. \quad (6.9)$$

Now, putting the value of  $k_{opt}$  from (6.6) in (6.9), we have

$$AMSE(t_i)_{\min} = \frac{1}{C_0} \left[ \sqrt{V_{1i}(c'_1 + c'_2)} + \sqrt{V_{0i}(c_1 + c_2W_1)} + \sqrt{c_3W_2V_{2i}} \right]^2 \quad (6.10)$$

In case of  $\bar{y}^*$ , the expected total cost is given by

$$C = E(C') = n(c_1 + c_2 W_1 + c_3 \frac{W_2}{k}) \quad (6.11)$$

For the fixed cost  $C_0$ , the expression for  $AMSE(\bar{y}^*)_{\min}$  will be given by

$$AMSE(\bar{y}^*)_{\min} = \frac{1}{C_0} \left[ \sqrt{V_0(c_1 + c_2 W_1)} + \sqrt{c_3 W_2 V_2} \right]^2 - \frac{S_y^2}{N} \quad (6.12)$$

Now, neglecting the term of  $O(N^{-1})$ , we have

$$AMSE(\bar{y}^*)_{\min} = \frac{1}{C_0} \left[ \sqrt{V_0(c_1 + c_2 W_1)} + \sqrt{c_3 W_2 V_2} \right]^2, \quad (6.13)$$

where  $V_0 = S_y^2 - W_2 S_{y(2)}^2$  and  $V_2 = W_2 S_{y(2)}^2$  are the coefficients of the terms of  $n^{-1}$  and  $k n^{-1}$  in the expression

$$V(\bar{y}^*) = \frac{f}{n} S_y^2 + \frac{W_2(k-1)}{n} S_{y(2)}^2. \quad (6.14)$$

For the fixed cost  $C_0$ , the expression for  $AMSE(T_i)_{\min}$  will be given by

$$AMSE(T_i)_{\min} = \frac{1}{C_0} \left[ \sqrt{c'_1 U_{1i}} + \sqrt{U_{0i}(c_1 + c_2 W_1)} + \sqrt{c_3 W_2 U_{2i}} \right]^2 - \frac{S_y^2}{N} \quad i=1, 2. \quad (6.15)$$

Now, neglecting the term of  $O(N^{-1})$ , we have

$$AMSE(T_i)_{\min} = \frac{1}{C_0} \left[ \sqrt{c'_1 U_{1i}} + \sqrt{U_{0i}(c_1 + c_2 W_1)} + \sqrt{c_3 W_2 U_{2i}} \right]^2, \quad (6.16)$$

where  $U_{0i}$ ,  $U_{1i}$  and  $U_{2i}$  are the coefficients of the terms of  $n^{-1}$ ,  $n'^{-1}$  and  $k n^{-1}$  in the expressions of  $AMSE(T_i)$ ,  $i=1, 2$ .

It is important to obtain the values of  $c'_1$  for which the estimator  $T_i$  will be more efficient than  $\bar{y}^*$ , i. e.  $AMSE(T_i)_{\min} < AMSE(\bar{y}^*)_{\min}$  is given by

$$c'_1 < \frac{1}{U_{1i}} \left[ \sqrt{(c_1 + c_2 W_1)}(\sqrt{V_0} - \sqrt{U_{0i}}) + \sqrt{c_3 W_2}(\sqrt{V_2} - \sqrt{U_{2i}}) \right]^2. \quad (6.17)$$

For the fixed cost  $C_0$ , the value of  $c'_2$  for which the estimator  $t_i$  will be more efficient than  $T_i$ , is given by  $\frac{c'_2}{c'_1} < \frac{(U_{1i} - V_{1i})}{V_{1i}}$ .

Sometimes it may be required to obtain an estimator with the desired degree of precision with minimum cost  $C$ , so for the specified variance of the estimator, say,  $V'_0$ , we obtain the optimum values of  $n'$ ,  $n$  and  $k$ , which have minimum cost.

## 7. Determination of $n'$ , $n$ and $k$ for the specified variance $V = V'_0$

Let  $V'_0$  be the variance of the estimator  $t_i$  fixed in advance, so we have

$$V'_0 = \frac{V_{0i}}{n} + \frac{V_{1i}}{n'} + \frac{kV_{2i}}{n} - \frac{1}{N}(S_y^2 + R^2 S_z^2 - 2RS_{yz}) \quad i=1, 2. \quad (7.1)$$

For minimizing the average total cost  $C$  for the specified variance ( $\cong AMSE(t_i)$ ,  $i=1,2$ ) of the estimator  $t_i$ , for obtaining the optimum values of  $n'$ ,  $n$  and  $k$ , we define a function  $\psi$  given as:

$$\psi = (c'_1 + c'_2)n' + n(c_1 + c_2W_1 + c_3 \frac{W_2}{k}) + \mu_i(AMSE(t_i) - V'_0) \quad i=1, 2, \quad (7.2)$$

where  $\mu_i$  is the Lagrange's multiplier. Now, differentiating  $\psi$  with respect to  $n'$ ,  $n$  and  $k$  and equal to zero, we have

$$n' = \sqrt{\frac{\mu_i V_{1i}}{(c'_1 + c'_2)}}, \quad n = \sqrt{\frac{\mu_i (V_{0i} + kV_{2i})}{(c_1 + c_2W_1 + c_3 \frac{W_2}{k})}}, \quad (7.3)$$

$$\text{and } k_{opt} = \sqrt{\frac{V_{0i}W_2c_3}{V_{2i}(c_1 + c_2W_1)}}. \quad (7.4)$$

Now, putting the value of  $n'$ ,  $n$  from (7.3) in (7.1) and using  $k_{opt}$  given by (7.4), we have

$$\sqrt{\mu_i} = \frac{\left[ \sqrt{(V_{0i} + k_{opt}V_{2i})(c_1 + c_2W_1 + c_3 \frac{W_2}{k_{opt}})} + \sqrt{V_{1i}(c'_1 + c'_2)} \right]}{V'_0 + \frac{1}{N}(S_y^2 + R^2 S_z^2 - 2RS_{yz})}. \quad (7.5)$$

The optimum values of  $n'$  and  $n$  are obtain by putting the value of  $k_{opt}$  and the optimum value of  $\sqrt{\mu_i}$  from (7.4) and (7.5) in (7.3). Further, it has been observed that the determinant of the matrix of the second order derivatives of  $\psi$  with respect to  $n'$ ,  $n$  and  $k$  is positive for the optimum values of  $n'$ ,  $n$  and  $k$ . The minimum expected total cost to be incurred on the use of  $t_i$  for the specified variance  $V'_0$  will be given by

$$C_{i\min} = \frac{\left[ \sqrt{(V_{0i} + k_{opt}V_{2i})(c_1 + c_2W_1 + c_3 \frac{W_2}{k_{opt}})} + \sqrt{V_{1i}(c'_1 + c'_2)} \right]^2}{V'_0 + \frac{1}{N}(S_y^2 + R^2 S_z^2 - 2RS_{yz})} \quad i=1, 2. \quad (7.6)$$

Now, neglecting the terms of  $O(N^{-1})$ , we have

$$C_{i\min} = \frac{\left[ \sqrt{(V_{0i} + k_{opt} V_{2i})(c_1 + c_2 W_1 + c_3 \frac{W_2}{k_{opt}})} + \sqrt{V_{1i}(c'_1 + c'_2)} \right]^2}{V'_0} \quad i=1, 2. \quad (7.7)$$

8. An empirical study

The data on physical growth of upper socio-economic group of 95 school going children of Varanasi under an ICMR study, Department of Pediatrics, B.H.U., during 1983-84 has been taken under study, (Khare and Sinha (2007)).The last 25% (i.e. 24 children) of units have been considered as non-responding units. The study variable ( $y$ ), auxiliary variable ( $x$ ) and the additional auxiliary variable ( $z$ ) are taken as follows:

- $y$  - weight (in kg) of the children,
- $x$  - chest circumference (in cm) of the children,
- $z$  - skull circumference (in cm) of the children.

The values of the parameters of the  $y, x$  and  $z$  variables for the given data are taken as follows:

$\bar{Y} = 21.9758, \bar{X} = 57.2116, \bar{Z} = 51.6084, C_y = 0.1905, C_x = 0.0705,$   
 $C_z = 0.0322, C_{y(2)} = 0.1856, C_{x(2)} = 0.0752, \rho_{yx} = 0.8338, \rho_{yz} = 0.4274,$   
 $\rho_{yx(2)} = 0.8426.$

**Table 1.** Absolute bias and Relative efficiency (with respect to  $\bar{y}^*$ ) of the estimators  $\bar{y}^*, T_1, T_2, t_1$  and  $t_2$  for the fixed values of  $n', n$  and different values of  $k$  ( $N=95, n' =70$  and  $n=50$ )

Estimators	1 / k		
	1/4	1/3	1/2
$\bar{y}^*$	00.00 100 (0.4181)*	00.00 100 (0.3340)	00.00 100 (0.2499)
$T_1$	0.00281 175 (0.2392)	0.00214 168 (0.1988 )	0.00146 158 (0.1583)
$T_2$	0.00078 113 (0.3699)	0.00078 117 (0.2859)	0.00078 124 (0.2019)
$t_1$	0.00268 180 (0.2316)	0.00201 175 (0.1912)	0.00133 166 (0.1507)
$t_2$	0.00065 115 (0.3623)	0.00065 120 (0.2783)	0.00065 127 (0.1942)

\*Figures in parenthesis give the AMSE(.).

From table 1, we observe that for the fixed value of  $n'$ ,  $n$  and  $k = 2, 3, 4$ , the absolute bias of the estimators  $t_1$  and  $t_2$  is less than the bias of the corresponding estimator  $T_1$  and  $T_2$ . The amount of absolute bias for the estimators  $T_1$  and  $t_1$  decreases as values of  $k^{-1}$  increase but the amount of absolute bias for the estimators  $T_2$  and  $t_2$  remains constant for different values of  $k$ . We also observe that the estimator  $t_1$  has maximum relative efficiency with respect to  $\bar{y}^*$ . Since the condition for preference of  $T_1$  over  $T_2$  is attained in the given data set, so we observe that the estimator  $T_1$  is more efficient than  $T_2$  and  $\bar{y}^*$ . However, the estimators  $t_1$  and  $t_2$  are more efficient than  $\bar{y}^*$  and correspondingly with  $T_1$  and  $T_2$ . In this case  $t_1$  is found to be more efficient than  $t_2$ .

## 9. Monte Carlo simulation study

**Case 1:** In the present Monte Carlo simulation study, we consider the same data set as described in the previous section-8. From the population of 95 schools going children of Varanasi, two preliminary samples of different sizes 70 and 60 are taken by simple random sampling without replacement and the values of  $\bar{x}'$  and  $\bar{z}'$  based on 70 and 60 units are calculated. Again, we take two subsample of different size 50 and 40 from each preliminary sample of size 70 and 60 respectively using simple random sampling without replacement scheme. In each subsample, the last 25% (12 and 10 children respectively) of units have been considered as non-responding units. We again take a subsample of  $r$  units from non-responding units with simple random sampling without replacement and collect all information on  $r = n_2 k^{-1}$  units. Here,  $n_2 = 12, 10$  is the non-responding unit in each subsample and  $k = 2, 3, 4$  respectively. The above process is replicated 1000 times.

Simulated absolute bias and simulated mean square error of  $t_i (i = 1, 2)$  are calculated as follows:

$$\text{Absolute Bias}(t_i) = \frac{1}{1000} \left| \sum_{j=1}^{1000} t_{ij} - \bar{Y} \right|,$$

$$MSE(t_i) = \frac{1}{1000} \sum_{j=1}^{1000} (t_{ij} - \bar{Y})^2.$$



**Table 2.** Simulated absolute bias and Simulated relative efficiency (with respect to  $\bar{y}^*$ ) of the estimators  $\bar{y}^*, T_1, T_2, t_1$  and  $t_2$  for the fixed values of  $n', n$  and different values of  $k$  ( $N = 95, n' = 70$  and  $n = 50$ )

Estimators	1 / k		
	1/4	1/3	1/2
$\bar{y}^*$	0.02313 100 (0.4248)*	0.02006 100 (0.3494)	0.01421 100 (0.2487)
$T_1$	0.02078 170 (0.2501)	0.01587 160 (0.2188 )	0.00865 153 (0.1627)
$T_2$	0.02268 114 (0.3738)	0.01843 116 (0.3015 )	0.01001 123 (0.2023)
$t_1$	0.02018 175 (0.2425)	0.01564 167 (0.2097)	0.00609 160 (0.1552)
$t_2$	0.02205 116 (0.3671)	0.01819 120 (0.2921)	0.00745 128 (0.1951)

\*Figures in parenthesis give the MSE(.).

From table 2, we observe that for the fixed value of  $n', n$  and  $k = 2, 3, 4$ , the simulated absolute bias of the estimators ( $t_1, t_2$ ) is less than the corresponding estimators ( $T_1, T_2$ ) and  $\bar{y}^*$ . We also observe that the estimators ( $t_1, t_2$ ) have less mean square error in comparison to the corresponding estimators ( $T_1, T_2$ ) and  $\bar{y}^*$ .

The estimator  $t_1$  has less mean square error in comparison to the estimator  $t_2$ . When  $k^{-1}$  increases, mean square error of all the estimators decreases.

From table 1-2, we observe that the amount of *Bias(.)* and *AMSE(.)* based on empirical study and the amount of *Bias(.)* and *MSE(.)* based on simulation study of different estimators considered under the study is almost same.

**Table 3.** Simulated absolute bias and Simulated relative efficiency (with respect to  $\bar{y}^*$ ) of the estimators  $\bar{y}^*, T_1, T_2, t_1$  and  $t_2$  for the fixed values of  $n', n$  and different values of  $k$  ( $N = 95, n' = 60$  and  $n = 40$ )

Estimators	1 / k		
	1/4	1/3	1/2
$\bar{y}^*$	0.03455 100 (0.6589)*	0.02174 100 (0.5005)	0.01381 100 (0.3761)
$T_1$	0.01758 167 (0.3939)	0.01447 159 (0.3144 )	0.00585 156 (0.2410)

**Table 3.** Simulated absolute bias and Simulated relative efficiency (with respect to  $\bar{y}^*$ ) of the estimators  $\bar{y}^*, T_1, T_2, t_1$  and  $t_2$  for the fixed values of  $n', n$  and different values of  $k$  ( $N=95, n' = 60$  and  $n=40$ ) (cont.)

Estimators	1 / k		
	1/4	1/3	1/2
$T_2$	0.02978 112 (0.5887)	0.01813 115 (0.4341)	0.01237 123 (0.3045)
$t_1$	0.01521 172 (0.3834)	0.01303 167 (0.2994)	0.00549 164 (0.2291)
$t_2$	0.02741 114 (0.5778)	0.01667 120 (0.4184)	0.01200 129 (0.2922)

\* Figures in parenthesis give the  $MSE(.)$ .

From table 3, we observe that for the fixed value of  $n', n$  and  $k=2, 3, 4$ , simulated absolute bias of the estimators ( $t_1, t_2$ ) is less than the corresponding estimators ( $T_1, T_2$ ) and  $\bar{y}^*$ . Simulated absolute bias of the estimators decreases as values of  $k^{-1}$  increase. We also observe that the estimators ( $t_1, t_2$ ) have less mean square error in comparison to the corresponding estimators ( $T_1, T_2$ ) and  $\bar{y}^*$ . The estimator  $t_2$  has more mean square errors in comparison to the estimator  $t_1$ . Mean square error of all the estimators decreases as  $k^{-1}$  increases.

**Case 2:** For the simulation study of the estimators  $R\hat{B}()$  and  $AM\hat{S}E()$ . In this case, we have taken the sample of 50 units from the population of 95 school going children of Varanasi with simple random sampling without replacement and calculated  $c_y^2, c_x^2, c_z^2, c_{xy}$  and  $c_{yz}$  based on 50 units. In sample of 50 units, the last 25% (12 children) of units have been considered as non-responding units. Again, we take a subsample of  $r=12k^{-1}$  units from non-responding units in sample of 50 units with simple random sampling without replacement and calculate  $c_{y(2)}^2, c_{x(2)}^2$  and  $c_{yx(2)}$  based on  $r$  units. Here, we take  $k = 2, 3, 4$ . After putting the values of  $c_y^2, c_x^2, c_z^2, c_{y(2)}^2, c_{x(2)}^2, c_{yx}, c_{yz}$  and  $c_{yx(2)}$  in the expression of the estimators  $R\hat{B}()$  and  $AM\hat{S}E()$  and replicating above process 1000 times, we find the simulated values of the estimators  $R\hat{B}()$  and  $AM\hat{S}E()$ .

**Table 4.** Simulated absolute value of the estimator  $R\hat{B}(\cdot)$  and Simulated relative efficiency (with respect to  $AM\hat{S}E(\bar{y}^*)$ ) of the estimator  $AM\hat{S}E(\cdot)$  for the fixed values of  $n'$ ,  $n$  and different values of  $k$  ( $N=95$ ,  $n'=70$  and  $n=50$ )

$R\hat{B}(\cdot)$ $AM\hat{S}E(\cdot)$	$1/k$		
	1/4	1/3	1/2
$R\hat{B}(\bar{y}^*)$ $AM\hat{S}E(\bar{y}^*)$	00.00 100 (0.4263)*	00.00 100 (0.3418)	00.00 100 (0.2482)
$R\hat{B}(T_1)$ $AM\hat{S}E(T_1)$	$1.1000 \times 10^{-4}$ 177 (0.2414)	$9.2073 \times 10^{-5}$ 171 (0.1999 )	$6.3361 \times 10^{-5}$ 157 (0.1579)
$R\hat{B}(T_2)$ $AM\hat{S}E(T_2)$	$3.5282 \times 10^{-5}$ 121 (0.3528)	$3.4978 \times 10^{-5}$ 122 (0.2794)	$3.4784 \times 10^{-5}$ 126 (0.1974)
$R\hat{B}(t_1)$ $AM\hat{S}E(t_1)$	$1.2000 \times 10^{-4}$ 183 (0.2336)	$9.8001 \times 10^{-5}$ 178 (0.1923)	$6.9112 \times 10^{-5}$ 165(0.1504)
$R\hat{B}(t_2)$ $AM\hat{S}E(t_2)$	$4.1314 \times 10^{-5}$ 124 (0.3451)	$4.0906 \times 10^{-5}$ 126 (0.2718)	$4.0535 \times 10^{-5}$ 131 (0.1899)

\*Figures in parenthesis give the  $AM\hat{S}E(\cdot)$ .

From table 4, we observe that for  $n'=70$ ,  $n=50$  and  $k=2,3,4$ , simulated values of the estimators ( $R\hat{B}(t_1)$ ,  $R\hat{B}(t_2)$ ) are more than the values of the corresponding estimators( $R\hat{B}(T_1)$ ,  $R\hat{B}(T_2)$ ). We also observe that the estimators ( $AM\hat{S}E(t_1)$ ,  $AM\hat{S}E(t_2)$ ) are more efficient in comparison to the corresponding estimators ( $AM\hat{S}E(T_1)$ ,  $AM\hat{S}E(T_2)$ ) and  $AM\hat{S}E(\bar{y}^*)$ . The values of all the estimators decrease as  $k^{-1}$  increases.

**Table 5.** Absolute bias of the estimators  $R\hat{B}(\cdot)$  and  $AM\hat{S}E(\cdot)$  for the fixed values of  $n'$ ,  $n$  and different values of  $k$  ( $N=95$ ,  $n'=70$  and  $n=50$ )

$  R\hat{B}(\cdot) - RB(\cdot)  $ $  AM\hat{S}E(\cdot) - AMSE(\cdot)  $	$1/k$		
	1/4	1/3	1/2
$  R\hat{B}(\bar{y}^*) - RB(\bar{y}^*)  $ $  AM\hat{S}E(\bar{y}^*) - AMSE(\bar{y}^*)  $	00.00 0.00829	00.00 0.00785	00.00 0.00173

**Table 5.** Absolute bias of the estimators  $R\hat{B}(\cdot)$  and  $AM\hat{S}E(\cdot)$  for the fixed values of  $n'$ ,  $n$  and different values of  $k$  ( $N=95$ ,  $n'=70$  and  $n=50$ ) (cont.)

$  R\hat{B}(\cdot) - RB(\cdot)  $ $  AM\hat{S}E(\cdot) - AMSE(\cdot)  $	1/k		
	1/4	1/3	1/2
$  R\hat{B}(T_1) - RB(T_1)  $ $  AM\hat{S}E(T_1) - AMSE(T_1)  $	$1.358 \times 10^{-5}$ 0.00212	$5.1711 \times 10^{-6}$ 0.00117	$3.0452 \times 10^{-6}$ 0.00043
$  R\hat{B}(T_2) - RB(T_2)  $ $  AM\hat{S}E(T_2) - AMSE(T_2)  $	$2.888 \times 10^{-7}$ 0.01708	$5.9268 \times 10^{-7}$ 0.00650	$7.857 \times 10^{-7}$ 0.00452
$  R\hat{B}(t_1) - RB(t_1)  $ $  AM\hat{S}E(t_1) - AMSE(t_1)  $	$1.562 \times 10^{-6}$ 0.00199	$6.7111 \times 10^{-6}$ 0.00115	$8.6591 \times 10^{-6}$ 0.00027
$  R\hat{B}(t_2) - RB(t_2)  $ $  AM\hat{S}E(t_2) - AMSE(t_2)  $	$1.1697 \times 10^{-5}$ 0.01721	$1.1289 \times 10^{-5}$ 0.00652	$1.0919 \times 10^{-7}$ 0.00436

From table 5, we observe that the differences between the estimated value and the actual value of  $R\hat{B}(\cdot)$  and  $AM\hat{S}E(\cdot)$  based on simulation technique are very small and likely to be neglected.

**Case 3:** For the simulation study of the estimators  $\bar{y}^*$ ,  $T_1$ ,  $T_2$ ,  $t_1$  and  $t_2$  when some fraction of  $r$  value of  $y$  and  $x$  variables is not observed in the last phase.

In this case, we assumed  $r/2$  units observed in the last phase and repeat all process according to case 1.

Simulated absolute bias and simulated mean square error of  $t_i = 1, 2$  are calculated as follows:

$$\text{Absolute Bias}(t_i) = \frac{1}{1000} \left| \sum_{j=1}^{1000} t_{ij} - \bar{Y} \right|,$$

$$MSE(t_i) = \frac{1}{1000} \sum_{j=1}^{1000} (t_{ij} - \bar{Y})^2$$

**Table 6.** Simulated absolute bias and Simulated relative efficiency (with respect to  $\bar{y}^*$ ) of the estimators  $\bar{y}^*$ ,  $T_1$ ,  $T_2$ ,  $t_1$  and  $t_2$  (when  $r/2$  units are observed) for the fixed values of  $n'$ ,  $n$  and different values of  $k$  ( $N=95$ ,  $n'=70$  and  $n=50$ )

Estimators	1/k		
	1/4	1/3	1/2
$\bar{y}^*$	0.04580 100 (1.1896)*	0.02465 100 (0.6566)	0.00918 100 (0.4187)
$T_1$	0.02859 197 (0.6042)	0.02396 179 (0.3664)	0.00391 169 (0.2480)
$T_2$	0.04191 105 (1.1325)	0.02272 108 (0.6064)	0.00379 113 (0.3718)
$t_1$	0.02821 200 (0.5950)	0.02359 183 (0.3595)	0.00275 175 (0.2387)
$t_2$	0.04149 106 (1.1218)	0.02237 109 (0.6007)	0.00264 115 (0.3627)

\* Figures in parenthesis give the  $MSE(.)$ .

From table 6, we observe that for the fixed value of  $n'$ ,  $n$  and  $k=2, 3, 4$ , simulated absolute bias of the estimators ( $t_1, t_2$ ) is less than the corresponding estimators ( $T_1, T_2$ ) and  $\bar{y}^*$ . Simulated absolute bias of the estimators decreases as values of  $k^{-1}$  increase. We also observe that the estimators ( $t_1, t_2$ ) have less mean square error in comparison to the corresponding estimators ( $T_1, T_2$ ) and  $\bar{y}^*$ .

**Table 7.** Simulated absolute bias and Simulated relative efficiency (with respect to  $\bar{y}^*$ ) of the estimators  $\bar{y}^*$ ,  $T_1$ ,  $T_2$ ,  $t_1$  and  $t_2$  (when  $r/2$  units are observed) for the fixed values of  $n'$ ,  $n$  and different values of  $k$  ( $N=95$ ,  $n'=60$  and  $n=40$ )

Estimators	1/k		
	1/4	1/3	1/2
$\bar{y}^*$	0.03782 100 (1.2869)*	0.02297 100 (1.1901)	0.02117 100 (0.6753)
$T_1$	0.02079 188 (0.6830)	0.01992 175 (0.6789)	0.01023 170 (0.3983)
$T_2$	0.03383 106 (1.2118)	0.02076 106 (1.1213)	0.01655 112 (0.6024)
$t_1$	0.02069 191 (0.6729)	0.01963 179 (0.6661)	0.00827 174 (0.3877)
$t_2$	0.03369 107 (1.2003)	0.02049 107 (1.1071)	0.01460 114 (0.5920)

\* Figures in parenthesis give the  $MSE(.)$ .

From table 7, we observe that for the fixed value of  $n'$ ,  $n$  and  $k=2,3,4$ , simulated absolute bias of the estimators  $(t_1, t_2)$  is less than the corresponding estimators  $(T_1, T_2)$  and  $\bar{y}^*$ . Simulated absolute bias of the estimators decreases as values of  $1/k$  increase. We also observe that the estimators  $(t_1, t_2)$  have less mean square error in comparison to the corresponding estimators  $(T_1, T_2)$  and  $\bar{y}^*$ . Mean square error of all the estimators decreases as  $k^{-1}$  increases.

**Case 4:** For the simulation study of the estimators  $R\hat{B}(\cdot)$  and  $AM\hat{S}E(\cdot)$  when some fraction of  $r$  value of  $y$  and  $x$  variables is not observed in the last phase.

In this case, we assumed  $r/2$  units observed in the last phase and repeat all process according to case 2.

**Table 8.** Simulated absolute value of the estimator  $R\hat{B}(\cdot)$  and Simulated relative efficiency (with respect to  $AM\hat{S}E(\bar{y}^*)$ ) of the estimator  $AM\hat{S}E(\cdot)$  for the fixed values of  $n'$ ,  $n$  and different values of  $k$  and  $r/2$  units is observed ( $N=95$ ,  $n'=70$  and  $n=50$ )

$R\hat{B}(\cdot)$ $AM\hat{S}E(\cdot)$	$1/k$		
	1/4	1/3	1/2
$R\hat{B}(\bar{y}^*)$ $AM\hat{S}E(\bar{y}^*)$	00.00 100 (0.4482)*	00.00 100 (0.3379)	00.00 100 (0.2524)
$R\hat{B}(T_1)$ $AM\hat{S}E(T_1)$	$1.2000 \times 10^{-4}$ 185 (0.2428)	$8.7005 \times 10^{-5}$ 173 (0.1955)	$6.2294 \times 10^{-5}$ 159 (0.1585)
$R\hat{B}(T_2)$ $AM\hat{S}E(T_2)$	$3.5567 \times 10^{-5}$ 125 (0.3585)	$3.5368 \times 10^{-5}$ 126 (0.2689)	$3.5054 \times 10^{-5}$ 128 (0.1967)
$R\hat{B}(t_1)$ $AM\hat{S}E(t_1)$	$1.2300 \times 10^{-4}$ 191 (0.2352)	$9.2965 \times 10^{-5}$ 180 (0.1879)	$6.8145 \times 10^{-5}$ 167 (0.1509)
$R\hat{B}(t_2)$ $AM\hat{S}E(t_2)$	$4.1547 \times 10^{-5}$ 128 (0.3508)	$4.1329 \times 10^{-5}$ 129 (0.2612)	$4.0904 \times 10^{-5}$ 133 (0.1891)

\*Figures in parenthesis give the  $AM\hat{S}E(\cdot)$ .

From table 8, we observe that for  $n'=70$ ,  $n=50$  and  $k=2,3,4$  simulated absolute value of the estimators  $(R\hat{B}(t_1), R\hat{B}(t_2))$  is more than the corresponding estimators  $(R\hat{B}(T_1), R\hat{B}(T_2))$ . We also observe that the estimators  $(AM\hat{S}E(t_1), AM\hat{S}E(t_2))$  are more efficient in comparison to the corresponding estimators  $(AM\hat{S}E(T_1), AM\hat{S}E(T_2))$  and  $AM\hat{S}E(\bar{y}^*)$ . The values of all the estimators decrease as  $k^{-1}$  increases.

**Table 9.** Absolute bias of the estimators  $R\hat{B}(\cdot)$  and  $AM\hat{S}E(\cdot)$  for the fixed values of  $n'$ ,  $n$  and different values of  $k$  and  $r/2$  units is observed ( $N=95$ ,  $n'=70$  and  $n=50$ )

$\begin{array}{c}   R\hat{B}(\cdot) - RB(\cdot)   \\   AM\hat{S}E(\cdot) - AMSE(\cdot)   \end{array}$	$1/k$		
	1/4	1/3	1/2
$\begin{array}{c}   R\hat{B}(\bar{y}^*) - RB(\bar{y}^*)   \\   AM\hat{S}E(\bar{y}^*) - AMSE(\bar{y}^*)   \end{array}$	$\begin{array}{c} 00.00 \\ 0.0302 \end{array}$	$\begin{array}{c} 00.00 \\ 0.0039 \end{array}$	$\begin{array}{c} 00.00 \\ 0.0025 \end{array}$
$\begin{array}{c}   R\hat{B}(T_1) - RB(T_1)   \\   AM\hat{S}E(T_1) - AMSE(T_1)   \end{array}$	$\begin{array}{c} 1.0891 \times 10^{-5} \\ 0.0036 \end{array}$	$\begin{array}{c} 1.0239 \times 10^{-5} \\ 0.0032 \end{array}$	$\begin{array}{c} 3.4624 \times 10^{-6} \\ 0.0002 \end{array}$
$\begin{array}{c}   R\hat{B}(T_2) - RB(T_2)   \\   AM\hat{S}E(T_2) - AMSE(T_2)   \end{array}$	$\begin{array}{c} 2.9000 \times 10^{-5} \\ 0.0114 \end{array}$	$\begin{array}{c} 2.0185 \times 10^{-7} \\ 0.0171 \end{array}$	$\begin{array}{c} 5.1701 \times 10^{-7} \\ 0.0052 \end{array}$
$\begin{array}{c}   R\hat{B}(t_1) - RB(t_1)   \\   AM\hat{S}E(t_1) - AMSE(t_1)   \end{array}$	$\begin{array}{c} 1.0695 \times 10^{-5} \\ 0.0035 \end{array}$	$\begin{array}{c} 1.6750 \times 10^{-6} \\ 0.0033 \end{array}$	$\begin{array}{c} 7.6916 \times 10^{-6} \\ 0.0002 \end{array}$
$\begin{array}{c}   R\hat{B}(t_2) - RB(t_2)   \\   AM\hat{S}E(t_2) - AMSE(t_2)   \end{array}$	$\begin{array}{c} 1.1931 \times 10^{-5} \\ 0.0115 \end{array}$	$\begin{array}{c} 1.1712 \times 10^{-5} \\ 0.0171 \end{array}$	$\begin{array}{c} 1.1287 \times 10^{-5} \\ 0.0052 \end{array}$

From table 9, we observe that the values of the estimator  $R\hat{B}(\cdot)$  based on simulation study are almost close to the  $RB(\cdot)$  of the estimators  $T_1, T_2, t_1$  and  $t_2$ . We also observe that the values of the estimator  $AM\hat{S}E(\cdot)$  based on simulation study are almost close to the  $AMSE(\cdot)$  of the estimators  $T_1, T_2, t_1$  and  $t_2$  based on empirical study.

**Table 10.** Relative efficiency of the estimators  $\bar{y}^*, T_1, T_2, t_1$  and  $t_2$  with respect to  $\bar{y}^*$  (for the fixed cost  $C \leq C_0 = Rs.100$ ,  $c'_1 = Rs. 0.50$ ,  $c'_2 = Rs. 0.10$ ,  $c_1 = Rs. 1$ ,  $c_2 = Rs. 2$ ,  $c_3 = Rs. 9$ )

Estimators	$k_{opt}$	$n'_{opt}$	$n_{opt}$	R.E. (.) in %
$\bar{y}^*$	1.70	-	26	100 (0.7839) *
$T_1$	1.79	49	20	111 (0.7054)
$T_2$	1.03	47	16	105 (0.7464)
$t_1$	1.79	39	20	114 (0.6897)
$t_2$	1.03	38	16	107 (0.7302)

\*Figures in parenthesis give the  $AMSE(\cdot)$ , Rs: Rupees (Indian currency)

From table 10, we observe that for the fixed cost the estimators  $t_1$  and  $t_2$  have smaller approximated mean square error than that of  $\bar{y}^*$  and  $(t_1, t_2)$  are also more efficient than the corresponding estimators  $(T_1, T_2)$ . The estimator  $t_1$  is more efficient than  $t_2$ .

**Table 11.** Expected cost of the estimators  $\bar{y}^*, T_1, T_2, t_1$  and  $t_2$  for the specified variance  $V'_0 = 0.4897$  : ( $c'_1 = \text{Rs. } 0.50$ ,  $c'_2 = \text{Rs. } 0.10$ ,  $c_1 = \text{Rs. } 1$ ,  $c_2 = \text{Rs. } 2$ ,  $c_3 = \text{Rs. } 9$ )

Estimators	$k_{opt}$	$n'_{opt}$	$n_{opt}$	Expected cost (Rs.)
$\bar{y}^*$	1.70	-	42	160
$T_1$	1.79	70	29	144
$T_2$	1.03	72	25	152
$t_1$	1.79	55	29	141
$t_2$	1.03	57	24	149

From table 11, we observe that for the specified variance the expected cost is minimum for  $t_1$  and  $t_2$  with corresponding cost for the estimators  $T_1$  and  $T_2$  and also  $\bar{y}^*$ . The estimator  $t_1$  has also less cost in comparison to  $t_2$ .

So, we conclude that the use of information on the additional auxiliary variable is useful in increasing the precision of the proposed estimators with respect to the relevant estimators for the fixed sample size ( $n', n$ ) based on empirical study and also on Monte Carlo simulation study. The results derived from Monte Carlo simulation study based on empirical data were also found to be in congruence to the results based on empirical study. On the basis of empirical study, the use of additional variable in the proposed estimators for population mean in the presence of non-response is also found to be more useful in increasing the precision of the proposed estimators with respect to the relevant estimators for the fixed cost  $C \leq C_0$ . For the specified variance, the total cost for the proposed estimators is also less than the corresponding relevant estimators.

## Acknowledgement

The authors are highly grateful to the referee for his encouraging suggestions which were very much useful in improving the quality of the paper.



## REFERENCES

- BOWELY, A. L. (1926). Measurements of precision attained in sampling. *Bull. Inter. Statist. Inst.*, 22: 1-62.
- CHAND, L. (1975). Some ratio-type estimators based on two or more auxiliary variables. Ph.D. Thesis submitted to Iowa State University, Ames, IOWA.
- COCHRAN, W. G. (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. *Jour. Agri. Science*, 30: 262-275.
- COCHRAN, W. G. (1942). Sampling theory when the sampling units are of unequal sizes. *Jour. Amer. Statist. Assoc.*, 37: 199-212.
- HANSEN, M. H. and HURWITZ, W. N. (1946). The problem of non-response in sample surveys. *Jour. Amer. Statist. Assoc.*, 41: 517-529.
- HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953). *Sample Survey methods and theory Vol. 2: John Wiley and Sons, New York.*
- KHARE, B. B. and SRIVASTAVA, S. (1993). Estimation of population mean using auxiliary character in presence of non-response. *Nat. Acad. Sc. Letters, India*, 16(3): 111-114.
- KHARE, B. B. and SRIVASTAVA, S. (1995). Study of conventional and alternative two phase sampling ratio, product and regression estimators in presence of nonresponse. *Proc. Nat. Acad. Sci., India*, 65(A) II: 195-203.
- KHARE, B. B. and SINHA, R. R. (2007). Estimation of the ratio of the two population means using multi-auxiliary characters in the presence of non-response. *Statistical Techniques in Life Testing, Reliability, Sampling Theory and Quality Control*, 163-171.
- KIREGYERA, B. (1980). A chain ratio type estimator in finite population double sampling using two auxiliary variables. *Metrika*, 27: 217-223.
- KIREGYERA, B. (1984). Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations. *Metrika*, 31: 215-226.
- LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with missing Data*, New York: Wiley.
- NEYMAN, J. (1934). On the two different aspects of representative method, the method of stratified sampling and the method of purposive selection. *Jour. Roy. Statist. Soc.*, 97: 558-625.
- NEYMAN, J. (1938). Contributions to the theory of sampling human population. *J. Amer. Statist. Assoc.*, 33: 101-116.

- RAO, C. R. and TOUTENBURG, H. (1995). *Linear Models: Least Squares and Alternatives*. New York: Springer.
- RAO, P. S. R. S. (1986). Ratio estimation with subsampling the nonrespondents. *Survey Methodology*, 12(2): 217-230.
- RAO, P. S. R. S. (1987). Ratio and regression Estimates with Subsampling the non-respondents. Paper presented at a special contributed session of the International Statistical Association Meetings, September, 2-16, Tokyo, Japan.
- ROBSON, D. S. (1957). Applications of multivariate polykeys to the theory of unbiased ratio-type estimators. *Jour. Amer. Statist. Assoc.*, 52: 511-522.
- SRIVASTAVA, S. RANI, KHARE, B. B. and SRIVASTAVA, S. R. (1990). A generalised chain ratio estimator for mean of finite population. *Jour. Ind. Soc. Ag. Statistics*, 42(1): 108-117.
- TOUTENBURG, H. and SRIVASTAVA, V. K. (1998). Estimation of ratio of population means in survey sampling when some observations are missing. *Metrika*, 48: 177-187.
- TOUTENBURG, H. and SRIVASTAVA, V. K. (2003). Efficient estimation of population mean using incomplete survey data on study and auxiliary characteristics. *Statistica*, LXIII, 2: 2003.
- WATSON, D. J. (1937). The estimation of leaf areas. *Jour. Agril. Sci.*, 27: 474-481.

## APPENDIX

## Relative Bias and Approximated Mean Square Error (AMSE)

$$\text{Let } \bar{y}^* = \bar{Y}(1 + \varepsilon_0), \quad \bar{x}^* = \bar{X}(1 + \varepsilon_1), \quad \bar{x} = \bar{X}(1 + \varepsilon_2), \quad \bar{x}' = \bar{X}(1 + \varepsilon_3), \\ \bar{z}' = \bar{Z}(1 + \varepsilon_4)$$

$$E(\varepsilon_i) = 0 \text{ and } |\varepsilon_i| < 1 \quad \forall \quad i=0,1,2,3,4.$$

$$E(\varepsilon_0^2) = \frac{f}{n} C_y^2 + \frac{W_2(k-1)}{n} C_{y(2)}^2, \quad E(\varepsilon_1^2) = \frac{f}{n} C_x^2 + \frac{W_2(k-1)}{n} C_{x(2)}^2, \quad E(\varepsilon_2^2) = \frac{f}{n} C_x^2$$

$$E(\varepsilon_3^2) = \frac{f'}{n'} C_x^2, \quad E(\varepsilon_4^2) = \frac{f'}{n'} C_z^2, \quad E(\varepsilon_0 \varepsilon_1) = \frac{f}{n} C_{yx} + \frac{W_2(k-1)}{n} C_{yx(2)},$$

$$E(\varepsilon_0 \varepsilon_2) = \frac{f}{n} C_{yx}, \quad E(\varepsilon_0 \varepsilon_3) = \frac{f'}{n'} C_{yx}, \quad E(\varepsilon_0 \varepsilon_4) = \frac{f'}{n'} C_{yz}, \quad E(\varepsilon_1 \varepsilon_3) = \frac{f'}{n'} C_x^2,$$

$$E(\varepsilon_1 \varepsilon_4) = \frac{f'}{n'} C_{xz}, \quad E(\varepsilon_2 \varepsilon_3) = \frac{f'}{n'} C_x^2, \quad E(\varepsilon_2 \varepsilon_4) = \frac{f'}{n'} C_{xz},$$

$$t_1 = \frac{\bar{y}^*}{\bar{x}^*} \frac{\bar{x}'}{\bar{z}'} \bar{Z} = \frac{\bar{Y}(1 + \varepsilon_0)}{\bar{X}(1 + \varepsilon_1)} \frac{\bar{X}(1 + \varepsilon_3)}{\bar{Z}(1 + \varepsilon_4)} \bar{Z}$$

$$t_1 = \bar{Y}(1 + \varepsilon_0)(1 + \varepsilon_3)(1 + \varepsilon_1)^{-1}(1 + \varepsilon_4)^{-1}$$

$$t_1 = \bar{Y}(1 + \varepsilon_0)(1 + \varepsilon_3)(1 - \varepsilon_1 + \varepsilon_1^2 \dots \dots \dots)(1 - \varepsilon_4 + \varepsilon_4^2 \dots \dots \dots)$$

After neglecting the terms of  $\varepsilon_i$  up to the second degree, we have

$$t_1 = \bar{Y}[1 + \varepsilon_0 - \varepsilon_1 + \varepsilon_3 - \varepsilon_4 - \varepsilon_0 \varepsilon_1 + \varepsilon_0 \varepsilon_3 - \varepsilon_0 \varepsilon_4 - \varepsilon_1 \varepsilon_3 + \varepsilon_1 \varepsilon_4 - \varepsilon_3 \varepsilon_4 + \varepsilon_1^2 + \varepsilon_4^2] \quad (\text{A.1})$$

$$t_2 = \frac{\bar{y}^*}{\bar{x}} \frac{\bar{x}'}{\bar{z}'} \bar{Z} = \frac{\bar{Y}(1 + \varepsilon_0)}{\bar{X}(1 + \varepsilon_2)} \frac{\bar{X}(1 + \varepsilon_3)}{\bar{Z}(1 + \varepsilon_4)} \bar{Z}$$

$$t_2 = \bar{Y}(1 + \varepsilon_0)(1 + \varepsilon_3)(1 + \varepsilon_2)^{-1}(1 + \varepsilon_4)^{-1}$$

$$t_2 = \bar{Y}(1 + \varepsilon_0)(1 + \varepsilon_3)(1 - \varepsilon_2 + \varepsilon_2^2 \dots \dots \dots)(1 - \varepsilon_4 + \varepsilon_4^2 \dots \dots \dots)$$

After neglecting the terms of  $\varepsilon_i$  up to the second degree, we have

$$t_2 = \bar{Y}[1 + \varepsilon_0 - \varepsilon_2 + \varepsilon_3 - \varepsilon_4 - \varepsilon_0 \varepsilon_2 + \varepsilon_0 \varepsilon_3 - \varepsilon_0 \varepsilon_4 - \varepsilon_2 \varepsilon_3 + \varepsilon_2 \varepsilon_4 - \varepsilon_3 \varepsilon_4 + \varepsilon_2^2 + \varepsilon_4^2] \quad (\text{A.2})$$

1. Relative bias of the estimators  $t_1$  and  $t_2$ 

$$RB(t_1) = \frac{E(t_1) - \bar{Y}}{\bar{Y}}$$

$$= E[\bar{Y} + \bar{Y}(\varepsilon_0 - \varepsilon_1 + \varepsilon_3 - \varepsilon_4 - \varepsilon_0 \varepsilon_1 + \varepsilon_0 \varepsilon_3 - \varepsilon_0 \varepsilon_4 - \varepsilon_1 \varepsilon_3 + \varepsilon_1 \varepsilon_4 - \varepsilon_3 \varepsilon_4 + \varepsilon_1^2 + \varepsilon_4^2) - \bar{Y}] / \bar{Y} \\ = -E(\varepsilon_0 \varepsilon_1) + E(\varepsilon_0 \varepsilon_3) - E(\varepsilon_0 \varepsilon_4) - E(\varepsilon_1 \varepsilon_3) + E(\varepsilon_1 \varepsilon_4) - E(\varepsilon_3 \varepsilon_4) + E(\varepsilon_1^2) + E(\varepsilon_4^2)$$

$$RB(t_1) = \left( \frac{1}{n} - \frac{1}{n'} \right) (C_x^2 - \rho_{yx} C_y C_x) + \frac{W_2(k-1)}{n} (C_{x(2)}^2 - \rho_{yx(2)} C_{y(2)} C_{x(2)}) + \frac{f'}{n'} (C_z^2 - \rho_{yz} C_y C_z) \quad (A.3)$$

$$RB(t_2) = \frac{E(t_2) - \bar{Y}}{\bar{Y}}$$

$$= E[\bar{Y} + \bar{Y}(\varepsilon_0 - \varepsilon_2 + \varepsilon_3 - \varepsilon_4 - \varepsilon_0\varepsilon_2 + \varepsilon_0\varepsilon_3 - \varepsilon_0\varepsilon_4 - \varepsilon_2\varepsilon_3 + \varepsilon_2\varepsilon_4 - \varepsilon_3\varepsilon_4 + \varepsilon_2^2 + \varepsilon_4^2) - \bar{Y}] / \bar{Y}$$

$$= -E(\varepsilon_0\varepsilon_2) + E(\varepsilon_0\varepsilon_3) - E(\varepsilon_0\varepsilon_4) - E(\varepsilon_2\varepsilon_3) + E(\varepsilon_2\varepsilon_4) - E(\varepsilon_3\varepsilon_4) + E(\varepsilon_2^2) + E(\varepsilon_4^2)$$

$$RB(t_2) = \left( \frac{1}{n} - \frac{1}{n'} \right) (C_x^2 - \rho_{yx} C_y C_x) + \frac{f'}{n'} (C_z^2 - \rho_{yz} C_y C_z) \quad (A.4)$$

## 2. Approximated Mean Square Error (AMSE) of the estimators $t_1$ and $t_2$

$$MSE(t_1) = E[t_1 - \bar{Y}]^2$$

$$= E[\bar{Y} + \bar{Y}(\varepsilon_0 - \varepsilon_1 + \varepsilon_3 - \varepsilon_4 \dots) - \bar{Y}]^2$$

$$= \bar{Y}^2 E[\varepsilon_0^2 + \varepsilon_1^2 + \varepsilon_3^2 + \varepsilon_4^2 - 2\varepsilon_0\varepsilon_1 + 2\varepsilon_0\varepsilon_3 - 2\varepsilon_0\varepsilon_4 - 2\varepsilon_1\varepsilon_3 + 2\varepsilon_1\varepsilon_4 - 2\varepsilon_3\varepsilon_4 \dots]$$

$$= \bar{Y}^2 [E(\varepsilon_0^2) + E(\varepsilon_1^2) + E(\varepsilon_3^2) + E(\varepsilon_4^2) - 2E(\varepsilon_0\varepsilon_1) + 2E(\varepsilon_0\varepsilon_3) - 2E(\varepsilon_0\varepsilon_4) - 2E(\varepsilon_1\varepsilon_3) + 2E(\varepsilon_1\varepsilon_4) - 2E(\varepsilon_3\varepsilon_4) \dots]$$

$$AMSE(t_1) = \bar{Y}^2 \left[ \left( \frac{1}{n} - \frac{1}{n'} \right) \{C_y^2 + C_x^2 - 2\rho_{yx} C_y C_x\} + \frac{W_2(k-1)}{n} \{C_{y(2)}^2 + C_{x(2)}^2 - 2\rho_{yx(2)} C_{y(2)} C_{x(2)}\} + \frac{f'}{n'} (C_y^2 + C_z^2 - 2\rho_{yz} C_y C_z) \right] \quad (A.5)$$

$$MSE(t_2) = E[t_2 - \bar{Y}]^2$$

$$= E[\bar{Y} + \bar{Y}(\varepsilon_0 - \varepsilon_2 + \varepsilon_3 - \varepsilon_4 \dots) - \bar{Y}]^2$$

$$= \bar{Y}^2 E[\varepsilon_0^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2 - 2\varepsilon_0\varepsilon_2 + 2\varepsilon_0\varepsilon_3 - 2\varepsilon_0\varepsilon_4 - 2\varepsilon_2\varepsilon_3 + 2\varepsilon_2\varepsilon_4 - 2\varepsilon_3\varepsilon_4 \dots]$$

$$= \bar{Y}^2 [E(\varepsilon_0^2) + E(\varepsilon_2^2) + E(\varepsilon_3^2) + E(\varepsilon_4^2) - 2E(\varepsilon_0\varepsilon_2) + 2E(\varepsilon_0\varepsilon_3) - 2E(\varepsilon_0\varepsilon_4) - 2E(\varepsilon_2\varepsilon_3) + 2E(\varepsilon_2\varepsilon_4) - 2E(\varepsilon_3\varepsilon_4) \dots]$$

$$AMSE(t_2) = \bar{Y}^2 \left[ \left( \frac{1}{n} - \frac{1}{n'} \right) \{C_y^2 + C_x^2 - 2\rho_{yx} C_y C_x\} + \frac{W_2(k-1)}{n} C_{y(2)}^2 + \frac{f'}{n'} (C_y^2 + C_z^2 - 2\rho_{yz} C_y C_z) \right] \quad (A.6)$$

STATISTICS IN TRANSITION-new series, December 2012  
Vol. 13, No. 3, pp. 519—536

## A CLASS OF CHAIN RATIO-CUM-DUAL TO RATIO TYPE ESTIMATOR WITH TWO AUXILIARY CHARACTERS UNDER DOUBLE SAMPLING IN SAMPLE SURVEYS

S. Choudhury<sup>12</sup>, B. K. Singh<sup>1</sup>

### ABSTRACT

This paper considers a chain ratio-cum-dual to ratio type estimator for estimating population mean of the study variate using two auxiliary variates under double (two-phase) sampling procedure, when the information on another additional auxiliary variate is available along with the main auxiliary variate. The asymptotically optimum estimators (AOEs) in the class are identified in two different cases with their bias and variances. The optimum values of the first phase and second phase sample sizes have been obtained for the fixed cost of survey. Theoretical and empirical studies have also been done to demonstrate the efficiency of the proposed estimator with respect to strategies which utilized the information on two auxiliary variates.

**Key words:** auxiliary variate, double sampling, chain ratio-cum-dual to ratio estimator, bias, variance, efficiency.

### 1. Introduction

The use of auxiliary information in the estimation of population values of the study variate has been a common phenomenon in sampling theory of surveys. Auxiliary information may be fruitfully utilized either at planning stage or at design stage or at the information stage to arrive at improved estimator compared to those, not utilizing auxiliary information. The use of ratio and product strategies in survey sampling solely depends upon the knowledge of population mean  $\bar{X}$  of the auxiliary character  $x$ . In many situations of practical importance, the population mean  $\bar{X}$  is not known before the start of a survey. In such a situation, the usual thing to do is to estimate it by the sample mean  $\bar{x}_1$  based on

---

<sup>1</sup> Department of Mathematics, North Eastern Regional Institute of Science and Technology, Nirjuli-791109, Arunachal Pradesh, India.

<sup>2</sup> E-mail: sanjibchy07@gmail.com.

a preliminary sample of size  $n_1$  of which  $n$  is a subsample ( $n < n_1$ ). If the population mean  $\bar{Z}$  of another auxiliary variate  $z$ , closely related to auxiliary variate  $x$  but compared to  $x$  remotely related to study variate  $y$  is known, it is advisable to estimate  $\bar{X}$  by  $\bar{X} = \bar{x}_1 \bar{Z} / \bar{z}'$ , which would provide better estimate of  $\bar{X}$  than  $\bar{x}_1$  to the terms of order  $o(n^{-1})$  if  $\rho_{xz} C_x / C_z > 1/2$ .

Chand (1975) and Sukhatme and Chand (1977) proposed a technique of chaining the available information on auxiliary characteristics with the main characteristics. Kiregyera (1980, 1984), Singh et al. (2006) also proposed some chain type ratio and regression estimators based on two auxiliary variables. Using prior information on parameters of auxiliary variate/variates, Srivastava and Jhaji (1980, 1995), Isaki (1983), Singh and Kataria (1990), Prasad and Singh (1990, 1992), Ahmed *et al.* (2000) defined estimators or classes of estimators of  $S_y^2$ . Using double sampling technique, Singh and Singh (2001) defined a ratio-type estimator of  $S_y^2$ . Al-Jararha and Ahmed (2002) defined two classes of estimators of  $S_y^2$  by using prior information on parameter of one of the two auxiliary variables under double sampling scheme. Ahmed *et al.* (2003) gave some chain ratio-type as well as chain product-type estimators of  $S_y^2$ , under two-phase sampling scheme. Singh and Tailor (2005) and Tailor and Sharma (2009) worked on ratio-cum-product estimators. Singh *et al.* (2006) proposed chain ratio and regression type estimators for median estimation. Use of a known coefficient of kurtosis of second auxiliary variable in double sampling Singh *et al.* (2009) defined a chain-type estimator of population variance.

Consider a finite population  $U = (U_1, U_2, \dots, U_N)$  of  $N$  units,  $y$  be the study variate,  $x$  and  $z$  are the two auxiliary variates. If  $\bar{X}$  is not known, but  $\bar{Z}$  the population mean of another cheaper auxiliary variate  $z$  closely related to  $x$  but compared to  $x$  remotely related to  $y$  (i.e.  $\rho_{yx} > \rho_{yz}$ ) is also available. In this case, Chand (1975) defined the chain ratio estimator

$$\bar{y}_R^{(dc)} = \bar{y} \frac{\bar{x}_1}{\bar{x}} \frac{\bar{Z}}{\bar{z}'}$$

where  $\bar{x}$  and  $\bar{y}$  are the sample mean of  $x$  and  $y$  respectively based on the sample size  $n$  out of the population  $N$  units and  $\bar{x}_1 = (1/n_1) \sum_{i=1}^{n_1} x_i$  denote the sample mean of  $x$  based on the first-phase sample of the size  $n_1$  and

$\bar{z}' = (1/n_1) \sum_{i=1}^{n_1} z_i$  denote the sample mean based on  $n_1 > n$  units of the auxiliary variate  $z$ .

Using the transformation  $\bar{x}_i^\sigma = (N\bar{X} - nx_i)/(N - n), (i = 1, 2, 3, \dots, N)$ , Srivenkataramana (1980) obtained dual to ratio estimator as

$$\bar{y}_R = \bar{y} \frac{\bar{x}^\sigma}{\bar{X}}$$

where  $\bar{x}^\sigma = (N\bar{X} - n\bar{x})/(N - n)$ .

Kumar *et al.* (2006) proposed a dual to ratio estimator in double sampling

$$\bar{y}_k^{(d)} = \bar{y} \frac{\bar{x}^\tau}{\bar{x}_1}$$

where  $\bar{x}^\tau = (n_1\bar{x}_1 - n\bar{x})/(n_1 - n)$  is an unbiased estimator of  $\bar{X}$

by using the transformation  $\bar{x}_i^* = (n_1\bar{x}_1 - nx_i)/(n_1 - n), (i = 1, 2, 3, \dots, N)$ .

Using an additional auxiliary variate  $z$  in dual to ratio estimator in double sampling,

Kumar *et al.* (2006) estimator reduces to chain dual to ratio estimator in double sampling as

$$\bar{y}_k^{(dc)} = \bar{y} \frac{\bar{x}^*}{\bar{x}_1} \frac{\bar{z}'}{\bar{Z}}$$

where  $\bar{x}^* = (n_1\bar{x}_1\bar{Z}/\bar{z}' - n\bar{x})/(n_1 - n)$ .

For estimating the population mean  $\bar{Y}$  of the study variate  $y$ , recently Sharma *et al.*

(2010) considered an estimator of ratio-cum-dual to ratio estimator given by

$$\bar{y}_{RdR} = \bar{y} \left[ \alpha \frac{\bar{X}}{\bar{x}} + (1 - \alpha) \frac{\bar{x}^\sigma}{\bar{X}} \right].$$

where  $\bar{X}$  is the known population mean of  $x$  and  $\alpha$  is suitably chosen constant.

Motivated by Sharma *et al.* (2010), we have proposed a class of chain ratio-cum-dual to ratio type estimator in double sampling for estimating population mean  $\bar{Y}$  using two auxiliary characters. Numerical illustrations are given in the support of the present study.

## 2. The proposed class of estimator

Let instead of  $\bar{X}$  the population mean  $\bar{Z}$  of another auxiliary variable  $z$  which has a positive correlation with  $x$  (that is,  $\rho_{xz} > 0$ ) be known. We further assume

that  $\rho_{yx} > \rho_{yz} > 0$ . Let  $\bar{x}_1$  and  $\bar{z}'$  be the sample means of  $x$  and  $z$  respectively based on a preliminary sample of size  $n_1$  drawn with simple random sampling without replacement (SRSWOR) strategy in order to get an estimate of  $\bar{X}$ . Then the proposed class of estimator for estimating  $\bar{Y}$  is

$$\bar{y}_{RdR}^{(dc)} = \bar{y} \left[ \alpha \frac{\bar{x}_1}{\bar{x}} \frac{\bar{Z}}{\bar{z}'} + (1-\alpha) \frac{\bar{x}^* \bar{z}'}{\bar{x}_1 \bar{Z}} \right] \quad (1)$$

where  $\alpha$  is determined so as to minimize the variance (V) of  $\bar{y}_{RdR}^{(dc)}$ .

To obtain the bias (B) and variance of  $\bar{y}_{RdR}^{(dc)}$ , we write

$$e_0 = (\bar{y} - \bar{Y})/\bar{Y}, \quad e_1 = (\bar{x} - \bar{X})/\bar{X}, \quad e'_1 = (\bar{x}_1 - \bar{X})/\bar{X} \text{ and } e'_2 = (\bar{z}' - \bar{Z})/\bar{Z}$$

Expressing  $\bar{y}_{RdR}^{(dc)}$  in terms of  $e$ 's, we have

$$\begin{aligned} \bar{y}_{RdR}^{(dc)} = \bar{Y} (1 + e_0) & \left[ \alpha (1 + e'_1) \left\{ 1 + (e_1 + e'_2 + e_1 e'_2) \right\}^{-1} \right. \\ & \left. + (1 - \alpha) \left\{ N^* - n^* (1 + e_1) (1 + e'_1 - e'_2 - e'_1 e'_2 + e_2'^2) \right\}^{-1} \right] \end{aligned} \quad (2)$$

where  $N^* = n_1/(n_1 - n)$  and  $n^* = n/(n_1 - n)$ .

We assume that  $|e_1 + e'_2 + e_1 e'_2| < 1$  and  $|e'_1 - e'_2 - e'_1 e'_2 + e_2'^2| < 1$  so that  $\left\{ 1 + (e_1 + e'_2 + e_1 e'_2) \right\}^{-1}$  and  $\left\{ 1 + (e'_1 - e'_2 - e'_1 e'_2 + e_2'^2) \right\}^{-1}$  are expandable.

Expanding the right hand side of (2), multiplying out and retaining terms of  $e$ 's up to the second degree we obtain

$$\begin{aligned} \bar{y}_{RdR}^{(dc)} - \bar{Y} \cong \bar{Y} & \left[ e_0 - n^* (e_1 - e'_1 + e'_2 + e_0 e_1 - e_0 e'_1 + e_0 e'_2 - e_1 e'_1 + e_1 e'_2 - e'_1 e'_2 + e_1'^2) \right. \\ & \left. + \alpha \{ e_1^2 + e_2'^2 + n^* e_1'^2 - N^* (e_1 - e'_1 + e'_2 + e_0 e_1 - e_0 e'_1 + e_0 e'_2) - N^{**} (e_1 e'_1 - e_1 e'_2 + e'_1 e'_2) \} \right] \end{aligned} \quad (3)$$

where  $N^{**} = (n_1 - 2n)/(n_1 - n)$ .

Taking the expectation in (3) and noting that

$$E(e_0) = E(e_1) = E(e'_1) = E(e'_2) = 0$$

and that the expectations of the second degree terms of order  $n^{-1}$ , we obtain

$$E \left\{ \bar{y}_{RdR}^{(dc)} \right\} = \bar{Y} + o(n^{-1}).$$

Thus, the bias of the estimator  $\bar{y}_{RdR}^{(dc)}$  is of the order  $n^{-1}$  and hence its contribution to the variance will be of the order of  $n^{-2}$ .



To find the bias and variance of  $\bar{y}_{RdR}^{(dc)}$ , let

$$C_y^2 = S_y^2 / \bar{Y}^2, \quad C_x^2 = S_x^2 / \bar{X}^2, \quad C_z^2 = S_z^2 / \bar{Z}^2, \quad \rho_{xy} = S_{xy} / S_x S_y, \quad \rho_{yz} = S_{yz} / S_y S_z \text{ and}$$

$$\rho_{zx} = S_{zx} / S_z S_x.$$

where

$$S_x^2 = \{1/(N-1)\} \sum_{i=1}^N (x_i - \bar{X})^2, \quad S_y^2 = \{1/(N-1)\} \sum_{i=1}^N (y_i - \bar{Y})^2, \quad S_z^2 = \{1/(N-1)\} \sum_{i=1}^N (z_i - \bar{Z})^2, \\ S_{yx} = \{1/(N-1)\} \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X}), \quad S_{yz} = \{1/(N-1)\} \sum_{i=1}^N (y_i - \bar{Y})(z_i - \bar{Z}) \text{ and} \\ S_{xz} = \{1/(N-1)\} \sum_{i=1}^N (x_i - \bar{X})(z_i - \bar{Z}).$$

The following two cases will be considered separately.

**Case I:** When the second phase sample of size  $n$  is a subsample of the first phase of size  $n_1$ . With this background, we get the following results

$$\left. \begin{aligned} E(e_0) &= E(e_1) = E(e'_1) = E(e'_2) = 0, \\ E(e_0^2) &= \{(1-f)/n\} C_y^2 = M_{0,0}, \quad E(e_1^2) = \{(1-f)/n\} C_x^2 = M_{1,0}, \\ E(e_1'^2) &= \{(1-f_1)/n_1\} C_x^2 = M_{2,0}, \quad E(e_2'^2) = \{(1-f_1)/n_1\} C_z^2 = M_{3,0}, \\ E(e_0 e_1) &= \{(1-f)/n\} C_{yx} C_x^2 = M_{0,1}, \quad E(e_0 e'_1) = \{(1-f_1)/n_1\} C_{yx} C_x^2 = M_{0,2}, \\ E(e_0 e'_2) &= \{(1-f_1)/n_1\} C_{yz} C_z^2 = M_{0,3}, \quad E(e_1 e'_1) = \{(1-f_1)/n_1\} C_x^2 = M_{1,2}, \\ E(e_1 e'_2) &= \{(1-f_1)/n_1\} C_{xz} C_z^2 = M_{1,3}, \quad E(e'_1 e'_2) = \{(1-f_1)/n_1\} C_{xz} C_z^2 = M_{2,3}. \end{aligned} \right\} \quad (4)$$

where  $f = n/N$ ,  $f_1 = n_1/N$ ,  $C_{yx} = \rho_{yx} C_y / C_x$ ,  $C_{yz} = \rho_{yz} C_y / C_z$  and  $C_{xz} = \rho_{xz} C_x / C_z$ .

Taking the expectation in (3) and using results in (4) we get the bias of the estimator  $\bar{y}_{RdR}^{(dc)}$  to the first order of approximation as

$$B\{\bar{y}_{RdR}^{(dc)}\}_I = \bar{Y}(\alpha P - \omega Q) \quad (5)$$

where  $\omega = n^* + \alpha N^{**}$ ,  $P = M_{1,0} - M_{2,0} + M_{3,0}$  and  $Q = M_{0,1} - M_{0,2} + M_{0,3}$ .

Again from (3), we have

$$\bar{y}_{RdR}^{(dc)} - \bar{Y} \cong \bar{Y} [e_0 - N_1^{**} (e_1 - e'_1 + e'_2)]. \quad (6)$$

Squaring both sides in (6), then taking the expectation and using the results in (4), we obtained the variance of the estimator  $\bar{y}_{RdR}^{(dc)}$  to terms of order  $n^{-1}$ , as

$$V\{\bar{y}_{RdR}^{(dc)}\}_I = \bar{Y}^2 (M_{0,0} - 2\omega Q + \omega^2 P) \quad (7)$$

The variance of  $\bar{y}_{RdR}^{(dc)}$  is minimum when

$$\alpha = \frac{Q}{PN^{**}} - n_1^* = \alpha_{lopt.} \text{ (say)} \quad (8)$$

where  $n_1^* = n/(n_1 - 2n)$ .

Putting (8) in (1) yields the 'asymptotically optimum estimator' (AOE) as

$$\{\bar{y}_{RdR}^{(dc)}\}_{lopt.} = \bar{y} \left[ \alpha_{lopt.} \frac{\bar{x}_1}{\bar{x}} \frac{\bar{z}}{\bar{z}'} + (1 - \alpha_{lopt.}) \frac{\bar{x}^* \bar{z}'}{\bar{x}_1 \bar{z}} \right]$$

Thus, the resulting variance of  $\{\bar{y}_{RdR}^{(dc)}\}_{lopt.}$  is given by

$$V\{\bar{y}_{RdR}^{(dc)}\}_{lopt.} = \bar{Y}^2 (M_{0,0} - Q^2/P). \quad (9)$$

**Remark 2.1.** For  $\alpha = 1$ , the estimator  $\bar{y}_{RdR}^{(dc)}$  in (1) boils down to the chain ratio estimator  $\bar{y}_R^{(dc)}$  suggested by Chand (1975) in double sampling. The bias and variance of  $\bar{y}_R^{(dc)}$  can be obtained by putting  $\alpha = 1$  in (5) and (7) respectively as

$$B\{\bar{y}_R^{(dc)}\}_I = \bar{Y}(P - Q).$$

and

$$V\{\bar{y}_R^{(dc)}\}_I = \bar{Y}^2 (M_{0,0} - 2Q + P). \quad (10)$$

**Remark 2.2.** Setting  $\alpha = 0$ , the estimator  $\bar{y}_{RdR}^{(dc)}$  in (1) reduces to the chain dual to ratio estimator  $\bar{y}_k^{(dc)}$  in double sampling. The bias and variance of  $\bar{y}_k^{(dc)}$  can be obtained by putting  $\alpha = 0$  in (5) and (7) respectively as

$$B\{\bar{y}_k^{(dc)}\}_I = -\bar{Y}n^*Q$$

and

$$V\{\bar{y}_k^{(dc)}\}_I = \bar{Y}^2 (M_{0,0} - 2n^*Q + n^{*2}P). \quad (11)$$

To the first degree of approximation, the variance of usual unbiased estimator  $\bar{y}$  is given by

$$V(\bar{y}) = \bar{Y}^2 M_{0,0} \quad (12)$$

and the variance of chain regression type estimator  $y_{reg.}^{(dc)} = \bar{y} + b_{yx} [\bar{x}_1 + b_{xz} (\bar{z} - \bar{z}') - \bar{x}]$ ,

where  $b_{yx}$  and  $b_{xz}$  are the regression coefficient of  $y$  on  $x$  and  $x$  on  $z$  respectively, suggested by Kiregyera (1984) is given by

$$V(y_{reg.}^{(dc)})_I = \bar{Y}^2 \left[ M_{0,0} - M_{0,1} C_{yx} + (M_{0,2} C_{yx} - 2M_{1,3} C_{yx} C_{yz} + M_{0,3} C_{yz} \rho_{xz}^2) \right] \quad (13)$$

From (10), (11), (12) and (13), it is envisaged that the proposed class of estimator  $\bar{y}_{RdR}^{(dc)}$  is better than:

- i. the usual chain ratio estimator  $\bar{y}_R^{(dc)}$  in double sampling if  
either  $1 < \alpha < 2Q/PN^{**} - N_1^*$   
or  $2Q/PN^{**} - N_1^* < \alpha < 1$

where  $N_1^* = n_1/(n_1 - 2n)$ .

- ii. the chain dual to ratio estimator  $\bar{y}_k^{(dc)}$  in double sampling if  
either  $0 < \alpha < 2Q/PN^{**} - 2n_1^*$   
or  $2Q/PN^{**} - 2n_1^* < \alpha < 0$ .

- iii. the usual unbiased estimator  $\bar{y}$  if  
either  $2Q/PN^{**} - n_1^* < \alpha < -n_1^*$   
or  $-n_1^* < \alpha < 2Q/PN^{**} - n_1^*$ .

- iv. the chain linear regression estimator  $y_{reg.}^{(dc)}$  if  
either  $(Q - \sqrt{Q^2 - PR})/PN^{**} - n_1^* < \alpha < (Q + \sqrt{Q^2 - PR})/PN^{**} - n_1^*$   
or  $(Q + \sqrt{Q^2 - PR})/PN^{**} - n_1^* < \alpha < (Q - \sqrt{Q^2 - PR})/PN^{**} - n_1^*$

where  $R = M_{0,1}C_{yx} - (M_{0,2}C_{yx} - 2M_{1,3}C_{yx}C_{yz} + M_{0,3}C_{yz}\rho_{xz}^2)$ .

Thus, it seems from the above results that the proposed estimator  $\bar{y}_{RdR}^{(dc)}$  may be made better than other estimators by making a suitable choice of the value of  $\alpha$ .

Also from (10), (11), (12) and (13), it is envisaged that the 'AOE'  $\{\bar{y}_{RdR}^{(dc)}\}_{lopt.}$  is better than:

- i. the usual chain ratio estimator  $\bar{y}_R^{(dc)}$ , since  
 $V\{\bar{y}_R^{(dc)}\}_I - V\{\bar{y}_{RdR}^{(dc)}\}_{lopt.} = \frac{\bar{Y}^2}{P}(P - Q)^2 > 0$ .
- ii. the chain dual to ratio estimator  $\bar{y}_k^{(dc)}$ , since  
 $V\{\bar{y}_k^{(dc)}\}_I - V\{\bar{y}_{RdR}^{(dc)}\}_{lopt.} = \frac{\bar{Y}^2}{P}(n^*P - Q)^2 > 0$ .
- iii. the usual unbiased estimator  $\bar{y}$ , since  
 $V\{\bar{y}\} - V\{\bar{y}_{RdR}^{(dc)}\}_{lopt.} = \frac{\bar{Y}^2}{P}Q^2 > 0$ .

- iv. the chain linear regression estimator  $y_{reg.}^{(dc)}$  if

$$M_{0,1}C_{yx} < (M_{0,2}C_{yx} - 2M_{1,3}C_{yx}C_{yz} + M_{0,3}C_{yz}\rho_{xz}^2).$$

Now, we state the following theorem:

**Theorem 2.1.** To the first degree of approximation, the proposed strategy under optimality condition (8), is always more efficient than  $V\{\bar{y}_R^{(dc)}\}$ ,  $V\{\bar{y}_k^{(dc)}\}$ ,  $V(\bar{y})$  and more efficient than  $V\{y_{reg}^{(dc)}\}$  if  $(M_{0,2}C_{yx} - 2M_{1,3}C_{yx}C_{yz} + M_{0,3}C_{yz}\rho_{xz}^2) > M_{0,1}C_{yx}$ .

**Case II:** When the second phase sample of size  $n$  is drawn independently of the first phase sample of size  $n_1$ , see Bose (1943). Under this condition, we get the following results

$$\left. \begin{aligned} E(e_0) &= E(e_1) = E(e'_1) = E(e'_2) = 0 \\ E(e_0^2) &= M_{0,0}, \quad E(e_1^2) = M_{1,0} \\ E(e_1'^2) &= M_{2,0}, \quad E(e_2'^2) = M_{3,0} \\ E(e_0e_1) &= M_{0,1}, \quad E(e'_1e'_2) = M_{2,3} \\ E(e_0e'_1) &= E(e_0e'_2) = E(e_1e'_1) = E(e_1e'_2) = 0 \end{aligned} \right\} \quad (14)$$

Taking the expectation in (3) and using the results in (14), we get the bias of  $\bar{y}_{RdR}^{(dc)}$  up to terms of order  $n^{-1}$  as

$$B\{\bar{y}_{RdR}^{(dc)}\}_{II} = \bar{Y} \left\{ -\omega M_{0,1} - (\alpha N^* - n^*) M_{2,3} + (\alpha - 1) n^* M_{2,0} + \alpha (M_{1,0} + M_{3,0}) \right\} \quad (15)$$

Squaring both sides in (6), then taking the expectation and using the results in (14), we obtain the variance of the estimator  $\bar{y}_{RdR}^{(dc)}$  to the terms of order  $n^{-1}$ , as

$$V\{\bar{y}_{RdR}^{(dc)}\} = \bar{Y}^2 \left\{ M_{0,0} - 2\omega M_{0,1} + \omega^2 (M_{1,0} + M_{2,0} + M_{3,0} - 2M_{2,3}) \right\} \quad (16)$$

Minimization of (16) with respect to  $\alpha$  yields its optimum value as

$$\alpha = M_{0,1} / SN^{**} - n_1^* = \alpha_{IIopt} \text{ (say)} \quad (17)$$

where  $S = M_{1,0} + M_{2,0} + M_{3,0} - 2M_{2,3}$ .

Thus, the resulting optimum variance of  $\bar{y}_{RdR}^{(dc)}$  is given by

$$V\{\bar{y}_{RdR}^{(dc)}\}_{IIopt.} = \bar{Y}^2 \left\{ M_{0,0} - \frac{(M_{0,1})^2}{S} \right\} \quad (18)$$

For simplicity we assume that the population size  $N$  is large enough as compared to the sample sizes  $n$  and  $n_1$  so that the finite population correction (FPC) terms  $1/N$  and  $2/N$  are ignored.

Ignoring the FPC in (16), the variance of  $\bar{y}_{RdR}^{(dc)}$  is given by

$$V\left\{\bar{y}_{RdR}^{(dc)}\right\}_{II} = \bar{Y}^2 \left\{M'_{0,0} - 2\omega M'_{0,1} + \omega^2 (M'_{1,0} + M'_{2,0} + M'_{3,0} - 2M'_{2,3})\right\} \quad (19)$$

where

$$(1/n)C_y^2 = M'_{0,0}, \quad (1/n)C_x^2 = M'_{1,0}, \quad (1/n_1)C_x^2 = M'_{2,0}, \quad (1/n_1)C_z^2 = M'_{3,0}, \\ (1/n)C_{yx}C_x^2 = M'_{0,1} \text{ and } (1/n_1)C_{xz}C_z^2 = M'_{2,3}.$$

Minimization of (19) with respect to  $\alpha$  yields its optimum value as

$$\alpha = M'_{0,1}/S'N^{**} - n_1^* = \alpha_{Ilopt.}^* \text{ (say).} \quad (20)$$

where  $S' = M'_{1,0} + M'_{2,0} + M'_{3,0} - 2M'_{2,3}$ .

Putting (20) in (1) yields the 'AOE' as

$$\left\{\bar{y}_{RdR}^{(dc^*)}\right\}_{Ilopt.} = \bar{y} \left[ \alpha_{Ilopt.}^* \frac{\bar{x}_1 \bar{Z}}{\bar{x} \bar{z}'} + (1 - \alpha_{Ilopt.}^*) \frac{\bar{x}^* \bar{z}'}{\bar{x}_1 \bar{Z}} \right]$$

Thus, the resulting variance of  $\left\{\bar{y}_{RdR}^{(dc^*)}\right\}_{Ilopt.}$  is given by

$$V\left\{\bar{y}_{RdR}^{(dc^*)}\right\}_{Ilopt.} = \bar{Y}^2 \left\{ M'_{0,0} - \frac{(M'_{0,1})^2}{S'} \right\}. \quad (21)$$

**Remark 2.3.** For  $\alpha = 1$ , the estimator  $\bar{y}_{RdR}^{(dc)}$  in (1) boils down to the chain ratio estimator  $\bar{y}_R^{(dc)}$  (Chand, 1975) in double sampling. Thus, putting  $\alpha = 1$  in (19), we get the variance of  $\bar{y}_R^{(dc)}$  to the first degree of approximation as

$$V\left\{\bar{y}_R^{(dc)}\right\}_{II} = \bar{Y}^2 \left\{ M'_{0,0} + M'_{1,0} + M'_{2,0} + M'_{3,0} - 2(M'_{0,1} + M'_{2,3}) \right\}. \quad (22)$$

**Remark 2.4.** For  $\alpha = 0$ , the estimator  $\bar{y}_{RdR}^{(dc)}$  in (1) boils down to the chain dual to ratio estimator  $\bar{y}_k^{(dc)}$  in double sampling. Setting  $\alpha = 0$  in (19), we get the variance of  $\bar{y}_k^{(dc)}$  to the first degree of approximation as

$$V\left\{\bar{y}_k^{(dc)}\right\}_{II} = \bar{Y}^2 \left( M'_{0,0} - 2n^* M'_{0,1} + n^{*2} S' \right). \quad (23)$$

Ignoring the FPC, the variance of sample mean  $\bar{y}$  under SRSWOR is given by

$$V(\bar{y}) = \bar{Y}^2 M'_{0,0}. \quad (24)$$

and chain regression type estimator  $y_{reg.}^{(dc)}$  suggested by Kiregyera (1984) in double sampling is

$$V(y_{reg.}^{(dc)})_{II} = \bar{Y}^2 \left\{ M'_{0,0} (1 - \rho_{yx}^2) + M'_{0,2} C_{yx} (1 - \rho_{xz}^2) \right\} \quad (25)$$

where  $(1/n_1) C_{yx} C_x^2 = M'_{0,2}$ .

From (22), (23), (24) and (25), it is observed that the proposed class of estimator  $\bar{y}_{RdR}^{(dc)}$  is better than:

- i. the usual chain ratio estimator  $\bar{y}_R^{(dc)}$  if  
 either  $(2M'_{0,1}/S'N^{**}) - N_1^* < \alpha < 1$   
 or  $1 < \alpha < (2M'_{0,1}/S'N^{**}) - N_1^*$ .
- ii. the chain dual to ratio estimator  $\bar{y}_k^{(dc)}$  if  
 either  $0 < \alpha < (2M'_{0,1}/S'N^{**}) - 2n_1^*$   
 or  $(2M'_{0,1}/S'N^{**}) - 2n_1^* < \alpha < 0$ .
- iii. the usual unbiased estimator  $\bar{y}$  if  
 either  $(M'_{0,1}/S'N^{**}) - n_1^* < \alpha < -n_1^*$   
 or  $-n_1^* < \alpha < (M'_{0,1}/S'N^{**}) - n_1^*$ .
- iv. the chain linear regression estimator  $y_{reg.}^{(dc)}$  if  
 either  $-n_1^* + \left( M'_{0,1} - \sqrt{(M'_{0,1})^2 - S'R'} \right) / S'N^{**} < \alpha < -n_1^* + \left( M'_{0,1} + \sqrt{(M'_{0,1})^2 - S'R'} \right) / S'N^{**}$   
 or  $-n_1^* + \left( M'_{0,1} + \sqrt{(M'_{0,1})^2 - S'R'} \right) / S'N^{**} < \alpha < -n_1^* + \left( M'_{0,1} - \sqrt{(M'_{0,1})^2 - S'R'} \right) / S'N^{**}$

where  $R' = M'_{0,0} \rho_{yx}^2 - M'_{0,2} C_{yx} (1 - \rho_{xz}^2)$ .

Thus, it seems from the above results that the proposed estimator  $\bar{y}_{RdR}^{(dc)}$  may be made better than other estimators by making a suitable choice of the value of  $\alpha$  in Case II.

Also from (22), (23), (24) and (25), it is envisaged that the 'AOE'  $\left\{ \bar{y}_{RdR}^{(dc*)} \right\}_{Ilopt.}$  is

better than

- i. the usual chain ratio estimator  $\bar{y}_R^{(dc)}$  if

$$M'_{1,0} + M'_{2,0} + M'_{3,0} > 2M'_{2,3}.$$

ii. the chain dual to ratio estimator  $\bar{y}_k^{(dc)}$  if

$$M'_{1,0} + M'_{2,0} + M'_{3,0} > 2M'_{2,3}.$$

iii. the usual unbiased estimator  $\bar{y}$  if

$$M'_{1,0} + M'_{2,0} + M'_{3,0} > 2M'_{2,3}.$$

iv. the chain linear regression estimator  $y_{reg.}^{(dc)}$  if

$$M'_{1,0} + M'_{2,0} + M'_{3,0} > 2M'_{2,3} \text{ and } M'_{0,0}\rho_{yx}^2 < M'_{0,2}C_{yx}(1 - \rho_{xz}^2).$$

Thus, we consider the following theorem:

**Theorem 2.2.** To terms of order  $n^{-1}$ , the proposed strategy under optimality condition (20) is always more efficient than  $V\{\bar{y}_R^{(dc)}\}$ ,  $V\{\bar{y}_k^{(dc)}\}$ ,  $V(\bar{y})$  if

$$M'_{1,0} + M'_{2,0} + M'_{3,0} > 2M'_{2,3} \text{ and more efficient than } V\{y_{reg.}^{(dc)}\} \text{ if}$$

$$M'_{1,0} + M'_{2,0} + M'_{3,0} > 2M'_{2,3} \text{ and } M'_{0,0}\rho_{yx}^2 < M'_{0,2}C_{yx}(1 - \rho_{xz}^2).$$

### 3. Cost aspect

The different estimators reported in this paper have so far been compared with respect to their variances. In practical applications, the cost aspect should also be taken into account. In the literature, therefore, convention is to fix the total cost of the survey and then to find optimum sizes of preliminary and final samples so that the variance of the estimator is minimized. In most of the practical situations, total cost is a linear function of samples selected at first and second phases.

In this section, we shall consider the cost of the survey and find the optimum sizes of the preliminary and second-phase samples in *Case I* and *Case II* separately.

**Case I:** When we use one auxiliary variate  $x$  then the cost function is given by

$$C = nC_1 + n_1C_2$$

Where  $C$ ,  $C_1$  and  $C_2$  are the total cost, cost per unit of collecting information on the study variate  $y$  and the cost per unit of collecting information on the auxiliary variate  $x$  respectively of the survey.

When we used additional auxiliary variate  $z$  to estimate  $\bar{y}_{RdR}^{(dc)}$ , then the cost function is given by

$$C = nC_1 + n_1(C_2 + C_3) \quad (26)$$

where  $C_3$  is the cost per unit collecting information on auxiliary variate  $z$ .

Ignoring FPC, the variance of  $\{\bar{y}_{RdR}^{(dc)}\}$  in (7) can be expressed as

$$V\{\bar{y}_{RdR}^{(dc)}\}_I = \frac{1}{n}V_1 + \frac{1}{n_1}V_2$$

Where

$$V_1 = \bar{Y}^2 (M'_{0,0} - 2\omega M'_{0,1} + \omega^2 M'_{1,0}), \quad V_2 = \bar{Y}^2 \{M'_{0,3} + 2\omega M'_{0,2} + \omega^2 (M'_{3,0} - M'_{2,0})\}$$

and  $(1/n_1)C_{yz}C_z^2 = M'_{0,3}$ .

It is assumed that  $C_1 > C_2 > C_3$ . The optimum values of  $n$  and  $n_1$  for fixed cost  $C = C_0$  which minimizes the variance of  $\bar{y}_{RdR}^{(dc)}$  at (7) under cost function are given by

$$n_{opt.} = \frac{C_0 \sqrt{V_1/C_1}}{\sqrt{V_1 C_1} + \sqrt{V_2 (C_2 + C_3)}} \quad \text{and} \quad n_{1opt.} = \frac{C_0 \sqrt{V_2/(C_2 + C_3)}}{\sqrt{V_1 C_1} + \sqrt{V_2 (C_2 + C_3)}}$$

Thus, the resulting variance of  $\bar{y}_{RdR}^{(dc)}$  is given by

$$V_{opt.} \{\bar{y}_{RdR}^{(dc)}\}_I = \frac{1}{C_0} \left\{ \sqrt{V_1 C_1} + \sqrt{V_2 (C_2 + C_3)} \right\}^2 \quad (27)$$

If all the resources were diverted towards the study variate  $y$  only, then we would have optimum sample size as below

$$n^{**} = C/C_1$$

Thus, the variance of sample mean  $\bar{y}$  for a given fixed cost  $C = C_0$  in case of large population is given by

$$V_{opt.}(\bar{y}) = \frac{C_1}{C_0} S_y^2. \quad (28)$$

From (27) and (28), the proposed sampling strategy would be profitable as long as

$$V_{opt.} \{\bar{y}_{RdR}^{(dc)}\}_I < V_{opt.}(\bar{y}).$$

or equivalently,

$$\frac{C_2 + C_3}{C_1} < \left[ \frac{S_y - \sqrt{V_1}}{\sqrt{V_2}} \right]^2.$$

**Case II:** we assume that  $y$  measured on  $n$  units,  $x$  and  $z$  are measured on  $n_1$  units. We consider a simple cost function

$$C = nC_1 + n_1(C'_2 + C'_3) \quad (29)$$

where  $C'_2$  and  $C'_3$  denote costs per unit of observing  $x$  and  $z$  values respectively.



The variance of  $\{\bar{y}_{RdR}^{(dc)}\}$  at (19) can be written as

$$V\{\bar{y}_{RdR}^{(dc)}\}_{II} = \frac{1}{n}V_1 + \frac{1}{n_1}V_3 \quad (30)$$

where  $V_3 = \bar{Y}^2 \omega^2 (M'_{3,0} + M'_{2,0} - 2M'_{2,3})$ .

To obtain the optimum allocation of sample between phases for a fixed cost  $C = C_0$ , we minimized (30) with condition (29). It is easily found that this minimum is attained for

$$n_{opt.} = \frac{C_0 \sqrt{V_1/C_1}}{\sqrt{V_1 C_1} + \sqrt{V_3 (C'_2 + C'_3)}} \text{ and } n_{1opt.} = \frac{C_0 \sqrt{V_3/(C'_2 + C'_3)}}{\sqrt{V_1 C_1} + \sqrt{V_3 (C'_2 + C'_3)}}.$$

Thus, the optimum variance corresponding to these optimum values of  $n$  and  $n_1$  are given by

$$V_{opt.}\{\bar{y}_{RdR}^{(dc)}\}_{II} = \frac{1}{C_0} \left\{ \sqrt{V_1 C_1} + \sqrt{V_3 (C'_2 + C'_3)} \right\}^2 \quad (31)$$

From (28) and (31) it is obtained that the proposed estimator  $\bar{y}_{RdR}^{(dc)}$  yields less variance than that of sample mean  $\bar{y}$  for the same fixed cost if

$$\frac{C'_2 + C'_3}{C_1} < \left[ \frac{S_y - \sqrt{V_1}}{\sqrt{V_3}} \right]^2.$$

#### 4. Empirical study

To examine the merits of the proposed estimator, we have considered the six natural population data sets. The description of the populations are given as follows.

Population I- Source: Cochran (1977)

$Y$  : Number of 'placebo' children,  $X$  : Number of paralytic polio cases in the placebo group,  $Z$  : Number of paralytic polio cases in the 'not inoculated' group.

$N = 34$ ,  $n = 10$ ,  $n_1 = 15$ ,  $\bar{Y} = 4.92$ ,  $\bar{X} = 2.59$ ,  $\bar{Z} = 2.91$ ,  $\rho_{yx} = 0.7326$ ,  $\rho_{yz} = 0.6430$ ,  $\rho_{xz} = 0.6837$ ,  $C_y^2 = 1.0248$ ,  $C_x^2 = 1.5175$ ,  $C_z^2 = 1.1492$ .

Population II- Source: Sukhatme and Chand (1977)

$Y$  : Apple trees of bearing age in 1964,  $X$  : Bushels of apples harvested in 1964,  $Z$  : Bushels of apples harvested in 1959.

$N = 200$ ,  $n = 20$ ,  $n_1 = 30$ ,  $\bar{Y} = 0.103182 \times 10^4$ ,  $\bar{X} = 0.293458 \times 10^4$ ,  $\bar{Z} = 0.365149 \times 10^4$ ,  $\rho_{yx} = 0.93$ ,  $\rho_{yz} = 0.77$ ,  $\rho_{xz} = 0.84$ ,  $C_y^2 = 2.55280$ ,  $C_x^2 = 4.02504$ ,  $C_z^2 = 2.09379$ .

Population III- Source: Murthy (1967)

$Y$  : Area under wheat in 1964,  $X$  : Area under wheat in 1963,  $Z$  : Cultivated area in 1961.

$N = 34$ ,  $n = 7$ ,  $n_1 = 10$ ,  $\bar{Y} = 199.44$  acre,  $\bar{X} = 208.89$  acre,  $\bar{Z} = 747.59$  acre,  
 $\rho_{yx} = 0.9801$ ,  $\rho_{yz} = 0.9043$ ,  $\rho_{xz} = 0.9097$ ,  $C_y^2 = 0.5673$ ,  $C_x^2 = 0.5191$ ,  $C_z^2 = 0.3527$ .

Population IV- Source: Srivastava *et al.* (1989, Page 3922)

$Y$  : The measurement of weight of children,  $X$  : Mid arm circumference of children

$Z$  : Skull circumference of children.

$N = 82$ ,  $n = 25$ ,  $n_1 = 43$ ,  $\bar{Y} = 5.60$  kg,  $\bar{X} = 11.90$  cm,  $\bar{Z} = 39.80$  cm,  $\rho_{yx} = 0.09$ ,  
 $\rho_{yz} = 0.12$ ,  $\rho_{xz} = 0.86$ ,  $C_y^2 = 0.0107$ ,  $C_x^2 = 0.0052$ ,  $C_z^2 = 0.0008$ .

Population V- Source: Srivastava *et al.* (1989, Page 3922)

$Y$  : The measurement of weight of children,  $X$  : Mid arm circumference of children

$Z$  : Skull circumference of children.

$N = 55$ ,  $n = 18$ ,  $n_1 = 30$ ,  $\bar{Y} = 17.08$  kg,  $\bar{X} = 16.92$  cm,  $\bar{Z} = 50.44$  cm,  $\rho_{yx} = 0.54$ ,  
 $\rho_{yz} = 0.51$ ,  $\rho_{xz} = -0.08$ ,  $C_y^2 = 0.0161$ ,  $C_x^2 = 0.0049$ ,  $C_z^2 = 0.0007$ .

Population VI- Source: Srivastava *et al.* (1990)

$N = 120$ ,  $n = 30$ ,  $n_1 = 60$ ,  $\bar{Y} = 2934.58$ ,  $\bar{X} = 1031.82$ ,  $\bar{Z} = 3651.49$ ,  $\rho_{yx} = 0.93$ ,  
 $\rho_{yz} = 0.84$ ,  $\rho_{xz} = 0.77$ ,  $C_y^2 = 4.02004$ ,  $C_x^2 = 2.55280$ ,  $C_z^2 = 2.09379$ .

To reflect the gain in the efficiency of the proposed estimator  $\bar{y}_{RdR}^{(dc)}$  over the estimators  $\bar{y}$ ,

$\bar{y}_R^{(dc)}$ ,  $\bar{y}_k^{(dc)}$  and  $\bar{y}_{reg}^{(dc)}$ , the effective ranges and optimum value of  $\alpha$  are shown in Table 1 with respect to the above population data sets.

To observe the relative performance of different estimators of  $\bar{Y}$ , we have computed the percent relative efficiencies (PRE) of the proposed estimator  $\bar{y}_{RdR}^{(dc)}$ , usual ratio estimator  $\bar{y}_R^{(dc)}$ , chain dual to ratio estimator  $\bar{y}_k^{(dc)}$  and linear regression estimator  $\bar{y}_{reg}^{(dc)}$  with respect to usual unbiased estimator  $\bar{y}$  in *Case I* and *Case II* and the findings are presented in Table 2.

**Table 1.** Effective ranges and optimum values of  $\alpha$  of  $\overline{y}_{RdR}^{(dc)}$

Popu- lation	Ranges of $\alpha$ in which $\overline{y}_{RdR}^{(dc)}$ is better than				Optimum value of $\alpha$
	$\overline{y}_R^{(dc)}$	$\overline{y}_k^{(dc)}$	$\overline{y}$	$y_{reg.}^{(dc)}$	
<b>Case I</b>					$\alpha_{lopt}^*$
I	(1.00, 1.79)	(0.00, 2.79)	(0.79, 2.00)	*	1.3956
II	(1.00, 1.42)	(0.00, 2.42)	(0.42, 2.00)	*	1.2079
III	(0.87, 1.00)	(0.00, 1.87)	(0.12, 1.75)	*	0.9331
IV	(1.00, 5.33)	(0.00, 6.33)	(2.76, 3.57)	*	3.1660
V	(0.56, 1.00)	(0.00, 1.56)	(-1.44, 3.00)	*	0.7819
VI	(1.00, 101.02)	(0.00, 102.02)	(-300.00, 402.02)	*	51.0103
<b>Case II</b>					$\alpha_{lopt}^*$
I	(1.00, 2.13)	(0.00, 3.13)	(1.13, 2.00)	*	1.5632
II	(1.00, 1.77)	(0.00, 2.77)	(0.77, 2.00)	*	1.3857
III	(1.00, 0.07)	(0.00, 1.07)	(-0.68, 1.75)	*	0.5366
IV	(1.00, 6.06)	(0.00, 7.06)	(3.49, 3.57)	*	3.5322
V	(1.00, 4.43)	(0.00, 5.43)	(2.43, 3.00)	*	2.7158
VI	(-600.99, 1.00)	(-599.99, 0.00)	(-300.00, -299.99)	*	-299.9968

\* Data is not applicable.

**Table 2.** Percent relative efficiencies of different estimators with respect to  $\overline{y}$

Popu- lation	Efficiencies of the proposed estimator $\overline{y}_{RdR}^{(dc)}$ belonging to the class								
	$\overline{y}$	$\overline{y}_R^{(dc)}$	$\overline{y}_R^{(dc)}$	$\overline{y}_k^{(dc)}$	$\overline{y}_k^{(dc)}$	$\overline{y}_{reg.}^{(dc)}$	$\overline{y}_{reg.}^{(dc)}$	$\left\{ \overline{y}_{RdR}^{(dc)} \right\}$	$\left\{ \overline{y}_{RdR}^{(dc)} \right\}$
								or	or
								$\left\{ \overline{y}_{RdR}^{(dc)} \right\}_I$	$\left\{ \overline{y}_{RdR}^{(dc^*)} \right\}$
Case I	Case II	Case I	Case II	Case I	Case II	Case I	Case II		
I	100.00	136.91	79.50	32.86	17.87	185.53	152.94	189.27	163.78
II	100.00	279.93	176.85	52.21	25.43	326.33	328.03	322.94	353.81
III	100.00	730.78	644.74	79.06	44.77	778.93	643.67	763.27	681.36
IV	100.00	81.92	66.85	67.54	49.40	101.00	100.69	100.81	100.64
V	100.00	131.91	107.73	127.21	77.79	118.31	113.35	132.32	120.39
VI	100.00	488.98	347.59	487.39	347.81	517.80	321.69	531.85	348.90

## 5. Conclusions

From Table 1, it is clear that the proposed estimator  $\bar{y}_{RdR}^{(dc)}$  is more efficient than the conventional estimators  $\bar{y}$ ,  $\bar{y}_R^{(dc)}$ ,  $\bar{y}_k^{(dc)}$  and  $\bar{y}_{reg.}^{(dc)}$  for both the cases under the effective ranges of  $\alpha$  as far as the variance criterion is considered. We have also computed the percent relative efficiencies of different estimators with respect to  $\bar{y}$  as shown in Table 2.

Table 2 clearly indicates that there is considerable gain in efficiency by using the estimators  $\bar{y}_{RdR}^{(dc)}$  or  $\left[ \left\{ \bar{y}_{RdR}^{(dc)} \right\}_{Iopt.} \text{ and } \left\{ \bar{y}_{RdR}^{(dc^*)} \right\}_{IIopt.} \right]$  over the estimators  $\bar{y}$ ,  $\bar{y}_R^{(dc)}$ ,  $\bar{y}_k^{(dc)}$  and  $\bar{y}_{reg.}^{(dc)}$  in both the cases except for the data set of population II and III for linear regression estimator  $\bar{y}_{reg.}^{(dc)}$  in *Case I*. It is due to poor correlation between  $y$  and  $x$ ,  $y$  and  $z$ ,  $x$  and  $z$ . It is also observed that overall performances of the proposed estimator  $\bar{y}_{RdR}^{(dc)}$  in *Case I* is better than *Case II*. Thus, it is preferred to use the proposed estimators  $\bar{y}_{RdR}^{(dc)}$  or  $\left[ \left\{ \bar{y}_{RdR}^{(dc)} \right\}_{Iopt.} \text{ and } \left\{ \bar{y}_{RdR}^{(dc^*)} \right\}_{IIopt.} \right]$  in practice.

## REFERENCES

- AL-JARARHA, J., AHMED, M. S., 2002. The class of chain estimators for a finite population variance using double sampling. *Information and Management Sciences* 13(2), 13–18.
- BOSE, C., 1943. Note on the sampling error in the method of double sampling. *Sankhya* 6, 330.
- CHAND, L., 1975. Some ratio type estimator based on two or more auxiliary variables. *Unpublished Ph. D. dissertation, Iowa State University, Ames, Iowa.*
- KADILAR, C., CINGI, H., 2003. A study on the chain ratio-type estimator. *Hacettepe Journal of Mathematics and Statistics* 32, 105-108.
- KIREGYERA, B., 1984. Regression type estimators using two auxiliary variables and the model of double sampling from finite population. *Metrika* 31, 215-226.
- KUMAR, M., BAHL, S., 2006. Class of dual to ratio estimators for double sampling. *Statistical Papers* 47, 319-326.
- PRADHAN, B. K., 2005. A Chain Regression Estimator in Two Phase Sampling Using Multi-auxiliary Information. *Bulletin of the Malaysian Mathematical Sciences Society* (2) 28(1), 81–86.
- SHARMA, B., TAILOR, R., 2010. A New Ratio-Cum-Dual to Ratio Estimator of Finite Population Mean in Simple Random Sampling. *Global Journal of Science Frontier Research* 10(1), 27-31.
- Singh, H. P., Espejo, M. R., 2007. Double sampling ratio-product estimator of a finite population mean in sampling surveys. *Journal of Applied Statistics* 34(1), 71-85.
- SINGH, H. P., MATHUR, N., CHANDRA, P., 2009. A chain-type estimator for population variance using auxiliary variables in two-phase sampling. *Statistics in Transition-new series* 10(1), 75-84.
- SINGH, S., SINGH, H. P., UPADHYAYA, L. N., 2006. Chain ratio and regression type estimators for median estimation in survey sampling. *Statistical Papers* 48, 23-46.
- SINGH, V. K., SINGH, B. K., SINGH, G. N., 1993. An efficient class of dual to ratio estimators using two auxiliary characteristics. *Journal of Scientific Research* 43, 219-228.

- SINGH, V. K., SINGH, G. N., SHUKLA, D., 1994. A class of chain ratio type estimators with two auxiliary variables under double sampling scheme. *Sankhya: The Indian Journal of Statistics* 56, Series B 2, 209-221.
- SRIVENKATARAMANA, T., 1980. A dual to ratio estimator in sample surveys. *Biometrika* 67 (1), 199-204.
- SRIVASTAVA, S. K., JHAJJ, H. S., 1980. A class of estimators using auxiliary information for estimating finite population variance. *Sankhya* 42, 87-96.
- SRIVASTAVA, S. R., KHARE, B. B., SRIVASTAVA, S. R., 1990. A Generalised Chain Ratio Estimator for Mean of Finite Population. *Journal of Indian Society of Agricultural Statistics* 42(I), 108-117.
- SUKHATME, B. V., CHAND, L., 1977. Multivariate ratio-type estimators, *Proceedings, Social Statistics section, American Statistical Association* 927-931.

STATISTICS IN TRANSITION-new series, December 2012  
Vol. 13, No. 3, pp. 537—550

## ALMOST UNBIASED RATIO AND PRODUCT TYPE EXPONENTIAL ESTIMATORS

Rohini Yadav<sup>1</sup>, Lakshmi N. Upadhyaya<sup>1</sup>,  
Housila P. Singh<sup>2</sup>, S. Chatterjee<sup>1</sup>

### ABSTRACT

This paper considers the problem of estimating the population mean  $\bar{Y}$  of the study variate  $y$  using information on auxiliary variate  $x$ . We have suggested a generalized version of Bahl and Tuteja (1991) estimator and its properties are studied. It is found that asymptotic optimum estimator (AOE) in the proposed generalized version of Bahl and Tuteja (1991) estimator is biased. In some applications, biasedness of an estimator is disadvantageous. So applying the procedure of Singh and Singh (1993) we derived an almost unbiased version of AOE. A numerical illustration is given in the support of the present study.

**Key words:** study variable, auxiliary variable, almost unbiased ratio-type and product-type exponential estimators, bias, mean squared error.

### 1. Introduction

It is well known that the use of auxiliary information provides efficient estimates of the population parameters. Ratio, regression and product methods of estimation are good illustrations in this context.

Let there be  $N$  units in population and information be available on auxiliary variable  $x$ . We draw a sample of size  $n$  using simple random sampling without replacement (SRSWOR) scheme and on the basis of which we can estimate the population mean of the character  $y$  under study. Let  $(\bar{y}, \bar{x})$  be the sample means of  $(y, x)$  respectively.

---

<sup>1</sup> Department of Applied Mathematics, Indian School of Mines, Dhanbad-826004, India.  
Email: rohiniyadav.ism@gmail.com, lnupadhyaya@yahoo.com, schat\_1@yahoo.co.in.

<sup>2</sup> School of Studies in Statistics, Vikram University, Ujjain-456 010, India. Email: hpsujn@gmail.com.

Assume that we have information about the population mean  $\bar{X}$  of the auxiliary character  $x$  and using  $\bar{X}$  it is desired to estimate the population mean  $\bar{Y}$  of the study character  $y$ .

When the correlation between the study variable  $y$  and the auxiliary variable  $x$  is positive, the classical ratio estimator for population mean  $\bar{Y}$  is defined by

$$\bar{y}_R = \bar{y} \frac{\bar{X}}{\bar{x}} \quad (1.1)$$

In such a situation, Bahl and Tuteja (1991) suggested a ratio-type exponential estimator  $t_1$  for the population mean  $\bar{Y}$  as

$$t_1 = \bar{y} \exp \left( \frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}} \right) \quad (1.2)$$

To the first degree of approximation, the biases and mean squared errors (MSEs) of  $\bar{y}_R$  and  $t_1$  are respectively given by

$$B(\bar{y}_R) = f_1 \bar{Y} C_x^2 (1-K) \quad (1.3)$$

$$B(t_1) = f_1 \bar{Y} \frac{C_x^2}{8} (3-4K) \quad (1.4)$$

$$MSE(\bar{y}_R) = f_1 \bar{Y}^2 [C_y^2 + C_x^2 (1-2K)] \quad (1.5)$$

$$MSE(t_1) = f_1 \bar{Y}^2 \left[ C_y^2 + \frac{C_x^2}{4} (1-4K) \right] \quad (1.6)$$

where  $f_1 = \left( \frac{1}{n} - \frac{1}{N} \right)$ ,  $C_y = \frac{S_y}{\bar{Y}}$ ,  $C_x = \frac{S_x}{\bar{X}}$ ,  $K = \rho_{yx} \frac{C_y}{C_x}$ ,

$$S_y^2 = \frac{1}{(N-1)} \sum_{i=1}^N (y_i - \bar{Y})^2, \quad S_x^2 = \frac{1}{(N-1)} \sum_{i=1}^N (x_i - \bar{X})^2,$$

$$S_{xy} = \frac{1}{(N-1)} \sum_{i=1}^N (x_i - \bar{X})(y_i - \bar{Y})$$

and  $\rho_{yx}$  is the correlation coefficient between  $y$  and  $x$ .

It is observed from (1.3) and (1.4) that the estimator  $t_1$  due to Bahl and Tuteja (1991) is less biased than the classical ratio estimator  $\bar{y}_R$  if

$$|B(t_1)| < |B(\bar{y}_R)|$$



$$\begin{aligned} \text{i.e. if } \left| \frac{1}{8}(3-4K) \right| &< |(1-K)| \\ \text{i.e. if } (48K^2 - 104K + 55) &> 0 \end{aligned} \quad (1.7)$$

From (1.5) and (1.6) it follows that the estimator  $t_1$  due to Bahl and Tuteja (1991) is more efficient than the classical ratio estimator  $\bar{y}_R$  if

$$\begin{aligned} \text{MSE}(t_1) &< \text{MSE}(\bar{y}_R) \\ \text{if } K &< \frac{3}{4} \quad \text{or} \quad \rho_{yx} < \frac{3}{4} \frac{C_x}{C_y} \end{aligned} \quad (1.8)$$

Murthy (1964) suggested a product-type estimator

$$\bar{y}_p = \bar{y} \frac{\bar{x}}{\bar{X}} \quad (1.9)$$

for the population mean  $\bar{Y}$  which is useful in the situation where the correlation between the study variable  $y$  and the auxiliary variable  $x$  is negative (high).

In negative correlation situation, Bahl and Tuteja (1991) suggested a product-type exponential estimator  $t_2$  for the population mean  $\bar{Y}$  is defined as

$$t_2 = \bar{y} \exp \left( \frac{\bar{x} - \bar{X}}{\bar{x} + \bar{X}} \right) \quad (1.10)$$

To the first degree of approximation, the biases and mean squared errors (MSEs) of  $\bar{y}_p$  and  $t_2$  are respectively given by

$$B(\bar{y}_p) = f_1 \bar{Y} C_x^2 K \quad (1.11)$$

$$B(t_2) = f_1 \bar{Y} \frac{C_x^2}{8} (4K - 1) \quad (1.12)$$

$$\text{MSE}(\bar{y}_p) = f_1 \bar{Y}^2 [C_y^2 + C_x^2 (1 + 2K)] \quad (1.13)$$

$$\text{MSE}(t_2) = f_1 \bar{Y}^2 \left[ C_y^2 + \frac{C_x^2}{4} (1 + 4K) \right] \quad (1.14)$$

It is observed from (1.11) and (1.12) that the product-type exponential estimator  $t_2$  is less biased than the usual product estimator  $\bar{y}_p$  if

$$|B(t_2)| < |B(\bar{y}_p)|$$

$$\begin{aligned} \text{i.e. if } \left| \frac{1}{8}(4K-1) \right| &< |(K)| \\ \text{i.e. if } (48K^2+8K-1) &> 0 \end{aligned} \quad (1.15)$$

From (1.13) and (1.14), it follows that the Bahl and Tuteja (1991) product-type exponential estimator  $t_2$  is more efficient than the product estimator  $\bar{y}_p$  if

$$\begin{aligned} \text{MSE}(t_2) &< \text{MSE}(\bar{y}_p) \\ \text{if } K &> -\frac{3}{4} \quad \text{or} \quad \rho_{yx} > -\frac{3}{4} \frac{C_x}{C_y} \end{aligned} \quad (1.16)$$

## 2. Generalized version of $t_1$ and $t_2$

We define a generalized version of  $t_1$  and  $t_2$  as

$$t_c = \bar{y} \exp \left[ c \left( \frac{\bar{X} - x}{\bar{X} + x} \right) \right] \quad (2.1)$$

where  $c$  is 'non-zero' constant. For  $c=1$ ,  $t_c$  reduces to  $t_1$  while for  $c=-1$  it reduces to  $t_2$ .

To obtain the bias and mean squared error (MSE) of the estimator  $t_c$  up to the first degree of approximation, we write

$$\bar{y} = \bar{Y}(1+e_0) \quad \text{and} \quad \bar{x} = \bar{X}(1+e_1)$$

such that  $E(e_0) = E(e_1) = 0$ ,

$$E(e_0^2) = f_1 C_Y^2, \quad E(e_1^2) = f_1 C_X^2 \quad \text{and} \quad E(e_0 e_1) = f_1 \rho_{yx} C_Y C_X.$$

Expressing  $t_c$  in terms of  $e$ 's, we have

$$t_c = \bar{Y}(1+e_0) \exp \left[ c \left( \frac{-e_1}{2+e_1} \right) \right]$$

Now, expanding the right hand side of the above, multiplying out and neglecting terms of  $e$ 's having power greater than two, we have

$$t_c = \bar{Y} \left[ 1 + e_0 - \frac{c}{2} e_1 - \frac{c}{2} e_0 e_1 + \frac{c}{4} e_1^2 + \frac{c^2}{8} e_1^2 \right]$$

or

$$(t_c - \bar{Y}) = \bar{Y} \left[ e_0 - \frac{c}{2} e_1 - \frac{c}{2} e_0 e_1 + \frac{c}{4} e_1^2 + \frac{c^2}{8} e_1^2 \right] \quad (2.2)$$

Taking expectations of both sides of (2.2), we get the bias of the estimator  $t_c$  up to the first order of approximation as

$$B(t_c) = f_1 \bar{Y} \frac{c}{2} C_x^2 (2 + c - 4K) \quad (2.3)$$

Squaring both sides of (2.2) and neglecting terms of  $e$ 's having power greater than two, we have

$$(t_c - \bar{Y})^2 = \bar{Y}^2 \left( e_0 - \frac{c}{2} e_1 \right)^2 \quad (2.4)$$

Taking expectation of both sides of (2.4), we get mean squared error (MSE) of the estimator  $t_c$  up to the first order of approximation as

$$MSE(t_c) = f_1 \bar{Y}^2 \left[ C_y^2 + c C_x^2 \left( \frac{c}{4} - K \right) \right] \quad (2.5)$$

Differentiating (2.5) w.r.t.  $c$  and equating it to zero, we get the optimum value of  $c$  as

$$c = 2K \quad (2.6)$$

Thus, the substitution of optimum value  $c=2K$  in (2.1) yields the asymptotic optimum estimator (AOE) in the class of estimator  $t_c$  as

$$t_K = \bar{y} \exp \left[ 2K \left( \frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}} \right) \right] \quad (2.7)$$

The bias and mean squared error (MSE) of the estimator  $t_K$  are respectively given by

$$B(t_K) = f_1 \bar{Y} C_x^2 \left\{ K(1-K)/2 \right\} \quad (2.8)$$

$$MSE(t_K) = f_1 \bar{Y}^2 C_Y^2 (1 - \rho^2) \quad (2.9)$$

It is observed from (2.9) that the MSE of the AOE  $t_K$  is equal to the approximate variance of the regression estimator  $\bar{y}_{lr} = \bar{y} + \hat{\beta}(\bar{X} - \bar{x})$ , which is biased, where  $\hat{\beta}$  is sample estimate of the population regression coefficient  $\beta$ . Expression (2.8) clearly indicates that the AOE is a biased estimator. So our objective is to obtain an almost unbiased estimator for the population mean  $\bar{Y}$ . In the following section, we meet with our objective using Singh and Singh (1993) approach.

### 3. Almost unbiased exponential estimator

We consider the estimators

$$t_1 = \bar{y} \exp \left[ 2K \left( \frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}} \right) \right] \quad (3.1)$$

$$t_2 = \bar{y} \exp \left[ 4K \left( \frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}} \right) \right] \quad (3.2)$$

$$t_3 = \bar{y} \exp \left[ 6K \left( \frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}} \right) \right] \quad (3.3)$$

such that  $t_1, t_2, t_3 \in H$ , where  $H$  denotes the set of all possible estimators for estimating the population mean  $\bar{Y}$ .

To the first degree of approximation, the biases and mean squared errors (MSEs) of the estimators  $t_1, t_2$  and  $t_3$  are respectively given by

$$B(t_1) = f_1 \bar{Y} (K/2) C_x^2 (1-K). \quad (3.4)$$

$$B(t_2) = f_1 \bar{Y} K C_x^2 \quad (3.5)$$

$$B(t_3) = f_1 \bar{Y} (3/2) K C_x^2 (1-K) \quad (3.6)$$

$$MSE(t_1) = f_1 \bar{Y}^2 C_y^2 (1-\rho^2) \quad (3.7)$$

$$MSE(t_2) = f_1 \bar{Y}^2 C_y^2 \quad (3.8)$$

$$MSE(t_3) = f_1 \bar{Y}^2 [C_y^2 + 3K^2 C_x^2] \quad (3.9)$$

Now, considering the estimators (3.1), (3.2) and (3.3), we suggest a class of exponential estimators for  $\bar{Y}$  as

$$t_{Kh} = \sum_{j=1}^3 h_j t_j \in H \quad (3.10)$$

$$\text{with } \sum_{j=1}^3 h_j = 1, h_j \in R, \quad (3.11)$$

where  $h_j$  ( $j=1, 2, 3$ ) denotes the statistical constants and  $R$  denotes the set of real numbers.

Expressing  $t_{kh}$  in terms of  $e$ 's, we have

$$\begin{aligned}
 t_{kh} &= \left[ h_1 \bar{Y} (1+e_0) \exp \left\{ -K e_1 \left( 1 + \frac{e_1}{2} \right)^{-1} \right\} + h_2 \bar{Y} (1+e_0) \exp \left\{ -2K e_1 \left( 1 + \frac{e_1}{2} \right)^{-1} \right\} \right. \\
 &\quad \left. + h_3 \bar{Y} (1+e_0) \exp \left\{ -3K e_1 \left( 1 + \frac{e_1}{2} \right)^{-1} \right\} \right] \\
 &= \bar{Y} \left[ h_1 (1+e_0) \left\{ 1 - K e_1 \left( 1 + \frac{e_1}{2} \right)^{-1} + \frac{K^2 e_1^2}{2} \left( 1 + \frac{e_1}{2} \right)^{-2} - \dots \right\} \right. \\
 &\quad \left. + h_2 (1+e_0) \left\{ 1 - 2K e_1 \left( 1 + \frac{e_1}{2} \right)^{-1} + 2K^2 e_1^2 \left( 1 + \frac{e_1}{2} \right)^{-2} - \dots \right\} \right. \\
 &\quad \left. + h_3 (1+e_0) \left\{ 1 - 3K e_1 \left( 1 + \frac{e_1}{2} \right)^{-1} + \frac{9}{2} K^2 e_1^2 \left( 1 + \frac{e_1}{2} \right)^{-2} - \dots \right\} \right] \\
 &= \bar{Y} \left[ h_1 (1+e_0) \left\{ 1 - K e_1 \left( 1 - \frac{e_1}{2} + \frac{e_1^2}{8} - \dots \right) + \frac{K^2 e_1^2}{2} (1 - e_1 + \dots) - \dots \right\} \right. \\
 &\quad \left. + h_2 (1+e_0) \left\{ 1 - 2K e_1 \left( 1 - \frac{e_1}{2} + \frac{e_1^2}{8} - \dots \right) + 2K^2 e_1^2 (1 - e_1 + \dots) - \dots \right\} \right. \\
 &\quad \left. + h_3 (1+e_0) \left\{ 1 - 3K e_1 \left( 1 - \frac{e_1}{2} + \frac{e_1^2}{8} - \dots \right) + \frac{9}{2} K^2 e_1^2 (1 - e_1 + \dots) - \dots \right\} \right] \\
 &= \bar{Y} \left[ h_1 \left\{ 1 + e_0 - K e_1 \left( 1 - \frac{e_1}{2} + \frac{e_1^2}{8} - \dots \right) + \frac{K^2 e_1^2}{2} (1 - e_1 + \dots) - K e_0 e_1 \left( 1 - \frac{e_1}{2} + \frac{e_1^2}{8} - \dots \right) + \dots \right\} \right. \\
 &\quad \left. + h_2 \left\{ 1 + e_0 - 2K e_1 \left( 1 - \frac{e_1}{2} + \frac{e_1^2}{8} - \dots \right) + 2K^2 e_1^2 (1 - e_1 + \dots) - 2K e_0 e_1 \left( 1 - \frac{e_1}{2} + \frac{e_1^2}{8} - \dots \right) + \dots \right\} \right. \\
 &\quad \left. + h_3 \left\{ 1 + e_0 - 3K e_1 \left( 1 - \frac{e_1}{2} + \frac{e_1^2}{8} - \dots \right) + \frac{9}{2} K^2 e_1^2 (1 - e_1 + \dots) - 3K e_0 e_1 \left( 1 - \frac{e_1}{2} + \frac{e_1^2}{8} - \dots \right) + \dots \right\} \right]
 \end{aligned}$$

Neglecting the terms of  $e$ 's having power greater than two, we have

$$t_{kh} = \bar{Y} \left[ 1 + e_0 - hK(e_1 + e_0 e_1) + \frac{Ke_1^2}{2} \{h + (h_1 + 4h_2 + 9h_3)K\} \right] \quad (3.12)$$

or

$$(t_{kh} - \bar{Y}) = \bar{Y} \left[ e_0 - hK(e_1 + e_0 e_1) + \frac{Ke_1^2}{2} \{h + (h_1 + 4h_2 + 9h_3)K\} \right] \quad (3.13)$$

$$\text{where } (h_1 + 2h_2 + 3h_3) = h \text{ (a constant).} \quad (3.14)$$

Taking expectation of both sides of (3.13), we get the bias of  $t_{kh}$  to the first degree of approximation as

$$B(t_{kh}) = f_1 \bar{Y} \left( \frac{KC_x^2}{2} \right) [(1-2K)h + K(h_1 + 4h_2 + 9h_3)] \quad (3.15)$$

Squaring both sides of (3.13) and neglecting terms of  $e$ 's having power greater than two, we have

$$(t_{kh} - \bar{Y})^2 = \bar{Y}^2 [e_0^2 + h^2 K^2 e_1^2 - 2hKe_0 e_1] \quad (3.16)$$

Taking expectation of both sides of (3.16), we get the MSE of  $t_{kh}$  to the first degree of approximation as

$$\text{MSE}(t_{kh}) = f_1 \bar{Y}^2 [C_y^2 + h^2 K^2 C_x^2 - 2hkpC_y C_x] \quad (3.17)$$

Minimizing (3.17) with respect to  $h$ , we get the optimum value of  $h$  as

$$(h_1 + 2h_2 + 3h_3) = h = 1 \quad (3.18)$$

Substitution of (3.18) in (3.17) yields minimum MSE of  $t_{kh}$  as

$$\min. \text{MSE}(t_{kh}) = f_1 \bar{Y}^2 C_y^2 (1 - \rho^2) \quad (3.19)$$

In order to get unique solution of  $h_j$ 's ( $j=1, 2, 3$ ), we shall impose the linear restriction as we have only two equations in three unknowns.

$$\sum_{j=1}^3 h_j B(t_j) = 0, \quad (3.20)$$

where  $B(t_j)$  represents the bias of the  $j^{\text{th}}$  estimator.

So, we have three equations (3.11), (3.18) and (3.20) with three unknowns. These can be written in the matrix form as

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ B(t_1) & B(t_2) & B(t_3) \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \quad (3.21)$$

Using (3.21), we get the unique values of  $h_j$ 's ( $j=1, 2, 3$ ) as

$$\left. \begin{aligned} h_1 &= \left( \frac{3}{4} + \frac{1}{4K} \right) \\ h_2 &= \left( \frac{1}{2} - \frac{1}{2K} \right) \\ h_3 &= \left( -\frac{1}{4} + \frac{1}{4K} \right) \end{aligned} \right\} \quad (3.22)$$

Using these  $h_j$ 's ( $j=1, 2, 3$ ), we can remove the bias of the estimator  $t_c$  up to the terms of order  $o(n^{-1})$ .

Thus, an almost unbiased exponential estimator for population mean  $\bar{Y}$  is defined as

$$t_k^{(u)} = \bar{y} \left[ \frac{(3K+1)}{4K} \exp \left\{ 2K \left( \frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}} \right) \right\} + \frac{(K-1)}{2K} \exp \left\{ 4K \left( \frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}} \right) \right\} + \frac{(1-K)}{4K} \exp \left\{ 6K \left( \frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}} \right) \right\} \right] \quad (3.23)$$

It can be shown to the first degree of approximation that the mean squared error of  $t_k^{(u)}$  is

$$MSE(t_k^{(u)}) = f_1 \bar{Y}^2 C_y^2 (1-\rho^2) \quad (3.24)$$

#### 4. Efficiency comparison

It is well known under SRSWOR that the variance of usual unbiased estimator  $\bar{y}$  is

$$\text{Var}(\bar{y}) = \text{MSE}(\bar{y}) = f_1 \bar{Y}^2 C_y^2 \quad (4.1)$$

From (1.5), (1.6), (1.13), (1.14), (3.24) and (4.1), we have

$$(i) \text{ MSE}(\bar{y}) - \text{MSE}(t_K^{(u)}) > 0 \text{ if } \rho > 0 \quad (4.2)$$

$$(ii) \text{ MSE}(\bar{y}_R) - \text{MSE}(t_K^{(u)}) > 0 \text{ if } (K^2 - 2K + 1) > 0 \\ K > 1 \text{ or } \rho > \frac{C_x}{C_y} \quad (4.3)$$

$$(iii) \text{ MSE}(t_1) - \text{MSE}(t_K^{(u)}) > 0 \text{ if } (4K^2 - 4K + 1) > 0 \\ K > \frac{1}{2} \text{ or } \rho > \frac{1}{2} \frac{C_x}{C_y} \quad (4.4)$$

$$(iv) \text{ MSE}(\bar{y}_P) - \text{MSE}(t_K^{(u)}) > 0 \text{ if } (K^2 + 2K + 1) > 0 \\ K > -1 \text{ or } \rho > -\frac{C_x}{C_y} \quad (4.5)$$

$$(v) \text{ MSE}(t_2) - \text{MSE}(t_K^{(u)}) > 0 \\ (4K^2 + 4K + 1) > 0 \\ K > -\frac{1}{2} \text{ or } \rho > -\frac{1}{2} \frac{C_x}{C_y} \quad (4.6)$$

## 5. Empirical study

To see the performance of the estimators  $\bar{y}_R$ ,  $\bar{y}_P$ ,  $t_1$ ,  $t_2$  and  $t_K^{(u)}$  over  $\bar{y}$ , we consider two population data sets. Using the formula

$$\text{PRE}(\cdot, \bar{y}) = \frac{\text{MSE}(\bar{y})}{\text{MSE}(\cdot)} \times 100, \quad (\cdot) = \bar{y}, \bar{y}_R, \bar{y}_P, t_1, t_2 \text{ and } t_K^{(u)}$$



We have computed the percent relative efficiencies (PREs) of the estimators  $\bar{y}_R$ ,  $\bar{y}_P$ ,  $t_1$ ,  $t_2$  and  $t_K^{(u)}$  over  $\bar{y}$  and compiled in Table 1.

The values of scalars  $h_j$ 's ( $j=1, 2, 3$ ) of the almost unbiased exponential estimator are calculated for different population data sets and compiled in Table 2.

The description of the populations is given below:

**Positive correlated variables:**

Population- I: [Source: Murthy (1967, pp. 228)]

It consists of 80 factories in a region, the characters  $x$  and  $y$  being fixed capital and output respectively. The variates are defined as follows:

$Y$ : output

$X$ : the number of fixed capital

$$C_y = 0.3519, \quad C_x = 0.7459, \quad \rho = 0.9413, \quad K = 0.4440$$

**Negative correlated variables:**

Population- II: [Source: Steel and Torrie (1960, pp. 282)]

$Y$ : log of leaf burn in secs,

$X$ : chlorine percentage

$$C_y = 0.4803, \quad C_x = 0.7493, \quad \rho = -0.4996, \quad K = -0.32$$

**Table 1.** Percent relative efficiencies (PREs) of the estimators

$\bar{y}$ ,  $\bar{y}_R$ ,  $\bar{y}_P$ ,  $t_1$ ,  $t_2$  and  $t_K^{(u)}$  with respect to  $\bar{y}$

S. No.	Estimator	PRE( $\bar{y}$ )	
		Population I	Population II
1.	$\bar{y}$	100.00	100.00
2.	$\bar{y}_R$	66.5233	20.0277
3.	$\bar{y}_P$	10.5433	53.2809
4.	$t_1$	783.5443	41.8739
5.	$t_2$	24.2792	120.5436
6.	$t_K^{(u)}$	878.0141	133.2177

**Table 2.** Values of  $h_j$ 's ( $j=1, 2, 3$ ) for almost unbiased exponential estimator

Scalars	Population I	Population II
$h_1$	1.3130	-0.0312
$h_2$	-0.6261	2.0625
$h_3$	0.3130	-1.0312

From Table 1, it is observed that the suggested almost unbiased exponential estimator  $t_K^{(u)}$  is more efficient than the usual unbiased estimator  $\bar{y}$ , classical ratio estimator  $\bar{y}_R$ , classical product estimator  $\bar{y}_P$  and Bahl and Tuteja (1991) ratio-type estimator  $t_1$  and product-type estimator  $t_2$  respectively.

From Table 2, we can say that by using these values of scalars  $h_j$ 's ( $j=1, 2, 3$ ), one can reduce the bias of the estimator  $t_K^{(u)}$  up to the first degree of approximation.

## 6. Conclusions

It is observed from (1.3) and (1.11) that the classical ratio estimator  $\bar{y}_R$  and the product estimator  $\bar{y}_P$  are biased. In some applications, biasedness of an estimator is disadvantageous. So keeping this in view, first we have suggested a generalized version of Bahl and Tuteja (1991) ratio-type and product-type estimators. It is observed that the suggested generalized estimator is also biased. So using the technique as adopted by Singh and Singh (1993), we have suggested an almost unbiased estimator for the population mean  $\bar{Y}$  with its variance formula. From Table 1 and Table 2, we have observed that the suggested almost unbiased exponential estimator  $t_K^{(u)}$  is more efficient than  $\bar{y}$ ,  $\bar{y}_R$ ,  $\bar{y}_P$ ,  $t_1$ ,  $t_2$  and  $t_K^{(u)}$ . We shall see that the suggested almost unbiased estimator depends only on the well known parameter  $K=\rho_{yx}(C_Y/C_X)$ , the value of which can be obtained quite accurately from some earlier survey or a pilot study.

**Acknowledgements**

The authors acknowledge the University Grants Commission, New Delhi, India for financial support in the project number F. No. 34-137/2008(SR). The authors are also thankful to Indian School of Mines, Dhanbad and Vikram University, Ujjain for providing the facilities to carry out the research work. The authors are also grateful to the referees for valuable suggestions regarding improvement of the paper.

## REFERENCES

- BAHL, S. and TUTEJA, R. K. (1991). Ratio and product type exponential estimator, *Information and Optimization Sciences*, vol. 12 (1), 159-163.
- MURTHY, M. N. (1964). Product method of estimation. *Sankhya, A*, 26, 69-74.
- MURTHY, M. N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta, India.
- SINGH, S. and SINGH, R. (1993). A new method: Almost separation of bias precipitates in sample surveys. *Jour. Ind. Stat. Assoc.*, 31, 99-105.
- STEEL, R. G. D. and TORRIE, J. H. (1960). *Principles and Procedures of Statistics*, Mc Graw Hill, New York.

## A BETTER ESTIMATOR OF POPULATION MEAN WITH POWER TRANSFORMATION BASED ON RANKED SET SAMPLING

Nitu Mehta (Ranka)<sup>1</sup>, V.L. Mandowara<sup>2</sup>

### ABSTRACT

Ranked set sampling (RSS) was first suggested by McIntyre (1952) to increase the efficiency of estimate of the population mean. It has been shown that this method is highly beneficial to the estimation based on simple random sampling (SRS). There has been considerable development and many modifications were done on this method. This paper presents a modified ratio estimator using prior value of coefficient of kurtosis of an auxiliary variable  $x$ , with the intention to improve the efficiency of ratio estimator in ranked set sampling. The first order approximation to the bias and mean square error (MSE) of the proposed estimator are obtained. A generalized version of the suggested estimator by applying the Power transformation is also presented.

**Key words:** ranked set sampling, ratio estimator, power transformation estimator, auxiliary variable.

### 1. Introduction

The traditional ratio estimator for the population mean  $\bar{Y}$  of the study variable  $y$  is defined by

$$\hat{Y}_R = \bar{y} \left( \frac{\bar{X}}{\bar{x}} \right) \quad (1.1)$$

in which it is assumed that the population mean  $\bar{X}$  of the auxiliary variable  $x$  is known. Here,  $\bar{y}$  is the sample mean of the study variable and  $\bar{x}$  is the sample mean of the auxiliary variable.

---

<sup>1</sup> Research Scholar, Dept. of Mathematics & Statistics, University College of Science, M. L. Sukhadia University, Udaipur-313001, Rajasthan, India. E-mail: nitumehta82@gmail.com.

<sup>2</sup> Dept. of Mathematics & Statistics, University College of Science, M. L. Sukhadia University, Udaipur-313001, Rajasthan, India. E-mail: mandowara\_vl@yahoo.co.in.

The bias and mean square error (MSE) of  $\bar{Y}_R$  are given by

$$B(\bar{Y}_R) = \theta \bar{Y} C_x^2 (1 - k) \quad (1.2)$$

$$\text{and } MSE(\bar{Y}_R) = \theta \bar{Y}^2 [C_y^2 + C_x^2 (1 - 2k)] \quad (1.3)$$

where  $\theta = \frac{1}{n}$ ,  $k = \rho \frac{C_y}{C_x}$ ,  $C_y$  and  $C_x$  are coefficients of variations of  $y$  and  $x$  respectively and  $\rho$  is correlation coefficient.

Singh, H.P (2004) proposed a modified ratio estimator as

$$\bar{Y}_M = \bar{y} \left[ \frac{\bar{X} + \beta_2(x)}{\bar{x} + \beta_2(x)} \right] \quad (1.4)$$

where  $\beta_2(x)$  is known value of the coefficient of kurtosis of an auxiliary variable.

The Bias and MSE of this estimator were given by

$$B(\bar{Y}_M) = \theta \bar{Y} \lambda C_x^2 (\lambda - k) \quad (1.5)$$

$$\text{and } MSE(\bar{Y}_M) = \theta \bar{Y}^2 [C_y^2 + \lambda C_x^2 (\lambda - 2k)] \quad (1.6)$$

$$\text{where } \lambda = \frac{\bar{X}}{\bar{X} + \beta_2(x)}$$

By applying the power transformation on  $\bar{Y}_{M\alpha}$  in (1.4), the generalized estimator is

$$\bar{Y}_{M\alpha} = \bar{y} \left[ \frac{\bar{X} + \beta_2(x)}{\bar{x} + \beta_2(x)} \right]^\alpha \quad (1.7)$$

where  $\alpha$  is a suitably chosen scalar.

The bias and MSE of the estimator  $\frac{\Lambda}{Y_{M\alpha}}$  to the first degree of approximation are respectively given by

$$B(\frac{\Lambda}{Y_{M\alpha}}) = \theta\alpha \left( \frac{\bar{Y}}{2} \right) \lambda C_x^2 \{ \lambda(\alpha + 1) - 2k \} \quad (1.8)$$

$$\text{and } MSE(\frac{\Lambda}{Y_{M\alpha}}) = \theta \bar{Y}^2 [C_y^2 + \alpha \lambda C_x^2 (\alpha \lambda - 2k)] \quad (1.9)$$

in which the value of  $\alpha = \frac{k}{\lambda}$  makes the MSE in (1.9) minimum. Comparing (1.9) when  $\alpha = \frac{k}{\lambda}$  with (1.6), Singh (2004) showed that  $\frac{\Lambda}{Y_{M\alpha}}$  is more efficient than  $\frac{\Lambda}{Y_M}$ .

## 2. The suggested estimator

In Ranked set sampling (RSS),  $m$  independent random sets, each of size  $m$ , are selected with equal probability and with replacement from the population. The members of each random set are ranked with respect to the characteristic of the study variable or auxiliary variable. Then, the smallest unit is selected from the ordered set and the second smallest unit is selected from the second ordered set. By this way, this procedure is continued until the unit with the largest rank is chosen from the  $m^{th}$  set. This cycle may be repeated  $r$  times, so  $mr$  units have been measured during this process.

When we rank on the auxiliary variable, let  $(y_{[i]}, x_{(i)})$  denote a  $i^{th}$  judgment ordering in the  $i^{th}$  set for the study variable and  $i^{th}$  set for the auxiliary variable.

Swami and Muttalak (1996) defined the estimator of the population ratio using RSS as

$$\frac{\Lambda}{R_{RSS}} = \frac{\bar{y}_{[n]}}{\bar{x}_{(n)}} \quad (2.1)$$

where  $\bar{y}_{[n]} = \frac{1}{n} \sum_{i=1}^n y_{[i]}$  and  $\bar{x}_{(n)} = \frac{1}{n} \sum_{i=1}^n x_{(i)}$

As Swami and Muttalak (1996) remind that this estimator can also be used for the population total and mean. We can write the following estimator for the population mean as

$$\frac{\Delta}{Y}_{R,RSS} = \bar{y}_{[n]} \left( \frac{\bar{X}}{\bar{x}_{(n)}} \right) \quad (2.2)$$

To obtain bias and MSE of  $\frac{\Delta}{Y}_{R,RSS}$ , we put  $\bar{y}_{[n]} = \bar{Y}(1 + \varepsilon_0)$  and  $\bar{x}_{(n)} = \bar{X}(1 + \varepsilon_1)$  so that  $E(\varepsilon_0) = E(\varepsilon_1) = 0$

$$\begin{aligned} V(\varepsilon_0) &= E(\varepsilon_0^2) = \frac{V(\bar{y}_{[n]})}{\bar{Y}^2} \\ &= \frac{1}{mr} \frac{1}{\bar{Y}^2} \left[ S_y^2 - \frac{1}{m} \sum_{i=1}^m \tau_{y[i]}^2 \right] = [\theta C_y^2 - W_{y[i]}^2] \end{aligned}$$

similarly,  $V(\varepsilon_1) = E(\varepsilon_1^2) = [\theta C_x^2 - W_{x(i)}^2]$

$$\begin{aligned} \text{and } Cov(\varepsilon_0, \varepsilon_1) &= E(\varepsilon_0, \varepsilon_1) = \frac{Cov(\bar{y}_{[n]}, \bar{x}_{(n)})}{\bar{X}\bar{Y}} \\ &= \frac{1}{\bar{X}\bar{Y}} \frac{1}{mr} \left[ S_{yx} - \sum_{i=1}^m \tau_{yx(i)} \right] = [\theta \rho_{yx} C_y C_x - W_{yx(i)}] \end{aligned}$$

where  $\theta = \frac{1}{mr}$ ,  $C_y^2 = \frac{S_y^2}{\bar{Y}^2}$ ,  $C_x^2 = \frac{S_x^2}{\bar{X}^2}$ ,  $C_{yx} = \frac{S_{yx}}{\bar{X}\bar{Y}} = \rho_{yx} C_y C_x$ ,

$$W_{x(i)}^2 = \frac{1}{m^2 r} \frac{1}{\bar{X}^2} \sum_{i=1}^m \tau_{x(i)}^2, \quad W_{y[i]}^2 = \frac{1}{m^2 r} \frac{1}{\bar{Y}^2} \sum_{i=1}^m \tau_{y[i]}^2 \quad \text{and}$$

$$W_{yx(i)} = \frac{1}{m^2 r} \frac{1}{\bar{X}\bar{Y}} \sum_{i=1}^m \tau_{yx(i)}.$$

Here, we would also like to remind that  $\tau_{x(i)} = \mu_{x(i)} - \bar{X}$ ,  $\tau_{y[i]} = \mu_{y[i]} - \bar{Y}$  and  $\tau_{yx(i)} = (\mu_{x(i)} - \bar{X})(\mu_{y[i]} - \bar{Y})$ .

Further, to validate first degree of approximation, we assume that the sample size is large enough to get  $|\varepsilon_0|$  and  $|\varepsilon_1|$  as small so that the terms involving  $\varepsilon_0$  and or  $\varepsilon_1$  in a degree greater than two will be negligible.



Bias and MSE of the estimator  $\bar{Y}_{R,RSS}^{\Lambda}$  to the first degree of approximation are respectively given by

$$B(\bar{Y}_{R,RSS}^{\Lambda}) = E(\bar{Y}_{R,RSS}^{\Lambda}) - \bar{Y}$$

$$\begin{aligned} \text{Here } \bar{Y}_{R,RSS}^{\Lambda} &= \bar{Y}(1 + \varepsilon_0)(1 + \varepsilon_1)^{-1} \\ &= \bar{Y}[1 + \varepsilon_0 - \varepsilon_1 - \varepsilon_0\varepsilon_1 + \varepsilon_1^2 + o(\varepsilon_1)] \end{aligned}$$

$$\begin{aligned} \text{Now } E(\bar{Y}_{R,RSS}^{\Lambda}) &= \bar{Y}[1 + E(\varepsilon_1^2) - E(\varepsilon_0\varepsilon_1)] \\ \Rightarrow B(\bar{Y}_{R,RSS}^{\Lambda}) &= \bar{Y}[\{\theta C_x^2 - W_{x(i)}^2\} - \{\theta \rho_{yx} C_y C_x - W_{yx(i)}\}] \\ \Rightarrow B(\bar{Y}_{R,RSS}^{\Lambda}) &= \bar{Y}[\theta C_x^2(1 - k) - (W_{x(i)}^2 - W_{yx(i)})] \end{aligned}$$

$$\text{where } k = \rho \frac{C_y}{C_x} \quad (2.3)$$

$$\begin{aligned} \text{Now } MSE(\bar{Y}_{R,RSS}^{\Lambda}) &= E(\bar{Y}_{R,RSS}^{\Lambda} - \bar{Y})^2 \\ &= \bar{Y}^2 E[\varepsilon_0 - \varepsilon_1 - \varepsilon_0\varepsilon_1 + \varepsilon_1^2]^2 = \bar{Y}^2 [\varepsilon_0^2 - \varepsilon_1^2 - 2\varepsilon_0\varepsilon_1] \\ &= \bar{Y}^2 [\theta C_y^2 - W_{y[i]}^2 + \theta C_x^2 - W_{x(i)}^2 - 2(\theta \rho_{yx} C_y C_x - W_{yx(i)})] \\ &= \bar{Y}^2 [\theta \{C_y^2 + C_x^2(1 - 2k)\} - \{W_{y[i]}^2 + W_{x(i)}^2 - 2W_{yx(i)}\}] \end{aligned}$$

$$MSE(\bar{Y}_{R,RSS}^{\Lambda}) = \bar{Y}^2 [\theta \{C_y^2 + C_x^2(1 - 2k)\} - \{W_{y[i]} - W_{x(i)}\}^2] \quad (2.4)$$

Adapting the estimator in (2.1) to the modified ratio estimator for the population mean suggested by Singh (2004), given in (1.4), we develop the following estimator

$$\bar{Y}_{M,RSS}^{\Lambda} = \bar{y}_{[n]} \left( \frac{\bar{X} + \beta_2(x)}{\bar{x}_n + \beta_2(x)} \right) \quad (2.5)$$

The Bias and MSE of  $\bar{Y}_{M,RSS}^{\Lambda}$  can be found as follows:

$$B(\bar{Y}_{M,RSS}^{\Lambda}) = E(\bar{Y}_{M,RSS}^{\Lambda}) - \bar{Y}$$

$$\text{Here } \bar{Y}_{M,RSS}^{\Lambda} = \bar{Y}(1 + \varepsilon_0)(1 + \lambda\varepsilon_1)^{-1} \text{ where } \lambda = \frac{\bar{X}}{\bar{X} + \beta_2(x)}$$

Suppose  $|\lambda\varepsilon_1| < 1$  so that  $(1 + \lambda\varepsilon_1)^{-1}$  is expandable.

$$\begin{aligned}\text{So } B(\bar{Y}_{M,RSS}) &= \bar{Y}[\lambda^2 E(\varepsilon_1^2) - \lambda E(\varepsilon_0 \varepsilon_1)] \\ &= \bar{Y}[\lambda^2 \{\theta C_x^2 - W_{x(i)}^2\} - \lambda \{\theta \rho_{yx} C_y C_x - W_{yx(i)}\}]\end{aligned}$$

$$\text{and } B(\bar{Y}_{M,RSS}) = \bar{Y}[\theta \lambda C_x^2 (1-k) - \lambda (W_{x(i)}^2 - W_{yx(i)})] \quad (2.6)$$

$$\text{where } k = \rho \frac{C_y}{C_x}$$

$$\begin{aligned}MSE(\bar{Y}_{M,RSS}) &= E(\bar{Y}_{M,RSS} - \bar{Y})^2 \\ &= \bar{Y}^2 [\varepsilon_0^2 + \lambda^2 \varepsilon_1^2 - 2\lambda \varepsilon_0 \varepsilon_1] \\ &= \bar{Y}^2 [\theta C_y^2 - W_{y[i]}^2 + \lambda^2 (\theta C_x^2 - W_{x(i)}^2) - 2\lambda (\theta \rho_{yx} C_y C_x - W_{yx(i)})] \\ &= \bar{Y}^2 [\theta \{C_y^2 + \lambda C_x^2 (\lambda - 2k)\} - \{W_{y[i]}^2 + \lambda^2 W_{x(i)}^2 - 2\lambda W_{yx(i)}\}]\end{aligned}$$

$$MSE(\bar{Y}_{M,RSS}) = \bar{Y}^2 [\theta \{C_y^2 + \lambda C_x^2 (\lambda - 2k)\} - \{W_{y[i]} - \lambda W_{x(i)}\}^2] \quad (2.7)$$

By applying the power transformation on  $\bar{Y}_{M,RSS}$ , the generalized estimator is given by

$$\bar{Y}_{M\alpha,RSS} = \bar{y}_{[n]} \left( \frac{\bar{X} + \beta_2(x)}{x_n + \beta_2(x)} \right)^\alpha \quad (2.8)$$

The bias and MSE of the estimator  $\bar{Y}_{M\alpha,RSS}$  to the first degree of approximation, are respectively given by

$$B(\bar{Y}_{M\alpha,RSS}) = E(\bar{Y}_{M\alpha,RSS}) - \bar{Y}$$

$$\text{Here } \bar{Y}_{M\alpha,RSS} = \bar{Y}(1 + \varepsilon_0)(1 + \lambda \varepsilon_1)^{-\alpha}$$

$$= \bar{Y} \left[ (1 + \varepsilon_0) \left\{ 1 - \lambda \alpha \varepsilon_1 + \frac{\alpha(\alpha+1)}{2} \lambda^2 \varepsilon_1^2 + o(\varepsilon_1) \right\} \right]$$

$$\begin{aligned}B(\bar{Y}_{M\alpha,RSS}) &= \bar{Y} \left[ \lambda^2 \frac{\alpha(\alpha+1)}{2} \{\theta C_x^2 - W_{x(i)}^2\} - \lambda \alpha \{\theta \rho_{yx} C_y C_x - W_{yx(i)}\} \right] \\ &= \theta \alpha \left( \frac{\bar{Y}}{2} \right) \lambda C_x^2 \{\lambda(\alpha+1) - 2k\} - \left( \frac{\bar{Y}}{2} \right) \lambda \alpha \{\lambda(\alpha+1) W_{x(i)}^2 - 2W_{yx(i)}\}\end{aligned}$$

$$\Rightarrow B(\bar{Y}_{M\alpha, RSS}) = \left(\frac{\bar{Y}}{2}\right) \lambda \alpha \left[ \theta C_x^2 \{ \lambda(\alpha + 1) - 2k \} - \{ \lambda(\alpha + 1) W_{x(i)}^2 - 2W_{yx(i)} \} \right] \quad (2.9)$$

$$\text{where } k = \rho \frac{C_y}{C_x}$$

$$\begin{aligned} \text{and } MSE(\bar{Y}_{M\alpha, RSS}) &= E(\bar{Y}_{M\alpha, RSS} - \bar{Y})^2 \\ &= \bar{Y}^2 E \left[ \varepsilon_0 - \lambda \alpha \varepsilon_1 + \lambda^2 \frac{\alpha(\alpha + 1)}{2} \varepsilon_1^2 - \lambda \alpha \varepsilon_0 \varepsilon_1 \right]^2 \\ &= \bar{Y}^2 E \left[ \varepsilon_0^2 + \lambda^2 \alpha^2 \varepsilon_1^2 - 2\lambda \alpha \varepsilon_0 \varepsilon_1 \right] \\ &= \bar{Y}^2 \left[ \theta C_y^2 - W_{y[i]}^2 + \lambda^2 \alpha^2 (\theta C_x^2 - W_{x(i)}^2) - 2\lambda \alpha (\theta \rho_{yx} C_y C_x - W_{yx(i)}) \right] \\ MSE(\bar{Y}_{M\alpha, RSS}) &= \bar{Y}^2 \left[ \theta \{ C_y^2 + \alpha \lambda C_x^2 (\alpha \lambda - 2k) \} - \{ W_{y[i]} - \lambda \alpha W_{x(i)} \}^2 \right] \quad (2.10) \end{aligned}$$

$$\text{let } A = (W_{y[i]} - \lambda \alpha W_{x(i)})^2$$

By this way, we can write (2.10) as

$$MSE(\bar{Y}_{M\alpha, RSS}) \cong MSE(\bar{Y}_{M\alpha}) - A \quad (2.11)$$

It is easily shown that MSE of the proposed estimator using ranked set sampling is always smaller than the estimator, suggested by Singh (2004) given in (1.7), because  $A$  is a non-negative value. As a result, it is shown that the proposed estimator  $\bar{Y}_{M\alpha, RSS}$  is more efficient than the estimator  $\bar{Y}_{M\alpha}$ .

### 3. Optimality of $\alpha$

The optimum value of  $\alpha$  to minimize the MSE of  $\bar{Y}_{M\alpha, RSS}$  can easily be found as follows:

$$\begin{aligned} \frac{\partial MSE(\bar{Y}_{M\alpha, RSS})}{\partial \alpha} &= 0 \\ \Rightarrow \theta C_x^2 (2\lambda^2 \alpha - 2k\lambda) - (2\lambda^2 \alpha W_{x(i)}^2 - 2\lambda W_{yx(i)}) &= 0 \end{aligned}$$

$$\begin{aligned}
&\Rightarrow \theta C_x^2 (2\lambda^2 \alpha - 2k\lambda) - (2\lambda^2 \alpha W_{x(i)}^2 - 2\lambda W_{yx(i)}) = 0 \\
&\Rightarrow \lambda \alpha (\theta C_x^2 - W_{x(i)}^2) + (W_{yx(i)} - \theta C_x^2 k) = 0 \\
&\Rightarrow \alpha' = \frac{\theta k C_x^2 - W_{yx(i)}}{\lambda (\theta C_x^2 - W_{x(i)}^2)} \quad (3.1)
\end{aligned}$$

When we replace  $\alpha$  by  $\alpha'$  in (2.10), we obtain minimum MSE of the proposed estimator as follows:

$$\begin{aligned}
\min .MSE(\bar{Y}_{M\alpha, RSS}) &\cong \bar{Y}^2 \left[ \theta \{C_y^2 + \alpha' \lambda C_x^2 (\lambda \alpha' - 2k)\} - \{W_{y[i]}^2 + \lambda \alpha' (\lambda \alpha' W_{x(i)}^2 - 2W_{yx(i)})\} \right] \\
\Rightarrow MSE(\bar{Y}_{M\alpha, RSS}) &= \bar{Y}^2 \left[ \theta \{C_y^2 + \alpha' \lambda C_x^2 (\alpha' \lambda - 2k)\} - \{W_{y[i]} - \lambda \alpha' W_{x(i)}\}^2 \right] \quad (3.2)
\end{aligned}$$

## REFERENCES

- KADILAR, C., UNYAZICI, Y. and CINGI, H. (2009). Ratio estimator for the population mean using ranked set sampling, *Stat. papers*, 50, 301-309.
- KOWALCZYK B. (2004). Ranked set sampling and its Applications in Finite population studies, *Statistics in transition*, Vol.6, No.7 pp.1031-1046.
- MCINTYRE, G. A. (1952). A method of unbiased selective sampling using ranked sets, *Australian Journal of Agricultural Research*, 3, 385-390.
- SAMAWI, H. M. and MUTTLAK, H. A. (1996). Estimation of Ratio Using Rank Set Sampling, *Biom. J.* 38, 6, 753-764.
- SINHA, A. K. (2005). On some recent developments in ranked set sampling, *Bulletin of Informatics & Cybernetics*, Research Association of Statistical Sciences, Vol. 37, 136-160.
- SINGH, H. P., TAILOR RAJESH, TAILOR RITESH and KAKRAN, M. S. (2004). An improved estimator of population mean using Power Transformation, *J. Ind. Soc. Agril. Statist.* 58(2), 223-230.
- SINGH, S. (2003). *Advanced Sampling Theory with Application*, Vol. I, Kluwer Academic Publishers, Netherlands.

STATISTICS IN TRANSITION-new series, December 2012  
Vol. 13, No. 3, pp. 559—668

## A KERNEL VERSION OF FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

Tomasz Górecki<sup>1</sup>, Mirosław Krzyśko<sup>2</sup>

### ABSTRACT

In this paper a new construction of functional principal components (FPCA) is proposed, based on principal components for vector data. A kernel version of FPCA is also presented. The quality of the two described methods was tested on 20 different data sets.

**Key words:** PCA, FPCA, kernel version of FPCA.

### 1. Introduction

Advances in modern technology, including computing environments, have facilitated the collection and analysis of high-dimensional data, or data that consist of repeated measurements of the same subject. If the repeated measurements are taken densely over a period of time, say on an interval  $I$ , often by machine, they are typically termed functional or curve data, with one observed curve (or function) per subject. This is often the case even if the data are observed with experimental error, since the operation of smoothing data recorded at closely spaced time points can greatly reduce the effects of noise. In such cases we may regard the entire curve for the  $i$ th subject, represented by the graph of the function  $X_i(t)$  say, as being observed in the continuum, even though in reality the recording times are discrete. The statistical analysis of a sample of  $n$  such graphs is commonly termed functional data analysis (see Ramsay and Dalzell, 1991). Functional data analysis whose main purpose is to provide tools for describing and modelling sets of curves is a topic of growing interest in the statistical community. The books by Ramsay and Silverman (2002, 2005) propose an interesting description of the available procedures dealing with functional observations. These functional approaches have been proved useful in various

---

<sup>1</sup> Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland.  
E-mail: tomasz.gorecki@amu.edu.pl.

<sup>2</sup> Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland.  
E-mail: mkrzysko@amu.edu.pl.

domains such as chemometrics, economics, climatology, biology and remote sensing. The statistician generally wants, as a first step, to represent as far as possible a set of random curves in a small space in order to get a description of the functional data that allows interpretation. Functional principal components analysis (FPCA) gives a small-dimension space which captures the main modes of variability of the data. The basic idea in functional principal components analysis is to find functions whose inner products with the data yield the maximum variation in the curves. The first principal component accounts for the most variation, the second principal component accounts for the largest variation orthogonal to the first principal component, and so on. In this way, much of the variation in the modern data can be captured using only a few principal components. A construction method for functional principal components is given in the monograph of Ramsay and Silverman (2005).

In this paper, we present a new way of constructing the functional principal components. This construction is described in Section 3. In Section 4 we present a kernel version of functional principal components analysis. In Section 5 we show how these two methods work on the example of 20 different real data sets. Section 6 contains results and conclusions.

## 2. Smoothing discrete data

Let  $y_{ij}$  denote the observed value of an investigated statistical property on the  $i$ th unit at the  $j$ th time point, where  $j = 1, 2, \dots, J_i$ ,  $i = 1, 2, \dots, N$ . The observation time points  $t_{ij}$  of a given statistical property may vary from unit to unit, and the intervals between these points need not be uniform. Our data then consist of pairs  $(t_{ij}, y_{ij})$ , where  $t_{ij} \in I$ ,  $j = 1, 2, \dots, J_i$ ,  $i = 1, 2, \dots, N$ .

The discrete data  $(t_{ij}, y_{ij})$  can be transformed into functional data (see Ramsay and Silverman (2005)):

$$\{x_i(t), i = 1, 2, \dots, N, t \in I\}.$$

Because the data transformation process is carried out separately for each  $i = 1, 2, \dots, N$ , our further considerations will relate to a single function  $x(t)$ ,  $t \in I$ .

One of the ways of smoothing discrete data  $\{t_j, y_j\}$ ,  $j \in J$ ,  $t_j \in I$ , to a continuous function  $x(t)$  on the interval  $I$  is to present that function as a linear combination  $N$  of orthonormal basis functions  $\varphi_k$ :

$$x(t) = \sum_{k=0}^{N-1} c_k \varphi_k(t), t \in I. \quad (2.1)$$

The coefficients  $c_k$  of this linear combination are selected by the least squares method, i.e. so as to minimize the function:

$$S(c_0, c_1, \dots, c_{N-1}) = \sum_{j=1}^J \left( y_j - \sum_{k=0}^{N-1} c_k \varphi_k(t_j) \right)^2.$$

In matrix notation, the function  $S$  takes the form:

$$S(\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})'(\mathbf{y} - \mathbf{\Phi c}),$$

where  $\mathbf{c} = (c_0, c_1, \dots, c_{N-1})'$  and  $\mathbf{\Phi}$  is a  $J \times N$  matrix containing the values  $\varphi_k(t_j)$ . By differentiating  $S(\mathbf{c})$  with respect to the vector  $\mathbf{c}$  we obtain a system of normal equations in the form:

Hence, the estimate of vector  $\mathbf{c}$  is equal to

$$\hat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{y}. \quad (2.2)$$

### 3. Construction of functional principal components

Suppose we observe a sample of the process  $X(t) \in L_2(I)$ , where  $L_2(I)$  is the Hilbert space of square integrable functions on an interval  $I$ , equipped with the scalar product  $\langle u, v \rangle = \int u(s)v(s)ds$ .

**Remark 1.** All integrals are taken over the interval  $I$ .

Moreover, suppose that  $EX(t) = 0$  and

$$\int E|X|^2 = E[\langle X, X \rangle] = E \int X^2(s)ds < \infty.$$

A principal component as defined for finite-dimensional vectors,  $X \in \mathbb{R}^k$ , and for a stochastic process  $X(t) \in L_2(I)$  is characterized as follows: Let  $H = \mathbb{R}^k$  in the vector case,  $H = L_2(I)$  in the functional case. Then, the first eigenvalue  $\lambda_1$  and associated weight function or vector  $u_1$  are defined as

$$\lambda_1 = \sup_{u \in H} \text{Var}(\langle u, X \rangle) = \text{Var}(\langle u_1, X \rangle), \quad (3.1)$$

subject to the constraint that

$$\|u\| = 1. \quad (3.2)$$

The condition (3.2) is imposed to ensure the uniqueness (except for sign) of the principal component.

The  $k$ th eigenvalue  $\lambda_k$  and weight function or vector  $u_k$ , for  $k > 1$ , are defined as

$$\lambda_k = \sup_{u \in H} \text{Var}(\langle u, X \rangle) = \text{Var}(\langle u_k, X \rangle),$$

where  $u$  is subject to (3.2) and the  $k$ th principal component  $U_k$  is uncorrelated with the first  $(k - 1)$  principal components  $U_i$ ,  $i = 1, \dots, k - 1$  where

$$U_k = \langle u_k, X \rangle. \quad (3.3)$$

We shall call  $(\lambda_k, u_k)$  the  $k$ th principal configuration.

Regarding the case where  $X(t)$  is a stochastic process, we will assume that the process  $X(t)$  can be represented by a finite number of orthonormal basis functions. Let

$$X(t) = \sum_{k=0}^{N-1} c_k \varphi_k(t), t \in I, \quad (3.4)$$

where  $\{\varphi_k\}$  are the first  $N$  elements of an orthonormal basis of  $L_2(I)$ , and  $\{c_k\}$  are random variables with zero means and finite variances. We adopt the notation

$$\begin{aligned} \varphi(t) &= (\varphi_0(t), \varphi_1(t), \dots, \varphi_{N-1}(t))', \\ c &= (c_0, c_1, \dots, c_{N-1})', 0 < N - 1 < \infty, \end{aligned}$$

with  $E(c) = 0$  and  $Var(c) = \Sigma$ . The process  $X(t)$  can be written in vector form as

$$X(t) = c' \varphi(t), t \in I. \quad (3.5)$$

**Theorem 1.** *The  $k$ th principal configuration of random vector  $c$ , defined by  $(\sigma_k, u_k)$ , is related to the  $k$ th principal configuration of stochastic process  $X(t)$ ,  $(\lambda_k, u_k(t))$ , as follows:*

$$u_k(t) = u_k' \varphi(t), \quad \lambda_k = \sigma_k. \quad (3.6)$$

**Proof.**

Each function  $u(t) \in L_2(I)$  can be written as

$$u(t) = \mathbf{u}' \boldsymbol{\varphi}(t), \text{ gdzie } \mathbf{u} \in \mathbb{R}^N.$$

Then

$$\begin{aligned} \langle u, X \rangle &= \langle \mathbf{u}' \boldsymbol{\varphi}, \mathbf{c}' \boldsymbol{\varphi} \rangle = \mathbf{u}' \langle \boldsymbol{\varphi}, \boldsymbol{\varphi}' \rangle \mathbf{c} = \mathbf{u}' I_N \mathbf{c} = \mathbf{u}' \mathbf{c}, \\ E[\langle u, X \rangle] &= \mathbf{u}' E(\mathbf{c}) = 0, \\ Var([\langle u, X \rangle]) &= \mathbf{u}' E(\mathbf{c} \mathbf{c}') \mathbf{u} = \mathbf{u}' \Sigma \mathbf{u}. \end{aligned}$$

Consider the first principal component of the process  $X(t)$ :

$$\lambda_1 = \sup_{u \in L(I)} Var(\langle u, X \rangle) = Var(\langle u_1, X \rangle),$$

where  $\langle u_1, u_1 \rangle = 1$ .



This is equivalent to stating that

$$\lambda_1 = \sup_{\mathbf{u} \in R^N} \mathbf{u}' \Sigma \mathbf{u} = \mathbf{u}_1' \Sigma \mathbf{u}_1,$$

where  $\mathbf{u}_1' \mathbf{u}_1 = 1$ .

This is the definition of the first principal component of the random vector  $\mathbf{c}$ .

On the other hand, if we begin with the first principal component of the random vector  $\mathbf{c}$  as given by the principal configuration  $(\sigma_1, \mathbf{u}_1)$ , we will obtain the first principal component of the process  $X(t)$  from the equation

$$(\lambda_1, u_1(t)) = (\sigma_1, \mathbf{u}_1 \boldsymbol{\varphi}(t)).$$

Similarly we can extend this reasoning to the second principal component and so on.

Principal components analysis for a random process with finite basis expansion is therefore equivalent to multivariate principal component analysis.

Since  $\Sigma$  is ordinarily unknown, we use the estimator  $\hat{\Sigma}$  based on  $N$  independent realizations of the random vector  $\mathbf{c}$ :

$$\hat{\mathbf{C}} = \begin{bmatrix} \hat{c}_{10} & \hat{c}_{11} & \dots & \hat{c}_{1,N-1} \\ \hat{c}_{20} & \hat{c}_{21} & \dots & \hat{c}_{2,N-2} \\ \dots & \dots & \dots & \dots \\ \hat{c}_{N0} & \hat{c}_{N1} & \dots & \hat{c}_{N,N-1} \end{bmatrix} = \begin{bmatrix} \hat{c}_1' \\ \hat{c}_2' \\ \dots \\ \hat{c}_N' \end{bmatrix} \quad (3.7)$$

where  $\hat{c}_{ik}$  are least squares estimates of the parameters  $c_{ik}$  in the representation

$$x_i(t) = \sum_{k=0}^{N-1} c_{ik} \varphi_k(t) \quad (3.8)$$

of the process  $X(t)$ ,  $t \in I$ ,  $i = 1, 2, \dots, N$ .

The unbiased estimator  $\hat{\Sigma}$  of the unknown matrix  $\Sigma$  has the following form

$$\hat{\Sigma} = \frac{1}{N-1} \hat{\mathbf{C}}' \hat{\mathbf{C}}, \quad (3.9)$$

where  $\hat{\mathbf{C}}$  is given by (3.7).

Then we find the nonzero eigenvalues  $\lambda_k$  and corresponding eigenvectors  $\mathbf{u}_k$  of matrix  $\hat{\Sigma}$ . Having determined the eigenvectors  $\mathbf{u}_k$  we determine its weight functions

$$u_k(t) = \mathbf{u}_k' \boldsymbol{\varphi}(t), t \in I. \quad (3.10)$$

Hence, the  $j$ th functional principal component  $X_i(t)$  is given by

$$\begin{aligned} U_{ij} = \langle u_j(t), X_i(t) \rangle &= \int u_j(t) X_i(t) dt = \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} c_{il} u_{jk} \int \varphi_k(t) \varphi_l(t) dt = \\ &= \sum_{k=0}^{N-1} c_{ik} u_{jk} = c'_i u_j, i = 1, 2, \dots, N, j = 1, 2, \dots \end{aligned} \quad (3.11)$$

#### 4. A kernel version of functional principal components

The space  $R^N$  of values of the random vector  $\mathbf{c} = (c_0, c_1, \dots, c_{N-1})'$  is transformed by the non-linear function  $\Psi$  into a Hilbert space  $H(k)$  with a nonnegative definite reproducing kernel  $k$ :

$$\Psi: R^N \rightarrow H(k).$$

Then, the Moore–Aronszajn theorem (see Aronszajn (1950)) guarantees a one-to-one correspondence between the functions  $\Psi$  and their scalar products.

Let  $\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, \dots, \hat{\mathbf{c}}_N$  be centred realizations of the random vector  $\mathbf{c} = (c_0, c_1, \dots, c_{N-1})'$ . In order to construct kernel principal components in space  $H(k)$  we find the eigenvalues  $\lambda > 0$  and the corresponding eigenvectors  $\mathbf{v} \in H(k)$  of the covariance matrix

$$\hat{\Sigma}_\Psi = \frac{1}{N} \sum_{k=1}^N \Psi(\hat{\mathbf{c}}_k) \Psi'(\hat{\mathbf{c}}_k) \quad (4.1)$$

constructed from the  $N$  centred and non-linear transformed vectors  $\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, \dots, \hat{\mathbf{c}}_N$ .

If  $\mathbf{v}$  is an eigenvector of the matrix  $\hat{\Sigma}_\Psi$  corresponding to the eigenvalue  $\lambda$ , then

$$\hat{\Sigma}_\Psi \mathbf{v} = \lambda \mathbf{v}$$

or equivalently

$$\Psi'(\hat{\mathbf{c}}_i) \hat{\Sigma}_\Psi \mathbf{v} = \lambda \Psi'(\hat{\mathbf{c}}_i) \hat{\Sigma}_\Psi \mathbf{v}, i = 1, 2, \dots, N. \quad (4.2)$$

Each eigenvector  $\mathbf{v}$  must lie in the subspace  $\text{span}\{\Psi(\hat{\mathbf{c}}_1), \Psi(\hat{\mathbf{c}}_2), \dots, \Psi(\hat{\mathbf{c}}_N)\}$  spanned by the vectors  $\Psi(\hat{\mathbf{c}}_1), \Psi(\hat{\mathbf{c}}_2), \dots, \Psi(\hat{\mathbf{c}}_N)$ , i.e. the eigenvector  $\mathbf{v}$  can be written as some linear combination of the vectors  $\Psi(\hat{\mathbf{c}}_1), \Psi(\hat{\mathbf{c}}_2), \dots, \Psi(\hat{\mathbf{c}}_N)$ . There therefore exist coefficients  $\alpha_j, j = 1, 2, \dots, N$ , such that

$$\mathbf{v} = \sum_{j=1}^N \alpha_j \Psi'(\hat{\mathbf{c}}_j) \quad (4.3)$$

By substituting (4.3) into (4.2) we obtain:

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^N \alpha_j \Psi'(\hat{c}_i) \sum_{k=1}^N \Psi(\hat{c}_k) \Psi'(\hat{c}_k) \Psi(\hat{c}_j) = \\ = \lambda \sum_{j=1}^N \alpha_j \Psi'(\hat{c}_i) \Psi(\hat{c}_j), \end{aligned} \quad (4.4)$$

for  $i = 1, 2, \dots, N$ . In matrix notation, equation (4.4) takes the form:

$$\mathbf{K}^2 \boldsymbol{\alpha} = N \lambda \mathbf{K} \boldsymbol{\alpha} \quad (4.5)$$

where  $\mathbf{K} = (k_{ij})$ ,  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_N)'$  or

$$\mathbf{K}^2 \boldsymbol{\alpha} = \tilde{\lambda} \mathbf{K} \boldsymbol{\alpha} \quad (4.6)$$

where  $\tilde{\lambda} = N\lambda$ .

Let us note that every vector  $\boldsymbol{\alpha} \neq \mathbf{0}$  being a solution to the equation is also a solution to equation (4.6), and that the solutions of (4.6) and (4.7) differ

$$\mathbf{K} \boldsymbol{\alpha} = \tilde{\lambda} \boldsymbol{\alpha} \quad (4.7)$$

only by the eigenvectors of matrix  $\mathbf{K}$  corresponding to zero eigenvalues, which is not significant for the problem of principal components (the principal components correspond only to non-zero eigenvalues).

We assumed earlier that the vectors  $\{\Psi(\hat{c}_i)\}, i = 1, 2, \dots, N$  are centred. In the general case we cannot centre the vectors  $\{\Psi(\hat{c}_i)\}$ , because we do not know the form of the function  $\Psi$ . Let

$$\tilde{\Psi}_i = \Psi_i - \frac{1}{N} \sum_{k=1}^N \Psi_k$$

and

$$\tilde{\mathbf{K}} = (\tilde{k}_{ij}) = (\langle \tilde{\Psi}_i, \tilde{\Psi}_j \rangle),$$

where  $\Psi_i = \Psi(c_i), i = 1, 2, \dots, N$ . We cannot compute the matrix  $\tilde{\mathbf{K}}$  directly, but we can express it in terms of the matrix  $\mathbf{K}$ :

$$\tilde{\mathbf{K}} = \mathbf{P} \mathbf{K} \mathbf{P},$$

where  $\mathbf{P} = \left( \delta_{ij} - \frac{1}{N} \right)$  and  $\delta_{ij}$  is the Kronecker delta. Hence, in the general case the construction of kernel principal components must be based on matrix  $\tilde{\mathbf{K}}$ . Because the kernel matrix  $\mathbf{K}$  is nonnegative definite, matrix  $\tilde{\mathbf{K}}$  is nonnegative definite also. This results from the fact (Seber (1984), p. 521) that if  $\mathbf{A} \geq 0$ , then  $\mathbf{C}' \mathbf{A} \mathbf{C} \geq 0$ . In our case  $\mathbf{C} = \mathbf{P}$ , where  $\mathbf{P}$  is a symmetric matrix, i.e.  $\mathbf{P}' = \mathbf{P}$ . Hence, all eigenvalues of  $\tilde{\mathbf{K}}$  are nonnegative.

Having determined the eigenvectors  $\boldsymbol{\alpha}_k$  we determine its weight functions

$$u_k(t) = \boldsymbol{\alpha}'_k \boldsymbol{\varphi}(t), t \in I. \quad (4.8)$$

## 5. Example

The quality of the two described methods (functional principal components analysis and the kernel version of functional principal components analysis) was tested on the 20 different data sets listed in Table 1. The data sets originate from the UCR Time Series Classification/Clustering Homepage (Keogh et al. (2006)).

**Table 1.** Data sets

Data set	Time series length	Number of classes	Number of observations
50Words	270	50	450
Adiac	176	37	390
Beef	470	5	30
CBF	128	3	30
Coffee	286	2	28
ECG200	96	2	100
Face (all)	131	14	560
Face (four)	350	4	24
Fish	463	7	175
Gun-Point	150	2	50
Lightning-2	637	2	60
Lightning-7	319	7	70
OliveOil	570	4	30
OSU Leaf	427	6	200
Swedish Leaf	128	15	500
Synthetic Control	60	6	300
Trace	275	4	100
Two Patterns	128	4	1000
Wafer	152	2	1000
Yoga	426	2	300

Elements from different classes were combined into one data set. For each data set separately, the discrete time series were centred, and then transformed into continuous functions of the form (2.1). As an orthonormal basis of  $L_2(I)$  we took the orthonormal system of Legendre polynomials in the space  $L_2([-1, 1])$ :

$$\tilde{P}_k(x) = \sqrt{\frac{2k+1}{2}} P_k(x),$$

where

$$P_{k+1}(x) = \frac{1}{k+1} [(2k+1)xP_k(x) - kP_{k-1}(x)], k \geq 1,$$

$$P_0(x) = 1, P_1(x) = x.$$

Each finite interval  $[a, b]$  can be transformed into the interval  $[-1, 1]$  by the substitution

$$x = \frac{2}{b-1}t - \frac{b+a}{b-a}, t \in [a, b], x \in [-1, 1].$$

In the case of the kernel version of functional principal components analysis we chose the kernel polynomial function of the form

$$k(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}'\mathbf{y})^2.$$

The most frequently considered objects are presented on a plot of the first two functional principal components. In this case the criterion of goodness of the constructed functional principal components is the expression:

$$\frac{\lambda_1 + \lambda_2}{\sum \lambda_i} 100\%, \quad (5.1)$$

where  $\lambda_1 \geq \lambda_2 \geq \dots$  are the nonzero eigenvalues of the matrix (3.9) or the matrix (4.1). The greater the value of expression (5.1), the greater is the variability shown by the first two functional principal components. The percentage of variability accounted for by the first two functional principal components is given in Table 2. Table 2 shows that better results are obtained in the case of the kernel version of functional principal components analysis.

**Table 2.** Values of the criterion (5.1)

Data set	FPCA	Kernel version of the FPCA
50Words	83.18	99.48
Adiac	95.85	96.04
Beef	94.89	99.39
CBF	71.92	94.48
Coffee	99.67	100.00
ECG200	99.99	100.00
Face (all)	70.58	97.74
Face (four)	75.14	96.59
Fish	98.66	100.00
Gun-Point	87.37	98.94
Lighting-2	66.79	95.90
Lighting-7	53.68	88.29
OliveOil	100.00	100.00
OSU Leaf	99.99	100.00
Swedish Leaf	91.22	99.86

**Table 3.** Values of the criterion (5.1) (cont.)

Data set	FPCA	Kernel version of the FPCA
Synthetic Control	79.77	92.16
Trace	100.00	98.98
Two Patterns	76.58	98.80
Wafer	92.76	99.95
Yoga	99.78	99.99
<b>Mean</b>	<b>85.93</b>	<b>97.78</b>

## 6. Conclusions

The effectiveness of functional principal components and their new kernel version was compared for 20 different data sets. Efficiency criterion was the ratio of the sum of the first two eigenvalues to the sum of all eigenvalues of the matrix (3.9) or the matrix (4.1). Average efficiency of functional principal components is 85.93%, while the average efficiency of the kernel version of functional principal components is 97.78%. This is a significant increase in efficiency, which supports the use of kernel version of functional principal components.

## REFERENCES

- ARONSZAJN, N. (1950). Theory of reproducing kernels, *Trans. Amer. Math. Soc.* 68, 337–404.
- KEOGH, E., XI, X., WEI, L. & RATANAMAHATANA, C. A. (2006). The UCR Time Series Classification/Clustering Homepage, [http://www.cs.ucr.edu/~eamonn/time\\_series\\_data/](http://www.cs.ucr.edu/~eamonn/time_series_data/).
- RAMSAY, J. O., DALZELL, C. J. (1991). Some tools for functional data analysis, *J. Royal Statist. Soc. B* 53, 539–572.
- RAMSAY, J. O., SILVERMAN, B. W. (2002). *Applied Functional Data Analysis*, Springer, New York.
- RAMSAY, J. O., SILVERMAN, B. W. (2005). *Functional Data Analysis*, Springer, New York.
- SEBER, G. A. F. (1984). *Multivariate Observations*, Wiley, New York.

STATISTICS IN TRANSITION-new series, December 2012  
Vol. 13, No. 3, pp. 569—580

## AN EMPIRICAL ANALYSIS OF THE EFFECTIVENESS OF WISHART AND MOJENA CRITERIA IN CLUSTER ANALYSIS

Artur Mikulec, Aleksandra Kupis-Fijalkowska

### ABSTRACT

Mojena and Wishart criteria are methods of selecting the optimal grouping result of agglomerative cluster analysis methods (hierarchical). Two criteria were proposed by Mojena in the 70's of the 20th century: the upper tail rule and moving average quality control rule, both based on an analysis of the fusion levels of objects in the dendrogram with the aim to determine the cut-off point of it, i.e. to choose the optimal clustering result. The third criterion: tree validation was created by Wishart and evaluates the randomness of the objects clustering in the dendrogram.

The purpose of this paper is to present the results of the empirical analysis of the effectiveness of Mojena and Wishart criteria for the number of clusters selection, in comparison to other applicable criteria in this area, including those proposed by: Baker and Hubert, Calinski and Harabasz, Davies and Bouldin, Hubert and Levine. The empirical analysis has been carried out in *ClustanGraphics 8* Program and selected packages in R environment for the generated data sets.

**Key words:** upper tail rule, moving average quality control rule, Mojena criteria, Wishart criterion (tree validation), *ClustanGraphics 8*.

### 1. Introduction

The assessment methods of the grouping result – in the broad sense – are related to the three issues of cluster analysis, respectively: determining the number of clusters, comparing two (or more) classification results and the grouping result quality assessment. The grouping result evaluation stage, i.e. selecting the number of clusters in the analysis which is based on the hierarchical clustering algorithms is one of the last steps here, however, it is extremely important in the classification process. In fact, when the classification sequence  $P_0, P_1, \dots, P_{n-1}$  is given, on the basis of some formal criteria a decision on the final grouping result selection should be taken.

The article aims to present the results of an empirical effectiveness analysis of the two Mojena criteria [Mojena 1977], both based on the distance analysis of the objects fusions in the graph tree – *best cut significance test (upper tail rule, moving average quality control rule)* and the Wishart criterion [Wishart 2006] which evaluates the randomness of objects clustering in the dendrogram – *tree validation*. The above mentioned criteria have been compared with other procedures for selecting the number of clusters, including the following: Baker and Hubert (BH), Caliński and Harabasz (CH), Davies and Bouldin (DB), and Hubert and Levine (HL). All criteria are based on determining the number of classes and the structure of the grouping result. The article discusses different cases of clusters generated on the basis of the same (identical) covariance matrix of variables.

## 2. Methods of determining the number of clusters

The cluster analysis literature widely presents and describes in detail many methods that aim to select the number of clusters [see Gan et al., 2007; Gatnar, Walesiak 2009; Mikulec 2012].

The two Mojena criteria and the Wishart criterion, which are the basis of this work, can be found among very few procedures for (determining) the number of clusters (in addition to the following indices: Beale, Duda and Hart, RMSSTD or RS) that are dedicated to the hierarchical classification methods, for example the agglomeration one. Nonetheless, also other procedures mentioned in the introduction can be used as the selection criteria for the number of clusters in the agglomeration methods (see Table 1) – they differ due to the construction of the internal criterion for the grouping result assessment.

**Table 1.** Chosen methods of determining the number of clusters in the data set <sup>a</sup>

CRITERION	Formula, confidence interval	Criterion of selecting the number of clusters
Baker & Hubert	$BH(u) = \frac{S_+ - S_-}{S_+ + S_-}, BH(u) \in \langle -1; 1 \rangle$	$\hat{u} = \arg \max_u [BH(u)]$
Caliński & Harabasz	$CH(u) = \frac{tr(B_u)/(u-1)}{tr(W_u)/(n-u)}, CH(u) \in R_+$	$\hat{u} = \arg \max_u [CH(u)]$
Davies & Bouldin	$BD(u) = \frac{1}{u} \sum_{q=1}^u \max_{r, q \neq r} \left( \frac{S_q + S_r}{d(q, r)} \right)$	$\hat{u} = \arg \min_u [BD(u)]$
Hubert & Lewine	$HL(u) = \frac{D(u) - l_w D_{\min}}{l_w D_{\max} - l_w D_{\min}}, HL(u) \in (0; 1)$	$\hat{u} = \arg \min_u [HL(u)]$



**Table 1.** Chosen methods of determining the number of clusters in the data set <sup>a</sup> (cont.)

CRITERION	Formula, confidence interval	Criterion of selecting the number of clusters
Upper tail rule (Mojena I)	$\alpha_{x+1} > \bar{\alpha} + k \cdot s_{\alpha}$	classification $P_x$ , where the corresponding step $x : x = 1, \dots, n-2$ is the first one which satisfies the given inequality
Moving average (Mojena II)	$\alpha_{x+1} > \bar{\alpha}_x + L_x + b_x + k \cdot s_x \text{ .where:}$ $L_x = \frac{(y-1)b_x}{2},$ $b_x = \frac{6 \left[ 2 \sum_{f=x-y+1}^x w_f \alpha_f - (y+1) \sum_{f=x-y+1}^x \alpha_f \right]}{y(y^2-1)},$ $w_f = w_{f-1} + 1, \quad f = (x-y+2), \dots, x,$ $w_{x-y+1} = 1$	classification $P_x$ , where the corresponding step $x : x = y, y+1, \dots, n-2$ is the first one which satisfies the given inequality
Tree validation – the criterion on the randomness of objects clustering in the dendrogram (Wishart)	A comparison of the classification sequence results obtained as a result of agglomeration methods application with the family of trees generated by a random permutation of the data set	$H_0$ where the structure of grouping objects as a given tree is random (no structure), $H_1 : \sim H_0$

$a$   $n$  – number of objects ( $i = 1, \dots, n$ );  $m$  – number of characteristics ( $j = 1, \dots, m$ );  
 $u$  – number of groups ( $q, r, s = 1, \dots, u$ );  $K_q$  –  $q$  cluster;  $S_+, S_-$  – number of the distance pairs, consistent and inconsistent respectively;  $tr(B_u), tr(W_u)$  – trace of the covariance matrix, between-groups ( $B_u$ ) and within-groups ( $W_u$ ) respectively;

$S_q = \sqrt{\left(1/n_q\right) \sum_{i \in K_q} \sum_{j=1}^m \left|x_{ij}^q - z_{qj}\right|^2}$  – measurement of dispersion for objects in the  $q$  group ( $K_q$ ), where for  $t=1$  it is the average distance of objects in the  $q$  cluster ( $K_q$ ) from the center of gravity, i.e. the medoid in the group, and for  $t=2$  it is a standard deviation of distance of objects in the  $q$  cluster ( $K_q$ ) from the center of gravity, i.e. the medoid in the group (for the  $r$  group the measurement  $S_r$  can be

analogically obtained);  $d(q, r) = \sqrt[p]{\sum_{j=1}^m |z_{qj} - z_{rj}|^p}$  – measurement of the distance between the gravity centers, i.e. medoids  $(z_{qj}, z_{rj})$  of the  $q$  and  $r$  groups, respectively Manhattan distance for  $p=1$  and the Euclidian distance for  $p=2$ ;  $D(u)$  – sum of all within-groups distances;  $l_w$  – number of within-groups distances;  $D_{\min}, D_{\max}$  – within-groups distance, respectively the smallest and the largest;  $\alpha_x = \min_{i < o} [d_{io}]$ ,  $(i, o = 1, \dots, n-x)$  – measure of dissimilarity (of distances) between the clusters;  $\alpha_{x+1}$  – level (distance) of groups fusion in the step  $x+1$ ,  $\bar{\alpha}$  – average level (distance) of groups fusion,  $s_\alpha$  – standard deviation of the level (distance) of group fusion; constant  $k$ ,  $k \in (2, 75; 3, 5)$ ;  $y$  – number of values of the level (distance) of classes fusion  $\alpha$  in a given case, which is used to calculate the moving average;  $\bar{\alpha}_x$  – moving average of the  $\alpha$  parameter value calculated in the step  $x$ ;  $L_x$  – correction for the delayed “trend” level (distance) of classes fusion calculated in the step  $x$ ;  $b_x$  – „moving” meansquare slope of the trend line for the level (distance) of classes fusion in the step  $x$ ;  $s_x$  – „moving” standard deviation of the  $\alpha$  parameter (distance) value.

Source: based on [Mojena 1977; Wishart 2006; Gatnar, Walesiak 2009].

### 3. Assumptions and the empirical analysis scheme

The empirical analysis of the effectiveness of two Mojena criteria and Wishart criterion against other four criteria, namely Baker and Hubert (BH), Caliński and Harabasz (CH), Davies and Bouldin (DB), and Hubert and Levine (HL), was conducted for:

- 2-5 clusters,
- 2-5 variables,
- clusters with the following structure (100 objects):
  - 2 clusters containing 50 and 50 objects respectively,
  - 3 clusters containing 30, 30 and 40 objects respectively,
  - 4 clusters containing 10, 20, 30 and 40 objects respectively,
  - 5 clusters containing 5, 10, 15, 30 and 40 objects respectively,
- clusters without noisy variables,
- clusters generated on the basis of the same (identical) covariance matrix of variables,
- Euclidian distance measure,
- the three most popular agglomeration methods – the complete linkage, the average linkage and the Ward’s method.

As a result, 16 data sets<sup>1</sup> were analyzed, the authors took into account 4 variants of the number of clusters and 4 variants of the number of variables. Three agglomeration methods were used here, one of the analyzed data sets is presented on the Figure 1.

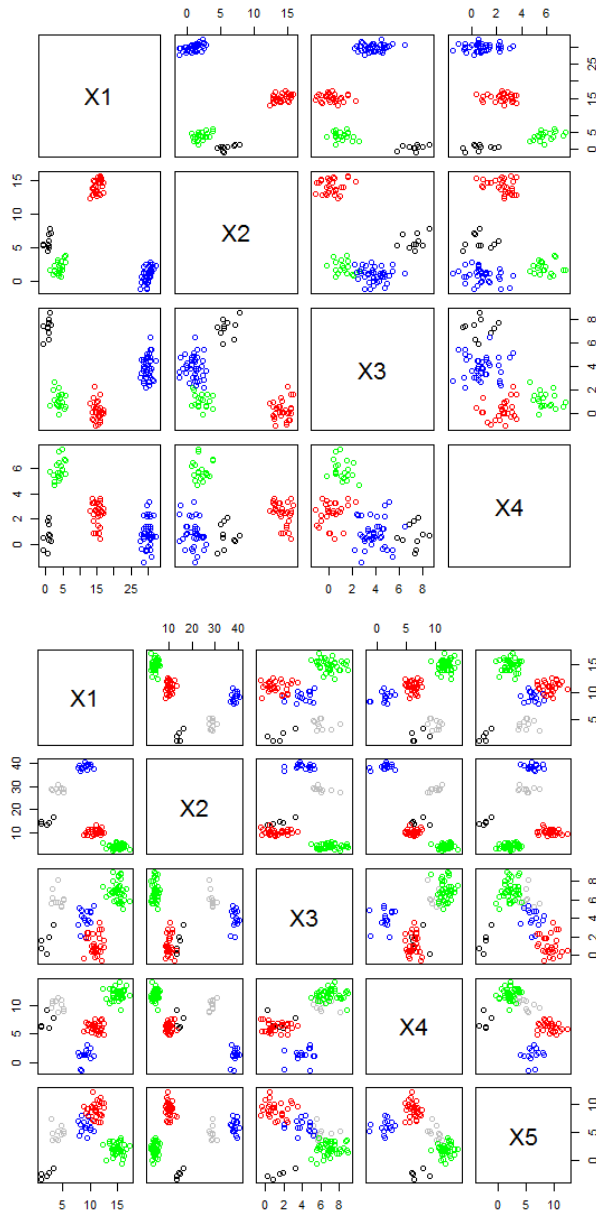
The calculations were performed on the data sets generated with `cluster.Gen` function of `clusterSim` package [Walesiak, Dudek 2012] working in the R environment and using the *ClustanGraphics* 8 software application [Wishart 2006], and the computations scheme looked as follows:

- Step 1, the data sets were generated on the basis of the assumptions (16 sets) for which the correct clusters structure was known (2-5 clusters),
- Step 2, in the *ClustanGraphics* 8 application the cluster analysis was carried out using three agglomerative clustering algorithms (48 results obtained), and each dendrogram result with all clustering sets of objects was written to a file,
- Step 3, in the *ClustanGraphics* 8 application the correct number of clusters was determined – as a result of grouping, according to two Mojena criteria and the Wishart criterion,
- Step 4, in the R environment (`clusterSim`) other indices for determining the number of clusters were calculated – Baker and Hubert (BH), Caliński and Harabasz (CH), Davies and Bouldin (DB), Hubert and Levine (HL) for divisions of 2 to 10 clusters, and the optimal result was chosen according to each criterion for selecting the number of classes: BH (max), CH (max), DB (min), HL (min),
- Step 5, with the given result of cluster analysis – the clusters structure indicated by the chosen criteria of determining the number of clusters for each analyzed dataset – adjusted Rand index was calculated in order to assess consistency of objects belonging to the cluster formed on the basis of each criterion, and the factual cluster of the analyzed set of objects (the known structure of the classes),
- Step 6, taking into account all clustering results, i.e. correct in the structure and objects belonging to the cluster – the Rand index values close to the unity, and incorrect in the structure and objects belonging to the cluster – the Rand index values close to zero, the assessment of the effectiveness of the analyzed criteria for selecting the number of clusters was done by averaging the Rand index values for each agglomeration method and each criterion.

---

<sup>1</sup> The complete characteristic of all analyzed data sets cannot be presented in the article as it is very wide.

**Figure 1.** The data set generated for 4 clusters and 4 variables, and for 5 clusters and 5 variables<sup>a</sup>



<sup>a</sup> In the case of sets with the structure 4 clusters vs. 4 variables and 5 clusters vs. 5 variables, the best criteria for determining the number of clusters (independently from the agglomeration method) were the following ones: BH, CH, DB, and the Wishart criterion, all determined the correct number of clusters and the grouping result structure (in 100%). The upper tail rule was effective for

the sets above only when the Ward's method was used. The moving average rule criterion determined incorrect number of clusters and incorrect clustering result structure for a set of clusters vs. 4 variables when the Ward's method was used. HL criterion incorrectly determined the number of clusters and clustering result structure for the sets above.

Source: own study using R environment `clusterSim` package.

#### 4. The results of the empirical analysis

The Table 2 below presents, for each analyzed criterion of the number of clusters selection, the number of correct and incorrect indications according to the agglomerative clustering methods for the 16 analyzed data sets, where the clusters were generated on the basis of the same (identical) covariance matrix of variables. The results were summarized to create a basis for the assessment of the validation of indications determining the number of clusters for different criteria.

**Table 2.** Assessment of the number of clusters according to the number of clusters selection criteria

Method	The indication			
	correct		incorrect	
Baker and Hubert (BH)				
Average linkage	10	62.50%	6	37.50%
Complete linkage	11	68.75%	5	31.25%
Ward's	12	75.00%	4	25.00%
Caliński and Harabasz (CH)				
Average linkage	13	81.25%	3	18.75%
Complete linkage	12	75.00%	4	25.00%
Ward's	12	75.00%	4	25.00%
Davies and Bouldin (DB)				
Average linkage	6	37.50%	10	<b>62.50%</b>
Complete linkage	10	62.50%	6	37.50%
Ward's	12	75.00%	4	25.00%

**Table 2.** Assessment of the number of clusters according to the number of clusters selection criteria (cont.)

Method	The indication			
	correct		incorrect	
Hubert and Levine (HL)				
Average linkage	1	6.25%	15	<b>93.75%</b>
Complete linkage	3	18.75%	13	<b>81.25%</b>
Ward’s	0	0.00%	16	<b>100.00%</b>
The upper tail rule (Mojena I)				
Average linkage	1	6.25%	15	<b>93.75%</b>
Complete linkage	1	6.25%	15	<b>93.75%</b>
Ward’s	7	43.75%	9	<b>56.25%</b>
The moving average rule (Mojena II)				
Average linkage	13	81.25%	3	18.75%
Complete linkage	13	81.25%	3	18.75%
Ward’s	3	18.75%	13	<b>81.25%</b>
Tree validation – the criterion on the randomness of the objects clustering in the dendrogram (Wishart)				
Average linkage	10	62.50%	6	37.50%
Complete linkage	12	75.00%	4	25.00%
Ward’s	14	87.50%	2	12.50%

*Source: own computations.*

Baker and Hubert criterion – independently of the agglomerative clustering method – in about 2/3 to 3/4 cases indicated the correct number of classes.

More correct results on the indications of the number of clusters were obtained on the basis of Harabasz and Caliński criterion, its accuracy level of the correct number of clusters selection, independently of the agglomerative clustering method, was at least 75.00%.

Davies and Bouldin index more often indicated the correct number of clusters for the analyzed data sets when the complete linkage and the Ward's methods were applied, whereas for the average linkage method the validation level of indications did not exceed 37.50%.

The next two criteria of Hubert and Levine, and the upper tail (Mojena I criterion) did not give the correct solutions for the number of clusters selection in the analyzed data sets, where the clusters were generated on the basis of the same (identical) covariance matrix of variables.

The level of incorrect clusters indications in the case of Hubert and Levine – independently of the agglomerative clustering method – exceeded 81.25%, and in the case of the upper tail rule misclassification rate was equal to 56.25% (the Ward's method) and 93.75% (average linkage and complete linkage methods).

On the basis of the correct indications results of clusters number for Mojena II criterion (moving average rule) it can be stated that in the case of average linkage and complete linkage methods in most cases (81.25%) it is able to determine the correct number of clusters for the analyzed data sets. It is specific that the moving average rule completely failed in the agglomerative clustering with the Ward's method.

The last of the analyzed criteria, the tree validation rule, which evaluates the randomness of the objects clustering in the dendrogram, indicated correctly the number of classes for the average linkage agglomeration method in about 2/3 cases, whereas in the case of complete linkage and the Ward's methods percentage of the correct number of clusters indications, where this decision rule was also used, it was equal respectively to 75.00% and 87.50%. Therefore, next to the Baker and Hubert and Harabasz and Caliński criteria, it should be recognized as the most stable and effective one.

It should be emphasized that the correctness (frequency) of the number of clusters indications by the different criteria is the main, however not the sufficient premise for the "quality" assessment, i.e. relevance of the considered criterion for selecting the number of clusters. The most important is the consistency of the grouping result in the term of objects belonging to the respective cluster, and therefore consistency of the clustering result with the known clustering structure in the generated data sets, for which the assessment based on the adjusted Rand index was taken.

Therefore, Table 3 presents the consistency assessment of the clustering result with the known clustering structure, which is based on the average of adjusted Rand index values with respect to each of the agglomeration methods and each criterion.

The averaging operation was made with taking into account clusters structure for all clustering results, i.e. the correct and incorrect number of clusters indications for each criterion, that are related to all 16 data sets (see Table 2).

Thus, in the effectiveness assessment two aspects should be included – the number of "good" and "bad" solutions (clustering results) indicated by different

criteria for selecting the number of clusters and the consistency of each grouping result with the known structure of the classes.

**Table 3.** Consistency of the grouping result according to the number of clusters selection criteria

METHOD	Criterion	Average value of adjusted Rand's index
Average linkage	Baker and Hubert (BH)	0.904
	Caliński and Harabasz (CH)	0.919
	Davies and Bouldin (DB)	0.871
	Hubert and Levine (HL)	0.713
	<b>The upper tail rule (Mojena I)</b>	<b>0.459</b>
	<b>The moving average rule (Mojena II)</b>	<b>0.945</b>
	<b>Tree validation – the criterion on the randomness of objects clustering in the dendrogram (Wishart)</b>	<b>0.946</b>
Complete linkage	Baker and Hubert (BH)	0.888
	Caliński and Harabasz (CH)	0.887
	Davies and Bouldin (DB)	0.859
	Hubert and Levine (HL)	0.778
	<b>The upper tail rule (Mojena I)</b>	<b>0.354</b>
	<b>The moving average rule (Mojena II)</b>	<b>0.905</b>
	<b>Tree validation – the criterion on the randomness of objects clustering in the dendrogram (Wishart)</b>	<b>0.890</b>
Ward's	Baker and Hubert (BH)	0.894
	Caliński and Harabasz (CH)	0.894
	Davies and Bouldin (DB)	0.905
	Hubert and Levine (HL)	0.622
	The upper tail rule (Mojena I)	0.771
	<b>The moving average rule (Mojena II)</b>	<b>0.602</b>
	<b>Tree validation – the criterion on the randomness of objects clustering in the dendrogram (Wishart)</b>	<b>0.943</b>

Source: own calculations.



## 5. Conclusions

Taking into account the results of the analyzes (see Table 2, Table 3) it can be stated that from the selected procedures aiming to determine the number of clusters, the most effective for the agglomeration methods are two of three criteria which are dedicated to the agglomerative clustering algorithms.

In the case of using medium linkage rule, high correctness in the indications of the number of clusters and the highest average consistency of the clustering result with the known clustering structure were obtained for the Wishart rule – the criterion of the randomness of the objects clustering in the dendrogram (0.946). Also, Mojena II criterion of the moving average gave high correctness and consistency (0.945). Moreover, in the cluster analysis using the complete linkage method high correctness of the grouping result and the highest average result consistency with the known classes structure was guaranteed by the same criterion (Mojena II, 0.905). The Wishart criterion on the randomness of the objects clustering in the dendrogram was characterized by a comparable high correctness and consistency (0.890). However, in the classification using the Ward's method, only Wishart criterion – tree validation (0.943) gave high correctness for the number of clusters indication and the highest average grouping result consistency with the known classes structure.

The conducted study shows that Mojena I criterion of the upper tail is absolutely ineffective in correct determination of the number of clusters, and recognition of the structure of objects belonging to the clusters. The grouping result consistence for the analyzed data sets and the mentioned criterion in the case of the medium linkage, complete linkage and the Ward's methods were equal to (0.459), (0.354) and (0.602) respectively.

Among other methods of selection of the number of clusters in a given data set, the following two criteria should be mentioned: Baker and Hubert (BH) and Caliński and Harabasz (CH), both characterized by high correctness of the number of clusters indications and consistency of the grouping result with the known class structure. Moreover, Hubert and Levin criterion usually erroneously determines the number of clusters.

## REFERENCES

- GAN, G., MA C., WU, J., Data clustering: theory, algorithms, and applications, SIAM, Philadelphia 2007.
- GATNAR, E., WALESIAK, M. (ed.), Statystyczna analiza danych z wykorzystaniem programu R, Wydawnictwo PWN, Warsaw 2009.
- MIKULEC, A., Metody oceny wyniku grupowania w analizie skupień, [in:] Jajuga K., Walesiak M. (ed.), Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2012.
- MOJENA, R., Hierarchical grouping methods and stopping rules: an evaluation, „Computer Journal” 1977, vol. 20 (4), p. 359-363.
- WISHART, D., Clustangraphics primer: a guide to cluster analysis, (the 4<sup>th</sup> edition), Edinburgh 2006.
- [www.clustan.com](http://www.clustan.com)

*STATISTICS IN TRANSITION-new series, December 2012*  
*Vol. 13, No. 3, pp. 581—600*

## **SYMBOLIC APPROACH IN REGIONAL ANALYSES**

**Justyna Wilk<sup>1</sup>**

### **ABSTRACT**

Regional studies cover a spectrum of diversified phenomena and problems including social, economic and environmental ones, which refer to territorial units. Owing to their specific characteristics they are most frequently of both multivariate and complex nature. Conducting regional research is associated with the need to consider such difficulties as large data sets, insufficient precision of phenomena description, disregarding territorial diversification of a given phenomenon, as well as incomplete description of problems.

The objective of the paper is to suggest solutions to these problems by means of symbolic approach application which basically consists in presenting phenomena in the form of symbolic data. The first part of the paper discusses specific nature of symbolic data, methods for collecting symbolic data and methods for these data analysis. The second part presents an empirical example referring to the assessment of labour market situation in Polish regions (NTS-2) using symbolic data and cluster analysis.

**Key words:** symbolic approach, symbolic data analysis, cluster analysis, regional research, labour market.

### **1. Introduction**

Regional studies cover an overall spectrum of diversified phenomena and problems including social, economic and environmental ones, referring to territorial units (TU) at international, national, regional and local level. Owing to their specific features they are most frequently of multivariate and complex nature.

Conducting regional studies is associated with the need to consider such problems as:

1. Large data set (many objects and variables) or incomplete description of the studied phenomena due to the description reduction requirement. For example, the research of local communities living standard cover a large group of TU and a

---

<sup>1</sup> Wrocław University of Economics, Department of Econometrics and Computer Science, 58-500 Jelenia Góra, Nowowiejska 3. E-mail: justyna.wilk@ue.wroc.pl.

significant spectrum of problems, including the financial situation of population, public safety level, natural environment condition, etc.

2. Insufficient precision of phenomena description and disregarding territorial diversification of a given phenomenon, in particular at higher territorial division levels. It happens as the result of data generalization, among others, due to descriptive statistics application (e.g. median, mean value). Registered unemployment rate may serve as the example, which in 2011 was specified at the level of 9.9% for Mazowieckie region (NTS-2), while in districts (NTS-4) covered by this region the range of registered values ranged from 3.8 up to 37.2%.

The objective of the paper is to suggest solutions to these problems by means of symbolic approach application which basically consists in presenting phenomena in the form of symbolic data. The application of symbolic data in regional studies allows for:

- description of higher level territorial units (e.g. NTS-1) including the situation of lower level units (e.g. NTS-4),
- characteristics of territorial units (e.g. NTS-2, NTS-5) applying not only metric and non-metric variables but also interval-valued, multivalued and modal variables.

The first part of the paper discusses the specific nature of symbolic data, methods for collecting symbolic data and methods for these data analysis. The second part presents an empirical example referring to the assessment of labour market situation in Polish regions (NTS-2) using symbolic approach and cluster analysis.

## **2. Symbolic data specification**

The approach most often applied in statistical analysis assumes that variables describing the set of objects are of metric nature (represented by a single value for each object) and/or non-metric (represented by a single category for each object) – see e.g. (Walesiak, 1993, Mynarski, 2000).

Symbolic data analysis (SDA) offers the possibility of including symbolic variables which take the form of (see Bock, Diday et al., 2000, Billard and Diday, 2006, Diday, Noirhomme-Fraiture et al., 2008):

- intervals of values, disjoint or non-disjoint (interval-valued variables),
- sets of categories or values (multivalued variables),
- sets of categories (values) with weights, frequencies, probabilities, etc. (modal variables).

The objects described by means of symbolic variables are defined as symbolic objects. With regard to aggregation level the following objects can be specified:

- first order symbolic objects which constitute objects following classical approach, i.e. primary units of the study (e.g. region, country, location)

which are described by means of symbolic and/or classical approach variables,

- second order symbolic objects which result from the aggregation of a set of first order objects (e.g. region made up of districts located in its territory) and they are described by symbolic variables only.

The set of observations referring to symbolic data is entered into a symbolic data table.

### **3. Symbolic data sources in regional studies**

#### **3.1. Aggregation of objects**

If symbolic approach is applied the set of classical data (objects described by metric and non-metric data) may be subject to aggregation process as the result of which symbolic data set is obtained. Based on the available literature studies two methods of data aggregation may be distinguished:

1. Aggregation of objects (territorial units) (see Bock, Diday et al., 2000, Diday, Noirhomme-Fraiture et al., 2008, Wilk, 2010b, pp. 86-88).
2. Aggregation of variables (descriptive characteristics for territorial units) (see Wilk, 2010b, pp. 88-91, Wilk, 2011).

The aggregation of objects consists in the representation of lower order objects by means of higher order objects, which results in obtaining second order symbolic objects. Such procedure is carried out to:

- reduce the description due to a very large set of objects, in particular local level territorial units (e.g. NTS-4, NTS-5),
- refine the description of higher order territorial units (e.g. NTS-0, NTS-2), i.e. consider their territorial diversification.

In this case symbolic data for a higher order territorial unit are obtained based on descriptive statistics values (e.g. minimum, maximum, frequency, etc.), assigned to lower order territorial units.

Interval-valued variables for a higher order territorial unit are obtained by specifying minimum and maximum values based on values obtained by lower order units. For example, on the basis of average gross monthly salary earned in sub-regions a synthetic variable may be obtained which illustrates salary spread interval in regions (see Table 1). In a similar way the following may be obtained:

- GDP *per capita* values interval in Polish regions (NTS-2), defined based on values obtained by sub-regions (NTS-3),
- unemployment rate size interval in EU countries (NTS-0), defined based on values obtained by NTS-4 level units.

**Table 1.** Average monthly gross wages and salaries in Poland in 2010 (PLN)

NTS-1 (region)		NTS-3 (sub-region)		Descriptive statistics for NTS-3		Symbolic interval-valued variable for NTS-1
Name	Value	Name	Value	Minimum value	Maximum value	
Central	4025.26	Łódzki	2705.93	2666.84	4694.47	[2666.84, 4694.47]
		City of Łódź	3243.15			
		Piotrkowski	3297.76			
		Sieradzki	2666.84			
		Skierniewicki	2810.76			
		Ciechanowsko-Płocki	3366.54			
		Ostrołęcko-Siedlecki	2950.04			
		Radomski	3054.40			
		City of Warsaw	4694.47			
		Warszawski Wschodni	3209.11			
		Warszawski Zachodni	3657.19			
Eastern	2987.67	Bialski	2841.15	2650.41	3410.28	[2650.41, 3410.28]
		Sandomiersko-Jędrzejowski	2982.23			
		.	.			
		.	.			

*Source: Author's estimation based on data provided by Local Data Bank of the Central Statistical Office in Poland.*

The construction of symbolic modal variable consists in defining the variable category (e.g. low, medium, high level; very favourable situation, moderately favourable, unfavourable, very unfavourable) and in specifying the percentage of objects meeting the criterion data. For example, in case of unemployment rate in Poland at NTS-4 level the following variable categories were defined:

- less than 10% – low level of unemployment,
- 10-20% – medium level of unemployment,
- more than 20% – high level of unemployment.

Next, the percentage of districts in each region meeting this criteria was calculated (see Table 2). Instead of defining levels the national average may be accepted as the determinant and on this basis it is possible to calculate the percentage of districts in a region ranked above and below national average level. In a similar way other economic aspects may be interpreted (e.g. average gross monthly salary), as well as environmental ones (e.g. the emission of gaseous air pollution from particularly noxious plants per 1 km<sup>2</sup>), etc.

**Table 2.** Registered unemployment rate in Poland in 2010 (%)

NTS-2 (voivodship/region)		NTS-4 (poviat/district)		Fraction of poviats/districts (NTS-4) due to phenomenon level			Symbolic modal variable for NTS-2
Name	Value	Name	Value	low	medium	high	
Mazowieckie	9.7	City of Warszawa	3.5	0.2	0.5	0.3	{low (0.2), medium (0.5), high (0.3)}
		Warszawski Zachodni	5.9				
		.	.				
		.	.				
		Radomski	30.8				
		Szydłowiecki	36.0				
Kujawsko-Pomorskie	17.0	City of Bydgoszcz	8.0	0.1	0.3	0.6	{low (0.1), medium (0.3), high (0.6)}
		.	.				
		Lipnowski	28.9				

*Source: Author's estimation based on data provided by Local Data Bank of the Central Statistical Office in Poland.*

### 3.2. Aggregation of variables

The second method for obtaining symbolic data is the aggregation based on variables describing the set of objects and, in fact, categories of these variables. Following this approach the territorial level of analyzed units is left unchanged, however, the form of describing phenomena is altered. The procedure consisting in variables aggregation may be used in situations when there is the need for:

- size reduction owing to an extensive set of variables describing territorial units at different level (e.g. counties, NTS-2, NTS-4, cities),
- synthetic description of the analyzed phenomena.

In this case the symbolic variable is constructed as the result of variable category combination following classical approach and, in particular, non-metric variable, e.g. variable presenting the availability of sport objects (swimming pools, tennis courts, sports grounds, etc.), the structure of monthly household expenditure, the occurrence of natural resources (lignite, rock salt, kaolin, etc).

Table 3 illustrates the method for multivalued variable construction which describes the availability of tertiary education institutions in Polish regions by the selected education type. For example, in Podkarpackie region only technical universities are available out of the analyzed educations profiles.

**Table 3.** Higher education institutions by type, in 2010 (facilities)

NTS-2 (voivodship / region)	Number of institutions (universities, academies) by type					Symbolic multivalued variable for NTS-2
	technical	agricultural	teacher education	medical	fine arts	
Kujawsko-Pomorskie	1	1	0	0	1	{technical, agricultural, fine arts}
Warmińsko-Mazurskie	0	0	2	0	0	{teacher education}
Śląskie	4	0	2	1	2	{technical, pedagogical, medical, fine arts}
Dolnośląskie	1	1	1	1	2	{technical, agricultural, teacher education, medical, fine arts}
Podkarpackie	1	0	0	0	0	{technical}

*Source: Author's compilation based on data from Local Data Bank of the Central Statistical Office in Poland.*

The combination of variable category illustrating groups of demographic age and their percentage share in Polish cities allowed for obtaining modal symbolic variable (see Table 4). The following factors were considered: pre-working age (up to the age of 17), working age (men aged 18-64, women aged 18-59) and post-working age (men aged 65 and more, women aged 60 and more).



**Table 4.** Population by economic age groups in Poland in 2010 (%)

NTS-5 (city)	Economic age group			Symbolic modal variable for NTS-5
Name	pre-working	working	post-working	
Łódź	13.9	64.3	21.8	{pre-working (0.139), working (0.643), post-working (0.218)}
Warszawa	15.0	63.9	21.1	{pre-working (0.150), working (0.639), post-working (0.211)}
Kraków	15.5	65.2	19.3	{pre-working (0.155), working (0.652), post-working (0.193)}
Wrocław	14.7	65.7	19.6	{pre-working (0.147), working (0.657), post-working (0.196)}

*Source: Author's compilation based on data from Local Data Bank of the Central Statistical Office in Poland.*

In a similar way the share structure may be obtained, which illustrates an economic profile of a territorial unit based on employment size in economic sectors (agricultural, industrial and service sector), or added value produced by a given sector, and also monthly household expenditure on food, clothes, footwear, house maintenance, etc.

#### 4. The symbolic data analysis methods

The complex structure of symbolic data prevents direct application of statistical methods dedicated to data following classical approach (see e.g. Hair et al., 2006, Everitt and Dunn, 2001). Two approaches are suggested in the professional literature for symbolic data analysis (see e.g. Bock, Diday et al., 2000, Wilk, 2010a).

The first approach consists in adapting algorithms originally designed for the analysis of data in classical approach and in particular distance matrix based methods. In such case distance measurement has to be performed using distance measures of symbolic objects (see e.g. Malerba et al., 2001, Malerba, Esposito, Monopoli, 2002, Wilk 2006a, 2006b).

The second approach is based on the application of methods dedicated to symbolic data, e.g. TREE algorithm from the family of classification trees, or Interscal in order to perform multivariate scaling. Selected methods, applied in symbolic data analysis, are presented in Table 5.

**Table 5.** Selected methods applied in symbolic data analysis

Name	Group	Selected algorithms	
		adopted	original
Decision trees	classification	–	TREE, BDT, SDT
	regression	–	–
Neural networks	supervised	multilayer perceptron	–
	unsupervised	self-organizing maps	–
Multidimensional scaling	based on distance matrix	bi-dimensional mapping	–
	based on symbolic data table	–	InterScal, SymScal, I-Scal
Cluster analysis	hierarchical	Ward's, complete linkage	Brito's, Chavent's
	non-hierarchical	<i>k</i> -medoids	SCLUST, DCLUST

Source: Author's compilation based on (Bock, Diday et al., 2000, Billard and Diday, 2006, Diday, Noirhomme-Fraiture et al., 2008, Wilk, 2009, 2010a).

## 5. An empirical example

The objective of the study is to define labour market situation in Polish regions (NTS-2) in 2010. Objects (regions) were described by means of variables characterizing the employment level and structure, as well as average wages and salaries. Symbolic approach was applied following which data aggregation was performed in line with classical approach. The approach based on objects aggregation as well as the one based on variables aggregation were considered.

As the result of aggregation 5 symbolic variables were obtained, including 3 interval-valued variables and 2 modal variables (see Table 6). Observations referring to variables are presented in the symbolic data table (see Appendix 1).

**Table 6.** The set of symbolic variables

No.	Variable name	Aggregation method	Type of symbolic variable	Set of variable implementation	Measure unit
1	Employment rate in sub-regions (NTS-3)	aggregation of objects	interval-valued	[36.4, 92.3]	%
2	Employment by economic sectors	aggregation of variables	modal	{agricultural [2.9, 28.2], industrial [20.9, 38.0], services [48.3, 66.2]}	%
3	Employment by education level	aggregation of variables	modal	{tertiary [21.8, 37.1], post-secondary and vocational secondary [26.3, 31.2], general secondary [6.8, 10.5], basic vocational [20.5, 34.0], lower secondary, primary and lower [4.0, 11.9]}	%
4	Job vacancy rate*	aggregation of variables	interval-valued	[0.29, 0.89]	%
5	Average monthly gross wages and salaries in poviats/districts (NTS-4)	aggregation of objects	interval-valued	[2106.84, 6012.95]	PLN

\* Job vacancy rate is the ratio of vacancies to occupied and unoccupied jobs (quarterly data).

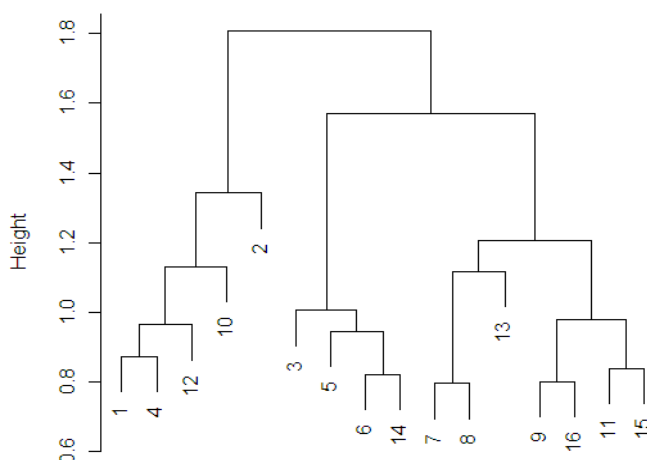
*Source: Author's compilation based on data provided by the Central Statistical Office in Poland (Local Data Bank, Labour Force Survey and elaboration on The demand for labour in 2010).*

Cluster analysis was applied in order to distinguish homogenous classes of regions. The first step of classification procedure focused on measuring symbolic objects distance using two componentwise distance measures, i.e. U\_3 Ichino-Yaguchi normalized measure for interval-valued variables and PU\_2 normalized measure for modal variables, as well as Euclidean distance as objectwise distance measure (see e.g. Bock, Diday et al., 2000). Distance measurement results are presented in Appendix 2.

The initial distant matrix analysis indicates that Mazowieckie region significantly differs from others. Measure values are higher than 1 for all regions and in particular for Lubelskie and Podkarpackie. The most similar ones, with regard to the accepted variables set, are Warmińsko-Mazurskie and Lubuskie regions.

In the next step hierarchical classification of objects set, using Ward's method, was performed. The obtained dendrogram is presented in Figure 1. Next, the number of classes was defined based on Hubert and Levine internal cluster quality index (see Figure 2). Finally, the 5 class structure was accepted for which index received the lowest value (0.396).

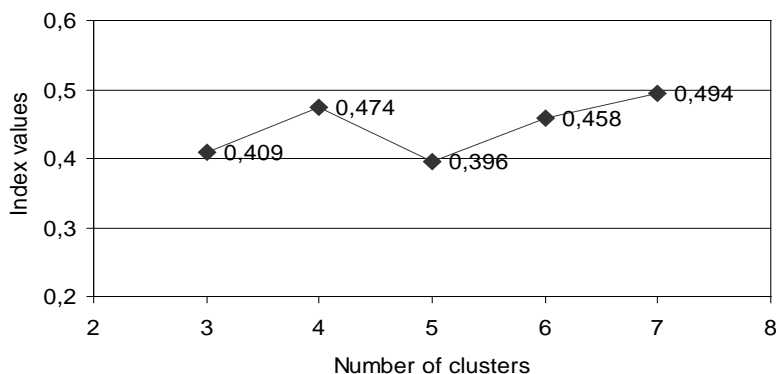
**Figure 1.** Dendrogram obtained using Ward's method



Explanations: 1 – Łódzkie, 2 – Mazowieckie, 3 – Małopolskie, 4 – Śląskie, 5 – Lubelskie, 6 – Podkarpackie, 7 – Podlaskie, 8 – Świętokrzyskie, 9 – Lubuskie, 10 – Wielkopolskie, 11- Zachodniopomorskie, 12 – Dolnośląskie, 13 – Opolskie, 14 – Kujawsko-Pomorskie, 15 – Pomorskie, 16 – Warmińsko-Mazurskie.

Source: Author's estimation in *clusterSim* and *cluster* package of R software v. 2.13.0.

**Figure 2.** Values of Hubert & Levine index for different number of clusters



Source: Author's estimation in *clusterSim* and *cluster* package of R software v. 2.13.0.

The final classification procedure step consisted in the description (interpretation) of the obtained classes using CLINT technique (see Brito, 2004). For interval-valued variables minimum and maximum variable value was defined as obtained by the given class objects (see Table 7). On the other hand, in case of modal variables the average weight obtained by objects included in the class for every variable category.

**Table 7.** Cluster description

No	Cluster members	Employment rate in NTS-3 (%)	Employment by economic sector (%)	Employment by education level (%)	Job vacancy rate (%)	Average monthly gross wages and salaries in NTS-4 (PLN)
1	Łódzkie, Śląskie, Dolnośląskie, Wielkopolskie	[40.0, 79.8]	{agricultural (9.3*), industrial (34.5), services (56.1)}	{tertiary (26.2), post-secondary and vocational secondary (28.7), general secondary (9.1), basic vocational (29.0), lower secondary, primary and lower (6.6)}	[0.48, 0.80]	[2106.84, 6012.95]
2	Mazowieckie	[43.2, 92.4]	{agricultural (11.3), industrial (22.3), services (66.2)}	{tertiary (37.1), post-secondary and vocational secondary (26.4), general secondary (9.7), basic vocational (20.5), lower secondary, primary and lower (6.3)}	[0.79, 0.89]	[2410.05, 4694.47]
3	Małopolskie, Lubelskie, Podkarpackie, Kujawsko-Pomorskie	[42.7, 73.4]	{agricultural (19.6*), industrial (27.8*), services (52.6)}	{tertiary (24.6), post-secondary and vocational secondary (28.5), general secondary (8.1), basic vocational (29.9), lower secondary, primary and lower (8.9)}	[0.29, 0.66]	[2500.93, 4428.15]

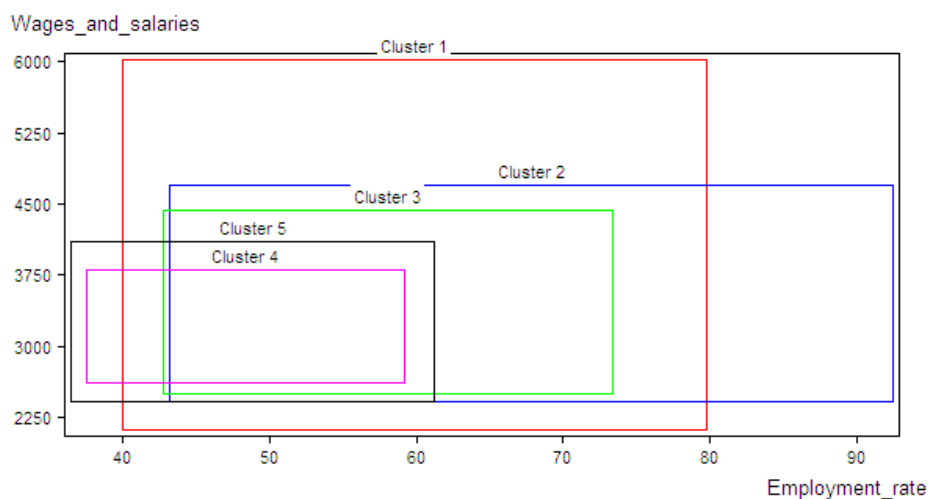
**Table 7.** Cluster description (cont.)

No	Cluster members	Employment rate in NTS-3 (%)	Employment by economic sector (%)	Employment by education level (%)	Job vacancy rate (%)	Average monthly gross wages and salaries in NTS-4 (PLN)
4	Podlaskie, Świętokrzyskie, Opolskie	[37.5, 59.2]	{agricultural (19.3*), industrial (29.6*), services (51.2)}	{tertiary (24.7), post-secondary and vocational secondary (28.7), general secondary (7.5), basic vocational (29.6*), lower secondary, primary and lower (9.4)}	[0.44, 0.86]	[2613.35, 3798.54]
5	Lubuskie, Zachodnio-Pomorskie, Pomorskie, Warmińsko-Mazurskie	[36.4, 61.2]	{agricultural (8.7), industrial (31.8), services (59.5)}	{tertiary (26.2), post-secondary and vocational secondary (27.6), general secondary (9.1), basic vocational (29.0), lower secondary, primary and lower (8.1)}	[0.45, 0.72]	[2414.19, 4108.37]

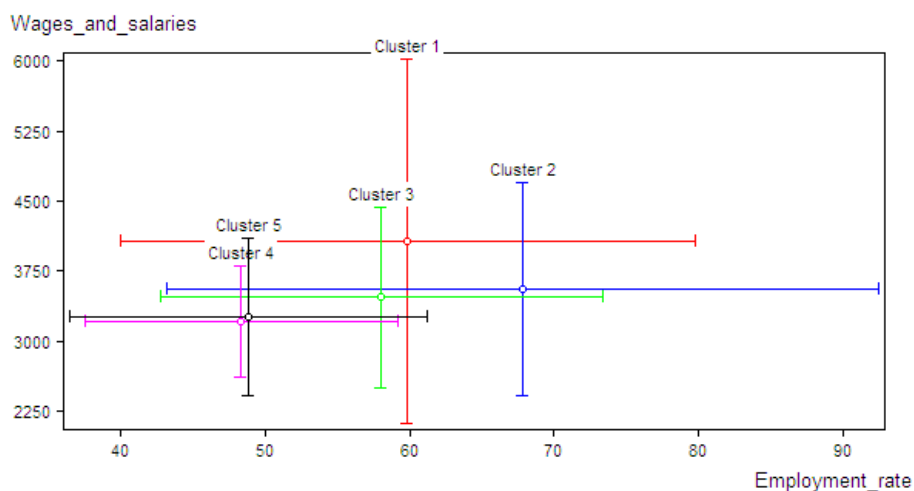
\* There is relatively high diversity of values (max – min = 10-15 percentage points).

Source: Author's estimation in *clusterSim* and *cluster* package of R software v. 2.13.0.

Picture 3 presents class characteristics with reference to employment level and average earnings. Squares in picture 4a show intervals of variables values, whereas perpendicular, intersecting lines in picture 4b illustrate centres and ranges of intervals. The smaller the squares the lower the territorial diversification of the analyzed qualities. As both pictures show the relatively low employment rate and low salaries, they are characteristic to regions included in classes 4 and 5. These are eastern and southern regions of the country, with the exception of Śląskie voivodship.

**Figure 3.** Cluster description regarding wages and salaries and employment rate

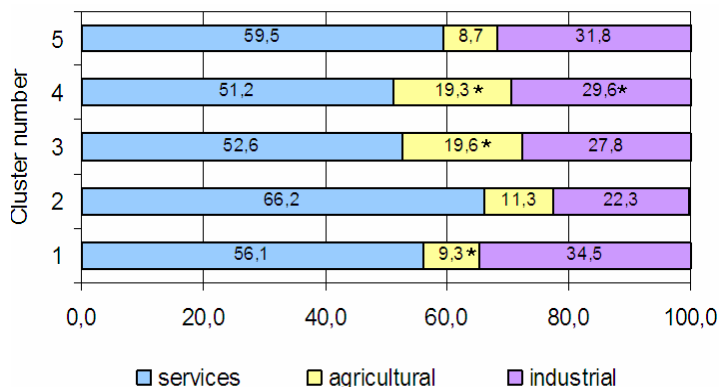
a) boxes



b) crosses

Source: Author's estimation in Sodas software v. 2.0.

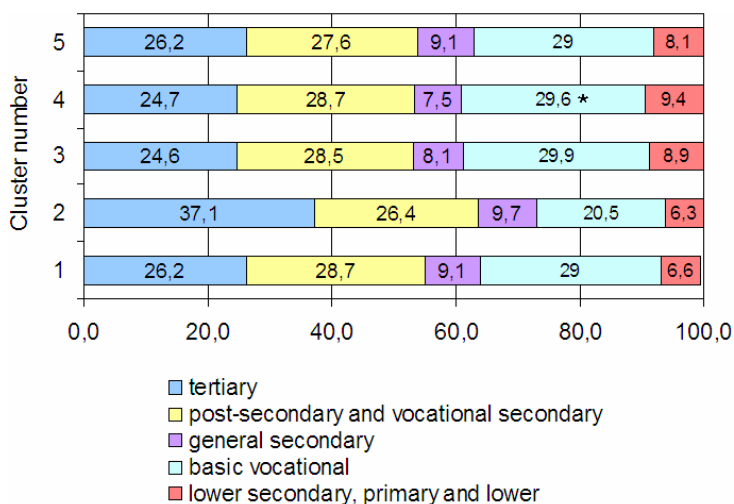
Service sector is the dominating one in all regions. In the second class represented by Mazowieckie region as many as 2/3 of labour force is employed in this sector and slightly more than 50% in the regions covered by classes 3 and 4 (see Figure 4). The regions listed in classes 4 and 5 may be referred to as the industry oriented since every third person is employed in industry sector.

**Figure 4.** Cluster description regarding employment by economic sector

\* There is relatively high diversity of values (max – min = 10-15 percentage points).

Source: Author's estimation.

Education structure among labour force is similar in all classes (see Figure 5). The occurring differences amount to about 5 percentage points. Class 2 (Mazowieckie region) is the only exception in which over 1/3 of labour force are tertiary education graduates which is responsible for 10 percentage points more than in other classes.

**Figure 5.** Cluster description regarding employment by education level

\* There is relatively high diversity of values (max-min = 10-15 percentage points).

Source: Author's estimation.



The first class includes the following regions: Łódzkie, Śląskie, Dolnośląskie and Wielkopolskie (see Figure 6). They are characterized by a very high rate of employment in industry sector at the background of other regions. Districts of these regions feature relatively high diversification regarding wages and salaries level. This group registers the lowest percentage of labour force characterized by lower secondary, primary and lower education.

**Figure 6.** Classes of regions



*Source: Author's compilation.*

The second cluster represents Mazowieckie region characterized by the highest employment rate level in services (over 66%) and the highest percentage of labour force recruiting from tertiary education graduates (over 37%). At the same time the region presents the highest diversification of districts (NTS-4) with regard to wages and salaries level, as well as many unoccupied jobs, which is indicated by high values of job vacancy rate.

The third class lists regions situated mainly in south-eastern part of the country, i.e. Małopolskie, Lubelskie, Podkarpackie and also Kujawsko-Pomorskie. They are featured by relatively low employment rate in services at sub-regional level. It has to be emphasized that labour supply in these regions corresponds to demand (job vacancy rate obtains relatively low values).

The fourth class is made up of the following regions: Podlaskie, Świętokrzyskie and Opolskie. Their characteristic feature is low employment rate at sub-regional level, as well as extensive seasonal fluctuations in this matter.

Very low earnings were registered in districts of these regions, however, at the same time the lowest territorial diversification was observed with regard to wages and salaries level.

The final, fifth class is composed of regions situated in northern (Zachodnio-Pomorskie, Pomorskie and Warmińsko-Mazurskie) and western (Lubuskie) part of Poland. The employment rate level in sub-regions of these regions is relatively low and additionally the lowest territorial diversification was also observed in this respect. Relatively high employment rate is typical for industry sector and low for agricultural sector. Districts in these regions feature relatively low level of wages and salaries.

## **6. Conclusions**

The objective of the paper is to present the proposal for symbolic approach application in regional studies. The suggested approach usefulness depends, mainly, on the goal underlying the performed study and the accepted research procedure. Symbolic approach is adequate in case of both small and large data sets when the need occurs for synthetic and detailed description of the analyzed phenomena or it is necessary to reduce the description.

Among difficulties resulting from this approach application of more laborious and time consuming research related to classical approach may be indicated, as well as more complicated interpretation of both data and the obtained results. The application of the procedure based on aggregation of objects is to a large extent determined by the availability of statistical data.

## **REFERENCES**

- BILLARD, L., DIDAY, E. (2006). *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Chichester: Wiley.
- BOCK, H. H., DIDAY, E. (Eds.) (2000). *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*, Berlin-Heidelberg: Springer-Verlag.
- BRITO, P. (2004). *Clustering Interpretation. Interpreting Clusters by using the module CLINT*, In: M. Noirhomme-Fraiture (Ed.), *User manual for SODAS 2 software, Software Report, Analysis System of Symbolic Official Data*, Project no. IST-2000-25161.
- DIDAY, E., NOIRHOMME-FRAITURE, M. (Eds.) (2008). *Symbolic data analysis and the Sodas software*, Chichester: John Wiley & Sons.
- EVERITT, B. S., DUNN, G. (2001). *Applied Multivariate Data Analysis*, London: Arnold.

- HAIR, J. F., BLACK, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L. (2006). *Multivariate Data Analysis*, New Jersey: Pearson Prentice Hall.
- MALERBA, D., ESPOSITO, F., GIOVALLE, V., TAMMA, V. (2001). Comparing Dissimilarity Measures for Symbolic Data Analysis, In: P. Nanopoulos (Ed.), *New Techniques and Technologies for Statistics and Exchange of Technology and Know-how (ETK-NTTS'01)*, conference materials, pp. 473-481.
- MALERBA, D., ESPOSITO, F., MONOPOLI, M. (2002). Comparing dissimilarity measures for probabilistic symbolic objects, In: A. Zanasi, C.A. Brebbia, N.F.F. Ebecken, P. Melli (Eds.), *Data Mining III, Series Management Information Systems*, vol. 6, Southampton: WIT Press, pp. 31-40.
- MYNARSKI, S. (2000). *Praktyczne metody analizy danych rynkowych i marketingowych [Practical methods for market and marketing data analysis]*, Kraków: Zakamycze.
- Popyt na pracę w 2010 r. [The demand for labour in 2010] (2011). *Statistical Information and Studies*, Central Statistical Office, Warsaw: Central Statistical Office Publishing House.
- WALESIK, M. (1993). Strategie postępowania w badaniach statystycznych w przypadku zbioru zmiennych mierzonych na skalach różnego typu [Procedures in statistical research in case of variables set measured in different types of scales], *Badania Operacyjne i Decyzje [Operational Studies and Decisions]* no. 1, pp. 71-77.
- WILK, J. (2010a). Metody analizy danych symbolicznych [Symbolic data analysis methods], In: J. Dziechciarz (Ed.), *Econometrics 29. The Application of quantitative methods*, Research Studies of Wrocław University of Economics no. 141, Wrocław: Wrocław University of Economics Publishing House, pp. 29-38.
- WILK, J. (2006b). Miary odległości obiektów opisanych zmiennymi symbolicznymi z wagami [Distance measures for objects described by symbolic variables with weights], In: K. Jajuga, M. Walesiak (Eds.), *Taxonomy 13. Data classification and analysis – theory and applications*, Research Studies of University of Economics in Wrocław no. 1126, Wrocław: University of Economics in Wrocław Publishing House, pp. 224-236.
- WILK, J. (2006a). Problemy klasyfikacji obiektów symbolicznych. Symboliczne miary odległości [Problems of symbolic objects classification. Symbolic distance measures], In: J. Garczarczyk (Ed.), *Quantitative and qualitative methods for market research. Measurement and its efficiency*, Research Bulletins University of Economics in Poznań no. 71, Poznań: University of Economics in Poznań Publishing House, pp. 69-83.

- WILK, J. (2010b.) Problemy segmentacji rynku z wykorzystaniem metod klasyfikacji i danych symbolicznych [Market segmentation problems applying classification methods and symbolic data], Wrocław University of Economics, doctoral dissertation (unpublished), Jelenia Góra.
- WILK, J. (2009). Przegląd metod wielowymiarowej analizy statystycznej wykorzystywanych w badaniach segmentacyjnych [The review of multivariate statistical analysis applied in segmentation studies], In: J. Dziechciarz (Ed.), *Econometrics* 23. The Application of quantitative methods, Research Studies of Wrocław University of Economics no. 37, Wrocław: University of Economics in Wrocław Publishing House, pp. 59-70.
- WILK, J. (2011). Taksonomiczna analiza rynku pracy województw Polski – podejście symboliczne [Taxonomic analysis of labour market in Polish regions – symbolic approach], In: J. Dziechciarz (Ed.), *Econometrics* 34. The Application of quantitative methods, Research Studies of Wrocław University of Economics no. 200, Wrocław: Wrocław University of Economics Publishing House, pp. 26-37.

## APPENDIX

## Appendix 1. Symbolic data table\*


	Employment_rate	Employment_by_economic_sectors	Employment_by_education_levels	Job_vacancy_rate	Wages_and_salaries
LÓDŹSKIE	[49.09 : 61.78]	agricult (0.13), industr (0.32), services (0.55)	tertiary (0.26), post-sec (0.28), general (0.19), basic vo (0.27), lower se (0.09)	[0.48 : 0.80]	[2379.47 : 4628.24]
MAZOWIECKIE	[43.17 : 52.36]	agricult (0.11), industr (0.22), services (0.66)	tertiary (0.37), post-sec (0.26), general (0.19), basic vo (0.20), lower se (0.06)	[0.79 : 0.89]	[2410.05 : 4694.47]
MAŁOPOLSKIE	[46.82 : 73.35]	agricult (0.14), industr (0.30), services (0.56)	tertiary (0.26), post-sec (0.28), general (0.09), basic vo (0.30), lower se (0.07)	[0.29 : 0.52]	[2562.74 : 3543.43]
ŚLĄSKIE	[40.91 : 63.47]	agricult (0.03), industr (0.38), services (0.59)	tertiary (0.26), post-sec (0.31), general (0.09), basic vo (0.29), lower se (0.04)	[0.51 : 0.72]	[2411.78 : 5213.89]
LUBELSKIE	[50.62 : 57.04]	agricult (0.26), industr (0.21), services (0.51)	tertiary (0.25), post-sec (0.31), general (0.07), basic vo (0.26), lower se (0.11)	[0.31 : 0.59]	[2548.07 : 4428.15]
PODKARPAŃSKIE	[51.64 : 59.31]	agricult (0.22), industr (0.28), services (0.50)	tertiary (0.25), post-sec (0.28), general (0.07), basic vo (0.31), lower se (0.10)	[0.32 : 0.64]	[2401.90 : 3428.12]
PODŁASKIE	[47.79 : 51.72]	agricult (0.23), industr (0.24), services (0.53)	tertiary (0.20), post-sec (0.30), general (0.06), basic vo (0.24), lower se (0.12)	[0.44 : 0.60]	[2610.35 : 3241.41]
ŚWIĘTOKRZYSKI	[50.48 : 59.18]	agricult (0.22), industr (0.29), services (0.48)	tertiary (0.25), post-sec (0.28), general (0.07), basic vo (0.31), lower se (0.09)	[0.41 : 0.83]	[2638.63 : 3231.66]
LUBUSKIE	[44.85 : 48.01]	agricult (0.08), industr (0.34), services (0.58)	tertiary (0.23), post-sec (0.30), general (0.09), basic vo (0.30), lower se (0.07)	[0.45 : 0.72]	[2414.19 : 3118.35]
WIELKOPOLSKIE	[46.73 : 78.80]	agricult (0.15), industr (0.34), services (0.51)	tertiary (0.25), post-sec (0.27), general (0.07), basic vo (0.34), lower se (0.07)	[0.56 : 0.77]	[2106.84 : 3814.03]
ZACHODNIOPOLSKIE	[36.35 : 55.09]	agricult (0.08), industr (0.30), services (0.62)	tertiary (0.28), post-sec (0.28), general (0.09), basic vo (0.28), lower se (0.08)	[0.56 : 0.68]	[2482.27 : 3586.82]
DOLNOŚLĄSKIE	[40.04 : 69.83]	agricult (0.06), industr (0.34), services (0.60)	tertiary (0.27), post-sec (0.30), general (0.19), basic vo (0.27), lower se (0.06)	[0.57 : 0.79]	[2598.06 : 6012.95]
OPOLSKIE	[37.53 : 47.64]	agricult (0.12), industr (0.36), services (0.52)	tertiary (0.22), post-sec (0.29), general (0.08), basic vo (0.34), lower se (0.07)	[0.68 : 0.83]	[2730.02 : 3788.54]
KUJAWSKO-POMIĘSKIE	[42.65 : 56.14]	agricult (0.14), industr (0.32), services (0.54)	tertiary (0.22), post-sec (0.27), general (0.09), basic vo (0.33), lower se (0.09)	[0.36 : 0.66]	[2500.93 : 3220.65]
POMORSKIE	[41.11 : 61.15]	agricult (0.07), industr (0.31), services (0.61)	tertiary (0.28), post-sec (0.26), general (0.19), basic vo (0.28), lower se (0.07)	[0.49 : 0.64]	[2557.33 : 4188.37]
WARMIŃSKO-MAZURSKIE	[39.05 : 45.20]	agricult (0.12), industr (0.31), services (0.57)	tertiary (0.26), post-sec (0.26), general (0.08), basic vo (0.29), lower se (0.11)	[0.50 : 0.70]	[2427.79 : 3443.21]


\* The names of the selected objects, variables and variables categories were shortened (for full names see Table 6 and Figure 6).

Source: Author's compilation.

## Appendix 2. Dissimilarity matrix

Object (territorial unit) number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	1.273														
3	1.193	1.553													
4	0.873	1.270	1.226												
5	1.106	<b>1.590</b>	0.973	1.148											
6	1.072	1.571	0.943	1.152	0.862										
7	0.986	1.387	1.242	1.172	1.130	1.040									
8	0.958	1.388	1.184	1.134	1.118	0.968	0.796								
9	0.996	1.455	1.148	1.062	1.109	0.977	0.886	0.916							
10	0.974	1.160	1.197	1.070	1.264	1.160	1.117	1.039	1.107						
11	1.013	1.347	1.193	0.990	1.136	1.049	1.056	1.048	0.910	1.092					
12	1.000	1.191	1.347	0.882	1.315	1.336	1.261	1.230	1.246	1.147	1.119				
13	1.078	1.334	1.430	1.134	1.365	1.289	1.041	1.052	1.057	1.151	0.988	1.163			
14	1.042	1.500	0.985	1.097	0.966	0.821	1.003	0.962	0.865	1.121	0.930	1.277	1.185		
15	0.897	1.327	1.059	0.898	1.012	0.958	1.064	1.017	0.948	1.043	0.839	1.055	1.111	0.909	
16	1.014	1.450	1.209	1.045	1.132	1.033	0.971	1.009	<b>0.800</b>	1.154	0.833	1.208	1.008	0.911	0.915

 low similarity (above 1.4)

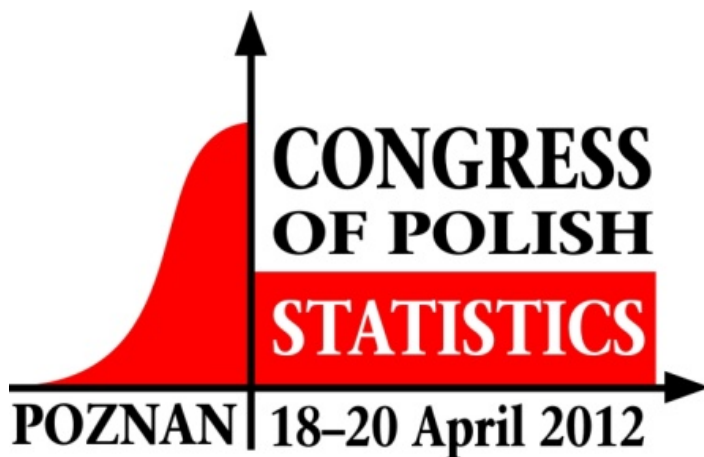
 high similarity (below 0.9)

Explanations: 1 – Łódzkie, 2 – Mazowieckie, 3 – Małopolskie, 4 – Śląskie, 5 – Lubelskie, 6 – Podkarpackie, 7 – Podlaskie, 8 – Świętokrzyskie, 9 – Lubuskie, 10 – Wielkopolskie, 11- Zachodniopomorskie, 12 – Dolnośląskie, 13 – Opolskie, 14 – Kujawsko-Pomorskie, 15 – Pomorskie, 16 – Warmińsko-Mazurskie.

Source: Author's compilation.

STATISTICS IN TRANSITION-new series, December 2012  
Vol. 13, No. 3, pp. 601–606

**The Congress of Polish Statistics to mark the 100th  
anniversary of The Polish Statistical  
Association, Poznań, 18–20 April 2012**



The Congress of Polish Statistics to mark the 100th anniversary of the Polish Statistical Association was held in Poznań between 18 and 20 April 2012. The year-long celebration featured exhibitions, seminars and conferences organized by regional sections of the Polish Statistical Association, including those in Wrocław, Lublin, Toruń and Bydgoszcz.

The Congress of Polish Statistics was a unique event with international participation, under the honorary patronage of the President of the Republic of Poland, Bronisław Komorowski. In a letter of congratulations, the President recognized the significance of the event and referred to the main goal that had inspired the founders of the Association. Stressing the role of the publication of *Statystyka Polska* in 1915, he noted its contribution as a valuable source of knowledge about the fatherland divided between the three partitioning states, the knowledge that was later used to determine the borders of the Second Polish Republic. The President noted that the following history of the PSA is marked by the dramas of Polish history, adding that ‘... *Members of the Association got involved in the struggle for the independence and defense of their country, doing their best to retain their national identity and create and develop civil society.*’

This great moment of Polish statistics was highlighted by the participation of eminent representatives of statistical institutions and organizations from Austria, Australia, Estonia, France, Holland, Germany, Norway, the USA and Italy. The

Congress was attended by numerous members of the statistical community from universities, institutes, Polish and international research organizations, the Central Statistical Office, regional offices and other national statistical institutes. Among special guests present during the anniversary session were representatives of state authorities and local and regional government, the business sector, administration and the student community. In total, about 600 participants attended the Congress.

The anniversary celebration was marked by the presence of representatives of international statistical organizations: Dr M. Bohata – Deputy Director General of Eurostat, Prof. I. Ritschelova – President of the Czech Statistical Office, Dr E. Laczka – Vice-President of the Hungarian Statistical Office and representatives of international statistical associations: Prof. R. Chambers – President of International Association of Survey Statisticians at the International Statistical Institute, Prof. R. Wasserstein – the Executive Director of the American Statistical Association, Prof. I. Traat – the President of the Estonian Statistical Association, Prof. W. Schmid – board member of the German Statistical Society. The Congress was also attended by numerous representatives of the Polish statistical community from universities, institutes, associations, the Central Statistical Office, regional offices and other institutions of public statistics, the business sector, administration and the student community.

The significance of the conference is reflected by the contents of the papers delivered during the three plenary sessions, one discussion panel, 35 parallel thematic sessions (including 12 in English) and 20 posters in the poster session. Altogether 155 papers co-authored by nearly 200 scientists were delivered. The conference program comprised a number of thematic sessions, including those devoted to survey methodology, regional statistics, population statistics, social and economic statistics, the problems of statistical data and statistics of health, sport and tourism.

The topic of the sessions on the first day of the Congress was *History and Development of Polish Statistics*. The first one featured presentations on major achievements of Polish statistics by eminent representatives of various fields. A historical perspective of the development of Polish statistics was presented by W. Ostasiewicz. J. Mielniczuk talked about Polish contribution to mathematical and applied statistics. T. Caliński's paper was an overview of Polish achievements in biometrics. The development of statistics in economic sciences was presented by K. Jajuga. J. Witkowski provided an overview of the historical development and current challenges facing public statistics. The second session was devoted to presenting the biographies and legacy of two most renowned Polish statisticians. T. Ledwina's talk focused on the life and achievements of Jerzy Neyman, while M. Krzyśko introduced the biography and legacy of Jan Czekanowski. The doyenne of Polish statistics, S. Bartosiewicz presented (together with E. Stańczyk) the socio-economic history of Poland between 1918 and 2008, while J. Kordos discussed the relationship between the development of theory and practice in survey methodology in Poland.



Another important event held on the first day of the Congress was the discussion panel on the future of statistics presided over by T. Caliński. The panel featured contributions from J. Józwiak in the area of demography, J. Witkowski and T. Panek in the area of social statistics, J. Wilkin in the area of economic statistics, J. Paradysz in the area of regional statistics, J. A. Moczko who addressed the problems of health statistics, M. Szreder on the quality of statistical data, and J. Koronacki who spoke about prospects for the methodology of statistical research.

The sessions on the second and third day of the Congress were very intensive. The program comprised a number of parallel thematic sessions chaired by respected and prominent statisticians of national and international renown. Keynote speakers included Polish and foreign representatives of various disciplines of statistical research. The present issue of *Statistics in Transition – new series*, just like the previous one, presents many of the papers delivered during the Congress. The Congress content can also be found in special issues of *Przegląd Statystyczny / Statistical Review* (two volumes) and *Studia Demograficzne / Demographic Studies* (No. 1/2012). In the *Wiadomości Statystyczne/ Statistical News*, starting from issue No. 7/2012, there is a section devoted to the Congress of Polish Statistics presenting a detailed overview of individual sessions and selected papers. Many eminent Polish and foreign statisticians took part in the Congress, making contributions that will provide enough valuable material for future analysis and reference. Listed below are topics of the Congress sessions, including invited papers. A detailed Congress program is available at [http://www.stat.gov.pl/pts/kongres2012/dok/ProgramPolski\\_18-04-2012.pdf](http://www.stat.gov.pl/pts/kongres2012/dok/ProgramPolski_18-04-2012.pdf).

Thematic sessions of the Congress of Polish Statistics, organizers and topics of invited papers:

1. **Mathematical statistics** – organized by Prof. Mirosław Krzyśko, Adam Mickiewicz University in Poznań
  - Tadeusz Bednarski (Wrocław University), *A statistical analysis of causes of bias in labour market surveys*
  - Jacek Koronacki (Institute of Computer Science of the Polish Academy of Sciences), *Analysing multivariate data given a small sample size*
  - Jana Jureckova (Charles University in Prague), *Regression quantiles and their two-step modifications*
  - Zbigniew Szkutnik (AGH University of Science and Technology in Cracow), *The EM algorithm and its modifications*
2. **Survey sampling and small area statistics** – organized by Prof. Janusz Wywił, University of Economics in Katowice
  - Malay Ghosh (University of Florida), *Finite population sampling: a model-design synthesis*
  - Ray Chambers, Gunky Kim (Wollongong University), *Regression analysis using data obtained by probability linking of multiple data sources*
  - Li-Chun Zhang (Statistics Norway), *Micro calibration for data integration*
  - Lorenzo Fattorini (University of Siena), *Design-Based Inference on Ecological Diversity*

- Jean-Claude Deville, Daniel Bonnéry, Guillaume Chauvet (ENSAE), *Neyman type optimality for marginal quota sampling*
- 3. **Population statistics** – organized by Prof. Janina Józwiak, Warsaw School of Economics
  - Nico Keilman (University of Oslo), *Challenges for statistics on households and families*
  - Irena E. Kotowska (Warsaw School of Economics), *Evolving population structures and their relevance for demographic and social change*
  - Francesco Billari (Bocconi University, Milan), *Challenges for new methods in demographic analysis*
- 4. **Social statistics** – organized by Prof. Tomasz Panek, Warsaw School of Economics
  - Achille Lemmi (Siena University, Dipartimento di Economia Politica) *Dimensions of Poverty. Theory, Models and New Perspectives*
  - Walenty Ostasiewicz (University of Economics in Wrocław), *Life quality as a subject of statistical research*
  - Anne Valia Goujon, Ramon Bauer, Samir K.C., Michaela Potančoková (Vienna Institute of Demography (VID) of the Austrian Academy of Sciences), *The Human Capital Puzzle: Why It Is so Hard to Find Good Data on Educational Attainment?*
  - Urszula Sztanderska (University of Warsaw), *Polish public statistics as a source of inspiration for and an obstacle to the development of labour market research*
- 5. **Economic statistics** – organized by Prof. Jerzy Wilkin, University of Warsaw
  - Włodzimierz Siwiński (University of Warsaw, L. Koźmiński Academy), *A statistical overview of the 2008 crisis*
  - Barbara Liberda (University of Warsaw), *Generational accounting by means of measuring the wealth of consecutive generations*
  - Elżbieta Mączyńska (Institute of Economic Sciences of the Polish Academy of Sciences, Polish Economic Society), *Statistics as a source of data in research. Dilemmas and incompatibilities*
  - Andrzej P. Wiatrak (University of Warsaw), *Contemporary problems of agricultural statistics*
- 6. **Regional statistics** – organized by Prof. Tadeusz Borys, University of Economics in Wrocław
  - Jan Paradysz (Regional Statistics Center, Poznań University of Economics), *Regional statistics: the current state, problems and directions*
  - Tadeusz Borys (University of Economics in Wrocław), Tomasz Potkański (The Polish Towns Union), *Monitoring regional and local development*
- 7. **Data analysis and classification** – organized by Prof. Krzysztof Jajuga, Prof. Marek Walesiak, University of Economics in Wrocław
  - Reinhold Decker (Universität Bielefeld), *Model-based analysis of online consumer reviews – methods and applications*

- Wojtek Krzanowski (University of Exeter), *Classification: some old principles applied to new problems*
- Patrick Groenen (Erasmus University Rotterdam), *The support vector machine as a powerful tool for binary classification*
- 8. **Statistical data** – organized by Prof. Mirosław Szreder, University of Gdańsk
- 9. **Health statistics** – organized by Prof. Jerzy Andrzej Moczko, Poznań University of Medical Sciences
- 10. **Statistics of sport and tourism** – organized by Prof Bogdan Sojkin, Poznań University of Economics

The Congress was accompanied by a series of anniversary publications, exhibitions and events. The open anniversary session of the Main Council of the PSA was an excellent opportunity to honour its most distinguished members and present the activities undertaken by regional sections. The medal of Jerzy Sława-Neyman was presented by Prof. C. Domański – the President of PSA – to five distinguished Professors: Sir A. Atkinson (London School of Economics), T. Caliński (University of Life Sciences in Poznań), M. Ghosh (University of Florida), J. Jurečková (Charles University in Prague), W. Krzanowski (University of Exeter). Honorary membership of the Polish Statistical Association was awarded by the Main Council of PSA to Dr K. Kruska. The honorary badge of merit “For services to the statistics of Poland” was presented by Prof. J. Witkowski – the President of Central Statistical Office – to M. Krzyśko, J. Pocięcha, B. Podolec, A.S. Barczak, A. Młodak and T. Klimanek.

Two anniversary publications were issued by ZWS GUS to mark the Congress. The book entitled „Statystycy polscy” (Polish statisticians) contains biographical notes about 85 Polish statisticians who played a major role in the history of Polish statistics. The book was compiled by a team consisting of W. Adamczewski, J. Berger, K. Kruska, M. Krzyśko (editor-in-chief) and B. Łazowska. The second book, entitled „Polskie Towarzystwo Statystyczne 1912–2012” (Polish Statistical Association 1912–2012) contains detailed information about the development of PSA and a description of its activities. The book was co-authored by J. Berger, J. Kordos, K. Kruska (editor-in-chief), W.W. Łagodziński and W. Ostasiewicz. The Congress was accompanied by two anniversary exhibitions: „Polish Statistical Association 1912–2012” and „Statistics in Wielkopolska”. They presented the statistical legacy of the past centuries, featuring archival results and publications that had contributed to the development of statistical thought and practice in Poland and Wielkopolska. The exhibitions also highlighted the people who had created this legacy. Both exhibitions are available in electronic form. The exhibition „Polish Statistical Association 1912–2012” prepared by J. Berger is available at the Congress website:

[http://www.stat.gov.pl/pts/kongres2012/dok/Polskie\\_Towarzystwo\\_Statystyczne.pdf](http://www.stat.gov.pl/pts/kongres2012/dok/Polskie_Towarzystwo_Statystyczne.pdf).

The exhibition „Statistics in Wielkopolska” prepared by the Statistical Office in Poznań was also presented at the Poznań Town Hall and at the Province Office of Wielkopolska as well as at the Poznań branch office of the National Bank of Poland. Its electronic version is available at:

[http://www.stat.gov.pl/pts/kongres2012/dok/statystyka\\_w\\_wielkopolsce.pdf](http://www.stat.gov.pl/pts/kongres2012/dok/statystyka_w_wielkopolsce.pdf).

Along with the Congress, two accompanying workshops were held between 16 and 17 April 2012, which were attended by about 60 participants. The workshops focused on two major areas of methodology: one was devoted to the integration of statistical data and small area estimation (conducted by Marcello D'Orazio and Fabrizio Solari, specialists from the Italian Statistical Office ISTAT) and the other one to data mining methods (prepared by the Statsoft company). The Congress also provided a fitting context for the closing ceremony of the provincial knowledge competition “Statistics concerns me” to mark the Day of Polish Statistics in 2012. The competition addressed to secondary school students is organized annually by the Statistical Office in Poznań and the Polish Statistical Association. The Congress cultural program featured a walk round the Old Town, including an organ concert in the parish church, a visit to the Old Brewery Commerce Art and Business Centre, the Welcome Party in the historic pub, the Jubilee Concert performed by students and graduates of the Academy of Music in Poznań and the chamber choir *Musica Viva* of the Poznań University of Economics conducted by Prof. M. Gandecki. The concert was prepared and hosted by Prof. H. Lorkowska, Rector of the Academy of Music in Poznań.

The celebration of the 100th anniversary of the Polish Statistical Association was organized as a joint enterprise of the whole statistical community represented by four institutions: the Polish Statistical Association, the Central Statistical Office, the Statistical Office in Poznań, and the Poznań University of Economics. In this way, the Congress was given a unique scope in the history of Polish statistics, as an event uniting communities representing both the practical and the theoretical dimension of statistical research and providing an opportunity for the exchange of ideas and experiences for representatives of public statistics, academic centres, administration authorities, entrepreneurs and other partners involved in the business of investigating economic, social and demographic processes. The Congress helped to direct methodological research conducted by the statistical community in the years to come. The priority research areas were formally recognised in the final resolution. It is already evident that the Congress was one of the most important and influential scientific events in the history of Polish statistical thought and practice.

Prepared by:  
Elżbieta Gołata

STATISTICS IN TRANSITION-new series, December 2012  
Vol. 13, No. 3, pp. 607–610

## REPORT ON SURVEY SAMPLING AND SMALL AREA STATISTICS SESSIONS DURING THE CONGRESS OF POLISH STATISTICS IN POZNAŃ, 18–20 APRIL 2012

Survey sampling is a field of statistics with a long time tradition in Poland starting from Jerzy Sława Neyman's papers on stratified random sampling. The considerations were continued, among others, by R. Zasępa and Z. Pawłowski. Nowadays, research of many Polish survey statisticians in Poland include studies on small area estimation. This is the reason why four sessions on both survey sampling and small area estimation were organized during the Congress of Polish Statistics. During the sessions fourteen papers in English were presented, including five invited lectures given by Ray Chambers from Wollongong University, Jean-Claude Deville, Lorenzo Fattorini from University of Siena, Malay Ghosh from University of Florida and Li-Chun Zhang from Statistics Norway.

Ray Chambers presented a paper prepared together with Gunky Kim entitled *Regression analysis using data obtained by probability linking of multiple data sources*. In the presentation authors described how the framework for the inference using probability-linked data can be extended to define methods for efficient bias-corrected regression analysis when three registers are linked, or when a sample is linked to two separate registers, enabling longitudinal analysis. Their development allows for correlated linkage errors as well as errors arising when not all records can be linked. It can also be extended when more than three data sets are linked. The authors presented, among other things, the results of simulation study on accuracy of regression parameters.

Jean-Claude Deville together with Daniel Bonnéry and Guillaume Chauvet prepared a presentation entitled *Neyman type optimality for marginal quota sampling*. They were seeking for a similar to Neyman optimization procedure leading to inclusion probabilities specific for each cell of the table. A simple model based method was used but the assumptions of the model were strong. When the sampling is near from maximal entropy, there exists a quite simple approximation of the variance which is model free. The authors tried to optimize this approximation. A natural iterative procedure was used and it was shown to converge to this optimum.

Lorenzo Fattorini presented a lecture on *Design-based inference on ecological diversity*. He considered three main issues of ecological diversity analysis. Firstly, he studied the problem of measuring ecological diversity by means of suitable

indexes. Secondly, the problem of estimating these indexes was considered when biological populations were sampled by means of sampling schemes actually adopted by biologists. Thirdly, the problem of comparing and ordering biological communities with respect to their diversity was analyzed.

The presentation of Malay Ghosh entitled *Finite population sampling: a model-design synthesis* was devoted to the problem of the choice of the approach in the survey sampling between the design-based and the model-based one. While the basic conceptual disagreement between the two approaches could not be resolved, from an operational point of view, the author found an agreement between the two. He derived some general results which provided this agreement exactly or asymptotically and illustrated the results with several examples.

Li-Chun Zhang presented a paper *Micro calibration for data integration*. In the paper he explained how micro calibration could be achieved for categorical as well as continuous data. Some possibilities of using micro calibrated data were discussed, such as for the purpose of small area estimation.

In their presentation Konrad Furmańczyk and Stanisław Jaworski from Medical University of Warsaw and Warsaw University of Life Sciences entitled *Change point detection in a sequence of independent observations* proposed a new method where the change was identified when some given threshold was exceeded. The method was based on minimax decision rule.

Sara Franceschi from Università di Firenze presented a paper prepared together with Lorenzo Fattorini from Università di Siena and Daniela Maffei from Università di Firenze entitled *Design-based treatment of unit nonresponse by the calibration approach*. They considered the non-response calibration weighting in a complete design-based framework. Approximate expressions of design-based bias and variance of the calibration estimator were derived, design-based consistency was investigated and some estimators of the sampling variance were proposed. The results of the simulation study demonstrated how the reliability of the procedure was mainly determined by the capability of selecting auxiliary variables in such a way that their relationship with the interest variable was similar for both the respondent and non-respondent sub-populations.

Wojciech Gamrot from University of Economics in Katowice prepared a presentation *On empirical inclusion probabilities*. The author studied the problem when unknown first order inclusion probabilities were unknown and had to be replaced by estimates obtained in a simulation study. Such estimates are called empirical inclusion probabilities. In the presentation the number of sample replications was analyzed which should have been drawn to ensure desired accuracy of the population total estimates. What is more, an attempt was made to review known solutions to this problem and to improve them for a certain sampling scheme.

Tomasz Klimanek from both University of Economics in Poznań and Statistical Office in Poznań presented a paper *Using indirect estimation with*

*spatial autocorrelation in social surveys in Poland* with an application of indirect estimation to estimate some characteristics of labour market in the population of people aged 15 and over at the level of NUTS4 in Wielkopolska region in 2008. What is more, he compared the precision of the direct estimator with those of the EBLUP estimator and the SEBLUP estimator (which takes into account spatial correlation).

Tomasz Józefowski from Statistical Office in Poznań presented a paper *Using a SPREE estimator to estimate the number of unemployed across subregions*. He used SPREE estimator to adjust values in the cells of an estimated contingency table to the totals obtained by means of survey sampling. The example of estimating the number of unemployed at the level of subregions of Wielkopolska province using data on registered unemployment and LFS was presented.

Marcin Szymkowiak from the University of Economics in Poznań presented a paper *Construction of calibration estimators of total for different distance measures*. The main goal of the paper was to present the construction of calibration estimators for the total for different types of distance measures. Its empirical part, based on a simulation study using real data, provided a comparison of different types of distance measures that could be used in the construction of calibration weights in the case of non-response.

Imbi Traat from University of Tartu in a paper *Domain Estimators Calibrated on Reference Survey* studied the situation when population totals, or totals of larger domains, for certain variables were estimated in one survey called the reference survey. Estimates based on the same variables but at a more detailed domain level were obtained from a second (later) survey, called the present survey. Consistency was required for the present survey estimates in the sense that domain totals for common study variables had to sum up to the corresponding estimated totals in the reference survey (or - in special case - known totals from registers). To solve the problem the author used calibration framework but in a more general setting.

Janusz L. Wywiał from University of Economics in Katowice presented a paper on *Estimation of population mean on the basis of a simple sample ordered by auxiliary variable*.

He considered estimation of the population average in a finite population by means of sampling strategies dependent on an order statistic of an auxiliary variable. The inclusion probabilities were dependent on the probability function of the order statistic of the auxiliary variable highly correlated with a variable of interest. The simple estimator of the population mean (total) of the variable under study was proposed. Its basic parameters were derived and the limit distribution of the estimator was also considered. Finally, on the basis of the simulation analysis, the accuracy of the estimator was compared with the precision of the well known estimators including the simple regression one.

Tomasz Żądło from University of Economics in Katowice in his paper *On prediction of totals for spatially correlated domains* studied small area estimation problems for longitudinal data. He proposed some special case of the General Linear Mixed Model including spatial and temporal correlation and assumed that the population could change in time and domains affiliation of population elements could change in time. Based on the assumption the author derived the empirical best linear predictor and the estimator of its MSE. Theoretical considerations were supported by simulation study.

Prepared by:

Janusz L. Wywiał, Tomasz Żądło



STATISTICS IN TRANSITION-new series, December 2012  
Vol. 13, No. 3, pp. 611—616

## REPORT

### **The XXXI International Conference on Multivariate Statistical Analysis, 12–14 November 2012, Łódź, Poland**

The 31st Edition of the International Annual Conference on **Multivariate Statistical Analysis** was held in **Łódź, Poland** from 12 to 14 November 2012. The MSA 2012 Conference was organized by the **Department of Statistical Methods** of the University of Łódź and the **Institute of Statistics and Demography** of the University of Łódź, with the cooperation of the **Polish Statistical Association**. The Organizing Committee was led by **Professor Czesław Domański**. The scientific secretaries functions were entrusted to Artur Mikulec, Ph.D. and Aleksandra Kupis-Fijałkowska, M.Sc.

The Multivariate Statistical Analysis MSA 2012 Conference was covered by the honorary patronage of the President of the Central Statistical Office of Poland, **Professor Janusz Witkowski**. Its organization was partially financed by the **National Bank of Poland**. The official partners of the MSA 2012 Conference were **StatSoft Polska Sp. z o.o.** and the **Polish Society of Market and Opinion Researchers - PTBRiO**.

The 2012 Edition, as all Multivariate Statistical Analysis conferences, aimed to create an opportunity for scientists and practitioners of Statistics to present and/or discuss the latest theoretical achievements in the field of the multivariate statistical analysis, its practical aspects and applications. A lot of presented and discussed statistical issues were based on questions identified during previous MSA conferences. The scientific programme covered various statistical problems, including multivariate distributions, statistical tests, non-parametric inference, discrimination analysis, Monte Carlo analysis, Bayesian inference, application of statistical methods for finance, economy, insurance, capital market, risk management, design of experiments and survey sampling methodology, mainly for the social sciences purposes. Briefly, the key statistics of the MSA 2012 Conference look as follows: 77 participants altogether (scientists, statisticians, econometricians, demographers, academic tutors, representatives of the National Bank of Poland and local Statistical Offices), 13 sessions, 44 papers and one invited lecture.

The Conference was opened by the Chairman of the Organizing Committee, **Professor Czesław Domański**. On the Opening session the speakers were, respectively:

**Professor Antoni Różalski**, Pro-Rector in Charge of Research, who was the representative of the University of Łódź Rector, **Professor Włodzimierz Nykiel**;

**Professor Paweł Starosta**, the Dean of the Faculty of Economics and Sociology of the University of Łódź;

**Professor Eugeniusz Gatnar**, Member of the Board of the National Bank of Poland;

**Professor Włodzimierz Okrasa**, the representative of the President of Central Statistical Office of Poland, **Professor Janusz Witkowski**.

The first session (SESSION I) was led by **Professor Krzysztof Jajuga** (Wrocław University of Economics) and it was started by **Professor Tadeusz Bednarski** (University of Wrocław), who presented his invited lecture on *The statistical paradigm of Jerzy Splawa-Neyman in causality analysis*. This session was dedicated to three eminent Polish scientists: **Professor Mirosław Krzyśko** (The Higher Vocational State School in Kalisz) recalled the work and profile of Karolina Iwaszkiewicz-Gintowt and **Professor Czesław Domański** presented the profiles of Rajmund Bułowski and Jan Piekalkiewicz.

The Chair of the session function was conferred respectively to:

- SESSION II A** Professor Krystyna Katulska (Adam Mickiewicz University in Poznań);
- SESSION II B** Professor Marek Walesiak (Wrocław University of Economics);
- SESSION III** Professor Bronisław Ceranka (Poznań University of Life Sciences);
- SESSION IV A** Professor Mirosław Krzyśko (the Higher Vocational State School in Kalisz);
- SESSION IV B** Professor Grażyna Trzpiot (Wrocław University of Economics);
- SESSION V** Professor Alina Jędrzejczak (University of Łódź);
- SESSION VI** Professor Eugeniusz Gatnar (The National Bank of Poland, Wrocław University of Economics);
- SESSION VII A** Professor Józef Dziechciarz (Wrocław University of Economics);

**SESSION VII B** Professor Wojciech Zieliński (Warsaw University of Life Sciences);

**SESSION VIII A** Professor Bronisław Ceranka (Poznań University of Life Sciences);

**SESSION VIII B** Professor Grażyna Dehnel (Poznań University of Economics);

**SESSION IX** Professor Grzegorz Kończak (Wrocław University of Economics).

The titles of all further MSA 2012 papers are listed in The Appendix attached below (in the order of the presented papers).

Additionally, on 12 November a meeting of **The Main Board of Polish Statistical Association** took place. The President of the Association, Professor Czesław Domański led the meeting. During the meeting, the Main Board planned the schedule for celebrations regarding The International Year of Statistics 2013 in Poland and discussed celebrations of the **Polish Statistical Association centenary** in 2012.

The MSA 2012 Conference was closed by the Chairman of the Organizing Committee, **Professor Czesław Domański**, who summarized the Conference as very effective and he added that all discussions and doubts should become inspirations and strong motivations for further work of both scientists and practitioners. Professor Domański pointed out that in 2012 Polish and foreign statisticians were celebrating the **100th anniversary of the Polish Statistical Association**, while the next year 2013 will be also an extraordinary one for all statisticians as it is announced to be **The International Year of Statistics**.

The next Multivariate Statistical Analysis Conference **MSA 2013** is planned on **18-20 November 2013** and will take place in **Łódź, Poland**. The Chairman of the Organizing Committee, **Professor Czesław Domański** informs this will be the 32nd edition of the Conference and kindly invites all interested Scientists, Researchers and Students to take part in it.

Prepared by:  
Artur Mikulec  
Aleksandra Kupis-Fijałkowska

## APPENDIX

1. Mirosław Krzyśko (The Higher Vocational State School in Kalisz), *Karolina Iwaszkiewicz-Gintowt (1902-1999). Jerzy Neyman collaborator of Warsaw period.*
2. Czesław Domański (University of Łódź), *Rajmund Buławski.*
3. Czesław Domański (University of Łódź), *Jan Piekalkiewicz.*
4. Grażyna Dehnel (Poznań University of Economics), Tomasz Klimanek (Poznań University of Economics), Jacek Kowalewski (Statistical Office in Poznań), *Indirect estimation accounting for spatial correlation in enterprise statistics.*
5. Tomasz Żądło (University of Economics in Katowice), *On MSE estimation of the misspecified predictor.*
6. Wojciech Gamrot (University of Economics in Katowice), *On some method of strength-borrowing in population total estimation.*
7. Jacek Stelmach (University of Economics in Katowice), *On an estimation of a quantity of base models with parametric and permutation tests.*
8. Daniel Kosiorowski (University of Economics in Cracow), Mateusz Bocian (University of Economics in Cracow), Anna Węgrzynkiewicz (University of Economics in Cracow), Zygmunt Zawadzki (University of Economics in Cracow), *{DepthProc} package in multivariate time series data mining.*
9. Justyna Brzezińska (University of Economics in Katowice), *Odds ratios in the analysis of contingency tables.*
10. Piotr Tarka (Poznań University of Economics), *Measurement scale of the consumers' attitudes in the context of one-parametric Rasch model and dichotomous data.*
11. Katarzyna Dębkowska (Białystok University of Technology), Marta Jarocka (Białystok University of Technology), *The impact of the methods of the data normalization on the result of linear ordering.*
12. Włodzimierz Okrasa (The advisor to the President of Central Statistical Office of Poland, Cardinal Stefan Wyszyński University in Warsaw), *Methodological remarks on interaction of methods and data in social statistics – the case of statistical-sociological research.*
13. Alina Jędrzejczak (University of Łódź), *Income inequality and poverty indicators in Poland by family type.*
14. Tadeusz Gerstenkorn (University of Łódź), Jacek Mańko (XXXI High School in Łódź), *Probability of the fuzzy events and its application in some economic problems.*

15. Dariusz Parys (University of Łódź), *Stepwise multiple tests procedure for discrete distributions.*
16. Justyna Wilk (Wrocław University of Economics), Marcin Pelka (Wrocław University of Economics), *Cluster analysis of symbolic data. A review.*
17. Małgorzata Szerszunowicz (University of Economics in Katowice), *Sequence of experimental trials realisation in design of experiment and production process costs.*
18. Marta Małecka (University of Łódź), *Experimental design in evaluating VaR forecasts*
19. Beata Bieszk-Stolorz (University of Szczecin), Iwona Markowicz (University of Szczecin), *Influence of the explanatory variable on the hazard and its changes in time in the Cox regression model.*
20. Jacek Białek (University of Łódź), *Remarks on general price index formulas.*
21. Eugeniusz Gatnar (The National Bank of Poland, University of Economics in Katowice), *Financial inclusion indicators in Poland (central bank perspective).*
22. Grzegorz Kończak (University of Economics in Katowice), *On the use of the extreme value distribution in monitoring processes.*
23. Wojciech Zieliński (Warsaw University of Life Sciences), *An application of the shortest confidence intervals for fraction in controls provided by Supreme Chamber of Control.*
24. Jerzy Korzeniewski (University of Łódź), *Modification of the HINoV method for selection of variables in multiple cluster structures.*
25. Mariusz Kubus (The Opole University of Technology), *Some remarks on feature ranking based wrappers*
26. Artur Zaborski (Wrocław University of Economics), *Gravity unfolding analysis for asymmetric similarities matrix.*
27. Katarzyna Cheba (The West Pomeranian University of Technology, Szczecin), Maja Kiba-Janiak (Wrocław University of Economics), *Conjoint analysis as a method of analysing consumer preferences on the example of a municipal transport market.*
28. Dominik Krężolek (University of Economics in Katowice), *Non-classical risk measures on the WSE – the application of alpha-stable distributions.*
29. Iwona Konarzewska (University of Łódź), *Optimal stock portfolio – application of multivariate statistical analysis.*

30. Joanna Trzęsiok (University of Economics in Katowice), *On some simulative procedures for comparing nonparametric methods of regression.*
31. Michał Trzęsiok (University of Economics in Katowice), *Extracting class description from Support Vector Machines.*
32. Anna Witaszczyk (University of Łódź), *The application of the random matrix theory in the statistical hypothesis testing concerning parameters of the multivariate normal distribution.*
33. Justyna Wilk (Wrocław University of Economics), Michał Bernard Pietrzak (Nicolaus Copernicus University in Toruń), *Internal migration analysis in the context of socio-economic aspects – two-step approach.*
34. Damian Gajda (University of Gdańsk), Tomasz Jurkiewicz (University of Gdańsk), *Attitudes of SMEs towards insurance – survey results for the period 2010-2012.*
35. Anna Maria Jurek (University of Łódź), *Insurance of medical events characteristic and analysis of existing claims.*
36. Katarzyna Bolonek-Lasoń (University of Łódź), Piotr Kosiński (University of Łódź), *Nash equilibria for unconstrained strategies*
37. Dorota Rozmus (University of Economics in Katowice), *Comparison of accuracy of affinity propagation method and cluster ensembles based on co-occurrence matrix.*
38. Czesław Domański (University of Łódź), Katarzyna Bolonek-Lasoń (University of Łódź), *Generalizations of Tukey-lambda distributions*
39. Małgorzata Misztal (University of Łódź), *Some remarks on using graphical methods in missing data analysis.*
40. Dominika Polko (University of Economics in Katowice), Grzegorz Kończak (University of Economics in Katowice), *On time series prediction based on control chart.*
41. Magdalena Chmieleńska (University of Economics in Katowice), *On influence of inaccuracy measurement on middle – operating control costs*
42. Grażyna Trzpiot (University of Economics in Katowice), *Bayesian spatial quantile regression.*
43. Bronisław Ceranka (Poznań University of Life Sciences), Małgorzata Graczyk (Poznań University of Life Sciences), *Optimal weighing designs for estimation total weight.*
44. Małgorzata Graczyk (Poznań University of Life Sciences), Bronisław Ceranka (Poznań University of Life Sciences), *Regular d-optimal spring balance weighing designs with diagonal variance matrix of errors.*

*STATISTICS IN TRANSITION*-new series, December 2012  
Vol. 13, No. 3, pp. 617—622

## **THE HISTORY OF THE MATHEMATICAL STATISTICS GROUP AT THE HORTICULTURAL FACULTY OF THE CENTRAL COLLEGE OF AGRICULTURE IN WARSAW, AND THE BIOMETRIC LABORATORY AT THE MARCELI NENCKI INSTITUTE OF THE WARSAW SCIENTIFIC SOCIETY**

**Mirosław Krzyśko<sup>1</sup>**

### **ABSTRACT**

The beginning of education of mathematical statistics at the Central College of Agriculture in Warsaw is connected with Jerzy Neyman, who was a lecturer at the Faculty of Horticulture since 1923. The Mathematical Statistics Group was established in 1928, under his leadership. In the same year he organized a Biometric Laboratory at the Marceli Nencki Institute. This paper introduces the history of these two Groups.

The Biometric Laboratory at the Marceli Nencki Institute of the Warsaw Scientific Society was formed in 1928. The Laboratory's property, consisting of the furnishings of two rooms, two Sunstrand electrical adding machines, two Odhner arithmometers and a set of books, was purchased out of a special grant from the National Culture Fund. Shortage of accommodation, a common problem in pre-war Poland, meant that the Biometric Laboratory had to rely constantly on the hospitality of more affluent institutions. In 1928–29 it was based in two rooms at ul. Nowowiejska 43, made available free of charge by the Central Statistical Office. The same rooms accommodated the Mathematical Statistics Group, also formed in 1928, at the Horticultural Faculty of Warsaw Agricultural College. Both Groups were led by Jerzy Neyman. From 1929 the two Groups used two or three rooms in buildings belonging to the Central College of Agricultural, initially at ul. Miodowa 23, and at ul. Rakowiecka 8 from 1935 onwards.

Apart from Neyman, the Mathematical Statistics Group had four employees. A founding member was Waclaw Pytkowski, who was an assistant of Professor Stefan Moszczeński in the Agricultural Economy Group at Warsaw Agricultural College since 1926, and on 30 June 1931 gained an agricultural engineering

---

<sup>1</sup> Adam Mickiewicz University, Faculty of Mathematics and Computer Science.

degree with a thesis entitled *Application of partial correlation to studies of the effect of certain factors on gross income on farms*. Between 1930 and 1931 he did practical work at a farm. In 1932 he returned to academic work with the Mathematical Statistics Group. In 1933 he waived his salary and worked as a volunteer assistant until the outbreak of the Second World War. At the same time he held other important posts: from 1 July 1933 he worked as an agricultural inspector and head of the Economics Department of the Polesie Chamber of Agriculture in Brest; from 10 December 1933 to 1 June 1935 he was deputy director of the Polesie Chamber of Agriculture; and on 26 June 1935 he became the director of the Wołyń Chamber of Agriculture in Łuck (Lutsk). He held that post until the start of the war. From 20 May 1936 to 2 July 1937 he chaired the Regional Markets Supervisory Commission in Łuck. In the course of his work at the Mathematical Statistics Group he published four papers.

On 1 November 1930 the Group was joined by Karolina Iwaszkiewicz, who became a deputy junior assistant. She had previously held the same position in the Meteorology and Climatology Group at the College's Agricultural Faculty from 1 December 1929 to 30 September 1930. It is interesting to consider the range of duties entrusted to Karolina Iwaszkiewicz:

*"The duties of Miss K. Iwaszkiewicz will include being of assistance to the lecturing professor in carrying out academic, teaching and administrative work.*

*In her professional capacity Miss K. Iwaszkiewicz will be obliged:*

- 1/ to fulfil the instructions of the professor to whom she reports,*
- 2/ at work and outside, to maintain the dignity of her post and avoid anything which might damage the integrity of her post and the trust which it requires,*
- 3/ to forward to the professor all requests, representations and complaints in professional matters,*
- 4/ to obtain the professor's permission for any other paid activity,*
- 5/ to comply with the obligations laid down in the Group's rules.*

*In case of a breach by Miss K. Iwaszkiewicz of her accepted obligations, the Academic Senate of the Agricultural College shall be entitled to terminate her employment without notice and without compensation."*

On 30 June 1931 she received an engineer's degree in horticulture with a thesis titled *Application of Poisson's Law to the counting of particles of a virus*, written under the supervision of Jerzy Neyman.

From 1 October 1932 to 30 September 1934 she was a senior assistant, and from 1 October 1934 to 28 February 1938 she worked as a lecturer.

On 5 July 1935 she gained the degree of a doctor of horticultural sciences, with a thesis titled *A generalization of the method of multiple correlation to the case where the eliminated variable is non-measurable*.

On 28 February 1938 she was released from her position at her own request, returned to her native Wilno (Vilnius), and from 1 March 1938 began working at the Central Statistical Office of that city's executive board.



During her work at the Mathematical Statistics Group she published 13 papers, of which five were written jointly with Jerzy Neyman.

The third person employed at the Mathematical Statistics Group was Stanisław Kołodziejczyk, who obtained a master's degree in mathematics on 7 October 1930 at Warsaw University's Faculty of Mathematics and Natural Science, with a thesis titled *Testing of a hypothesis on the constancy of probability based on the principles of probability calculations*.

During his studies, in 1929–1930, he worked as a laboratory technician for the Biometric Laboratory of the Nencki Institute. From August 1930 to September 1931 he did military service. On 1 October 1931 he received an annual scholarship from the National Culture Fund, for special studies in probability and mathematical statistics.

From 1 April 1932 to 30 September 1932 he worked at the Mathematical Statistics Group as a volunteer assistant (without pay). On 1 October 1932 he obtained the post of senior assistant, and from 1 October 1934 until the start of the war he was a lecturer. In 1939 he won a doctoral degree in mathematics from Warsaw University, with a thesis titled *On a certain class of statistical hypotheses relating to the method of least squares*. This degree was conferred on 12 June 1939. He published ten papers, including one written jointly with Jerzy Neyman and Karolina Iwaszkiewicz.

On 1 November 1933 the Group was joined by Waław Kozakiewicz, a pupil of Professor Stefan Mazurkiewicz, who received a master's degree in mathematics on 7 November 1933 at the Faculty of Mathematics and Natural Sciences at Warsaw University. In 1935 he gained a doctorate at the same faculty. In November 1938 he travelled to France on a scholarship from the National Culture Fund, and was still there when the war broke out. In the years 1933–1938 he published six papers in probability theory.

The Group's members gave lectures and classes in Higher Mathematics, Theory of Statistics, Applications of Statistics to Special Areas, and Methodology of the Design of Field Experiments, within the Agricultural College's horticultural and agricultural faculties. Jerzy Neyman also gave seminars for degree students.

Let us now consider the activities of the Biometric Laboratory at the Marcei Nencki Institute.

There were four classes of researchers at the Institute: heads of research groups, senior and junior research assistants, laboratory technicians, and collaborators. The Institute had a permanent staff of just 20 persons, and a larger number of collaborators. The latter were not employed by the Institute – they were employees of other institutions who carried out research in cooperation with groups at the Institute, making use of their resources and equipment. A stay of a year or two, or even more, at the Institute provided conditions to carry out research which would simply not have been possible at universities and colleges.

The permanent members of the Biometric Laboratory were Jerzy Neyman, its head, Karolina Iwaszkiewicz, initially a technician and later a junior assistant, and

the other technicians Janina Hosiasson, Stanisław Kołodziejczyk, I. Staniewska and M. Stępkowska.

Collaborators included Michał Alpern, Maria Iwaszkiewicz (Karolina's sister), Waław Kozakiewicz, W. Lewitska, Tadeusz Matuszewski, Jan Mydlarski, E. Proskurowska, Józef Przyborowski, Waław Pytkowski and Bogumiła Tokarska-Kozakowa. The full Christian names of some of these persons have not been recorded.

The inseparability of the two Groups in terms of personnel and accommodation meant that their activities were also hard to separate. The Groups were created in order to provide academic centres devoted to the applications of mathematical statistics to a range of problems connected with biology in a broad sense. However, work in this direction could not take place without the simultaneous development of a theory of statistics. For this reason the theory of mathematical statistics was always the main subject of the work done by the Groups, with applications playing a somewhat secondary role. The work in statistical theory mainly related to the theory of verification of statistical hypotheses. Research in this area was carried out in constant contact with the Department of Applied Statistics at the University of London, which led to a number of publications produced jointly with E. S. Pearson. The work relating to the applications of mathematical statistics was not limited to biological problems. Because these were the only two groups in Poland specializing in mathematical statistics, which has an extremely wide range of applications, it was a consequence of everyday life and contacts with other institutions that a number of works were produced relating to areas outside biology. The general direction taken by the work of the two Groups can be seen in the following breakdown of papers published in the years 1928–1935. Out of 50 papers published in that period, 22 concern the theory of statistics, whereas the remainder relate to applications in the fields of agriculture (8), microbiology and serology (8), heredity theory (4), social security (4), economics (3), and engineering (1). There were also three books by Jerzy Neyman. A bibliography of these works, with a description of their contents, was included in the report of the Biometric Laboratory for 1928–1935 (see [2]). In total, in the period from 1928 until the start of the war, the two Groups produced around 80 publications.

The Mathematical Statistics Group together with the Biometric Laboratory collected prints of the papers produced by the two Groups, making them into volumes titled *Statistica*, and sent them on an exchange basis to people and institutions with related interests. Six volumes of *Statistica* appeared successively in 1930, 1932, 1933, 1934, 1935 and 1938. These were sent to 142 institutions, publishers and private individuals, including 67 in Poland, 22 in the USA, 11 in France, 6 in the Soviet Union, 3 in Germany, 3 in Switzerland, 2 in India, 2 in Ireland, 2 in Italy, 2 in Sweden, and one each in Bulgaria, Czechoslovakia, the Philippines, the Netherlands, Japan, Norway and Hungary.

The library of the Biometric Laboratory received, by subscription or in exchange for *Statistica*, the following foreign journals: The Annals of Eugenics,

The Annals of Mathematical Statistics, The Bell System Technical Journal, Eugenical News, International Labour Review, Journal of the American Statistical Association, Journal of the Institute of Actuaries, Journal of the Royal Statistical Society, Sankhyā. The Indian Journal of Statistics, Social Science Abstracts, Bulletin de l'Association des Actuaries Suisses, Bulletin de la Société Mathématique de France, Archiv für Rassen- und Gesellschafts- Biologie, Handbuch der Vererbungswissenschaft, Zeitschrift für Angewandte Mathematik und Mechanik, Zeitschrift für Induktive Abstammungs- und Vererbungslehre, and Giornale dell' Istituto Italiano degli Attuari. By the end of 1935 the library of the Biometric Laboratory held a total of 1048 volumes.

Due to his continued lack of a permanent post at any university or college in Poland, in 1934 Jerzy Neyman took up the position of lecturer at University College, London. From then on he divided his time between London and Warsaw. The circumstances of his taking up employment in London are described by Klonecki and Zonn (see [3]):

*“His social and material situation remained nonetheless difficult, chiefly due to Neyman’s engagement in the political life of contemporary Poland. He took part in meetings of the anticlerical Polish Association of Free Thought. At meetings of that association Neyman spoke on the subject of the expenditure incurred by the Polish state on maintaining the Catholic Church within the country. Based on Neyman’s report, someone else printed an article in the journal Racjonalista under the title “The Telling Numbers”. It should be noted that Neyman had taken all of his numerical data from publicly available official documents concerning the national budget. However, these activities made it difficult for Neyman to gain further promotion, and so when he received an invitation to London in 1934, he left Poland. Thus began Neyman’s life as an emigrant.”*

Starting from the 1934–35 academic year, Neyman’s classes were taken over by Karolina Iwaszkiewicz and Stanisław Kołodziejczyk. Over time Neyman’s links with Poland became weaker, and in 1936 he left Poland permanently. In the 1936–37 academic year the position of head of the Mathematical Statistics Group at the Horticultural Faculty of Warsaw Agricultural College was given to Dr Karolina Iwaszkiewicz, and later its supervisor became Professor Michał Korczewski, the dean of the faculty. At the end of 1937 the activity of the Biometric Laboratory was suspended, and all of its members left the Institute.

Over the nine years from 1928 to 1936, that is from his arrival from a secondment in Paris until he left for England, Jerzy Neyman worked intensively on the foundations of statistics. This was one of the most creative periods in his academic career, and the results which he obtained in those years in Poland (he did a lot of his research work at that time at Małdralin near Otwock) set the course for the further development of statistics over the next decades. With his pupils he presented numerous works, both theoretical and relating to applications of statistical methods in agriculture, microbiology, anthropology, genetics, economics and sociology.

## REFERENCES

- Archives of Warsaw University of Life Sciences-SGGW (1936). *Instytut imienia Nenckiego Towarzystwa Naukowego Warszawskiego. 1928-1935. Organizacja-Działalność-Środki*, published by the Institute, Warsaw.
- KLONECKI, W. and ZONN, W. (1973). *Jerzy Sława-Neyman*, Roczniki Polskiego Towarzystwa Matematycznego. Series II: Wiadomości Matematyczne XVI, 55–70.
- KUŹNICKI, L. (2008) *Instytut Biologii Doświadczalnej im. Marcelego Nenckiego. Historia i Teraźniejszość, Tom I, 1918-2007*, PAN M. Nencki Institute of Experimental Biology, Warsaw.
- LAUDAŃSKI, Z. and MĄDRY, W., *Historia Katedry Doświadczalnictwa i Bioinformatyki*,  
<http://agrobiol.sggw.waw.pl/biometria/pages/strona-glowna/bhistoriab.php>