STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 1—2

FROM THE EDITOR

Welcome to our new issue of *Statistics in Transition new series.* It contains a set of seventeen items grouped according to their contents into five sections. Its major part (first) is composed of six articles gathered under heading *Sampling methods and estimation* that traditionally constitutes one of the key points of interest of the journal. This time, its topics range from innovative approaches to estimating basic population parameters to consumer weights for small subgroups of the reference population by R. B. Papalia (concerning estimating item weights of consumer price indexes for small subgroups of the reference populations). Means estimation problems are discussed by Shukla and Thakur (with use of imputation of missing data); by Singh and Agnihotri *(*using auxiliary information in sample surveys)*;* and by Singh, Chauhan and Sawan (an exponential estimator for estimating the finite population mean). Variance estimating procedures are in focus of papers by Dubey and Sharma *(*using auxiliary information)*;* by Singh and Chandra (an alternative to ratio estimator of the population variance); while W. Gamrot discusses an asymptotic properties of some standard deviation estimates.

Next section, *Multivariate Analyses*, consists of three articles each of them treating on different types of problems: Begum, Dwivedi and Pandey compare case *versus* cohort medical data sets in analysis aimed at tooling a more efficient birth control instrument to detect pregnancy (especially, unwanted pregnancies). A. Pollastri compares power of a new stepwise test with Duncan's test of the hypotheses on the means of a Bivariate Correlated Normal (B.C.N.). And Singh and Chander discuss a class of shrunken estimators for estimating the α-th power of the standard deviation of a normal distribution, using prior information. With M. Kozak's remarks on educational implications of different conceptualizations of correlation and regression (in section three), the substantive sections are concluded and followed by items relating to certain affairs taking place in some public statistical systems.

In order to make a subject matter (a problem or an occurrence) that gains increasing importance in a country's data system, or internationally, a more noticeable and appealing to community of producers and users of official statistics, as well as to other parties concerned about the overall quality of public data, including policy makers and practitioners, we are launching a special forum section devoted to *Current Issues in Public Statistics.* It is commenced by addressing in the journal's present issue one of the most quintessential and vital for public statistics question of protecting individual entities' data file against disclosure for non-statistical purposes (*Confidentiality and Data Access).*

Having a real reason to address this problem — as outlined in the letter from the Polish Statistical Association (opening this section) — we hope to draw on a vast work that has been done so far in many countries and under auspice of international organizations in our efforts to bring to public awareness the complex nature of the problem. And to assist in a way the national statistical agencies that are in need to develop effective measures against the increased risk from a third party's attempts - especially, a public authority in newly emerged democracies (successors of the formerly socialist economies) — to access such microdata for administrative action. We plan to devote a special issue of the journal to this problem — see our call for papers (in this volume).

I take this opportunity to thank all our collaborators for their continued, various types of contributions to the fulfillment of our journal's mission.

WŁODZIMIERZ OKRASA
Editor-in-Chief

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 3—4

CALL FOR PAPERS: *CONFIDENTIALITY AND DATA ACCESS — THEORY AND EVIDENCE FROM A GLOBAL PERSPECTIVE. SPECIAL ISSUES OF STATISTICS IN TRANSITION NEW SERIES*

Deadline for Manuscritps: November 30, 2008

The Editor of the Polish Statistical Association and the Central Statistical Office's international journal ***Statistics in Transition new series*** invites contributions to a Special Issue of the journal entitled "Confidentiality and Data Access — Theory and Evidence from a Global Perspective".

Papers submitted should be located in relation to current problems in the area of statistical confidentiality and data access, including the relevant principles and practices as elaborated in *theory* and *evidence-based* knowledge developed in different contexts, in an international perspective. Any of the four major aspects of the problem (or any configuration of them) - legislation, administrative policies, statistical disclosure limitations, and "ethical issues" [1] — are an object of interest. We hope to include papers from a range of viewpoints, grounded in studies and experiences from different country's data systems worldwide.

One of the key aims of this special issue, and a suggested point of interest, is lessons to be learned from research findings and applications in different regions for dealing with the distinctive challenges being faced in this respect by public statistics in newly emerged democracies of Central and Eastern Europe.

It means that in addition to such urging problems that become common to all national statistical agencies, like increased risks of breaching data confidentiality by growing demand for and technological capabilities of accessing microdata — including constantly excelled procedures employed in inter-agency data sharing, multi-source data linkage, spatial distribution of population data, mining census data and other official statistics, etc., — there are region-specific problems too. Of particular interest would be hazardous consequences that apparently allowable claims from some countries' administrative authorities, to have access to individual entities information for administrative purposes, might have for the data integrity and quality.

[1] As specified in *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics* (Committee on National Statistics, NRC and the Social Science Research Council; National Academy Press, Washington D.C.,1993.)

Please submit your manuscript at sit@stat.gov.pl by November 30, 2008, with a notation that you would like to have it included into this Special Issue.

Please find specific submission instructions on the SiT Guidelines Web site at: http://www.stat.gov.pl/gus/45_2638_ENG_HTML.htm

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 5—6

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 the respective 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. Wlodzimierz 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/gus/45_2638_ENG_HTML.htm

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 7—18

ON ESTIMATING POPULATION VARIANCE USING AUXILIARY INFORMATION

V. Dubey and H. K. Sharma

ABSTRACT

The paper deals with a difference type estimator of population variance using auxiliary information. It is found that proposed estimator is more efficient than various estimators under stringent conditions. After estimating the constants involved in the a estimator, modified regression estimator has been suggested. A numerical illustration has also been made.

Key words: Auxiliary Variable, Bias, Mean Square Error (MSE), Relative Efficiency, Simple Random Sampling, Coefficient of Kurtosis.

1. Introduction

The problem of estimating population variance arises in many practical situations like genetical, biological and medical studies. [Bland and Altman (1986)]. The problem has been well dealt in literature in simple random sampling. Wakimoto (1971) considered this problem in stratified sampling while Tripathi (1977), Liu (1974), Chaudhury (1978), Mukhopadhyay (1978), Swain and Mishra (1994) have paid their attention in PPS and general sampling designs. Padmawar and Mukhopadyay (1981). Mukhopadhyay (1982) made a significant contribution for estimating population variance under super-population models. Chaudhury and Adhikari (1990) made an attempt for its estimation, using randomized response technique while Dutta and Ghosh (1993) extended their view under Bayesian approach. Taking advantage of high correlation between study and auxiliary variables, Isaki (1983) proposed ratio and regression type estimators of population variance. Birader and Singh (1998), Agrawal and Panda (1999) explored their discussion under prediction approach.

Assume that the finite population consists of N identifiable units ($U_1, U_2, U_3, \ldots, U_N$) taking the values ($Y_1, Y_2, Y_3, \ldots, Y_N$) on study variable y. Let $\overline{Y} = N^{-1}\sum_{i=1}^{N} Y_i$ and $\sigma_y^2 = N^{-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$ be population mean and variance

of y. Let $\bar{y} = n^{-1} \sum_{j=1}^{n} y_j$ and $s_y^2 = (n-1)^{-1} \sum_{j=1}^{n} (y_j - \bar{y})^2$ be sample mean and variance of y based on a sample s = (1,2,…j,..,n) taken from U by simple random sampling. Again, let K_i ; i=1,2,..,6 be suitable constants. Singh et al (1973), Searls and Intarapanich (1990) proposed estimator of σ_y^2

$$s_{y1}^2 = K_1 \; s_y^2 \tag{1.1}$$

The optimum value of K_1 depends upon the coefficient of kurtosis of y. The estimator s_{y1}^2 is found to be more efficient than s_y^2, if sample size n is small.

Let x be an auxiliary variable highly correlated with y; $\bar{x}$, s_x^2, $\bar{X}$ and σ_x^2 be defined similarly as $\bar{y}$, s_y^2, $\bar{Y}$ and σ_y^2 respectively. Das and Tripathi (1978) proposed estimators of σ_y^2 if $\bar{X}$ or σ_x^2 or $C_x = (\sigma_x / \bar{X})$ is known. Further, Srivastava and Jhajj (1980) proposed a class of estimators of σ_y^2 under the knowledge of $\bar{X}$ and σ_x^2 and studied its properties under certain regularity conditions. This estimator is more efficient than usual estimators for skewed populations but for symmetrical populations, it is equally efficient as difference type estimator

$$s_{yd}^2 = s_y^2 + K_2 \, (\sigma_x^2 - s_x^2) \tag{1.2}$$

Replacing s_y^2 by ratio type estimator $s_{yr}^2 = (s_y^2 s_x^{-2}) \sigma_x^2$ in (1.1), Prasad and Singh (1990) proposed modified ratio estimator of σ_y^2 which is almost equally efficient as s_{yr}^2 for large samples. The quest for improvement over s_{yd}^2 lead Singh et al (1988) to consider a marginally more efficient estimator

$$s_{yd_1}^2 = K_3 s_y^2 + K_4 \, (\sigma_x^2 - s_x^2) \tag{1.3}$$

Noting that sum of coefficients of s_y^2, s_x^2 and σ_x^2 in s_{yd}^2 is unity, Dubey and Kant (2001) suggested estimator of σ_y^2

$$s^2_{yd_2} = K_5 s^2_y + K_6 s^2_x + (1 - K_5 - K_6)\sigma^2_x \qquad (1.4)$$

The estimator s^2_{yd2} is more efficient than all the above estimators if σ^2_y is near to σ^2_x. This is possible if past data is used as auxiliary variable where lag period is not so long.

In section 2, a generalized estimator has been proposed which is more efficient than all the above estimators under very obvious conditions.

2. Proposed estimator

Let $\hat{\sigma}^2_y$ and $\hat{\sigma}^2_x$ be the unbiased estimator of σ^2_y and σ^2_x under any sampling design. We propose the estimator of σ^2_y as

$$\hat{\sigma}^2_{yg} = \lambda_1 \hat{\sigma}^2_y + \lambda_2 (\sigma^2_x - \hat{\sigma}^2_x) + (1 - \lambda_1)\lambda_3 \sigma^2_x \qquad (2.1)$$

where λ_i; i=1,2,3 are suitable constants. The estimator $\hat{\sigma}^2_{yg}$ includes following estimators as its particular cases

$$\hat{\sigma}^2_{yg1} = \hat{\sigma}^2_y + \lambda_2 (\sigma^2_x - \hat{\sigma}^2_x) \qquad (2.2)$$

$$\hat{\sigma}^2_{yg2} = \lambda_1 \hat{\sigma}^2_y + \lambda_2 (\sigma^2_x - \hat{\sigma}^2_x) \qquad (2.3)$$

$$\hat{\sigma}^2_{yg3} = \lambda_1 \hat{\sigma}^2_y + \lambda_2 (\sigma^2_x - \hat{\sigma}^2_x) + (1 - \lambda_1)\sigma^2_x \qquad (2.4)$$

The proposed estimator $\hat{\sigma}^2_{yg}$ has bias

$$B(\hat{\sigma}^2_{yg}) = (\lambda_1 - 1)(\sigma^2_y - \lambda_3 \sigma^2_x) \qquad (2.5)$$

and mean square error

$$M(\hat{\sigma}^2_{yg}) = (1-\lambda_1)^2(\sigma^2_y - \lambda_3\sigma^2_x)^2 + \lambda_1^2 V(\hat{\sigma}^2_y) + \lambda_2^2 V(\hat{\sigma}^2_x) + \lambda_1\lambda_2 Cov(\hat{\sigma}^2_y, \hat{\sigma}^2_x) \qquad (2.6)$$

Let $\eta_g = \dfrac{Cov(\hat{\sigma}^2_y, \hat{\sigma}^2_x)}{\sqrt{V(\hat{\sigma}^2_y)V(\hat{\sigma}^2_x)}}$, $B_g = \dfrac{Cov(\hat{\sigma}^2_y, \hat{\sigma}^2_x)}{V(\hat{\sigma}^2_x)}$, $\phi = \dfrac{\sigma^2_x}{\sigma^2_y}$, $C^2(\hat{\sigma}^2_y) = \dfrac{V(\hat{\sigma}^2_y)}{\sigma^4_y}$,

$$Q_{1g}=\frac{C^2(\hat{\sigma}_y^2)(1-\eta_g^2)}{(1-\lambda_3\phi)^2}, Q_{2g}=C^2(\hat{\sigma}_y^2)(1-\eta_g^2),\ Q_{3g}=\frac{C^2(\hat{\sigma}_y^2)(1-\eta_g^2)}{(1-\phi)^2}.$$

The best values of λ_1 and λ_2 for which $M(\hat{\sigma}_{yg}^2)$ is minimized, are given by

$$\lambda_{01}=\frac{1}{1+Q_{1g}} \tag{2.7}$$

$$\lambda_{02}=\lambda_{01}\,B_g \tag{2.8}$$

For such a choice of λ_i, i =1,2, the minimum MSE of $\hat{\sigma}_{yg}^2$ is given by

$$M_0(\hat{\sigma}_{yg}^2)=\frac{V(\hat{\sigma}_y^2)(1-\eta_g^2)}{1+Q_{1g}} \tag{2.9}$$

We find that $M_0(\hat{\sigma}_{yg}^2)\to 0$ if λ_3 is chosen such that $\phi\lambda_3\to 1$. In this case $\lambda_{01}\to 0$ (and hence $\lambda_{02}\to 0$).

The difference type estimator $\hat{\sigma}_{yg1}^2$ has minimum variance

$$V_0(\hat{\sigma}_{yg1}^2)=V(\hat{\sigma}_y^2)(1-\eta_g^2) \tag{2.10}$$

while $\hat{\sigma}_{yg2}^2$ and $\hat{\sigma}_{yg2}^2$ have minimum MSE

$$M_0(\hat{\sigma}_{yg2}^2)=\frac{V(\hat{\sigma}_y^2)(1-\eta_g^2)}{1+Q_{2g}} \tag{2.11}$$

$$M_0(\hat{\sigma}_{yg3}^2)=\frac{V(\hat{\sigma}_y^2)(1-\eta_g^2)}{1+Q_{3g}} \tag{2.12}$$

From (2.9) and (2.10), we find that $M_0(\hat{\sigma}_{yg}^2)$ is always smaller than $V_0(\hat{\sigma}_{yg1}^2)$. Further

$$M_0(\hat{\sigma}_{yg}^2)<M_0(\hat{\sigma}_{yg2}^2), \qquad \text{if } 0<\lambda_3<2\,\phi^{-1} \tag{2.13}$$

$$M_0(\hat{\sigma}^2_{yg}) < M_0(\hat{\sigma}^2_{yg3}), \qquad \text{if } 0 < \lambda_3 < \{2(\phi^{-1}) - 1\} \tag{2.14}$$

Pandey and Singh (1977) used shrinkage type estimator of σ^2_y using its guessed value σ^2_0. Let $\sigma^2_{(1)y}$ be minimum value of σ^2_y, which may be guessed from past data or repeated surveys. Hence, let $\phi_{(1)} = \dfrac{\sigma^2_x}{\sigma^2_{(1)y}}$ be minimum value of ϕ. Therefore, if λ_3 be taken such that

$$0 < \lambda_3 < 2\,\phi^{-1}_{(1)} \tag{2.15}$$

the proposed estimator $\hat{\sigma}^2_{yg}$ would record its efficiency over $\hat{\sigma}^2_{yg2}$. Further, the value of λ_3 between 1 and $(2\phi^{-1}_{(1)} - 1)$, would enable the proposed estimator to be more efficient than $\hat{\sigma}^2_{yg3}$.

3. Choice of λ_{01} and λ_{02}

The optimum value λ_{01} could also be expressed as

$$\lambda_{01} = 1 - \frac{V(\hat{\sigma}^2_{yg_{10}})}{E(\hat{\sigma}^2_{yg_{10}} - \lambda_3\,\sigma^2_x)^2} \tag{3.1}$$

where

$$\hat{\sigma}^2_{yg_{10}} = \hat{\sigma}^2_y + B_g(\sigma^2_x - \hat{\sigma}^2_x) \tag{3.2}$$

Let $V(\hat{\sigma}^2_{yg_{10}}) = \dfrac{V_g}{n}$; $\hat{V}_g$ and $\hat{B}_g$ be consistent estimates of V_g and B_g; and

$$\hat{\sigma}^2_{ygr_1} = \hat{\sigma}^2_y + \hat{B}_g(\sigma^2_x - \hat{\sigma}^2_x) \tag{3.3}$$

be regression type estimator of σ^2_y. We estimate of λ_{01} and, hence, λ_{02} as

$$\hat{\lambda}_{01}=1-\frac{\hat{V}_g}{n(\hat{\sigma}^2_{ygr_1}-\lambda_3\,\sigma_x^2)^2} \tag{3.4}$$

$$\hat{\lambda}_{02}=-\hat{B}_g\,\hat{\lambda}_{01} \tag{3.5}$$

respectively. Therefore, estimate of σ_y^2 is obtained as

$$\hat{\sigma}^2_{ygr_2}=\hat{\lambda}_{01}\hat{\sigma}^2_{ygr_1}+(1-\hat{\lambda}_{01})\lambda_3\sigma_x^2 \tag{3.6}$$

Let

$$(\hat{\sigma}^2_{ygr1}-\lambda_3\sigma_x^2)^2=E(\hat{\sigma}^2_{yrg1}-\lambda_3\sigma_x^2)^2+u_1 \tag{3.7}$$

$$\hat{V}_g=V_g+u_2 \tag{3.8}$$

where $E(u_1)=0$ and u_2 is of order (n^{-s}), $s>0$. We find that

$$M(\hat{\sigma}^2_{ygr_2})=M_0(\hat{\sigma}^2_{yg}) \tag{3.9}$$

up to the first order of approximation.

4. Special case: SRSWR

Whenever units are taken by simple random sampling with replacement procedure, we have $\hat{\sigma}_y^2=s_y^2$, $\hat{\sigma}_x^2=s_x^2$. In this case estimators in (2.2), (2.3) and (2.4) are similarly defined as s^2_{yd}, $s^2_{yd_1}$ and $s^2_{yd_2}$ respectively. For expressing MSE of estimators, we use following notations -

$$\mu_{rs}=E(y-\overline{Y})^r(x-\overline{X})^s,\ \beta_2(x)=\frac{\mu_{04}(y,x)}{\sigma_x^4},\ \beta_2(y)=\frac{\mu_{40}(y,x)}{\sigma_y^4},$$

$$l(y,x)=\frac{\mu_{22}(y,x)}{\sigma_y^4\,\sigma_x^4},\ \beta_2^*(y)=\beta_2(y)-1,\ \beta_2^*(x)=\beta_2(x)-1,$$

$$l^*(y,x)=l(y,x)-1,\ \ \eta_s=\frac{l^*(y,x)}{\sqrt{\beta_2^*(y)\beta_2^*(x)}},$$

$$Q_{1s}=\frac{\beta_2^*(y)(1-\eta_s^2)}{(1-\delta\phi)^2},\ Q_{2s}=\beta_2^*(y)(1-\eta_s^2),\ Q_{3s}=\frac{\beta_2^*(y)(1-\eta_s^2)}{(1-\phi)^2}.$$

Thus minimum MSEs obtained in (2.9), (2.10), (2.11) and (2.12) correspondingly reduce to

$$M_0(\hat{\sigma}^2_{yg})_{rs}=\frac{\beta_2^*(y)(1-\eta_s^2)\sigma_y^4}{(n-1)+Q_{1s}} \tag{4.1}$$

$$M_0(s^2_{yd})=\frac{\beta_2^*(y)(1-\eta_s^2)\sigma_y^4}{(n-1)} \tag{4.2}$$

$$M_0(s^2_{yd1})=\frac{\beta_2^*(y)(1-\eta_s^2\}\sigma_y^4}{(n-1)+Q_{2s}} \tag{4.3}$$

$$M_0(s^2_{yd2})=\frac{\beta_2^*(y)(1-\eta_s^2)\sigma_y^4}{(n-1)+Q_{3s}} \tag{4.4}$$

For getting estimates of λ_{01} and λ_{02}, we consider

$$\hat{B}=\frac{\sum_{j=1}^{n}[\{(y_j-\bar{y})^2-s_y^2\}\{(x_j-\bar{x})^2-s_x^2\}]}{\sum_{j=1}^{n}\{(x_j-\bar{x})^2-s_x^2\}^2}$$

and $\hat{V}(y,x)=\frac{1}{n}\sum_{j=1}^{n}[(y_j-\bar{y})^2-s_y^2-\hat{B}\{(x_j-\bar{x})^2-s_x^2\}]^2$

as consistent estimates of B and $\beta_2^*(y)(1-\eta_s^2)\sigma_y^4$ respectively. Let

$$s^2_{ylr}=s_y^2+\hat{B}(\sigma_x^2-s_x^2) \tag{4.5}$$

be regression-type estimator of σ_y^2 with MSE equal to $M_0(s_{yd}^2)$ up to the first order of approximation. The estimates of λ_{01} and λ_{02} are given by

$$\hat{\lambda}_{10R} = 1 - \frac{\hat{V}(y,x)}{(n-1)(s_{ylr}^2 - \lambda_3 \sigma_x^2)^2} \tag{4.6}$$

$$\hat{\lambda}_{02R} = -\hat{B}\hat{\lambda}_{01R} \tag{4.7}$$

For the case of bi-variate normal populations, where $\beta_2(y) = 3 = \beta_2(x)$ and $l(y,x) = 1 + 2\rho^2$, expressions (4.1) to (4.4) reduce to

$$M_0(\hat{\sigma}_{yg}^2)^* = \frac{2(1-\rho^4)\sigma_y^4}{(n-1) + 2(1-\rho^4)(1-\lambda_3\phi)^{-2}} \tag{4.11}$$

$$M_0(s_{yd}^2)^* = \frac{2(1-\rho^4)\sigma_y^4}{(n-1)} \tag{4.12}$$

$$M_0(s_{yd_1}^2)^* = \frac{2(1-\rho^4\}\sigma_y^4}{(n-1) + 2(1-\rho^4\}} \tag{4.13}$$

$$M_0(s_{yd2}^2)^* = \frac{2(1-\rho^4)\sigma_y^4}{(n-1) + 2\,(1-\rho^4)(1-\phi)^{-2}} \tag{4.14}$$

From (4.12) and (4.13), we find that s_{yd}^2 and s_{yd1}^2 are approximately equally efficient while $\hat{\sigma}_{yg}^2$ will be better than all under conditions discussed in section 2.

Further, in this situation, the value of B reduces to $B^* = \beta^2$, β is regression coefficient of y on x. Let r= sample correlation coefficient between y and x, $b = (r s_y / s_x)$ be sample estimate of β. An estimate of B^* is given by $\hat{B}^* = b^2$. Thus (3.3) reduces to

$$s^2_{ylr_1} = s^2_y + b^2(\sigma^2_x - s^2_x) \tag{4.15}$$

with MSE same as given in (4.12) [See Isaki (1983)]. Similarly, estimating $\sigma^4_y(1-\rho^4)$ by $\hat{V}^*(y,v) = s^4_y(1-r^4)$, we have estimate of λ_{01} as

$$\hat{\lambda}^*_{10R} = 1 - \frac{2\hat{V}^*(y,x)}{(n-1)(s^2_{ylr1} - \lambda_3\,\sigma^2_x)^2} \tag{4.16}$$

Therefore for bivariate normal populations, an estimator of σ^2_y is given by

$$s^2_{ylr_2} = \hat{\lambda}^*_{01R}\, s^2_{ylr_1} + (1-\hat{\lambda}^*_{01})\lambda_3\,\sigma^2_x \tag{4.17}$$

which has MSE (4.11).

5. Empirical study

Source: Tripathi et al (2002). The data consists of 278 villages or towns/ward under Gijole police station of Malda district of West Bengal, India where x and y are number of agricultural laborers in 1961 and 1971. For this data

N = 278, S_y =56.46, S_x =40.67, ρ = 0.72,

$\beta_2(y)$ = 40.8536, $\beta_2(x)$ = 38.89, $\beta_1(x)$ = 25.90, $l(y,x)$ =26.81,

The relative efficiencies (R.E.) of different estimators of S^2_y with respect to conventional estimator s^2_y, defined by $[V(s^2_y)/M(\hat{\sigma}^2_{yg})]$ are displayed in the following table:

Table 1.

δ	n = 30	n =50	n=100
0.0	4.26	3.91	3.65
0.5	4.97	4.33	3.86
1.0	7.11	5.59	4.49
1.5	20.87	13.74	8.52
2.0	605.59	359.80	79.80
2.5	13.12	9.15	6.25
3.0	6.17	5.04	4.21
4.0	4.14	3.84	3.62
6.0	3.59	3.52	3.46
8.0	3.48	3.45	3.43
10	3.45	3.43	3.42

R. E. of s^2_{yd} with respect to s^2_y is 3.40 for all sample sizes. The R. E. of s^2_{y1} with respect to s^2_y for sample sizes 30, 50, 100 are respectively 1.86, 1.51 and 1.25. R. e. of $\hat{\sigma}^2_{yg_1}$ and $\hat{\sigma}^2_{yg2}$ are given in Table 1 for δ = 0 and δ =1 respectively.

From Table 1 it is clear that $\hat{\sigma}^2_{yg}$ is more efficient than $s^2_{y_1}$. It has superiority over $\hat{\sigma}^2_{yg_1}$ if $1 < \delta < 4$. For $1 < \delta < 2.5$, it is more efficient than all existing estimators.

REFERENCES

AGRAWAL, M. C. and PANDA, K. B. (1999): A predictive justification for variance estimation using auxiliary information. Jour. Ind. Soc. Ag. Stat., 52(2), 192—200.

BIRADAR, R. S. and SINGH, H. P. (1998): Predictive estimation of finite population variance. Cal. Statist. Assoc. Bull., 48, 229—235.

BLAND, J. M. and ALTMAN, D. G. (1986): Statistical method for assessing agreement between two methods of clinical measurement., Lance, 1(8476), 307—310.

CHAUDHURY, A. (1978): On estimating the variance of a finite population. Metrika, 25, 66—67.

CHAUDHURY, A. and ADHIKARI, A. K. (1990): Variance estimation with randomized response., Commun. Statist - Theory Meth., 19(3), 1119—1125.

DAS, A. K. and TRIPATHI, T. P. (1977): Admissible estimators for quadratic forms in finite populations., Bull. Inter. Stat. Inst., 47(4), 132—135.

DAS, A. K. and TRIPATHI, T. P. (1978): Use of auxiliary information in estimating the finite population variance, Sankhya, 40C, 139—148.

DUBEY, V. and KANT, S. (2001): A weighted estimator of population variance using auxiliary information, Abstract, International conference on Statistical Inference and Reliability to honour Prof J. V. Despande, XXI Annual Conference of ISPS and Annual Conference of Indian Chapter of Indian Society of Bayesian Analysis, Dec 21—24, Chandigarh University.

DUTTA, G. S. and GHOSH, MALAY (1993)**:** Bayesian estimation of a finite population variance with auxiliary information., Sankhya, 55(2), Ser B, 156—170.

ISAKI, C. T. (1983): Variance estimation using auxiliary information, Jour. Amer. Stat. Assoc., 78, 117—123.

LIU, T. P (1974): A generalized unbiased estimator for the variance of a finite population, Sankhya, 36, C, 23—32.

MUKHOPADHYAY, P. (1978): Estimating a finite population variance under a super population model., Metrika, 25, 115—122.

MUKHOPADHYAY, P. (1982): Optimum Strategies for estimating the variance of a finite population under a super population model., Metrika, 29, 143—158.

PADMWAR, V. R. and MUKHOPADHYAY, P. (1981): Estimation of symmetric functions of a finite population., Metrika, 31, 89—97.

PANDEY, B. N. and SINGH, J. (1977): Estimation of variance of normal population using prior information., Jour. Ind. Statist. Assoc., 15, 141—150.

PRASAD, B., and SINGH, H. P. (1990): Some improved ratio-type estimator of finite population variance in sample surveys. Commun. Statist.-Theory Math, 19(3), 1127—1139.

SEARLS, D. T. and INTARAPANICH, P. (1990): A note on an estimator for the variance that utilizes the kurtosis. Amer. Statistician, 44(4), 295—296.

SINGH, H. P., UPADHYAY, L. N. and NAMJOSH, U. D. (1988): Estimation of finite population variance., Current Science, 57, 24, 1331—1334.

SINGH, J. PANDEY, B. N. and HIRANO, K. (1973): On the utilization of a known coefficient of kurtosis in the estimation procedure of variance, Ann. Inst. Stat. Math., 25, 51—55.

SRIVASTAVA, S. K. and JHAJJ, H. S. (1980): A class of estimators using auxiliary information for estimating finite population variance, Sankhya, 42, C, 87—96.

SWAIN, A. K. P. C. and MISHRA, G. (1994): Estimation of population variance under unequal probability sampling., Sankhya, Ser B, 56, 374—384.

TRIPATHI, T. P., SINGH, H. P. and UPADHYAYA, L. N. (2002): A general method of estimation and its application to the estimation of coefficient of variation, Statistics in Transition, 5(6), 1081—1102.

WAKIMOTO, K. (1971): Stratified random sampling (I): Estimation of Population variance., Ann. Inst. Stat. Math., 23, 233—252.

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 19—31

ON ASYMPTOTIC PROPERTIES OF STANDARD DEVIATION ESTIMATES UNDER DOUBLE SAMPLING FOR NONRESPONSE

Wojciech Gamrot

ABSTRACT

The phenomenon of nonresponse in a sample survey reduces the precision of parameter estimates and introduces the bias. Several procedures have been developed to reduce this bias. A well-known technique is the two-phase (or double) sampling scheme which relies on subsampling the nonrespondents and re-approaching them in order to obtain the missing data. In this paper two estimators of the finite population standard deviation are proposed for the general two-phase sampling procedure involving arbitrary sampling designs in both phases. Their asymptotic properties are derived. A simulation study is also provided.

Key words: nonresponse, double sampling, two-phase sampling, finite population standard deviation

1. Introduction

Consider a finite population U containing N population elements (units), and some population characteristic X taking fixed values $x_1,...,x_N$. Several population parameters may be defined, including the population total of X:

$$t_x = \sum_{i \in U} x_i \tag{1}$$

population mean of X:

$$\overline{X} = \frac{t_x}{N} \tag{2}$$

population variance of X:

$$S_U^2(X) = \frac{1}{N-1}\sum_{i\in U}(x_i - \overline{X})^2 \tag{3}$$

and population standard deviation of X:

$$S_U(X) = \sqrt{S_U^2(X)} \tag{4}$$

Our interest in this paper is in the estimation of the population standard deviation as defined above. The study will be based on known results concerning estimation of population totals.

Consider the following two-phase sampling procedure, dedicated to nonresponse. In the first phase of the survey a random sample s of size n is drawn from U, according to some sampling design $p(s)$, characterized by inclusion probabilities of the first and second order respectively denoted by:

$$\pi_i = \sum_{s\ni i} p(s) \tag{5}$$

and:

$$\pi_{ij} = \sum_{s\ni i,j} p(s) \tag{6}$$

for $i\neq j\in U$. We assume that n may vary from sample to sample. When nonresponse appears in the survey some units fail to provide responses. Denote the sample subset of responding units by s_1 and the sample subset of nonresponding units by s_2, so that $s = s_1 \cup s_2$ and $s_1 \cap s_2 = \varnothing$. Following Cassel et al. (1983) we assume that the nonresponse mechanism may be described in stochastic terms, or in other words that there exists some probability distribution $q(s_1 \mid s)$ such that $q(s_1 \mid s) \geq 0$ and $\sum_{s_1\subseteq s} q(s_1 \mid s) = 1$. Särndal et al. (1992) call $q(s_1 \mid s)$ the *response distribution*. In general, it is defined conditionally with respect to s in order to reflect the interactions between sampled individuals in the survey. The response distribution determines individual probability of i-th unit responding in the survey:

$$\rho_{i\mid s} = \sum_{s_1\ni i} q(s_1 \mid s) \tag{7}$$

and joint probability of any i-th and j-th unit responding in the survey:

$$\rho_{ij\mid s} = \sum_{s_1\ni i,j} q(s_1 \mid s) \tag{8}$$

The behaviour of the random set s_2 is also determined by the distribution $q(s_1 \mid s)$ because $s_2=s-s_1$ and s is constant with respect to the response distribution. Consequently, we have:

$$q(s_1 \mid s) = q(s_2 \mid s) = q(s_1, s_2 \mid s) \tag{9}$$

To obtain some information about non-responding units a second phase of the survey is implemented. A subsample s' of size n' is drawn from the nonrespondent subset s_2 according to another sampling design $p'(s'|s,s_2)$ which is characterized by the set of inclusion probabilities of the first order:

$$\pi_{i|s,s_2} = \sum_{s \ni i} p'(s' \mid s, s_2) \tag{10}$$

and second order:

$$\pi_{ij|s,s_2} = \sum_{s \ni i,j} p'(s' \mid s, s_2) \tag{11}$$

It is often assumed (see. Lessler and Kalsbeek 1992) that appropriate efforts are undertaken in the second phase of the survey (e.g. personal interview instead of mail or telephone contact, financial incentives) that allow to obtain complete response in the second phase. This assumption is crucial in order to obtain unbiased estimates of population parameters.

In the framework above, three sources of sample randomness were defined, each of them associated with respective probability distribution: $p(s)$, $q(s_1|s)$ and $p'(s'|s,s_2)$. In the following study all expectations will be computed jointly with respect to these three probability distributions unless otherwise stated.

2. Estimation

Under complete response the population total t_x is unbiasedly estimated by the Horvitz-Thompson (1952) statistic:

$$\hat{t}_x = \sum_{i \in s} \frac{x_i}{\pi_i}; \tag{12}$$

In a special case of $x_i=1$ for $i \in U$ we have $t_x = N$ and the statistic above becomes an unbiased estimator of the population size in the form:

$$\hat{N} = \sum_{i \in s} \frac{1}{\pi_i}; \tag{13}$$

As it has been shown by Nargundkar and Joshi (1975), under nonresponse the Horvitz-Thompson estimator computed on the basis of the respondent subset s_1

instead of the sample s is badly biased. However, it is possible to eliminate the bias by introducing the statistic (see Särndal et al 1992):

$$\hat{t}_x^{\bullet} = \sum_{i \in s_1} \frac{x_i}{\pi_i} + \sum_{i \in s'} \frac{x_i}{\pi_i \pi_{i|s,s_2}}; \tag{14}$$

Again, by assuming $x_i=1$ for $i \in U$ we obtain an unbiased estimator of the population size:

$$\hat{N}^{\bullet} = \sum_{i \in s_1} \frac{1}{\pi_i} + \sum_{i \in s'} \frac{1}{\pi_i \pi_{i|s,s_2}}; \tag{15}$$

The same principle may be used to construct nonresponse-corrected estimators of more complicated population parameters. Consider the population variance $S_U^2(X)$ given by expression (3). It may also be expressed in the equivalent form:

$$S_U^2(X) = \frac{1}{N-1} t_{x^2} - \frac{1}{N(N-1)} t_x^2; \tag{16}$$

where $t_{x^2} = \sum_{i \in U} x_i^2$. Under complete response, the unknown population totals t_x and t_{x^2} are often replaced with corresponding Horvitz-Thompson estimators. This leads to two alternative variance estimators:

$$\hat{S}_1^2(X) = \frac{1}{N-1} \hat{t}_{x^2} - \frac{1}{N(N-1)} \hat{t}_x \hat{t}_x; \tag{17}$$

and

$$\hat{S}_2^2(X) = \frac{1}{\hat{N}-1} \hat{t}_{x^2} - \frac{1}{\hat{N}(\hat{N}-1)} \hat{t}_x \hat{t}_x; \tag{18}$$

Consequently, two alternative estimators of the population standard deviation respectively take the form:

$$\hat{S}_1(X) = \sqrt{\hat{S}_1^2(X)}; \tag{19}$$

and

$$\hat{S}_2(X) = \sqrt{\hat{S}_2^2(X)}; \tag{20}$$

Under nonresponse these estimators have to be computed on the basis of the respondent set s_1 instead of s and they are generally biased. To correct for the nonresponse bias we replace Horvitz-Thompson estimators with their unbiased double-sampling counterparts and obtain two new estimators of the population variance:

$$\hat{S}^2_{\bullet 1}(X) = \frac{1}{N-1}\hat{t}^{\bullet}_{x^2} - \frac{1}{N(N-1)}\hat{t}^{\bullet}_{x}\hat{t}^{\bullet}_{x}; \tag{21}$$

$$\hat{S}^2_{\bullet 2}(X) = \frac{1}{\hat{N}^{\bullet}-1}\hat{t}^{\bullet}_{x^2} - \frac{1}{\hat{N}^{\bullet}(\hat{N}^{\bullet}-1)}\hat{t}^{\bullet}_{x}\hat{t}^{\bullet}_{x}; \tag{22}$$

and corresponding two estimators of the population standard deviation.

$$\hat{S}_{\bullet 1}(X) = \sqrt{\hat{S}^2_{\bullet 1}(X)}; \tag{23}$$

$$\hat{S}_{\bullet 2}(X) = \sqrt{\hat{S}^2_{\bullet 2}(X)}; \tag{24}$$

Using the Taylor linearization the approximate variance of $\hat{S}_{\bullet 1}(X)$ may be expressed in the form:

$$AV(\hat{S}_{\bullet 1}(X)) = \frac{1}{4S^2_U(X)(N-1)^2}\left(\sum_{i,j\in U} u_i u_j\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right) + E_{pq}\left(\sum_{i,j\in s_2}\frac{u_i u_j}{\pi_i\pi_j}\left(\frac{\pi_{ij|s,s_2}}{\pi_{i|s,s_2}\pi_{j|s,s_2}}-1\right)\right)\right); \tag{25}$$

where $u_i = (x_i - \overline{X})^2 - \overline{X}^2$. The first-order approximate bias of $\hat{S}_{\bullet 1}(X)$ is null. The second-order approximate bias may be expressed in the form:

$$AB(\hat{S}_{\bullet 1}(X)) = -\frac{1}{8S^3_U(X)(N-1)^2}\left(\sum_{i,j\in U} u_{ij}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right) + E_{pq}\left(\sum_{i,j\in s_2}\frac{u_{ij}}{\pi_i\pi_j}\left(\frac{\pi_{ij|s,s_2}}{\pi_{i|s,s_2}\pi_{j|s,s_2}}-1\right)\right)\right); \tag{26}$$

where $u_{ij} = x_i x_j(u_i u_j + 4x_i x_j \tilde{S}^2_U(X))$ and $\tilde{S}^2_U(X) = ((N-1)/N)\cdot S^2_U(X)$.

Using the same technique we obtain the approximate variance of $\hat{S}_{\bullet 2}(X)$:

$$AV(\hat{S}_{\bullet 2}(X)) = \frac{1}{4S^2_U(X)(N-1)^2}\left(\sum_{i,j\in U} u^*_i u^*_j\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right) + E_{pq}\left(\sum_{i,j\in s_2}\frac{u^*_i u^*_j}{\pi_i\pi_j}\left(\frac{\pi_{ij|s,s_2}}{\pi_{i|s,s_2}\pi_{j|s,s_2}}-1\right)\right)\right); \tag{27}$$

where $u^*_i = (x_i - \overline{X})^2 - S^2_U(X)$. Its first-order approximate bias is null. The second order approximate bias takes the form:

$$\mathrm{AB}(\hat{\mathrm{S}}_{\bullet 2}(\mathrm{X})) = -\frac{1}{8\mathrm{S}_{\mathrm{U}}^{3}(\mathrm{X})(\mathrm{N}-1)^{2}}\left(\sum_{\mathrm{i,j}\in \mathrm{U}} \mathrm{u}_{\mathrm{ij}}^{*}\left(\frac{\pi_{\mathrm{ij}}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}}-1\right)+\mathrm{E}_{\mathrm{pq}}\left(\sum_{\mathrm{i,j}\in \mathrm{s}_2}\frac{\mathrm{u}_{\mathrm{ij}}^{*}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}}\left(\frac{\pi_{\mathrm{ij|s,s_2}}}{\pi_{\mathrm{i|s,s_2}}\pi_{\mathrm{j|s,s_2}}}-1\right)\right)\right); \quad (28)$$

where $\mathrm{u}_{\mathrm{ij}}^{*} = \mathrm{u}_{\mathrm{ij}} + 2(\mathrm{x}_{\mathrm{i}} - \overline{\mathrm{X}})^{2}(\mathrm{S}_{\mathrm{U}}^{2}(\mathrm{X}) + \overline{\mathrm{X}}^{2}) - \overline{\mathrm{X}}^{4} - 3\mathrm{S}_{\mathrm{U}}^{4}(\mathrm{X})$. The symbol $\mathrm{E}_{\mathrm{pq}}(\cdot)$ in expressions above represents the expectation with respect to the sampling design p(s), and the response distribution $q(s_1|s)$. As indicated by Särndal et al (1992), additional assumptions concerning inclusion probabilities in the second phase are needed to get rid of this operator. In the next paragraph this possibility will be illustrated by the example.

On the other hand, the approximate variances may be estimated from the sample without any additional assumptions by respective statistics:

$$\hat{\mathrm{V}}(\hat{\mathrm{S}}_{\bullet 1}(\mathrm{X})) = \frac{1}{4\mathrm{S}_{\mathrm{U}}^{2}(\mathrm{X})(\mathrm{N}-1)}\left(\sum_{\mathrm{i,j}\in \mathrm{s}_1\cup \mathrm{s}'} \frac{\hat{\mathrm{u}}_{\mathrm{i}}\hat{\mathrm{u}}_{\mathrm{j}}}{\pi_{\mathrm{ij}}^{*}}\left(\frac{\pi_{\mathrm{ij}}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}}-1\right)+\sum_{\mathrm{i,j}\in \mathrm{s}'}\frac{\hat{\mathrm{u}}_{\mathrm{i}}\hat{\mathrm{u}}_{\mathrm{j}}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}\pi_{\mathrm{ij|s,s_2}}}\left(\frac{\pi_{\mathrm{ij|s,s_2}}}{\pi_{\mathrm{i|s,s_2}}\pi_{\mathrm{j|s,s_2}}}-1\right)\right); \quad (29)$$

$$\hat{\mathrm{V}}(\hat{\mathrm{S}}_{\bullet 2}(\mathrm{X})) = \frac{1}{4\mathrm{S}_{\mathrm{U}}^{2}(\mathrm{X})(\mathrm{N}-1)}\left(\sum_{\mathrm{i,j}\in \mathrm{s}_1\cup \mathrm{s}'} \frac{\hat{\mathrm{u}}_{\mathrm{i}}^{*}\hat{\mathrm{u}}_{\mathrm{j}}^{*}}{\pi_{\mathrm{ij}}^{*}}\left(\frac{\pi_{\mathrm{ij}}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}}-1\right)+\sum_{\mathrm{i,j}\in \mathrm{s}'}\frac{\hat{\mathrm{u}}_{\mathrm{i}}^{*}\hat{\mathrm{u}}_{\mathrm{j}}^{*}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}\pi_{\mathrm{ij|s,s_2}}}\left(\frac{\pi_{\mathrm{ij|s,s_2}}}{\pi_{\mathrm{i|s,s_2}}\pi_{\mathrm{j|s,s_2}}}-1\right)\right); \quad (30)$$

where

$$\hat{\mathrm{u}}_{\mathrm{i}} = \left(\mathrm{x}_{\mathrm{i}} - \frac{\hat{\mathrm{t}}_{\mathrm{x}}^{\bullet}}{\mathrm{N}}\right)^{2} - \frac{\hat{\mathrm{t}}_{\mathrm{x}}^{\bullet}\hat{\mathrm{t}}_{\mathrm{x}}^{\bullet}}{\mathrm{N}^{2}}; \quad (31)$$

$$\hat{\mathrm{u}}_{\mathrm{i}}^{*} = \left(\mathrm{x}_{\mathrm{i}} - \frac{\hat{\mathrm{t}}_{\mathrm{x}}^{\bullet}}{\hat{\mathrm{N}}^{\bullet}}\right)^{2} - \hat{\mathrm{S}}_{\bullet 2}^{2}(\mathrm{X}); \quad (32)$$

and

$$\pi_{\mathrm{ij}}^{*} = \begin{cases} \pi_{\mathrm{ij}}\pi_{\mathrm{ij\,s,s_2}} & \text{for} \quad \mathrm{i,j}\in \mathrm{s}_2 \\ \pi_{\mathrm{ij}}\pi_{\mathrm{i|s,s_2}} & \text{for} \quad \mathrm{i}\in \mathrm{s}_2, \mathrm{j}\in \mathrm{s}_1 \\ \pi_{\mathrm{ij}}\pi_{\mathrm{j\,s,s_2}} & \text{for} \quad \mathrm{i}\in \mathrm{s}_1, \mathrm{j}\in \mathrm{s}_2 \\ \pi_{\mathrm{ij}} & \text{for} \quad \mathrm{i,j}\in \mathrm{s}_1 \end{cases} \quad (33)$$

If the statistics $\hat{\mathrm{u}}_{\mathrm{i}}$ and $\hat{\mathrm{u}}_{\mathrm{i}}^{*}$ estimated constants u_{i} and $\mathrm{u}_{\mathrm{i}}^{*}$ without error, then variance estimators above would respectively be unbiased for $\mathrm{AV}(\hat{\mathrm{S}}_{\bullet 1}(\mathrm{X}))$ and $\mathrm{AV}(\hat{\mathrm{S}}_{\bullet 2}(\mathrm{X}))$. Obviously, they do not and some bias appears, but we may hope that it remains modest and tends to zero for large samples.

3. A special case

Assume, as in the papers of Srinath (1971) or Rao (1986), that the simple random sampling without replacement (SRSWOR) is used in both phases of the survey. Consequently, inclusion probabilities take the form:

$$\pi_i = \frac{n}{N}; \tag{34}$$

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}; \tag{35}$$

$$\pi_{i|s,s_2} = \frac{n'}{n_2} \tag{36}$$

$$\pi_{ij|s,s_2} = \frac{n'(n'-1)}{n_2(n_2-1)} \tag{37}$$

for $i \neq j \in U$. Assume that the subsample size is a linear function of the nonrespondent subset size according to the formula:

$$n' = cn_2; \tag{38}$$

where 0<c<1. Also assume that nonresponse is deterministic, which means that the population is divided into two disjoint subsets (strata): U_1 and U_2, of sizes N_1 and N_2 such that:

$$\rho_{i|s} = \begin{cases} 1 & \text{for} \quad i \in U_1 \\ 0 & \text{for} \quad i \in U_2 \end{cases} \tag{39}$$

Under these assumptions we have $\hat{N} \equiv N$. Consequently both estimators $\hat{S}_{\bullet 1}(X)$ and $\hat{S}_{\bullet 2}(X)$ are equivalent and take common form:

$$\hat{S}_+(X) = \sqrt{\frac{1}{N-1}\hat{t}^+_{x^2} - \frac{1}{N(N-1)}\hat{t}^+_x\hat{t}^+_x}; \tag{40}$$

where

$$\hat{t}^+_{x^2} = n_1\overline{X^2_{s_1}} + n_2\overline{X^2_{s'}}; \tag{41}$$

$$\hat{t}_x^{+} = n_1 \overline{X}_{s_1} + n_2 \overline{X}_{s'}; \tag{42}$$

$$\overline{X_{s_1}^2} = \frac{1}{n_1} \sum_{i \in s_1} x_i^2 ; \tag{43}$$

$$\overline{X_{s'}^2} = \frac{1}{n'} \sum_{i \in s'} x_i^2 ; \tag{44}$$

$$\overline{X}_{s_1} = \frac{1}{n_1} \sum_{i \in s_1} x_i ; \tag{45}$$

$$\overline{X}_{s'} = \frac{1}{n'} \sum_{i \in s'} x_i ; \tag{46}$$

The approximate variance may be expressed in the form:

$$AV(\hat{S}_+(X)) = \frac{N^2}{4S_U^2(X)(N-1)^2} \left(\frac{N-n}{Nn} S_U^2(u) + \frac{N_2}{N} \frac{1}{n} \frac{1-c}{c} S_{U_2}^2(u) \right); \tag{47}$$

where

$$S_U^2(u) = \frac{1}{N-1} \sum_{i \in U} (u_i - \overline{u})^2 ; \tag{48}$$

$$S_{U_2}^2(u) = \frac{1}{N_2 - 1} \sum_{i \in U_2} (u_i - \overline{u}_{U_2})^2 ; \tag{49}$$

$$\overline{u} = \frac{1}{N} \sum_{i \in U} u_i ; \tag{50}$$

$$\overline{u}_{U_2} = \frac{1}{N_2} \sum_{i \in U_2} u_i ; \tag{51}$$

The first-order bias is nil. The second-order bias takes the form:

$$AB(\hat{S}_+(X)) = -\frac{N^2}{8S^3(N-1)^2} \left(\frac{N-n}{Nn} \left(S_U^2(u) + 4\tilde{S}_U^2(X) S_U^2(X) \right) + \right.$$

$$+\frac{N_2}{Nn}\frac{1-c}{c}\left(S^2_{U_2}(u)+4\tilde{S}^2_{U_2}(X)S^2_{U_2}(X)\right)\Bigg); \tag{52}$$

Both the approximate bias and the approximate variance decrease when initial sample size n grows, which suggests consistency. From (29) we also obtain the variance estimator:

$$\hat{V}\left(\hat{S}_+(X)\right)=\frac{1}{4S^2_U(X)(N-1)}N\frac{N-n}{n(n-1)}\Bigg(\ (n_1-1)S^2_{s_1}(u^+)+$$

$$+\frac{N(n_2-1)-cn_2(n-1)+n_1}{c(N-n)}S^2_{s'}(u^+)+\frac{n_1n_2}{n}(\bar{u}^+_{s_1}-\bar{u}^+_{s'})^2\ \Bigg) \tag{53}$$

where

$$\bar{u}^+_{s_1}=\frac{1}{n_1}\sum_{i\in s_1}\hat{u}^+_i\ ; \tag{54}$$

$$\bar{u}^+_{s'}=\frac{1}{n'}\sum_{i\in s'}\hat{u}^+_i\ ; \tag{55}$$

$$S^2_{s_1}(u^+)=\frac{1}{n_1-1}\sum_{i\in s_1}(\hat{u}^+_i-\bar{u}^+_{s_1})^2\ ; \tag{56}$$

$$S^2_{s'}(u^+)=\frac{1}{n'-1}\sum_{i\in s'}(\hat{u}^+_i-\bar{u}^+_{s'})^2\ ; \tag{57}$$

and

$$\hat{u}^+_i=\left(x_i-\frac{\hat{t}^+_x}{N}\right)^2-\frac{\hat{t}^+_x\hat{t}^+_x}{N^2}; \tag{58}$$

4. Simulation study

As pointed out by Lessler and Kalsbeek (1992), the deterministic nonresponse model considered in the previous paragraph is often viewed as too simplistic. Consequently, a simulation study was carried out to explore the properties of the proposed standard deviation estimator in the case of stochastic nonresponse.

The data on 2420 farms obtained during the 1996' agricultural census in certain municipalities of the Dąbrowa Tarnowska district represented the population under study. Total annual farm sales were chosen as the variable under study. The experiment was carried out by repeatedly drawing samples from the population using SRSWOR, simulating the stochastic nonresponse according to a pre-determined model and drawing using SRSWOR a 30% subsample from the nonrespondent subset. It was assumed that population units respond independently, with individual response probabilities depending on the variable under study and given by expression:

$$\rho_{i|s} = \frac{1}{1+\exp(\beta_0 + \beta_1 x_i)} \tag{59}$$

where the constants β_0=-1 and β_1=-1 were chosen arbitrarily in such a way that average response probability is equal to 0.71 and it diminishes with X. The experiment was repeated for initial sample size n=40,60,...,200. Each time a total of 100000 sample-subsample pairs were drawn. From the empirical distribution of the estimator $\hat{\mathbf{S}}_+(\mathbf{X})$ its properties were assessed. The dependency between initial sample size n and the variance of the estimator is shown on the figure 1. Apparently, the variance tends to zero when n grows. The recorded bias is shown on the figure 2. It takes negative values and tends to zero when n grows. The share of bias in the total MSE is shown on figure 3. Its recorded values fall between 1% and 5% and it also tends to zero with increasing n. These results suggest that the estimator retains its attractive properties under stochastic-type nonresponse model.

Figure 1. The variance of the estimator $\hat{\mathbf{S}}_{+}\mathbf{(X)}$ as a function of initial sample size n.

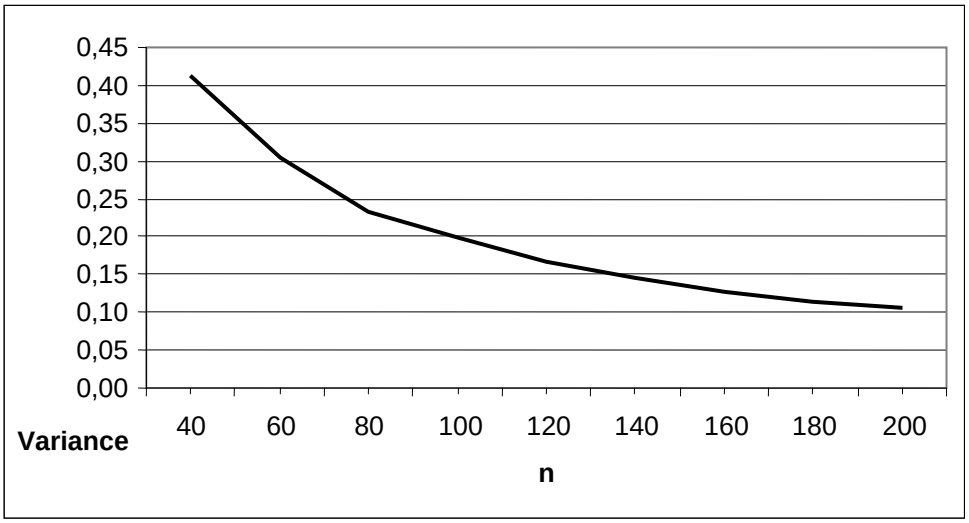


Figure 2. The bias of the estimator $\hat{\mathbf{S}}_{+}\mathbf{(X)}$ as a function of initial sample size n.

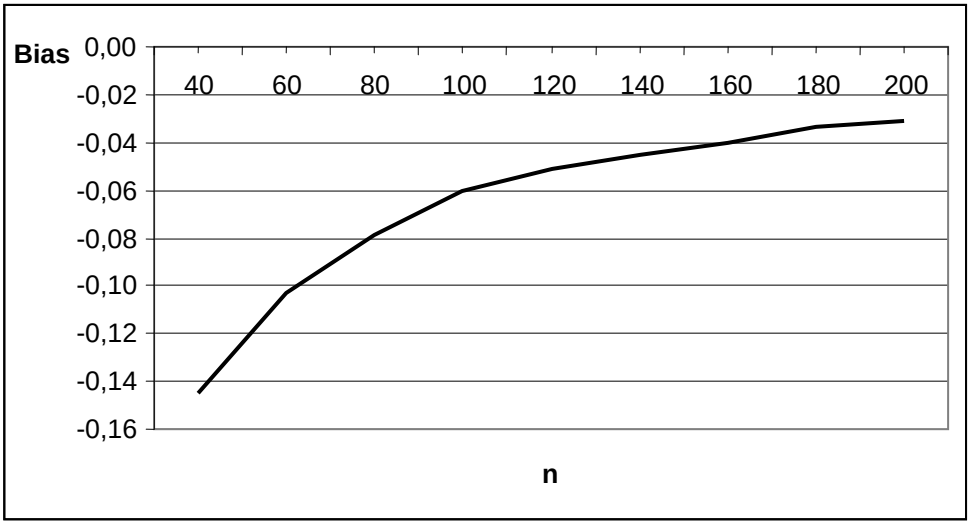
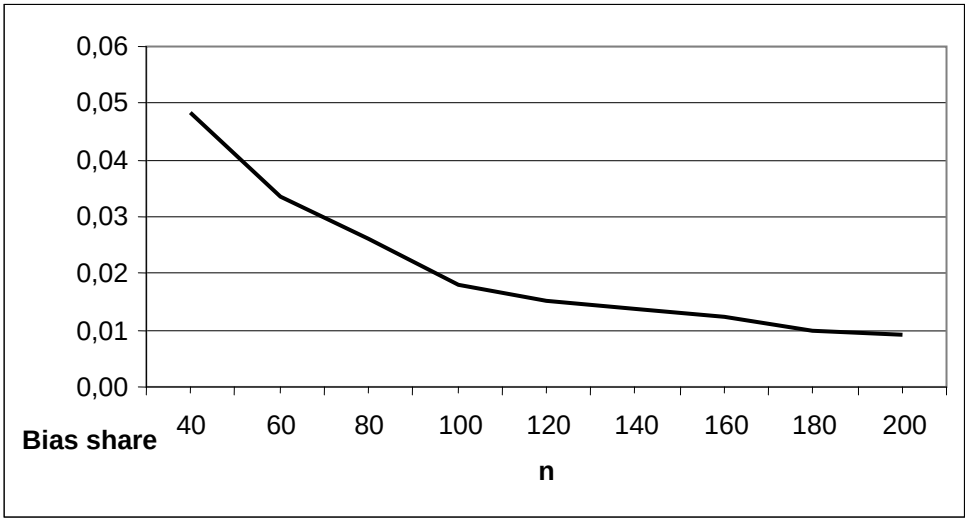


Figure 3. The share of squared bias in the total mean square error of the estimator $\hat{\mathbf{S}}_{+}(\mathbf{X})$ as a function of initial sample size n.



5. Conclusions

In this paper a nonresponse correction for two finite population standard deviation estimators is proposed. The corrected estimators are computed based on the data obtained using a general two-phase sampling procedure involving arbitrary sampling designs in both phases. Their approximate variances and biases are derived for general stochastic nonresponse mechanism governed by arbitrary response distribution. An important special case of simple random sampling without replacement and deterministic nonresponse mechanism is considered. Simulation results exploring estimator's properties under stochastic nonresponse are also presented. Proposed nonresponse adjustment seems to retain the consistency of estimators when the assumption of complete response in the second phase is satisfied. Further research is needed to shed some light on their properties in case of this assumption not being satisfied.

Acknowledgements

The research was financed in the year 2006 by the grant from Polish Ministry of Education and Science (research project No.: 1H02B 022 30).

REFERENCES

CASSEL C. M., SÄRNDAL C. E., WRETMAN J. H. (1983) Some Uses of Statistical Models in Connection with the Nonresponse Problem, in: *Incomplete Data in Sample Surveys* W.G. Madow I. Olkin (eds.) Academic Press New York.

HANSEN M. H., HURWITZ W. N. (1946) The Problem of Nonresponse in Sample Surveys, *Journal of the American Statistical Society*. 41, 517—529.

Horvitz D. G., Thompson D. J. (1952) A Generalization of Sampling Without Replacement from a Finite Universe, *Journal of the American Statistical Association* No 47, 663—685.

LESSLER J. T., KALSBEEK W. D. (1992) *Nonsampling Error in Surveys* John Wiley & Sons, New York.

NARGUNDKAR M. S., JOSHI G. B. (1975) *Non-response in Sample Surveys* 40-th Session of International Statistical Institute - Warsaw 1975, Contributed Papers, 626—628.

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

SÄRNDAL C. E., SWENSSON B. WRETMAN J. H. (1992) *Model Assisted Survey Sampling* Springer-Verlag, New York.

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 33—48

ESTIMATION OF MEAN WITH IMPUTATION OF MISSING DATA USING FACTOR TYPE ESTIMATOR

Diwakar Shukla, Narendra Singh Thakur

ABSTRACT

In sample surveys, the problem of non-response is one of the most frequent and wide whose solution is required to obtain using statistical techniques. The imputation is one such methodology, which uses available data as a tool for the replacement of missing observations. This paper presents the use of factor-type (F-T) estimator as a tool of imputation for dealing with non-responding units in sample surveys. Proposed estimator is found efficient and bias controlled than the other similar estimation procedures. In support of facts a simulation study is performed over multiple data sets showing the numerical based efficiency of the proposed technique.

Key words: Estimation, missing data, imputation, bias, mean squared error (m.s.e.), Factor-Type (F-T).

1. Introduction

Imputation is a technique used for computation of missing values in the sample obtained as consequence of survey procedure. In literature, several imputation techniques are described, some of them are better over others. Rubin (1976) addressed three concepts: MAR, OAR and PD. In what follows MCAR is used. Let $\overline{Y} = N^{-1}\sum_{i=1}^{N} Y_i$ be the mean of a finite population under consideration for estimation. A simple random sample S without replacement (SRSWOR), of size n is drawn from population $\Omega = \{1, 2, ..., N\}$ to estimate $\overline{Y}$. Sample S of n units contains r responding units ($r < n$) forming a set R and ($n - r$) non-responding with the sub-space ($n - r$) having symbol R^C in the space. The variable Y is of main interest and X an auxiliary variable correlated with Y. For every unit $i \in R$, the value y_i is observed available. However, for the units

$i \in R^C$, the y_i values are missing and imputed values are to be derived. The i^{th} value x_i of auxiliary variate is used as a source of imputation for missing data when $i \in R^C$. This is to assume that for sample *S*, the data $x_s = \{x_i : i \in S\}$ are known and $S = R \bigcup R^C$. Under this setup, some well known imputation methods are given below :

(1) RATIO METHOD OF IMPUTATION

For sampled values y_i and x_i, define $y_{\bullet i}$ as

$$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \hat{b} x_i & if \quad i \in R^C \end{cases} \tag{1.1}$$

where $\hat{b} = \sum_{i \in R} y_i \Big/ \sum_{i \in R} x_i$

Using above, the imputation-based estimator of population mean $\bar{Y}$ is:

$$\bar{y}_S = \frac{1}{n} \sum_{i \in S} \bar{y}_{\bullet i} = \bar{y}_r \left(\frac{\bar{x}_n}{\bar{x}_r} \right) = \bar{y}_{RAT} \tag{1.2}$$

where $\bar{y}_r = \frac{1}{r} \sum_{i \in R} \bar{y}_i$, $\bar{x}_r = \frac{1}{r} \sum_{i \in R} \bar{x}_i$ and $\bar{x}_n = \frac{1}{n} \sum_{i \in S} \bar{x}_i$

(2) MEAN METHOD OF IMPUTATION

For y_i and x_i, define $y_{\bullet i}$ as

$$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \\ \bar{y}_r & if \quad i \in R^C \end{cases} \tag{1.3}$$

Using above, the imputation-based estimator of $\bar{Y}$ is :

$$\bar{y}_m = \frac{1}{r} \sum_{i \in R} \bar{y}_i = \bar{y}_r \tag{1.4}$$

(3) COMPROMISED METHOD OF IMPUTATION

Singh and Horn (2000) proposed compromised imputation procedure

$$y_{\bullet i} = \begin{cases} (\alpha n/r) + (1-\alpha)\ \hat{b}x_i & if \quad i \in R \\ (1-\alpha)\ \hat{b}x_i & if \quad i \in R^C \end{cases} \tag{1.5}$$

where α is a suitably chosen constant, such that the resultant variance of the estimator is minimum. The imputation-based estimator, for this case, is

$$\bar{y}_{COMP} = \left[\alpha \bar{y}_r + (1-\alpha)\bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \right] \tag{1.6}$$

(4) AHMED METHODS OF IMPUTATION

For the case where y_{ji} denotes the i^{th} available observation for the j^{th} imputation method Ahmed et al. (2006) suggested:

$$\textbf{(A)}\ y_{1i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\bar{y}_r \left(\dfrac{\bar{X}}{\bar{x}_n} \right)^{\beta_1} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \tag{1.7}$$

Under this, the point estimator of $\bar{Y}$ is

$$t_1 = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_n} \right)^{\beta_1} \tag{1.8}$$

$$\textbf{(B)}\ y_{2i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\bar{y}_r \left(\dfrac{\bar{x}_n}{\bar{x}_r} \right)^{\beta_2} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \tag{1.9}$$

The point estimator of $\bar{Y}$ is

$$t_2 = \bar{y}_r \left(\frac{\bar{x}_n}{\bar{x}_r} \right)^{\beta_2} \tag{1.10}$$

$$\textbf{(C)}\ y_{3i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\bar{y}_r \left(\dfrac{\bar{X}}{\bar{x}_n} \right)^{\beta_3} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \tag{1.11}$$

and point estimator of $\bar{Y}$ is

$$t_3 = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_r} \right)^{\beta_3} \tag{1.12}$$

Terms β_1, β_2 and β_3 are suitably chosen constants, so as to keep the variance of the resultant estimator minimum.

As special cases:

when $\beta_3 = 1$, then $t_{Ratio} = \bar{y}_r \left(\dfrac{\bar{X}}{\bar{x}_r} \right)$ (1.13)

and when $\beta_3 = -1$, then $t_{\mathrm{Pr}oduct} = \bar{y}_r \left(\dfrac{\bar{x}_r}{\bar{X}} \right)$ (1.14)

This is natural analogue of the ratio estimator called the product estimator used when an auxiliary variate *x* has negative correlation with *y*.

2. Proposed imputation methods

Singh and Shukla (1987) have discussed a family of factor-type (F-T) ratio estimator for estimating population mean. In one more contribution, Singh and Shukla (1993) derived efficient factor-type estimator for estimating the same population parameter. Deriving motivation from both of these, some proposed imputation methods for missing data are:

$$\textbf{(D)}\ (y_{FT1})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{\bar{y}_r}{(n-r)}\left[n\phi_1(k) - r\right] & if \quad i \in R^C \end{cases} \tag{2.1}$$

$$\textbf{(E)}\ (y_{FT2})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{\bar{y}_r}{(n-r)}\left[n\phi_2(k) - r\right] & if \quad i \in R^C \end{cases} \tag{2.2}$$

$$\textbf{(F)}\ (y_{FT3})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{\bar{y}_r}{(n-r)}\left[n\phi_3(k) - r\right] & if \quad i \in R^C \end{cases} \tag{2.3}$$

where $\phi_1(k) = \left[\dfrac{(A+C)\bar{X} + fB\bar{x}_n}{(A+fB)\bar{X} + C\bar{x}_n}\right]$; $\phi_2(k) = \left[\dfrac{(A+C)\bar{x}_n + fB\bar{x}_r}{(A+fB)\bar{x}_n + C\bar{x}_r}\right]$;

$$\phi_3(k) = \left[\frac{(A+C)\bar{X} + fB\bar{x}_r}{(A+fB)\bar{X} + C\bar{x}_r}\right];$$

$$A = (k-1)(k-2)\,; \qquad B = (k-1)(k-4);$$

$$C = (k-2)(k-3)(k-4); \qquad f = \frac{n}{N} \text{ and } 0 < k < \infty \text{ is a constant.}$$

Under (2.1), (2.2) and (2.3), point estimators of $\bar{Y}$ are :

$$\left.\begin{aligned} T_{FT1} &= \bar{y}_r\phi_1(k) \\ T_{FT2} &= \bar{y}_r\phi_2(k) \\ T_{FT3} &= \bar{y}_r\phi_3(k) \end{aligned}\right\} \tag{2.4}$$

2.1. Special cases

Following relate to Ahmed et al. (2006) :

$$\text{when } k=1,\ \beta_l=1 \text{ then } T_{FTl}=t_l$$

$$\text{when } k=2,\ \beta_l=-1 \text{ then } T_{FTl}=t_l$$

$$\text{and when } k=4,\ \beta_l=0 \text{ then } T_{FTl}=t_l=\bar{y}_r;\ (l=1,2,3)$$

3. Properties of proposed imputation methods

Let $\varepsilon=\frac{\bar{y}_r}{\bar{Y}}-1;\ \delta=\frac{\bar{x}_r}{\bar{X}}-1$ and $\eta=\frac{\bar{x}_n}{\bar{X}}-1$. Using the concept of two-phase sampling, following Rao and Sitter (1995) and the mechanism of MCAR, for given r and n, we have:

$$E(\varepsilon)=E(\delta)=E(\eta)=0$$

$$E(\varepsilon^2)=M_1C_Y^2;\quad E(\delta^2)=M_1C_X^2 \qquad E(\eta^2)=M_2C_X^2;$$

$$E(\varepsilon\delta)=M_1\rho C_Y\,C_X;\quad E(\varepsilon\eta)=M_2\rho C_Y\,C_X \qquad E(\delta\eta)=M_2C_X^2;$$

$$E(\varepsilon^i\delta^j)=E(\varepsilon^i\eta^j)=E(\delta^i\eta^j)=0 \quad \text{where } i+j>2,\quad j,i=0,1,2,3,4.....$$

$$M_1=\left(\frac{1}{r}-\frac{1}{N}\right);\ M_2=\left(\frac{1}{n}-\frac{1}{N}\right);$$

$$C_Y^2=\frac{S_Y^2}{\bar{Y}^2};\ C_X^2=\frac{S_X^2}{\bar{X}^2};\ \rho=\frac{S_{XY}}{S_XS_Y}$$

where S_Y^2, S_X^2 and S_{XY} have their usual meanings related to population mean squares of Y and X.

REMARK 3.1.

Define $\theta_1=\frac{fB}{A+fB+C};\qquad \theta_2=\frac{C}{A+fB+C};$

$$\theta_3 = \frac{A+C}{A+fB+C}; \qquad \theta_4 = \frac{A+fB}{A+fB+C};$$

$$P = (\theta_1 - \theta_2) = -(\theta_3 - \theta_4);$$

$$V = \rho \frac{C_Y}{C_X}; \qquad M_3 = (M_1 - M_2) = \left(\frac{1}{r} - \frac{1}{n}\right).$$

THEOREM 3.1.

[a_1]: The estimator T_{FT1} in terms of ε, δ and η, up to first order of approximation, could be expressed as:

$$T_{FT1} = \overline{Y}\left[1 + \varepsilon + P\left(\eta + \varepsilon\eta - \theta_2\eta^2\right)\right] \tag{3.1}$$

[a_2]: Bias of T_{FT1} up to first order of approximation is:

$$B(T_{FT1}) = -PM_2\overline{Y}\left(\theta_2 C_X^2 - \rho C_Y C_X\right) \tag{3.2}$$

[a_3]: Mean squared error of T_{FT1} up to first order is :

$$M(T_{FT1}) = \overline{Y}^2\left[M_1 C_Y^2 + PM_2\left(PC_X^2 + 2\rho C_Y C_X\right)\right] \tag{3.3}$$

[a_4]: The minimum m.s.e. of T_{FT1} occurs when $P = -V$ and expression is:

$$M(T_{FT1})_{\min} = \left(M_1 - M_2\rho^2\right)S_Y^2 \tag{3.4}$$

PROOF:

[a_1]: $T_{FT1} = \overline{y}_r\phi_1(k)$ and use of approximation of section 3.0,

$$= \overline{Y}(1+\varepsilon)\left[\frac{(A+C)\overline{X} + fB\overline{X}(1+\eta)}{(A+fB)\overline{X} + C\overline{X}(1+\eta)}\right]$$

$$= \overline{Y}(1+\varepsilon)(1+\theta_1\eta)(1+\theta_2\eta)^{-1}$$

$$= \overline{Y}(1+\varepsilon)(1+\theta_1\eta)(1-\theta_2\eta+\theta_2^2\eta^2 - .....)$$

$$= \overline{Y}\left[1 + \varepsilon + P\left(\eta + \varepsilon\eta - \theta_2\eta^2\right)\right]$$

[a$_2$]: $B(T_{FT1}) = E(T_{FT1} - \overline{Y})$

$$= E\left[\overline{Y}\{1 + \varepsilon + P(\eta + \varepsilon\eta - \theta_2\eta^2)\} - \overline{Y}\right]$$

$$= -PM_2\overline{Y}(\theta_2 C_X^2 - \rho C_Y C_X)$$

[a$_3$]: $M(T_{FT1}) = E(T_{FT1} - \overline{Y})^2$

$$= E\left[\overline{Y}\{1 + \varepsilon + P(\eta + \varepsilon\eta - \theta_2\eta^2)\} - \overline{Y}\right]^2$$

$$= \overline{Y}^2\left[M_1 C_Y^2 + PM_2(PC_X^2 + 2\rho C_Y C_X)\right]$$

[a$_4$]: Minimum m.s.e. occurs when

$$\frac{d}{dP}M(T_{FT1}) = 0$$

$$\Rightarrow 2PC_X^2 + 2\rho C_Y C_X = 0 \qquad \Rightarrow P = -V$$

$$\Rightarrow M(T_{FT1})_{\min} = (M_1 - M_2\rho^2)\, S_Y^2$$

THEOREM 3.2.

[a$_5$]: The estimator T_{FT2} in terms of ε, δ and η, up to first order of approximation, is:

$$T_{FT2} = \overline{Y}\left[1 + \varepsilon + P\{\delta - \eta + \varepsilon\delta - \varepsilon\eta + (\theta_2 - \theta_4)\delta\eta - \theta_2\delta^2 + \theta_4\eta^2\}\right] \quad (3.5)$$

[a$_6$]: Bias of estimator T_{FT2} is :

$$B(T_{FT2}) = -PM_3\overline{Y}(\theta_2 C_X^2 - \rho C_Y C_X) \quad (3.6)$$

[a$_7$]: Mean squared error of T_{FT2} is :

$$M(T_{FT2}) = \overline{Y}^2\left[M_1 C_Y^2 + PM_3(PC_X^2 + 2\rho C_Y C_X)\right] \quad (3.7)$$

[a$_8$]: The minimum m. s. e. of T_{FT2} is at $P = -V$

$$M(T_{FT2})_{\min} = (M_1 - M_3\rho^2)S_Y^2 \quad (3.8)$$

PROOF:

[a$_5$]: $T_{FT2} = \bar{y}_r \phi_2(k)$

$$= \bar{Y}(1+\varepsilon)\left[\frac{(A+C)(1+\eta)\bar{X} + fB(1+\delta)\bar{X}}{(A+fB)(1+\eta)\bar{X} + C(1+\delta)\bar{X}}\right]$$

$$= \bar{Y}(1+\varepsilon)(1+\theta_1\delta+\theta_3\eta)(1+\theta_2\delta+\theta_4\eta)^{-1}$$

$$= \bar{Y}\left[1+\varepsilon+P\left\{\delta-\eta+\varepsilon\delta-\varepsilon\eta+(\theta_2-\theta_4)\delta\eta-\theta_2\delta^2+\theta_4\eta^2\right\}\right]$$

[a$_6$]: $B(T_{FT2}) = E(T_{FT2} - \bar{Y})$

$$= E\left[\bar{Y}\left\{1+\varepsilon+P\left\{\delta-\eta+\varepsilon\delta-\varepsilon\eta+(\theta_2-\theta_4)\delta\eta-\theta_2\delta^2+\theta_4\eta^2\right\}\right\}-\bar{Y}\right]$$

$$= -PM_3\bar{Y}(\theta_2 C_X^2 - \rho C_Y C_X)$$

[a$_7$]: $M(T_{FT2}) = E(T_{FT2} - \bar{Y})^2$

$$= E\left[\bar{Y}\left\{1+\varepsilon+P\left\{\delta-\eta+\varepsilon\delta-\varepsilon\eta+(\theta_2-\theta_4)\delta\eta-\theta_2\delta^2+\theta_4\eta^2\right\}\right\}-\bar{Y}\right]^2$$

$$= \bar{Y}^2\left[M_1 C_Y^2 + PM_3(PC_X^2 + 2\rho C_Y C_X)\right]$$

[a$_8$]: To obtain minimum m. s. e., let

$$\frac{d}{dP}M(T_{FT2}) = 0 \quad \Rightarrow P = -V$$

and substitution provides

$$M(T_{FT2})_{\min} = (M_1 - M_3\rho^2)S_Y^2$$

THEOREM 3.3.

[a$_9$]: Estimator T_{FT3} in terms of ε, δ and η, up to first order of approximation, is :

$$T_{FT3} = \bar{Y}\left[1+\varepsilon+P(\delta+\varepsilon\delta-\theta_2\delta^2)\right] \tag{3.9}$$

[a$_{10}$]: Bias expression is:

$$B(T_{FT3}) = -PM_1\bar{Y}(\theta_2 C_X^2 - \rho C_Y C_X) \tag{3.10}$$

[a11]: The mean squared error of T_{FT3} is :

$$M(T_{FT3}) = M_1 \overline{Y}^2 \left[C_Y^2 + P^2 C_X^2 + 2P\rho C_Y C_X \right] \quad (3.11)$$

[a12]: Minimum mean squared error is:

$$M(T_{FT3})_{\min} = (1-\rho^2) M_1 S_Y^2 \qquad \text{at } P = -V \quad (3.12)$$

PROOF:

[a9]: $T_{FT3} = \overline{y}_r \phi_3(k)$

$$= \overline{Y}(1+\varepsilon)\left[\frac{(A+C)\overline{X} + fB(1+\delta)\overline{X}}{(A+fB)\overline{X} + C(1+\delta)\overline{X}} \right]$$

$$= \overline{Y}(1+\varepsilon)(1+\theta_1\delta)(1+\theta_2\delta)^{-1}$$

$$= \overline{Y}\left[1 + \varepsilon + P(\delta + \varepsilon\delta - \theta_2\delta^2)\right]$$

[a10]: $B(T_{FT3}) = E(T_{FT3} - \overline{Y})$

$$= E\left[\overline{Y}\{1 + \varepsilon + P(\delta + \varepsilon\delta - \theta_2\delta^2)\} - \overline{Y}\right]$$

$$= -PM_1\overline{Y}(\theta_2 C_X^2 - \rho C_Y C_X)$$

[a11]: $M(T_{FT3}) = E(T_{FT3} - \overline{Y})^2$

$$= E\left[\overline{Y}\{1 + \varepsilon + P(\delta + \varepsilon\delta - \theta_2\delta^2)\} - \overline{Y}\right]^2$$

$$= M_1\overline{Y}^2\left[C_Y^2 + P^2 C_X^2 + 2P\rho C_Y C_X\right]$$

[a12]: Using $\frac{d}{dP} M(T_{FT3}) = 0$

the minimum m.s.e. is

$$M(T_{FT3})_{\min} = (1-\rho^2) M_1 S_Y^2 \qquad \text{at } P = -V$$

3.1. Multiple choices of k :

The optimality condition $P = -V$ provides the equation

$$AV + (V+1)fB + (V-1)C = 0 \quad (3.1.1)$$

which is cubic in terms of k. One can get at most three values of k like k_1, k_2, k_3 for which m. s. e. is optimal. The best choice criteria for k is :

STEP I: Compute $\left|B(T_{FTi})_{k_j}\right|$ for i, $j = 1, 2, 3$.

STEP II: For given values of i, choose k_j as

$$\left|B(T_{FTi})_{k_j}\right| = \min_{j=1,2,3}\left[\left|B(T_{FTi})_{k_j}\right|\right]$$

This ultimately gives bias control at the optimal level of m. s. e.

NOTE 3.1: For given pair of values of (V, f), $0 < V < \infty$; $0 < f < 1$, one can generate a trivariate table for k_1, k_2, k_3 so as to achieve solution quickly.

4. Comparison

[b$_1$]: $D_1 = \left[M(T_{FT1})_{\min} - M(T_{FT2})_{\min}\right] = (M_3 - M_2)\rho^2 S_Y^2$ (4.1)

So, T_{FT2} is better than T_{FT1}

$$\text{if} \quad r < \frac{Nn}{2N-n} = \frac{n}{2-f}, \quad (0 < f < 1)$$

$$\left.\begin{array}{l} \text{when } f = 0, r < \dfrac{n}{2} \\ \text{when } f = 0, \ r < n \end{array}\right\}$$

Hence, $D_1 > 0$ occurs most of time while $r < \dfrac{n}{2}$ which is usual.

[b$_2$]: $D_2 = \left[M(T_{FT1})_{\min} - M(T_{FT3})_{\min}\right] = M_3\rho^2 S_Y^2$ (4.2)

which is always positive, therefore T_{FT3} is always better than T_{FT1}.

[b$_3$]: $D_3 = \left[M(T_{FT2})_{\min} - M(T_{FT3})_{\min}\right] = M_2\rho^2 S_Y^2$ (4.3)

which is always positive providing T_{FT3} is better than T_{FT2}.

5. Empirical study

An attached appendix A has a generated artificial population of size $N = 200$ containing values of main variable Y and auxiliary variable X. Parameters of this are given below:

$$\overline{Y} = 42.485; \quad \overline{X} = 18.515; \quad S_Y^2 = 199.0598; \quad S_X^2 = 48.5375;$$

$$\rho = 0.8652; \; C_X = 0.3763; \quad C_Y = 0.3321; \quad V = \rho \frac{C_Y}{C_X} = 0.7635;$$

Using random sample of size $n = 20$; $f = 0.1$ by SRSWOR some computed optimum values of constants are

$$\alpha = 0.2365; \quad \beta_1 = \beta_2 = \beta_3 = V = 0.7635;$$

By solving optimum condition $P = -V$ as in (3.1.1), the cubic equation in k provides three k-values

$$k_1 = 1.5206; \qquad k_2 = 2.4505; \qquad k_3 = 8.9456.$$

The condition of bias and m.s.e. of the existing and proposed estimator are computed based of 30,000 repeated samples drawn by SRSWOR from population $N = 200$. These computations, with respect to $\overline{y}_r$, are given in tables 5.1, 5.2, 5.3 and 5.4, where efficiency measurement is considered as

$$e(\hat{y}) = \frac{M(\overline{y}_r)}{M(\hat{y})}$$

with $M(\hat{y})$ the mean squared error of any estimator $\hat{y}$.

The simulation procedure contains following steps:

STEP 1: Draw a random sample of size 20 from the population of $N = 200$ by SRSWOR.

STEP 2: Drop down 5 units randomly from each sample corresponding to Y.

STEP 3: Compute and impute the dropped units of Y with the help of proposed methods and available methods.

STEP 4: Repeat the above steps 30,000 times, which provides multiple sample based estimates $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$,…….. $\hat{y}_{30000}$.

STEP 5: Bias of $\hat{y}_1$ is obtained by

$$B(\hat{y}) = \frac{1}{30000} \sum_{i=1}^{30000} \left[(\hat{y}_i) - \overline{Y} \right]$$

STEP 6: M. S. E. of $\hat{y}$ is computed by

$$M(\hat{y}) = \frac{1}{30000} \sum_{i=1}^{30000} \left[(\hat{y}_i) - \overline{Y} \right]^2$$

Table 5.1.

Estimator	Min M(.)	Efficiency e(.)	Bias (.)
$\overline{y}_r$	10.5462	1	0.2683
$\overline{y}_{RAT}$	9.4185	1.1197	0.3216
$\overline{y}_{COMP}$	8.5039	1.2401	0.4119

Table 5.2.

Estimator	*Min M(.)*	*Efficiency e(.)*	*Bias (.)*
t_1	7.5314	1.4003	0.4117
$(T_{FT1})_{k_1}$	7.2523	1.4541	0.3989
$(T_{FT1})_{k_2}$	7.2638	1.4518	0.3368
$(T_{FT1})_{k_3}$	7.2436	1.4559	0.3904

Table 5.3.

Estimator	*Min M(.)*	*Efficiency e(.)*	*Bias(.)*
t_2	8.6355	1.2213	0.3103
$(T_{FT2})_{k_1}$	8.1399	1.2956	0.2947
$(T_{FT2})_{k_2}$	7.8758	1.2991	0.2508
$(T_{FT2})_{k_3}$	8.1858	1.2884	0.2572

Table 5.4.

Estimator	*Min M(.)*	*Efficiency e(.)*	*Bias (.)*
t_3	3.4158	3.0875	0.4863
$\left(T_{FT3}\right)_{k_1}$	3.1192	3.3811	0.4043
$\left(T_{FT3}\right)_{k_2}$	3.1043	3.3973	0.3363
$\left(T_{FT3}\right)_{k_3}$	3.1029	3.3988	0.3948

6. Almost unbiased imputation methods

In terms of expression (3.2), (3.6) and (3.10), the bias of T_{FTi} ; i = 1, 2, 3 could be made zero to the first order of approximation. This provides three equations :

$$PM_2\overline{Y}\left(\theta_2 C_X^2 - \rho C_Y C_X\right) = 0 \tag{6.1}$$

$$PM_3\overline{Y}\left(\theta_2 C_X^2 - \rho C_Y C_X\right) = 0 \tag{6.2}$$

$$PM_1\overline{Y}\left(\theta_2 C_X^2 - \rho C_Y C_X\right) = 0 \tag{6.3}$$

Using data of appendix A, in all above (6.1), (6.2) and (6.3), the solution appears either

$$\text{(i)} \quad P = 0 \tag{6.4}$$

$$\text{or (ii)} \quad AV + fBV + (V-1)C = 0 \tag{6.5}$$

The equation (6.4) provides choice either $k = k' = 4$ or $k = k_1' = 1.9156$, $k = k_2' = 3.1844$ where proposed estimators are almost unbiased. By (6.5) which is cubic equation in k, values $k = k_1'' = 1.5206$; $k = k_2'' = 2.4505$ and $k = k_3'' = 8.9456$ make the estimator imputed almost unbiased to the first order of approximation.

7. Conclusions

The Factor - type (F-T) estimator of Singh and Shukla (1987) has been used as a imputation tool for missing data, which is found effective and efficient over

Ahmed et al. (2006). In fact, procedures of Ahmed et al. (2006) are some what special cases of F-T imputed estimators when

$$k = 1; \beta_1 = \beta_2 = \beta_3 = 1$$

$$k = 1; \beta_1 = \beta_2 = \beta_3 = -1$$

This implies the factor-type could be looked upon as a general class over Ahmed et al. (2006) for some choices. An specific property with F-T imputed estimator is there are three values constant *k* for which m. s. e. attains the optimum. The choice of *k* among these having the lowest bias. In this way, the factor-type imputation based estimator reduces the bias along with maintaining the optimum level of mean squared error, which are not in Singh and Horn (2000) and Ahmed et al (2006). Table 5.1 to 5.4 shows better efficiency and bias of imputed F-T estimators over estimators t_1, t_2 and t_3; over Ahmed et al. (2006), $\bar{y}_r$, $\bar{y}_{RAT}$ and $\bar{y}_{COMP}$. More importantly, the imputed factor-type estimators could be made almost unbiased also by an appropriate choice of multiple available *k* values. One can choose that *k* as best having almost unbiased level F-T estimator with lowest possible m.s.e. One can generate a trivariate table for easy optimum choices of *k* values for different pair of (*f, V*) values. The proposed methodology is an improved, efficient and more practicable version than Singh and Horn (2000) and Ahmed et al. (2006).

REFERENCES

AHMED, M. S., AL-TITI, O., AL-RAWI, Z. and ABU-DAYYEH, W. (2006): *Estimation of a population mean using different imputation methods,* Statistics in Transition, 7, 6, 1247—1264.

COCHRAN, W. G. (1977): *Sampling Techniques,* John Wiley and Sons, New York.

RAO, J. N. K. and SITTER, R. R. (1995): *Variance estimation under two-phase sampling with application to imputation for missing data,* Biometrika, 82, 453—460.

RUBIN, D. B. (1976): *Inference and missing data,* Biometrika, 63, 581—593.

SINGH, S. and HORN, S. (2000): *Compromised imputation in survey sampling,* Metrika, 51, 266—276.

SINGH, V. K. and SHUKLA, D. (1987): *One parameter family of factor-type ratio estimator,* Metron, 45, 1—2, 273—283.

SINGH, V. K. and SHUKLA, D. (1993): *An efficient one parameter family of factor-type estimator in sample survey,* Metron, 51, 1—2, 139—159.

APPENDIX A

Population (N = 200)

Y_i	45	50	39	60	42	38	28	42	38	35
X_i	15	20	23	35	18	12	8	15	17	13
Y_i	40	55	45	36	40	58	56	62	58	46
X_i	29	35	20	14	18	25	28	21	19	18
Y_i	36	43	68	70	50	56	45	32	30	38
X_i	15	20	38	42	23	25	18	11	09	17
Y_i	35	41	45	65	30	28	32	38	61	58
X_i	13	15	18	25	09	08	11	13	23	21
Y_i	65	62	68	85	40	32	60	57	47	55
X_i	27	25	30	45	15	12	22	19	17	21
Y_i	67	70	60	40	35	30	25	38	23	55
X_i	25	30	27	21	15	17	09	15	11	21
Y_i	50	69	53	55	71	74	55	39	43	45
X_i	15	23	29	30	33	31	17	14	17	19
Y_i	61	72	65	39	43	57	37	71	71	70
X_i	25	31	30	19	21	23	15	30	32	29
Y_i	73	63	67	47	53	51	54	57	59	39
X_i	28	23	23	17	19	17	18	21	23	20
Y_i	23	25	35	30	38	60	60	40	47	30
X_i	07	09	15	11	13	25	27	15	17	11
Y_i	57	54	60	51	26	32	30	45	55	54
X_i	31	23	25	17	09	11	13	19	25	27
Y_i	33	33	20	25	28	40	33	38	41	33
X_i	13	11	07	09	13	15	13	17	15	13
Y_i	30	35	20	18	20	27	23	42	37	45
X_i	11	15	08	07	09	13	12	25	21	22
Y_i	37	37	37	34	41	35	39	45	24	27
X_i	15	16	17	13	20	15	21	25	11	13
Y_i	23	20	26	26	40	56	41	47	43	33
X_i	09	08	11	12	15	25	15	25	21	15
Y_i	37	27	21	23	24	21	39	33	25	35
X_i	17	13	11	11	09	08	15	17	11	19
Y_i	45	40	31	20	40	50	45	35	30	35
X_i	21	23	15	11	20	25	23	17	16	18
Y_i	32	27	30	33	31	47	43	35	30	40
X_i	15	13	14	17	15	25	23	17	16	19
Y_i	35	35	46	39	35	30	31	53	63	41
X_i	19	19	23	15	17	13	19	25	35	21
Y_i	52	43	39	37	20	23	35	39	45	37
X_i	25	19	18	17	11	09	15	17	19	19

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 49—70

ESTIMATING ITEM WEIGHTS OF CONSUMER PRICE INDEXES FOR SMALL REFERENCE POPULATIONS

Rosa Bernardini Papalia[1]

ABSTRACT

This paper discusses issues of estimation methods used to compute consumer price indexes for small subgroups of the reference population. Similar situations arise when estimates are needed for many domains obtained by classifying the population according to various characteristics (i.e. low/high income households). In this view, the price movement measurement is weighted by the importance of the item in the spending patterns of the appropriate population group and each index can illustrate and explain the impact of local economic conditions on consumers' experience with price change. The study aims to illustrate the use of some suitable estimators of the expenditure shares and to propose an alternative estimator based on the maximum entropy principle. The proposed estimation procedure is applied to compute regional consumer price indexes for small population subgroups.

Key words: Expenditure weights, small-domain estimation, Generalized Maximum Entropy estimation, mean square error.

1. Introduction

This paper discusses issues of estimation methods used to compute consumer price indexes for small subgroups of the reference population. We refer to special group indices which are computed using the same methodology as for the reference population index and which have the same item coverage of the reference population. The price series used for the subgroups are based on data collected for the whole consumer price index (CPI). Each of the indices is computed using a set of expenditure weights representing different basket, but all are drawn from the same reference period. Potential dissimilarities between the

[1] Rosa Bernardini Papalia,University of Bologna, Department of Statistics, Via Belle arti 41, 40126 Bologna – Italy, tel: +39 051 2098275, fax: +39 051 232153, e-mail: rossella.bernardini@unibo.it

indices for the special groups and the index for the reference population should result solely from different sets of weights.

In this view, the price movement measurement for the specific subpopulation group is weighted by the importance of the item in the spending patterns of the appropriate population group and each index can illustrate and explain the impact of local economic conditions on consumers' experience with price change.

For subpopulations comprising only a small fraction of the total population, there is a small number of sampled households and, the direct estimation method could yield CPIs which are not accurate enough. More specifically, the task of estimate accurate consumer price index weights for domains with a small number of sampled households cannot be achieved by the traditional design-based procedure which use survey data only from the subgroup of interest because of the availability of a smaller sample relative to the total sample. Similar situations arise when estimates are needed for many domains obtained by classifying the population according to various characteristics (i.e. low/high income households).

The study aims to illustrate the use of some suitable estimators of the expenditure shares of CPI's for small reference populations and to propose an alternative estimator based on the maximum entropy principle. The proposed estimator is derived by the estimation of a demand system which consists of a set of budget share equations.

The paper is organized as follows: in Section 2 composite and Bayesian estimators of CPI weights for small reference subpopulations are presented. An alternative estimation procedure, based on the maximum entropy principle, is proposed and discussed in section 3. Section 4 illustrates a procedure for decomposing the relative difference between two price indexes obtained from alternative population groups. Results of an application relative to the proposed estimation technique are presented in section 5. Finally, concluding remarks are provided in Section 6.

2. Estimation of expenditure-population weights in Laspeyres Consumer Price Indexes

The Laspeyres CPI is defined as a fixed-quantity price index able to measure the price change in a fixed market basket of consumption goods and services that are purchased by the reference population (Turvey, 2002).

Let $\hat{e}_g^0$ be an estimator of the expenditure on commodity group g ($\hat{e}^0 = \sum_g \hat{e}_g^0$), the Laspeyres CPI at time t with base year 0 is given by:

$$
\begin{aligned}
&I^t = \sum_g I_g^t w_g^0 = \sum_g I_g^t \hat{e}_g^0 / \hat{e}^0 \\
&\sum_g w_g^0 = 1, \\
&w_g^0 = \frac{\sum_{i \in H^0} e_i^0 w_{ig}^0}{\sum_{i \in H^0} e_i^0}, \\
&e_i^0 = \sum_g e_{ig}^0
\end{aligned}
\tag{1}
$$

where I^t_g is the price index of commodity group (or expenditure item) g, ($g=1,..,G$), e^0_i denotes household i total expenditures (i=1,...,H), w^0_g is the budget shares for expenditure category g in the CPI, and w^0_{ig} denotes household i budget share for expenditure category g. The budget shares for expenditure category g in the CPI, w^0_g, equals the share of the expenditure on g in the total base year expenditure on all commodities for a specific group of households.

The methodology suggested in this paper to estimate expenditure weights is based on a similar weighted structure than in the official indexes which are used as a general measure of inflation or as a deflator in national accounts. Expenditure shares for each good are treated as if they were those of an aggregate "super-household" representative of the specific population group. More specifically, the CPIs use weights which reflect the composition of the estimate aggregate values of the reference population. Each household contributes to these weights by an amount proportional to its expenditure. Such weighting is named "plutocratic" in contrast with the "democratic" type of weighting which gives equal importance to all households by averaging consumption value proportions over the whole reference population. The aggregate CPI is computed with the weights which reflect the expenditure of an average household. This resulting Laspeyres price index, known as plutocratic Laspeyres price index (Prais, 1959), measures the price change of total base period consumption and it may be interpreted as the price change of the "representative" household's base period consumption relative to the reference population group.

In summary, the price movement measurement is weighted by the importance of the item in the spending patterns of the appropriate population group. After estimating these weights for each of the items of each population groups, the final CPI for each reference population group is computed by applying price data for each item to each expenditure weight.

Considering the partial price indices, and a weight estimator $\hat{w}^0_g = \hat{e}^0_g / \hat{e}^0$, the relative change over time (relative to period t with respect to period s) of the CPI estimator ($\hat{I}^t$) is given by:

$$\left(\hat{I}^t \Big/ \hat{I}^s\right) - 1 = \left(\frac{\sum_g I_g^t \hat{w}_g^0}{\sum_g I_g^s \hat{w}_g^0}\right) - 1 = \left(\frac{\sum_g I_g^t \hat{e}_g^0}{\sum_g I_g^s \hat{e}_g^0}\right) - 1 \quad (t>s) \tag{2}$$

Linearizing the previous expression in (2) by a first order Taylor series approximation produces the following variance formula (Särndal et al., 1992):

$$\begin{aligned} &\operatorname{var}\left[\left(\hat{I}^t \Big/ \hat{I}^s\right) - 1\right] = \operatorname{var}\left(\hat{I}^t \Big/ \hat{I}^s\right) \approx \left(I^t \Big/ I^s\right)^2 \sum_g \sum_h B_{gh}^{st} \operatorname{cov}\left(\hat{e}_g^0, \hat{e}_h^0\right) \Big/ \left(\bar{e}^0\right)^2, \\ &B_{gh}^{st} = \left[\left(I_g^t / I^t\right) - \left(I_g^s / I^s\right)\right]\left[\left(I_h^t / I^t\right) - \left(I_h^s / I^s\right)\right]. \end{aligned} \tag{3}$$

Therefore, the variance of the CPI change, conditional on the partial price indices, can be expressed as a weighted sum of the variances and covariances of expenditures on all commodity groups (Boon and Haan, 1997).

In this view, we are interested in improving the accuracy of the CPI estimator by using an appropriate estimator of the expenditure shares for each small reference population.

2.1. Composite estimation methods

Composite estimation methods can be used to decrease the mean square error (MSE) of the expenditure shares of the subgroup of the reference population for which very little information is obtained from the sample surveys by using data from the corresponding reference population.

After computing expenditure shares for each commodity class within each subgroup and major group of the reference population, the composite estimated expenditure share of a particular commodity class is obtained as a weighted average of the two preliminary expenditure share estimators one for the subpopulation group and one for the major population group.

Let $w^0{}_g$ be an expenditure share estimate of the commodity group g for a major population group and $w^0{}_{g,j}$ be its estimate for the subpopulation group j within that major population of interest. Then, $w^0{}_{g,j}$ is replaced by $w^{0*}{}_{g,j}$:

$$w_{gj}^{0*} = \alpha w_g^0 + (1-\alpha) w_{gj}^0, \tag{4}$$

where $0 \le \alpha \le 1$. The weight α is chosen to be the number between 0 and 1 that minimizes the mean square error of $w^{0*}{}_{g,j}$ and it takes into consideration covariances between population groups. The weighted average of the subpopulation and major population expenditure share estimates is expected to be a more accurate estimate of the subpopulation group's expenditure share because the subpopulation group and major population group have similar expenditure

patterns, but the major population has a larger sample size in terms of sampled households.

To minimize the variance of $w^{0*}_{g,j}$, the derivative of $Var(w^{0*}_{g,j})$ is taken with respect to α and the value of α that minimize the variance of $w^{0*}_{g,j}$ is given by:

$$\alpha=\frac{Var(w^0_{gj})-Cov(w^0_g,w^0_{gj})}{Var(w^0_{gj}-w^0_g)}. \tag{5}$$

where: $Var(w^0_{g,j})$ is the estimated variance of the expenditure share estimate of the commodity group g for the subpopulation j, $Cov(w^0_{g,},w^{0*}_{g,j})$ is the estimated covariance of the expenditure share estimate of the commodity group g for the major population group, w^0_g and the expenditure share estimate of the commodity group g for the subpopulation j, $w^0_{g,j}$, and $Var(w^0_{gj}, w^{0*}_g)$ is the estimated variance of expenditure share estimate of the commodity group g for the subpopulation group j and the expenditure share estimate of the commodity group g for the major population group.

When the MSE is minimized the formula of α becomes the following:

$$\alpha = \frac{Var(w^0_{gj}) - Cov(w^0_g, w^0_{gj})}{E\left[\left(w^0_{gj} - w^0_g\right)^2\right]}. \tag{6}$$

where: $E[(w^0_{gj,},w^0_g)^2]$ is the estimated expected squared difference of the expenditure share estimate for the population group j, and the major population group.

Multivariate procedure of this method may be used with the aim to take into account also covariances between item strata. We refer to multivariate shrinkage (composite) estimators which combine information also across subpopulations and outcome variables and which are particularly effectives when the local area level means are highly correlated, and the sample means of one or few components have small sampling and between-area variances. In this perspective, estimation for subpopulations based on small samples can be greatly improved by incorporating information from subpopulations with larger sample sizes.

2.2. Bayesian estimation methods

Model-based estimation techniques can be used to increase precision of CPI weights for domains with a small number of sampled households. In this context, Bayesian estimation techniques have received a considerable attention in many applications of small-domain estimation (Lahiri, 1992; Datta and Lahiri, 1992; Arora and Lahiri, 1997; Cohen and Sommers, 1984; Lahiri and Wang, 1992).

A linear empirical Bayes model can be used to estimate the expenditure of the commodity group *g* with reference to each population group of interest *j* as:

$$e^{o}_{gj}=\sum_{k}\beta_{gkj}x_{kj}+\varepsilon_{gj} \quad (g=1,..,G) \tag{7}$$

where e^{o}_{gj} is the expenditure of commodity class *g* relative to the population group *j*, the x_{kj}'s are *K* exogenous variables relative to the population group *j*, the β_{gkj}'s are coefficients of those variables (k=1,..,K), and ε_{gj} is a random error term.

The true expenditure of commodity class *g* relative to the population group *j* is then estimated as a weighted average of the direct survey estimator $\hat{e}^{o}_{gj}(D)$ and the model-based estimator $\tilde{e}^{o}_{gj}$:

$$\hat{\tilde{e}}^{o}_{gj}=\alpha e^{o}_{gj}(D)+(1-\alpha)\tilde{e}^{o}_{gj} \tag{8}$$

where $\alpha=\dfrac{Var(e^{o}_{gj})}{Var(e^{o}_{gj})+Var(e^{o}_{gj}(D)\left|e^{o}_{gj}\right.)}$.

This estimation strategy assumes that there are some outside variables that can be used to produce good estimates of expenditures.

In addition, a hierarchical Bayes model can be used to estimate expenditure shares. Let e^{o}_{gj} be the true total expenditure for commodity class g (g=1,..,G) in population group j (j=1,..,J), let $e^{o}_{gj}(D)$ be the direct survey estimator of the total expenditure, and let $e_{j}=\left[e_{1j},..,e_{Gj}\right]'$ and $e(D)_{j}=\left[e(D)_{1j},..,e(D)_{Gj}\right]'$ be vectors of those quantities.

Then, under the following assumptions:

1. conditional on the true expenditure e_j, the $e(D)_j$'s are independent with $e(D)_j\left|e_j\right. \square N\left(e_j,V_j\right)$;

2. conditional on $\mu \in R^k$ and $0 < r_0 < \infty$, the e_j's are independent with $e_j \left| \mu, r_0 \right. \square N\left(\mu, r_0^{-1} I\right)$;
3. the prior density function of μ and r_0 is given by: $\pi\left(\mu, r_0\right) \propto r_0^{\frac{1}{2} f-1} \exp\left(-\frac{1}{2} c r_0\right)$;
4. the values of V_j, *f*, and *c* are assumed to be known;

the posterior distribution of $e = \left[\theta_1', .., \theta_J'\right]'$ given $e(D) = \left[e(D)_1', .., e(D)_J'\right]'$ and r_0 is a multivariate normal distribution with mean vector $E\left(e \left| e(D), r_0 \right.\right) = \left[Q_1', .., Q_J'\right]'$,

where $Q_J = \left(I - V_J W_J\right) e(D)_J + V_J W_J \left(\sum_{j=1}^{J} W_J \right)^{-1} \sum_{j=1}^{J} W_{J e(D)_J}$, and where $W_J = \left(V_J + r_0^{-1} I\right)^{-1}$.

The hierarchical Bayes estimator of the true total expenditure for commodity class *g*, e_{gj}^{o}, is then obtained using the known iterated formula:

$$\hat{e} = E\left(e \left| e(D)\right.\right) = E\left[E\left(e \left| e(D), r_0 \right.\right) \left| e(D) \right.\right] \tag{9}$$

where the expectation is computed using the posterior density of r_0 given *e(D)*, which is:

$$g\left(r_0 \left| e(D)\right.\right) \propto \left| \sum_{j=1}^{J} W \right|^{\frac{1}{2}} \prod_{j=1}^{J} \left|W_j\right|^{\frac{1}{2}} \exp\left[-\frac{1}{2} \left(e r_0 + \sum_{j=1}^{J} e(D)_j' W_j e(D)_j - \left(\sum_{j=1}^{J} W_j e(D)_j \right)' \left(\sum_{j=1}^{J} W_j \right)^{-1} \sum_{j=1}^{J} W_j e(D)_j \right) \right]. \tag{10}$$

From a general point of view, the properties of the model-based estimators depend on both (i) the sampling design of the Households Expenditure Surveys, and (ii) on the model that is assumed to have generated individual expenditures. Finally, model-based estimation techniques may result more complex to implement.

3. The generalized maximum entropy estimation procedure

In this section we introduce a Generalized Maximum Entropy (GME) estimator of expenditure shares for the computation of CPI weights in small groups of the reference population. The idea is to introduce a model which is

coherent with the empirical evidence and with the incomplete data information while displaying consistency with consumer theory assumptions.

We start by considering a nonlinear version of a complete demand system which consists of a set of budget-share equations relating the budget share w^0_{ig}, of the class of goods g (g=1,..,G) by household i (i=1,..,H), to the logarithm of total expenditure e_i, of household i, and to a set of K household-specific demographic and economic variables x_{ki}, (k=1,..,K) (Deaton-Muellbauer, 1980).

Assuming that: (i) demographic household characteristics are used to capture the effect of household preferences on demand; (ii) the household-specific price level is identical for all households so that all households face the same increases for each class of goods, the share equations are given by:

$$w^0_{ig} = \sum_k \beta_{gk} x_{ki} + \alpha_g \ln e_i + \varepsilon_{ig}, \tag{11}$$

where β_{gk} and α_g are unknown parameters, and ε_{ig} are error terms.

Moreover, to account for the presence of zero expenditures, we then introduce a system of censored demand equations:

$$\begin{cases} w^0_{ig} = \sum_k \beta_{gk} x_{ki} + \alpha_g \ln e_i + \varepsilon_{ig}, & w^0_{ig} > 0, \\ w^0_{ig} > \sum_k \beta_{gk} x_{ki} + \alpha_g \ln e_i + \varepsilon_{ig}, & w^0_{ig} = 0. \end{cases} \tag{12}$$

Our objective is to simultaneously recover the signal (β,α) and the noise ε without imposing distributional assumption and assumptions regarding the exact relationship between sample and population moments. To this end, using the GME reparameterization we express all coefficients and errors in the form of probabilities and we assume that the signal and the noise can be expressed as linear combinations of the unknown probabilities (p^β, p^α) and r respectively, as follows:

$$\begin{cases} w^0_{ig} = \sum_k \sum_m z^\beta_{gkm} p^\beta_{gkm} x_{ki} + \sum_m z^\alpha_m p^\alpha_{gm} \ln e_i + \sum_s v_s r_{igs} & w^0_{ig} > 0 \\ w^0_{ig} > \sum_k \sum_m z^\beta_{gkm} p^\beta_{gkm} x_{ki} + \sum_m z^\alpha_m p^\alpha_{gm} \ln e_i + \sum_s v_s r_{igs} & w^0_{ig} = 0 \end{cases} \tag{13}$$

The GME estimator is obtained by maximizing the objective function under:

(i) a set of constraints for budget shares equations;

(ii) additivity constraints for unknown probabilities;

(iii) consumer-theory restrictions.

We specify a set of support points for each unknown parameter ($m=1,..,M$) and error (s=$1,..,S$) and use maximum entropy to estimate the unknown probabilities associated with the support points. The parameter support is based on prior information or economic theory while for the error term a symmetric representation centered on zero is used, assuming a discrete uniform distribution.

The error support is also defined according the "three sigma" rule of Pukelsheim (1994). Alternately, an "Empirical Bayes" type procedure can be used (Efron-Morris (1973), Casella (1985)). Assuming the unknown weights on the parameters and the errors supports are independent, we jointly estimate the unknown parameters and errors by solving the following constrained optimization problem:

$$\text{Max}_{p,r} \quad \text{H}(p,r)=-p'\ln p'-r'\ln r' \tag{14}$$

with $p=(p^{\beta},p^{\alpha})$,
subject to:

budget-share equations:

$$\begin{cases} w_{ig}^{0}=\sum_{k}\sum_{m} z_{gkm}^{\beta} p_{gkm}^{\beta} x_{ki}+\sum_{m} z_{m}^{\alpha} p_{gm}^{\alpha} \ln e_i+\sum_{s} v_s r_{igs} & w_{ig}^{0}>0 \\ w_{ig}^{0}>\sum_{k}\sum_{m} z_{gkm}^{\beta} p_{gkm}^{\beta} x_{ki}+\sum_{m} z_{m}^{\alpha} p_{gm}^{\alpha} \ln e_i+\sum_{s} v_s r_{igs} & w_{ig}^{0}=0 \end{cases} \tag{15}$$

additivity constraints

$$\sum_{m} p_{gkm}^{\beta}=\sum_{m} p_{gm}^{\alpha}=1; \tag{16}$$

consumer-theory restrictions (adding-up and homogeneity conditions, and the restrictions for the shares to add to one)

$$\sum_{g}\beta_{g1}=1, \quad \sum_{g}\beta_{gk}=0,(k=2,..,K) \qquad \sum_{g}\sum_{s} v_s r_{igs}=0, \quad \sum_{s} r_{igs}=1. \tag{17}$$

Forming the Lagrangean and solving for the first-order conditions yields the optimal solution $\hat{p}_{gkm}^{\beta}$, $\hat{p}_{gm}^{\alpha}$ and $\hat{r}_{igj}$, from which we derive the following GME point estimates:

$$\hat{\beta}_{gk}=\sum_{m} z_{gkm}^{\beta}\hat{p}_{gkm}^{\beta}, \qquad \hat{\alpha}_{g}=\sum_{m} z_{m}^{\alpha}\hat{p}_{gm}^{\alpha}, \qquad \hat{\varepsilon}_{ig}=\sum_{s} v_s \hat{r}_{igs}. \tag{18}$$

In this formulation we introduce a balanced approach which gives equal importance to precision and prediction objectives.

The GME estimation procedure appears useful in the following regards. First, goods results are produced in the case of small-sized samples. Second, restrictions expressed in terms of inequality can be introduced to deal with the problem of zero expenditures. Third, it is possible to estimate systems with a large number of demand equations. This approach also present several advantages over the traditional maximum likelihood methods: (i) it is not necessary to introduce hypotheses regarding the form of the error term distribution; (ii) it is possible to introduce demand relationships that are subject to a high degree of collinearity amongst the explanatory variables. Estimates are efficient, consistent and robust

when the error term distribution is not normal and the explanatory variables are highly correlated (Golan et al., 1997).

4. Decomposition of the total relative discrepancy between indices for small reference populations

In this section we introduce a procedure for decomposing the relative difference between two price indexes obtained from alternative reference populations. The objective is to investigate the effect of varying consumption patterns for subpopulation groups on the measured relative differences of their price indexes with respect to a reference price index.

Consider two price indexes in period t, I^t_C, and I^t_R, computed with reference to the subgroup and the reference populations, respectively.

The relative difference between the "comparison index" relative to a specific subpopulation group, I^t_C, and the reference index, I^t_R, can be decomposed into the separate impacts of the expenditure weights and of the rate of commodity-specific price increases relative to the average.

Under the assumption, introduced for the computation of the special group indexes, that prices are similar for both the comparison and reference population groups, $(p^C_\cdot = p^R_\cdot)$, the following total relative discrepancy between the two indexes, TRD_1, is obtained:

$$
\begin{aligned}
TRD_1 &= \left(\frac{I_C^t - I_R^t}{I_R^t}\right)\times 100 = \left(\frac{100}{I_R^t}\right)\left(I_C^t - I_R^t - I_R^t + I_R^t\right) = \\
&= \frac{100}{I_R}\sum_{g=1}^{G}\left[\left(\frac{p_{gt}}{p_{g0}}\right)\frac{p_{i0}q^C_{i0}}{\sum_i p_{i0}q^C_{i0}}\right] - \left[\sum_{g=1}^{G} I_{Rt}\frac{p_{i0}q^C_{i0}}{\sum_i p_{i0}q^C_{i0}}\right] - \\
&- \left[\sum_{g=1}^{G}\left(\frac{p_{gt}}{p_{g0}}\right)\frac{p_{i0}q^R_{i0}}{\sum_i p_{i0}q^R_{i0}}\right] + \left[\sum_{g=1}^{G} I_{Rt}\frac{p_{i0}q^R_{i0}}{\sum_i p_{i0}q^R_{i0}}\right] = \\
&= \frac{100}{I_R}\sum_{g=1}^{G}\left[\left(\frac{p_{gt}}{p_{g0}}\right) - I_{Rt}\right]\left[\frac{p_{i0}q^C_{i0}}{\sum_i p_{i0}q^C_{i0}}\frac{\sum_i p_{i0}q^R_{i0}}{p_{i0}q^R_{i0}}\frac{p_{i0}q^R_{i0}}{\sum_i p_{i0}q^R_{i0}} - \frac{p_{i0}q^R_{i0}}{\sum_i p_{i0}q^R_{i0}}\right] = \\
&= \sum_{g=1}^{G}\left[\frac{\left(\frac{p_{gt}}{p_{g0}}\right)}{\frac{\sum_i p_{it}q^R_{i0}}{\sum_i p_{i0}q^R_{i0}}} - 1\right]\left[\frac{\frac{p_{i0}q^C_{i0}}{\sum_i p_{i0}q^C_{i0}}}{\frac{p_{i0}q^R_{i0}}{\sum_i p_{i0}q^R_{i0}}} - 1\right]\left[\frac{p_{i0}q^R_{i0}}{\sum_i p_{i0}q^R_{i0}}\times 100\right] = \\
&= \sum_{g=1}^{G}\left[\frac{I_{gt}}{I_{Rt}} - 1\right]\left[\frac{w_{gC}}{w_{gR}} - 1\right]\left[w_{gR}\times 100\right],
\end{aligned}
\tag{19}
$$

where g=*1*,..,*G* and *i=1*,..,*I* identify the commodities (commodity group) and are indices over the same goods *G=I*, $I_{g,t}$ is the component-specific price index for commodity *g* in period *t*, and w_{gc} (w_{gr}) refers to the weight of commodity *g* in the comparison (reference) population group.

If there is some difference between the two population group in the relative importance of commodity *g*, its impact on the total discrepancy TRD_1 will be influenced by the relative divergence between the rate of change in the price movement of commodity *g* and the average rate change, I^t_R, as well as by the difference in the relative importance of commodity *g*.

This relative discrepancy is decomposed into two separate components: (i) the relative differences between the rates of price change of each commodity and the average rate of change, and (ii) the relative differences in the weights of each commodity in the two reference populations.

The decomposition can be performed at a low level of expenditure detail or at a higher level of expenditure aggregation for the main commodity groups.

The procedure here presented is important in determining: (i) which expenditure items (group of items) cause the comparison index to be relatively larger or smaller than the reference average household's; and (ii) the cause of each item's (group of item's) positive or negative contribution in terms of differential price change or weight difference. This analysis can reveal which

items are the most important determinants of the divergence between indexes and also the direction of their contribution by showing which items contribute most toward raising the CPI for the reference population relative to the CPI for the subgroup population.

Moreover, the classification of the expenditures in terms of their total impact on the divergence of the two index, and in terms of the cause of their impact is important to policymakers to focus the attention on specific commodities whose price increase can most severely impact the subgroup of the reference population.

5. An application of the proposed estimation procedure

The GME estimation technique is applied to estimate weights of regional consumer price indexes for subgroups of the reference population.

The regional consumer price index is calculated for the Umbria region by using monthly data over the period January to June 2001 with regard to the 2000 basket and prices, and by assuming the month of December 2000 as the time base. The price indices for each product in different four cities (Perugia, Terni, Città di Castello, Orvieto) are combined in the regional price index for product using weights which represent relative importance of the expenditure in each cities and since expenditure data are not available at city level, population weights are used. Then, the price indices for all products are combined to obtain the regional index using weights which represent the relative importance of the expenditure for each product in the region. The data base used to estimate the expenditure shares for subgroups of the reference population are derived from the Italian Expenditures Survey and contains detailed information about the expenditure, together with a great number of household characteristics of a sample of households in 2000.

In synthesis, the regional index based on the Eq. (1), is computed by calculating: *(i)* city indices of product; *(ii)* regional indices of product, as weighted average of the elementary city indices of product with weights equal to the population in the cities; *(iii)* regional index, as weighted average of the regional indices of product with weights equal to the expenditures in the region. Weights are obtained from a Family Budget Survey and from other sources in combination with National Accounts data on regional household consumption within the Italian economic territory.

Results of the computation of the Consumer Price Index for the Umbria region are presented in Table 1.

From January to June 2001 the CPI for the Umbria region rose 1.9%; it rose at a slower pace than the Italian CPI and it is above the Italian CPI for all the period; the Italian CPI seems to have escalated at a faster rate than the RCPI for all main consumption groups. Over the entire period, the greatest price increases at the main component level are in Food and Beverages, Alcohol and Tobacco, Transportation, Restaurant and Hotels, while the greatest declines are for Housing and Communication.

An examination of Table 2 shows that the subgroups' weights at the level of *Housing, Communication,* and *Medical care* are higher than those of the reference population for low-income households, nokids and single type families. In the case of high-income households, the relative share of total expenditures for the *Clothing, Furnishings* and *Transportation* categories are 8.2, 9.5 and 15.8 percent as compared to 7.5, 7.8 and 14.6 percent for the reference population and are 9.3%, 22% and 8.2% higher. Low-income households' average expenditures on Food, Housing, and Medical care categories are 23.1, 38.2 and 3.7 percent as compared to 17.98, 27.84 and 3.49 percent for the reference population and are 28.5%, 37.2% and 6.0% higher.

After establishing that the consumption expenditure of some special groups of the reference population is quite different from the average expenditure of the general population (see Tab. 2), we also compute the Regional Consumer Price Index for some specific subgroups of the reference population: *(i)* low-income household (FamL), *(ii)* high-income household (FamH), *(iii)* one-person type family (Single), and *(iv)* family without children (Nokids). To make them comparable, the special group indices are computed using the same methodology as for the reference population index. As said in section 1, each of the indices is computed using a set of expenditure weights representing different basket, but all are drawn from the same reference period and the price series used for the subgroups are based on regional data collected for the Regional Consumer Price index, RCPI.

Table 1. Consumer Price Index of the Umbria region (December 2000=100, January—June 2001)

Expenditure Category	Jan-01	Feb-01	Mar-01	Apr-01	May-01	Jun-01
Food and Beverage	101.1	101.6	102.1	102.6	103.3	103.7
Alcoholic bev. Tobacco	100.1	100.1	100.3	103.2	103.2	103.3
Clothing and Footwear	100.0	100.0	100.3	100.7	100.9	100.9
Housing. Fuels	100.1	99.7	99.9	100.1	98.8	98.7
Household furnishing	100.1	100.6	100.7	100.8	101.8	101.8
Medical care	100.6	100.6	100.6	100.7	100.7	100.7
Transportation	99.2	99.4	99.5	99.9	101.0	101.2
Communication	99.5	99.2	99.1	99.0	98.7	98.7
Recreation	102.0	102.0	102.0	102.1	102.3	102.2
Education	100.0	100.0	100.0	100.0	100.0	100.0
Hotels. Restaurants	102.7	103.2	103.5	105.1	105.2	105.4
Others goods and serv.	101.1	101.2	101.4	102.0	102.1	102.1
All items	100.6	100.8	101.0	101.5	101.8	101.9

Table 2. Distribution of total average expenditures for the major components of the Consumer Price Index **(**Umbria region. relative share in percentage. Household Survey 2000)

Expenditure Category	W Umbria	W H-F	W L-F	W Nokids	W Single
Food and Beverage	18.0	16.1	23.1	18.4	17.4
Alcoholic bev. Tobacco	1.6	1.5	1.9	1.6	1.5
Clothing and Footwear	7.6	8.3	5.6	7.5	6.7
Housing. Fuels	27.8	24.2	38.2	29.5	34.9
Household furnishing	7.8	9.5	3.0	7.0	4.5
Medical care	3.5	3.4	3.7	3.9	4.8
Transportation	14.6	15.8	11.1	13.8	11.2
Communication	2.6	2.4	3.1	2.6	3.1
Recreation	5.3	6.0	3.4	5.2	5.5
Education	0.3	0.3	0.2	0.2	0.3
Hotels. Restaurants	2.4	2.7	1.7	2.3	3.4
Others goods and serv.	8.6	9.7	5.2	8.0	6.8
All items	100	100	100	100	100

The proposed Generalized Maximum Entropy procedure is also applied to estimate the expenditure weights for the selected subpopulations. A system of the budget-share equations relative to the following expenditure categories: *Food* and *Beverages, Alcohol and Tobacco, Clothing* and *Footwear, Housing, Furnishings* and *Household operations, Medical care, Transportation, Communications, Recreations, Educations, Restaurant* and *Hotels, Other goods* and *Service* is estimated by solving the constrained optimizing problem (14). The explanatory variables introduced into the model are divided into four groups: *(i)* demographic variables (number of kids, number of components, household has zero children, head of household aged 0-24, head of household aged 45-64, head of household aged >64, household has more than one child, household consists of a single adult, household has one child, sex of head of household); *(ii)* economic variables (per capita total household expenditure (expressed in logarithmic form), household with no working members, households having a total expenditure below the median of the distribution of all sampled households); *(iii)* occupational dummies for two standard classifications (unoccupied, retired and out of labor force); *(iv)* indicators for the presence of stocks (households pay to rent the house, households make use of the house but do not pay for rent).

Results of the computation of sub-indexes for: (i) low-income household (FamL), *(ii)* high-income household (FamH), *(iii)* one-person type family (Single), *and (iv)* family without children (Nokids), *are reported in Figure 1 and 2.*

Even if the differences in weights between each special group and the reference population are significant at regional level, an analysis of the movements of the corresponding indices indicates that, nevertheless, differences of indices with respect to the RCPI are less evident. This is partly due to the relationships between relative prices and weights that exist simultaneously and therefore tend to offset each other.

From January to June 2001 all the indexes stayed within one percentage point of each others; the indexes for famH, famL, nokids, single rose 1.2%, 0.8%, 1.1%, 0.9%, respectively. During this period the sub-indexes rise more slowly than the regional consumer price index; they follow the same trend but their paths move apart.

From January to April 2001, all the indexes moved up at a very similar rate, their paths tend to merge; starting in April the famH and Umbria indexes rice more slowly but the other indexes decrease.

The rise in the H-fam index closely follows that for the Umbria reference population and it is all of the time lower, especially at the end of the period. The famH index has the highest monthly increase with respect to the other sub-indexes while the famL index follows the movement of the single index.

The importance of these index differences in euro terms emerges if we deflate mean family income using the arithmetic mean over six months of each index. For example, deflating income with the national (regional) CPI yields a real income of 38959 (44381) while deflating with the index for families without children yields a real income of 44646 euro. The result is a -5687 (-265) difference in purchasing power, or approximately a -12.7 (-0.6) percent difference. The results for the other indexes are between a difference of -0.5 (-12) and -1 (-13) percent.

Figure 1. Consumer Price Indexes for the subgroups of the population (Umbria region. January—June 2001)

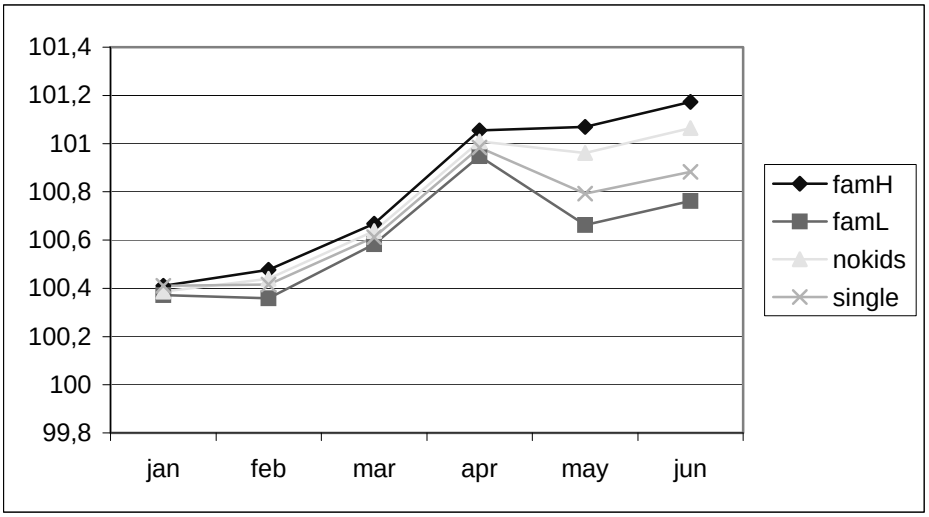


Figure 2. Relative change of the Consumer Price Indexes for the subgroups of the population (Umbria region. January—June 2001)

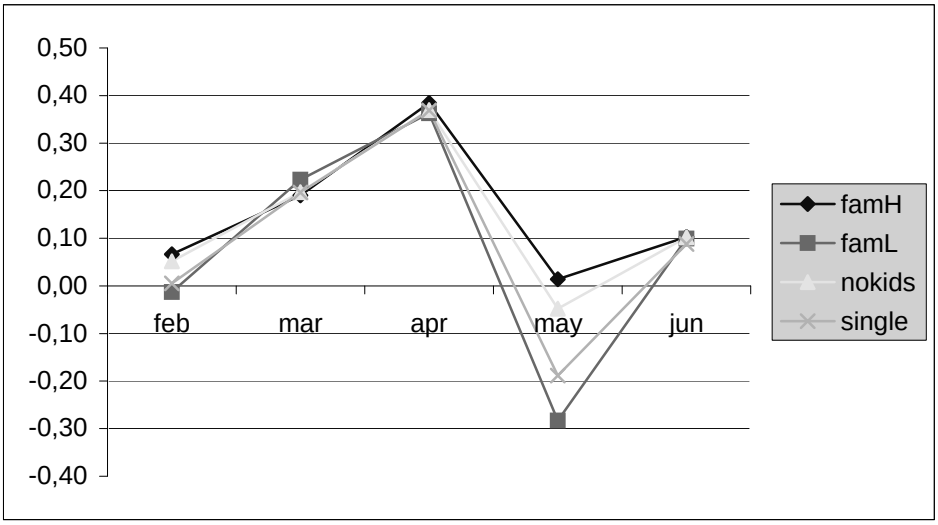


Table 3. Classification of Expenditure Categories in terms of the source of their impact on the relative discrepancy (between CPI-R and CPI-FamH)

Expenditure weight in CPI-FamH relative to weight in CPI-R	*Price increase of the expenditure Category relative to CPI-R*	
	Higher	Lower
Higher	*Positive contribution to difference*	*Negative contribution to difference*
	Food and Beverages	Alcohol and Tobacco
	Medical care	Clothing and Footwear
	Recreations	Housing
	Restaurant and Hotels	Furnishings and Household operations
	Other goods and Service	Transportation
		Communications
		Educations
Lower	*Negative contribution to difference*	*Positive contribution to difference*

Table 4. Classification of Expenditure Categories in terms of the source of their impact on the relative discrepancy (between CPI-R and CPI-FamL)

Expenditure weight in CPI-FamL relative to weight in CPI-R	*Price increase of the expenditure Category relative to CPI-R*	
	Higher	Lower
Higher	*Positive contribution to difference*	*Negative contribution to difference*
Lower	*Negative contribution to difference*	*Positive contribution to difference*
	Food and Beverages	Alcohol and Tobacco
	Medical care	Clothing and Footwear
	Recreations	Housing
	Restaurant and Hotels	Furnishings and Household operations
	Other goods and Service	Transportation
		Communications
		Educations

Table 5. Classification of Expenditure Categories in terms of the source of their impact on the relative discrepancy (between CPI-R and CPI-Nokids)

Expenditure weight in CPI-Nokids relative to weight in CPI-R	*Price increase of the expenditure Category relative to CPI-R*	
	Higher	Lower
Higher	*Positive contribution to difference*	*Negative contribution to difference*
	Food and Beverages	Housing
	Medical care	
Lower	*Negative contribution to difference*	*Positive contribution to difference*
	Recreations	Alcohol and Tobacco
	Restaurant and Hotels	Clothing and Footwear
	Other goods and Service	Furnishings and Household operations
		Transportation
		Communications
		Educations

Table 6. Classification of Expenditure Categories in terms of the source of their impact on the relative discrepancy (between CPI-R and CPI- Single)

Expenditure weight in CPI-Single relative to weight in CPI-R	*Price increase of the expenditure Category relative to CPI-R*	
	Higher	Lower
Higher	*Positive contribution to difference*	*Negative contribution to difference*
Lower	*Negative contribution to difference*	*Positive contribution to difference*
	Food and Beverages	Alcohol and Tobacco
	Medical care	Clothing and Footwear
	Recreations	Housing
	Restaurant and Hotels	Furnishings and Household operation
	Other goods and Services	Transportation
		Communication
		Education

In order to determine which expenditure groups are driving the difference between the sub-indexes and the reference index, an analysis based on the decomposition of the relative difference among CPIs indexes obtained from alternative groups of the reference population is performed.

The regional consumer price index, RCPI, is used as the "reference" index and the relative difference between it and each small population subgroup, from January to June 2001, are calculated by using equation (19). Results are summarized in Tabb. 3-6. The upper-left and lower-right quadrants of each table show the categories which tend to raise the index of the population subgroup relative to the reference regional price index. In the lower-left and upper-right quadrants are identified the expenditure categories which tend to lower the index of population subgroup relative to that of the average household.

An examination of the tables 3-6 reveals that, except for single and low-income household subgroups, differences emerge for the other population subgroups CPIs.

More specifically, for the high-income household group (FamH), the categories which contribute most toward raising the subgroup index are: *Food* and *Beverages, Medical care, Recreations, Educations, Restaurant* and *Hotels, Other goods* and *Service.* In all of the cases, these categories are relatively more important in the high-income households' budget and their prices have been rising

at a relatively faster rate than the average rate of price change. All of the other categories tend most to lower the CPI-FamH relative to the RCPI.

For the single and low-income population subgroups, (Single and FamL), *Food* and *Beverages, Medical care, Recreations, Educations, Restaurant and Hotels, Other goods* and *Service*, whose prices have increased more rapidly than the average rate, are relatively less important in the subpopulation' budget and they contribute negatively toward the difference. All of the other categories contribute positively toward the difference but while the categories are relatively less important in the subpopulation' budget they have increased at a relatively slower rate than the average.

Finally, for the subgroup of households without children (Nokids), the categories Food and Beverages, Medical care, Alcohol and Tobacco, Clothing and Footwear, Furnishings and Household operations, Transportation, Communications and Education contribute most toward raising the difference between the two CPIs but the reasons for this positive contribution are not the same. In the case of Food and Beverages, Medical care, not only they are more important in the subpopulation' budget but their prices have increased more rapidly than the average rate. For the other categories the opposite is true: while they are relatively less important in the budget of the households without children they have increased at a relatively slower rate than the average. All of the other categories contribute negatively toward the difference but Recreations, Restaurant and Hotels, Other goods and Service are relatively less important in the subpopulation' budget, and they have increased at a relatively slower rate than the average while Housing, whose prices have increased more rapidly than the average rate, is relatively more important in the subpopulation' budget.

6. Concluding remarks

In this paper we have been concerned with the problem of estimating accurate item weights of Consumer Price Indexes that are computed for small reference populations. The methodology of computing the special group indices here presented is based on the following assumptions: (i) outlets selected are identical for all indices so that they do not necessarily reflect the consumption habits of each subgroup; (ii) the basket of goods and services is the one consumed by the general population, items with zero expenditure are removed but items that may specifically represent the consumption of each subgroup are not considered; (iii) the prices used are identical to the prices for the general population; prices tend to follow the same trend in a local area so that relative prices do not greatly vary from one group to another.

A Generalized maximum entropy model-based estimator has been proposed to estimate expenditure shares of small subgroups of the reference population. The GME formalism represents a new basis to recover the unknown parameter and the unknown variables when: (i) surveys used to adjust weights for the special groups

population are sensitive to sampling and response errors; (ii) and the amount of sample units used in determining the weights is significantly reduced. Within this framework, minimal distributional assumption are necessary and a dual loss function is used to take into account both the estimation precision and prediction objectives.

A procedure for decomposing the relative discrepancy between the subgroup and the reference population indexes has been also presented in order to isolate the impact of each component of consumption to the difference in indexes. Using this analysis it is possible to identify the commodities whose prices have been increasing at a relatively higher rate than the average and to verify if they contribute positively or negatively to the differences in indexes. These evidences can provide a suitable basis for policymakers to identify items or groups of commodities that can produce the differential price experiences of subgroups of the reference population.

The working of the proposed estimation procedure has been illustrated by applying it on the computation of regional consumer price indexes for small population subgroups.

Our results have shown the differences in group specific price indexes and the effect of those differences in analyzing relative discrepancy of each subpopulation index and the whole population price index. Over the entire period analyzed (January-June 2001), the results of our application seem to have shown that for the high-income households the price escalation rose at a higher rate than price change for the low-income households. In this direction, additional work needs to be conducted in order to determine whether the CPI for low-income households will remain lower in the long run.

An analysis based on the decomposition of the relative difference among the subpopulation indexes and the regional index have indicated that: (i) except for single and low-income household subgroups, differences emerge for the other population subgroups CPIs; (ii) different specific expenditure groups are driving the difference between each sub-indexes and the reference index.

The GME estimation procedure seem to be a promising alternative model-based estimation technique because it is easy to implement, it does not depend on any hypotheses regarding the form of the error distribution in the model, and it produces goods results for small-sized samples. In particular, theoretical and other nonsample information may be directly imposed on the GME estimates much easier than with classic Maximum likelihood and Bayesian estimation techniques. Further work is needed to: (i) investigate the effect of the introduction of different functional forms for the commodity groups into our GME formulation, (ii) and to combine direct estimates with our GME model-based estimates by defining a composite generalized maximum entropy estimator (Bernardini Papalia, 2005) in order to obtain a bias lower than the bias of the model-based estimator, and a standard error lower than the standard error of the direct estimator.

REFERENCES

ARORA V., and P. LAHIRI (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica, 7*, 1053—1063.

BOON M., and J. DE HAAN (1997). Estimating Consumer Price Indices for small reference populations. J*ournal of Official Statistics, 13, 2,* 143—158.

BERNARDINI PAPALIA R. (2005). *A Composite Generalized Cross Entropy formulation in small samples estimation*, Proceedings of "The 2nd Conference on Information and Entropy Econometrics: Theory, Methods, and Applications", Washington DC.

CASELLA G., (1985). An introduction to empirical Bayes data analysis. *The American Statistician 39*, 83—87.

COHEN M. P., and J.P. SOMMERS (1984). *Evaluation of methods of composite estimation of cost weights for the CPI*. Proceedings of the American Statistical Association, 466—471.

DATTA G. S., and P. LAHIRI (1992). *Robust Hierarchical Bayes Estimation of small area characteristics in presence of covariates*. Technical Report, University of Georgia.

DEATON A. and J. MUELLBAUER (1980). An Almost Ideal Demand System. *American Economic Review 70*, 312—326.

EFRON B., and C. MORRIS (1973). Stein's estimation rule and its competitors-an empirical Bayes approach. *Journal of the American Statistical Association 68,* 117—130.

GOLAN A., JUDGE G., and D. MILLER (1996). *Maximum entropy econometrics: robust estimation with limited data*, Wiley.

LAHIRI P. (1992). *Estimation of Consumer expenditures for small areas: the hierarchical Bayes approach*. Technical Report, University of Nebraska.

LAHIRI P., and W. Wang (1992). A Multivariate Procedure towards composite Estimation of Consumer Expenditure for the US Consumer Price Index Numbers. *Survey Methodology,18.2*, 279—292.

OWEN A. B., (1991). Empirical Likelihood for linear models. *The Annals of Statistics 19*, 1725—1747.

PRAIS S. (1959). Whose Cost of Living?. *Review of Economic Studies*, 26, 126—134.

PUKELSHEIM F. (1994). The Three Sigma Rule. *American Statistician* 48, 88—91.

SARNDAL C. E., SWENSSON B., and J. WRETMAN (1992). *Model assisted Survey Sampling*. New York, Springer Verlag.

TURVEY R. (2002). CPI Manual. True Cost of Living Indexes.

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 71—87

A GENERAL PROCEDURE OF ESTIMATING POPULATION MEAN USING AUXILIARY INFORMATION IN SAMPLE SURVEYS

Housila. P. Singh and Neha Agnihotri

ABSTRACT

In this paper we have suggested a general procedure for estimating the population mean through defining a class of estimators. A large number of estimators are identified as members of the suggested family. The expressions for bias and mean squared error (MSE) of the suggested family of estimators have been obtained under large sample approximation. Asymptotic optimum estimator (AOE) in the class is identified with its approximate MSE formula. It is observed that the proposed class of estimators attains its minimum MSE when the exact optimum value of the scalar(s) involved in the estimator is known. In practice the optimum value of the scalar depends on the unknown population parameters and hence the optimum estimator. It restricts the use of the optimum estimator in practice. Keeping this in view we replace the unknown population parameters (involved in the optimum value of the scalar) by their consistent estimators obtained from the sample to obtain an estimator based on estimated optimum value. It has been shown to the first degree of approximation that the MSE of the estimator based on the estimated optimum value is same as the minimum MSE of the proposed class of estimators (or the variance of the optimum estimator in the class). Numerical illustrations are given in support of the present study. Both theoretical and numerical findings are encouraging and useful in practice.

Key words: Study variate, Auxiliary variate, Bias, Mean squared error.

1. Introduction

The problem of estimating the population mean in the presence of an auxiliary variable has been widely discussed in finite population sampling literature. Out of many ratio, product, difference and regression methods of estimation are good examples in this context. It is well known result that the ratio estimator is the best among a wide class of estimators when the relation between y and x, the variate under study and the auxiliary variate respectively, is a straight line through the

origin and the variance of y about this line proportional to x [Cochran (1977)]. In many practical situations the regression line does not pass through the origin. This fact led various authors to propose modified ratio and product estimators for instance see, Singh (1965, 67), Srivastava (1967, 1971, 1980), Walsh (1970), Reddy (1973, 74), Sahai (1979), Sahai and Ray (1980), Vos (1980), Srivenkatramana and Tracy (1981), Sahai and Sahai(1985), Naik and Gupta (1991) and Upadhyaya and Singh (1999) etc.

In this paper we have suggested a general procedure for estimating the population mean through defining a class of estimators. A large number of estimators are identified as members of the suggested family. The properties of the suggested family are studied under large sample approximation. Empirical studies are carried out to judge the merits of the suggested estimators over conventional estimators.

2. The proposed family of estimators

Let the population consist of N identified sampling units, the i[th] unit labeled as $U_i\,(i=1,2,...,N)$. Let (y,x) be the character under study and the auxiliary character, respectively. Further, let y_i be the unknown real variable value of $y\,and\,x_i$ be the known variable value of x associated with $U_i\,(i=1,2,...,N)$. Let $(\bar{Y},\bar{X})$ be the population means of the variates (y,x) respectively. It is assumed that the population mean $\bar{X}$ of the auxiliary variate x is known. It is desired to estimate the population mean $\bar{Y}$ of the study variate y using information on the population mean $\bar{X}$ of the auxiliary variate x. Let a sample of size n be drawn from population using simple random sampling without replacement (SRSWOR). We define a family of ratio-product estimators of population mean $\bar{Y}$ as

$$T_{RP}=\delta\bar{y}\left(\frac{a\bar{X}+b}{a\bar{x}+b}\right)+(1-\delta)\bar{y}\left(\frac{a\bar{x}+b}{a\bar{X}+b}\right), \qquad (2.1)$$

where $\left(\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,\ \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\right)$ are unbiased estimators of the population means $(\bar{Y},\bar{X})$respectively, 'a' and 'b' are known characterizing positive scalars and δ is a real constant to be determined such that the mean squared error of T_{RP} is minimum. The family of estimators T_{RP} reduces to the following set of known estimators:

(i) for $(a,b,\delta)=(0,1,\delta)$, $T_{RP} \to \bar{y}$ (usual unbiased estimator),

(ii) for $(a,b,\delta)=(1,0,\delta)$, $T_{RP} \to T_1 = \left[\delta \bar{y}\left(\frac{\bar{X}}{\bar{x}}\right)+(1-\delta)\bar{y}\left(\frac{\bar{x}}{\bar{X}}\right)\right]$

which is due to Singh and Ruiz Espejo (2003), and

(iii) for

$$(a,b,\delta)=(1,C_x,\delta),\ T_{RP} \to T_2 = \left[\delta \bar{y}\left(\frac{\bar{X}+C_x}{\bar{x}+C_x}\right)+(1-\delta)\bar{y}\left(\frac{\bar{x}+C_x}{\bar{X}+C_x}\right)\right] \qquad (2.2)$$

envisaged by Singh and Tailor (2005), where C_x is the known population coefficient of variation C_x of the auxiliary variable x. Many other ratio-product estimators can be generated from T_{RP} by putting suitable values of (a,b,δ).

To obtain the bias and mean square error (MSE) of the proposed family of estimators T_{RP} in (2.1), we write

$$\bar{y} = \bar{Y}(1+e_0),\ \bar{x} = \bar{X}(1+e_1)$$

such that

$$E(e_0) = E(e_1) = 0$$

and

$$E(e_0^2) = \frac{(1-f)}{n}C_y^2,\ E(e_1^2) = \frac{(1-f)}{n}C_x^2,\ E(e_0e_1) = \frac{(1-f)}{n}KC_x^2,$$

where $f = \frac{n}{N}, K = \rho\frac{C_y}{C_x}, \rho = \frac{S_{xy}}{S_xS_y}, C_y = \frac{S_y}{\bar{Y}}, C_x = \frac{S_x}{\bar{X}},$

$$S_{xy} = \frac{\sum_{i=1}^{N}(x_i-\bar{X})(y_i-\bar{Y})}{(N-1)}, S_x^2 = \frac{\sum_{i=1}^{N}(x_i-\bar{X})^2}{(N-1)}, S_y^2 = \frac{\sum_{i=1}^{N}(y_i-\bar{Y})^2}{(N-1)}.$$

Expressing (2.1) in terms of e's, we have

$$T_{RP} = \bar{Y}(1+e_0)\left[\delta(1+\theta e_1)^{-1} + (1-\delta)(1+\theta e_1)\right], \qquad (2.3)$$

where $\theta = \frac{a\bar{X}}{(a\bar{X}+b)}$.

We assume that $|\theta e_1| < 1$ so that $(1+\theta e_1)^{-1}$ is expandable. From (2.3) we have

$$T_{RP} = \bar{Y}(1+e_0)\left[\delta(1-\theta e_1+\theta^2 e_1^2-\theta^3 e_1^3+\theta^4 e_1^4-\ldots)+(1-\delta)(1+\theta e_1)\right]$$

$$= \bar{Y}\left[\delta(1+e_0)(1-\theta e_1+\theta^2 e_1^2-\theta^3 e_1^3+\theta^4 e_1^4-\ldots)+(1-\delta)(1+e_0)(1+\theta e_1)\right]$$

$$= \bar{Y}\left[\delta(1+e_0-\theta e_1+\theta^2 e_1^2-\theta e_0 e_1+\theta^2 e_0 e_1^2-\theta^3 e_1^3+\theta^4 e_1^4-\theta^3 e_0 e_1^3+\ldots)\right.$$
$$\left.+(1-\delta)(1+e_0+\theta e_1+\theta e_1 e_0)\right]$$

$$= \bar{Y}\left[1+e_0+\theta e_1+\theta e_0 e_1\right.$$
$$\left.+\delta(1+e_0-\theta e_1-\theta e_0 e_1+\theta^2 e_1^2+\theta^2 e_0 e_1^2-\theta^3 e_1^3+\theta^4 e_1^4-\theta^3 e_0 e_1^3+\ldots-1-e_0-\theta e_1-\theta e_1 e_0)\right]$$
$$= \bar{Y}\left[1+e_0+\theta e_1+\theta e_0 e_1\right.$$
$$\left.+\delta(-2\theta e_1-2\theta e_0 e_1+\theta^2 e_1^2+\theta^2 e_0 e_1^2-\theta^3 e_1^3+\theta^4 e_1^4-\theta^3 e_0 e_1^3+\ldots)\right]$$

$$= \bar{Y}\left[1+e_0+(1-2\delta)\theta e_1+(1-2\delta)\theta e_0 e_1\right.$$
$$\left.+\delta\theta^2 e_1^2+\delta(\theta^2 e_0 e_1^2-\theta^3 e_1^3+\theta^4 e_1^4-\theta^3 e_0 e_1^3+\ldots)\right]$$

We assume that the contribution of terms involving powers in e_0 and e_1 higher than the second is negligible, being of order $1/n^{\upsilon}$ *where* $\upsilon > 1$. Thus, from the above expression we write to a first approximation,

$$T_{RP} \cong \bar{Y}\left[1+e_0-(1-2\delta)\theta e_1+(1-2\delta)\theta e_0 e_1+\ \delta\theta^2 e_1^2\right]$$

or

$$\left(T_{RP}-\bar{Y}\right) = \bar{Y}\left[e_0-(1-2\delta)\theta e_1+(1-2\delta)\theta e_0 e_1+\ \delta\theta^2 e_1^2\right] \qquad (2.4)$$

Taking expectation of both the sides of (2.4) we obtain the bias of T_{RP} to the first degree of approximation, as

$$B(T_{RP}) = \frac{(1-f)}{n}\theta\bar{Y}\left[K+\delta\ (\theta-2K)\right]C_x^2 \qquad (2.5)$$

which will vanish if

$$\delta = \frac{K}{(2K-\theta)}$$

Thus, for $\delta = K/(2K-\theta)$, T_{RP} is almost unbiased.

Squaring both sides of (2.4) and neglecting terms of e's having power greater than two we have

$$\left(\mathrm{T_{RP}} - \bar{Y}\right)^2 = \bar{Y}^2\left[e_0^2 + (1-2\delta)\,\theta\left\{(1-2\delta)\theta e_1^2 + 2\,e_0 e_1\right\}\right] \tag{2.6}$$

Taking expectation of both sides of (2.6) we get the mean squared error (MSE) of T_{RP} to the first degree of approximation as

$$\mathrm{MSE}\left(\mathrm{T_{RP}}\right) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2 + \theta(1-2\delta)\mathrm{C_x^2}\left\{(1-2\delta)\theta + 2K\right\}\right] \tag{2.7}$$

which is minimized for

$$\delta = \frac{1}{2}\left(1 + \frac{\mathrm{K}}{\theta}\right) = \delta_0 \ (\text{say}) \tag{2.8}$$

Putting (2.8) in (2.1) we get the "asymptotically optimum estimator" (AOE) as

$$\mathrm{T_{RP0}} = \frac{\bar{y}}{2}\left[\left(1 + \frac{\mathrm{K}}{\theta}\right)\left(\frac{a\bar{X} + b}{a\,\bar{x} + b}\right) + \left(1 - \frac{\mathrm{K}}{\theta}\right)\left(\frac{a\bar{x} + b}{a\,\bar{X} + b}\right)\right] \tag{2.9}$$

Substitution of (2.8) in (2.7) yields the minimum MSE of $\mathrm{T_{RP}}$ (or the MSE of AOE T_{RP0}) as

$$\begin{aligned}\min.\mathrm{MSE}\left(\mathrm{T_{RP}}\right) &= \frac{(1-f)}{n}S_y^2\left(1-\rho^2\right) \\ &= MSE\left(T_{RP0}\right)\end{aligned} \tag{2.10}$$

which equals to the approximate MSE of the regression estimator

$$\bar{y}_{lr} = \bar{y} + \hat{\beta}\left(\bar{X} - \bar{x}\right), \tag{2.11}$$

where $\hat{\beta} = s_{xy} \big/ s_x^2$ is the sample estimate of the population regression coefficient β of y on x, $\mathrm{s_{xy}} = \sum_{\mathrm{i=1}}^{\mathrm{n}}\left(\mathrm{x_i} - \bar{\mathrm{x}}\right)\left(\mathrm{y_i} - \bar{\mathrm{y}}\right) \Big/ (\mathrm{n}-1)$ and $\mathrm{s_x^2} = \sum_{\mathrm{i=1}}^{\mathrm{n}}\left(\mathrm{x_i} - \bar{\mathrm{x}}\right)^2 \Big/ (\mathrm{n}-1)$.

It is to be noted that the AOE $\mathrm{T_{RP0}}$ in (2.9) depends on K and θ, so the AOE $\mathrm{T_{RP0}}$ can be used in practice only when K and θ are known. Here it should be mention that θ is a function of known quantities $\left(a, b, \bar{X}\right)$. So only the value of K should be known for making the use of AOE $\mathrm{T_{RP0}}$ in practice. The value of K can be made known quite accurately either from pilot study or past data or experience gathered in due course of time. This problem has been discussed

among others by Murthy (1967, pp.96-99), Reddy (1973, 1974, 1978) and Srivankataramana and Tracy (1980). Thus, the value of K can be guessed quite accurately and such an estimator can be used in practice.

3. Allowable departure

Let K_0 be an estimate (or guessed value) of K with

$$K_0 = K(1+\eta), \text{ then}$$

$$\delta = \frac{1}{2}\left(1+\frac{K_0}{\theta}\right) = \frac{1}{2}\left(1+\frac{K}{\theta}+\frac{1}{\theta}(K_0 - K)\right)$$

$$= \delta_0 + \frac{\eta K}{2\theta} \tag{3.1}$$

Putting (3.1) in (2.7) we obtain the MSE of T_{RP} as

$$MSE(T_{RP}) = MSE(T_{RP0}) + \left(\frac{1-f}{n}\right)\rho^2\eta^2 S_y^2$$

$$\Rightarrow MSE(T_{RP}) - MSE(T_{RP0}) = \left(\frac{1-f}{n}\right)\rho^2\eta^2 S_y^2$$

$$\Rightarrow \frac{MSE(T_{RP}) - MSE(T_{RP0})}{MSE(T_{RP0})} = \frac{\eta^2\rho^2}{(1-\rho^2)} \tag{3.2}$$

It follows from (3.2) that the proportional increase in MSE of T_{RP} over that of AOE T_{RP0} is less than γ if

$$\frac{\eta^2\rho^2}{(1-\rho^2)} < \gamma$$

i.e. if $|\eta| < \sqrt{\frac{(1-\rho^2)}{\rho^2}\gamma}$ (3.3)

which clearly shows that to ensure only a small relative increase in MSE of T_{RP}, $|\eta|$ must be in the neighborhood of "zero" if ρ is high but can depart substantially from "zero" if ρ is moderate.

4. Efficiency comparisons

It is well known under SRSWOR that

$$\mathrm{Var}(\bar{y}) = \left(\frac{(1-f)}{n}\right) S_y^2 \tag{4.1}$$

From (2.7) and (4.1) we have

$$\mathrm{Var}(\bar{y}) - MSE(T_{RP}) = \left(\frac{1-f}{n}\right) \bar{Y}^2 C_x^2 \theta(2\delta - 1)\left[(1-2\delta)\theta + 2K\right]$$

which is non-negative if

$$\min\left\{\frac{1}{2}, \frac{1}{2}\left(1+\frac{2K}{\theta}\right)\right\} < \delta < \max\left\{\frac{1}{2}, \frac{1}{2}\left(1+\frac{2K}{\theta}\right)\right\} \tag{4.2}$$

It is to be noted that for $\delta = 1$, T_{RP} reduces to the ratio-type estimator

$$T_R = \bar{y}\left(\frac{a\bar{X}+b}{a\bar{x}+b}\right) \tag{4.3}$$

while for $\delta = 0$ the estimator T_{RP} turns out to be the product–type estimator

$$T_P = \bar{y}\left(\frac{a\bar{x}+b}{a\bar{X}+b}\right) \tag{4.4}$$

To the first degree of approximation the mean squared errors of T_R and T_P are respectively given by

$$\mathrm{MSE}(\mathrm{T_R}) = \frac{(1-f)}{n} \bar{Y}^2 \left[C_y^2 + \theta \mathrm{C_x^2}\{\theta - 2K\}\right] \tag{4.5}$$

$$\mathrm{MSE}(\mathrm{T_P}) = \frac{(1-f)}{n} \bar{Y}^2 \left[C_y^2 + \theta \mathrm{C_x^2}\{\theta + 2K\}\right] \tag{4.6}$$

From (2.7), (4.5) and (4.6) we have

$$\mathrm{MSE}(\mathrm{T_R}) - \mathrm{MSE}(\mathrm{T_{RP}}) = \frac{4(1-\mathrm{f})}{\mathrm{n}} \bar{\mathrm{Y}}^2 \theta \mathrm{C_x^2}(1-\delta)(\delta\theta - \mathrm{K}) \tag{4.7}$$

$$\mathrm{MSE}(\mathrm{T_P}) - \mathrm{MSE}(\mathrm{T_{RP}}) = \frac{4(1-\mathrm{f})}{\mathrm{n}} \bar{\mathrm{Y}}^2 \delta\theta \mathrm{C_x^2}\left[\theta(1-\delta) + \mathrm{K}\right] \tag{4.8}$$

It follows from (4.7) and (4.8) that the ratio-product estimator T_{RP} is more efficient than

(i) the ratio type estimator T_R if

$$\min\left(\frac{K}{\theta}, 1\right) < \delta < \max\left(\frac{K}{\theta}, 1\right) \tag{4.9}$$

(ii) the product-type estimator if

$$\min\left(\left(1+\frac{\mathrm{K}}{\theta}\right), 0\right) < \delta < \max\left(\left(1+\frac{\mathrm{K}}{\theta}\right), 0\right) \tag{4.10}$$

Further, if we set (a,b)=(1,0) in (4.3) and (4.4) the ratio- type estimator T_R and the product-type estimator T_P respectively reduce to

$$T_R \to \bar{y}_R = \bar{y}\frac{\overline{X}}{\bar{x}} \text{ (usual ratio estimator)} \tag{4.11}$$

and

$$T_P \to \bar{y}_P = \bar{y}\frac{\bar{x}}{\overline{X}} \text{ (usual product estimator)} \tag{4.12}$$

Putting (a,b)=(1,0) in (4.5) and (4.6) we get the mean squared errors of usual ratio and product estimators respectively as

$$MSE(\bar{y}_R) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left[C_y^2 + C_x^2(1-2K)\right] \tag{4.13}$$

$$MSE(\bar{y}_P) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left[C_y^2 + C_x^2(1+2K)\right] \tag{4.14}$$

From (2.7), (4.13) and (4.14) we have

$$\mathrm{MSE}(\bar{y}_R) - \mathrm{MSE}(T_{RP}) = \left(\frac{1-f}{n}\right)\overline{Y}^2(1+\theta-2\theta\delta)(1-\theta-2K+2\delta\theta) \tag{4.15}$$

$$\mathrm{MSE}(\bar{y}_P) - \mathrm{MSE}(T_{RP}) = \left(\frac{1-f}{n}\right)\overline{Y}^2 C_x^2(1-\theta+2\theta\delta)(1+\theta+2K-2\delta\theta) \tag{4.16}$$

From (4.15) and (4.16) we note that the ratio-product estimator T_{RP} is better than

(i) the usual ratio estimator $\bar{y}_R$ if

$$\min\left\{\left(\frac{1+\theta}{2\theta}\right), \left(\frac{2K+\theta-1}{2\theta}\right)\right\} < \delta < \max\left\{\left(\frac{1+\theta}{2\theta}\right), \left(\frac{2K+\theta-1}{2\theta}\right)\right\} \tag{4.17}$$

(ii) the usual product estimator $\bar{y}_P$ if

$$\min\left\{\left(\frac{\theta-1}{2\theta}\right),\left(\frac{2K+\theta+1}{2\theta}\right)\right\}<\delta<\max\left\{\left(\frac{\theta-1}{2\theta}\right),\left(\frac{2K+\theta+1}{2\theta}\right)\right\} \qquad (4.18)$$

5. Estimator based on estimated optimum value

The optimum value of δ at (2.8) is

$$\delta_0=\frac{1}{2}\left(1+\frac{K}{\theta}\right) \qquad (5.1)$$

where θ is known quantity and

$$K=\rho\frac{C_y}{C_x}=\frac{\rho S_y\overline{X}}{\overline{Y}S_x}=\frac{S_{xy}}{RS_x^2}=\frac{\beta}{R}$$

Replacing β and R by their consistent estimators

$$\hat{\beta}=\frac{s_{xy}}{s_x^2} \text{ and } \hat{R}=\frac{\overline{y}}{\overline{X}}$$

respectively in (5.1) we get a consistent estimate of δ_0 as

$$\hat{\delta}_0=\frac{1}{2}\left(1+\frac{\hat{K}}{\theta}\right) \qquad (5.2)$$

where $\hat{K}=\hat{\beta}/\hat{R}$.

If the experimenter is unable to guess the value of K then it is worth advisable to replace K by $\hat{K}$ in (2.9). Thus, the estimator based on estimated 'optimum' value is

$$\hat{T}_{RP0}=\left(\frac{\overline{y}}{2}\right)\left[\left(1+\frac{\hat{K}}{\theta}\right)\left(\frac{a\overline{X}+b}{a\overline{x}+b}\right)+\left(1-\frac{\hat{K}}{\theta}\right)\left(\frac{a\overline{x}+b}{a\overline{X}+b}\right)\right] \qquad (5.3)$$

To obtain the MSE of $\hat{T}_{RP0}$ we write

$$\hat{\mathrm{K}}=\mathrm{K}(1+\mathrm{e}_2)$$

with $\mathrm{E}(\mathrm{e}_2)=\mathrm{K}+\mathrm{O}(\mathrm{n}^{-1})$. Expressing (5.3) in terms of e's we have

$$\hat{T}_{RP0}=\left(\frac{\overline{Y}}{2}\right)(1+e_0)\left[\left\{1+\frac{K}{\theta}(1+e_2)\right\}(1+\theta e_1)^{-1}+\left\{1-\frac{K}{\theta}(1+e_2)\right\}(1+\theta e_1)\right] \qquad (5.4)$$

From (5.4) we have

$$\hat{T}_{RPO} = \frac{\overline{Y}}{2}(1+e_0)\left[\{1+(K/\theta)(1+e_2)\}(1-\theta e_1+\theta^2 e_1^2+...)+\{1-(K/\theta)(1+e_2)\}(1+\theta e_1)\right]$$

$$= \frac{\overline{Y}}{2}(1+e_0)\left[1+(K/\theta)(1+e_2)-\theta e_1-K(e_1+e_1e_2)+K\theta(e_1^2+e_2e_1^2)+...\right.$$
$$\left.+1-(K/\theta)(1+e_2)+\theta e_1-K(e_1+e_1e_2)\right]$$

$$= \frac{\overline{Y}}{2}(1+e_0)\left[2-2K(e_1+e_1e_2)+K\theta(e_1^2+e_2e_1^2)+...\right]$$

$$= \overline{Y}(1+e_0)\left[1-K(e_1+e_1e_2)+\frac{K\theta}{2}(e_1^2+e_2e_1^2)+...\right]$$

$$= \overline{Y}\left[1+e_0-K(e_1+e_0e_1+e_1e_2+e_0e_0e_2)+\frac{K\theta}{2}(e_1^2+e_2e_1^2+e_0e_1^2+e_1^2e_0e_2)+...\right]$$

Neglecting terms of e's having power greater than two we have

$$\hat{T}_{RPO} = \overline{Y}\left[1+e_0-Ke_1-K(e_0e_1+e_1e_2)+\frac{K\theta}{2}e_1^2\right]$$

or

$$\left(\hat{T}_{RPO}-\overline{Y}\right) = \overline{Y}\left[e_0-Ke_1-K(e_0e_1+e_1e_2)+\frac{K\theta}{2}e_1^2\right] \qquad (5.5)$$

Now squaring both sides of (5.5) and neglecting terms of e's having power than the second, we have

$$\left(\hat{T}_{RPO}-\overline{Y}\right)^2 = \overline{Y}^2\left[e_0^2-K^2e_1^2-2Ke_0e_1\right]$$

Taking expectation of both sides of the above expression we get the mean squared error of the estimator $\hat{T}_{RPO}$ as

$$MSE\left(\hat{T}_{RPO}\right) = E\left(\hat{T}_{RPO}-\overline{Y}\right)^2$$

$$= \overline{Y}^2 E\left[e_0^2+K^2e_1^2-2K(e_0e_1)\right]$$

$$= \left(\frac{1-f}{n}\right)\overline{Y}^2\left[C_y^2 + K^2C_x^2 - 2K\rho C_yC_x\right]$$

$$= \left(\frac{1-f}{n}\right)\overline{Y}^2\left[1+\left(\rho\frac{C_y}{C_x}\right)^2 C_x^2 - 2\left(\rho\frac{C_y}{C_x}\right)\rho C_yC_x\right]$$

$$= \left(\frac{1-f}{n}\right)\overline{Y}^2\left(C_y^2 - \rho^2C_y^2\right)$$

$$= \left(\frac{1-f}{n}\right)\overline{Y}^2C_y^2\left(1-\rho^2\right)$$

$$= \left(\frac{1-f}{n}\right)S_y^2\left(1-\rho^2\right) \qquad (5.6)$$

which is equal to the minimum MSE of T_{RP} (or the MSE of T_{RP0}) given by (2.10). Thus, we established the result that the MSE of the estimator $\hat{T}_{RP0}$ in (5.3) based on 'estimated optimum value' to the first degree of approximation is same as that of AOE T_{RP0} in (2.9). So it is interesting to note that the estimator $\hat{T}_{RP0}$ in (5.3) can be used as an alternative to the AOE T_{RP0} in (2.9) if the value of the parameter K is not known.

6. Particular case

In literature it appears that Searls(1964) was the first to propose an estimator of population mean $\overline{Y}$ using the knowledge of coefficient of variation C_y of the study variate y. Later Khan(1968), Govindarajula and Sahai (1972), Gleser and Healy (1976), Upadhyaya and Singh (1984), Singh (1986) and Singh and Katiyar (1988) have also used the knowledge of coefficient of variation C_y of the study variate y in estimating the population mean $\overline{Y}$. Motivated by Searls (1964), Sisodia and Dwivedi (1981), Pandey and Dubey (1988) used the known coefficient of variation C_x of the auxiliary variable x for estimating population mean $\overline{Y}$ of the study variate y. Upadhyay and Singh (1999) have proposed some ratio and product estimators using known coefficient of variation C_x and

coefficient of kurtosis $\beta_2(x)$ of an auxiliary character x. Das and Tripathi (1980) have advocated that in many situations of practical importance the value of an auxiliary variable may be available for each unit in the population. In such situations knowledge on $\overline{X}, C_x, \beta_1(x)$(coefficient of skewness), $\beta_2(x)$ (coefficient of kurtosis) and possibly on some other parameters may be utilized. It is to be noted that few authors including Sisodia and Dwivedi (1981) and Upadhyaya and Singh(1999) have used the coefficient of variation (CV) and coefficient of kurtosis (CK) of auxiliary variable x in additive form to the sample and population mean of the auxiliary character x in estimating the population mean $\overline{Y}$ of the study variate y. Singh (2003) have pointed out that as CV and CK are unit free constants therefore their additions may not be justified. Also if CV and population mean of the auxiliary variable x are known, standard deviation in additive form is more justified. Inspired by the above points Singh (2003) suggested some ratio and product type estimators using information $(\sigma_x, \overline{X}), (\sigma_x, \overline{X}, \beta_1(x))$ *and* $(\sigma_x, \overline{X}, \beta_2(x))$ of auxiliary variate x and studied their properties. Here we have considered the problem of estimating population mean $\overline{Y}$ in situations where $\overline{X}, \sigma_x, \beta_1(x)$ *and* $\beta_2(x)$ of the auxiliary variable x are known. The proposed ratio and product type estimators for population mean $\overline{Y}$ are respectively defined by

$$T_{R1} = \overline{y}\left(\frac{\Delta\overline{X} + \sigma_x}{\Delta\overline{x} + \sigma_x}\right) \tag{6.1}$$

and

$$T_{P1} = \overline{y}\left(\frac{\Delta\overline{x} + \sigma_x}{\Delta\overline{X} + \sigma_x}\right) \tag{6.2}$$

where $\Delta = (\beta_2(x) - \beta_1(x) - 1)$.

Putting $(a,b) = (\Delta, \sigma_x)$ in (2.1) we obtain the ratio-product estimator for $\overline{Y}$ as

$$T_{RP}^{*} = \delta\overline{y}\left(\frac{\Delta\overline{X} + \sigma_x}{\Delta\overline{x} + \sigma_x}\right) + (1-\delta)\overline{y}\left(\frac{\Delta\overline{x} + \sigma_x}{\Delta\overline{X} + \sigma_x}\right) \tag{6.3}$$

Substitution of $(a,b) = (\Delta, \sigma_x)$ in (2.5) and (2.6) yield the bias and MSE of T_{RP}^{*} to the first degree of approximation, respectively as

$$B(T_{RP}^{*}) = \left(\frac{1-f}{n}\right)\theta^{*}\overline{Y}\left[K + \delta(\theta^{*} - 2K)\right]C_x^2 \tag{6.4}$$

and

$$MSE\left(T_{RP}^{*}\right)=\left(\frac{1-f}{n}\right)\overline{Y}^{2}\left[C_{y}^{2}+\theta^{*}\left(1-2\delta\right)\left\{\left(1-2\delta\right)\theta^{*}+2K\right\}C_{x}^{2}\right] \qquad (6.5)$$

where $\theta^{*}=\dfrac{\Delta\overline{X}}{\Delta\overline{X}+\sigma_{x}}$.

The expression of bias in T_{RP}^{*} at (6.4) will vanish if

$$\delta=\frac{K}{\left(2K-\theta^{*}\right)} \qquad (6.6)$$

Thus, for $\delta=K/\left(2K-\theta^{*}\right)$ the estimator T_{RP}^{*} is almost unbiased.

For $\left(a,b\right)=\left(\Delta,\sigma_{x}\right)$ it can be easily seen from (4.2), (4.9), (4.10), (4.17) and (4.18) that the proposed estimator T_{RP}^{*} is more efficient than

(i) usual unbiased estimator $\overline{y}$ if

$$\min.\left\{\frac{1}{2},\frac{1}{2}\left(1+\frac{2K}{\theta^{*}}\right)\right\}<\delta<\max.\left\{\frac{1}{2},\frac{1}{2}\left(1+\frac{2K}{\theta^{*}}\right)\right\}$$

(ii) ratio-type estimator T_{R1} if

$$\min.\left\{1,\frac{K}{\theta^{*}}\right\}<\delta<\max.\left\{1,\frac{K}{\theta^{*}}\right\}$$

(iii) product-type estimator T_{P1} if

$$\min.\left\{\left(1+\frac{\mathrm{K}}{\theta^{*}}\right),0\right\}<\delta<\max.\left\{\left(1+\frac{\mathrm{K}}{\theta^{*}}\right),0\right\}$$

(iv) usual ratio estimator $\overline{y}_{R}$ if

$$\min.\left\{\frac{\left(1+\theta^{*}\right)}{2\theta^{*}},\frac{\left(2K+\theta^{*}-1\right)}{2\theta^{*}}\right\}<\delta<\max.\left\{\frac{\left(1+\theta^{*}\right)}{2\theta^{*}},\frac{\left(2K+\theta^{*}-1\right)}{2\theta^{*}}\right\}$$

(v) the product estimator $\overline{y}_{P}$ if

$$\min.\left\{\frac{\left(\theta^{*}-1\right)}{2\theta^{*}},\frac{\left(2K+\theta^{*}+1\right)}{2\theta^{*}}\right\}<\delta<\max.\left\{\frac{\left(\theta^{*}-1\right)}{2\theta^{*}},\frac{\left(2K+\theta^{*}+1\right)}{2\theta^{*}}\right\}$$

6.1. Optimum Estimator

The MSE of T_{RP}^{*} at (6.5) is minimized for

$$\delta_0^{*} = \frac{1}{2}\left(1+\frac{K}{\theta^{*}}\right) \tag{6.7}$$

Thus, the resulting optimum estimator in the class of estimators T_{RP}^{*} is given by

$$\mathrm{T}_{\mathrm{RP0}}^{*} = \frac{\bar{y}}{2}\left[\left(1+\frac{\mathrm{K}}{\theta^{*}}\right)\left(\frac{\Delta\bar{X}+\sigma_x}{\Delta\bar{x}+\sigma_x}\right)+\left(1-\frac{\mathrm{K}}{\theta^{*}}\right)\left(\frac{\Delta\bar{x}+\sigma_x}{\Delta\bar{X}+\sigma_x}\right)\right] \tag{6.8}$$

and the estimator based on estimated optimum value

$$\hat{\delta}_0^{*} = \frac{1}{2}\left(1+\frac{\hat{K}}{\theta^{*}}\right) \tag{6.9}$$

is defined by

$$\hat{\mathrm{T}}_{\mathrm{RP0}}^{*} = \frac{\bar{y}}{2}\left[\left(1+\frac{\hat{\mathrm{K}}}{\theta^{*}}\right)\left(\frac{\Delta\bar{X}+\sigma_x}{\Delta\bar{x}+\sigma_x}\right)+\left(1-\frac{\hat{\mathrm{K}}}{\theta^{*}}\right)\left(\frac{\Delta\bar{x}+\sigma_x}{\Delta\bar{X}+\sigma_x}\right)\right] \tag{6.10}$$

where $\hat{K} = \hat{\beta}/\hat{R}$.

It can be easily shown to the first degree of approximation that

$$MSE\left(T_{RP0}^{*}\right) = MSE\left(\hat{T}_{RP0}^{*}\right) = \left(\frac{1-f}{n}\right)S_y^2\left(1-\rho^2\right) \tag{6.11}$$

7. Empirical study

To see the merits of the suggested estimator T_{RP}^{*} over $\bar{y}, \bar{y}_R, \bar{y}_P, T_{R1}$ *and* T_{P1} we consider a natural population data earlier considered by Das (1988).

$$\overline{\mathrm{Y}} = 39.0680, \overline{\mathrm{X}} = 25.1110, \mathrm{C_y} = 1.4451, \mathrm{C_x} = 1.6198, \rho = 0.7213,$$
$$\beta_1(\mathrm{x}) = 5.1869, \beta_2(\mathrm{x}) = 38.8898, \sigma_{\mathrm{x}}^2 = 1654.4,, \mathrm{N} = 278, \mathrm{n} = 30.$$

We have computed ranges of δ in which the proposed estimator T_{RP}^{*} is better than $\bar{y}, \bar{y}_R$ *and* T_{R1} and findings are given in Table 7.1.

Table 7.1. Ranges of δ for which T_{RP}^{*} is better than $\bar{y}, \bar{y}_R$ *and* T_{R1}

Estimator	**Ranges of δ in which T_{RP}^{*} is better than:**
$\bar{y}$	$0.50 < \delta < 1.1753$
T_{R1}	$0.6753 < \delta < 1$
$\bar{y}_R$	$0.6506 < \delta < 1.0247$

Table 7.1 exhibits that there is enough scope of selecting the values of scalar δ to get better estimators than $\bar{y}, \bar{y}_R$ *and* T_{R1}.

We have also computed the percent relative efficiencies (PREs) of $T_{RP0}^{*}\left(or\ \hat{T}_{RP}^{*}\right), T_{R1}$ *and* $\bar{y}_R$ with respect to usual unbiased estimator $\bar{y}$ and findings are given in Table 7.2.

Table 7.2. Percent relative efficiencies (PREs) of $\bar{y}_R, T_{R1}$ *and* $T_{RP0}^{*}\left(or\ \hat{T}_{RP}^{*}\right)$ with respect to $\bar{y}$

Estimator	$\bar{y}$	$\bar{y}_R$	T_{R1}	$T_{RP0}^{*}\left(or\ \hat{T}_{RP0}^{*}\right)$
$PRE(.,\bar{y})$	100.00	156.39	166.38	208.45

Table 7.2 clearly indicate that the proposed estimator $T_{RP0}^{*}\left(or\ \hat{T}_{RP0}^{*}\right)$ is better than the usual unbiased estimator $\bar{y}$, usual ratio estimator $\bar{y}_R$ and T_{R1}. Thus, the estimator T_{RP0}^{*} is to be preferred in practice.

Acknowledgement

Authors are thankful to the referee for critically examining the paper and pointing out the mistakes in earlier draft of the paper.

REFERENCES

COCHRAN, W. G. (1977): Sampling techniques. 3rd edition, *John Wiley, New York.*

DAS A. K. (1988): Contributions to the theory of sampling strategies based on auxiliary information. Ph.D. Thesis submitted to BCKV, Mohanpur, Nadia, West Bengal, India.

DAS, A. K. and TRIPATHI, T. P. (1980): Sampling strategies for population mean when the coefficient of variation of an auxiliary character is known. Sankhya, C,42, 76—86.

GLESER, L. J. and HEALY, J. D. (1976): Estimating the mean of a normal distribution with coefficient of variation. Jour. Amer. Statist. Assoc., 71, 977—981.

GOVINDARAJULU, Z. and SAHAI, H. (1972): Estimation of parameters of a normal distribution with known coefficient of variation. Rep. Stat. App. Res., JUSU, 19, 86—98.

KHAN, R. A. (1968): A note on estimating the mean of a normal distribution with known coefficient of variation. Jour. Amer. Statist. Assoc., 63, 1039—1041.

MURTHY, M. N. (1967): Sampling theory and methods. Statistical Publishing Society, Calcutta.

NAIK, V. D. and GUPTA, P. C. (1991): A General class of estimators for estimating population mean using auxiliary information. *Metrika, 38, 11—17.*

PANDEY, B. N. and DUBEY, V. (1988): Modified product estimator using coefficient of variation of auxiliary variate. Assam. Statist. Rev., 2(2), 64—66.

REDDY, V. N. (1978): A study on the use of prior knowledge on certain population parameters in estimation. Sankhya, C, 40, 29—37.

REDDY, V. N. (1974): On a transformed ratio method of estimation; *Sankhya*. 36(C), 59—70.

SAHAI, A. (1979): An efficient variant of the product and ratio estimators. *Stat. Neerl.* 33, 27—35.

SAHAI, A. and RAY, S. K (1980): An efficient estimator using auxiliary information; *Metrika*, 27, 271—275.

SAHAI, A. and SAHAI, A. (1985): On efficient use of auxiliary information. *Jour. Statist. Plann. Inf.*, 12, 203—212.

SEARLS D. T. (1964): The utilization of a known coefficient of variation in the estimation procedure. Jour. Amer. Statist. Assoc. 59, 1225—1226.

SINGH, G. N. (2003): On the improvement of product method of estimation in sample surveys. *Jour. Ind. Soc. Agril. Statist.*, 56, 3, 267—275.

SINGH, H. P. and KATYAR, N. P. (1988): A generalized class of estimators for common parameters of two normal distribution with coefficient of variation. Jour. Indian. Soc. Agril. Statist., 40(2), 127—149.

SINGH, M. P. (1965): On the estimation of ratio and product of the population parameters. *Sankhya, B, 27, 321—328.*

SINGH, M. P. (1967): Ratio cum Product method of estimation. Metrika, 12, 34—42.

SISODIA, B. V. S. and DWIVEDI, V. K. (1981): A modified ratio estimator using coefficient of variation of an auxiliary variable. *Ind. Soc. Agri. Stat., 33, 13—18.*

SRIVASTAVA, S. K. (1967): An estimator using auxiliary information, Cal. Stat Assoc. Bull., 16, 121—132.

SRIVASTAVA, S. K. (1971): A generalized estimator for the mean of a finite population using multi-auxiliary information. *Jour. Amer. Stat. Assoc., 66, 404—407.*

SRIVASTAVA, S. K. (1980): A class of estimators using auxiliary information in sample surveys. *Canad. Jour. Stat., 8, 253—254.*

SRIVENKATARAMANA, T. and TRACY, D. S. (1980): An alternative to ratio method in sample surveys. Ann. Inst. Statist. Math., A, 32, 111—120.

SRIVENKATARAMANA, T. and TRACY, D. S. (1981): Extending product method of estimation to positive correlation case in surveys. Austarl. Jour. Statist., 23, 95—100.

UPADHYAYA, L. N. and SINGH, H. P. (1984): On the estimation of the population mean with known with coefficient of variation. Biometrical Jour., 8, 915—922.

UPADHYAYA, L. N. and SINGH, H. P. (1999): Use of transformed auxiliary variable in estimating the finite population mean. Biometrical Jour., 41(5), 627—636.

VOS, J. W. E. (1980): Mixing of direct, ratio and product method estimators. *Stat. Neerl., 34, 209—218.*

WALSH, J. E. (1970): Generalization of ratio estimates for population total. *Sankhya, A, 32, 99—106.*

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 89—103

AN ALTERNATIVE TO RATIO ESTIMATOR OF THE POPULATION VARIANCE IN SAMPLE SURVEYS

Housila P. Singh and Prem Chandra[1]

ABSTRACT

This paper deals with the problem of estimating the population variance σ_y^2 of the study variable y using information on σ_x^2, the population variance of the auxiliary variable x. An alternative to ratio estimator of σ_y^2 is defined and its properties are studied. It has been shown that the suggested estimator is more efficient than usual unbiased estimator s_y^2 and the ratio estimator reported by Isaki (1983). An empirical study is carried out in support of the proposed estimator.

Key words: Population variance, Study variable, Auxiliary Variable, Bias and Variance.

1. Introduction

It is well known that the auxiliary information in the theory of sampling is used to increase the efficiency of estimator of population parameters. Out of many ratio, regression and product methods of estimation are good examples in this context. The problem of estimating population mean $\bar{Y}$ of the study variable y using supplementary information on an auxiliary variable x has been considered extensively in the survey literature. Important references on an indirect methods such as ratio, regression, product and their modifications are available in Singh (1986), Krisnaiah and Rao (1988) and Chaudhuri and Vos (1988).

In many situations, the problem of estimating the population variance σ_y^2 of y assumes importance. The use of auxiliary information in estimating the population variance σ_y^2 was first considered by Das and Tripathi (1978). Later

[1] Department of Biostatistics, All India Institute of Medical Sciences, New Delhi-110029, India

various authors including Isaki (1983), Prasad and Singh (1990,92), Rueda Garcia and Arcos Cebrian (1990), Singh and Biradar (1994), Arcos Ceberian and Rueda Garcia (1997), Ahmed et al (2003), Biradar and Singh (1994,98), Singh and Joarder (1998), Srivastava and Jhaji (1980), Upadhyaya and Singh (1983,86,99), Singh et al (1988), Singh and Singh (1984,2001), Singh (1988), Singh and Biradar (1994), Singh and Sahoo (1996-2000) and Tripathi et al., (1988) etc. have paid their attention towards the estimation of population variance σ_y^2 using auxiliary information and have proposed estimators with their properties.

Consider a finite population with N units U_1, U_2, …..U_N. The study variate y and the auxiliary variate x related to y assumes values (y_i, x_i) on the unit U_i , (i=1,2,……..,N). We denote

$$S_y^2 = \frac{1}{(N-1)}\sum_{i=1}^{N-1}(y_i - \bar{y})^2 = \frac{N}{(N-1)}\sigma_y^2 \qquad \text{and}$$

$$S_x^2 = \frac{1}{(N-1)}\sum_{i=1}^{N-1}(x_i - \bar{x})^2 = \frac{N}{(N-1)}\sigma_x^2 \qquad \text{(assumed to be known),}$$

where $\bar{Y} = (1/N)\sum_{i=1}^{N} y_i$ and $\bar{X} = (1/N)\sum_{i=1}^{N} x_i$ are the population means of y and x respectively.

When the population is very large (or infinite)

$$S_y^2 \cong \sigma_y^2, \quad S_x^2 \cong \sigma_x^2.$$

It is desired to estimate S_y^2 using information on S_x^2 (known). Let a sample of size n be drawn from this population using simple random sampling without replacement (SRSWOR) scheme. Then a usual unbiased estimator of S_y^2 is define by

$$s_y^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad (1.1)$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ is sample mean of y.

When the population variance S_x^2 is known, Isaki (1983) suggested a ratio estimator for S_y^2 as

$$t_R = (s_y^2 / s_x^2)\ S_x^2, \qquad (1.2)$$

where $s_x^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is an unbiased estimator of S_x^2.

For simplicity, suppose that the population size N is large enough relative to sample size n, assuming that the finite population correction factor can be ignored.

To the first degree of approximation, the bias and variance of t_R, are respectively given by

$$B(t_R) = (S_y^2 / n)\ (\beta_2(x) - 1)\ (1 - C) \tag{1.3}$$

and

$$Var\ (t_R) = (S_y^4 / n)[(\beta_2(y) - 1) + (\beta_2(x) - 1)\ (1 - 2C)], \tag{1.4}$$

where, $\beta_2(y) = \mu_{40} / \mu_{20}^2$, $\beta_2(x) = \mu_{04} / \mu_{02}^2$, $h = \mu_{22} / (\mu_{20}\mu_{02})$,

C= (*h*-1) / ($\beta_2(x) - 1$) and $\mu_{rs} = \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^r \left(x_i - \bar{x}\right)^s / N$,

(r, s) being non-negative integers.

To the first degree of approximation, the variance of s_y^2 is given by

$$Var\left(s_y^2\right) = \left(S_y^4 / n\right)\ \left(\beta_2(y) - 1\right) \tag{1.5}$$

As noted by Murthy (1967, p.365) in case of ratio estimator for population mean $\bar{Y}$, we also note that the closeness of the expressions (1.3) and (1.4) respectively to the exact bias and variance (/mean square error) of t_R depends much on the composition of the population and sample size. Hence, these expressions must be taken with reservation.

2. Product method as alternative to ratio method

In contrast to the above, exact expressions for the bias and variance (/mean square error) can be obtained if the product method of estimation is employed. Thus motivated by Srivenkataramana and Tracy (1980), we define a product estimator alternative to Isaki's ratio estimator as

$$t_p = s_y^2 \left(\frac{A - s_x^2}{A - S_x^2} \right), \tag{2.1}$$

where A is a scalar to be chosen suitably.

To obtain the bias and mean square error of t_p, we write

$$s_y^2 = S_y^2(1+e_0)$$

$$s_x^2 = S_x^2(1+e_1)$$

such that

$$E(e_0) = E(e_1) = 0.$$

Expressing t_p in terms of e's , we have

$$t_p = S_y^2(1+e_0)\frac{\{A - S_x^2(1+e_1)\}}{(A - S_x^2)}$$

$$= S_y^2(1+e_0)\ (1-\theta\ e_1)$$

$$= S_y^2(1+e_0-\theta\ e_1-\theta\ e_0\ e_1) \qquad (2.2)$$

where $\theta\ = S_x^2/(A - S_x^2)$.

Taking expectation of both sides of t_p and then subtracting S_y^2 from both sides, we get the exact bias of t_p as

$$B(t_p) = -\theta\ \frac{\mathrm{cov}(s_y^2, s_x^2)}{S_x^2} \qquad (2.3)$$

and

$$\mathrm{MSE}\ (t_p)$$

$$= [(S_y^4/n)\{(\beta_2(y)-1)+\theta\ (\beta_2(x)-1)(\theta-2C)\}+\theta\ S_y^4\{\theta\ E(e_0^2\ e_1^2)-2\ E(e_0^2\ e_1)+2\theta\ E(e_0\ e_1^2)\}] \qquad (2.4)$$

Thus, to the first degree of approximation, (ignoring finite population correction terms) the bias and variance of t_p are respectively given by

$$\mathrm{B}\ (t_p) = -\theta\ (S_y^2/n)\ (\beta_2(x)-1)\ C \qquad (2.5)$$

and

$$\mathrm{Var}\ (t_p) = (S_y^4/n)[(\beta_2(y)-1)+\theta(\beta_2(x)-1)(\theta-2C)], \qquad (2.6)$$

The variance of t_p at (2.6) is minimized for

$$\theta_{Opt} = C$$

$$\Rightarrow A_{Opt} = \frac{S_x^2(1+\theta_{Opt})}{\theta_{Opt}} \tag{2.7}$$

Substituting (2.7) in (2.6) we get the minimum variance of t_p as

$$\min \text{Var}\,(t_p) = \left(S_y^4 / n\right)\left[\left(\beta_2(y)-1\right) - \left(\beta_2(x)-1\right) C^2\right] \tag{2.8}$$

From (1.4), (1.5) and (2.8) we have

$$Var\left(s_y^2\right) - \min.Var(t_p) = \left(S_y^4 / n\right)\left(\beta_2(x)-1\right) C^2 \geq 0 \tag{2.9}$$

$$Var\ (\mathrm{t}_R) - \min .\ Var\,(t_p) = \left(S_y^4 / n\right)\left(\beta_2(x)-1\right)\ (1-C)^2\ \geq 0 \tag{2.10}$$

It follows from (2.9) and (2.10) that the proposed estimator t_p is more efficient than usual unbiased estimator s_y^2 and Isaki's ratio estimator t_R at optimum condition. The value of S_x^2 is known, but the exact value of 'C' is rarely known. However, in repeated surveys or studies based on multiphase sampling, where information regarding the same variates is collected on several occasions, it is possible to guess quite accurately the values of certain parameters [see, Srivenkataramana and Tracy (1980)]. Hence we assume that 'C' can be guesed. In turn a good approximation A_0 of A_{opt} can be obtained.

3. Allowable departure from optimum

In this section we discus to what extent approximate value A_0 may depart form the optimum A_{opt} and yet give an estimator better than t_R or s_y^2.

Suppose θ_0 be the guessed value of θ_{opt} such that

$$\theta = \theta_0 = C\ (1+\in)\text{, when A = A}_0.$$

Thus,

$$Var(t_p) = \min.Var(t_p)\left[1 + \frac{C^2\left(\beta_2(x)-1\right)\in^2}{\left\{\left(\beta_2(y)-1\right) - C^2\left(\beta_2(x)-1\right)\right\}}\right]$$

$$\text{or } \frac{Var\ (t_p) - \min.Var(t_p)}{\min.Var(t_p)} = \frac{C^2\ \in^2 \left(\beta_2(x)-1\right)}{\left\{\left(\beta_2(y)-1\right)-C^2\left(\beta_2(x)-1\right)\right\}}, \qquad (3.1)$$

Thus, it follows that the proportional increase in $Var(t_p)$ over that of $\min.Var(t_p)$ is less than α if

$$|\in| < \left[\frac{\left\{\left(\beta_2(y)-1\right)-C^2\left(\beta_2(x)-1\right)\right\}\alpha}{C^2\left(\beta_2(x)-1\right)}\right]^{1/2}$$

or equivalently,

$$|\in| < \left[\frac{\left(1-\rho_{**}^2\right)\alpha}{\rho_{**}^2}\right]^{1/2} \qquad (3.2)$$

where ρ_{**} is the correlation coefficient between $(y-\overline{Y})^2$ and $(x-\overline{X})^2$.

Thus to ensure only a small relative increase in variance, $|\in|$ must be close to zero if ρ_{**} is high but can depart substantially form 'zero' if ρ_{**} is just moderate.

For $\theta = \theta_0$ in (2.6), we have

$$Var(t_P) = (S_y^4/n)\left[\beta_2(y)-1) + \theta_0(\beta_2(x)-1)(\theta_0 - 2C)\right] \qquad (3.3)$$

From (1.4) and (3.3), we have

$$Var(t_R) - Var(t_P) = (S_y^4/n)(\beta_2(x)-1)(1-\theta_0)(\theta_0 - 2C + 1)$$

$$> 0 \text{ if}$$

$$\text{either} \quad 1 < \theta_0 < (2C\text{-}1) \qquad (3.4)\ (a)$$

$$\text{or} \quad (2C\text{-}1) < \theta_0 < 1 \qquad (3.4)\ (b)$$

Further from (1.5) and (3.3) we have

$$Var(s_y^2) - Var(t_P) = -(S_y^4/n)\theta_0(\beta_2(x)-1)(\theta_0 - 2C)$$

$$> 0 \text{ if}$$

$$\text{either} \quad 0 < \theta_0 < 2C \qquad (3.5)\ (a)$$

$$\text{or } 2C < \theta_0 < 0 \qquad (3.5)\text{ (b)}$$

As discussed in Srivenkataramana and Tracy (1980), to investigate where (3.4) (b) and (3.5) (a) are satisfied simultaneously, we distinguish between the cases $0 \le C \le 1$

and C >1.

Case (i): $0 \le C \le 1$

Here, choose $A_0 > 2S_x^2$ so that θ_0 is in (0, 1). If C happens to be in (0, ½) condition (3.4) (b) is automatically met since 2C-1 < 0, but (3.5) (a) needs

$$A_0 > (1+\frac{1}{2C})S_x^2 \qquad (3.6)$$

On the other hand, if C is in (½, 1), then (3.5) (a) always satisfied since 2C >1, but (3.4) (b) requires

$$A_0 < \left\{1+\frac{1}{(2C-1)}\right\}S_x^2 \qquad (3.7)$$

Thus, any $A_0 > 2S_x^2$, satisfying (3.6) or (3.7) as the case may be, will make t_p an improved estimator of S_y^2.

Case (ii): C>1

Here, choose $A_0 < 2 S_x^2$. In addition we need only that $A_0 > \left\{1+\frac{1}{(2c-1)}\right\} S_x^2$ for t_p to be more precise than t_R or s_y^2.

To get a clearer idea for t_p to be more precise than t_R or s_y^2 some typical situations are given in Table 3.1.

Table 3.1. Optimum A, and lower and upper bounds on the choice of A for different values of 'C'

C	Lower bound on A	Optimum A	Upper bound on A
0.05	$11.00\ S_x^2$	$21.00\ S_x^2$	∞
0.15	$4.33\ S_x^2$	$7.67\ S_x^2$	∞
0.25	$3.00\ S_x^2$	$5.00\ S_x^2$	∞
0.35	$2.43\ S_x^2$	$3.86\ S_x^2$	∞
0.45	$2.11\ S_x^2$	$3.22\ S_x^2$	∞
0.50	$2.00\ S_x^2$	$3.00\ S_x^2$	∞
0.55	$2.00\ S_x^2$	$2.82\ S_x^2$	$11.00\ S_x^2$
0.65	$2.00\ S_x^2$	$2.54\ S_x^2$	$4.33\ S_x^2$
0.75	$2.00\ S_x^2$	$2.33\ S_x^2$	$3.00\ S_x^2$
0.85	$2.00\ S_x^2$	$2.17\ S_x^2$	$2.43\ S_x^2$
0.95	$2.00\ S_x^2$	$2.05\ S_x^2$	$2.11\ S_x^2$
1.00	$2.00\ S_x^2$	$2.00\ S_x^2$	$2.00\ S_x^2$
1.25	$1.67\ S_x^2$	$1.80\ S_x^2$	$2.00\ S_x^2$
1.50	$1.50\ S_x^2$	$1.67\ S_x^2$	$2.00\ S_x^2$
1.75	$1.40\ S_x^2$	$1.57\ S_x^2$	$2.00\ S_x^2$
2.00	$1.33\ S_x^2$	$1.50\ S_x^2$	$2.00\ S_x^2$
2.25	$1.29\ S_x^2$	$1.44\ S_x^2$	$2.00\ S_x^2$
2.50	$1.25\ S_x^2$	$1.40\ S_x^2$	$2.00\ S_x^2$
2.75	$1.22\ S_x^2$	$1.36\ S_x^2$	$2.00\ S_x^2$
3.00	$1.20\ S_x^2$	$1.33\ S_x^2$	$2.00\ S_x^2$
3.50	$1.17\ S_x^2$	$1.29\ S_x^2$	$2.00\ S_x^2$
4.00	$1.14\ S_x^2$	$1.25\ S_x^2$	$2.00\ S_x^2$

Remark 3.1: In case of bivariate normal populations, the variance of s_y^2, t_p and t_R are respectively given by

$$Var(s_y^2) = 2S_y^4 / n \tag{3.8}$$

$$Var(t_R) = 4S_y^4(1-\rho^2)/n \tag{3.9}$$

and

$$Var(t_p) = 2S_y^4\left[1+\theta(\theta-2\rho^2)\right]/n \tag{3.10}$$

From (3.8), (3.9) and (3.10) we have,

$$Var(s_y^2) - Var(t_p) = -(2s_y^4/n)\theta(\theta-2\rho^2)$$

$$>0 \quad \text{if}$$

$$0 < \theta < 2\rho^2 \tag{3.11}$$

$$Var(t_R) - Var(t_p) = (2S_y^4/n)(1-\theta)(1-2\rho^2+\vartheta)$$

$$> 0 \quad \text{if}$$

$$(2\rho^2-1) < \theta < 1 \tag{3.12}$$

Thus, from (3.11) and (3.12) we note that the suggested estimator t_p is more efficient than t_R or s_y^2 if

$$\text{either } 0 < \theta < 2\rho^2 \;; \qquad \rho^2 < \frac{1}{2} \tag{3.13}$$

$$\text{or } (2\rho^2-1) < \theta < 1 \;; \quad \rho^2 > \frac{1}{2} \tag{3.14}$$

The minimization of (3.10) yields the optimum value of θ as

$$\theta_{opt} = \rho^2 \tag{3.15}$$

$$\Rightarrow A_{opt} = \frac{S_x^2(1+\theta_{opt})}{\theta_{opt}} = \frac{S_x^2(1+\rho^2)}{\rho^2} \tag{3.16}$$

Thus, the resulting minimum variance of t_p is given by

$$\min.Var(t_p) = \left(\frac{2S_y^4}{n}\right)(1-\rho^4) \tag{3.17}$$

From (3.8), (3.9) and (3.17) we have

$$Var(s_y^2) - \min.Var(t_p) = \left(\frac{2S_y^4}{n}\right)\rho^4 > 0 \tag{3.18}$$

$$Var(t_R) - \min.Var(t_P) = \left(\frac{2S_y^4}{n}\right)(1-\rho^2)^2 \tag{3.19}$$

> 0 provided $|\rho| \neq 1$

It follows from (3.18) and (3.19) that the proposed estimator t_p at its optimum condition is better than s_y^2 and t_R in bivariate normal population too. It is interesting to note that only the knowledge of correlation coefficient between y and x is sufficient to get better estimators from t_p. The value of ρ can be easily obtained quite accurately from past experiences or past data.

4. Estimator based on estimated optimum

Substitution of (2.7) in t_p yields the optimum estimator of S_y^2 as

$$t_p = s_y^2 \frac{\left\{S_x^2 + C(S_x^2 - s_x^2)\right\}}{S_x^2} \tag{4.1}$$

with the variance given at (2.8).

We note that the optimum estimator t_{p0} is used in practice only when 'C' is exactly known. In practice it is not so. Then, it is advisable to replace it with its consistent estimator

$$\hat{C} = \frac{(\hat{h}-1)}{(\beta_2(\hat{x})-1)}, \tag{4.2}$$

$$\text{where } \hat{h} = \frac{\hat{\mu}_{22}}{\hat{\mu}_{20}\,\hat{\mu}_{02}}, \quad \hat{\beta}_2(x) = \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} \quad \text{and}$$

$$\hat{\mu}_{rs} = \sum_{i=1}^{n} (y_i - \bar{y})^r (x_i - \bar{x})^s / n \ ,$$

(r, s) being non –negative integers, see Singh and Joarder (1998).

Replacing C in (4.1) by $\hat{C}$, we get an estimator based on 'estimated optimum' as

$$\hat{t}_{p0} = s_y^2 \frac{\left\{ S_x^2 + \hat{C}(S_x^2 - s_x^2) \right\}}{S_x^2} \tag{4.3}$$

To find the variance of the estimator $\hat{t}_{p0}$, let us define

$e_2 = (\hat{C} - C)/C$, where E (e_2) = O (n^{-1}), then, variance of $\hat{t}_{p0}$ is given by

$$Var(\hat{t}_{p0}) = E\left(\hat{t}_{p0} - E(\hat{t}_{p0}) \right)^2$$

$$= E\left[S_y^2 (1+e_0)\{1 - C(1+e_2)e_1\} - \left\{ S_y^2 + B(\hat{t}_{p0}) \right\} \right]^2$$

As it can easily be seen that the bias of $\hat{t}_{p0}$ will contain the terms of order $n^{-\nu} (\nu \geq 1)$, therefore, to terms of order n^{-1},

$$Var(\hat{t}_{p0}) \cong S_y^4 E(e_0 - Ce_1)^2$$

$$= S_y^4 E\left[e_0^2 + C^2 e_1^2 - 2Ce_0 e_1 \right]$$

$$= (S_y^4 / n)\left[(\beta_2(y) - 1) + C^2(\beta_2(x) - 1) - 2C(h - 1) \right]$$

$$= (S_y^4 / n)\left[(\beta_2(y) - 1) - C^2(\beta_2(x) - 1) \right]$$

$$= \min . Var(t_p)$$

which is approximately same as min. $Var(t_p)$ in (2.8).

5. Empirical study

To observe the performance of newly formulated estimator t_p over s_y^2 and t_R we have chosen a natural population data considered by Das (1988).

The variates and the required parameters are:

x: number of agricultural labourers in 1961,

y: number of agricultural labourers in 1971,

S_x^2 =1654.44, S_y^2 =3187.42, h = 26.8142,

$\beta_2(x) = 38.8898$, $\beta_2(y) = 25.8969$.

The percent relative efficiencies of the suggested estimator t_p with respect to s_y^2 and t_R have been computed for different values of A_0 (θ_0) and presented in Table 5.1.

Table 5.1. Percent relative efficiencies of t_p with respect to s_y^2 and t_R for different values of A_0

θ_0	A_0	PRE (t_p, s_y^2)	PRE (t_p, t_R)
0	∞	100.00	44.82
0.25	5.00 S_x^2	173.40	77.71
0.40	3.50 S_x^2	241.53	108.25
0.50	3.00 S_x^2	291.02	130.43
θ_{opt} =0.6813	A_{opt} = 2.46778 S_x^2 = 4082.79	340.61	152.65
0.75	2.33 S_x^2	332.38	149.01
0.85	2.18 S_x^2	296.82	133.03
1.00	2.00 S_x^2	223.13	100.00
1.25	1.80 S_x^2	130.46	57.03
Range of θ_0	⟶	(0.00, 1.3626)	(0.3626, 1.00)
Range of A_0	⟶	(1.7339 S_x^2 , ∞) = (2868.62 , ∞)	(2.00 S_x^2 , 3.7579 S_x^2) = (3308.88, 6217.15)

Table 5.1 clearly indicates that the proposed estimator t_p (at its optimum condition) is more efficient than usual unbiased estimator s_y^2 and Isaki's (1983) ratio estimator t_R with substantial gain in efficiency. It is also seen that even if A_0 (or θ_0) departs from its exact optimum value, the proposed estimator t_p is more efficient than s_y^2 and t_R with considerable gain in efficiency. Thus there is enough scope of selecting A in t_p to give better estimator than s_y^2 and t_R.

REFERENCES

ARCOS CEBRIAN A. and RUEDA GARCIA M. (1997): Variance estimation using auxiliary information ; an almost unbiased multivariate ratio estimator. Metrika, 45: 171—178.

AHMED M. S., ABU-DAYYEH W. and HURAIRAH, A. A. O. (2003): Some estimators for population variance under two phase sampling. Statistics in Transition, 6 (1), 143-150.

BIRADAR R. S. and SINGH H. P. (1994): An alternative to ratio estimator of population variance. Assam Statistical Review, 8, (2), 18—33.

BIRADAR R. S. and SINGH H. P. (1998): Predictive estimation of finite population variance. Cal. Statist. Assoc., 48, (191—192), 229—235.

CHAUDHURI A. and VOS JWE (1988): Unified theory and strategies of survey sampling. North Holland.

DAS A. K. (1988): Contributions to the theory of sampling strategies based on auxiliary information. Ph.D. Thesis submitted to BCKV, Mohanpur, Nadia, West Bengal, India.

DAS A. K. and TRIPATHI T. P. (1978): Use of auxiliary information in estimating the finite population variance. Sankhya, C, 40, 139—148.

ISAKI C. T. (1983): Variance estimation using auxiliary information, Jour. Amer. Statist. Assoc., 78, 117—123.

KRISNAIAH P. R. and RAO C. R. (1988): Hand book of Statistics, Vol. 6, North Holland.

MURTHY M. N. (1967): Sampling Theory and Methods. Statistical Publishing Society, Calcutta.

PRASAD B. and SINGH H. P. (1990): Some improved ratio-type estimators of finite population variance in sample surveys. Communications in Statistics-Theory and Methods, 19 (3), 1127—1139.

PRASAD B. and SINGH H. P. (1992): Unbiased estimators of finite population variance using auxiliary information in sample surveys. Communications in Statistics- Theory and Methods, 21(5), 1367—1376.

SINGH H. P. and UPADHYAYA L. N. and NAMJOSHI U. D. (1988): Estimation of finite population variance. Current Science, 57, (24), 1331—1334.

SINGH H. P. and SINGH R. (2001): Improved ratio type estimator for variance using auxiliary information. Jour. Ind. Soc. Agril. Statist., 53 (3), 276—287.

SINGH H. P., and SAHOO L. N. (1996—2000): A class of estimators for population variance using two auxiliary variates. The Vikram, 24, (1 & 3), 41—48.

SINGH H. P. (1988): Estimation of variance when the coefficient of variation is known in normal parent. Proc. Nat. Acad. Sci., India, 58(A), II, 247—250.

SINGH H. P. and BIRADAR R. S. (1994): Estimation of finite population variance using auxiliary information. Jour. Indian Soc. Statist. Opers. Res., 15, (1—4), 47—53.

SINGH H. P. (1986): A generalized class of estimators of ratio, product and mean using supplementary information on an auxiliary character in PPSWR sampling scheme. Gujarat Statist. Rev., 39 (3), 280—288.

SINGH S. and JOARDER A. H. (1998): Estimation of finite population variance using random non -response in survey sampling. Metrika, 47, 241—249.

SINGH R. K. and SINGH G. (1984): A class estimators for population variance using information on two auxiliary variates. Aligarh Jour. Statistics, (3&4), 43—49.

SRIVASTAVA S. K. and JHAJI H. S. (1980): A class of estimators using auxiliary information for estimating finite population variance. Sankhya, C, 42, 87—96.

SRIVENKATARAMANA T. and TRACY D. S. (1980): An alternative to ratio method in sample surveys. Ann. Inst. Statist. Maths., 32, 1, A, 111—120.

TRIPATHI T. P., SINGH H. P. and UPADHYAYA L. N. (1988): A generalized method of estimation in double sampling. Jour. Ind. Statist. Assoc., 26, 91—101.

UPADHYAYA L. N. and SINGH H. P. (1986): On a dual to ratio estimator for estimating finite population variance. Nep. Math. Sc. Rep., 11(1), 37—42.

UPADHYAYA L. N. and SINGH H. P. (1999): An estimator of population variance that utilizes the kurtosis of an auxiliary variable in sample surveys. Vikram Mathematical Journal, 19, 14—17.

UPADHYAYA L. N. and SINGH H. P. (1983): Use of auxiliary information in the estimation of population variance. Mathematical Forum, 6, (2), 33—36.

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 105—115

ON LINEAR COMBINATION OF RATIO AND PRODUCT TYPE EXPONENTIAL ESTIMATOR FOR ESTIMATING THE FINITE POPULATION MEAN

Rajesh Singh, Pankaj Chauhan and Nirmala Sawan

ABSTRACT

A ratio-product type exponential estimator for estimating the finite population mean is proposed. Under simple random sampling without replacement (SRSWOR) scheme, the expressions of bias and mean-squared error (MSE) up to the first order of approximation are derived. The estimator is compared for its precision with usual mean per unit, ratio, product and Bahl and Tuteja (1991) estimators and is found to be more efficient in many practical situations.

Key words: Auxiliary information, bias, mean-squared error, ratio-product estimator, exponential estimator.

1. Introduction

Consider a finite population $U=(U_1, U_2...U_N)$ of N units. Let y and x stand for the variable under study and auxiliary variable respectively. Let (y_i, x_i), i=1,2...N, denote the values of the population units for (y, x) respectively. Further let (y_i, x_i), i=1,2...N, denote the values of the units included in a sample s_n of size n drawn by SRSWOR. It is common practice to use the auxiliary variable for improving the precision of the estimate of a parameter. Out of many, ratio and product methods of estimation are good illustrations in this context when the correlation between the study variate and the auxiliary variate is positive (high) ratio method of estimation is quite effective. On the other hand, when this correlation is negative (high) product method of estimation is employed effectively.

In order to have a survey estimate of the population mean $\overline{Y}$ of the study character y, assuming the knowledge of the population mean $\overline{X}$ of the auxiliary character x, the well-known ratio estimator is

$$t_1 = \bar{y}\left(\frac{\bar{X}}{\bar{x}}\right) \tag{1.1}$$

The conventional product estimator for $\bar{Y}$ is defined as

$$t_2 = \bar{y}\left(\frac{\bar{x}}{\bar{X}}\right) \tag{1.2}$$

Bahl and Tuteja (1991) suggested an exponential ratio and product type estimators

$$t_3 = \bar{y}\exp\left(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right) \tag{1.3}$$

$$t_4 = \bar{y}\exp\left(\frac{\bar{x}-\bar{X}}{\bar{x}+\bar{X}}\right) \tag{1.4}$$

In this paper, under SRSWOR, we have suggested a ratio-product type exponential estimator for estimating the finite population mean. The expressions of bias and MSE, up to the first order of approximation, have been obtained. The case of double sampling is also discussed.

2. Proposed estimator

It has been theoretically established that, in general, the linear regression estimator is more efficient than the ratio (product) estimator except when the regression line of y on x passes through the neighborhood of the origin, in which case the efficiencies of these estimators are almost equal. Also in many practical situations the regression line does not pass through the neighborhood of the origin. In these situations, the ratio estimator does not perform as good as the linear regression estimator. The ratio estimator does not perform well as the linear regression estimator does.

Following Singh and Espejo (2003), we propose following class of ratio-product estimators:

$$t_5 = \bar{y}\left[\alpha\exp\left(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right)+(1-\alpha)\exp\left(\frac{\bar{x}-\bar{X}}{\bar{x}+\bar{X}}\right)\right] \tag{2.1}$$

where α is a real constant to be determined such that the MSE of t_5 is minimum.

For $\alpha=1$, t_5 reduces to Bahl and Tuteja (1991) estimator $t_3 = \bar{y}\exp\left(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right)$, and for $\alpha=0$, it reduces to $t_4 = \bar{y}\exp\left(\frac{\bar{x}-\bar{X}}{\bar{x}+\bar{X}}\right)$.

3. Bias and MSE of t_5

To obtain the bias and MSE of t_5 to the first degree of approximation, we write

$$\bar{y} = \bar{Y}(1+e_0),\ \bar{x} = \bar{X}(1+e_1)$$

such that

$$E(e_0)=E(e_1)=0.$$

$$E(e_0^2) = \frac{1-f}{n}C_y^2,$$

$$E(e_1^2) = \frac{1-f}{n}C_x^2,$$

$$E(e_0e_1) = \frac{1-f}{n}\rho C_y C_x.$$

Expressing (2.1) in terms of e's, we have-

$$t_5 = \bar{Y}(1+e_0)\left[\alpha\exp\left(\frac{\bar{X}-\bar{X}(1+e_1)}{\bar{X}+\bar{X}(1+e_1)}\right)+(1-\alpha)\exp\left(\frac{\bar{X}(1+e_1)-\bar{X}}{\bar{X}(1+e_1)+\bar{X}}\right)\right]$$

$$= \bar{Y}(1+e_0)\left[\alpha\exp\left(\frac{-e_1}{2+e_1}\right)+(1-\alpha)\exp\left(\frac{e_1}{2+e_1}\right)\right] \qquad (3.1)$$

Expanding the right hand side of (3.1) and retaining terms up to second powers of e's, we have

$$t_5 = \bar{Y}\left[1+e_0+\frac{e_1}{2}-\alpha e_1+\frac{e_1^2}{4}+\frac{e_0e_1}{2}-\alpha e_0e_1\right] \qquad (3.2)$$

Taking expectations of both sides of (3.2) and then subtracting $\bar{Y}$ from both sides, we get the bias of the estimator t_5, up to the first order of approximation, as

$$\text{Bias}(t_5) = \left(\frac{1-f}{n}\right)\overline{Y}\left[\frac{C_x^2}{4} + \rho C_y C_x\left(\frac{1}{2} - \alpha\right)\right] \tag{3.3}$$

From (3.2), we have

$$\left(t_5 - \overline{Y}\right) \cong \overline{Y}\left[e_0 + e_1\left(\frac{1}{2} - \alpha\right)\right] \tag{3.4}$$

Squaring both sides of (3.4) and then taking expectation, we get MSE of the estimator t_5, up to the first order of approximation, as

$$\text{MSE}(t_5) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left[C_y^2 + C_x^2\left(\frac{1}{4} + \alpha^2 - \alpha\right) + 2\rho C_y C_x\left(\frac{1}{2} - \alpha\right)\right] \tag{3.5}$$

Minimization of (3.5) with respect to α yields its optimum value as

$$\alpha = \frac{2K+1}{2} = \alpha_0 \,(\text{say}) \tag{3.6}$$

where $K = \rho\dfrac{C_y}{C_x}$.

Substitution of (3.6) in (3.5) yields the minimum value of MSE (t_5) as

$$\min.\text{MSE}(t_5) = \text{M}(t_5)_o, (\text{say}) = \left(\frac{1-f}{n}\right)\overline{Y}^2 C_y^2(1-\rho^2) \tag{3.7}$$

which is same as that of traditional linear regression estimator.

4. Efficiency comparisons

In this section, the conditions for which the proposed estimator t_5 is better than $\overline{y}$, t_1, t_2, t_3 and t_4 have been obtained. The MSE's of these estimators up to the order $O(n^{-1})$ are derived as

$$\text{Var}(\overline{y}) = \left(\frac{1-f}{n}\right)\overline{Y}^2 C_y^2 \tag{4.1}$$

$$\text{MSE}(t_1) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left(C_y^2 + C_x^2 - 2\rho C_y C_x\right) \tag{4.2}$$

$$MSE(t_2) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left(C_y^2 + C_x^2 + 2\rho C_y C_x\right) \tag{4.3}$$

$$MSE(t_3) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left(C_y^2 + \frac{C_x^2}{4} - \rho C_y C_x\right) \tag{4.4}$$

$$MSE(t_4) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left(C_y^2 + \frac{C_x^2}{4} + \rho C_y C_x\right) \tag{4.5}$$

To compare the efficiency of the proposed estimator t_5 with the existing estimator, from (3.7) and (4.1)-(4.5), we have

$$Var(\bar{y}) - M(t_5)_o = \rho^2 \geq 0 \tag{4.6}$$

$$MSE(t_1) - M(t_5)_o = (C_x - \rho C_y)^2 \geq 0 \tag{4.7}$$

$$MSE(t_2) - M(t_5)_o = (C_x + \rho C_y)^2 \geq 0 \tag{4.8}$$

$$MSE(t_3) - M(t_5)_o = \left(\frac{C_x^2}{2} - \rho C_y\right)^2 \geq 0 \tag{4.9}$$

$$MSE(t_4) - M(t_5)_o = \left(\frac{C_x^2}{2} + \rho C_y\right)^2 \geq 0 \tag{4.10}$$

Using (4.6)-(4.10), we conclude that the proposed estimator t_5 outperforms $\bar{y}$, t_1, t_2, t_3 and t_4.

5. Double sampling

A number of sampling strategies depend upon the possession of advanced information about auxiliary information (i.e. on the known population mean $\overline{X}$ of the auxiliary variable x). However, in certain practical situations when $\overline{X}$ is not known a priori, the technique of two-phase or double sampling is used. The two-phase sampling happens to be a powerful and cost effective (economical) procedure for finding the reliable estimate in first phase sample for the unknown parameters of the auxiliary variable x and hence has eminent role to play in survey sampling, for instance, see Hidiroglou and Sarndal (1998).

Allowing SRSWOR design in each phase, the two-phase sampling scheme is as follows:

i. The first phase sample $s_{n'}(s_{n'} \subset U)$ of a fixed size n' is drawn to measure only x in order to formulate a good estimate of population mean $\overline{X}$,

ii. Given $s_{n'}$, the second phase sample $s_n(s_n \subset s_{n'})$ of a fixed size n is drawn to measure y only.

Let $\overline{x} = \frac{1}{n}\sum_{i\in s_n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i\in s_n} y_i$ and $\overline{x}' = \frac{1}{n'}\sum_{i\in s_{n'}} x_i$.

When $\overline{X}$ is not known, two-phase ratio and product type exponential estimator are

$$t_6 = \overline{y}\exp\left[\frac{\overline{x}' - \overline{x}}{\overline{x}' + \overline{x}}\right] \tag{5.1}$$

$$t_7 = \overline{y}\exp\left[\frac{\overline{x} - \overline{x}'}{\overline{x} + \overline{x}'}\right] \tag{5.2}$$

To obtain the bias and MSE of t_6 and t_7, we write

$$\overline{y} = \overline{Y}(1+e_0),\ \overline{x} = \overline{X}(1+e_1),\ \overline{x}' = \overline{X}(1+e_1')$$

such that

$$E(e_0) = E(e_1) = E(e_1') = 0$$

and

$$E(e_0^2) = f_1 C_y^2,\ E(e_1^2) = f_1 C_x^2,\ E(e_1'^2) = f_2 C_x^2,\ E(e_0 e_1') = f_2 \rho C_y C_x,$$

$$E(e_1 e_1') = f_2 C_x^2$$

where $f_1 = \frac{1}{n} - \frac{1}{N}$, $f_2 = \frac{1}{n'} - \frac{1}{N}$.

Expressing t_6 and t_7 in terms of e's and then taking expectation, we get the expression of bias and MSE up to the first order of approximation as –

$$\text{Bias}(t_6) = \overline{Y} f_3 \frac{C_x^2}{4}(1-2K) \tag{5.3}$$

$$\text{MSE}(t_6) = \overline{Y}^2\left[f_1 C_y^2 + f_3 \frac{C_x^2}{4}(1-4K)\right] \tag{5.4}$$

$$\text{Bias}(t_7) = \overline{Y} f_3 \frac{C_x^2}{4}(1+2K) \tag{5.5}$$

$$\text{MSE}(t_7) = \overline{Y}^2 \left[f_1 C_y^2 + f_3 \frac{C_x^2}{4}(1+4K) \right] \tag{5.6}$$

where $f_3 = \frac{1}{n} - \frac{1}{n'}$.

6. Proposed two-phase estimator

For estimating $\overline{Y}$, following Singh and Espejo (2003), we propose the following class of ratio-product estimators:

$$t_8 = \overline{y}\left[\alpha_1 \exp\left(\frac{\overline{x}' - \overline{x}}{\overline{x}' + \overline{x}} \right) + (1-\alpha_1) \exp\left(\frac{\overline{x} - \overline{x}'}{\overline{x} + \overline{x}'} \right) \right] \tag{6.1}$$

where α_1 is a real constant to be determined such that MSE of t_8 is minimum.

For α_1=1, t_8 reduce to estimator $t_6 = \overline{y} \exp\left(\frac{\overline{x}' - \overline{x}}{\overline{x}' + \overline{x}} \right)$ and for α_1=0, it reduces to $t_7 = \overline{y} \exp\left(\frac{\overline{x} - \overline{x}'}{\overline{x} + \overline{x}'} \right)$.

Expressing equation (6.1) in terms of e's, we have –

$$t_8 = \overline{Y}(1+e_0)\left[\alpha_1 \exp\left(\frac{e_1' - e_1}{e_1' + e_1 + 2} \right) + (1-\alpha) \exp\left(\frac{e_1 - e_1'}{e_1' + e_1 + 2} \right) \right] \tag{6.2}$$

Expanding right hand side of (6.2) and retaining terms up to second powers of e's, we have

$$t_8 = \overline{Y}[1 + e_0 + \left(\frac{e_1 - e_1'}{2} \right) + \alpha_1 (e_1' - e_1) + \alpha_1 e_0 e_1' - \alpha_1 e_0 e_1 + \frac{(e_1 - e_1')^2}{4}$$

$$+ \frac{e_0 e_1}{2} - \frac{e_0 e_1'}{2}] \tag{6.3}$$

Taking expectations of both sides of (6.3) and then subtracting $\overline{Y}$ from both sides, we get the bias of the estimator t_8, up to the first order of approximation, as

$$\text{Bias}(t_8) = f_3\overline{Y}\frac{C_x^2}{8}\left[1-8K\left(\alpha_1-\frac{1}{2}\right)\right] \tag{6.4}$$

From (6.3), we have

$$\left(t_8-\overline{Y}\right)\cong\overline{Y}\left[e_0+\alpha_1\left(e_1'-e_1\right)+\left(\frac{e_1-e_1'}{2}\right)\right] \tag{6.5}$$

Squaring both sides of (6.5) and then taking expectations, we get the MSE of the estimator t_8, up to the first order of approximation, as

$$\text{MSE}(t_8) = \overline{Y}^2\left[f_1C_y^2+f_3C_x^2\left(\alpha_1-\frac{1}{2}\right)\left\{\left(\alpha_1-\frac{1}{2}\right)-2K\right\}\right] \tag{6.6}$$

Minimization of (6.6) with respect to α_1 yields its optimum value as-

$$\alpha_1 = \left(\frac{2K+1}{2}\right) = \alpha_{10}(\text{say}) \tag{6.7}$$

Substitution of (6.7) in (6.6) yields the minimum value of MSE (t_8) as

$$\min.\text{MSE}(t_8) = \text{M}(t_8)_0,(\text{say}) = \overline{Y}^2C_y^2\left[f_1-f_3\rho^2\right] \tag{6.8}$$

which is same as that of usual two-phase linear regression estimator.

7. Efficiency comparisons

The MSE of usual two-phase ratio and product estimator is given by

$$\text{MSE}(t_9) = \overline{Y}^2\left[f_1C_y^2+f_3C_x^2(1-2K)\right] \tag{7.1}$$

$$\text{MSE}(t_{10}) = \overline{Y}^2\left[f_1C_y^2+f_3C_x^2(1+2K)\right] \tag{7.2}$$

From (6.8) and (4.1), (5.4), (5.6), (7.1) and (7.2), we have

$$\text{Var}(\overline{y})-\text{M}(t_8)_o = f_3\rho^2 \geq 0 \tag{7.3}$$

$$\text{MSE}(t_6)-\text{M}(t_8)_o = f_3\left(\frac{C_x}{2}-\rho C_y\right)^2 \geq 0 \tag{7.4}$$

$$\mathrm{MSE}(t_7) - \mathrm{M}(t_8)_o = f_3\left(\frac{C_x}{2} + \rho C_y\right)^2 \geq 0 \tag{7.5}$$

$$\mathrm{MSE}(t_9) - \mathrm{M}(t_8)_o = f_3\left(C_x - \rho C_y\right)^2 \geq 0 \tag{7.6}$$

$$\mathrm{MSE}(t_{10}) - \mathrm{M}(t_8)_o = f_3\left(C_x + \rho C_y\right)^2 \geq 0 \tag{7.7}$$

From (7.3) and (7.7), we conclude that our proposed estimator t_8 is better than $\bar{y}$, t_6, t_7, t_9 and t_{10}.

8. Empirical study

The various results obtained in previous sections are now examined with the help of following data:

Population I: Cochran (1977)
The required parameters are

$$C_y = 1.4177, C_x = 1.4045, \rho = 0.887.$$

Population II: Rao (1983)
Parameters of the population are

$$C_y^2 = 0.1818,\ C_x^2 = 0.0165,\ \rho = -0.7036.$$

Population III: Anderson (1958)
The required parameters are

$$\overline{Y} = 183.84, \overline{X} = 185.72, \overline{Z} = 151.12, C_y = 0.0546, C_x = 0.0526,$$
$$\rho = 0.7108,$$

Take n=7and $n' = 10$.

Table 8.1. Percent Relative Efficiency of various estimators (Single Sampling) with respect to $\bar{y}$.

Estimator	PRE's (., $\bar{y}$)	
	Population	
	I	II
$\bar{y}$	100	100
t_1	446.46	65.99
t_2	26.75	114.12
t_3	272.76	80.98
t_4	47.07	123.42
$(t_5)_o$	468.98	198.04

Table 8.2. Percent Relative Efficiency of various estimators (Double Sampling) with respect to $\bar{y}$.

Estimator	PRE's(., $\bar{y}$)
	Population III
$\bar{y}$	100
t_6	123.27
t_7	72.35
$(t_8)_0$	126.68
t_9	122.56
t_{10}	57.08

9. Conclusion

From table 8.1 we see that our proposed estimator t_5 outperforms all other estimators considered. Also, it is unaffected whether study and auxiliary variates are positively or negatively correlated.

From table 8.2, it is clear that the two-stage version of proposed estimator t_5 i.e. t_8 also performs well than other stated estimators.

REFERENCES

ANDERSON, T. W. (1958): An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, Inc., New York.

BAHL, S. and TUTEJA, R. K., (1991): Ratio and Product type exponential estimator, Information and Optimization sciences, Vol.XII, I, 159—163.

COCHRAN, W. G., (1977): Sampling techniques. Third U.S. edition. Wiley eastern limited, 325.

HIDIROGLOU, M. A. and SARNDAL, C. E., (1998): Use of auxiliary information for two-phase sampling. Survey Methodology, 24(1), and 11—20.

RAO, T. J., (1983): A new class of unbiased product estimators, Tech. Rep. No. 15/83, Stat.-Math., ISI, Calcutta, India.

SINGH, H. P. and ESPEJO, M. R., (2003): On linear regression and ratio-product estimation of a finite population mean. The statistician, 52, 1, 59—67.

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 117—128

ON OPTIMUM NUMBERS OF CONTROLS UNDER CASE CONTROL STUDIES ON UNWANTED PREGNANCIES: AN ANALYTICAL APPRAISAL

Shahina Begum[1], **Sada Nand Dwivedi**[2], **Arvind Pandey**[3]

ABSTRACT

Case-Control study design is very much in medical statistics. But decision to take number of control per case is difficult in a limited resources and time period. The objective of the study is to decide how many controls per case were needed in a case control study. The data were extracted from the second round of National Family Health Survey. Unwanted pregnancy was considered as case and wanted pregnancy as control. Controls were selected in the ratio of 1:1, 1:2, 1:3 and 1:4 from the total available wanted pregnancies using simple random sampling. Socio-economic-demographic and programmatic variables were considered as independent variable. Multiple logistic regression analysis was carried out to find out the factors associated in each data set. Hosmer & Lemeshow for goodness of fit was used to validate the model. It was concluded that one control per case might be sufficient to provide result in part with the cohort approach in unwanted pregnancies.

Key words: cohort approach, case control approach, unintended pregnancies,

Introduction

Under the non-experimental analytical studies, there are commonly used three types of study designs, namely cross-sectional studies, cohort studies, and case control studies. Under the cross-sectional studies, the data on outcome variable as well as various potential covariates are collected as prevailing on the day of data collection. Cohort study may be either retrospective cohort or prospective cohort. The major emphasis under such design is on selection of the population with

[1] Asst Prof, Department of Community Medicine, RD Gardi Medical College, Surasa, Ujjain, Madhya Pradesh-456006, India.
[2] Addl Prof, Department of Biostatistics, All India Institute of Medical Sciences, New Delhi-110 029, India.
[3] Director, National Institute of Medical Sciences, ICMR, Ansari Nagar, New Delhi-110029, India.

consideration of important exposure variable. Under retrospective design the data are collected on outcome variable as well as exposure variable along with the other covariates once the exposure and outcome occurred before the start of study. But, exposure precedes the outcome variable. The data collection may start either from exposure to outcome variable (historical cohort) or from outcome to exposure variable (case-control study). Under the prospective cohort, a group of exposed population is followed to observe the occurrence of the outcome. (Rothman, 2002; Silman & Macfarlane, 2002; Brownson & Petitti, 1998; Gordis, 1996; Schlesselman, 1982). Obviously, the prospective cohort studies involve additional cost and time. To overcome these problems, case control study design is generally a preferable study design in medical research, where the selection of population depends on the person with outcome variable as a case whereas a person without outcome variable as a control (Schlesselman, 1982). Also, as mentioned earlier case control analysis can also be carried out using the data available under cohort studies. The case control studies may be preferred over cohort studies not only because it is comparatively easier to conduct, less costly and less time taking but it also involves comparatively smaller sample size and there is no loss to follow up. In addition, the case control studies are preferred under either time bound program to complete the studies and report the findings or rare outcomes (Schlesselman, 1982). In view of advantages of case control design over cohort design, with due care of desired issues related to this design, it may be justified to deal with case-control approach within a limited time period. After planning a case control design it is difficult to decide how many numbers of controls per case will be needed to achieve the objective of the study. The objective of the paper is to assess what would be the appropriate ratio of cases and controls under a case-control analysis to provide the results in part with cohort approach. Both the study designs are used to find out the factors associated with the outcome variables. The present study also was to find out the factors associated with unwanted pregnancy. Unwanted pregnancy is the pregnancy that is not wanted at all by the respondent. It was reported the age, education level, exposure to mass media, birth interval were associated with unwanted pregnancy (Eggleston, 1999; Adetunji, 1998). In addition some programmatic variables were also considered in the present study assuming that programmatic variables in the community may affect the prevalence of unwanted pregnancies.

Method and Materials

To achieve the first objective, data has been extracted from second round of National Family Health Survey (NFHS). It was conducted in 1998-99 under the auspices of the Ministry of Health and Family welfare, India and funded by United States Agency for International Development (USAID) through ORC MACRO, USA. The main objective of NFHS-2 was to provide national and state-level estimates on fertility, infant and child mortality, the practice of family

planning, maternal and child health, utilization of health services available to mothers and children, quality of health and family welfare services, women reproductive health problems, status of women and domestic violence. It also aimed to provide quality data for analytical research in these areas.

Sampling method used under NFHS-2 was multistage systematic random sampling. Two-stage sampling procedure was followed in rural area, with selection at the first stage of primary sampling units (PSU) i.e. selection of village with probability proportional to population size (PPS), followed by systematic selection of households within each selected village. Three staged sampling procedure was followed in urban area, wards were selected using PPS sampling at first stage, one Census Enumeration Block (CEB) was selected from each selected ward at second stage, then at third stage households were selected from each selected CEB using systematic random sampling. On an average 30 households were targeted for selection in each PSU of rural and urban areas. All ever-married women aged 15-49 years were interviewed from each selected household. The survey had standardized questionnaires, training, data collection, and data processing. Three questionnaires were used, namely, household questionnaire, women questionnaire, and village questionnaire (administered only in rural area). More details of the sample design, methodology, data processing and questionnaires are available in the report for India (IIPS & ORC MACRO, 2000).

The present study is based on the pregnancies among a sample of currently married women at country level (except northeastern states excluding Assam) who were pregnant at the time of interview. A total of 5650 complete records of pregnant women were available for analysis.

Dependent Variable

The pregnancy intention, as mentioned earlier, was the dependent variable in this study. The related information were collected by asking women directly as "At the time you became pregnant, did you want to become pregnant then, did you want to wait until later, or did you want no (more) children at all?" Pregnancy was classified as unwanted if women did not want any more children, mistimed if women wanted to wait until later and wanted if women wanted to become pregnant then. Therefore, a total records on pregnancy intention (n=5650) was classified as unwanted (n=561), mistimed (n=922) and wanted pregnancy (n=4167). In this study mistimed pregnancies was excluded from the study.

Explanatory Variables

On the basis of review of literature and subject knowledge, a set of independent/explanatory variables was selected for the analysis such as age (<25/≥25 years), place of residence (rural/urban), and husband lives in household (no/yes), women's occupation (working/not working), women's/husband's education level (high school complete & `/illiterate/literate, <middle school complete/middle school complete), religion-caste (SC-ST-OBC Hindu/other Hindu/ non-Hindu), standard of living index (high/low/medium), type of family (nuclear/extended), no son/one sons/more than one son), the interval from last live birth to index pregnancy (<18, ≥18), exposure to mass media (no/yes), ever contraceptive use (no/yes), ever terminated pregnancy (no/yes), and ever physically mistreated by husband (no/yes), exposure to IEC activities (no/yes), distance to the nearest health facility (<2/≥2 km), and distance to all season road (<2/≥2 km). Some of the community level variables such as percent of women exposed to the health facility within 2 km, percent of literate women and percent of women exposed to family planning messages on mass media were included. Governmental actions, policies, awareness and information are communicated to people through mass media. Exposure to FP messages on mass media widens the source of knowledge and opinion and may enhance an individual's capability for making informed choices given the health facility. Education may be interpreted as measure of capacity to make informed decision. It is expected that an educated women would not have a pregnancy that she was not prepared for, if she had access to the means of preventing it.

Cohort Approach

In this approach, total available recodes on unwanted (561) and wanted (4167) pregnancies were considered for analysis.

Case-Control Approach

Definition of Case and Control

Case: Unwanted pregnancy was considered as one type of case
Control*:* Wanted pregnancy was considered as control.

Data management

As reported in cohort approach, there were 561 unwanted pregnancies and 4167 wanted pregnancies (controls). To carry out case-control analysis 561, 1122, 1683, and 2244 wanted pregnancies (controls) were randomly selected from a total of 4167 available wanted pregnancies in the ratio of 1:1, 1:2, 1:3, and 1:4 respectively. Random selection was done using SPSS software (SPSS 11.5).

Therefore, as such 5 different data sets were made, saved and kept separately for analysis.

Statistical Analysis

To begin with, for each data set i.e. under cohort as well as case-control approach, percentage distribution of women in various categories related to each covariate was worked out for unwanted and wanted pregnancy. Chi-square test was used to assess the difference in proportion between cases and controls. Since the outcome variable is dichotomous, stepwise logistic regression analysis with inclusion criteria of p=0.10 and exclusion criteria of p=0.15 have been performed to determine the association of some selected variables to unwanted pregnancies. Further, Hosmer and Lemeshow goodness of fit was used to validate the developed model. Bootstrapping technique is also used to validate the developed model.

Result

Characteristics of Unwanted and Wanted Pregnancies

Under cohort approach:

Table 1 presents the percentage distribution of total unwanted (n=561) and total wanted pregnancies (n=4167) in relation to selected background characteristics of women. As evident from table, demographic & socioeconomic characteristics such as, age, education level, religion-caste, partner's education level, living of husband in household, standard of living index and type of family were associated ($p<0.05$) with unwanted pregnancies. Higher proportion of women aged over or equal to 25 years, with middle SLI, and from extended family reported unwanted pregnancy. It was found that place of residence and working status of women were not associated ($p>0.05$) with unwanted pregnancy. Regarding reproductive characteristics of women such as number of living sons, interval between last live birth and index pregnancy, and ever experience of MTP were associated ($p<0.05$) with unwanted pregnancy. It was evident that as the number of surviving sons increased, the percentage of unwanted pregnancies also increased. Further, women's contraceptive behaviour such as ever contraceptive use and exposure to family planning message on mass media were associated ($p<0.05$) with unwanted pregnancy. Significantly ($p<0.05$) higher proportion of women who were physically mistreated by their husband reported unwanted pregnancy.

Programmatic variable such as IEC was found to be associated ($p<0.05$) with unwanted pregnancies. State level variables such as percent of women having access to health facility within 2 km, percentage of literate women and percentage

of women exposed to family planning messages on mass media were also considered to examine the variation among states.

Under case control approach

Table 1 presents the percentage distribution of unwanted and wanted pregnancies in relation to selected background characteristics of women under all the four data sets (1:1, 1:2, 1:3, & 1:4). As evident from the table, demographic & socioeconomic characteristics such as age, education level, religion-caste, partner's education level, living of husband in household, standard of living index and type of family were associated ($p<0.05$) with unwanted pregnancies. It was true under all the four data sets. However, place of residence was associated ($p<0.05$) with unwanted pregnancy for data set 1:1 only. Further, reproductive characteristics of women such as number of surviving sons, interval between last live birth & index pregnancy, and ever experience of MTP were associated ($p<0.05$) with unwanted pregnancy under all the four data sets. It was evident that as the number of surviving sons increased, the percentage of unwanted pregnancies also increased. Further, women contraceptive behaviour such as, ever contraceptive use, and exposure to family planning message on mass media were significantly ($p<0.05$) associated with unwanted pregnancy under all the four data sets. Significantly ($p<0.05$) higher proportion of women who were physically mistreated by their husband reported unwanted pregnancy. This association was true under each of the four data sets. Among the considered programmatic variables such as IEC and all season road, only IEC was found to be associated ($p<0.05$) with unwanted pregnancies under each of the four data sets.

Table 1. Percentage Distribution of Unwanted and Wanted Pregnancy by Selected Background Characteristics under Cohort and Case Control Approach

Background characteristics	Total women with pregnancies		Controls (wanted) in the ratio of			
	Unwanted (561)	Wanted (4167)	1:1	1:2	1:3	1:4
Age						
15-24	138 (24.6)	2685 (64.4)	361 (64.3)	732 (65.2)	1103 (65.5)	1434 (63.9)
25 & over	423 (75.4)	1482 (35.6)	200 (35.7)	390 (34.8)	580 (34.5)	810 (36.1)
Residence						
Urban	134 (23.9)	1088 (26.1)	161 (28.7)	290 (25.8)	430 (25.5)	585 (26.1)
Rural	427 (76.1)	3079 (73.9)	400 (71.3)	832 (74.2)	1253 (74.5)	1659 (73.9)
Education level						
Illiterate	389(69.3)	2256 (54.1)	310 (55.3)	604 (53.8)	896 (53.2)	1234 (55.0)
Less than middle school	81(14.4)	749 (18.0)	86 (15.3)	199(17.7)	305 (18.1)	389 (17.3)
Middle school	47 (8.4)	401 (9.6)	58 (10.3)	119 (10.6)	156 (9.3)	208 (9.3)
High school & above	44 (7.8)	761 (18.3)	107 (19.1)	200 (17.8)	326 (19.4)	413 (18.4)
Religion- Caste						
Non-Hindu	144 (25.7)	798 (19.2)	111 (19.8)	211 (18.8)	296 (17.6)	422 (18.8)
Hindu other	118 (21.0)	1065 (25.6)	159 (28.3)	296 (26.4)	453 (26.9)	565 (25.2)
Hindu SC/ST/OBC	299 (53.3)	2304 (55.3)	291 (51.9)	615 (54.8)	934 (55.5)	1257 (56.0)
Working status*						
Not working	406 (72.4)	2999 (72.0)	404 (72.0)	801 (71.4)	1230 (73.1)	1609 (71.7)
Working	155 (27.6)	1168 (28.0)	157 (28.0)	321 (28.6)	453 (26.9)	635 (28.3)
Partner's education						
Illiterate	193 (34.4)	1176 (28.2)	150 (26.7)	305 (27.2)	469 (27.9)	637 (28.4)
Less than middle school	171 (30.5)	964 (23.1)	129 (23.0)	269 (24.0)	375 (22.3)	502 (22.4)
Middle school	70 (12.5)	664 (15.9)	87 (15.5)	169 (15.1)	278 (16.5)	366 (16.3)
High school & above	127 (22.6)	1363 (32.7)	195 (34.8)	379 (33.8)	561 (33.3)	739 (32.9)
Husband lives in house						
Yes	544 (97.0)	3946 (94.7)	530 (94.5	1056 (94.1	1597 (94.9)	2112 (94.1)
No	17 (3.0)	221 (5.3)	31 (5.5)	66 (5.9)	86 (5.1)	132 (5.9)
Standard of living index						
Low	210 (37.4)	1275 (30.6)	181 (32.3	338 (30.1	504 (29.9)	706 (31.5)
Medium	290 (51.7)	2054 (49.3)	266 (47.4)	576 (51.3)	834 (49.6)	1100 (49.0)
High	61 (10.9)	838 (20.1)	114 (20.3)	208 (18.5)	345 (20.5)	438 (19.5)
Type of family						
Nuclear	261 (46.5)	1202 (28.8)	160 (28.5)	308 (27.5)	467 (27.7)	660 (29.4)
Extended	300 (53.5)	2965 (71.2)	401 (71.5)	814 (72.5)	1216 (72.3)	1584 (70.6)

Table 1. Percentage Distribution of Unwanted and Wanted Pregnancy by Selected Background Characteristics under Cohort and Case Control Approach (Contd.)

Background characteristics	Total women with pregnancies		Controls (wanted) in the ratio of			
	Unwanted (561)	Wanted (4167)	1:1	1:2	1:3	1:4
Number of surviving sons						
No sons	66 (11.8)	2662 (63.9)	350 (62.4)	723 (64.4)	1074 (63.8)	1426 (63.5)
One son	222 (39.6)	1162 (27.9)	167 (29.8)	305 (27.2)	473 (28.1)	647 (28.8)
At least 2 sons	273 (48.7)	343 (8.2)	44 (7.8)	94 (8.4)	136 (8.1)	171 (7.6)
Interval between last live birth to index pregnancy						
<18 months	98 (17.5)	373 (9.0)	51 (9.1)	104 (9.3)	143 (8.5)	224 (10.0)
>=18 months	450 (80.2)	2218 (53.2)	293 (52.2)	583 (52.0)	869 (51.6)	1182 (52.7)
No child	13 (2.3)	1576 (37.8)	217 (38.7)	435 (38.8)	671 (39.9)	838 (37.3)
Ever contraceptive use						
Ever used	188 (33.5)	716 (17.2)	93 (16.6)	201 (17.9)	282 (16.8)	396 (17.6)
Never used	373 (66.5)	3451 (82.8)	468 (83.4)	921 (82.1)	1401 (83.2)	1848 (82.4)
Expose to FP messages on mass media						
No	261 (46.5)	1678 (40.3)	232 (41.4)	443 (39.5)	667 (39.6)	920 (41.0)
Yes	300 (53.5)	2489 (59.7)	329 (58.6)	679 (60.5)	1016 (60.4)	1324 (59.0)
Ever experienced MTP						
Never	538 (95.9)	4088 (98.1)	548 (97.7	1098 (97.9	1649 (98.0)	2204 (98.2)
Ever	23 (4.1)	79 (1.9)	13 (2.3)	24 (2.1)	34 (2.0)	40 (1.8)
Ever physically mistreated by husband						
Never	410 (73.1)	3617 (86.8)	500 (89.1)	967 (86.2)	1449 (86.1)	1926 (85.8)
Ever	151 (26.9)	550 (13.2)	61 (10.9)	155 (13.8)	234 (13.9)	318 (14.2)
IEC						
No	283 (50.4)	1868 (44.8)	243 (43.3)	502 (44.7)	755 (44.9)	1015 (45.2)
Yes	278 (49.6)	2299 (55.2)	318 (56.7)	620 (55.3)	928 (55.1)	1229 (54.8)
All season road*						
<2 K.M.	317 (56.5)	2331 (55.9)	328 (58.5)	627 (55.9)	919 (54.6)	1227 (54.7)
≥2 K.M.	244 (43.5)	1836 (44.1)	233 (41.5)	495 (44.1)	764 (45.4)	1017 (45.3)

Note: * indicate p>0.05, and @ p>0.05 except under the data set 1:1

Multiple Logistic Regression Analysis

Under cohort approach:

As evident from the Table 2, age, educational level, standard of living index, number of surviving sons, interval between last live birth to index pregnancy, ever contraceptive use, exposure to FP messages on mass media, ever experience of MTP, and ever physically mistreated by her husband were found to be significantly associated with unwanted pregnancies. None of the community level (state) variables were retained in the final model.

Under case control approach:

As evident from the Table 2, age, education, number of surviving sons, and ever contraception users were retained in the model under each of the four data sets, whereas covariates standard of living index and type of family remained in the model under data sets 1:1 and 1:2 respectively. Further, covariates exposed to family planning messages on mass media and ever physically mistreated by her husband were retained in the model except under data sets 1:3 and 1:1 respectively. Among community level covariates, only percentage of women who had access to health facility within 2 km were associated with unwanted pregnancies under data sets 1:3 & 1:4. Further, percentage of literate women was retained only in the model for data set 1:4.

It was found that women aged 25 and over were more likely to have experienced unwanted pregnancy than their counterparts under each of the four data sets. Illiterate women were more likely to have experienced unwanted pregnancy than women with high school & above under each of the four data sets. Women with middle school education were more likely to report their index pregnancy as unwanted under each of the four data sets, whereas women with less than middle school were more likely to report their index pregnancy as unwanted under data set 1:1 only. Women from extended family were 25 percent less likely to have experienced unwanted pregnancy than their counterparts under

Table 2. Adjusted OR (with 95% CI) of Unwanted Pregnancy Using Multiple Logistic Regression Analysis under Cohort and Case Control Approach

Background characteristics	Cohort approach	Case -Controls approach in the ratio of			
	All	1:1	1:2	1:3	1:4
-2loglikelihood	2504.44	1027.33	1466.78	1732.90	1984.56
p-value under Hosmer-Lemeshow test	0.67	0.93	0.68	0.50	0.62
Age					
15-24	1.00	1.00	1.00	1.00	1.00
25 & over	2.38 (1.88-3.01)	2.13 (1.53-2.97)	2.19 (1.65-2.90)	2.43 (1.87-3.16)	2.21 (1.72-2.85)
Education level					
Illiterate	1.55 (1.03-2.34)	2.08 (1.21-3.55)	1.75 (1.11-2.77)	1.64 (1.08-2.49)	1.78 (1.17-2.72)
Less than middle school	1.26 (0.81-1.96)	1.99 (1.08-3.66)	1.58 (0.95-2.61)	1.40 (0.87-2.25)	1.50 (0.95-2.37)
Middle school	2.18 (1.34-3.54)	2.48 (1.27-4.87)	1.97 (1.11-3.47)	2.42 (1.41-4.16)	2.10 (1.26-3.52)
High school & above	1.00	1.00	1.00	1.00	1.00
Standard of living index					
Low	1.37 (0.93-2.02)	1.06 (0.62-1.80)			
Medium	1.55 (1.09-2.20)	1.54 (0.95-2.50)			
High	1.00	1.00			
Type of family					
Nuclear			1.00		
Extended			0.75 (0.57-0.98)		
No. of surviving sons					
0	1.00	1.00	1.00	1.00	1.00
1	3.71 (2.70-5.09)	3.12 (2.09-4.65)	3.6 2 (2.55-5.15)	3.55 (2.53-4.99)	3.48 (2.51-4.83)
2+	12.38 (8.82-17.39)	12.10 (7.41-19.75)	11.81 (7.89-17.69)	12.60 (8.61-18.44)	13.01 (9.04-18.72)
Interval from last live birth to index pregnancy					
<18 months	6.68 (3.47-12.87)	7.95 (3.79-16.67)	6.45 (3.22-12.90)	7.99 (4.05-15.77)	6.22 (3.19-12.12)
>=18 months	3.52 (1.89-6.58)	4.05 (2.06-7.98)	3.75 (1.97-7.15)	3.88 (2.05-7.34)	3.71 (1.97-6.97)
No child	1.00	1.00	1.00	1.00	1.00
Ever contra-ceptive use					
Never used	1.00	1.00	1.00	1.00	1.00
Ever used	1.99 (1.57-2.52)	2.44 (1.68-3.57)	1.85 (1.36-2.52)	2.11 (1.60-2.80)	1.85 (1.42-2.41)
Expose to FP messages on mass media					
No	1.00	1.00	1.00		1.00
Yes	1.31 (1.04-1.66)	2.12 (1.41-3.19)	1.37 (1.02-1.84)		1.29 (1.00-1.67)
Ever physically mistreated by husband					
Never	1.00		1.00	1.00	1.00
Ever	1.56 (1.23-2.00)		1.54 (1.12-2.10)	1.54 (1.15-2.05)	1.41 (1.08-1.84)
% women who had access to health facility within 2 km				1.01 (1.01-1.02)	1.02 (1.00-1.03)
% literate women					0.99 (0.98-1.00)

data set 1:2 only. As the number of surviving sons increased the odds of unwanted pregnancy also increased under each of the four data sets. Ever users of contraceptive methods were more likely to report their index pregnancy unwanted than never users. This significant association was true under each of the four data sets. Women exposed to mass media were more likely to report their current pregnancy as unwanted pregnancy than their counterparts under two data sets 1:1 & 1:2. When the other factors considered in the analysis were controlled statistically, women ever physically mistreated by their husband were significantly more likely to have experienced unwanted pregnancy than their counterparts except under data set 1:1. Community level variable, percentage of women exposed to FP messages on mass media was found to be significantly associated with unwanted pregnancy under the data sets 1:3 & 1:4.

Validation of the Developed Model Hosmer-Lemeshow Goodness of Fit

The value of level of significance (i.e. p) related to Hosmer & Lemeshow test for goodness of fit is presented in Table 2. As evident from the table, p-value (p=0.67) indicates that the developed model using cohort approach describes the data satisfactorily. The Table also indicate the p-value (p=0.93, p=0.68, p=0.50 & p=0.62 respectively) under each of the four data sets (1:1, 1:2, 1:3, & 1:4) under case control approach of unwanted pregnancy indicates that the developed models describe the data satisfactorily. The higher p-value indicates the best fit of the model i.e. in case of data set considering one control per case.

Discussion and conclusion

Using the second round of NFHS-2 data it was found that factors age, education, number of surviving sons, and ever contraception use were significantly associated with unwanted pregnancy under each of the four data sets under case control approach as well as under whole data set under cohort approach. Covariates standard of living index remained in the model under data set 1:1 but found significantly associated with unwanted pregnancy under cohort approach. Type of family remained in the model for data set 1:2 only. Further, covariates exposed to family planning messages on mass media was significantly associated with unwanted pregnancy under cohort approach and under case control approach except for data sets 1:3. Also, other variable, physical violence was associated with unwanted pregnancy for all the data sets except data set 1:1. Among community level covariates, only percentage of women having access to health facility within 2 km was found to be significantly associated with unwanted pregnancy for data sets 1:3 & 1:4. Further, the odds ratio of the variables under consideration was in same direction in all the five sets of data.

The higher p-value of Hosmer & Lemeshow was found for data set 1:1 under case control approach of unwanted pregnancy indicates that the developed models describe the data satisfactorily. Hence, it was concluded that one control per case

may be sufficient to provide results in part with the cohort approach in unwanted pregnancies.

REFERENCE

ADETUNJI JA. (1998). Unintended Childbearing in Developing Countries: Levels, Trends, and Determinants. *Demographic and Health Surveys Analytical Reports Series*. Macro International Inc. Calverton, Maryland USA.

BROWNSON RC, and PETITTI D. (1998). *Applied Epidemiology: Theory and Practice*. Oxford University Press. New York.

EGGLESTON E. (1999). Determinant of unintended pregnancy among women in Ecuador. *International Family Planning Perspectives*; 25 (1): 27—33.

GORDIS L. 1996. *Epidemiology.* Saunders. Philadelphia

ROTHMAN KJ. (2002) *Epidemiology: An Introduction*. Oxford University Press. New York.

SILMAN AJ, and MACFARLANE GJ. (2002). Epidemiological Studies: A practical Guide. Cambridge University Press. Cambridge UK.

SCHLESSELMAN JJ. (1982). Case Control Studies- Design, Conduct, Analysis. Oxford University Press. New York.

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 129—138

TEST OF HYPOTHESIS ON THE MEANS OF A BIVARIATE CORRELATED NORMAL

Angiola Pollastri[1] [2]

ABSTRACT

The purpose of this paper is to improve the procedure proposed by Duncan to test the hypotheses on the means of a Bivariate Correlated Normal (B.C.N.) that are both equal to fixed values against all the possible alternatives.

The procedure proposed here is based on the exact distribution of absolute maximum and absolute minimum of the components of a Standardized Bivariate Correlated Normal (S.B.C.N.) r.v.. The critical values are reported in appendix.

The new stepwise test allows to accept the null hypotheses with a fixed probability error rate while Duncan's procedure is conservative. Moreover, the procedure here considered is more powerful than the Duncan's one.

Key words: Bivariate Correlated Normal, Hypothesis on the means, distribution of absolute maximum and absolute minimum.

1. Introduction

Very often we have to face the problem to compare two treatments or the means of a variable in two different periods of time. We are interested in verifying if the mean of the variable in the second period or in another situation is changed and the direction of the change.

First of all, in § 2, we shall consider a stepwise procedure proposed by Duncan reported in Miller.[3] Duncan observed that if one variable is very high, the statistic $\max\left\{\left|X_1\right|,\left|X_2\right|\right\}$ is greater than the critical value, even if the mean of the

[1] Dipartimento di Metodi Quantitativi per le Scienze Economico-Aziendali, Università degli Studi di Milano-Bicocca, Italy. e-mail: angiola.pollastri@unimib.it

[2] The present paper is financially supported by MURST.

[3] In Miller (1981), the year in which Duncan procedure was published is not specified. Probably it is 1947.

another variable is equal to 0. To face this problem, Duncan proposed the two stages procedure reported in § 2.

To improve the procedure proposed by Duncan we use the exact distribution of the absolute maximum, proposed by Zenga in 1979, and of the absolute minimum, studied by Pollastri and Tornaghi in 2004, of the components of the S B.C.N. r.v..

2. Two stage test based on maximum modulus

Suppose we draw a simple random sample of size *n* from a r.v. (Y_1, Y_2) ~*B.C.N.*.

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right]$$

Suppose that the variances σ_1^2 and σ_2^2 are known or the sample is very large and it is possible to use the estimates of the variances instead of the real values.

Suppose we want to verify the hypotheses

$$H_0: \ (\mu_1 = \gamma_1) \cap (\mu_2 = \gamma_2)$$

against the alternatives

a) $(\mu_1 = \gamma_1) \cap (\mu_2 > \gamma_2)$ *b*) $(\mu_1 = \gamma_1) \cap (\mu_2 < \gamma_2)$ *c*) $(\mu_1 > \gamma_1) \cap (\mu_2 > \gamma_2)$

d) $(\mu_1 > \gamma_1) \cap (\mu_2 < \gamma_2)$ *e*) $(\mu_1 > \gamma_1) \cap (\mu_2 = \gamma_2)$ *f*) $(\mu_1 < \gamma_1) \cap (\mu_2 = \gamma_2)$

g) $(\mu_1 < \gamma_1) \cap (\mu_2 > \gamma_2)$ *h*) $(\mu_1 < \gamma_1) \cap (\mu_2 < \gamma_2)$

From the observations (y_{1i}, y_{2i}) *(i=1,…,n)* obtained from the sample we can compute the estimates of the means indicated by $\overline{y_1}$ and $\overline{y}_2$.

The statistics

$$(X_1 = \frac{\overline{Y_1} - \gamma_1}{\sigma_1 / \sqrt{n}}, X_2 = \frac{\overline{Y_2} - \gamma_2}{\sigma_2 / \sqrt{n}})$$

under the null hypotheses have Standardized Bivariate Correlated Normal (SBCN) Distribution having correlation coefficient equal to ρ, that is

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]$$

Duncan, in order to verify the hypotheses that $\mu_1=\mu_2=0$ against the alternatives from a) to h) in which $\gamma_1 = \gamma_2 = 0$ proposed the following two-stage procedure. The first stage is based on

Stage 1

The critical value c' is computed in case of independence of $|X_1|$ *and* $|X_2|$. Hence $P[\max\{|X_1|,|X_2|\} \le c'] = P[|X_1| \le c'] \cdot P[|X_2| \le c'] = 1-\alpha$.

It follows $(1-\alpha')^2 = 1-\alpha$, $\alpha' = 1- \sqrt{1-\alpha}$

F. i., having fixed $1-\alpha = 0.95$, $\alpha' = 0.0253$, the critical value is $c' = 2.236$.

Having fixed a size equal to α, compare *max* $\{|X_1|,|X_2|\}$ with the *c'*

If *max* $\{|X_1|,|X_2|\} \le c'$ decide in favour of H_0 : $\mu_1=\mu_2=0$ and the procedure stops.

If *max* $\{|X_1|,|X_2|\} = |X|_{(i)} > c'$*(i=1 or 2)* decide $\mu_{(i)}>0$ or $\mu_{(i)}<0$ respectively if $X_{(i)} > 0$ or if $X_{(i)} < 0$ and proceed to stage 2.

Stage 2

Compare *min* $\{|X_1|,|X_2|\} = |X|_{(j)}$ with $z_{1-\alpha/2}$, the $(1-\alpha/2)$th quantile of the standardized Normal (j=1 or 2, $j \neq i$).

a) If $|X|_{(j)} < z_{1-\alpha/2}$, decide $|\mu|_{(j)}=0$.

b) If $|X|_{(j)} > z_{1-\alpha/2}$, decide $\mu_{(j)} >0$ or $\mu_{(j)}<0$ respectively if $X_{(j)} > 0$ or if $X_{(j)} < 0$.

3. Improvements of Duncan's procedure

Duncan's stepwise test reported in § 2 is based on Bonferroni inequality. Here, we intend to use the exact distribution of maximum and minimum modulus of the components of a SBCN (X_1, X_2).

The density function (d.f.) of the r.v. T= *max* $\{|X_1|,|X_2|\}$ is a mixture of two Arctangent d.f. with parameters $a_1 = \sqrt{\frac{1+\rho}{1-\rho}}$ and $a_2 = \sqrt{\frac{1-\rho}{1+\rho}}$ and with

proportions $\pi_1 = \frac{2}{\pi} arc\tan(a_1)$ and $\pi_2 = \frac{2}{\pi} arc\tan(a_2)$. (See Pollastri(1979), Pollastri et al.(2004)).

The Arctangent r.v. was proposed by Zenga in 1979 and has the following d.f.:

$$g(x;a) = \begin{cases} \dfrac{e^{-\frac{1}{2}x^2}}{arc\tan(a)} \int_0^{ax} e^{-\frac{1}{2}y^2} dy, & for \quad x \geq 0, \qquad a>0 \\ 0 & elsewhere. \end{cases}$$

Hence: $f_T(t) = g(t;a_1)\frac{arc\tan(a_1)}{\pi/2} + g(t;a_2)\frac{arc\tan(a_2)}{\pi/2}$

The d.f. and the Cumulative Distribution Function (C.D.F.) of the r.v. *T* for some values of $|\rho|$ are reported in Fig. 3 and in Fig. 4 respectively.

The d.f. of the r.v. *V*= *min* $\{|X_1|,|X_2|\}$ (see Pollastri et al., 2004) is given by $f_V(x) = 2(2\varphi(x)) - f_T(x)$ (for $x \geq 0$), that is a linear combination of the d.f. of a Folded Standard Normal. and of the d.f. of the v.c. *T*. . The d.f. and the C.D.F.of the r.v. *V* for some values of $|\rho|$ are reported in Fig. 3 and in Fig. 4 respectively.

Let $h(\alpha,|\rho|)$ be the $(1-\alpha)100^{th}$ percentile of the r.v. *T* that is $P\{T < h(\alpha,|\rho|\} = 1-\alpha$. The values of $h(\alpha,|\rho|)$ for some α and $|\rho|$ (Pollastri, 2003) are reported in table 1.

Let $k(\alpha,|\rho|)$ be the $(1-\alpha)100^{th}$ percentile of the r.v. *V* that is $P\{\min[|X_1|,|X_2|] < k(\alpha,|\rho|\} = 1-\alpha$ .The values of $k(\alpha,|\rho|)$ for some α and $|\rho|$ (Pollastri, 2003) are reported in table 2.

We compare the $max\{|X_1|,|X_2|\}$ with the value $h(\alpha,|\rho|)$ .If $max\{|X_1|,|X_2|\} < h(\alpha,|\rho|)$ we accept H_0 and the procedure stops.

If it happens that $\max\{|X_1|,|X_2|\} = |X|_{(i)} \geq h(\alpha,|\rho|$ $(i = 1 or 2)$ we accept the Hypotheses that $|\mu|_{(j)} \neq \gamma_{(j)}$ and we must check if *min* $\{|X_1|,|X_2|\} = |X|_{(j)} > k(\alpha,|\rho|)$ (*j* = 1 or 2).

If this is true, we accept that $\mu_{(j)} < \gamma_{(j)}$ if $X_{(j)}$ is negative and that $\mu_{(j)} > \gamma_{(j)}$ otherwise.

Let us note that when ρ is unknown, it is necessary to estimate it.

The procedure here proposed is of size equal to the pre-assigned value α (Duncan's procedure was of size greater or equal to α).

From table 2 it is possible to observe that $k(\alpha,|\rho|)$ is always less then $z_{1-\alpha/2}$ and $k(\alpha,|\rho|)$ tends to $z_{1-\alpha/2}$ as $|\rho|$ tends to zero. Then it follows that

$$P\left\{\min\left[|X_1|,|X_2|\right] > k(\alpha,|\rho|)/H_1\right\} \geq P\left\{\min\left[|X_1|,|X_2|\right] > z_{1-\alpha/2}/H_1\right\}$$

Then, following the procedure considered here we can obtain a test with a probability of accepting H_1 rightly in a higher proportion of cases than will Duncan's procedure.

4. Conclusions

The paper presents a two stages procedure to test hypothesis about the means of a Bivariate Correlated Normal which improve the two stages test proposed by Duncan. The new procedure is concerned with the exact distribution of absolute maximum and absolute minimum of the components of a Bivariate Correlated Normal. The critical values are reported. The new procedure is of exact first type error and more powerful of the one proposed by Duncan.

REFERENCES

MILLER R.G.JR (1981), *Simultaneous Statistical Inference*, Springer–Verlag, Heidelberg.

POLLASTRI A. (1979), Intervalli di confidenza simultanei asintotici per le probabilità marginali in una tabella 2x2, *Quaderni di Statistica e matematica applicata alle Scienze Economico-Sociali*, Vol. II , 45—58.

POLLASTRI A. (2003), Verifica di ipotesi sulle medie di una Distribuzione Normale Bivariata Correlata, *Working paper n. 60 Dipartimento di Metodi Quantitativi* per le *Scienze Economico e Aziendali- Università Milano-Bicocca.*

POLLASTRI A., TORNAGHI F. (2004), Some properties of the Arctangent Distribution, *Statistica e Applicazioni*, Vol.II, n.1, 3—18.

ZENGA M. (1979), L'impiego della funzione arcotangente incompleta nello studio della distribuzione asintotica dello scarto assoluto massimo di una trinomiale, *Statistica*, Vol. XXXIX, n.2, 269—286.

Table 1. Values of the $(1-\alpha/2)th$ quantile $h(\alpha,|\rho|)$ of the maximum modulus

α \ |ρ|	0,01	0,02	0,05	0,10	0,15
0,0000	2,8018	2,5717	2,2354	1,9483	1,7618
0,0250	2,8018	2,5717	2,2354	1,9482	1,7617
0,0500	2,8017	2,5716	2,2352	1,9480	1,7614
0,0750	2,8016	2,5714	2,2350	1,9476	1,7610
0,1000	2,8015	2,5712	2,2346	1,9471	1,7603
0,1250	2,8013	2,5709	2,2341	1,9464	1,7595
0,1500	2,8011	2,5706	2,2335	1,9456	1,7585
0,1750	2,8008	2,5702	2,2328	1,9446	1,7574
0,2000	2,8005	2,5697	2,2320	1,9434	1,7560
0,2250	2,8001	2,5691	2,2311	1,9421	1,7545
0,2500	2,7996	2,5684	2,2300	1,9407	1,7527
0,2750	2,7991	2,5677	2,2288	1,9390	1,7508
0,3000	2,7985	2,5668	2,2275	1,9372	1,7487
0,3250	2,7978	2,5659	2,2260	1,9352	1,7463
0,3500	2,7970	2,5648	2,2244	1,9330	1,7438
0,3750	2,7961	2,5636	2,2226	1,9307	1,7410
0,4000	2,7952	2,5623	2,2207	1,9281	1,7381
0,4250	2,7940	2,5608	2,2186	1,9254	1,7349
0,4500	2,7928	2,5592	2,2163	1,9224	1,7314
0,4750	2,7914	2,5574	2,2138	1,9192	1,7277
0,5000	2,7898	2,5554	2,2111	1,9157	1,7237
0,5250	2,7881	2,5532	2,2081	1,9120	1,7195
0,5500	2,7861	2,5508	2,2049	1,9080	1,7149
0,5750	2,7839	2,5482	2,2014	1,9037	1,7101
0,6000	2,7815	2,5452	2,1977	1,8991	1,7049
0,6250	2,7788	2,5420	2,1936	1,8941	1,6993
0,6500	2,7758	2,5384	2,1891	1,8888	1,6934
0,6750	2,7724	2,5345	2,1842	1,8830	1,6870
0,7000	2,7686	2,5301	2,1788	1,8767	1,6801
0,7250	2,7643	2,5252	2,1730	1,8699	1,6726
0,7500	2,7595	2,5198	2,1665	1,8625	1,6645
0,7750	2,7541	2,5137	2,1594	1,8544	1,6558
0,8000	2,7479	2,5068	2,1514	1,8455	1,6461
0,8250	2,7409	2,4990	2,1425	1,8355	1,6355
0,8500	2,7327	2,4900	2,1324	1,8244	1,6237
0,8750	2,7230	2,4796	2,1208	1,8118	1,6103
0,9000	2,7115	2,4673	2,1072	1,7971	1,5949
0,9250	2,6973	2,4521	2,0908	1,7796	1,5767
0,9500	2,6787	2,4327	2,0700	1,7577	1,5540
0,9750	2,6518	2,4048	2,0407	1,7272	1,5228
0,9900	2,6253	2,3775	2,0126	1,6983	1,4934
0,9999	2,5790	2,3306	1,9650	1,6502	1,4449

Table 2. Values of the $(1-\alpha/2)th$ quantile $k(\alpha,|\rho|)$of the distribution of the minimum modulus

α \ $\lvert\rho\rvert$	0,01	0,02	0,05	0,10	0,15
0,0000	1,6449	1,4705	1,2170	1,0022	0,8645
0,0250	1,6457	1,4712	1,2174	1,0025	0,8647
0,0500	1,6483	1,4731	1,2186	1,0032	0,8652
0,0750	1,6527	1,4763	1,2205	1,0044	0,8660
0,1000	1,6588	1,4809	1,2233	1,0061	0,8672
0,1250	1,6665	1,4867	1,2269	1,0083	0,8687
0,1500	1,6760	1,4939	1,2314	1,0110	0,8706
0,1750	1,6871	1,5023	1,2366	1,0143	0,8729
0,2000	1,6997	1,5120	1,2428	1,0181	0,8755
0,2250	1,7139	1,5231	1,2498	1,0225	0,8786
0,2500	1,7295	1,5353	1,2578	1,0274	0,8820
0,2750	1,7464	1,5488	1,2666	1,0330	0,8860
0,3000	1,7644	1,5635	1,2765	1,0393	0,8904
0,3250	1,7836	1,5793	1,2873	1,0463	0,8953
0,3500	1,8036	1,5961	1,2990	1,0540	0,9007
0,3750	1,8244	1,6139	1,3117	1,0624	0,9068
0,4000	1,8459	1,6325	1,3254	1,0717	0,9135
0,4250	1,8679	1,6519	1,3400	1,0818	0,9208
0,4500	1,8903	1,6719	1,3556	1,0928	0,9289
0,4750	1,9131	1,6924	1,3720	1,1047	0,9377
0,5000	1,9361	1,7135	1,3892	1,1175	0,9474
0,5250	1,9593	1,7349	1,4071	1,1312	0,9579
0,5500	1,9828	1,7567	1,4257	1,1458	0,9694
0,5750	2,0065	1,7788	1,4449	1,1614	0,9818
0,6000	2,0304	1,8012	1,4647	1,1779	0,9953
0,6250	2,0545	1,8238	1,4850	1,1952	1,0098
0,6500	2,0788	1,8468	1,5057	1,2133	1,0253
0,6750	2,1035	1,8701	1,5269	1,2322	1,0419
0,7000	2,1285	1,8937	1,5486	1,2518	1,0595
0,7250	2,1539	1,9178	1,5707	1,2720	1,0780
0,7500	2,1798	1,9424	1,5934	1,2930	1,0975
0,7750	2,2062	1,9676	1,6167	1,3147	1,1179
0,8000	2,2334	1,9934	1,6407	1,3372	1,1393
0,8250	2,2613	2,0202	1,6657	1,3606	1,1617
0,8500	2,2904	2,0480	1,6918	1,3852	1,1853
0,8750	2,3208	2,0773	1,7193	1,4112	1,2104
0,9000	2,3531	2,1084	1,7487	1,4392	1,2375
0,9250	2,3880	2,1422	1,7809	1,4700	1,2674
0,9500	2,4271	2,1803	1,8175	1,5052	1,3016
0,9750	2,4746	2,2268	1,8624	1,5488	1,3444
0,9900	2,5134	2,2652	1,9000	1,5856	1,3806
0,9999	2,5670	2,3194	1,9537	1,6389	1,4337

Figure 1. Density functions of the maximum modulus of S.B.C.N. with $|\rho|$=0, 0.3, 0.6, 0.9

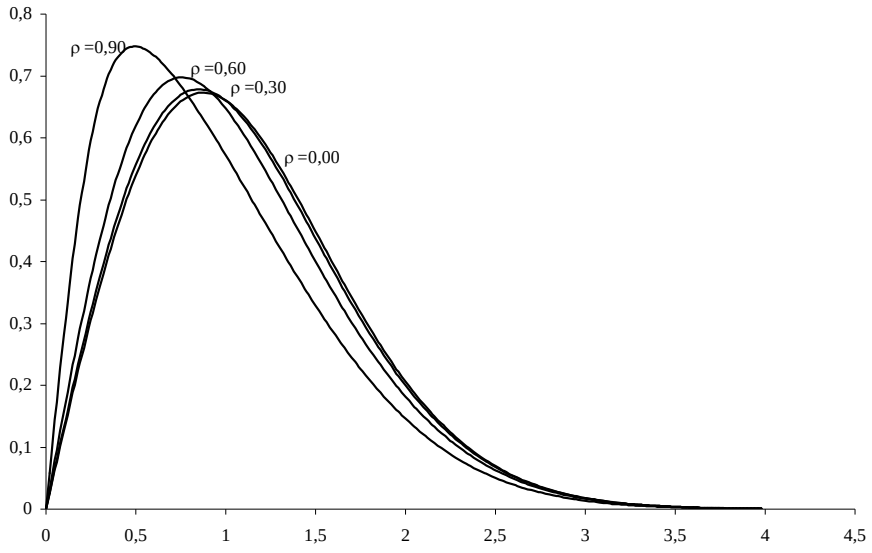


Figure 2. Cumulative Distribution Functions of the maximum modulus of the components of a S.B.C.N.

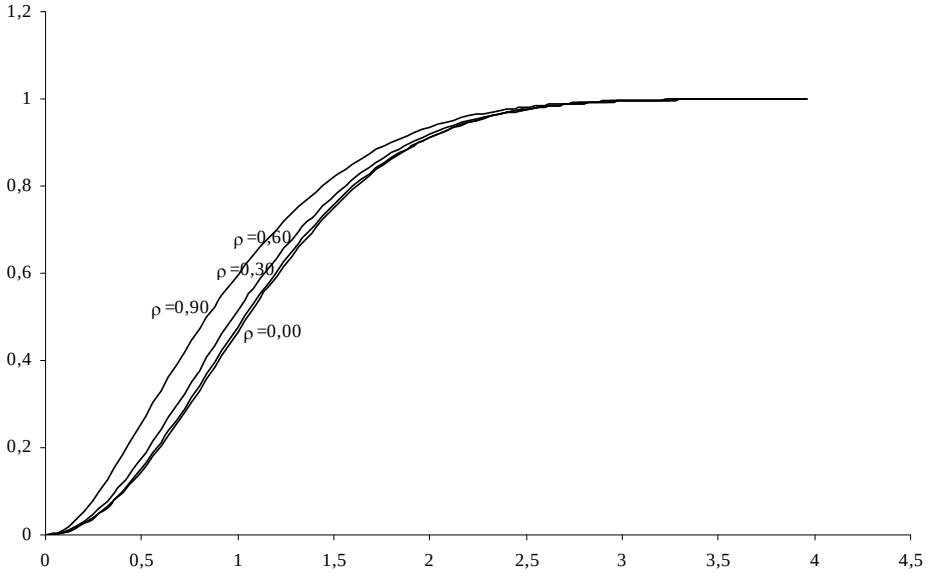


Figure 3. Density functions of the minimum modulus of S.B.C.N. with $|\rho|$ =0, 0.3, 0.6, 0.9

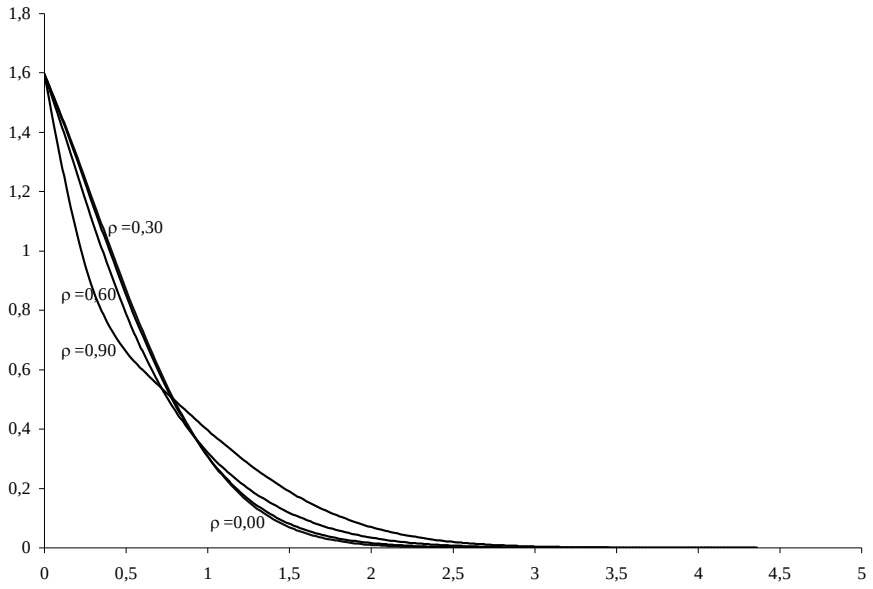
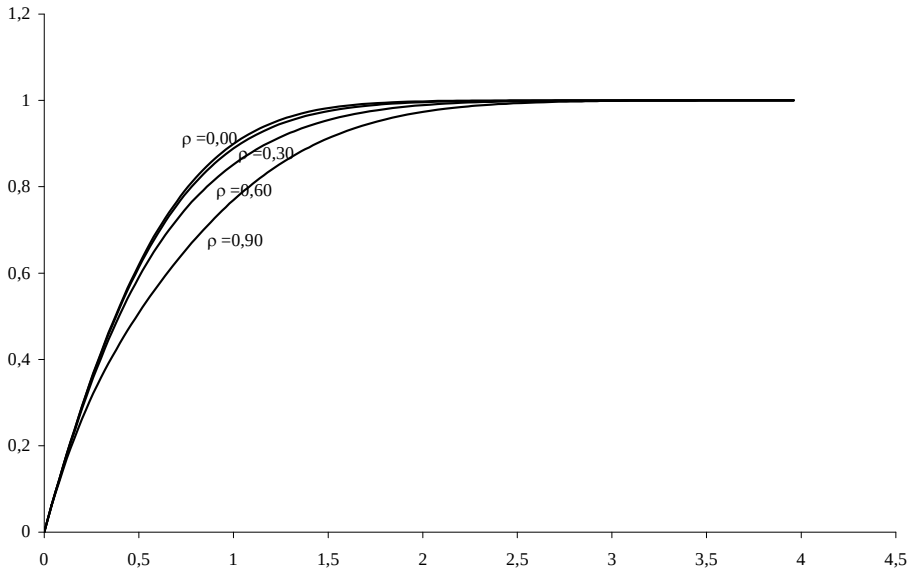


Figure 4. Cumulative Distribution Functions of the minimum modulus of the components of a S.B.C.N.



	ro= 0,00	ro= 0,30	ro= 0,60	ro= 0,90
0	0	0	0	0
0,02	0,02546	0,02669	0,03182	0,05834
0,04	0,05088	0,05333	0,06356	0,11619
0,06	0,07621	0,07987	0,09514	0,17307
0,08	0,10143	0,10628	0,12648	0,22853
0,1	0,12648	0,1325	0,15751	0,28214
0,12	0,15133	0,15849	0,18815	0,33353
0,14	0,17594	0,1842	0,21833	0,38235
0,16	0,20027	0,20959	0,24798	0,42833
0,18	0,22429	0,23462	0,27702	0,47124
0,2	0,24795	0,25925	0,3054	0,51091
0,22	0,27123	0,28343	0,33306	0,54725
0,24	0,29408	0,30713	0,35993	0,58019
0,26	0,31647	0,33031	0,38597	0,60975
0,28	0,33838	0,35293	0,41112	0,63597
0,3	0,35976	0,37496	0,43534	0,65895
0,32	0,38059	0,39637	0,45859	0,67881

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 139—158

A GENERAL PROCEDURE FOR ESTIMATING VARIOUS MEASURES OF NORMAL DISTRIBUTION USING PRIOR KNOWLEDGE OF STANDARD DEVIATION

Housila P. Singh[1] **and Vankim Chander**[2]

ABSTRACT

This paper presents a class of shrunken estimators for estimating the α-th power of σ, the standard deviation of a normal distribution (i.e σ^{α}, α being an integer) using prior information (or guessed value) (say σ_0^{α}) of σ^{α} and hence considers the estimation of the general parameter $G(\delta,\alpha)=\delta\sigma^{\alpha}$ (δ (>0) being a known parameter). A class of shrunken estimators for σ^{α} using σ_0^{α} has been suggested with its properties. In particular, the problem of estimating the inverse of variance has been considered. Numerical illustrations are given in support of the present study. Simulation study confirms the high efficiency of the developed classes of shrunken estimators when compared with the usual unbiased estimators and minimum mean squared error (MMSE) estimators.

Key words: Bias, Mean Squared Error, Percent Relative Efficiency (PRE), Prior information, Guass-laguerre integration method, Normal distribution.

1. Introduction

The Guassian (Normal) distribution has dominated statistical practice as well as theory. It is known that any variable whose expression results from the additive contribution of many small effects will tend to be normally distributed. Also for measurements whose distributions are not normal, a simple transformation of the scale of measurement may induce approximate normality. The square root, and logarithm, ln(x) are often used as such transformations. Thus, from statistical

[1] hpsujn@rediffmail.com
[2] vcupadhyay06@yahoo.co.in

point of view, Normal distribution is very important and hence the estimation of its parameters deserves special attention. Let $x_{1,}\, x_{2,\,\ldots\,,}\, x_n$ be a random sample of size n, drawn from normal distribution, the probability distribution function (pdf) of which is given by

$$f(x;\mu,\sigma^2)=\begin{cases}\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, & -\infty<x<\infty,\ -\infty<\mu<\infty,\ \sigma>0\\ 0, & \textit{otherwise}.\end{cases} \quad (1.1)$$

where μ is the mean and σ is the standard deviation.

The problem under investigation is to estimate the general parameter defined by

$$\tau\left(\omega^*,\alpha\right)=\omega^*\,\sigma^{\alpha} \quad (1.2)$$

where ω^* is a known real positive constant and α is an integer (may be positive or negative).

Table 1.1. Showing some unknown parameters which are the function of σ

Parametric Functions		Values of (δ, α)
Standard Deviation	$\left(\sigma\right)$	(1, 1)
Fisher Information	$\left(\frac{1}{\sigma^2}\right)$	(1, -2)
Fourth Moment about Mean	$\left(3\sigma^4\right)$	(3, 4)
Variance	$\left(\sigma^2\right)$	(1, 2)
Mean Deviation about Mean	$\left(\sigma\sqrt{\frac{2}{\pi}}\right)$	$\left(\sqrt{\frac{2}{\pi}},1\right)$
* Process Capability Index	$\left(C_p\right)$	$\left(\left\{\frac{(USL-LSL)}{6}\right\},\ -1\right)$
Precision of the Sample Mean	$\left(\frac{n}{\sigma^2}\right)$	(n, -2)

*Note : * USL and LSL stand for Upper specification Limit and Lower Specification Limit.*

Using (1.2) for different values of ω^* and α we get different parametric functions some of which given by Table 1.1.

It is obvious from (1.2) that the general parameter $\tau\left(\omega^*,\alpha\right)$ depends on σ^{α} which is unknown thus the problem of estimation of σ^{α} deserves special attention.

The usual unbiased estimator of σ^{α} is given by

$$J_{(u,\alpha)} = A_{(n,\alpha)}\, s^{\alpha} \quad , \tag{1.3}$$

where $A_{(n,\alpha)} = \left\{\frac{n-1}{2}\right\}^{\frac{\alpha}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n+\alpha-1}{2}\right)}$ (1.4)

and $s = \sqrt{\left\{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}}$ with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x$. (1.5)

The relative variance (RV) of $J_{(u,\alpha)}$ is given by

$$RV\left\{J_{(u,\alpha)}\right\} = \left\{B_{(n,\alpha)} - 1\right\}$$

$$= \nu^{(n,\alpha)} \quad , \tag{1.6}$$

where $\nu^{(n,\alpha)} = \left\{B_{(n,\alpha)} - 1\right\}$ (1.7)

and $B_{(n,\alpha)} = \frac{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n+2\alpha-1}{2}\right)}{\Gamma^2\left(\frac{n+\alpha-1}{2}\right)}$. (1.8)

The minimum mean squared error (MMSE) estimator of σ^{α} in the class $P\,J_{(u,\alpha)}$ (*P*, being suitably chosen constant such that mean squared error (MSE) of $P\,J_{(u,\alpha)}$ is minimum) is given by

$$J_{(MMSE,\alpha)} = W_{(n,\alpha)}\; s^{\alpha} \quad , \tag{1.9}$$

where $W_{(n,\alpha)} = \left\{\frac{n-1}{2}\right\}^{\frac{\alpha}{2}} \frac{\Gamma\left(\frac{n+\alpha-1}{2}\right)}{\Gamma\left(\frac{n+2\alpha-1}{2}\right)}$. (1.10)

The absolute relative bias (ARB) and relative mean squared error (RMSE) be respectively given as

$$ARB\left\{J_{(MMSE,\alpha)}\right\} = \left|\left\{B_{(n,\alpha)}^{-1} - 1\right\}\right| \tag{1.11}$$

and

$$RMSE\left\{J_{(MMSE,\alpha)}\right\} = \left\{1 - B_{(n,\alpha)}^{-1}\right\}$$

$$= \left\{ \frac{\nu^{(n,\alpha)}}{1+\nu^{(n,\alpha)}} \right\}, \qquad (1.12)$$

where $\nu^{(n,\alpha)}$ is given by (1.7).

In the estimation of unknown parameter σ^{α} there often exists some form of prior knowledge about the parameter due to long association with the experimental material or from other empirical investigations from extraneous sources, which one would like to utilize in order to get better estimate. The estimation problem using some a priori information regarding some population parameters have been investigated, for instance see Thompson (1968), Mehta and Srinivasan (1971), Pandey and Singh (1979), Pandey (1979), Singh and Singh(1997), Singh et al. (1999), Singh and Saxena (2003), Saxena and Singh (2004). The present investigation is speculated to propose a class of shrinkage estimators for estimating the general parameter σ^{α} on the line of Thompson (1968) when a prior or guessed value, say point estimate σ_0 of standard deviation σ is available and discuss its properties. Numerical computations in term of percent relative efficiency have been made. Subsequently, modified class of estimators is also developed by shrinkage towards a point and its properties are discussed. In particular, a class of shrinkage estimators of inverse of variance is defined along with its properties.

2. The new suggested class of shrinkage estimators for estimating σ^{α}

When a prior point estimate or guessed value σ_0 of the parameter σ is available, we suggest a class of shrinkage estimators $\hat{T}_{(s,\alpha)}$ (say) for the parameter σ^{α} in the model (1.1) as

$$\hat{T}_{(s,\alpha)} = \ddot{\psi} J_{(u,\alpha)} + \{1-\ddot{\psi}\}\sigma_0^{\alpha} = \ddot{\psi}\left\{J_{(u,\alpha)} - \sigma_0^{\alpha}\right\} + \sigma_0^{\alpha}, \qquad (2.1)$$

where $\ddot{\psi}$ is a scalar ($0 \le \ddot{\psi} \le 1$) specified by the experimenter according to his belief in σ_0^{α} (i.e. in σ_0). A value of $\ddot{\psi}$ near to 'zero' implies a strong belief in σ_0 and a value of $\ddot{\psi}$ near to 'one' shows a strong belief in sample estimate $J_{(u,\alpha)}$.

The ARB and RMSE of $\hat{T}_{(s,\alpha)}$ are respectively given by

$$ARB\left\{\hat{T}_{(s,\alpha)}\right\} = \left|(1-\ddot{\psi})\{\Lambda_{(\alpha)} - 1\}\right| \qquad (2.2)$$

and $RMSE\left\{\hat{T}_{(s,\alpha)}\right\} = \ddot{\psi}^2\, v^{(n,\alpha)} + \left(1-\ddot{\psi}\right)^2\left\{\Lambda_{(\alpha)}-1\right\}^2$, (2.3)

where $\Lambda_{(\alpha)} = \dfrac{\sigma_0^{\alpha}}{\sigma^{\alpha}}$, $v^{(n,\alpha)} = \left\{B_{(n,\alpha)} - 1\right\}$ and $B_{(n,\alpha)} = \dfrac{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n+2\alpha-1}{2}\right)}{\Gamma^2\left(\frac{n+\alpha-1}{2}\right)}$.

The quantity $\Lambda_{(\alpha)}\left(=\frac{\sigma_0^{\alpha}}{\sigma^{\alpha}}\right)$ shows the departure ratio of natural origin σ_0^{α} from its true value. From (1.6) and (2.3), the suggested class of estimators $\hat{T}_{(s,\alpha)}$ are better than the usual unbiased estimator $\hat{T}_{(u,\alpha)}$ if

$$\frac{\left\{\left(\Lambda_{(\alpha)}-1\right)^2 - v^{(n,\alpha)}\right\}}{\left\{\left(\Lambda_{(\alpha)}-1\right)^2 + v^{(n,\alpha)}\right\}} < \ddot{\psi} \le 1 \quad . \tag{2.4}$$

Further, from (1.12) and (2.3), we have

$$RMSE\left\{J_{(MMSE,\alpha)}\right\} - RMSE\left\{\hat{T}_{(s,\alpha)}\right\} =$$

$$\left[\left\{\frac{v^{(n,\alpha)}}{1-v^{(n,\alpha)}}\right\} - \ddot{\psi}^2\, v^{(n,\alpha)} - \left(1-\ddot{\psi}\right)^2\left\{\Lambda_{(\alpha)}-1\right\}^2\right] > 0$$

$$\left[\frac{\left\{\Lambda_{(\alpha)}-1\right\}^2}{\varpi^{**}} - \frac{v^{(n,\alpha)}}{\varpi^{**}}\sqrt{\frac{\Lambda_{(\alpha)}(2-\Lambda_{(\alpha)})}{\left(1+v^{(n,\alpha)}\right)}}\right] < \ddot{\psi} < \left[\frac{\left\{\Lambda_{(\alpha)}-1\right\}^2}{\varpi^{**}} + \frac{v^{(n,\alpha)}}{\varpi^{**}}\sqrt{\frac{\Lambda_{(\alpha)}\{2-\Lambda_{(\alpha)}\}}{\left(1+v^{(n,\alpha)}\right)}}\right] \tag{2.5}$$

where $\varpi^{**} = \left\{v^{(n,\alpha)} + \{\Lambda_{(\alpha)}-1\}^2\right\}$, $\Lambda_{(\alpha)} \in (0,\, 2)$.

2.1. Percent Relative Efficiency

To elucidate the performance of the proposed family of estimator $\hat{T}_{(s,\alpha)}$ with usual unbiased estimator $J_{(u,\alpha)}$ and MMSE estimator $J_{(MMSE,\alpha)}$, the percent relative efficiencies (PREs) of $\hat{T}_{(s,\alpha)}$ with respect to $J_{(u,\alpha)}$ and $J_{(MMSE,\alpha)}$ have been computed by respectively using the following formulae:

$$PRE\left\{\hat{T}_{(s,\alpha)}, J_{(u,\alpha)}\right\} = \frac{MSE\left\{J_{(u,\alpha)}\right\}}{MSE\left\{\hat{T}_{(s,\alpha)}\right\}} * 100$$

$$= \left[\ddot{\psi}^2 + \frac{(1-\ddot{\psi})^2 \left\{ \Lambda_{(\alpha)} - 1 \right\}^2}{v^{(n,\alpha)}} \right]^{-1} *100 \qquad (2.6)$$

and

$$PRE\left\{ \hat{T}_{(s,\alpha)}, J_{(MMSE,\alpha)} \right\} = \frac{MSE\left\{ J_{(MMSE,\alpha)} \right\}}{MSE\left\{ \hat{T}_{(s,\alpha)} \right\}} *100$$

$$= \frac{v^{(n,\alpha)}}{\left\{ v^{(n,\alpha)} \ddot{\psi}^2 + (1-\ddot{\psi})^2 \left\{ \Lambda_{(\alpha)} - 1 \right\}^2 \right\} \left\{ 1 + v^{(n,\alpha)} \right\}} *100 \ . \qquad (2.7)$$

3. Estimation of inverse of variance

In particular we have considered the problem of estimating the inverse of variance. The usual unbiased estimator $\hat{T}_{(u,-2)}$ of $\frac{1}{\sigma^2}$ is given by

$$\hat{T}_{(u,-2)} = \left\{ \frac{n-3}{n-1} \right\} \frac{1}{s^2} \qquad (3.1)$$

Putting $\alpha = -2$ in (1.6) the RV of $\hat{T}_{(u,-2)}$ as

$$RV\left\{ \hat{T}_{(u,-2)} \right\} = \left\{ \frac{2}{n-5} \right\} \ . \qquad (3.2)$$

The MMSE estimator $\hat{T}_{(MMSE,-2)}$ of $\frac{1}{\sigma^2}$ in the class $P' \hat{T}_{(u,-2)}$ (P' , being suitably chosen constant such that MSE of $P' \hat{T}_{(u,-2)}$ is minimum) is given by

$$\hat{T}_{(MMSE,-2)} = \left\{ \frac{n-5}{n-1} \right\} \frac{1}{s^2} \qquad (6.3.3)$$

with RMSE

$$RMSE\left\{ \hat{T}_{(MMSE,-2)} \right\} = \left\{ \frac{2}{n-3} \right\} \ , \qquad (3.4)$$

[see, Mishra (1985)].

Putting $\alpha = -2$ in (2.1), we get a class of shrinkage estimators of $\frac{1}{\sigma^2}$ as

$$\hat{T}_{(s,-2)} = \ddot{\psi}\hat{T}_{(u,-2)} + \{1 - \ddot{\psi}\}\sigma_0^{-2}$$

$$= \ddot{\psi}\left\{\hat{T}_{(u,-2)} - \sigma_0^{-2}\right\} + \sigma_0^{-2} \ ,$$

$$= \ddot{\psi}\left\{\frac{n-3}{n-1}\right\}\frac{1}{s^2} + (1 - \ddot{\psi})\frac{1}{\sigma_0^2} \qquad (3.5)$$

Putting $\alpha = -2$ in (2.2) and (2.3) we get the ARB and RMSE of $\hat{T}_{(s,-2)}$ are respectively given by

$$ARB\left\{\hat{T}_{(s,-2)}\right\} = \left|(1 - \ddot{\psi})\{\Lambda_{(-2)} - 1\}\right| \quad (3.6)$$

and

$$RMSE\left\{\hat{T}_{(s,-2)}\right\} = \ddot{\psi}^2 v^{(n,-2)} + \{1 - \ddot{\psi}\}^2\{\Lambda_{(-2)} - 1\}^2 \ . \qquad (3.7)$$

3.1.Empirical Study

The range of dominance of $\ddot{\psi}$ has been computed for different values of n and $\Lambda_{(-2)}$ in which the proposed estimator $\hat{T}_{(s,-2)}$ is better than the usual unbiased estimator $\hat{T}_{(u,-2)}$ and MMSE estimator $\hat{T}_{(MMSE,-2)}$ using formulae (2.4) and(2.5) for $\alpha = -2$ respectively (see, Table 3.1). The percent relative efficiencies of $\hat{T}_{(s,-2)}$ with respect to $\hat{T}_{(u,-2)}$ and $\hat{T}_{(MMSE,-2)}$ have also been computed using the formula (2.6) and (2.7) respectively for different values of n = 11(2)23 and $\Lambda_{(-2)}$ = 0.3(0.1)1.7 , (see, Tables 3.2 and 3.3).

It is observed from Tables 3.2 and 3.3 and extended computation that the proposed estimator is better than the usual unbiased estimator $\hat{T}_{(u,-2)}$:

(i) When $\ddot{\psi} = \frac{1}{4}$, $\frac{1}{2} \le \Lambda_{(-2)} \le 1.5$ and $6 \le n \le 17$.

(ii) When $\ddot{\psi} = \frac{1}{2}$, $0.3 \le \Lambda_{(-2)} \le 1.7$ and $6 \le n \le 17$.

(iii) When $\ddot{\psi} = \frac{3}{4}$, $0 \le \Lambda_{(-2)} \le 2$ and $6 \le n \le 19$.

For fixed $\ddot{\psi}$ decreasing trend is observed in the range of dominance of $\Lambda_{(-2)}$ as n increases. For fixed $\ddot{\psi}$ and n the value of PRE decreases as $\Lambda_{(-2)}$ goes away from unity. A common range of $\Lambda_{(-2)}$ for which the estimator $\hat{T}_{(s,-2)}$ is more efficient than $\hat{T}_{(u,-2)}$ is $0.52 \le \Lambda_{(-2)} \le 1.48$ for $6 \le n \le 17$ and $\frac{1}{4} \le \ddot{\psi} \le \frac{3}{4}$. The larger gain in efficiency is observed when $\ddot{\psi}$ is closer to 'zero' but for a smaller range of $\Lambda_{(-2)}$. When $\ddot{\psi}$ approaches to unity there is moderate gain in efficiency but for a wider range of $\Lambda_{(-2)}$ even for large values of n.

A similar trend is observed (Table 3.3) for $PRE\left\{\hat{T}_{(s,\alpha)}, \hat{T}_{(MMSE,\alpha)}\right\}$. The gain in efficiency by using $\hat{T}_{(s,-2)}$ over MMSE estimator $\hat{T}_{(MMSE,-2)}$ is fewer than by using $\hat{T}_{(s,-2)}$ over unbiased estimator $\hat{T}_{(u,-2)}$.

Thus, there is enough scope for selecting the scalar $\ddot{\psi}$ to get better estimator than the unbiased estimator $\hat{T}_{(u,-2)}$ and MMSE estimator $\hat{T}_{(MMSE,-2)}$

4. Estimation of the departure ratio $\Lambda_{(\alpha)}$

The quantity $\Lambda_{(\alpha)}$ depends upon the unknown parameter σ^{α} , so the unbiased estimator of $\Lambda_{(\alpha)}$ is to be recommended in practice and is define as

$$\tilde{\Lambda}_{(\alpha)} = \frac{\sigma_0^{\alpha}}{s^{\alpha}} \left(\frac{2}{n-1}\right)^{\frac{\alpha}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-\alpha-1}{2}\right)} \tag{4.1}$$

for different α. For the situation of estimating the inverse of variance, we suggest the following unbiased estimator for $\Lambda_{(-2)}$

$$\tilde{\Lambda}_{(-2)} = \frac{s^2}{\sigma_0^2} \tag{4.2}$$

Table 3.1. Range of $\ddot{\psi}$ for which the proposed estimator $\hat{T}_{(s,-2)}$ is better than $\hat{T}_{(u,-2)}$ and MMSE estimator $\hat{T}_{(MMSE,-2)}$

$\Lambda_{(-2)}$ $\downarrow$n$\rightarrow$	11	13	15	17	19	21
0.3	(0.19,1.00) **(0.34,0.85)**	(0.32,1.00) **(0.45,0.88)**	(0.42,1.00) **(0.52,0.90)**	(0.49,1.00) **(0.58,0.91)**	(0.55,1.00) **(0.62,0.93)**	(0.59,1.00) **(0.66,0.93)**
0.4	(0.04,1.00) **(0.19,0.85)**	(0.18,1.00) **(0.30,0.88)**	(0.29,1.00) **(0.38,0.90)**	(0.37,1.00) **(0.45,0.92)**	(0.43,1.00) **(0.50,0.93)**	(0.48,1.00) **(0.55,0.94)**
0.5	(0.00,1.00) **(0.00,0.86)**	(0.00,1.00) **(0.11,0.89)**	(0.11,1.00) **(0.20,0.91)**	(0.20,1.00) **(0.28,0.92)**	(0.27,1.00) **(0.34,0.93)**	(0.33,1.00) **(0.39,0.94)**
0.6	(0.00,1.00) **(0.00,0.86)**	(0.00,1.00) **(0.00,0.89)**	(0.00,1.00) **(0.00,0.91)**	(0.00,1.00) **(0.06,0.92)**	(0.06,1.00) **(0.12,0.93)**	(0.12,1.00) **(0.18,0.94)**
0.7	(0.00,1.00) **(0.00,0.86)**	(0.00,1.00) **(0.00,0.89)**	(0.00,1.00) **(0.00,0.91)**	(0.00,1.00) **(0.00,0.92)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.94)**
0.8	(0.00,1.00) **(0.00,0.86)**	(0.00,1.00) **(0.00,0.89)**	(0.00,1.00) **(0.00,0.91)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.94)**
0.9	(0.00,1.00) **(0.00,0.86)**	(0.00,1.00) **(0.00,0.89)**	(0.00,1.00) **(0.00,0.91)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.94)**
1.0	(0.00,1.00) **(0.00,0.86)**	(0.00,1.00) **(0.00,0.89)**	(0.00,1.00) **(0.00,0.91)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.94)**
1.1	(0.00,1.00) **(0.00,0.86)**	(0.00,1.00) **(0.00,0.89)**	(0.00,1.00) **(0.00,0.91)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.94)**
1.2	(0.00,1.00) **(0.00,0.86)**	(0.00,1.00) **(0.00,0.89)**	(0.00,1.00) **(0.00,0.91)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.94)**
1.3	(0.00,1.00) **(0.00,0.86)**	(0.00,1.00) **(0.00,0.89)**	(0.00,1.00) **(0.00,0.91)**	(0.00,1.00) **(0.00,0.92)**	(0.00,1.00) **(0.00,0.93)**	(0.00,1.00) **(0.00,0.94)**
1.4	(0.00,1.00) **(0.00,0.86)**	(0.00,1.00) **(0.00,0.89)**	(0.00,1.00) **(0.00,0.91)**	(0.00,1.00) **(0.06,0.92)**	(0.06,1.00) **(0.12,0.93)**	(0.12,1.00) **(0.18,0.94)**
1.5	(0.00,1.00) **(0.00,0.86)**	(0.00,1.00) **(0.11,0.89)**	(0.11,1.00) **(0.20,0.91)**	(0.20,1.00) **(0.28,0.92)**	(0.27,1.00) **(0.34,0.93)**	(0.33,1.00) **(0.39,0.94)**
1.6	(0.04,1.00) **(0.19,0.85)**	(0.18,1.00) **(0.30,0.88)**	(0.29,1.00) **(0.38,0.90)**	(0.37,1.00) **(0.45,0.92)**	(0.43,1.00) **(0.50,0.93)**	(0.48,1.00) **(0.55,0.94)**
1.7	(0.19,1.00) **(0.34,0.85)**	(0.32,1.00) **(0.45,0.88)**	(0.42,1.00) **(0.52,0.90)**	(0.49,1.00) **(0.58,0.91)**	(0.55,1.00) **(0.62,0.93)**	(0.59,1.00) **(0.66,0.93)**
1.8	(0.32,1.00) **(0.48,0.84)**	(0.44,1.00) **(0.57,0.87)**	(0.52,1.00) **(0.63,0.79)**	(0.59,1.00) **(0.68,0.91)**	(0.64,1.00) **(0.72,0.92)**	(0.67,1.00) **(0.74,0.93)**
1.9	(0.42,1.00) **(0.60,0.82)**	(0.53,1.00) **(0.67,0.86)**	(0.60,1.00) **(0.72,0.88)**	(0.66,1.00) **(0.76,0.91)**	(0.70,1.00) **(0.79,0.91)**	(0.73,1.00) **(0.81,0.92)**

*The **bold** figures in parentheses show the range of $\ddot{\psi}$ in which the proposed estimator is better than MMSE estimator.*

Table 3.2. PRE of developed class of shrinkage estimator $\hat{T}_{(s,-2)}$ with respect to $\hat{T}_{(u,-2)}$ for different $\Lambda_{(-2)}$, $\ddot{\psi}$ =0.25,0.5,0.75 and n = 11(2)23

$\ddot{\psi}$	$\Lambda_{(-2)}\downarrow$ n$\rightarrow$	11	13	15	17	19	21	23
0.25	0.3	112.44	85.84	69.41	58.27	50.2	44.10	39.32
	0.4	149.25	114.61	93.02	78.28	67.57	59.44	53.05
	0.5	206.45	160	130.61	110.34	95.52	84.21	75.29
	0.7	466.47	377.36	316.83	273.04	239.88	213.90	193.0
	0.8	769.23	655.74	571.43	506.33	454.55	412.37	377.36
	0.9	1259.84	1176.47	1103.45	1038.96	981.6	930.2	883.98
	1.0	1600	1600	1600	1600	1600	1600	1600
	1.1	1259.84	1176.47	1103.45	1038.96	981.6	930.2	883.98
	1.2	769.23	655.74	571.43	506.33	454.55	412.37	377.36
	1.3	466.47	377.36	316.83	273.04	239.88	213.90	193.0
	1.5	206.45	160	130.61	110.34	95.52	84.21	75.29
	1.6	149.25	114.61	93.02	78.28	67.57	59.44	53.05
	1.7	112.44	85.84	69.41	58.27	50.2	44.10	39.32
	Range of $\Lambda_{(-2)}$	[0.26,1.74]	[0.36,1.64]	[0.43,1.57]	[0.48,1.52]	[0.52,1.48]	[0.55,1.45]	[0.6,1.4]
0.5	0.3	161.94	135.14	115.94	101.52	90.29	81.30	73.94
	0.4	192.31	163.93	142.86	126.58	113.64	103.09	94.34
	0.5	228.57	200	177.78	160	145.45	133.33	123.08
	0.7	314.96	294.12	275.86	259.74	245.4	232.56	220.99
	0.8	357.14	344.83	333.33	322.58	312.5	303.03	294.12
	0.9	314.96	294.12	275.86	259.74	245.4	370.37	220.99
	1.0	400	400	400	400	400	400	400
	1.1	388.35	384.62	380.95	377.36	373.83	370.37	220.99
	1.2	357.14	344.83	333.33	322.58	312.5	303.03	294.12
	1.3	314.96	294.12	275.86	259.74	245.4	232.56	123.08
	1.5	228.57	200	177.78	160	145.45	133.33	94.34
	1.6	192.31	163.93	142.86	126.58	113.64	103.09	94.34
	1.7	161.94	135.14	115.94	101.52	90.29	81.30	73.94
	Range of $\Lambda_{(-2)}$	[0,2]	[0.14,1.86]	[0.23,1.77]	[0.3,1.7]	[0.35,1.65]	[0.4,1.6]	[0.45,1.55]
0.75	0.3	152.82	145.99	139.74	134	128.72	123.84	119.31
	0.4	158.73	153.26	148.15	143.37	138.89	134.68	130.72
	0.5	164.1	160	156.1	152.38	148.84	145.45	142.22
	0.7	172.6	170.94	169.31	167.71	166.15	164.61	163.10
	0.8	175.44	174.67	173.91	173.16	172.41	171.67	170.947
	0.9	177.19	176.99	176.8	176.6	176.41	176.21	176.02
	1.0	177.78	177.78	177.78	177.78	177.78	177.78	177.78
	1.1	177.19	176.99	176.8	176.6	176.41	176.21	176.02
	1.2	175.44	174.67	173.91	173.16	172.41	171.67	170.947
	1.3	172.6	170.94	169.31	167.71	166.15	164.61	163.10
	1.5	164.1	160	156.1	152.38	148.84	145.45	142.22
	1.6	158.73	153.26	148.15	143.37	138.89	134.68	130.72
	1.7	152.82	145.99	139.74	134	128.72	123.84	119.31
	Range of $\Lambda_{(-2)}$	[0, 2.5]	[0, 2.3]	[0, 2.1]	[0, 2]	[0, 2]	[0.1,1.9]	[0.15,1.85]

Table 3.3. PRE of developed class of shrinkage estimator $\hat{T}_{(s,-2)}$ with respect to $\hat{T}_{(MMSE,-2)}$ for different $\Lambda_{(-2)}$, $\ddot{\psi}$ =0.25,0.5,0.75 and n = 11(2)23

$\ddot{\psi}$	$\Lambda_{(-2)}\downarrow$ n$\rightarrow$	11	13	15	17	19	21	23
0.25	0.3	84.33	68.67	57.85	49.94	43.93	39.20	35.39
	0.4	111.94	91.69	77.52	67.1	59.12	52.83	47.75
	0.5	154.84	128	108.84	94.58	83.58	74.85	67.76
	0.7	349.85	301.89	264.03	234.03	209.9	190.14	173.70
	0.8	576.92	524.59	476.19	434	397.73	366.55	339.62
	0.9	944.88	941.18	919.54	890.54	858.9	826.87	795.58
	1.0	1200	1280	1333.33	1371.43	1400	1422.22	1440
	1.1	944.88	941.18	919.54	890.54	858.9	826.87	795.58
	1.2	576.92	524.59	476.19	434	397.73	366.55	339.62
	1.3	349.85	301.89	264.03	234.03	209.9	190.14	173.70
	1.5	154.84	128	108.84	94.58	83.58	74.85	67.76
	1.6	111.94	91.69	77.52	67.1	59.12	52.83	47.75
	1.7	84.33	68.67	57.85	49.94	43.93	39.20	35.39
	Range of $\Lambda_{(-2)}$	[0.37,1.63]	[0.43,1.57]	[0.48,1.52]	[0.52,1.48]	[0.55,1.45]	[0.58,1.42]	[0.6,1.4]
0.5	0.3	121.46	108.11	96.62	87.02	79.01	72.27	66.54
	0.4	144.23	131.15	119.05	108.5	99.43	91.64	84.91
	0.5	171.43	160	148.15	137.14	127.27	118.52	110.77
	0.7	236.22	235.29	229.89	222.63	214.72	206.72	198.90
	0.8	267.86	275.86	277.78	276.5	273.44	269.36	264.71
	0.9	291.26	307.69	317.46	323.45	327.1	329.22	330.28
	1.0	300	320	333.33	342.86	350	355.56	360
	1.1	291.26	307.69	317.46	323.45	327.1	329.22	330.28
	1.2	267.86	275.86	277.78	276.5	273.44	269.36	264.71
	1.3	236.22	235.29	229.89	222.63	214.72	206.72	198.90
	1.5	171.43	160	148.15	137.14	127.27	118.52	110.77
	1.6	144.23	131.15	119.05	108.5	99.43	91.64	84.91
	1.7	121.46	108.11	96.62	87.02	79.01	72.27	66.54
	Range of $\Lambda_{(-2)}$	[0.19,1.81]	[0.26,1.74]	[0.32,1.68]	[0.37,1.63]	[0.41,1.59]	[0.44,1.64]	[0.47,1.53]
0.75	0.3	114.61	116.79	116.45	114.86	112.63	110.08	107.38
	0.4	119.05	122.61	123.46	122.89	121.53	119.72	117.65
	0.5	123.08	128	130.08	130.61	130.23	129.29	128
	0.7	129.45	136.75	141.09	143.76	145.38	146.32	146.29
	0.8	131.58	139.74	144.93	148.42	150.86	152.6	153.85
	0.9	132.89	141.59	147.33	151.37	154.36	156.63	158.42
	1.0	133.33	142.22	148.15	152.38	155.56	158.02	160
	1.1	132.89	141.59	147.33	151.37	154.36	156.63	158.42
	1.2	131.58	139.74	144.93	148.42	150.86	152.6	153.85
	1.3	129.45	136.75	141.09	143.76	145.38	146.32	146.29
	1.5	123.08	128	130.08	130.61	130.23	129.29	128
	1.6	119.05	122.61	123.46	122.89	121.53	119.72	117.65
	1.7	114.61	116.79	116.45	114.86	112.63	110.08	107.38
	Range of $\Lambda_{(-2)}$	[0.01, 1.99]	[0.03, 1.97]	[0.09, 1.91]	[0.12, 1.88]	[0.16, 1.84]	[0.2,1.8]	[0.23,1.77]

5. Improved classes of estimators for σ^{α} when point prior estimate σ_0 is available

Minimization of (2.3) with respect to scalar $\ddot{\psi}$ we get the optimum value of $\ddot{\psi}$ as

$$\ddot{\psi}_{opt} = \frac{\left\{1-\Lambda_{(\alpha)}\right\}^2}{\left\{1-\Lambda_{(\alpha)}\right\}^2 + v^{(n,\alpha)}} = \frac{\left(\sigma^{\alpha}-\sigma_0^{\alpha}\right)^2}{\left\{\left(\sigma^{\alpha}-\sigma_0^{\alpha}\right)^2 + \sigma^{2\alpha}\, v^{(n,\alpha)}\right\}} \qquad (5.1)$$

Now replacing σ^{α} and $\sigma^{2\alpha}$ by their unbiased estimator $\hat{T}_{(u,\alpha)}$ and $\hat{T}_{(u,2\alpha)}$ respectively in (5.1), we get a consistent estimate of $\ddot{\psi}_{opt}$ as

$$\hat{\ddot{\psi}}_{opt}^{(1)} = \frac{\left\{J_{(u,\alpha)}-\sigma_0^{\alpha}\right\}^2}{\left\{\left(J_{(u,\alpha)}-\sigma_0^{\alpha}\right)^2 + v^{(n,\alpha)}\, A_{(n,2\alpha)}\, s^{2\alpha}\right\}} \; . \qquad (5.2)$$

Replacing σ^{α} by its unbiased estimator $\hat{T}_{(u,\alpha)}$ in (5.1), we get

$$\hat{\ddot{\psi}}_{opt}^{(2)} = \frac{\left\{J_{(u,\alpha)}-\sigma_0^{\alpha}\right\}^2}{\left\{\left(J_{(u,\alpha)}-\sigma_0^{\alpha}\right)^2 + v^{(n,\alpha)}\left[A_{(n,\alpha)}\right]^2 s^{2\alpha}\right\}} \; , \qquad (5.3)$$

and replacing σ^{α} by its unbiased estimator $\hat{T}_{(u,\alpha)}$ and $\sigma^{2\alpha}$ by its MMSE estimator $T_{(MMSE,2\alpha)} = \left\{\frac{n-1}{2}\right\}^{\alpha}\left\{\frac{\Gamma\left(\frac{n+2\alpha-1}{2}\right)}{\Gamma\left(\frac{n+4\alpha-1}{2}\right)}\right\} s^{2\alpha}$ in (5.1), we get

$$\hat{\ddot{\psi}}_{opt}^{(3)} = \frac{\left\{\hat{T}_{(u,\alpha)}-\sigma_0^{\alpha}\right\}^2}{\left\{\left(\hat{T}_{(u,\alpha)}-\sigma_0^{\alpha}\right)^2 + v^{(n,\alpha)}\, W_{(n,2\alpha)}\, s^{2\alpha}\right\}} \qquad (5.4)$$

Many more consistent estimators of $\ddot{\psi}_{opt}$ can be obtained. However, a more flexible estimator of $\ddot{\psi}_{opt}$ can be obtained as :

$$\hat{\ddot{\psi}}_{opt}^{(4)} = \frac{\left\{J_{(u,\alpha)} - \sigma_0^{\alpha}\right\}^2}{\left\{\left(J_{(u,\alpha)} - \sigma_0^{\alpha}\right)^2 + L\, s^{2\alpha}\right\}} \quad , \tag{5.5}$$

where $L(\geq 0)$ is a suitably chosen constant.

Thus, the resulting modified shrinkage estimators of σ^{α} are given by

$$\hat{T}_{(s,\alpha)}^{(1)} = \sigma_0^{\alpha} + \frac{\left\{J_{(u,\alpha)} - \sigma_0^{\alpha}\right\}^3}{\left\{\left(J_{(u,\alpha)} - \sigma_0^{\alpha}\right)^2 + \nu^{(n,\alpha)}\, A_{(n,2\alpha)}\, s^{2\alpha}\right\}} \quad , \tag{5.6}$$

$$\hat{T}_{(s,\alpha)}^{(2)} = \sigma_0^{\alpha} + \frac{\left\{J_{(u,\alpha)} - \sigma_0^{\alpha}\right\}^3}{\left\{\left(J_{(u,\alpha)} - \sigma_0^{\alpha}\right)^2 + \nu^{(n,\alpha)}\left[A_{(n,\alpha)}\right]^2 s^{2\alpha}\right\}} , \tag{5.7}$$

$$\hat{T}_{(s,\alpha)}^{(3)} = \sigma_0^{\alpha} + \frac{\left(\hat{T}_{(u,\alpha)} - \sigma_0^{\alpha}\right)^3}{\left\{\left(J_{(u,\alpha)} - \sigma_0^{\alpha}\right)^2 + \nu^{(n,\alpha)}\, W_{(n,2\alpha)}\, s^{2\alpha}\right\}} \tag{5.8}$$

and $\hat{T}_{(s,\alpha)}^{(4)} = \sigma_0^{\alpha} + \frac{\left\{J_{(u,\alpha)} - \sigma_0^{\alpha}\right\}^3}{\left\{\left(J_{(u,\alpha)} - \sigma_0^{\alpha}\right)^2 + L\, s^{2\alpha}\right\}}$. (5.9)

The biases and MSEs of the estimators $\hat{T}_{(s,\alpha)}^{(i)}$, i = 1 to 4 are respectively given by

$$Bias\left\{\hat{T}_{(s,\alpha)}^{(i)}\right\} = \left\{\sigma_0^{\alpha} - \sigma^{\alpha}\right\} + \int_0^{\infty}\left[\frac{\left\{J_{(u,\alpha)} - \sigma_0^{\alpha}\right\}^3}{\left\{\left(J_{(u,\alpha)} - \sigma_0^{\alpha}\right)^2 + M\, s^{2\alpha}\right\}}\right] f(s^{\alpha})\, ds^{\alpha} \tag{5.10}$$

and $MSE\left\{\hat{T}_{(s,\alpha)}^{(i)}\right\} = \left\{\sigma_0^{\alpha} - \sigma^{\alpha}\right\}^2 +$

$$2\left\{\sigma_0^{\alpha} - \sigma^{\alpha}\right\}\int_0^{\infty}\left[\frac{\left\{J_{(u,\alpha)} - \sigma_0^{\alpha}\right\}^3}{\left\{\left(J_{(u,\alpha)} - \sigma_0^{\alpha}\right)^2 + M\, s^{2\alpha}\right\}}\right] f(s^{\alpha})\, ds^{\alpha}$$

$$+\int_0^{\infty}\left[\frac{\left\{J_{(u,\alpha)}-\sigma_0^{\alpha}\right\}^6}{\left\{\left(J_{(u,\alpha)}-\sigma_0^{\alpha}\right)^2+Ms^{2\alpha}\right\}^2}\right]f(s^{\alpha})\,ds^{\alpha}\quad , \tag{5.11}$$

where $M=\nu^{(n,\alpha)}A_{(n,2\alpha)}$, $\nu^{(n,\alpha)}\left[A_{(n,\alpha)}\right]^2$, $\nu^{(n,\alpha)}W_{(n,2\alpha)}$ and $L(\geq 0)$.

Now, transform the above expressions in the form of $\Lambda_{(\alpha)}$ and X for this we are using transformation $X=\frac{(n-1)s^2}{2\sigma^2}$, the ARB and RMSE be respectively given by

$$ARB\left\{\hat{T}^{(i)}_{(s,\alpha)}\right\}=$$

$$\left|\left\{\Lambda_{(\alpha)}-1\right\}+\frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\int_0^{\infty}\left\{\frac{\left\{\hat{P}^*(n,\alpha)X^{\frac{1}{2}}-\Lambda_{(\alpha)}\right\}^3}{\left(\hat{P}^*(n,\alpha)X^{\frac{1}{2}}-\Lambda_{(\alpha)}\right)^2+\left(\frac{2}{n-1}\right)^{\alpha}M\,X}\right\}e^{-X}\,X^{\left(\frac{n-3}{2}\right)}dX\right| \tag{5.12}$$

and

$$RMSE\left\{\hat{T}^{(i)}_{(s,\alpha)}\right\}=$$

$$\left[\left\{\Lambda_{(\alpha)}-1\right\}^2+\frac{2\left\{\Lambda_{(\alpha)}-1\right\}}{\Gamma\left(\frac{n-1}{2}\right)}\int_0^{\infty}\left\{\frac{\left\{\hat{P}^*(n,\alpha)X^{\frac{1}{2}}-\Lambda_{(\alpha)}\right\}^3}{\left(\hat{P}^*(n,\alpha)X^{\frac{1}{2}}-\Lambda_{(\alpha)}\right)^2+\left(\frac{2}{n-1}\right)^{\alpha}M\,X}\right\}e^{-X}X^{\left(\frac{n-3}{2}\right)}dX\right.$$

$$\left.+\frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\int_0^{\infty}\left\{\frac{\left\{\hat{P}^*(n,\alpha)X^{\frac{1}{2}}-\Lambda_{(\alpha)}\right\}^6}{\left\{\left(\hat{P}^*(n,\alpha)y^{\frac{1}{2}}-\Lambda_{(\alpha)}\right)^2+\left(\frac{2}{n-1}\right)^{\alpha}M\,X\right\}^2}\right\}e^{-X}\,X^{\left(\frac{n-3}{2}\right)}dX\right] \tag{5.13}$$

where $\hat{P}^*(n,\alpha)=\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n+\alpha-1}{2}\right)}$.

5.1.Numerical Illustrations With Application to Measure Inverse of Variance

The percentage relative efficiencies (PREs) of modified shrinkage estimators $\hat{T}^{(i)}_{(s,\alpha)}$ (i = 1, 2, 3) with respect to the unbiased estimator $J_{(u,\alpha)}$ and MMSE estimator $J_{(MMSE,\alpha)}$ are respectively defined by

$$PRE\left\{\hat{T}^{(i)}_{(s,\alpha)}, J_{(u,\alpha)}\right\} = \frac{RV\left\{J_{(u,\alpha)}\right\}}{RMSE\left\{\hat{T}^{(i)}_{(s,\alpha)}\right\}} *100 \qquad (5.14)$$

and

$$PRE\left\{\hat{T}^{(i)}_{(s,\alpha)}, J_{(u,\alpha)}\right\} = \frac{RMSE\left\{J_{(MMSE,\alpha)}\right\}}{RMSE\left\{\hat{T}^{(i)}_{(s,\alpha)}\right\}} *100 \qquad (5.15)$$

for different values of $\Lambda_{(\alpha)}$ n and α ,

Where $RV\left\{J_{(u,\alpha)}\right\}$, $RMSE\left\{J_{(MMSE,\alpha)}\right\}$ and $RMSE\left\{\hat{T}^{(i)}_{(s,\alpha)}\right\}$ are respectively given by (1.6), (1.12), and (5.13).

Since we have considered the problem of estimating $\frac{1}{\sigma^2}$ (i.e. for α = -2) as a particular case to study the behavior of the developed modified class of estimators so the PRE of $\hat{T}^{(i)}_{(s,-2)}$ w.r.t. unbiased estimator $\hat{T}_{(u,-2)}$ and MMSE estimator $\hat{T}_{(MMSE,-2)}$ for different values of i = 1, 2, 3 and $M = A_{(n,-4)}$, and $W_{(n,-4)}$ are given in the Tables 5.3, 5.4 and 5.5 respectively. It is observed that the proposed modified estimators $\hat{T}^{(i)}_{(s,-2)}$(i=1,2) are better than $\hat{T}_{(u,-2)}$ and $\hat{T}_{(MMSE,-2)}$ when $\Lambda_{(-2)} \in [0.7, 1.3]$ and $6 \le n \le 19$, where as estimator $\hat{T}^{(3)}_{(s,-2)}$ is better than $\hat{T}_{(u,-2)}$ and $\hat{T}_{(MMSE,-2)}$ when $\Lambda_{(-2)} \in [0.7, 1.3]$ and $10 \le n \le 19$.

For fixed $\Lambda_{(-2)}$, the $PRE\left\{\hat{T}^{(1)}_{(s,-2)}, \hat{T}_{(u,-2)}\right\}$ decreases as n increases and for fixed n these values decrease as $\lambda_{(-2)}$ goes away from unity where PRE's attain its maximum. Similar trend found for $\hat{T}^{(2)}_{(s,-2)}$ and $\hat{T}^{(3)}_{(s,-2)}$. We also observe that the gain in efficiency with respect to $\hat{T}_{(MMSE,-2)}$ is lesser than $\hat{T}_{(u,-2)}$ (see, Table 5.1, 5.2 and 5.3). Tables 5.1,5.2 and 5.3 show that the estimator $\hat{T}^{(2)}_{(s,-2)}$ is the best

(in the sense of having smallest MSE) among $\hat{T}^{(1)}_{(s,-2)}$, $\hat{T}^{(2)}_{(s,-2)}$ and $\hat{T}^{(3)}_{(s,-2)}$ followed by $\hat{T}^{(3)}_{(s,-2)}$. Thus the modified estimator $\hat{T}^{(2)}_{(s,-2)}$ is to be preferred in practice when $\Lambda_{(-2)}$ moves in the vicinity of unity and larger efficiency are observed when sample size n is small.

Table 5.1. Showing the $PRE\left\{\hat{T}^{(1)}_{(s,-2)}, \hat{T}_{(u,-2)}\right\}$ and $PRE\left\{\hat{T}^{(1)}_{(s,-2)}, \hat{T}_{(MMSE,-2)}\right\}$ for different **n=11(2)23 and** $\Lambda_{(-2)}$ **=0.3(0.1)1.7**

$\Lambda_{(-2)} \downarrow$ n→	11	13	15	17	19	21	23
0.3	127.03	117.82	112.43	108.98	106.64	105	103.81
	95.27	**94.26**	**93.69**	**93.41**	**93.31**	**93.33**	**93.43**
0.4	133.25	121.05	113.62	109.14	105.76	103.11	101.15
	99.94	**96.84**	**94.68**	**93.55**	**92.54**	**91.65**	**91.04**
0.5	145.87	133.08	121.47	114.3	111.13	107.4	102.76
	109.41	**106.46**	**101.22**	**97.97**	**97.24**	**95.47**	**92.49**
0.6	166.86	151.26	142.82	128.03	121.2	121.9	118.8
	125.14	**121.01**	**119.02**	**109.74**	**106.05**	**108.36**	**106.92**
0.7	198.81	172.62	171.6	161.75	140.79	136.95	146.88
	149.1	**138.1**	**143**	**138.64**	**123.19**	**121.73**	**132.19**
0.8	229.29	202.39	194.16	205.83	184.37	159.91	166.05
	171.96	**161.91**	**161.8**	**176.42**	**161.32**	**142.14**	**149.44**
0.9	245.13	231.37	212.45	230.84	232.81	199.11	185.42
	183.85	**185.1**	**177.04**	**197.86**	**203.71**	**176.98**	**166.88**
1.0	247.94	239.64	220.57	234.97	248.67	219.27	198.77
	185.96	**191.71**	**183.81**	**201.4**	**217.58**	**194.91**	**178.89**
1.1	240.65	231.98	205.52	220.01	240.58	209.44	178.57
	180.49	**185.58**	**171.26**	**188.58**	**210.51**	**186.17**	**160.72**
1.2	223.16	220.75	179.72	185.91	216.19	196.76	153.08
	167.37	**176.6**	**149.76**	**159.35**	**189.16**	**174.9**	**137.77**
1.3	197.35	207.63	160.77	148.32	178.57	182.09	140.91
	148.01	**166.11**	**133.98**	**127.13**	**156.25**	**161.85**	**126.82**
1.4	169.34	190.69	150.41	122.14	139.18	160.24	135.14
	127.01	**152.55**	**125.34**	**104.69**	**121.79**	**142.43**	**121.63**
1.5	145.34	169.62	144.44	108.12	109.61	133.07	128.44
	109	**135.7**	**120.37**	**92.68**	**95.91**	**118.28**	**115.6**
1.6	127.9	146.67	139.17	101.84	91.95	107.77	117.59
	95.92	**117.33**	**115.97**	**87.29**	**80.46**	**95.8**	**105.83**
1.7	116.56	125.16	132.41	99.56	82.92	89.43	103.62
	87.42	**100.13**	**110.34**	**85.34**	**72.55**	**79.5**	**93.26**
Range of $\Lambda_{(-2)}$	[0, 2]	[0,1.86]	[0, 2]	[0,1.62]	[0,1.55]	[0.25,1.6]	[0.28,1.7]
	[0.37,1.61]	[0.42,1.71]	[0.48,1.81]	[0.52,1.45]	[0.53,1.49]	[0.57,1.5]	[0.58,1.6]

Bold *figures denote the PRE with respect to MMSE estimator*

Table 5.2. Showing the $PRE\left\{\hat{T}^{(2)}_{(s,-2)}, \hat{T}_{(u,-2)}\right\}$ and $PRE\left\{\hat{T}^{(2)}_{(s,-2)}, \hat{T}_{(MMSE,-2)}\right\}$ for different n=11(2)23 and $\Lambda_{(-2)}$ =0.3(0.1)1.7

$\Lambda_{(-2)}\downarrow$ n$\rightarrow$	11	13	15	17	19	21	23
0.3	129.72	118.38	112.20	108.43	105.97	104.30	103.12
	97.29	**94.71**	**93.50**	**92.94**	**92.73**	**92.71**	**92.80**
0.4	138.51	123.14	114.25	109.06	105.41	102.61	100.54
	103.88	**98.51**	**95.21**	**93.48**	**92.23**	**91.21**	**90.48**
0.5	155.82	137.43	124.01	115.44	111.42	107.53	102.78
	116.87	**109.94**	**103.34**	**98.95**	**97.49**	**95.58**	**92.51**
0.6	184.17	159.93	147.82	131.75	123.22	122.78	119.66
	138.13	**127.94**	**123.18**	**112.93**	**107.81**	**109.14**	**107.70**
0.7	225.50	189.01	181.04	168.97	146.35	140.38	149.09
	169.12	**151.21**	**150.86**	**144.83**	**128.05**	**124.79**	**134.18**
0.8	267.10	228.29	211.27	218.48	195.06	167.81	171.98
	200.33	**182.63**	**176.06**	**187.27**	**170.67**	**149.16**	**154.79**
0.9	291.97	264.98	237.23	250.04	249.30	212.25	195.90
	218.98	**211.98**	**197.69**	**214.32**	**218.14**	**188.66**	**176.31**
1.0	298.07	273.74	248.01	257.48	266.85	233.97	211.18
	223.55	**218.99**	**206.68**	**220.70**	**233.50**	**207.97**	**190.06**
1.1	287.98	259.85	228.12	240.20	256.21	220.06	187.65
	215.98	**207.88**	**190.10**	**205.88**	**224.19**	**195.61**	**168.88**
1.2	262.71	241.23	194.50	199.96	227.64	202.87	157.58
	197.03	**192.98**	**162.08**	**171.39**	**199.19**	**180.32**	**141.82**
1.3	227.18	222.17	168.91	156.71	185.84	185.23	142.24
	170.39	**177.73**	**140.76**	**134.32**	**162.61**	**164.65**	**128.01**
1.4	190.43	200.83	153.91	126.50	143.57	161.76	134.67
	142.82	**160.67**	**128.26**	**108.43**	**125.62**	**143.79**	**121.20**
1.5	159.60	176.69	144.98	109.66	112.11	133.97	127.16
	119.70	**141.35**	**120.82**	**94.00**	**98.10**	**119.08**	**114.45**
1.6	136.93	151.75	137.99	101.41	92.98	108.46	116.24
	102.70	**121.40**	**114.99**	**86.92**	**81.36**	**96.41**	**104.62**
1.7	121.54	128.94	130.45	97.77	82.74	89.79	102.63
	91.16	**103.15**	**108.71**	**83.80**	**72.40**	**79.81**	**92.37**
Range of $\Lambda_{(-2)}$	[0, 2]	[0, 2.0]	[0, 2]	[0,1.62]	[0,1.55]	[0.25,1.6]	[0.28,1.7]
	[0.37,1.61]	**[0.42,1.71]**	**[0.48,1.81]**	**[0.52,1.45]**	**[0.53,1.49]**	**[0.57,1.5]**	**[0.58,1.6]**

Bold *figures denote the PRE with respect to MMSE estimator.*

Table 5.3. Showing the $PRE\left\{\hat{T}^{(3)}_{(s,-2)},\hat{T}_{(u,-2)}\right\}$ and $PRE\left\{\hat{T}^{(3)}_{(s,-2)},\hat{T}_{(MMSE,-2)}\right\}$ for different n=11(2)19 and $\Lambda_{(-2)}$ =0.3(0.1)1.7

$\Lambda_{(-2)}\downarrow$ n→	11	13	15	17	19	21	23
0.3	106.59	108.52	108.30	107.46	106.52	105.64	104.87
	79.94	**86.82**	**90.25**	**92.11**	**93.21**	**93.90**	**94.38**
0.4	107.47	107.93	107.83	106.71	105.00	103.52	102.37
	80.60	**86.35**	**89.86**	**91.47**	**91.87**	**92.02**	**92.14**
0.5	108.67	113.06	108.98	108.46	108.50	105.37	101.78
	81.51	**90.45**	**90.82**	**92.97**	**94.93**	**93.66**	**91.60**
0.6	109.41	119.62	121.31	112.58	112.97	117.43	113.99
	103.16	**102.11**	**101.09**	**101.01**	**100.00**	**104.38**	**102.59**
0.7	119.21	120.99	137.17	133.36	120.20	123.98	137.20
	117.23	**115.26**	**114.31**	**114.31**	**105.17**	**110.21**	**123.48**
0.8	127.45	127.85	140.62	160.60	146.60	132.79	144.58
	119.23	**116.67**	**117.19**	**137.65**	**128.27**	**118.03**	**130.12**
0.9	128.47	139.86	142.38	168.78	177.21	156.01	150.99
	122.33	**111.89**	**118.65**	**144.67**	**155.06**	**138.68**	**135.89**
1.0	128.23	145.40	144.84	166.77	187.19	170.87	158.73
	125.36	**116.32**	**120.70**	**142.95**	**163.79**	**151.89**	**142.86**
1.1	154.33	147.53	140.02	158.88	185.22	171.38	147.90
	130.23	**118.02**	**116.68**	**136.18**	**162.07**	**152.34**	**133.11**
1.2	152.59	148.65	133.03	141.27	173.76	170.59	136.64
	122.89	**118.92**	**110.86**	**121.09**	**152.04**	**151.64**	**122.97**
1.3	151.86	148.22	131.36	119.91	150.99	165.21	134.54
	121.93	**118.57**	**109.47**	**102.78**	**132.12**	**146.85**	**121.09**
1.4	150.25	144.82	132.82	107.15	122.18	150.61	134.54
	113.44	**115.86**	**110.68**	**103.84**	**106.91**	**133.88**	**121.09**
1.5	140.56	136.63	134.50	103.20	100.95	127.41	130.96
	111.85	**109.31**	**112.09**	**101.46**	**100.45**	**113.26**	**117.86**
1.6	128.34	123.15	134.64	103.69	100.63	103.82	121.03
	75.56	**98.52**	**99.34**	**88.88**	**77.55**	**92.28**	**90.12**
1.7	108.11	107.68	105.20	105.70	84.78	87.57	85.23
	76.20	**86.15**	**88.12**	**90.60**	**74.18**	**77.84**	**95.55**
Range of $\Lambda_{(-2)}$	[0.2, 1.89]	[0.2,1.85]	[0.2, 1.77]	[0.2,1.71]	[0.2,1.6]	[0.25,1.6]	[0.28,1.6]
	[0.58,1.57]	**[0.58,1.57]**	**[0.59,1.7]**	**[0.59,1.6]**	**[0.6,1.55]**	**[0.6,1.55]**	**[0.6,1.52]**

Bold *figures denote the PRE with respect to MMSE estimator.*

6. Simulation study

For simulation study a random sample of 13 observations is generated from a normal population with mean $\mu = 20$ and standard deviation $\sigma = 5$ (see, Table 6.1), if the prior estimate of standard deviation σ is available from the past behavior of the population as $\sigma_0 = 4$, i.e. $\Lambda_{(-2)}$ =1.5625.

Table 6.1. Ordered simulated data from N(20,25)

11.82	15.97	15.98	17.50	17.97	18.09	19.76
20.36	21.69	22.90	24.34	26.77	28.16	-

The inverse of variance is estimated using their usual estimators and developed modified estimators $\hat{T}^{(1)}_{(s,-2)}$, $\hat{T}^{(2)}_{(s,-2)}$ and $\hat{T}^{(3)}_{(s,-2)}$, and the gain in efficiency with respect to conventional unbiased estimator, MMSE estimator and themselves (i.e. $\hat{T}^{(1)}_{(s,-2)}$, $\hat{T}^{(2)}_{(s,-2)}$ and $\hat{T}^{(3)}_{(s,-2)}$) are given in Table 6.2

Table 6.2. Presents the simulation results

Parametric Function	True Value	Estimator used	Estimate	RV or RMSE	PRE of $\hat{T}^{(1)}_{(s,-2)}$	PRE of $\hat{T}^{(2)}_{(s,-2)}$	PRE of $\hat{T}^{(3)}_{(s,-2)}$
Inverse of Variance $\tau(1,-2) = \left(\frac{1}{\sigma^2}\right)$	0.04	Unbiased	0.0391	0.25	155.28	160.98	128.73
		MMSE	0.03126	0.20	124.22	128.80	102.97
		$\hat{T}^{(1)}_{(s,-2)}$	0.04745	0.1610	100	103.67	82.90
		$\hat{T}^{(2)}_{(s,-2)}$	0.04869	0.1553	96.46	100	79.97
		$\hat{T}^{(3)}_{(s,-2)}$	0.04243	0.1942	120.62	125.05	100

REFERENCES

MEHTA, J. S. and SRINIVASAN, R. (1971). "Estimation of the mean by shrinkage to a point", Journal of the American Statistical Association, 66, 86—90.

MISHRA, A. (1985). "A note on estimation of amount of information in normal samples", Journal of Indian Society of Agricultural Statistics, 37, 3, 226—230.

PANDEY, B. N. (1979). "On shrinkage estimation of normal population variance", Communications in Statistics – Theory and Methods, 8, 359—365.

PANDEY, B. N. and SINGH. J. (1977). "Estimation of the variance of normal population using prior information", Journal of the Indian Statistical Association, 15, 141—150.

SEXENA, S. and SINGH, H. P. (2004). "Estimating various measures in normal population through a single class of estimators". Journal of the Korean Statistical Society, 33:3, pp 323—337.

SINGH, H. P. and SAXENA, S. (2003). "An improved class of shrinkage estimators for the variance of a normal population", Statistics in Transition, 6, 119—129.

SINGH, H. P. and SINGH, R. (1997). "A class of shrinkage estimators for the variance of a normal population", Microelectronics and Reliability, 37, 863—867.

SINGH, H. P., SHUKLA, S. K. and KATYAR, N. P. (1999)."Estimation of standard deviation in normal distribution with prior information", Proceedings of the National Academic Sciences India", 69, 183—189.

THOMPSON, J. R. (1968). "Some shrinkage techniques for estimating the mean:, Journal of the American Statistical Association, 63, 113—122.

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 159—162

CORRELATION AND REGRESSION: SIMILAR OR DIFFERENT CONCEPTS?

Marcin Kozak[1]

ABSTRACT

Correlation and regression are different statistical methodologies. In spite of this, quite often they are taught, or explained, simultaneously, which may be a cause of students' confusion. Statistical packages also have their contribution to this problem by presenting a regression output supported by correlation. I show that linking correlation and regression should always be done with great caution.

Key words: teaching; correlation; regression.

1. Introduction

When teaching correlation and regression, many teachers link these two methodologies. The former seems to be an ideal introduction to the latter. The same can be seen in statistics books: correlation and regression are usually presented in the same section—sometimes they have their own subsections, sometimes they do not. After a course a student usually thinks of correlation in terms of regression: both of them describe relations between two variables. In this paper I show it is not appropriate and may do more harm than good.

2. Correlation and regression

Detailed description of correlation and regression is beyond the scope of this paper. However, main concepts of these methods that I use later to discuss the main points of this papers are worth pointing out. For the sake of simplicity let us assume that correlation and regression are to be taught in relation to normal variable/variables, which is common for young non-statistical students. Thus we discuss Pearson's correlation between two normal variables, X_1 and X_2, and linear simple regression between two variables, Y and X, Y being a normal variable and

[1] Department of Biometry, Warsaw Agricultural University, e-mail: m.kozak@omega.sggw.waw.pl

X not yet specified. I will take no notice of multiple regression, although what will be said about simple regression also relates to the multiple one.

Correlation aims to study a linear relationship between two normal variables, X_1 and X_2. It does not count whether the variables are in a cause-effect relation or are co-related. Simply, based on correlation, one may study whether the relationship exists, and if it does, one may interpret it in terms of its direction (positive or negative) and magnitude (strong, weak, etc.). Both variables are assumed to be random.

Regression, in its classical form, aims to study a relationship between two variables, *Y* and *X*. We may distinguish two main problems regression can be applied to, namely, prediction and exploration. The former aims to predict *Y* based on *X* values. For this problem we do not have to assume that *Y* is a result of *X* and *X* is a cause of *Y*—in fact, the underlying process may be opposite or the variables may be correlated in nature; the only thing we aim here to obtain is a strong relationship between *Y* and *X*. The latter problem, exploration, aims to study a cause-effect relationship between *Y* and *X*, this time the former being the dependent variable whereas the latter, the cause (usually called independent) variable.

In case of prediction the aim is to determine a regression line of high quality. In second case, that of exploration, which is the most important for, e.g., biologists (Quinn and Keough, 2002), we would like to learn whether the relationship indeed exists and what its nature is. And here all the statistical essence comes out. The regression assumptions should be met in order to make the estimation and testing correct, and to make the interpretation unbiased and reflect the true relationship one wants to study.

3. Correlation versus regression

There is no need to remind here all the regression assumptions. However, there is one assumption that has to be mentioned to help us come back to the main topic of this paper. In a classical linear regression analysis a dependent variable is assumed to be normally distributed, and the observed values of an independent variable, to be a set of known constants (e.g., Sokal and Rohlf 1995, Rawlings et al. 1998, Quinn and Keough 2002). The latter assumption simply means that *X* is to be a deterministic, not random, variable. If so, how would we want to link correlation, in which both variables are random and are assumed to follow bivariate normal distribution, with regression? No, we should not do that: correlation and regression should be thought of as two distinct methods that may be applied in specific situations. Therefore, even mentioning a correlation coefficient when there is only one random variable (which is “natural” because standard statistical packages do provide this information in their basic regression output) may be seen inappropriate.

The main point of this paper is that students should always be kept aware of this situation. They should not be taught that correlation and regression are strictly related methods. They should not be taught how to use a correlation coefficient in regression. They should be taught about correlation *and* about regression, and these two philosophies should be taught separately.

There is a very clear way out of this situation, shown quite a long time ago. There is Model I regression, also called fixed effects model, and there is Model II regression, also called random effects model (Sokal and Rohlf 1995, Quinn and Keough 2002). The former deals with what I called the classical regression, that is, regression with a deterministic independent variable, whereas the latter, with a bi-normal distribution of dependent and independent variables.

In Model I regression, the ordinary least squares (OLS) estimation is unbiased, but in Model II regression, it is not. Sokal and Rohlf (1995) and Quinn and Keough (2002) suggest that for prediction, OLS are appropriate for Model II regression; however, where the regression is to explore the true relationship between the two variables, OLS estimation might be inappropriate so should be applied with great caution. In such case the maximum likelihood (ML) estimation should be applied to model the relationship (e.g., Anderson 1958, Kendall and Stuart 1963, Rao 1973). Note that in case of a bivariate normal distribution both ML and OLS provide the same results although the former is and the latter is not correct. There are also other methods to handle estimation in such a situation, such as major axis regression, reduced major axis regression, and a slope-range method. Discussion about these methods is beyond the scope of this paper; it can be found, e.g., in Quick and Keough (2002) and the citations therein.

4. Conclusion

Correlation and regression are two distinct statistical philosophies that sometimes are intermingled. But for students at the beginning stages of statistical education, they should be taught separately. If the use of correlation is to be incorporated into a course, it should be done with great caution to keep the students aware that correlation is not just a part of regression.

Of help should be Model I and II regressions, which are different from philosophical, methodological, and application points of view. They may provide similar results, but this will not be the case in many applications. Therefore, it is of importance to distinguish the two methodologies in statistical teaching, consultancy, and statistical practice. Mixing them up may cause results of an analysis to be false and the interpretation of the results to be incorrect. A correlation coefficient in its classical, Pearson's form, on the other hand, is concerned with a bi-normal distribution of the variables so should be associated with Model II regression. In Model I regression a correlation coefficient measures the quality of fit for Y regressed by X, but it should not be thought of as

correlation in its natural way of thinking, in which it is a measure of association between two random variables.

Despite their usefulness, the terms Model I and Model II regression are unfortunately not commonly known among statisticians. Therefore, they should be brought to the statistical audience attention and incorporated into statistical courses because of their importance, too much ignored until now.

Acknowledgments

I wish to thank Dr. Stan Lipovestky from GFK Custom Research North America for his valuable comments and discussions on the topic of this paper.

REFERENCES

ANDERSON, T. W. (1958) An Introduction to Multivariate Statistical Analysis. Wiley and Sons, New York.

KENDALL, M. G. and STUART, A. (1966). *The Advanced Theory of Statistics*. Griffin, London.

QUINN, G. P. and KEOUGH, M. J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.

RAO, C. R. (1973) *Linear Statistical Inference and its Applications*. Wiley and Sons, New York.

RAWLINGS, J. O., PANTULA, S. G. and DICKEY, D. A. (1998). *Applied regression analysis: A research tool*. Springer Verlag.

SOKAL, R. R. and ROHLF, F. J. (1995). *Biometry*. 3rd edition. W.H. Freeman, New York.

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 163—164

FROM THE EDITORIAL OFFICE

We are pleased to announce the launch of the journal's new section, *Current Issues in Public Statistic.* It is envisaged as a forum for discussing the issues that emerge as challenging to official statistics in a country or region, and we hope to inspire all parties concerned about its status, data producers and users, to share with us some evidence on problem-solving approaches, or 'best practices', relating to any facets of public statistical system.

Since one of the fundamental principles of the governmental statistics in Poland — unconditional protection of individual data files against disclosure for non-statistical purposes — has recently been challenged by a public authority's attempt to use them for administrative actions, we decided to address it in this issue. Below, we present a summary of the letter we received from the Polish Statistical Association. Written in the face of repeated accidents of breaching the principle of statistical confidentiality, the letter calls for attention to the problem of the integrity and credibility of public statistics.

As some misinterpretations or even 'wrong practice' on behalf of the administrative policies toward accessing statistical data present still a problem in some of the formerly socialists economies, we wish to open a discussion about the dual issue of confidentiality and data access, to which we plan to devote a special issue soon (see our 'call for papers' in this issue).

Letter from the Polish Statistical Association

The National Council of the Polish Statistical Association expresses its serious concern about the irreversible consequences that the legal-type actions which have recently been undertaken by some of the regional public prosecutor's offices against representatives of statistical offices — for their statutory efforts to protect individual data against disclosure for non-statistical purposes — will have for the public statistics. In times of weakening the willingness of data providers to deliver adequate data on regular basis, such actions would inevitably damage the integrity of official statistics and diminish the overall quality of the produced information.

More importantly, the trust-based willingness to provide data is vital for the functioning of the system of public statistics. Therefore, such a kind of actions can destroy the fundamentals on which the whole system is built on, including image of statistical institution as an independent and leading unit of the state informational system (blamed for its apparent inability to assure confidentiality and anonymity of data subjects).

This letter calls all parties interested in obtaining, and valuing as a public good, the well-timed and reliable statistical information for supporting the Polish Statistical Association and the Central Statistical Office's joint efforts to protect the principles of confidentiality and privacy, along with rules of accessibility of the collected data for strictly analytical and statistical purposes.

Statistical confidentiality is the legal, ethical and methodological duty of the official statistics, and the professional virtue of its services, in Poland and all over the world. As established by the Law on official statistics, dated 29th June 1995, it guarantees that the collected and gathered data are subject to particular protection and no disclosure of any of them can be made to anybody for purposes other than statistics. This law standard is obligatory in such provisions of law like UNO or OECD systems, has special position in European Code of Statistical Practices and is a standard base of official statistics accepted all over the world.

The evidence from Poland and all over the world clearly proves that any violations of the principle of statistical confidentiality would cause long-lasting damages in terms of image of, and public support to public statistics. Without adequate and reliable statistical information, there are no means for assessing different aspects of reality, making evidence-based decision or forecasting and building information society.

On the other hand, we fully share the need to increase access to non-identifiable statistical data for statistical and analytical purposes. Accordingly, we follow the recommendations of, and actively collaborate with the leading international organizations and agendas, such as the EUROSTAT, European Commission and the Council of Europe, OECD, UN, IMF, World Bank, etc., that explicitly address the complexity of the dual problem of confidentiality and data access. We do our best to keep our practice in these areas to the standards adopted by those organizations and their member states.

Calling for public attention to this subject, the members of the Polish Statistical Association (that is affiliated to the International Statistical Institute gathering numerous representatives of official statistics and scientific circles, and remaining one of the biggest international organisation of professional statisticians) hope for involving the international community of statisticians (both producers and users of official statistics) in discussion on this crucially important issues, as well as for inspiring a systematic reflection and, eventually, a recommendation or common action towards strengthening the foundations of the country's system of public statistics.

President of the Polish Statistical Association
Dr. Kazimierz Kruszka

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 165—173

STATISTICAL DISCLOSURE LIMITATIONS ISSUES IN RESEARCH AND PRACTICE
LESSONS FOR A COUNTRY'S DATA SYSTEM FROM INTERNATIONAL STUDIES

Włodzimierz Okrasa[1]

1. Introduction

The fact that public statistical systems in some formerly socialist countries, including Poland, suffer from deficiencies of the kinds that are practically absent in well established democracies — such as lack of an effective mechanism of protection against administrative actions that can put in jeopardy the integrity of official statistics — prompts some reflections on their origin and persistence. Why has that happened? Especially, given that majority of those countries follow the legal principles governing statistical activities along the international standards since the early stage of transformation process (e.g., in Poland in 1994 and 1995). And they have placed particularly strong emphasis on ensuring unconditional confidentiality of individual data files — containing identifiable information about persons, households, farms, businesses, and governmental bodies — against disclosure for non-statistical purposes.

One may hypothesize that the gap that still exists between the *de jure* principles and *de facto* practices in this respect has in a way been pre-designed. Namely, that it is a remnant of some public servants' attitudes rooted in the pre-transition era's mentality, when the concept of confidentiality of statistical data has been linked with the policy of secrecy of official (state's) informational resources, rather than with a concern about individual entities' informational privacy.

Another line of reasoning that seems, however, to be totally compatible with the above one, leads to research infrastructure of public statistical system — very weak or not existing compared to that in advanced OECD countries (e.g., to a long tradition of such research in USA — Okrasa, 1991). As noted by the authors of the report of the survey devoted to the state of disclosure procedures in

[1] University of Cardinal Stefan Wyszynski, Warsaw and Central Statistical Office of Poland, Warsaw

formerly socialist economies (Felsö et al., 2001), despite addressing this problem and adopting the legally defined confidentiality principles, not much progress has been done in most of them in practice due to difficulties with their organizational and technical implementation.

The insufficiency of efforts to ensure statistical data confidentiality is typically paralleled by a lack of research-informed public debate on vital importance that the appropriate rules of confidentiality of responses *at work* have for the statistical agency's credibility to its providers. And, subsequently, for its relationship of mutual respect and trust with respondents, which is essential for public statistics at large, and a prerequisite of fulfillment of its mission to produce and disseminate high quality data.

The problem of designing effective modes of statistical disclosure limitations procedures turns out to be pretty complicated, however, by the trade-off that exist between confidentiality and accessibility of data. Since data and statistical information are being compiled in order to be widely disseminated to others — users from either outside of the public statistical system (people, business, government, etc.) or within it (e.g., inter-agency data sharing or matching data from different sources, etc) — risks of disclosure are increased either when data are tabulated for small groups or when detailed public-use microdata files are released.

In general, there are two main options for protecting the confidentiality of the data being released by a statistical agency in tabular or microdata files: (i) one is to restrict the data through the use of statistical disclosure limitation procedures; (ii) another is to restrict access to data through imposing conditions on who may have access, for what purpose, at what locations, and so forth (Jabin 1993). As the latter seems to be predominantly of a policy rather than a research matter, and was the type of solution used on discretionary basis in the pre-transition era in Central and Eastern Europe, the former only remains an object of interest in the context discussed here.

2. Confidentiality and Data Access

Much of the work for the advancement of methodology and implementation of procedures dealing with the twofold issue of confidentiality and data access was done during the previous decade (90s), both in America and Europe. An institutional expression of that interest in the form of creation (in the mid of 90s) of a special interest subcommittee of the US Federal Committee on Statistical Methodology (FCSM), the Confidentiality and Data Access Committee (CDAC), proves the high importance being attached in official statistics to the issues relating to protecting data confidentiality while providing selective and controlled access to confidential data.

The major recommendations of CDAC (*Report on Statistical Disclosure Limitation Methodology* 2005), made large use of the results of one of the most

systematic studies ever conducted on those issues, namely, the study performed by the Panel on Confidentiality and Data Access organized by the National Research Council's Committee on National Statistics (CNSTAT) and the Social Science Research Council (SSRC), (Duncan et al., 1993). The Panel's comprehensive approach to the data confidentiality-accessibility dilemma did set a kind of a paradigm helpful for further explorations of this complex issue, including clarifying nomenclature[1] and sorting priorities for research and practical applications.

The question of the relevance that the Panel's report might have for undergoing transition statistical systems in Central and Eastern Europe — discussed already at the time of its release (Okrasa, 1994) — seems still to be valid. Therefore, the remarks below draw extensively on this work.

They need to be preceded, however, by mentioning the work being conducted in this area by Eurostat, especially on preparation of a guide on disclosure limitations (Eurostat 1996). As a follow-up, the Eurostat also supported a project on implementation of alternative procedures listed in its *Manual* (in selected countries) along with discussion of suitability of procedural solutions for particular types of data and data protection requirements (Holvast, 1999). In most of EU countries, a policy of monitoring changes in public statistics in the form of "Peer Review on the Implementation of the European Statistics Code of Practice," on country-by-country basis, has already resulted in greater attention to, and better explication of the issue of statistical confidentiality.

It should also be noted that a special survey was conducted in 1998 by the United Nations Economic Commission for Europe (UNECE) on the transition economies, embracing 14 Central and Eastern European countries and 5 countries from the Commonwealth of Independent States (CIS), (Luige and Meliskova 1999). Authors admit that despite the substantial progress that took place over the 90s, both methodological and application aspects of the issues need further development. According to Felsö et al, (2001), in face of challenges posed by increasing demand for microdata and growing software capabilities, a move is needed toward implementing more sophisticated disclosure (probability model based) techniques. By the way, Poland, with two simple disclosure limitations techniques employed at that time in the case of releasing demographic microdata (*geographic or population thresholds* and *re-coding variables***) ,** extended by additional one for the case of business data (*microaggregation*), was placed at the bottom half among the ten *countries in accession* included into the above mentioned study.

---

[1] For instance, the term *disclosure* itself is understood, after the Panel's definition, as related to "inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a data subject is identified from a released file ***(****identity disclosure****),*** sensitive information about a data subject is revealed through the released file (*attribute disclosure),* or the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible ***(****inferential disclosure).*" Duncan et al., p. 23—24.

There are several good reasons for recalling the basic recommendations of, and clarifications made by the Panel on Confidentiality and Data Access. One is its explicit view of the statistical disclosure limitation issue as two sides of the same coin, through addressing concurrently confidentiality and data access. Another is their relevance for methodological and practical considerations in countries characterized by different culture of the relationships between data producers (statistical agencies) and users, including public administration.

3. The relevance of the CNSTAT/SSRC *Panel on Confidentiality and Data Access' Report* for redesigning information systems in Central and Eastern Europe [1]

Despite that the major goal of the Panel was to develop recommendations intended to aid federal statistical agencies in their stewardship of data for policy decisions and research, the fundamental importance of its primary concern is common to all statistical agencies — namely: (1) protecting the interests of data providers through procedures that insure privacy and confidentiality; (2) enhancing public confidence in the integrity of the data; and (3) facilitating the responsible dissemination of data to users.

The aim of the remaining part of this article is twofold. The first is to present some of the major issues involved in the trade-off between confidentiality and data access, as well as the panel's major recommendations for dealing with these issues. The second is to consider the validity and relevance of such an international study for the different organizational contexts of the highly centralized statistical system of the formerly socialist economies of Central and Eastern Europe.

Special emphasis needs to be placed on how suitable the panel's recommendations are for work in countries that seek to redesign their frameworks within which to treat controversies surrounding the relationships between data providers, data producers, and data users. Confidentiality and accessibility of data — the matters that have been systematically neglected by the state-monopolized statistical institutions — constitute jointly two of the most important qualities of these relationships: mutual trust and collaboration. For this reason, accessibility and confidentiality are crucial to the functioning of the entire information system, which, in turn, affects significantly the process of the development of market-based democracy.

The fact that interest in confidentiality and data access is so persistent is a consequence of the very nature of these issues; they derive from challenges to

[1] The author served as an SSRC program director to the Committee on Confidentiality and Data Access. Support for this study was provided by the National Science Foundation, the Bureau of the Census, the Bureau of Labor Statistics, the Internal Revenue Service's Statistics of Income Division, the National Institute on Aging, the National Center for Education Statistics, and other federal agencies through their contributions to the work of the Committee on National Statistics

statistical agencies brought about by continual changes in the environment. Specifically, while both computing power and the analytical capabilities of data users are ever expanding, there is a diminishing willingness on the part of individual data providers to participate in statistical surveys, thus lowering the completeness and quality of the collected data.

The increasing exposure of persons and organizations to a dizzying variety of information-gathering activities not only represents an ever-larger claim on their time, but more importantly raises fears about who will have access to information about them.

One reason why the public increasingly feels that its privacy is being eroded by the organizations that develop and use microdata bases is the fact that many statistical agencies lack adequate legal authority to protect identifiable statistical records from mandatory disclosures for non-statistical use. At the same time, the linkage of data from different sources, perceived by the public as a particular threat, increasingly attracts both government and non-government data users, who have at their command ever greater technological advances.

From the point of view of data users, existing restrictions on inter-agency sharing of data — for instance, the prohibition against access to such administrative records as Social Security earnings reports (in USA) — prevent some important analyses and leads to costly duplication of certain statistical programs. Policy analysts and other nongovernmental users who receive smaller and smaller amounts of detailed information on individual units from the statistical agencies, especially in such forms as public-use microdata files, feel that their ability to contribute to the understanding and resolution of significant economic and social problems is also narrowed.

Another key point of interest in the Panel's work is that the panelists take for granted that statistical agencies have legal and ethical responsibilities toward the data subjects and data providers, as well as to data users and to other agencies involved in the development of data bases. Hence, their analyses and recommendations were guided by three principles of information: *democratic accountability*, *constitutional empowerment*, and *individual autonomy*. They have targeted their recommendations to four areas: legislation, administrative policies, statistical disclosure limitations, and "ethical issues."

There is not enough room here to discuss the recommendations in detail. However, some brief examples may provide insight into the general nature of the topics covered, while allowing for a demonstration of certain key terms. [As mentioned earlier, establishing definitions for such terms is another way in which the panel has contributed to systematic knowledge of data protection and use.]

1. In light of the fact that many statistical agencies lack the legal authority to protect identifiable statistical records from mandatory disclosures for non-statistical use, the panel proposes that the ***principle of functional separation*** (as enunciated in 1977 by the Privacy Protection Study Commission) be consistently employed: **Data collected for research or statistical purposes**

should not be made available for administrative action about a particular data subject.

2. Because of the cost and excessive burden on data providers implied by some barriers to inter-agency sharing of data, the panel suggests that in specific instances (where it is significantly beneficial) data sharing should occur, given that protection of confidentiality is assured.
3. The panel expresses concern about the consequences of the reduction in the amount of detailed data provided by statistical agencies to nongovernmental users in tabulations and public-use microdata files. To aid researchers and policy analysts, the panel recommends the continuation and expansion of efforts to provide more detailed data to users, and suggests that users do everything within their power to prevent disclosure, under legal sanction.
4. Given that privacy is being eroded by authorized organizations (e.g., through linkage of data from different sources), the panel suggests application of a multi-stage procedure: exploration of the possibilities for greater use of administrative record sources, on condition that the consequences of participation in surveys be consistent with public opinion regarding *informed consent* and related issues (e.g., on data sharing for statistical purposes); and active solicitation of the views of advocacy groups concerned with privacy issues.
5. Because of the publicly perceived threat from technological advances in computers and communications, the panel suggests that all agencies be responsible for developing effective statistical disclosure limitation techniques for all forms of data dissemination. This should be of particular concern with regard to the release of new public-use microdata files.

4. Specific points of relevance of the Panel's work to the statistics sector in Eastern Europe

"You can't have a democratic society without having a good data base." This statement by Janet Norwood — former commissioner of the Bureau of Labor Statistics, whose remarks were taken by the authors of the report as a motto for one of its chapters — might serve as the credo of the information systems policy in all market-based democracies. This means that questions such as how a country's data system contributes to pursuing two essential goals of a democratic polity, accountability and the representation of diverse interests, should be among the important targets of the efforts toward strengthening the infrastructure of the statistics sector in countries redesigning their public statistical systems.

It was recognized at the beginning of transition to markets and democracy that it is an information-intensive process. Also, that there is no good decision without good information, and that the latter depends crucially upon whether or not there are legal and operational rules to protect identifiable statistical records from mandatory disclosure for non-statistical use. Neglect of the issues of confidentiality

and data access, the key aspects of data collection and dissemination, has contributed greatly to dysfunctionality in the information system in the previously centrally-planned economies.

The problem is not only what must be done in order for data to be appropriate for market conditions. Even more important is the problem of users' "perceptual bias" rooted in suspicion of previously dominant practices. There are also important differences in the sociocultural characteristics of the environment in which the trade-off between confidentiality and data access is being considered, such as the contradictions between professional standards and some of the routine manners and attitudes that developed in the statistical and administrative staffs under the former regimes, and how these might be resolved.

5. Summing-up

It is increasingly recognized in each of the newly emerged democracy in Europe that confidentiality must become an essential part of legislation. But also, as noted at the beginning of the transformation (Blades, 1993), it will take years to convince respondents that their answers are indeed held in confidence.

So, confidentiality and accessibility, which jointly affect the quality and usability of the data produced, depend directly upon the relationship between the three major players on the statistical scene: data providers (individuals and organizations), data producers (statistical offices), and data users (including nongovernmental units and the general public).

As they restructure their statistical systems, the reforming governments have originally focused primarily on their systems' "technological" capabilities, including hardware and software upgrades, and on new programs to meet the informational obligation imposed by the international organizations. At the same time, they have generally failed to recognize the fundamental importance of the relationships among the above-mentioned players for maintaining the wide range of principles and practices involved in the effective operation of any statistical system.

Therefore, given the panel's focus on the ethics of information and its emphasis on the legislative aspects of statistical activity, which encompasses all of the three major parties in the data process, its findings seem especially pertinent to the problems specific to the formerly socialist economies, including providing a useful framework for dealing with such urgent issues in the restructuring of their statistical systems as issues of making the actual policies and institutional circumstances conducive to the enhancement of mutual trust and respect between statistical agencies, data subjects, and data users.

There is a lot to be learned from the panel's accomplishments in this respect. Specifically, from its focus on the institutional aspects of the production and use of "good data" that might help to guide the efforts of all parties involved as well

as researchers toward improving the public information systems appropriate for markets and democracy.

REFERENCES

BLADES, DEREK, Comments on "Statistics has changed in the Transition Countries: How to counvince the Users" by Rodocea and L. Dumitrescu in Statistics in Transition: Journal of the Polish Statistical Association, Vol. 1, No. 2, December 1993.

DUNCAN, GEORGE T., JABINE, THOMAS B., and VIRGINIA A. de WOLF, eds., *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics,* National Academy Press, Washington D.C.,1993.

Eurostat (1996) *Manual on Disclosure Control Methods,* Luxemburg: Office for Publications of the European Communities.

FELSÖ, F., THEEUWES, J., WAGNER, G. (2001) **"**Disclosure Limitation Methods in Use: Results of a Survey" Chapt. 2 in Doyle, P., Lane, J., Theeuwes, J., Zayatz, l., (eds) Confidentiality, Disclosure And Data Access: Theory And Practical Applications For Statistical Agencies. Elsevier, New York Amsterdam.

HOLVAST, J., (1999) 'Statistical Dissemination, Confidentiality and Disclosure', in Eurostat: Statistical Data Confidentiality—Proceedings of the Joint Eurostat/UNECE Work Session on Statistical Data Confidentiality Held in Thessaloniki in March 1999.

JABINE, T. B. (1993), "Procedures for Restricted Data Access," Journal of Official Statistics, Vol. 9, No. 2, p. 537—589.

LUIGE, T., and JANA MELISKOVA (1999) 'Confidentiality Practices in the Transition Countries', in *Eurostat: Statistical Data Confidentiality—Proceedings of the Joint Eurostat /UNECE Work Session on Statistical Data Confidentiality Held in Thessaloniki in March 1999.*

OKRASA, W., (1991) Research Infrastructure of the Federal Statistical System in USA" ("Infrastruktura Badawcza Federalnego Systemu Statystycznego w U.S.A") Biblioteka Wiadomości Statystycznych, No. 2(369):36—38.

OKRASA, W., (1994) "Private Lives, Public Policies. Report of the Panel on Confidentiality and Data Access and its relevance for designing information systems in Central and Eastern Europe", Items, Vol. 48 Nr.1 (March 1994), Social Science Research Council, New York.

Statistical Policy Working Paper 22 (Second version, 2005) Report on Statistical Disclosure Limitation Methodology; Statistical and Science Policy Office of Information and Regulatory Affairs, Office of Management and Budget December 2005

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 175—181

SOME LEGAL ASPECTS OF STATISTICAL CONFIDENTIALITY OF MICRO-DATA COLLECTED BY OFFICIAL STATISTICS IN POLAND — AN EXPERT'S VIEW[1]

Tadeusz Walczak

1. The official statistics provides reliable and objective information on economic development, demographic and social situation in Poland for the society, the state and public administration bodies and other interested institutions. The thorough execution of this duty needs to implement and observe **many methodological, organizational and legal principles. The violation of these principles could cause a deterioration of information quality and decrease of confidence to statistics, very difficult to recover.**
2. One of the most important conditions of preparation a complete and reliable information is to provide reliable and consistent with the actual true values data received from respondents (persons, households, business entities including farms) according to survey programs. The official statistics services spare no effort to establish appropriate co-operation with respondents to acquire their active participation in surveys and to receive true and relevant data. They rely on precise explanation of survey targets, appropriate methodology and documentation development as well a proper survey organization, trainings of the personnel etc.
3. The active participation of respondents in statistical surveys and receiving from them true data can be only possible under full confidence of respondents to the official statistics services in charge of surveys. The confidence can be gained on the ground of full guaranty that **data supplied by respondents to statistical services in a survey shall be used exclusively for statistical compilations and analyses and shall not be used for any other purpose as well individual data with characteristics allowing respondents identification shall be protected with great care and will be given nobody any access.** Even single case of microdata disclosure can undermine the respondents' trust to statistics services for a

[1] The text is a translation from Polish of an expertise written for the Central Statistical Office of Poland. Author is grateful for the CSO's permission to submit it for publication in Statistics in Transition.

long time. It will cause increased percentage of refusals to participate in voluntary surveys and it will be reflected in the data quality. In case of compulsory surveys. Such a situation will cause increased quantity of inaccurate and unreliable answers. **Breach of the ban of access to microdata[1] allowing the identification of the reporting units could make impossible the existence of the objective and reliable official statistics with difficult to estimate damages for economy and for the authority of public institutions.** At the turn of the 1980s, taking account of the fundamental changes of the political, economic and social situation of the Country in Poland new statistical principles foundations have been built. Among them many far-reaching changes were introduced in the all system of collecting, processing and dissemination of information. The changes were expressed in the new Law on official statistics issued on 29th June, 1995. The Law was prepared on the base of the best experiences of a dozen or so countries and international statistical organizations. Owing to this the Law has implemented many rules typical for democratic societies, i.e. a significant lessening of the administrative compulsion to execution of duties for the benefit of statistics and stronger stressing the public nature of statistics[2]. Respecting companies' autonomy in the field of their internal information systems and guaranteeing information arrangement in the country by use of homogeneous nomenclatures and classifications, the Law issued in 1995 did not include i.a. a duty to adjust the internal information companies systems to the requirements of the reporting obligations, **the power to inspect the internal documentation, reliability and the timeliness of statistical reporting.** The rules of the new Law aimed to base the official statistics system on confidence of the official statistics services and respondents understanding the needs of the modern information system and undertaking to provide reliable statistical data on their economic activity. On the other hand the Central Statistical Office and its services were obligated to guarantee the respondents that **their data shall be used exclusively for statistical compilations and analyses and nobody will be given an access to data in a form making it possible to identify the respondent.** The legal provision included in the article 10 of the Law serves the purpose[3]. During the parliamentary debate on 15th December, 1994 the CSO President underlined: "The Law introduces explicitly **protection of information, collected during statistical**

[1] Microdata - personal data related to a concrete natural person or individual data related to a business entity or other legal entity or an entity without a legal status

[2] Previous Laws were entitled „on administrative statistics" than „on state statistics" and the Law issued in 1995 is entitled „on public statistics"

[3] The statistical services can give an access to microdata for scientific institutions for analyses and studies. In each case the data should be deprived of any attribute making possible to identify the unit concerned and the data cannot be used for other purpose than analyses.

surveys, and being subject to the statistical confidentiality. There is an absolute obligation to observe the statistical confidentiality. Due to the fact it is the confidentiality similar to a medical secret. It means that among different confidentialities/secrets: company, office, trade, the statistical confidentiality doesn't accept any exemption, is uncompromising and its observance must be explicitly organizationally, technically and legally guaranteed".

4. The observance obligation of microdata confidentiality protection results from the Polish Law issued on the 29th June 1995 on official statistics as well from many international legal and legal-ethical documents. The rules concerning protection obligations of microdata are included in some articles of **the Law on official statistics. Article 10** defines: "The collected and gathered in the statistical surveys of official statistics individual and personal data shall be confidential and subject to particular protection; the data shall be used exclusively for statistical calculations, compilations and analyses and for the creation by the statistical services of official statistics sampling frames for statistical surveys conducted by those services; providing or use of individual and personal data for other than specified above purposes shall be prohibited (statistical confidentiality)". **Article 12** imposes particular obligations of the statistical confidence observance on statistics services staff: "The staff of the official statistical services, the census enumerators, statistical interviewers and other persons performing activities in the name and on the behalf of official statistics, having direct access to individual and personal data shall be obliged to observe without exceptions the statistical confidentiality and shall be allowed to perform those activities only after delivering an oath in a written form, at a statistical office or other units of official statistical services".**Article 38: "**1. It shall not be allowed to publish or disseminate individual data obtained in the statistical services of official statistics. 2. It shall not be allowed to publish or disseminate obtained in statistical surveys of official statistics statistical information which can be linked or can identify natural persons or individual data characterizing business entities, especially if the aggregated data consist of less than three entities or the share of one entity in the compilation is higher than the three-fourths of the total". **Article 39:** The President of the Central Statistical Office shall ensure that the storing of collected statistical data guarantees observing the principles of statistical confidentiality". **Article 54:** "Who violates statistical confidentiality shall be subject to imprisonment up to 3 years".
5. International rules regulating the functioning of official statistics systems devote much attention to the protection of microdata. The purpose of these rules is to provide reliable data comparable on the international scale. **The rules aim at the protection of the units reporting in good faith their**

business data to statistical services as well as the creation of confidence atmosphere of the units to statistical services. This atmosphere constitutes the necessary condition of the credible and reliable statistics. Among the more important international documents regulating the statistical services duties in the area of the microdata protection are:

1) **Convention No. 108 of the Council of Europe** on the Protection of Individuals with regard to Automatic Processing of Personal Data[1]. It includes general principles of the personal data protection concerning different areas: medicine, social insurance, banking, employment, advertising, statistics, telecommunications, police etc. Their entry into force takes place on grounds of areas recommendations addressed to governments prepared by project groups under the protectorate of the Consulting Committee of the European Council. The principles of the personal data protection collected and processed for statistical purposes were formulated in Recommendation adopted in September, 1997[2]. The Recommendation stresses a necessity to observe the personal privacy during all surveys using personal data. It stresses too that data collected and processed for statistical purposes must be used exclusively for these purposes. They may not be used for other purposes other than purposes defined by surveys. Particularly they may not be used to make decisions or to take steps against persons, who the data concern, as well as to completion or correction data files which data are processed for other than statistical purposes (stressing T.W.).[3]
2) After over seven years work within international statistical groups in 1985 the General Assembly of the International Statistical Institute[4] approved "Declaration on Professional Ethic for Statisticians". The Declaration stresses that statisticians must be aware that nearly each survey involves a certain burden for entities selected to surveys (persons, households, economic units, institutions). The degree of

[1] Signed on the 28th January, 1981 in Strasburg came into force on the 1st October, 1985. Poland signed the Convention in 1999 and ratified it on the 24th May, 2002 as a condition of the access to the EU.

[2] European Council, Ministries Committee, Recommendation, No. R(97) 18 of Ministries Committee for Member States concerning the protection of personal data gathered and processed for statistical purposes adopted by Ministries Committee on the 30th September, 1997 during 602nd Delegated Ministries Meeting.

[3] There is allowed and legal to process for statistical purposes the data which were primarily collected for other than statistical purposes. It concerns i.a. administrative data used by official statistics services. Appropriate guarantees must be maintained that the data will not be used to take decisions against the persons or other entities which the data concern.

[4] International Statistical Institute (ISI) established in 1885 is one of the oldest scientific associations all over the world, including eminent representatives of statistical science and practice. Nowadays the Institute has over 2000 members elected and representing 133 countries. Including members of specialized associations there are about 5000 ISI members coming from 156 countries.

burden can be different. Different can be the feelings and reactions of each entity. Some types of questions included to the survey program can cause negative reactions of respondents and their reluctance to answer the questions. Statisticians can receive a part of data on inquired units from other sources without knowledge and consent of interested entities. Statisticians shall extend the use of data from administrative sources to limit respondents' burden. They shall remember to protect the individual data and to use so collected data exclusively for analytical, statistical and not for administrative purposes. The Declaration emphasizes the obligation of protection of individual data gathered during statistical surveys and the ban on their dissemination to anybody and in any form. It obliges statisticians not only to the effective protection and to non-dissemination of individual data but also to prevent the direct or indirect disclosure of such data (item 4.6 of the "Declaration").

3) Very important document for statistical services are "Fundamental Principles of Official Statistics" adopted during 47th session of the UN Economic Commission for Europe as the document regulating the activity principles of statistical services in the ECE. This very document was considered of a universal importance by Statistical Committee of Economic and Social Council of Asia and Pacific as well as by Common Conference of African Planners, Statisticians and Demographers. It prepared grounds to adopt the document by UN Statistical Commission and UN Economic and Social Council on 15th April 1994 as official document of the United Nations. "The Fundamental Principles" constitute a peculiar statistical Decalogue including "ten commandments" which shall be kept by statistical services of each country. One of the principles of particular importance for the formulated hereby expert opinion concerning responsibility of the microdata protection, namely **the 6th principle** states that: "**Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes**."[1]
4) Many legal rules in European Union regulate the question of protection of microdata collected by statistical services. The most important is **Council Regulation (EC) No 322/97 of 17 February 1997 on**

[1] In the course of a round-table conference discussion on the implementation of "Fundamental Principles" organized by CSO Poland under the protectorate of UN Statistical Commission, European Economic Commission and Eurostat resulted legal rules regulating principles of official statistics activities in all countries ensure the full protection of microdata. Especially they do not make possible to access to individual statistical data for law enforcement agencies and tax offices.

Community Statistics (OJ No. L. 52.22.2.97). This very legal document in item 13) of the preamble states: "Whereas it is important to protect the confidential information which the national and Community statistical authorities must collect for the production of Community statistics, in order to gain and maintain the confidence of the parties responsible for providing this information; whereas the confidentiality of statistical information must satisfy the same set of principles in all the Member States (stressing T.W). **Article 10** defines "confidentiality" as "the protection of data related to single statistical units which are obtained directly for statistical purposes or indirectly from administrative or other sources against any breach of the right to confidentiality. It implies the prevention of non-statistical utilization of the data obtained and unlawful disclosure". **Article 13** states: "**Data used by the national authorities and the Community authority for the production of Community statistics shall be considered confidential when they allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information.** To determine whether a statistical unit is identifiable, account shall be taken of all the means that might reasonably be used by a third party to identify the said statistical unit". **Article 15** stipulates that "Confidential data obtained exclusively for the production of Community statistics shall be used by national authorities and by the Community authority exclusively for statistical purposes unless the respondents have unambiguously given their consent to the use for any other purposes".

5) The European Statistics Code of Practice adopted by the Statistical Programme Committee in February 2005 formulates principles of the independence, reliability and liability of National and Community statistical institutions. Among the 15 most important principles being obligatory for Community and National institutions the Principle 5 "Statistical Confidentiality" stresses that "The privacy of data providers (households, enterprises, administrations and other respondents), the confidentiality of the information they provide and its use only for statistical purposes must be absolutely guaranteed".[1]

In order to ensure the uniform implementation of principles adopted in the Code, Eurostat, responsible to the Commission, conduct many training and inspection activities. Within the inspection activities the so-called peer reviews in each country are organized. During the reviews the activities of National

[1] The importance of the Code to ensure the quality and reliability of statistical information as well as the necessity to observe it was stressed in the European Commission Communiqué to European Parliament and Council of 25 May 2005 on the independence, reliability and liability of National and Community statistical institutions.

Statistical Offices are analyzed in detail especially the implementation of 15 Principles formulated in the Code.[1]

Conclusions

1. The protection of confidential microdata received from respondents by statistical services shall be absolutely guaranteed. If the legal regulations are not clear enough in the light of the present legal interpretations, they shall be modified in the meaning as the CSO President underlines during Parliamentary debate (see item 3 of this expert opinion).
2. The breach of the statistical confidentiality by statistical services can cause an incalculable harm for the reliability of the statistical information. Even single messages concerning the danger of microdata disclosure by CSO (such messages were published in the press last time) can be used by respondents to excuse their reluctance to participate in surveys or to give inaccurate information.
3. We cannot ignore international reactions too. The worldwide public opinion, especially of statisticians, international organizations including European Union are particularly sensitive to observe the protection obligations of confidential microdata. Information on a breach of this principle can cause adverse consequences of international circles on positive opinion on Polish statistics as well on authorities and institutions being responsible for the observance of these principles.

[1] The peer review was conducted on the 25 – 27 April 2007 in Central Statistical Office in Poland. Among issues concerning realization of the Code Principles by Polish statisticians, the protection of the statistical confidence was detailed analyzed.

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 183—186

ON THE IMPORTANCE OF STATISTICAL CONFIDENTIALITY FOR THE QUALITY OF SURVEY DATA

Mirosław Szreder

1. Introduction

The increasing role of information in modern societies (information societies), which serves both public institutions and market enterprises to make decisions, is the reason why the price of reliable and high-quality information is rising. Credible information creates favourable conditions for taking right decisions. Incorrect or low-quality information distorts reality and leads to wastage of efforts, time and money. It concerns both the world of great politics, where secret services' wrong information can cause unnecessary wars and disasters, as well as, in smaller scale, societies, local communities, consumer groups exposed to wrong decisions of those who make them basing on false premises or low-quality information. In most cases, the problem is not individual registered data but sets of numbers, i.e. information of statistical nature. At present, what is regarded as the most exact description of wide areas of social and economic reality are sets of numbers and information on different characteristics of surveyed populations themselves. The importance and the price of reliable statistical information are increasing because, at the opposite side, the costs of wrong decisions (or not taking decision) are rising as a consequence of low-quality information received by decision-makers.

2. The role of respondent in statistical surveys

In order to present the real and possibly complete picture of reality by means of statistical data one needs a well-prepared survey project with an appropriate measuring instrument (a questionnaire) as well as a positive attitude and trust that the surveyed persons have in survey workers. The willingness of persons selected for a survey (respondents) to cooperate is particularly important in all kinds of social and economic issues. Unlike experimental fields, measurements in such surveys are not taken by statisticians responsible for preparing a measure

instrument in a proper way, but by respondents who provide a statistician with necessary information. A statistician does not count persons employed in a factory, does not estimate the property of a businessman, does not control the amount of money spent by a consumer for his shopping. Such information is provided by respondents themselves. A statistician's duty is to persuade the respondent to give him reliable and precise information which can be used to make right decisions. What in turn depends on a respondent is his willingness to share his knowledge with a pollster, to give true information, but also to distort the part of reality he knows or even to refuse to participate in a survey at all. As statisticians, sociologists and researchers of public opinions' long-standing experience shows, the respondent's trust in the researcher, his deep conviction that facts and opinions disclosed in a survey will remain anonymous, are of fundamental importance for his attitude towards the researcher. Each respondent wants to be certain that no person and no institution will have access to the individual information given by him. In any publications based on such kinds of surveys, there can be also no risk that the respondent will be identified by anybody. The respondent's conviction that all his information is protected by statistical confidentiality is the basis of the cooperation between the researcher and the respondent and the quality of information collected in a survey depend on it to the highest degree. We, statisticians, observe repeatedly that respondents do not wish to reveal any facts and opinion provided by them. They fear for their image in their circles, for the competitors' behaviour on the market, for intervention of public authorities and sometimes for their safety.

3. The statistical confidentiality and the quality of quantitative surveys

Each case of a violation of the statistical confidentiality can seriously damage the respondents' trust to mass surveys of all types. The respondent anxiety over expressing his opinion or information results simply in either refusing to participate in a survey or giving unreliable or false data. Both attitudes lower the quality of statistical surveys and directly affect the creation of a false picture of reality for users of surveys' results. In such circumstances, diminishing the quality of surveys is acute to such an extent that it is difficult to estimate an error that the above mentioned respondents' attitudes affect the final results of the surveys with.

As recipients of opinion polls, we are accustomed to getting the statistical error estimate, usually at 3 per cent level, that should be attributed to poll's results. In fact, this error is the smallest component of the whole error with which the results of many quantity surveys are affected. Apart from this error, there are 4 other types of errors, *i.e.:*

- coverage error, which can be a result of using out-of-date or incomplete population list in a survey;
- non-response error;

- measurement error, related to registration of false information on a surveyed respondent;
- post-survey processing error.

As many as two of above mentioned errors are directly connected with the respondents' attitude. The non-response error can result from the loss of the respondent's trust in a survey worker or insufficient conviction that he remains anonymous. This kind of error category is the one which still strongly affects the general quality of surveys despite ever more advanced missing data imputation techniques. While societies in many countries are observed to become increasingly tired of the number of different questionnaire surveys, it is important not to deepen such tendency through, among other things, the lack of consequence in keeping statistical confidentiality as it leads to undermining a general trust to statistical surveys. The trust in this matter once lost can be difficult to regain. Official statistics institutions as well as the whole statisticians, sociologists, market and public opinion researchers circles should care for it. Another error category, which is influenced by respondents' attitude, is measurement error. The source of this kind of error can be a question in a questionnaire wrongly formulated or wrongly understood by respondent, but also purposely giving false information. The reason for the last mentioned behaviour is the respondent's fear of revealing the true information to the third parties that is the lack of trust in survey worker.

In Poland, like in other countries, there are many institutions and organizations conducting different statistical surveys. They all should be concerned with strengthening public trust in a survey worker and the conviction that such surveys are useful. The confidentiality of individual data collected in surveys and the observance of the respondents' anonymity principle are two of the most important rules which all statistical institutions should follow. The Central Statistical Office (CSO) and the regional statistical offices are obliged to keep statistical confidentiality by law. The full observance of this principle builds society's trust in these institutions. The breach of, or even a single exception from the rule, damage a delicate bond between a researcher and a respondent, which is necessary to obtain full and true information from respondents on a given subject.

4. Persons concerned with the observance of confidentiality principle in statistical surveys

Who, in the first place, should seek the full observance of statistical confidentiality guaranteed by law? Certainly, statisticians are interested in this matter, but, as it appears, the receivers of surveys' results should care for it to the same or higher degree since they are the persons most concerned with high-quality data. To a large extent, statistical data are used by ministries, other central bodies, local administration offices and many other units, in which the most important decisions for society or local community are made. These very

organizations, being convinced that the right decisions can be taken only on the ground of reliable and possibly full information should demand protection for statistical confidentiality when it is threatened. It would be tantamount to demand high-quality statistical data. According to the basic principles of the official statistics published by the United Nation Organization, the confidentiality principle should be obligatory for all statistical organizations. This principle states: "Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes"[1]. All institutions *in a* democratic state *should* respect th*is* principle *having* regard *to well* understood social interest in shorter and longer time *perspective*.

[1] An Extract of the „Fundamental Principles of Official Statistics" adopted by the United Nations Statistical Commission in its Special Session of 11-15 April 1994.

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 187—188

CONFERENCE ON THE OCCASION OF THE 90TH ANNIVERSARY OF THE CSO, KRAKOW, 28TH —30TH JANUARY, 2008

The scientific conference inaugurating celebrations of the 90th Anniversary of the Central Statistical Office took place from 28th to 30th January 2008 in Krakow. The beginnings of the CSO date from the 3rd January 1918, when the Regent Council issued a decree on the temporary organization of general authorities in the Polish Kingdom.

On the force of this decree the Statistical Department was created and included into the Ministry of the Interior. It was the origin of the future Central Statistical Office, because on 13th July 1918 the Regent Council of the Polish Kingdom issued a Rescript on the formation and organization of the Central Statistical Office, which was formally created then as an independent unit of Polish state administration.

The conference was held under the patronage of the President of the CSO, Prof. Józef Oleński, PhD. It was organized by the Statistical Office in Krakow. The Director of the Office, Krzysztof Jakóbik, managed works of the organizational committee.

The participants of the conference were employees of the CSO and Statistical Offices from all over Poland as well as representatives of public administration units, scientific circles and other receivers of information in the field of social research. There were also invited representatives of local self-government authorities, senators and deputies of the Krakow region, directors of public administration units and non-governmental organizations interested in social research issues located in Małopolska.

The conference, apart its jubilee character, had a main scientific goal. It was entitled “Social statistics – accomplishments – chances – prospects”. And these were predominating issues during the conference. Speeches presented social research issues, their new directions and development of methodology in reply to a wide demand in this field.

Accomplishments and challenges of public statistics were discussed as regards the satisfaction of information needs of modern society.

For making the most of this conference, and to improve the information service according to expectations and demands of receivers, not only statisticians were invited to discuss, but also the most important representatives of researchers and receivers of the results of surveys themselves.

During the conference over twenty broad-spectrum speeches were presented, ranging from general methodological problems to selected research areas and concrete results of surveys in detail. Issues proposed by speakers covered a wide range of fields: demography, research on life conditions with special consideration of poverty, social exclusion problems and quality of life, condition of public health and evaluation of individual health, problems of social life, development of local communities, activities of public institutions and non-governmental organizations as regards social politics, issues of migration and depopulation, social and cultural transformation as well as forecasts of further directions of social changes.

Krakow, with its deep and strong roots in scientific circles, which guarantees its credibility, is a special place to emphasize a reliable and objective character of public statistics institution. Therefore, the conference was held just in this place where these roots of institution of Polish public statistics reach.

First of them is the Jagiellonian University and the second one the Municipality of Krakow – already before the formation of the Central Statistical Office, on the 3rd January 1884, the Council of Krakow appointed the Jagiellonian University professor, Dr Józef Kleczyński as the Director of the Municipal Statistical Bureau.

These best traditions are continued up to now, what was proved once again by the conference. As mentioned, scientists from leading scientific centres all over Poland participated in the conference, among them so well-known seniors as Prof. Kazimierz Zając, PhD or Prof. Jan Małecki, PhD.

This meeting will surely have a good influence on further development of research in field of social statistics in Poland and will contribute to deepening the cooperation between all people interested in this issue. Collection of speeches presented during the conference will be published in special series of Statistical News in order to make them available to a wide circle of interested people.

The conference was the first of those organized by the CSO in the jubilee year. Their subject matter concerns consecutive fields in which public statistics serve the society in respect of the creation of information order.

STATISTICS IN TRANSITION-new series, June 2008
Vol. 9, No. 1, pp. 189—192

PUBLIC STATISTICS AS ONE OF THE FUNDAMENTS OF SELF-GOVERNANCE IN A DEMOCRATIC STATE, TORUŃ, 16—18 APRIL 2008

The Scientific Conference "Public Statistics as one of the Fundaments of Self-Governance in a Democratic State" took place from 16th to 18th April 2008 in Toruń. It was held under the patronage of Prof. Józef Oleński, the President of the Central Statistical Office (CSO) of Poland, and Piotr Całbecki, the Marshal of Kujawsko-Pomorskie Voivodship. The main goal of the conference, aside from celebrating the 90th Anniversary of the CSO, was to emphasize the necessity of creating an enduring platform of agreement between local governments, scientific research units (the theoreticians of regional development) and public statistics as well as indicating the important role of regional statistics as a useful instrument in the realization of tasks connected to regional development.

The conference was solemnly inaugurated by Piotr Stolarczyk, the Director of Statistical Office of Bydgoszcz. The opening speech was given also by the Vice-President of the CSO Dr Halina Dmochowska, who, after greeting all of the guests, presented the role of public statistics in service of self-government and democracy, in particular improving the statistical system at the local level with the help of modern techniques of data gathering, processing and dissemination.

All of the participants were also given warm welcome by Franciszek Złotnikiewicz, the member of the Voivodship Board, who was representing the Marshal of Kujawsko-Pomorskie Voivodship. He emphasized that without trustworthy statistical information it is impossible for a civil society and democratic state to work properly. On behalf of the Rector Magnificus of Nicolaus Copernicus University in Toruń (NCU) Prof. Grzegorz Jarzembski, the Vice-Rector for Development and Computerisation gave his concerns over scientific support, which is necessary for a further development of statistics. The president of Toruń Michał Zaleski reported a short paper about taking advantage of statistical data in the process of city management.

Amid the participants of the conference there were representatives of academic and scientific sphere, particularly Polish universities, public administration and local governments, including representatives of the Offices of the Marshal of Kujawsko-Pomorskie, Łódzkie, Mazowieckie and Opolskie Voivodships, as well as employees of the CSO and regional statistical offices.

The programme of the conference included 26 presentations covering very wide scope of topics, from general methodological issues, through public

statistics frameworks, to specific case analysis. Given that, the scientific session of the conference was divided into 6 blocks of topics.

I part: "Statistics, Self-Governance, Democracy" covered speeches presenting existing and possible relations between public statistics and local governments. Prof. Czesław Domański (University of Łódź) dealt with a question, whether it is efficient statistics that generates nation's prosperity or, on the contrary, it is prosperity that determines better statistics? Next lecture, from Dr Kazimierz Kruszka, the President of the Polish Statistical Association (PTA) characterised forms and levels of cooperation between public statistics and Polish cities governments. Prof. Ludosław Drelichowski (University of Technology and Life Sciences in Bydgoszcz) presented the conception of creating knowledge base system which is necessary for realisation of supervising function in area of city's municipal services.

II part: "Statistical methods in regional and local analysis" was opened by the distinguished scientist Prof. Kazimierz Zając, who gave an overview of the connections between statistics theory and development of natural sciences. Next speaker, Prof. Jan Paradysz (The Poznań University of Economics) characterised an innovative research project in domain of Polish public statistics which assumes the use of indirect estimation method during national censuses planned for 2010 and 2011. Prof. Tadeusz Kufel (professor of NCU in Toruń) demonstrated a computer software that allows direct use of statistical data from CSO's Data Regional Bank for analysis and building econometric models. Dr Joanna Bruzda (NCU in Toruń) presented a model in which national economies work as evolving systems and their development can be described as a process with elements changeable in time. This part of the session ended with a joint lecture from Dr Iwona Markowicz and Dr Beata Stolorz (University of Szczecin) concerning the impact that quality of the statistical data has on results of research on unemployment.

III session: „The level, dynamics and innovation of socio-economic regional development". Mr. Roman Chmielewski from the Ministry of Regional Development analysed a spatial diversity of the level of socio-economic development of Poland, which was presented on NTS-2 and NTS-3 level, using data from CSO's Data Regional Bank. Ms. Dorota Doniec (the Manager of Regional Accounts Center in Regional Office of Katowice) took up similar issues in her speech. She talked about regional accounts which provide full, essential information for making economic analyses. Ms. Małgorzata Góralczyk, (The Polish Academy of Sciences) presented possibilities of making economic analyses, also working out rates which describe economy sectors with the use of a research device — matrix NAMEA. This session was finished by Dr. Irena Łącka's speech (University of Agriculture in Szczecin) which concerned the measurement of innovation in Poland.

IV session: „The cooperation of public statistics with the local administration" was started by Dr Ewa Wędrowska (University of Warmia and Mazury in

Olsztyn). Dr Ewa Wędrowska made an attempt at drawing up a suitable method of filtering data, so that to receive ultimate tables. These tables constituted the biggest and the most important collection of information. Dr Andrzej Potoczek from Marshal Office Kujawsko-Pomorskie Voivodship reported forms of cooperation between Marshal Office and Regional Offices of Bydgoszcz. He stressed that all common activities led to shaping regional politics foundations, moreover, to regional politics cohesion. Mr. Marek Olszewski (the Vice-President of Union Rural Communes of Republic of Poland, and the Administrative Officer of the Lubicz Commune) took up analogous issues. On completion of the session Mr. Paweł Strzelecki and MR Andrzej Tomaszewski from The Office of Geodesist of Mazovian Voivodship acquainted participants with Mazovian Spatial Information System.

V session: „The public statistics utilization by local administration". The achievements and directions planned of the research of public statistics were presented by The President of CSO Prof. Józef Oleński. Next, Ms. Dominika Rogalińska (the Deputy Director of Department of Regional Research and Environment CSO) analysed the system of information of public statistics. In further part, Mr. Wiesław Łagodziński (the Public Relation Officer of CSO) described statistic information as a potential base of information for the local governments. Dr Dominik Rozkrut (the Director of Regional Office of Szczecin) presented the protection of statistic confidentiality. Next, Ms. Bożena Łazowska (the Director of Central Statistic Library) reminded an important role of public statistics, which is that of educating people. The session was finished by Dr Dusana Bogdanova's speech (University of Opole). He presented the results of research into an interest in public information, published on the Public Information Bulletin websites.

VI session: „Modern public finances management — budget task implementation" was begun by dr Tomasz Strąk (Szczecin University), who reminded the definition of budget task as a budget method, in which expenditure are included in formulating objective tasks. He mentioned the necessity of introducing a public system finances to adapt to public administration unit management system. This issue was developed by Mr Tadeusz Dobek, the President of Regional Chamber of Financial in Bydgoszcz, who presented it in connection with the budget expenses structure in socio-economic situation of Kujawsko-Pomorskie Voivodship.

An uncontested value of the conference was the discussions which closed blocks of topics. During the debates, conclusions from results of all sessions were presented. The problem of recognizing the needs for information in economic range of local government unit was regarded as the most important one. Another issue was organization of information infrastructure and use of administrative data sources for statistical purposes. A different matter was the necessity of Polish official statistic to take part in defining the terms in administrative system of information.

Local government representatives paid attention to the fact, that without a reliable description of reality it is not possible to make proper decisions enabling the region management.

The conference was the second of series of jubilee conferences organized by Central Statistical Office of Poland. The subject matter of the following conferences will be closely related to the scope of public statistics studies.

Conference papers will be published soon on the conference website http://www.stat.gov.pl/bydgosz/konf/index.html.