STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 179—180

FROM THE EDITOR

Following the well-established (in eyes of the journal's readers and authors alike) formula for arranging the articles through starting from those devoted to sample and estimation related procedures, the first part of this issue consists of six papers addressing such types of problems. They range, however, from sampling to variety of nonsampling factors and relevant distributional measures. M. S. Ahmed and Atsu S. S. Dorvlo derive the general estimators for finite population mean using multivariate auxiliary information (under multiphase sampling); they supplement it by designing an optimum sample size using a cost function (*A General Class Of Estimators Under Multi Phase Sampling*). P. Maiti presents a simple estimation procedure for the purpose of estimating response variances - interviewer variance and supervisor variance **(***Estimation of Nonsampling Variance Components under the linear model Approach*). A. Ibrahim Al-Omari and others propose a ranked set sampling method for estimating the population median, and compare its efficiency with simple random sampling (*Multistage Balanced Groups Ranked Set Samples For Estimating The Population Median*). G. Prakash and D. C. Singh propose classes of estimators and the shrinkage testimators, and discuss their efficiencies in terms of relative bias and risk under different conditions using exponential data (*Double stage shrinkage testimation in exponential type – II censored data*). N. Salvati and M. Pratesi address the issue of large sampling variability in small area estimates and seek to improve them through using semi-parametric approach to small area estimation, based on M-quantile models, with intention to account for spatial correlation between small areas (*Spatial M-quantile Models for Small Area Estimation***) .** G.C. Tikkiwal and P. K. Rai also search for better (composite) estimates based on small area without requiring estimation of optimum weights and relying on the sensible interval of involved weights (*Composite Estimators for Small Domains and its Sensitivity Interval for Weights* α)

Each of the three other papers that constitute the second part of this issue (*Other topics*) illustrates a different type of statistical approaches to empirical phenomena. Canal and Ostasiewicz discuss four measurement and modeling methods (item response models, factorial models, latent classification, and paired comparison) to analyzing the social stress (*Statistical Models for Social Stress Analysis*). Zhanjun Xing examines the psychometric properties and expandability to the nation-wide scope of a scale of subjective well-being that was originally proposed ('normalized') for one province only (*Development of the Revised Well-being Scale for Chinese Citizens*). N. Farmakis proposes a new method - along with an algorithm and its illustrating application - for searching and mining latent

periodicities from labeled data (time series, DNA sequences, etc.) used by biologists, environmentalists, financial and management researchers, etc. (*Searching for Periodicities In Data Series*).

Report section – with a note by C. Domański on the occasion of the hundredths birthday of one of the most influential Polish and American mathematician (in the middle of XX century), Stanisław Ulam – concludes this issue.

Włodzimierz Okrasa
Editor-in-Chief

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 181

SUBMISSION INFORMATION FOR AUTHORS

Statistics in Transition – new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and the Central Statistical Office of Poland, on a quarterly basis (during 1993–2006 it was issued twice and since 2006 three times a year). Also, it has extended its scope of interest beyond its originally primary focus on statistical issues pertinent to transition from centrally planned to a market-oriented economy through embracing questions related to systemic transformations of and within the national statistical systems, world-wide.

The SiT-*ns* seeks contributors that address the full range of problems involved in data production, data dissemination and utilization, providing international community of statisticians and users – including researchers, teachers, policy makers and the general public – with a platform for exchange of ideas and for sharing best practices in all areas of the development of statistics.

Accordingly, articles dealing with any topics of statistics and its advancement – as either a scientific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series*.

Demonstration of the role played by statistical research and data in economic growth and social progress (both locally and globally), including better-informed decisions and greater participation of citizens, are of particular interest.

Each paper submitted by prospective authors are peer reviewed by internationally recognized experts, who are guided in their decisions about the publication by criteria of originality and overall quality, including its content and form, and of potential interest to readers (esp. professionals).

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl., followed by a hard copy addressed to
Prof. 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/5840_2638_ENG_HTML.htm

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 183—192

A GENERAL CLASS OF ESTIMATORS UNDER MULTI PHASE SAMPLING

M. S. Ahmed[1] & Atsu S. S. Dorvlo[2]

ABSTRACT

This paper derives the general estimators for finite population mean using multivariate auxiliary information under multiphase sampling. Here a number of auxiliary variables are considered in each phase under general sampling design. The properties of these estimators are studied and the results are presented for simple random sampling without replacement (SRSWOR) scheme. Using a modified cost function the optimum sample sizes are also derived.

Key words and Phrases: Multivariate auxiliary information; Multiphase sampling; Chain estimators.

1. Introduction

Multi-phase sampling scheme is very useful for estimating finite population parameters when a number of auxiliary variables are available cheaply related to the survey variable. The estimation of population mean of a survey variable under the partial knowledge of the auxiliary means has been considered by Chand (1975), Kiregyera (1980, 1984), Mukerjee *et al*. (1987) and Srivastava *et al*. (1990). However, their results are confined to the use of two auxiliary variables only, the mean of one is known while the other is unknown. They considered the two-phase sampling scheme only for simple random sampling without replacement (SRSWOR) at both the phases. Ahmed (1995) and Tripathi & Ahmed (1995) extended these results considering more than two auxiliary variables in two-phase sampling. Some recent works related to multiphase sampling are Ahmed *et al*. (1995&96), Ahmed (2003), and Diana *et al*.2004).

Suppose that $S_0 = (S_{01}, S_{02}, ..., S_{0n_0})$ is a finite population of size n_0 (given) and Y denotes the study variable with population mean $\overline{Y}_{(0)}$ ($\overline{Y}_{(0)} = \sum_{S_0} Y_i / n_0$).

[1] M S Ahmed, email: msahmed@squ.edu.om (Corresponding author).
[2] Atsu S. S. Dorvlo, email: atsu@squ.edu.om, Department of Mathematics and Statistics, Sultan Qaboos University, P.O. Box 36, Al-Khoudh PC 123, Muscat, Sultanate of Oman.

Suppose that $\underline{X}_i^{(r)} = (X_i^{(0)'}, X_i^{(1)'}, X_i^{(2)'}, ..., X_i^{(r)'})'$, where $X_i^{(r)} = (X_{1i}, X_{2i} ..., X_{k_r i})'$ are k-auxiliary variables ($k = \sum_{r=0}^{m} k_r$) and they are available from r-th phase source (r=1,2,...,m) with moderate cost to estimate the population mean $\overline{Y}_{(0)}$ of the study variable Y.

Suppose that X_{jr} is available for all $j \in S_{r-1}$, where S_r is a sub-sample drawn from S_{r-1} under the sampling design D_r (r = 1,2,...,m). The study variable Y_i is observed for all $i \in S_m$, a comparatively small sample with moderate cost.

Suppose that $\overline{Y}_{(m)}$ is an unbiased estimate of the population mean $\overline{Y}_{(0)}$ under the sampling design D_m at the m-th phase. Further, $\overline{X}_{j(r)}$ is an unbiased estimate of the population mean $\overline{X}_{j(0)}$ ($\overline{X}_{j(0)} = \sum_{S_0} X_{ji} / n_0$ under the sampling design D_r (r = 1,2,...,m) at the r-th phase.

The layout of auxiliary variables and study variable for generalized multiphase sampling are given as follows:

Source		Size	Auxiliary Variables	No. of variables
Population	S_0	n_0	$(X_{01}, X_{02}, \ldots, X_{0k_0})' = \underline{X}_0$	k_0
1st phase	S_1	n_1	$(X_{11}, X_{12}, \ldots, X_{1k_1})' = \underline{X}_1$	k_1
2nd phase	S_2	n_2	$(X_{21}, X_{22}, \ldots, X_{2k_2})' = \underline{X}_2$	k_2
…	…	…	…	…
m-1th phase	S_{m-1}	n_{m-1}	$(X_{m-11}, X_{m-12}, \ldots, X_{m-1k_m})' = \underline{X}_{m-1}$	k_{m-1}

$n_i \leq n_{i-1}$

Total auxiliary variables gather	Size	No. of variables
$\underline{X}_0 = \underline{X}^{(0)}$	n_0	k_0
$(\underline{X}_1, \underline{X}_0)' = \underline{X}^{(1)}$	n_1	$k_0 + k_1$
$(\underline{X}_0, \underline{X}_1, \underline{X}_2)' = \underline{X}^{(2)}$	n_2	$k_0 + k_1 + k_2$
	…	
$(\underline{X}_0, \underline{X}_1, \underline{X}_2, \ldots, \underline{X}_i)' = \underline{X}^{(i)}$	n_i	$k_0 + k_1 + k_2 + \ldots + k_i$
…		…
$(\underline{X}_0, \underline{X}_1, \underline{X}_2, \ldots, \underline{X}_i, \ldots, \underline{X}_{m-1})' = \underline{X}^{(m-1)}$	n_{m-1}	$k_0 + k_1 + k_2 + \ldots + k_{m-1}$
$(\underline{X}_0, \underline{X}_1, \underline{X}_2, \ldots, \underline{X}_i, \ldots, \underline{X}_m)' = \underline{X}^{(m)}$	n_m	$k_0 + k_1 + k_2 + \ldots + k_m$

Some particular cases of generalized multiphase sampling are given as follows:

k_0=1 and $k_1 = k_2 = ... = k_{m-1} = 0$ Classical Ratio and Regression estimators	Cochran (1977)
k_0>1 and $k_1 = k_2 = ... = k_{m-1} = 0$ Multiple Ratio and Regression estimators	Cochran (1977)
k_1=1 and $k_0 = k_2 = k_3 = ... = k_{m-1} = 0$ Classical Two-phase or Double Sampling Ratio and Regression estimators	Cochran (1977)
k_1>1 and $k_0 = k_2 = k_3 = ... = k_{m-1} = 0$ Multivariate Two-phase or Double Sampling Ratio and Regression estimators	Cochran (1977)
$k_0 = k_1$=1 and $k_2 = k_3 = ... = k_{m-1} = 0$ Two-phase or Double Sampling Ratio and Regression estimators with two auxiliary variables	Chand (1975), Kiregyera (1980,1984), Srivastava *et al.* (1990), Mukerjee *et al.* (1987), Singh *et al.* (1994), Ahmed and Ali (1996), Ahmed (1997), Ahmed *et al.* (1998).
$k_0 > 1$, k_1>1 and $k_2 = k_3 = ... = k_{m-1} = 0$ Two-phase or Double Sampling Ratio and Regression estimators with more auxiliary variables	Ahmed *et al.* (1994) and Tripathi and Ahmed (1995).
$k_1 = k_2 = ... = k_{m-1} = 1$ Multiphase Ratio and Regression estimators with more auxiliary variables	Diana *et al.* (2004)
$k_1 \geq 1, k_2 \geq 1 ... = k_{m-1} \geq 1$ Multiphase Ratio and Regression estimators with more auxiliary variables	Ahmed *et al*. (1995&1996), Ahmed (2003)

2. The proposed estimators and their properties

For estimating $\overline{Y}_{(0)}$ we may consider the estimation procedure,

$$G_m = g\left(\overline{Y}_{(m)}, \underline{\overline{X}}^{(0)}_{(0)}, \underline{\overline{X}}^{(0)}_{(1)}, \underline{\overline{X}}^{(1)}_{(1)}, \underline{\overline{X}}^{(1)}_{(2)}, ..., \underline{\overline{X}}^{(m-1)}_{(m-1)} \underline{\overline{X}}^{(m-1)}_{(m)}\right)$$

$$G_m = g\left(\underbrace{\underbrace{\overline{Y}_{(m)}, \overbrace{\overline{X}_{(0)}^{(0)}, \overline{X}_{(1)}^{(0)}}^{k_0}}_{1st}, \overbrace{\overline{X}_{(1)}^{(1)}, \overline{X}_{(2)}^{(1)}}^{k_0+k_1}}_{2nd}, \ldots, \overbrace{\overline{X}_{(m-1)}^{(m-1)} \overline{X}_{(m)}^{(m-1)}}^{k_0+k_1+\ldots+k_{m-1}}}_{m-th}\right) \quad m>0 \qquad (2.1)$$

$$H_m = \overline{Y}_{(m)} h\left(\overline{X}_{(0)}^{(0)}, \overline{X}_{(1)}^{(0)}, \overline{X}_{(1)}^{(1)}, \overline{X}_{(2)}^{(1)}, \ldots, \overline{X}_{(m-1)}^{(m-1)} \overline{X}_{(m)}^{(m-1)}\right) \qquad (2.2)$$

Using linearization, Ahmed *et al*. (1995&1996) showed that

$$G_m \approx H_m \approx \overline{Y}_{(m)} + \sum_{r=1}^{m} (\underline{\overline{X}}_{(r-1)}^{(r-1)} - \underline{\overline{X}}_{(r)}^{(r-1)})' \underline{\boldsymbol{T}}^{(r)} \qquad (2.3)$$

Where $\underline{\boldsymbol{T}}^{(r)} = \left(\underline{T}^{(0)'}, \underline{T}^{(1)'}, \ldots, \underline{T}^{(r)'}\right)'$ and $\underline{T}^{(r)} = \left(T_1^{(r)}, T_2^{(r)}, \ldots, T_{k_0}^{(r)}\right)'$

Two special cases of H_m as Ahmed (2003) defined

$$H_m^{(1)} = \overline{Y}_{(m)} \cdot \prod_{r=1}^{m} \prod_{j=1}^{k_r} \left(\frac{\overline{X}_{j(r-1)}}{\overline{X}_{j(r)}}\right)^{a_{j(r)}} ; \overline{X}_{j(0)} = \overline{X}_j \qquad (2.4)$$

Another estimation procedure based on fixed weights defined as

$$H_m^{(2)} = \overline{Y}_{(m)} \cdot \prod_{r=1}^{m} \prod_{j=1}^{k_r} u_{j(r)} \left(\frac{\overline{X}_{j(r-1)}}{\overline{X}_{j(r)}}\right)^{a_{j(r)}} ; \overline{X}_{j(0)} = \overline{X}_j \qquad (2.5)$$

where $a_{j(r)}$ are suitably chosen constants and $u_{j(r)}$ are fixed weights.

We assume that $\overline{X}_{j(r)}$ is defined such as to be conditionally unbiased for $\overline{X}_{j(r)}$ i.e. the conditional expectation given S_{r-1}

$$E_r\left(\overline{X}_{j(r)}\right) = \overline{X}_{j(r-1)} \quad \text{for all } j=1,2,\ldots,k_r \text{ and } r=1,2,\ldots,m \qquad (2.6)$$

Further we have the conditional covariance given S_r

$$C_r\left(\overline{X}_{j(r)} \overline{X}_{j(r-1)}\right) = 0 \text{ and } C_r\left(\overline{X}_{j(r)} \overline{Y}_{(r-1)}\right) = 0 \qquad (2.7)$$

where E_r and C_r stand for conditional expectation and covariance respectively at r-th phase given S_{r-1}, r = 1,2, ... ,m.

It is always possible to define unbiased estimates $\overline{X}_{j(r)}$ and $\overline{Y}_{(m)}$ provided each $S_{0i} \in S_0$, $S_0 = (S_{01}, S_{02}, \ldots, S_{0n_0})$ has possible probability of selecting for

fixed size sampling scheme. If $\pi_1^{(r)}$, $\pi_2^{(r)}$, ... , $\pi_{n_0}^{(r)}$; $\pi_1^{(r)} > 0$ for all $S_{0i} \in S_0$, ($\sum_{S_0} \pi_i^{(r)} = n_r$) are the inclusion probabilities for S_r, where $\pi_i^{(r)} = \pi_i^{(r-1)} . \pi_{i|S_{(r-1)}}^{(r-1)}$. Here $\pi_{i|S_{(r-1)}}^{(r-1)}$ is the conditional inclusion probability of i-th unit in S_r given S_{r-1} then

$$\bar{X}_{j(r)} = \frac{1}{n_r} \sum_{S_r} \frac{X_{j(r)}}{\pi_i^{(r)}} \text{ and } \bar{Y}_{(r)} = \frac{1}{n_r} \sum_{S_r} \frac{Y_i}{\pi_i^{(r)}} \tag{2.8}$$

Define,

$R_j = \bar{Y}_{(0)} / \bar{X}_{j(0)}$, $m_{jj'}^{(r)} = E_1 E_2 ... E_{r-1} C_r \left(\bar{X}_{j(r)}, \bar{X}_{j'(r)} \right)$

and $\alpha_j^{(r)} = E_1 E_2 .... E_{r-1} C_r \left(\bar{X}_{j(r)}, \bar{Y}_{(r)} \right)$.

Define,

$a^{(r)} = \left(a_1^{(r)}, a_2^{(r)}, ..., a_{k_r}^{(r)} \right)'$, $a_-^{(r)} = diag\left(a_1^{(r)}, a_2^{(r)}, ..., a_{k_r}^{(r)} \right)$,

$b^{(r)} = \left(b_1^{(r)}, b_2^{(r)}, ..., b_{k_r}^{(r)} \right)'$, $b_-^{(r)} = diag\left(b_1^{(r)}, b_2^{(r)}, ..., b_{k_r}^{(r)} \right)$,

$u^{(r)} = \left(u_1^{(r)}, u_2^{(r)}, ..., u_{k_r}^{(r)} \right)'$, $u_-^{(r)} = diag\left(u_1^{(r)}, u_2^{(r)}, ..., u_{k_r}^{(r)} \right)$,

$R_-^{(r)} = diag\left(R_1^{(r)}, R_2^{(r)}, ..., R_{k_r}^{(r)} \right)$, $\alpha^{(r)} = \left(\alpha_1^{(r)}, \alpha_2^{(r)}, ..., \alpha_{k_r}^{(r)} \right)'$, $m^{(r)} = \left(m_{jj'}^{(r)} \right)$

and $m_{(r)}^{-1} = \left(m_{(r)}^{jj'} \right)$; $\forall j, j' = 1,2,..., k_r$.

For large samples, the contribution of the third and higher order products and central moments can be ignored. Taking the expectation both sides and after some simplification, we have stated the following theorems.

Theorem 2.1: For large samples, the biases and mean square errors of G_1, H_1 and H_2 for estimating the population mean of the study variable Y are given respectively by,

$$M\left(G_1\right) = V\left(\bar{Y}_{(m)}\right) + \sum_{r=1}^{m} \underline{\boldsymbol{T}}^{(r)'} \left(m^{(r)} \underline{\boldsymbol{T}}^{(r)} - 2\alpha^{(r)} \right) \tag{2.9}$$

$$B\left(H_1\right) = \frac{1}{2\bar{Y}} \sum_{r=1}^{m} \left(a_-^{(r)} R_-^{(r)} m^{(r)} R_-^{(r)} a_-^{(r)} + trace\ a_-^{(r)} R_-^{(r)} m^{(r)} R_-^{(r)} - 2 a_-^{(r)} R_-^{(r)} a_-^{(r)} \right) \tag{2.10}$$

$$M\left(H_1\right) = V\left(\bar{Y}_{(m)}\right) + \sum_{r=1}^{m} a_-^{(r)} R_-^{(r)} \left(m^{(r)} R_-^{(r)} a_-^{(r)} - 2\alpha^{(r)} \right) \tag{2.11}$$

$$B(H_2)=\frac{1}{2\bar{Y}}\sum_{r=1}^{m}\Big(b_{-}^{(r)}R_{-}^{(r)}m^{(r)}R_{-}^{(r)}b_{-}^{(r)}u_{-}^{(r)}+trace\,b_{-}^{(r)}R_{-}^{(r)}m^{(r)}R_{-}^{(r)}u_{-}^{(r)}$$

$$-2b_{-}^{(r)}u_{-}^{(r)}R_{-}^{(r)}\alpha^{(r)}) \qquad (2.12)$$

$$M(H_2)=V(\bar{Y}_{(m)})+\sum_{r=1}^{m}b_{-}^{(r)}u_{-}^{(r)}R_{-}^{(r)}\left(m^{(r)}R_{-}^{(r)}u_{-}^{(r)}b_{-}^{(r)}-2\alpha^{(r)}\right) \qquad (2.13)$$

Theorem 2.2: The optimum choices of $\underline{\boldsymbol{T}}^{(r)}$, $a^{(r)}$ and $b^{(r)}$ which minimizes $M(G_1)$, $M(H_1)$ and $M(H_2)$ are given respectively by $\underline{\boldsymbol{T}}^{(r)}=m^{(r)^{-1}}\alpha^{(r)}$, $a_0^{(r)}=R_{-}^{(r)^{-1}}m^{(r)^{-1}}\alpha^{(r)}$ and $b_0^{(r)}=R_{-}^{(r)^{-1}}m^{(r)^{-1}}u_{-}^{(r)^{-1}}\alpha^{(r)}$ for all r and for large samples, the minimum mean square errors of G_1, H_1 and H_2 are same and given by

$$M_0=V(\bar{Y}_{(m)})-\sum_{r=1}^{m}\alpha^{(r)'}m^{(r)^{-1}}\alpha^{(r)} \qquad (2.14)$$

3. Results under SRSWOR scheme

Now, we will give the results when the SRSWOR scheme is adopted for selecting S_r for all $r = 1,2,...,m$.

Define,

$$\bar{X}_{j(r)}=\frac{1}{n_r}\sum_{S_r}X_{ji},\bar{Y}_{(r)}=\frac{1}{n_r}\sum_{S_r}Y_i,\ \sigma_{jj'}=\frac{1}{n_0}\sum_{U}\left(X_{ji}-\bar{X}_j\right)\left(X_{j'i}-\bar{X}_{j'}\right)$$

$$\sigma_{jy}=\frac{1}{n_0}\sum_{U}\left(X_{ji}-\bar{X}_j\right)\left(Y_i-\bar{Y}\right) \text{ and } \sigma_y^2=\frac{1}{n_0}\sum_{U}\left(Y_i-\bar{Y}\right)^2$$

for all $j,j'=1,2,....,k_r$ and $r = 0, 1,2,....,m$.

Also define $\Lambda=\begin{pmatrix}\Lambda_{11} & \Lambda_{12} & ... & \Lambda_{1m}\\ \Lambda_{12}' & \Lambda_{22} & ... & \Lambda_{2m}\\ .. & .. & ... & ..\\ \Lambda_{1m}' & \Lambda_{2m}' & ... & \Lambda_{mm}\end{pmatrix}$, $\Lambda_{rr'}=(\sigma_{jj'})$,

$\phi=(\phi_{(1)}',\phi_{(2)}',....,\phi_{(r)}')'=(\sigma_{jy})$ and $\phi_{(r)}=(\sigma_{jy})$ for all $j,j'=1,2,....,k_r$ and $r = 1,2,....,m$.

If SRSWOR is used at all the phases then

$\alpha^{(r)}=n(r,r-1)\phi_{(r)}$, $m^{(r)}=n(r,r-1)\Lambda_{rr'}$ and

$$V\left(\overline{Y}_{(m)}\right) = n(r, r-1)\sigma_y^2 \tag{3.1}$$

where $n(a,b) = \left(\frac{n_0}{n_0 - 1}\right)\left(\frac{1}{n_a} - \frac{1}{n_b}\right)$

For SRSWOR, the biases and mean square errors of T_1 and T_2 are respectively, given by

$$B\left(H_1\right) = \frac{1}{2\overline{\overline{Y}}}\sum_{r=1}^{m} n(r, r-1)(a_{-}^{(r)} R_{-}^{(r)} \Lambda_{rr'} R_{-}^{(r)} a_{-}^{(r)} + Trace\ a_{-}^{(r)} R_{-}^{(r)} \Lambda_{rr'} R_{-}^{(r)}$$

$$-2a_{-}^{(r)} R_{-}^{(r)} a_{-}^{(r)}) \tag{3.2}$$

$$M\left(H_1\right) = n(m,0)\sigma_y^2 + \sum_{r=1}^{m} n(r, r-1) a_{-}^{(r)} R_{-}^{(r)} \left(\Lambda_{rr'} R_{-}^{(r)} a_{-}^{(r)} - 2\phi_{(r)}\right) \tag{3.3}$$

$$M\left(G_1\right) = n(m,0)\sigma_y^2 + \sum_{r=1}^{m} n(r, r-1) \underline{\boldsymbol{T}}^{(r)} \left(\Lambda_{rr'} \underline{\boldsymbol{T}}^{(r)} - 2\phi_{(r)}\right) \tag{3.3}$$

$$B\left(H_2\right) = \frac{1}{2\overline{\overline{Y}}}\sum_{r=1}^{m} n(r, r-1)(b_{-}^{(r)} R_{-}^{(r)} \Lambda_r R_{-}^{(r)} b_{-}^{(r)} u_{-}^{(r)} + Trace\ b_{-}^{(r)} R_{-}^{(r)} \Lambda_r R_{-}^{(r)} u_{-}^{(r)}$$

$$- 2b_{-}^{(r)} u_{-}^{(r)} R_{-}^{(r)} \phi_{(r)} \tag{3.4}$$

$$M\left(H_2\right) = n(m,0)\sigma_y^2 + \sum_{r=1}^{m} n(r, r-1) b_{-}^{(r)'} u_{-}^{(r)} R_{-}^{(r)} (\Lambda_{r'r} R_{-}^{(r)} u_{-}^{(r)} b^{(r)} - 2\phi_{(r)}) \tag{3.5}$$

For SRSWOR, the optimum choices of $\underline{\boldsymbol{T}}^{(r)}$, $a^{(r)}$ and $b^{(r)}$ which minimize $M(G_1)$, $M(H_1)$ and $M(H_2)$ are given respectively by

$\underline{\boldsymbol{T}}_{\boldsymbol{0}}^{(r)} = \Lambda_{r'r}^{-1}\phi_{(r)}$, $a_0^{(r)} = R_{-}^{(r)^{-1}} \Lambda_{r'r}^{-1} \phi_{(r)}$ and $b_0^{(r)} = R_{-}^{(r)^{-1}} \Lambda_{r'r}^{-1} u_{-}^{(r)^{-1}} \phi_{(r)}$ for all $r, r' = 1,2,...,m$

For large samples, the minimum mean square error of H_1 and H_2 are same and given by

$$M\left(H_1\right)_{\min} = \sigma_y^2 \left[n(m,0)\left(1 - \rho_{y.12....k_m}^2\right) + \sum_{r=1}^{m} n(r, r-1)\left(1 - \rho_{y.12....k_{r-1}}^2\right) \right] \tag{3.6}$$

where $\rho_{y.12....k_r}$ is multiple correlation coefficient.

4. Selection of optimum sample sizes for a fixed cost

For designing a sample survey efficiently, it is essential to have some broad information about the variability in the population and on the cost of different steps involved in carrying out the survey. A design-based measure of the sampling error in the results of the survey is given by the mean square error of the estimator used, which reduces to its variance in the case of an unbiased estimator. In sample surveys, the variance of the estimator usually decreases with increase in sample size. Further, variance and cost would also depend on the nature of the sampling unit. Hence, for the above sampling scheme it becomes necessary to take both these aspects into account in arriving at the optimum sampling unit and the optimum sizes n_r which would provide the maximum information per unit of cost.

We will choose the values of n_r such that the minimum mean squared error for fixed cost function.

Suppose C_r and $C_{(y)}$ is the per unit cost for $\sum k_j$ auxiliary variables X_j and the study variable Y respectively. Suppose C_h is the overhead cost and C_t is the total cost, then a cost function may be defined as,

$$C_t = C_h + \sum_{r=0}^{m-1} C_r n_r + C_{(y)} n_m \tag{4.1}$$

The optimum choices of n_r (r = 1,2,...,m) for which mean sum of squares of errors of T_2 and T_3 are minimum for the cost function are given by

$$n_r^0 = \frac{1}{g\sqrt{C_r}}\left(C_t - C_h\right)\sqrt{\left(\rho^2_{y.12\ldots k_m} - \rho^2_{y.12\ldots k_r}\right)}$$

$$\text{and } n_m^0 = \frac{1}{g\sqrt{C_m}}\left(C_t - C_h\right)\sqrt{\left(1-\rho^2_{y.12\ldots k_m}\right)} \tag{4.2}$$

The minimum mean square error is

$$M_0 = \frac{n_0\sigma_y^2}{n_0 - 1}\sum_{r=1}^{m}\left(\frac{g^2}{C_t - C_h} - \frac{1-\rho^2_{y.12,\ldots,k_1+k_2+\ldots+k_r}}{n_0}\right) \tag{4.3}$$

where $g = \sum_{r=0}^{m-1}\sqrt{\left(\rho^2_{y.12,\ldots,k_m} - \rho^2_{y.12,\ldots,k_r}\right)C_r} + \sqrt{\left(1-\rho^2_{y.12,\ldots,k_m}\right)C_m}$

REFERENCES

AHMED, M. S. AND ALI, M.A.1996). The general class of chain estimators for the product of two means using double sampling. *Journal of Statistical Studies,* 16, 65–68.

AHMED, M. S.1995). *Some estimation procedure using multivariate auxiliary information in sample surveys*. Unpublished Ph.D thesis, Department of Statistics & Operations Research, Aligarh Muslim University, Aligarh-202002, India.

AHMED, M. S.1997). The general class of chain estimators for the ratio of two means using double sampling. *Communication in Statistics-Theory and Methods,* 26(9), 2247–2254.

AHMED, M. S.1998). A note on regression type estimators using multiple auxiliary information. *Australian & New Zealand Journal Statistics* 40(3), 373–376.

AHMED, M. S.2003). General chain estimators under multi phase sampling. *Journal of Applied Statistical Science*, Vol. 12, No. 4, pp. 243–250.

AHMED, M. S., KHAN, S.U. AND TRIPATHI, T.P.1994). Two general class of chain ratio and product estimators for a finite population mean based on two-phase sampling and multivariate information. *Journal of Statistical Studies*, 14, 86–99.

AHMED, M. S., RAHMAN, M.S. AND AHMED, R.1998). The general class of chain estimators for a finite population mean using double sampling. *Journal of Applied Statistical Science,* 7(4), 185–190.

AHMED, M.S., KHAN, S.U. AND TRIPATHI, T.P. (1995&96). Model based regression estimators using multiphase sampling. *Aligarh Journal of Statistics* 15&16, 69–74.

CHAND, L. (1975). Some ratio-type estimators based on two or more auxiliary variables. Ph.D. thesis submitted to Iowa State University, Ames, Iowa.

COCHRAN, W.G. (1977). Sampling Techniques (*First edition 1953, Second edition 1963*). John Wiley and Sons, New York.

DIANA,G.,TOMMASI,C.,AND PREO,P.2004). Estimation for finite population mean under multi-phase sampling. Atti della XLII Riunione Scientifica SIS, Bari, 525–528.

KIREGYERA, B. (1980). A chain ratio-type estimator in finite population two phase sampling using two-auxiliary variables. *Metrika*, 27, 217–223.

KIREGYERA, B. (1984). Regression type estimators using two-auxiliary variables and the model of double sampling from finite populations". *Metrika*, 31, 215–226.

MUKERJEE, R., RAO, T. J. AND VIJAYAN, K. (1987). Regression-type estimators using multiple auxiliary information. *Aust. Jour. Stat*., 29, 3, 244–254.

SINGH,V.K.; SINGH, HARI P.; SINGH, HOUSILA P., AND SHUKLA, D. (1994). A general class of chain estimators for ratio and product of two means of a finite population. *Comm.Stat.-Theo. Math*., 23(5), 1341–1355.

SRIVASTAVA, S.RANI., KHARE, B. B., AND SRIVASTAVA, S. R. (1990). A generalized chain ratio estimator for mean of a finite population. *Jour. Ind. Sco. Ag. Stat*., 42, 1990, 108–117.

TRIPATHI, T.P. AND AHMED, M. S. (1995). A class of estimators for a finite population mean based on multivariate information and two-phase sampling". *Cal. Stat. Asso. Bull*., 45, 179-180, 203–218.

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 193—222

ESTIMATION OF NONSAMPLING VARIANCE COMPONENTS UNDER THE LINEAR MODEL APPROACH

Pulakesh Maiti[1]

ABSTRACT

The importance of nonsample or measurement errors has long been recognized. [for numerous references see e.g., the comprehensive papers by Mahalanobis (1946), Hansen et.al. (1961), Bailar and Dalenius (1970), Dalenius (1974)]. Attempts have been made for estimating components due to nonsampling errors. The work in this area starts developing **surveys, specifically_designed** to incorporate features which can facilitate the estimation of non sampling components such as **reinterviews and/or interpenetrating samples**. However most of the survey designs so far developed, though few, are very complex in nature [Fellegi (1964, 1974), Biemen et al. (1985), Folsom(1980), Nelson(1974)]. Here, a very simple survey design as well as a simple estimation procedure have been developed for the purpose of estimating **simple as well as correlated response variances**, namely **interviewer variance** and **superviser variance**.

Key words: Simple Response Variance, Correlated response variance, Measurement error.

1 Introduction

The various models developed for such errors have assumed that a survey record (a recorded content item) differs from **its true value** (Zarkovich 1966) by a systematic bias and various additive error contributions associated with various sources such as interviewers, supervisors, coders etc. These models indicate that the errors made by a specified error source (say a particular interviewer) are usually correlated. These correlated errors contribute to the additive components of the total mean square error of a survey estimate. As a result of these correlated components, the usual unbiased estimations of variances of estimators of total or mean appear to be negatively biased. The models also indicate that these biases can be eliminated or reduced, if estimates of correlated response variance are

[1] Indian Statistical Institute, Kolkata.

available. This necessitates the estimation of the components due to correlated variances.

1.1 Survey Measurements

We start with a set $U = \{U_1, U_2, \ldots, U_N\}$ of N objects and a set $\{Y\}$ of real numbers corresponding to the objects. Each object is assigned one and only one number and two objects may be assigned the same number [(Dalenius, (1974)].

Some of the essential conditions for having a **measurement** may be identified as follows:

(a) the value of the characteristic to be measured should be precisely defined for every population unit in a manner consonant with the users to which the data are to be put;
(b) for any given population unit, this value known as the **true value** should be unique and should exist;
(c) there should exist procedures for obtaining information on the true value and although this procedure may be costly and very difficult to use (Sukhatme and Sukhatme (1970);

1.2 Measurement Error

The approaches to define measurement error in surveys vary according to a particular researcher's view on **true values**. One approach considers the true value to exist on the survey condition, while the other takes a strict operational approach in relation to the survey condition [Hansen et al. (1951)].

Under the assumption that it is meaningful to talk about a true value Y_i of the study variable for the i^{th} population unit, then measurement error is defined as

$$(Y_i - y_i), \tag{1.1}$$

where y_i is observed measurement for the i^{th} individual using a specific measurement technique.

1.3 Three Distinct views on the Nature of Measurement Variability

Three different views in existence may be described as follows:

(a) Measurements are random variables having a mean and a finite variance [Hansen (1951); Raj (1956), (1968)]; (1.2)
(b) Measurements as random variables are generated by a conceptual sequence of repeated independent trials of a generating process [Hansen (1951)]; (1.3)

(c) A third point of view does not allow the variability at the elementary level, but assumes that the variability results from **interviewers** and **subsequent handling of data**; (1.4)

1.4 Nonsampling Bias and Nonsampling Variance

Nonsampling bias is a measure of the difference between the expected value of their repeated observations and the corresponding true value.

Nonsampling variance measures the variation of the observed values for fixed samples in **hypothetical repetitions** of the survey process, if it is agreed upon that the survey is conceptually repeatable under identical conditions. More precisely, it is assumed that a measurement derived has a well defined, though quite likely unknown, probability distribution.

Nonsampling variance has two components, namely (a) simple or uncorrelated response variance and correlated response variance.

a) **Simple response variance**: Uncorrelated responses are those that are not affected by the particular interviewer or superviser or coder or any other survey personnel who happen to be associated with a particular element of the sample.
b) **Correlated response variance**: In so far as individual interviewers have different average effects on their work loads, they introduce response errors which are correlated for all elements of the assignment included in the work load of the investigator. The correlated errors thus arising give rise to correlated response variance. The correlated response variance may be categorized as
   - interviewer variance;
   - superviser variance;
   - coder variance etc.

For different kinds of correlation of response deviation, one can refer to the paper by Fellegi (1964).

Another way of looking at correlated response variance, say, interviewer variance is that it results from bias effects that differ from one interviewer to the other.

2 General Measurement Model

A measurement model wants to specify the joint probability distribution of the measurement y_i for the i^{th} unit conditional on a samples "s" in sample surveys or for every unit i of the population U in case, a census or complete enumeration is conducted.

From a frequentist's point of view, given a particular sample selection procedure leading to a simple "s" and a specific measurement technique adopted,

the process generates an observed value for every $i \in s$, and given independent observations many times on the same sample "s", a long series of data $\left\{y_i^t; t = 1,2,\ldots\right\}$ for each $i \in s$ is generated. The observed value of a specified element $i \in s$ would vary in a random fashion around a long term mean value μ_i and long term variance σ_i^2. These moments may or may not depend on the sample. The same thing applies to every $i \in U$, when complete enumeration is conducted.

2.1 Some Specific Error Models

2.1.1 Based on the views expressed in (1.1) and (1.2)

Non sampling error models are essential for understanding the effects of measurement errors on statistics and statistical inference. All such models developed assume that observed value differs from the true value by a systematic bias and additive error terms.

The basic model developed in the U.S. Bureau of census was first introduced by Hansen et al. (1951). The basic model assumes conceptually repeated trials and possesses the views on the nature of measurement variability expressed in (1.2) and (1.3). The measurement for the i^{th} unit at t^{th} trial, y_{it} was thus modeled a

$$y_{it} = Y_i + \beta_i + e_{it} \tag{2.1}$$

where β_i is a systematic bias and e_{it} is the variable error. Under repeat measurement for the same unit i, e_{it} is taken as a mean zero random error. Subsequent elaborations of the basic model was made by Hansen, Hurwitz, Bershad (1961), Hansen, Hurwitz and Pritzkar (1964).

2.1.2 Based on the views expressed in (1.3)

Let y_{ij} be an observation from a randomly selected population unit i which is thought of as the sum of two components, the true value Y_i and an error d_j; the error d_j may be attributed to the measurement processes (including the interviewer, questionnaire, the interviewer setting and so on).

In its most general form, the structure of the error d_j provides for essential correlations amongst the different measurement errors due to interviewers, supervisers, coders etc., or due to any other survey operators.

The observation collected from the i^{th} respondent by the j^{th} investigator i.e., y_{ij} may be modeled as

$$y_{ij} = Y_i + d_{ij}$$

$$y_{ij} = Y_i + \beta_j + e_{ij} \tag{2.2}$$

Y_i being the true value, β_j being the j^{th} operator bias and $e_{ij's}$ are elementary errors. β_j may be **fixed or random**. For random effects, $\beta_j's$ constitute a random sample from an infinite population of operator effects having mean μ_b and variance σ_b^2. $e_{ij's}$ are random variables with mean 0, variance σ_e^2. The following covariance structure for d_{ij} and $d_{i'j'}$ may be mentioned as follows.

$$Cov\left(d_{ij}, d_{i'j'}\right) = \begin{cases} \sigma_b^2, \text{ for } j = j', i \neq i' \\ 0, \text{for } j \neq j', i \neq i \\ 0, \text{for } j \neq j', i = i' \\ \sigma_b^2 + \sigma_e^2 \text{ for } j = j', i = i' \end{cases} \tag{2.3}$$

Under the assumption of **fixed operator effects**, the covariance structure is modified by Letting $\sigma_b^2 = 0$. Another special case is the case of no-operator effects i.e., $\beta_j = 0$ for all j. This model is referred to as **uncorrelated model**.

2.2 Measurement Models taking care of a specific measurement Technique

2.2.1 Personal Interview Method

Measurements have been described as being realized under a model which specifies the joint distribution of $y_i's$ .The model is specified in terms of its moments namely, μ_i, σ_i^2 , σ_{ij} .

Introduction of the model did not require any specific measurement procedure. We now consider situations when data are collected by interviewers. They may introduce bias, variance and correlations in to the measurements (being reflected through σ_i^2, σ_{ij}). Such interviewer effects have been detected in many empirical studies.

Comprehensive and elaborate discussions on different kinds of interviewer settings and associated models taking interviewers effects into account are available in any standard text book.

3 Mean Square Error in the Presence of Combined Effects of Total Error

Decomposition and Linear model

The general models for variability in the literature are expressed either as **Mean Square Error Decomposition Model or Mixed Linear Models**. The net bias is assumed to be zero, so that the model deals only with variability. The major difference between the two models is that decomposition approach often has a component attributable to the interaction between sampling and measurement error, whereas the linear model approach omits this component. The linear model approach defines response variability about the true value. However, both the approaches merge on a specific occasion.

The Situation, When Both Variance Decomposition and Linear Model Approach Merge

The variance decomposition approach and the linear model formulation tend to merge, when the variance decomposition approach focuses on a particular source of error-mostly the error due to the interviewer. If, in the variance decomposition model with the interviewer setting, the interviewers influence in the response deviation independently, then the **additive model would be appropriate**. The additive model would not be applicable, if, in the hypothetical repition of the (original or repeat) survey, such factors as common training or supervision etc., have a correlating effect on the response deviations obtained by different enumerators/interviewers (Fellegi 1964). It has been observed by others also.

It may be mentioned that the model in (2.1) and that in (2.2) can be thought of as variance decomposion approach and linear model formulation respectively.

3.1 NonSampling Variances under: Mean Square Error Decomposition Model

Hansen, Hurwitz and Bershad Model (1961) decomposes the variance into three components namely, **(a) sampling variance, (b) measurement variance and (c) covariance between response and sampling deviation**. Their model has implications for the total survey design in so far it relates to the accuracy of survey results [Jabine and Tepping (1973)]. In our discussion we are mainly presenting the expressions for nonsampling variances.

Let $\hat{t}_\pi = \sum_{i \in S} y_i \Big/ \pi_i$ be an unbiased estimator for *Y*, the population total, then it may be shown that,

Sampling

$$\text{Variance(SV)} = \sum_{i \in U} \frac{(1-\pi_i)}{\pi_i} \mu_i^2 + \sum_{i \neq j} \sum_{\in U} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} \mu_i \mu_j \qquad (3.1A)$$

Measurement

$$\text{Variance (MV)} = \sum_{i,j} \sum_{\in U} \frac{\sigma_{ij}(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} + \sum_i \sum_j \sigma_{ij} \qquad (3.1B)$$

where,

$\mu_i, \sigma_i^2, \sigma_{ij}$ are model parameters and, (3.2)

π_i, π_{ij} the inclusion probabilities are sampling design parameters.

We have,

simple response variance = $\sum_{i \in U} \sigma_i^2 \Big/ \pi_i$ and (3.3A)

correlated response variance = $\sum_{i \neq j} \sum_{\in U} \frac{\sigma_{ij}\, \pi_{ij}}{\pi_i \pi_j}$ (3.3B)

However, correlated response variance can alternatively be expressed as

$$\sum \sum \frac{\sigma_{ij}(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} + \sum \sum \sigma_{ij} \qquad (3.4)$$

3.2 Remarks

(1) Contrary to what generally accepted name suggests, the measurement variance depends both on measurement model and the sampling design $\left(\sigma_i^2, \sigma_j, \pi_i, \pi_{ij}\right)$. It would be therefore interesting to isolate a component that is unaffected by sampling i.e., **a term that would remain unaffected, even if the sampling were pushed to the ultimate limit of complete enumeration.**

(2) Therefore, under complete enumeration,

$$MV = \sum_{i \in U} \sigma_i^2 \quad + \quad \sum_{i \neq j \in U} \sum \sigma_{ij} \qquad (3.5)$$

3.2 Measurement Variances under Variance Decomposition Approach with different Measurement Models and Different Measurement Techniques

Here we present only the expressions for measurement variances under the following different situations.

3.2.1 Measurement variances without taking care of interviewer's effect:

a) Sampling Design: $SRS(N,n)$;
b) For model specification, we refer to the model specified by equation (2.1) of the basic model introduced by Hansen et al. (1951);

Let $\bar{y}_t$ be an estimator for population mean $\bar{Y}$. From equation (3.1) or otherwise also, it follows that,

$$MV(\bar{y}_t) = \frac{1}{n}\left[\sigma_d^2 + (n-1)\zeta\sigma_d^2\right], \text{ where,}$$

$$\begin{aligned} \sigma_d^2 &= E_t\left(y_{it} - E_t(y_{it})\right)^2 \\ \zeta\sigma_d^2 &= E_t\left[\left(y_{it} - E_t(y_{it})\right)\left(y_{i't} - E_t(y_{i't})\right)\right] \end{aligned} \tag{3.6}$$

This was originally derived by Hansen et al and also later by Bailar and Dalenius (1969).

3.2.2 Measurement variances with taking care of interviewer effect:

a) General Sampling Design: $[\pi:(N,n)]$;
b) For **Survey Design** with interviewer settings of deterministic as well as random assignments. We consider the following situation.

there is a fixed set of *J* interviewers labelled $j = 1,2,\ldots,J$, and prior to the survey, the population is partitioned in to *j* responding groups $U_1, U_2, \ldots, U_J$, so that each interviewer *J* is linked a unique or a number of groups according to the specified survey design (3.7)

For the estimator $\hat{t}_\pi$ of the population total, we have from equation (3.1A,B),

$$MV(\hat{t}_\pi) = \begin{cases} \sum_{j=1}^{J}\left(\sum_{i\in Uj} 1/\pi_i\right)v_j + \sum_{j=1}^{J}\left(\sum_{i\neq j\in Uj}\sum \pi_{ik}/\pi_i\pi_k\right)e_j v_j; & (3.8) \\ (v_\beta + v_e)\sum_{i\in U} 1/\pi_i + v_\beta \sum_{j=1}^{J}\left(\sum_{i\neq j\in Uj}\sum \pi_{ik}/\pi_i\pi_k\right), & (3.9) \end{cases}$$

where, (3.8) and (3.9) refer to deterministic and random assignments respectively and $\upsilon_j, \upsilon_\beta, \upsilon_e$ are model parameters.

3.3 Remark

Under complete enumeration, the expressions in (3.8), and (3.9) take the respective forms as

$$\sum_{j=1}^{J} N_j \nu_j + \sum_{j=1}^{J} N_j (N_j - 1)\ e_j \nu_j \qquad (3.10)$$

$$\nu_\beta \sum_{j=1}^{J} N_j^2 + N \nu_e ;$$

and N_j is the number of responding units in the j^{th} group U_j

3.3 A General Expression For the Measurement Variance of an Estimator taking Interviewer and Superviser Effects.

Let $y_{ijk}^{(t)}$ be a measurement made by the j^{th} investigator on the i^{th} respondent under the supervision of k^{th} superviser at the t^{th} trial $(i = 1,2,\ldots,I; j = 1,2,\ldots J \text{ and } k = 1,2,\ldots K)$;

Let us define the following indicator variables.

$$u_i = \begin{cases} 1, \text{if } i^{th} \text{ element is included into the sample}; \\ 0, \text{otherwise}; \end{cases} \qquad (3.11)$$

$$v_j = \begin{cases} 1, \text{if } j^{th} \text{ interviewer is selected for the survey}; \\ 0, \text{otherwise}; \end{cases}$$

$$\delta_k = \begin{cases} 1, \text{if } j^{th} \text{ superviser is selected for supervising the job}; \\ 0, \text{otherwise}; \end{cases}$$

$$c_{ij} = \begin{cases} 1, \text{if the } i^{th} \text{ element is assigned to the } j^{th} \text{ interviewer}; \\ 0, \text{otherwise}; \end{cases}$$

$$\gamma_{k(i,j)} = \begin{cases} 1, \text{if the schedule filled up by the } j^{th} \text{ interviewer from the } i^{th} \text{ respondent} \\ \quad \text{is alloted to be } k^{th} \text{ supervisor for supervision;} \\ 0, \text{otherwise;} \end{cases}$$

Let $\bar{y}_t = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} u_i v_j c_{ij} \delta_k \gamma_{k(i,j)} / n$ be the estimator for $\bar{Y}$. (3.12)

Measurement variance, M.V. $(\bar{y}_t)$

$$= \left(\frac{1}{n}\right)^2 \sum_i \sum_j \sum_k E\left\{ u_i v_j c_{ij} \delta_k \gamma_{k(i,j)} E_t (y_{ijk}^{(t)} - y_{ijk})^2 \middle| u_i v_j c_{ij} \delta_k \gamma_{k(i,j)} = 1 \right\}$$

$$+ \sum_i \sum_{j \neq} \sum_{j'} \sum_k E\left\{ u_i v_j v_{j'} c_{ij} c_{ij'} \delta_k \gamma_{k(i,j)} \gamma_{k(i,j')} E_t \left(y_{ijk}^{(t)} - y_{ijk}\right)\left(y_{ij'k}^{(t)} - y_{ij'k}\right) \right|$$

$$\left| u_i v_j v_{j'} \delta_k c_{ij} c_{ij'} \gamma_{k(i,j)} \gamma_{k(i,j')} = 1 \right\}$$

$$+ \sum_i \sum_j \sum_{k \neq} \sum_{k'} E\left\{ u_i v_j c_{ij} \delta_k \delta_{k'} \gamma_{k(i,j)} \gamma_{k'(i,j)} E_t \left(y_{ijk}^{(t)} - y_{ijk}\right) \right.$$

$$\left. \left(y_{ijk'}^{(t)} - y_{ijk'}\right) \middle| u_i v_j \delta_k \delta_{k'} c_{ij} \gamma_{k(i,j)} \gamma_{k'(i,j)} = 1 \right\}$$

$$+ \sum_i \sum_{j \neq} \sum_{j'} \sum_{k \neq} \sum_{k'} E\left\{ u_i v_{j'} v_j c_{ij} c_{ij'} \delta_k \delta_{k'} \gamma_{k(i,j)} \gamma_{k'(i,j')} E_t (y_{ijk}^{(t)} - y_{ijk}) \right.$$

$$\left. y_{ij'k'}^{(t)} - y_{ij'k'}) \middle| u_i v_{j'} v_j c_{ij} c_{ij'} \delta_k \delta_{k'} \gamma_{k(i,j)} \gamma_{k'(i,j')} = 1 \right\}$$

$$+ \sum_{i \neq} \sum_{i'} \sum_j \sum_k E\left(u_i u_{i'} v_j c_{ij} c_{i'j} \delta_k \gamma_{k(i,j)} \gamma_{k(i',j)} \left\{ E_t \left(y_{ijk}^{(t)} - y_{ijk} \right)\left(y_{i'jk}^{(t)} - y_{i'jk} \right) \right| \right.$$

$$\left. u_i u_{i'} v_j c_{ij} c_{i'j} \delta_k \gamma_{k(i,j)} \gamma_{k(i',j)} = 1 \right\}$$

$$+\sum_{i\neq}\sum_{i'}\sum_{j\neq}\sum_{j'}\sum_{k}E\Bigg(u_iu_{i'}v_jv_{j'}c_{ij}c_{i'j'}\delta_k\gamma_{k(i,j)}\gamma_{k(i',j')}\Bigg\{E_t\left(y_{ijk}^{(t)}-y_{ijk}\right)\left(y_{i'j'k}^{(t)}-y_{i'j'k}\right)\Bigg|$$

$$u_iu_{i'}v_jv_{j'}c_{ij}c_{i'j'}\delta_k\gamma_{k(i,j)}\gamma_{k(i',j')}=1\Bigg\} \tag{3.13}$$

$$+\sum_{i\neq}\sum_{i'}\sum_{j}\sum_{k}\sum_{k''}E\Bigg(u_iu_{i'}v_jc_{ij}c_{i'j}\delta_k\delta_{k'}\ \gamma_{k(i,j)}\gamma_{k'(i',j)}\Bigg\{E_t\left(y_{ijk}^{(t)}-y_{ijk}\right)\left(y_{i'jk'}^{(t)}-y_{i'jk'}\right)\Bigg|$$

$$u_iu_{i'}v_jc_{ij}c_{i'j}\delta_k\delta_{k'}\ \gamma_{k(i,j)}\gamma_{k'(i',j)}=1\Bigg\}$$

$$+\sum_{i\neq}\sum_{i'}\sum_{j\neq}\sum_{j'}\sum_{k\neq}\sum_{k'}E\Bigg(u_iu_{i'}v_jv_{j'}c_{ij}c_{i'j'}\gamma_{k(i,j)}\delta_k\delta_{k'}\ \gamma_{k'(i',j')}$$

$$\Bigg\{E_t\left(y_{ijk}^{(t)}-y_{ijk}\right)\left(y_{i'j'k'}^{(t)}-y_{i'j'k'}\right)\Bigg|$$

$$u_iu_{i'}v_jv_{j'}\delta_k\delta_{k'}\ c_{ij}c_{i'j'}\gamma_{k(i,j)}\gamma_{k(i',j')}\gamma_{k'(i',j')},\gamma_{k'(i,j)}=1\Bigg\}$$

The above result is an extension of the similar type of result, obtained by Lessler, (1992).

3.4 Remark

1) Both under Census Bureau Model/Cochran Model and the model due to Raj (1968), simple as well as the correlated response/measurement variance of the sample mean takes the form of

$$SRV=\frac{1}{n}\frac{1}{JI}\sum_{j=1}^{J}\sum_{i=1}^{I}V_t(y_{ijt}) \tag{3.14}$$

$$\text{and}\quad CRV=\frac{1}{n}\frac{(m-1)}{JI(1-1)}\sum_{j=1}^{J}\sum_{i\neq i'}^{I}Cov(y_{ijt},y_{i'jt}) \tag{3.15}$$

where, m is the number of assignments for an investigator j and n is the sample size, y_{ijt} is the observation collected from the i^{th} respondent by

the j^{th} investigator at the t^{th} trial and $V_t(y_{ijt}), Cov(y_{ijt}, y_{i'jt})$ are measurement variance and measurement co-variances.

2) Under Census Bureau Model, it has been assumed that a fixed number of investigators is available. No sampling of investigators is made, i.e., $\upsilon_j = 1$ for all j in our above setting.
3) Here only the covariance between the observations of a particular investigator was considered, but not that between investigators.

3.4 Linear Model in the Context of Variance Decomposition Approach

$$\text{Let } Y_{ijt} = Y_i + b_j + e_{ij} \tag{3.16}$$

where, Y_i, b_j and e_{ij} are mutually independent and that the b_j and e_{ij} arises respectively from an infinite population of interviewer effects and an infinite population of random effects. The b_j are independently distributed with $E_t(b_j) = 0$ and $V(d_j) = \sigma_b^2$; similarly, e_{ij} are independently distributed with $E_t(e_{ij}) = 0$ and $V(e_{ij}) = \sigma_e^2$.

It may be shown that

$$MV(\bar{y}_t) = \frac{(\sigma_b^2 + \sigma_e^2)}{n}\left[1 + (m-1)\zeta\right], \tag{3.17}$$

m being the number of elements assigned to an investigator and $\zeta = \sigma_b^2 / \left[\sigma_b^2 + \sigma_e^2\right]$;

It may be noted that $\left(\sigma_b^2 + \sigma_e^2\right)$ is the simple response variance and σ_b^2 is the correlated response variance.

4 Existing estimators of NonSampling Variance components

4.1 Need of Randomisation of the assignments and of repeat measurements

If the survey arrangement is such that each investigator is assigned to work in only one sample cluster, then the effect of interviewers will be completely confounded with the effect of clustering on the sample variance, and usual methods of estimating sampling variance will automatically include the correlated interviewer variance. By contrast, if all interviewers work as a team in each cluster or if the interviewer's work loads are distributed at random, the usual

estimation of sampling variance would not include the effect of additional variability due to interviewers.

To make separate estimates of interviewer variance or other types of correlated variance, it is necessary to introduce some degree of randomization of interpenetration of work loads to the particular category of the survey personnel.

Repeat measurement technique has been advocated as a tool in measurement variance estimation. The measurement model using repeat measurement so far developed has one crucial assumption that all repeat measurements have to be uncorrelated with original measurements. There are several discussions on the implication of the assumption of lack of independence between the two surveys and change in the distribution of y_{it} [for example, see Hansen et al. (1964)].

4.2 Estimation of NonSampling Variance in the Variance Decomposition Approaches:

Different methods under this approach can be broadly categorized into one of the following categories.

a) Survey Design: Interpenetrated sample, but no repeat measurement: This method is primarily due to Mahalanobis (1946).

Let $\bar{y}_{jt} = \sum_{i \in S_j} y_{ijt} \Big/ n(s_j)$, $j = 1,2,\ldots,J$ and $n(s_j) = m$;

Let Between-interviewer mean Square(BIMS) be defined as

$$BIMS = \sum_{j=1}^{J} (\bar{y}_{jt} - \bar{y}_t)^2 /(J-1);\ \bar{y}_t = \sum_{j=1}^{J} \bar{y}_{jt} / J \qquad (4.1)$$

Then, (*BIMS/J*), as expected, is an estimate of **total variance** i.e.,
Thus, *E* (*BIMS/J*) = Sampling Variance + Measurement Variance.

b) **Survey Design: No interpenetration, but repeat measurements are available**.

(b.1) Repeat measurements of the entire sample:

Let, Mean Square within element (MSWE), Square Mean Deference (SMD) and Square of the difference Between Measures (BMWE) be defined as

$$MSWE = \frac{1}{2n} \sum_{m=1}^{2} \sum_{i=1}^{n} (y_{imt} - \bar{y}_{it})^2 \qquad (4.2)$$

$$SMD = \frac{1}{n}\left(\bar{y}_{1t} - \bar{y}_{2t}\right)^2 \tag{4.3}$$

$$BMWE = \frac{2}{n}\sum_{i=1}^{N}\sum_{j=1}^{J} c_{ij}\left(y_{ij1t} - y_{ij2t}\right)^2 \tag{4.4}$$

It may be noted that the estimators in (4.2), (4.3) and (4.4) estimates simple response variance only.

(b. 2) Survey Design: Repeat measurements of the sub-sample

Let an original samples of n_s be drawn from a population with a sampling design $p(\cdot)$ having the inclusion probabilities π_i, π_{ij}; From S, a sub-sample of size $n_r (< n_s)$ is drawn by SRSWOR. After two stages, we have the number of observations $n_s + n_r$, as

$$\left\{y_i^{(1)}, i \in n(s)\right\} \text{ and } \left\{y_i^{(2)}, i \in n(r)\right\};$$

Let $Z_i = \left(y_i^{(1)} - y_i^{(2)}\right)$ *for* $i \in S(n_r)$

Then, unbiased Estimate of SRV = $\left(\sum_r \frac{Z_i^2}{\pi_i}\right)\cdot\frac{n_s}{2n_r}$, and

unbiased Estimate of CRV = $\left(\sum_r \frac{Z_i Z_j}{\pi_i \pi_j}\right)\cdot\frac{n_s(n_s-1)}{2n_r(n_r-1)}$; (4.5)

It may be noted that (4.5) can be used to estimate SRV and CRV of (3.8) and (3.9).

However, in case of correlated response variance for other categories inclusive with interviewer's variance, (4.5) can not estimate separate correlated response variances due to all the categories.

c) **Survey Design: Methods that use a combination of interpenetrated samples and repeat measurements**.

Methods that use a combination of interpenetrated/ replicated samples and repeat measurements are reflected in Fellegi's work [(1964), (1974)]. Survey design developed by him with the use of both interpenetration and repeat measures is a very complex one. He, in his paper (1964) extended the model of Hansen, Hurwitz and Bershad (1961) to provide a frame work for joint application of two devices namely interpenetration and interviewer traditionally used to measure response variance. He built up some estimating equations of the parameters involved and these equations help one provide estimator for non-sampling variances, though biased. However, Fellegi (1974) came up with a relatively simple design

compared to the previous one [Fellegi (1964)]. But this survey design is also not simple from the operational point of the experimental survey design.

4.3 Estimation of Non Sampling Variance in the linear Model frame work

Under linear model, we have $Y_{ijt} = Y_i + b_j + e_{ij}$ (4.7)

Let the statistic within interview Measurement Square (WIMS) be defined as

$$WIMS = \sum_{j=1}^{J} v_j \sum_{i=1}^{N} c_{ij} \left(y_{ijt} - \bar{y}_{it}\right) / \ (m-1)J,$$

v_j, c_{ij} being defined as in (3.11), then under the assumption of fixed population of interviewers and $Y_{ij} = Y_i$; we have,

$$E(WIMS) = \sigma_e^2 + S_Y^2 \tag{4.8}$$

However, if one uses BMWE based on **repeat measurement** without the device of **interpenetration**, estimate of simple response variance, but not correlated response variance would be made possible, as one may observed that

$$E(BMWE) = \sigma_b^2 + \sigma_e^2 . \tag{4.9}$$

Hartley et al (1977) provided estimates of variance of **only elementary errors** under a two-stage sampling design using **linear model approach**. The linear model structure used by them was to capture the interviewer and Coders effect along with elementary errors. However they only estimated **variance of elementary errors**, by synthesis based method, which is a MINQUE estimate in component variance estimation problem.

It may be noted that, most of the available methods in estimating measurement variance fall under the category of **variance decomposition approach**; [For references, see papers by Koop (1974), Koch (1973), Nathan (1973), Chai (1971), Folsom (1980) etc.]. All the models developed so far are based on very complex survey designs.

Compared to estimation of non sampling variances under decomposition approach, the work of estimation following linear model formulation are not too many, except the early paper by Sukhatme and Seth (1952), followed by the work of Hartley and Rao (1978), Biemer (1978), Biemer's paper can be considered as an extension of the paper by Hartley and Rao (1978).

In the sections to follow, we have provided a simple estimation procedure for obtaining simple as well as separate correlated response variances following the linear model formulation.

5 Interactive Linear Model

Considered is the problem of simple response and correlated response variances, where a set of investigators are employed to extract responses from a set of respondents and additionally, a set supervisers are also employed to oversee the entire process and take corrective measures. Not withstanding the investigator and superviser bias, a **mixed effect model** has been developed which, in turn, has been used for the purpose of estimation of simple as well as correlated response variances, due to investigators and supervisors.

$$\text{Let } Y_{ijkt} = Y_i + b_j + S_k + e_{ijk} \tag{5.1}$$

where, Y_i, b_j, S_k being true value, investigator effect and superviser effect respectively. It is assumed that Y_i, b_j, S_k and e_{ijk} are mutually independent and that the b_j, S_k arise from an infinite population of random effects. The $b_j's$ are independently distributed with $E(b_j) = 0$, and $V(b_j) = \sigma_b^2$; $S_k'S$ are independently distributed with $E(s_k) = 0$ and $V(s_k) = \sigma_s^2$; Similarly, $e_{ijk's}$ are independently distributed with $E(e_{ijk}) = 0$ and $V(e_{ijk}) = \sigma_e^2$. In this case, there is no overall bias in the measurement process and

$$E_t\left(Y_{ijkt}\right) = Y_i \tag{5.2}$$

However, it may be noted that a bias could be introduced by letting either the expected value of b_j, s_k, e_{ijk} to be non-zero.

5.1 Interviewer Setting

We consider the same setting as in (3.7).

5.2 Survey Design with the help of a symmetric BIBD

Let *J*, the number of investigators be of the form $J = 4t + 3,\ t \geq 1$.

Then, we can have r, the number of responding groups to which is assigned an investigator and λ, the number of responding groups to which is assigned every pair of investigator for work of the form

$$\left\{J = 4t + 3; r = 2t + 1,\ \lambda = t\right\} \tag{5.3}$$

$$\text{and } \left\{J = 4t + 3; r = 2t + 2,\ \lambda = t\right\} \tag{5.4}$$

(complement of 5.3)

Series (5.3) and (5.4) exist and can be constructed easily be using Galois field (J), wherever, J is a prime or power of prime. The method of construction is due to R.C. Bose. However, for BIBD'S with different values of r in a given range, extensive tables are available, which may be consulted to construct the BIBDS (Raghava Rao, pp. 91–95).

The method of construction is bared on "difference sets". One will have the initial block as: $I = \left\{x^0, x^2, x^4, \ldots, x^{4t}\right\}$, where, x is a primitive root of GF ($J = 4$ t + 3), J being a prime or power of a prime.

5.2.1 Illustration Distribution of Seven investigators into Seven Responding Groups

Here $J = 7$ with $t = 1$ in $J = 4\ t + 3$ and $r = 3$, $\lambda = 1$; We have, $I = \left\{x^0, x^2, x^4\right\}$; Using $x = 3$, as a primitive root,

$$I = \left\{3^0, 3^2, 3^4\right\} = \{\ 1, 2, 4\ \}, \tag{5.5}$$

and the sets would be as:

[1, 2, 4], [2, 3, 5], [3, 4, 6], [4, 5, 7], [5, 6, 1], [6, 7, 2], [7, 1, 3] (5.6)

Table 5.1. Distribution of Interviewer Assignment into Responding Groups

Responding Groups Investigators	U_1	U_2	U_3	U_4	U_5	U_6	U_7
1	✔	✔		✔			
2		✔	✔		✔		
3			✔	✔		✔	
4				✔	✔		✔
5	✔					✔	
6		✔				✔	✔
7	✔		✔				✔

5.2.2 Distribution of Investigator work to Supervisers

Let there be two supervisers S_1 and S_2 ; The investigators work are assigned randomly in to two supervisers according to the following lay out.

Table 5.2. Distribution of Investigator's work into two Supervisers

Investigators Supervisers	1	2	3	4	5	6	7
S_1	✔	✔	✔	✔			
S_2				✔	✔	✔	✔

5.3 Estimation of Simple and Correlated Response Variances

$$\left(\sigma_b^2, \sigma_s^2, \sigma_e^2 \text{ and } \sigma_b^2 + \sigma_s^2 + \sigma_e^2\right)$$

For simplicity of calculation. It has been assumed that each responding unit has only one respondent. However our method of estimation would remain unchanged, even if every $U_j\,(j = 1,2,\ldots,7)$ has more than one respondents. In that case, only the size of the data matrix would be larger.

5.3.1 Acquisition and modeling of Data

Following the methods of data collection and of supervision (Ref. Tables 5.1, 5.2) we would have 24 recorded and supervised values. Thus we shall have,

$$\underset{\sim}{Y}_{24\times 1} = \left\{y_{ijk}; i = 1,2,\ldots 7; j = 1,2,\ldots 7 \text{ and } k = 1,2\right\}$$

For example, data after collection and after supervision, we would have four observations from each of the respondents namely 4th, 5th and 7th. In fact data from the 5th respondent would read as follows:

$$\begin{pmatrix} y_{521} \\ y_{541} \\ y_{542} \\ y_{552} \end{pmatrix} = \begin{pmatrix} y_5 + b_2 + s_1 + e_{521} \\ y_5 + b_4 + s_1 + e_{541} \\ y_5 + b_4 + s_2 + e_{542} \\ y_5 + b_5 + s_2 + e_{552} \end{pmatrix} \qquad (5.7)$$

Similarly, there will be 3 observations each from 1st, 2nd, 3rd and 6th respondent, thus totaling to 24 observations.

5.3.2 Canonical Reduction of the Data

Theorem 5.2.1. The $\underset{\sim}{Y}$ can be partitioned as $\underset{\sim}{Y} = \begin{pmatrix} \underset{\sim 7\times 1}{U} \\ \underset{\sim 17\times 1}{V} \end{pmatrix}$ with the dispersion matrix $\sum_{24\times 24}$ which can be partitioned also as $\sum = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}$, where, $\sum_{11}$ and $\sum_{22}$ are the variance. Covariance matrices of $\underset{\sim}{U}$ and $\underline{V}$.

$$\underset{\sim}{U} = \begin{bmatrix} \frac{1}{\sqrt{3}}(Y_{111} + Y_{152} + Y_{172}) \\ \frac{1}{\sqrt{3}}(Y_{211} + Y_{221} + Y_{262}) \\ \frac{1}{\sqrt{3}}(Y_{321} + Y_{331} + Y_{372}) \\ \frac{1}{\sqrt{4}}(Y_{411} + Y_{431} + Y_{441} + Y_{442}) \\ \frac{1}{\sqrt{4}}(Y_{521} + Y_{541} + Y_{542} + Y_{552}) \\ \frac{1}{\sqrt{3}}(Y_{631} + Y_{652} + Y_{662}) \\ \frac{1}{\sqrt{4}}(Y_{741} + Y_{742} + Y_{762} + Y_{772}) \end{bmatrix};$$

$$
\underset{\sim}{V} = \begin{bmatrix}
\frac{1}{\sqrt{2}} y_{111} - \frac{1}{\sqrt{2}} y_{152} \\
\frac{1}{\sqrt{6}} y_{111} + \frac{1}{\sqrt{6}} y_{152} - \frac{2}{\sqrt{6}} y_{172} \\
\frac{1}{\sqrt{2}} y_{211} - \frac{1}{\sqrt{2}} y_{221} \\
\frac{1}{\sqrt{6}} y_{211} + \frac{1}{\sqrt{6}} y_{221} - \frac{2}{\sqrt{6}} y_{262} \\
\frac{1}{\sqrt{2}} y_{321} - \frac{1}{\sqrt{2}} y_{331} \\
\frac{1}{\sqrt{6}} y_{321} + \frac{1}{\sqrt{6}} y_{331} - \frac{1}{\sqrt{6}} y_{372} \\
\frac{1}{\sqrt{2}} y_{411} - \frac{1}{\sqrt{2}} y_{431} \\
\frac{1}{\sqrt{6}} y_{411} + \frac{1}{\sqrt{6}} y_{431} - \frac{2}{\sqrt{6}} y_{441} \\
\frac{1}{\sqrt{12}} y_{411} + \frac{1}{\sqrt{12}} y_{431} + \frac{1}{\sqrt{12}} y_{441} - \frac{3}{\sqrt{12}} y_{422} \\
\frac{1}{\sqrt{2}} y_{521} - \frac{1}{\sqrt{2}} y_{541} \\
\frac{1}{\sqrt{6}} y_{521} + \frac{1}{\sqrt{6}} y_{541} - \frac{2}{\sqrt{6}} y_{542} \\
\frac{1}{\sqrt{12}} y_{521} + \frac{1}{\sqrt{12}} y_{541} + \frac{1}{\sqrt{12}} y_{542} - \frac{3}{\sqrt{12}} y_{552} \\
\frac{1}{\sqrt{2}} y_{631} - \frac{1}{\sqrt{2}} y_{652} \\
\frac{1}{\sqrt{6}} y_{631} + \frac{1}{\sqrt{6}} y_{652} - \frac{2}{\sqrt{6}} y_{662} \\
\frac{1}{\sqrt{2}} y_{741} - \frac{1}{\sqrt{2}} y_{742} \\
\frac{1}{\sqrt{6}} y_{741} + \frac{1}{\sqrt{6}} y_{742} - \frac{2}{\sqrt{6}} y_{762} \\
\frac{1}{\sqrt{12}} y_{741} + \frac{1}{\sqrt{12}} y_{742} + \frac{1}{\sqrt{12}} y_{762} - \frac{3}{\sqrt{12}} y_{772}
\end{bmatrix}
$$

Proof: The reduction follows immediately by Helmet transformation.

It may be shown that, $\sum_{11}$ Dispersion matrix of $\underset{\sim}{U}$ and $\sum_{22}$ the dispersion matrix of $\underset{\sim}{V}$ can be represented as

$$\sum_{11} = A_1' \sigma_b^2 + A_2' \sigma_s^2 + I \sigma_e^2$$

and (5.8)

$$\sum_{11} = A_1 \sigma_b^2 + A_2' \sigma_s^2 + I \sigma_e^2$$

where, A_1, A_2, A_2', A_2' are all symmetric matrices of real numbers. After calculation, the matrices A_1, A_2, A_2', A_2' in $\sum_{11}$ and $\sum_{22}$ have been found to be as given in the tables (5.3), (5.4) & (5.5).

5.4 Estimation of σ_b^2, σ_s^2 and σ_e^2

5.4.1 Estimation of σ_e^2

We have, $Y_{ijk} = Y_i + b_j + S_k + e_{ijk}$;

$$(i = 1,2,\ldots,I;\ j = 1,2,\ldots,J \text{ and } k = 1,2,\ldots,k)$$

$$= \overline{Y} + (Y_i - \overline{Y}) + b_j + S_k + e_{ijk} \quad (5.9)$$

$$= \overline{Y} + \alpha_i + b_j + S_k + e_{ijk} \quad \text{with } \Sigma\alpha_i = 0.$$

Let $SSE = \sum_i \sum_j \sum_k \left[y_{ijk} - y_i \ldots - y_{.j.} - y_{..k} + 2y_{\ldots} \right]^2$, then it can be shown that

$$E(SSE) = \left[IJK - \{(I-1) + (J-1) + (K-1)\} - 1 \right] \sigma_e^2$$

$$\text{i.e., } E\ \{ [SSE / \ [IJK - \{(I-1) + (J-1) + (K-1)\} - 1\ ]] = \sigma_e^2$$

$$\text{Therefore, } \hat{\sigma}_e^2 = \left[SSE / \ [IJK - \{I-1) + (J-1) + (K-1)\} - 1] \right] \quad (5.10)$$

5.4.2 Estimation of σ_e^2

Let $\underset{\sim}{\varepsilon_1}$ and λ_1 be the eigen vector corresponding to maximum eigen value λ_1 of the matrix A_1. Now, from (5.8), we have,

$$\varepsilon_1' \hat{\sum}_{22} \varepsilon_1 = \varepsilon_1' A_1 \varepsilon_1 \hat{\sigma}_b^2 + \varepsilon_1' A_2 \varepsilon_1 \hat{\sigma}_b^2 + \sigma_e^2 \qquad (5.11)$$

After calculaton, all eigen values appeared to be non negative, as they are expected. On computation, maximum eigen value λ_1 becomes 3.5 and $\varepsilon_1' A_2 \varepsilon_1$ becomes – 7.85704 e – 005 which is almost 0.

Thus, from (5.11), we have

$$\varepsilon_1' \hat{\sum}_{22} \varepsilon_1 - \hat{\sigma}_e^2 = 3.5 \hat{\sigma}_b^2 - .000078570 47 \hat{\sigma}_s^2,$$

Thus, $$\hat{\sigma}_b^2 = \left[\varepsilon_1' \hat{\sum}_{22} \varepsilon_1 - \hat{\sigma}_e^2 \right] \Big/ 3.5 \qquad (5.12)$$

Table. 5.3 A_1, the Coefficient Matrix of σ_b^2 in $\sum 22$

1	0	0.5	.28867513	0	0	0.5	.28867513	-.204124	0	0	.612372	0.5	.288675	0	0	0
	1	.28867513	.166666	0	0	.28867513	.16666	.11785113	0	0	-.3535539	-.28867513	.16666	0	0	.70710678
		1	0	-0.5	-28867513	0.5	.28867513	.2041241659	-0.5	-.2886751	-.20412416	0	0	0	0	0
			1	.28867513	.166666	.28867513	.166666	.1178511309	.28867513	.1666	.11785133	0	.6666	0	.6666	-.23570226
				1	0	0.5	-.28867513	-.204124165	0.5	.28867513	.204124165	-0.5	-.28867513	0	0	0
					1	-.28867513	.16666	.17785113019	.28867513	.16666	.1178511309	.28867513	.16666	0	0	.7071067
						1	0	0	0	0	0	-0.5	-.28867513	0	0	0
							1	.70710678	.57735	.33333	-.47140452	.28867513	.1666	0	.66666	.471405
								0.5	.408248333	.2357023	-.333333	.204124166	.11785113	0	-.471405	-.16666
									1	.57735026	-204124165	0	0	0	-.5773503	-.4082483
										.33333	-.11785113	0	0	0	-.33333	-.2357023
											1.16666	.61237249	-.35355339	0	.4714045	.33333
												1	0	0	0	0
													1	0	.666666	-.23570226
														0	0	0
															1.33333	.235702
																1.16666

Table. 5.4 A_2, the Coefficient Matrix of σ_s^2 in $\sum 22$

1	.5773502	0	1.154705	0	1.154705	0	0	1.224745	0	1.154705	.8164966	1	.57735	1	.57735026	4082483
	.33333	0	.666666	0	.66666	0	0	.70710678	0	.6666	.4714045	.57735026	.33333	.57735026	.33333	.2357024
		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
			1.33333	0	1.3333	0	0	1.4142135	0	1.3333	.94280904	1.1547005	.66666	1.1547005	.666666	.4714045
				0	0	0	0	0	0	0	0	0	0	0	0	0
					1.33333	0	0	1.41421356	0	1.3333	.942809	1.1547005	.66666	1.1547005	.66666	.4714045
						0	0	0	0	0	0	0	0	0	0	0
							0	0	0	0	0	0	0	0	0	0
								1.5	0	1.41421356	1	1.2247449	.70710678	1.2247449	.707106678	0.5
									0	0	0	0	0	0	0	0
										1.33333	.942809	1.1547005	.66666	1.1547005	.66666	.47140452
											.666666	.8164966	.4714045	.8164966	.4140452	.333333
												1	.57735026	1	.57735026	.40824833
													.333333	.57735026	.33333	.23570266
														1	.57735026	.40824833
															.333333	.23570226
																.1666666

Table. 5.5. Coefficient Matrix of σ_b^2, σ_s^2 in $\sum_{11}$

$A_1' =$

1	.33333	.33333	.28867513	.28867513	.33333	.28867513
	1	.33333	.28867513	.28867513	.33333	.28867513
		1.5	.28867513	.28867513	.33333	.28867513
			1.5	1	.28867513	1
				1.5	.28867513	1
					1	.28867513
						1.5

$A_2' =$

1.66666	1.33333	1.33333	1.44337	1.732051	1.66666	2.02073
	1.66666	1.66666	2.02073	1.732051	1.33333	1.443375
		1.66666	2.02073	1.732051	1.33333	1.443375
			2.5	2	1.443375	1.5
				2	1.732051	2
					1.66666	2.02073
						2.5

5.4.3 Estimation of σ_s^2:

Method – I

Simple response variance, i.e., $\left(\sigma_b^2 + \sigma_s^2 + \sigma_e^2\right)$ can be estimated from the repeat measurements. This can be obtained through our survey design.

Hence, $\hat{\sigma}_S^2 = \left(\sigma_b^2 + \sigma_s^2 + \sigma_e^2\right) - \hat{\sigma}_b^2 - \hat{\sigma}_e^2$ (5.13)

Method – II

Let ε_2 and λ_2 be the eigen vector corresponding to maximum eigen value $\tilde{\lambda}_2$ of the matrix A_2. All the eigen values appeared to be non-negative as they are expected and the maximum eigen value is found to be 1.03332.

Now, from (5.8), we have ,

$$\varepsilon_2^{'} \hat{\sum}_{22} \varepsilon_2 = \varepsilon_2^{'} A_1 \varepsilon_2 \hat{\sigma}_b^2 + \varepsilon_2^{'} A_2 \varepsilon_2 \hat{\sigma}_b^2 + \hat{\sigma}_e^2$$

(5.14)

$$\varepsilon_2^{'} \hat{\sum}_{22} \varepsilon_{-2} \hat{\sigma}_e^2 = 1.650562751 \hat{\sigma}_b^2 + 1.03332 \hat{\sigma}_s^2$$

Now, from (5.10), (5.12) and (5.13), $\hat{\sigma}_S^2$ can be obtained.

6 Estimation of Simple and Correlated Response variances associated with measurement process in estimating population total

Let y_{ijkt} be the observation collected from the i^{th} respondent by the j^{th} investigator and supervised by the k^{th} superviser at t^{th} trial following our survey design (Ref. Tables 5.1, 5.2). The investigator's and superviser's assignment rule has been defined earlier in section 5 and under the assumption of the model parameters,

$$\text{we have, } E_t\left(y_{ijkt}\right) = Y_i, \tag{6.1}$$

using the notation similar to those in (3.11),

Let $C_{ij} = \begin{cases} 1, \text{ if } j^{th} \text{ investigator is assigned to the } i^{th} \text{ respondent} \\ 0, \text{ otherwise} \end{cases}$

and $C_{ij(k)} = \begin{cases} 1, \text{if } k^{th} \text{ superviser superviser the job to } j^{th} \text{ investigator} \\ 0, \text{otherwise} \end{cases}$

Let $\hat{Y} = \sum_i \sum_j \sum_k C_{ij}\ C_{ijk}\ y_{ijkt}$

Then we have the following

Theorem 6.1.: $\hat{Y}$ is unbiased with simple response $\left(\sigma_b^2 + \sigma_s^2 + \sigma_e^2\right)$ and correlated response variances $\sigma_b^2, \sigma_s^2, \sigma_e^2$ (6.2)

Proof: $E(\hat{Y}) = E_R\ E_{t|R}\ (\hat{Y})$,

where *t* denotes the measurement variable and *R* stands for the random allocation of the investigators to the respondents. Therefore,

$$E(\hat{Y}) = \sum_i \sum_j \sum_k E_{t|R}\left(y_{ijkt} \middle| C_{ij}\ C_{ij(k)} = 1\right) E_R\left(C_{ij}\ C_{ij(k)} = 1\right)$$

$$= \sum_i \sum_j \sum_k Y_i \frac{1}{JK} = \sum_{i=1}^{I} Y_i \sum_j \sum_k \frac{1}{JK},$$

Since, $E\left(C_{ij} = 1\right) = \frac{1}{J}$ and $E\left(C_{ij(k)} = 1 \middle| C_{ij} = 1\right) = \frac{1}{K}$

Therefore, $E(\hat{Y}) = \sum_i Y_i = Y$ (6.3)

Now, measurement variance

$$\text{MV}(\hat{Y}) = E\left(\sum_i \sum_j \sum_k C_{ij}\, C_{ij(k)} \left(y_{ijkt} - E_t\left(y_{ijkt}\right)\right)\right)^2$$

Proceeding same as before in (3.13), it may be shown that the corresponding terms in (3.13) under this situation can be found to be

1st term = $I\left(\sigma_b^2 + \sigma_s^2 + \sigma_e^2\right)$

2nd term = $I\sigma_S^2$

3rd term = $I \cdot \sigma_S^2$

4th term = 0 (6.4)

5th term = $I(I-1)\left(\sigma_b^2 + \sigma_s^2\right)$

6th term = $I(I-1)\sigma_s^2$

7th term = $I(I-1)\sigma_b^2$

8th term = 0

Combining all the terms of (6.4), we have,

$$MV(\hat{Y}) = I^2\left(2\sigma_b^2 + 2\sigma_s^2 + \frac{\sigma_e^2}{I}\right) \quad (6.5)$$

Let, $\zeta_1^* = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_s^2 + \sigma_e^2}$ (6.6)

$$\zeta_2^* = \frac{\sigma_s^2}{\sigma_b^2 + \sigma_s^2 + \sigma_e^2}$$

and $\zeta_3^* = \frac{\sigma_e^2}{\sigma_b^2 + \sigma_s^2 + \sigma_e^2}$

then, $MV(\hat{Y}) = I^2 \left(\sigma_b^2 + \sigma_s^2 + \sigma_e^2\right) \quad \left[2\zeta_1^* + 2\zeta_2^* + \frac{\zeta_e^*}{I}\right]$ (6.7)

and $MV(\bar{Y}) = 2\left(\sigma_b^2 + \sigma_s^2 + \sigma_e^2\right) \quad \left(\zeta_1^* + \zeta_2^*\right) + \frac{\sigma_e^2}{I}$;

The form of the expressions in (6.7) are similar to that of Hansen, Hurwitz and Bershad (1961).

Following the methodology already discussed in the section in the section 5, $\sigma_b^2, \sigma_s^2, \sigma_e^2$ can be estimated.

7 Some Concluding Remarks and Discussions

1. Symmetric BIBD was needed to be constructed for this specific interview setting, where the number of responding groups into which the population was partitioned was equal to the number of interviewers. However, the design need not necessarily be always that of a symmetric BIBD. Depending on the interviewer setting, it could be non-symmetric BIBD also. In fact, intensive tables are available for construction of symmetric and non-symmetric BIBD for any specied value of J.
2. The proposed survey design is very simple to construct and to operate compared to the earlier methods, where both **interpenetration** and **repeat measurement techniques** have been used [(Fellegi, (1964), (1974); Beimer (1985)]. In fact, from the user point of view, he can simply consult the available tables for constructing the BIBDS for allocation of the interviewer assignment, without having in depth knowledge on Galois field etc. Our proposed method is a very user friendly one.
3. The method of estimation procedure also involves easy computation. It requires calculation of only eigen values and eigen vectors and for that standard computer package is readily available. The method is very flexible also in the sense that it can accommodate any number of effects arising from different categories of survey personnel from survey management group.
4. For illustration purpose, we have taken J = 7 and number of respondents from each U_j to be one. However, our estimation procedure would remain

unchanged, even if we have more number of investigators and more number of respondents from each responding group. Only the size of the data matrix will be larger, but the estimation procedure would remain unaltered.

REFERENCES

BAILAR, BARBARA A. and TORE DALENIUS (1969): Estimating the Response Variance Components of the U.S. Bureau of the Census Survey Model, Sankhȳa , 31B, 341–360.

BIEMER, PAUL P. and S. LYNNE STokes (1985): Optimal Design of Interviewer Variance Experiments in Complex Surveys, JASA, 80, 158–166.

CHAI, JOHN J (1971): Correlated measurement errors and the least Square Estimators of the Regression Coefficient, JASA, 66, 478–483.

DALENIUS TORE (1974): The Ends and Means of Total Survey Design. Stockholm: The University of Stockholm.

FELLEGI, IVAN P. (1964): Response variation and estimation, JASA, 59, 1016–1041.

---------- (1974): An Improved method of Estimating the Correlated Response Variance, JASA, 69, 496–501.

FOLSOM, RALPH E. Jr. (1980): U-Statistics estimation of Variance Components for unequal probability samples with non additive Interviewer and response errors, American Statistical Association 1980 proceedings of the Survey Research Methods Section, 137–142.

HANSEN, MORRIS H. WILLIAM N. HURWITZ, ETES. MARKS, and W. PARKER MAULDIN (1951): Response Errors in Survey, JASA, 46, 147–190.

---------- and MAX A. BERSHAD (1961): Measurement Errors in Censuses and Surveys, Bulletin of the International Statistical Institute, 38, 359–374.

HANSEN, MORRIS H, WILLIAM N. HURWITZ and LEAN PRITZKER (1964): The Estimation and Interpretation of Gross differences and the simple response variance: In C. R. Rao with D.B. Lahiri, K.R. Nair, P. Pant and S.S. Shrikhande eds. Contributions to Statistics Presented to Professor P.C. Mahalanobis on the Occasion of his 70th Birthday, Oxford, England: Pergamen Calcutta Statistical Publishing society, 111–136.

HARTLEY, H.O. and RAO, J.N.K. (1978): The Estimation of Non sampling Variance Components in Sample Surveys, N. Krishnan Namboodiri, ed., Survey Sampling and Measurement, New York; Academic.

JABINE, J.B. and B.J. TEPPING (1973): Controlling the quality of occupation and industry data, Bulletin of the International Statistical Institute, 45 (3), 36–389.

KOOP, J.C. (1974): Notes for a unified theory of Estimation for Sample Surveys taking into account Response Errors, Metrika 21, 19–39.

KOCH GORY G. (1973): An alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications to estimators involving Subclass Means, JASA, 68, 906 – 913.

LESSLER, JUDITH T. and KALSBEEK, WILLIAM, D. (1992): Non sampling Errors in Surveys. John Wiley and Sons. Inc.

MAHALANOBIS, P.C. (1946): Recent Experiments in Statistical Sampling in the Indian Statistical Institute, JRSS, 109, 327–328.

NATHAN, GAD (1973): Response Errors Based on Different Samples, Sankhȳa , 35 A, 205–220.

NELSON, FORREST D. (1977): Censored Regression Models with unobserved Stochastic Censoring Thresholds, Journal of Econometrics, 6, 581–592.

RAGHAVA RAO, D. (1971): Constructions and Combinatorial Problems in Design of Experiments. John Wiley and Sons. Inc.

RAJ DES (1956): Some Estimators in sampling with varying probabilities without replacement, JASA, 51, 269–284.

RAJ DES (1968): Sampling Theory, New York: Mc Graw Hill.

SUKHATME, PANDURANG VASUDEO and G.R. SETH (1952): Non sampling Errors in Surveys, Journal of the Indian Society of Agricultural Statistics, 4, 5–41.

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 223—233

MULTISTAGE BALANCED GROUPS RANKED SET SAMPLES FOR ESTIMATING THE POPULATION MEDIAN

Amer Ibrahim Al-Omari[1], Kamarulzaman Ibrahim[2], Abdul Aziz Jemain[2], and Said Ali Al-Hadhrami[3]

ABSTRACT

A multistage balanced groups ranked set samples (MBGRSS) method and its properties for estimating the population median is considered. The suggested estimator is compared to those obtained based on simple random sampling (SRS) and the ranked set sampling (RSS) methods. The MBGRSS estimator of the population median is found to be unbiased if the underlying distribution is symmetric and has a small bias if the underlying distribution is asymmetric, the bias is decreasing in r (r is the number of stage). It is found that, MBGRSS is as efficient as RSS when $m = 3$ and $r = 1$, and it is more efficient than RSS for $r > 1$. However, the efficiency of MBGRSS is increasing in r for specific value of the sample size whether the underlying distribution is symmetric or asymmetric. Real data is used to illustrate the method.

Key words: Ranked set sampling; simple random sampling; balanced groups ranked set samples, symmetric distribution; asymmetric distribution.

1. Introduction

McIntyre (1952) introduced the ranked set sampling method to estimate mean pasture and forage yields. Takahasi and Wakimoto (1968) provided the necessary mathematical theory of RSS. Dell and Clutter (1972) considered the case in which the ranking may be done with errors. Samawi et al. (1996) investigated the extreme ranked set samples (ERSS) for estimating a population mean. Muttlak (1997) suggested using median ranked set sampling (MRSS) to estimate the population mean. Al-Saleh and Al-Omari (2002) proposed the multistage ranked set sampling (MSRSS) method to increase the efficiency when estimating the

[1] Department of Mathematics, Faculty of Science and Nursing, Jerash Private University, Jordan.
[2] School of Mathematical Sciences, Faculty of Science and Technology, University Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia.
[3] Department of Mathematics, College of Applied Sciences, Nizwa, Oman.

population mean for specific value of the sample size. Muttlak (2003a, 2003b) suggested quartile ranked set sampling (QRSS) and percentile ranked set sampling (PRSS) for estimating the population mean and showed that PRSS and QRSS produced unbiased estimators of the population mean when the underlying distribution is symmetric. Jemain and Al-Omari (2006) proposed multistage median ranked set sampling (MMRSS) method for estimating the population mean. They found that MMRSS is more efficient than the commonly used SRS based on the same sample size. Jemain et al. (2007) suggested multistage extreme ranked set sampling (MERSS) method for estimating the population mean. Al-Omari and Jaber (2008) investigated estimation the population mean using percentile double ranked set sampling. For more details about RSS see Zheng and Modarres (2006), Tseng and Shao (2007), and Ozturk and Deshpande (2006), and Tiensuwan et al., (2007)

1.1. Ranked set sampling

The RSS involves randomly selecting m^2 units from the population. These units are randomly allocated into m sets, each of size m. The m units of each sample are ranked visually or by any inexpensive method with respect to the variable of interest. From the first set of m units, the smallest unit is measured. From the second set of m units, the second smallest unit is measured. The process is continued until from the mth set of m units the largest unit is measured. Repeat the process n times to have a set of size mn from the initial m^2n units.

1.2. Multistage balanced groups ranked set samples

The MBGRSS procedure consists of the following steps:

Step 1: Randomly select $(3k)^{r+1}$ $(k = 1, 2, ...)$ units from the target population. These units are randomly allocated into $(3k)^r$ sets, each of size $3k$.

Step 2: The $3k$ units of each set are ranked visually or by any inexpensive method with respect to the variable of interest. Then the $(3k)^r$ sets are divided into three groups, each of $3^{r-1}k^r$ sets.

Step 3: From each set in the first group, the smallest rank unit is selected; from each set in the second group; the median rank unit is selected, and from each set in the third group, the largest rank unit is selected. This step yields $(3k)^{r-1}$ sets, $3^{r-2}k^{r-1}$ sets in each group.

Step 4: Without doing any actual quantification, from the $3^{r-2}k^{r-1}$ sets in the first group, smallest rank unit is selected, from the $3^{r-2}k^{r-1}$ sets in the second group; the median rank unit is selected, and from the $3^{r-2}k^{r-1}$ sets in the third

group; the largest rank unit is selected. This step yields $(3k)^{r-2}$ sets, $3^{r-3}k^{r-2}$ sets in each group each set of size $3k$.

Step 5: The process is continued using Steps (3) and (4) until we end up with one rth stage balanced groups ranked set sample of size $3k$.

The procedure is repeated n times if needed to get a sample of size $3kn$ from initial $(3k)^{r+1}n$ units. It is clear that, based on MBGRSS method, the measured units at the last stage involves the same number of minimums, medians and maximums, which is different from the usual RSS where we select the ith smallest ranked unit from the ith sample. These features make the MBGRSS more practical in the field since it is easy to identify the minimums or the maximums and the medians of the sample particularly when sample size is odd. Indeed the MBGRSS, which is, based on $(3k)^{r+1}$ units, is more informative than both RSS and SRS, since these later two methods are based on $3k$ and $9k^2$ units respectively. For even and odd sample sizes, the suggested method is denoted by MBGRSSE and MBGRSSO, respectively.

Let us consider the following example to illustrate MBGRSS for estimating the population median.

2. Example

Consider the case of $r=3$ and $k=1$, then $m=3k=3$. So we will select $(3k)^{r+1}=81$ units, which are $X_1, X_2, ..., X_{81}$. Allocate the 81 selected units into $(3k)^r=27$ sets each of size 3. The 3 units of each set are ranked with respect to the variable of interest as follows: $\{X_{i(1:3)}, X_{i(2:3)}, X_{i(3:3)}\}$, $(i=1,2,...,27)$. Now, allocate the 27 sets into 3 groups, each of 9 sets as:

First group: $\{X_{i(1:3)}, X_{i(2:3)}, X_{i(3:3)}\}$, $(i=1,2,...,9)$,

Second group: $\{X_{i(1:3)}, X_{i(2:3)}, X_{i(3:3)}\}$, $(i=10,11,...,18)$,

Third group: $\{X_{i(1:3)}, X_{i(2:3)}, X_{i(3:3)}\}$, $(i=19,20,...,27)$.

Then, for $r=1$, we select the smallest ranked unit from each set in the first group, and the median ranked unit from each set in the second group, and the largest ranked unit from each set in the third group. We then allocate these selected units into 27 sets, 9 sets in each group.

First group: $\{X^{(1)}_{3(i-1)+1(1:3)}, X^{(1)}_{3(i-1)+2(1:3)}, X^{(1)}_{3(i-1)+3(1:3)}\}$, $(i=1,2,3)$

Second group: $\{X^{(1)}_{3(i-1)+1(2:3)}, X^{(1)}_{3(i-1)+2(2:3)}, X^{(1)}_{3(i-1)+3(2:3)}\}$, $(i=4,5,6)$

Third group: $\{X^{(1)}_{3(i-1)+1(3:3)}, X^{(1)}_{3(i-1)+2(3:3)}, X^{(1)}_{3(i-1)+3(3:3)}\}$, $(i=7,8,9)$.

For $r=2$, rank the units within each set in each group,

$$\left\{X_{i(1:3)}^{(2)}, X_{i(2:3)}^{(2)}, X_{i(3:3)}^{(2)}\right\},\ \left(i=1,2,...,9\right).$$

Then, we select the smallest ranked unit from each set in the first group, and the median ranked unit from each set in the second group, and the largest ranked unit from each set in the third group. We then allocate these selected units into 9 sets, 3 sets in each group.

First group: $\left\{X_{1(1:3)}^{(2)}, X_{2(1:3)}^{(2)}, X_{3(1:3)}^{(2)}\right\}$,

Second group: $\left\{X_{4(2:3)}^{(2)}, X_{5(2:3)}^{(2)}, X_{6(2:3)}^{(2)}\right\}$,

Third group: $\left\{X_{7(3:3)}^{(2)}, X_{8(3:3)}^{(2)}, X_{9(3:3)}^{(2)}\right\}$.

For $r=3$, rank the units within each set in each group

$$\left\{X_{i(1:3)}^{(3)}, X_{i(2:3)}^{(3)}, X_{i(3:3)}^{(3)}\right\},\ \left(i=1,2,3\right).$$

Then, we select the smallest ranked unit from each set in the first group, and the median ranked unit from each set in the second group, and the largest ranked unit from each set in the third group. The median of these units is considered as an estimator of the population median as

$$\hat{\eta}_{MBGRSSO}^{(r)} = \text{median}\left\{X_{1(1:3)}^{(3)}, X_{2(2:3)}^{(3)}, X_{3(3:3)}^{(3)}\right\}$$

It is of interest to note here that for $m=3$ at the first stage, MBGRSS is the same as RSS for estimating the population mean.

3. Estimation of the population median

Let $X_1, X_2, ..., X_m$ be a random sample with pdf $f(x)$, cdf $F(x)$, a finite mean μ and variance σ^2. Let $X_{11h}, X_{12h}, \ldots, X_{1mh}; X_{21h}, X_{22h}, \ldots, X_{2mh}; \ldots; X_{m1h}, X_{m2h}, \ldots, X_{mmh}$ be independent random variables all with the same cumulative distribution function $F(x)$ in the hth cycle $\left(h=1,2,...,n\right)$. Let $X_{i(1:m)h}, X_{i(2:m)h}, \ldots, X_{i(m:m)h}$ be the order statistics of the ith sample $X_{i1h}, X_{i2h}, ..., X_{imh}$, $\left(i=1,2,...,m\right)$. Then $X_{1(1:1)h}, X_{2(2:2)h}, \ldots, X_{m(m:m)h}$ denote the measured RSS.

The SRS estimator of the population median η is defined as:

$$\hat{\eta}_{SRS} = \begin{cases} X_{\left(\frac{m+1}{2}:m\right)h} & \text{, if } m \text{ is odd} \\ \frac{1}{2}\left(X_{\left(\frac{m}{2}:m\right)h} + X_{\left(\frac{m+2}{2}:m\right)h} \right) & \text{, if } m \text{ is even.} \end{cases} \quad (1)$$

The RSS estimator of the population median η from a sample of size *m* is given by

$$\hat{\eta}_{RSS} = \text{median}\left\{X_{i(i:m)h}, i = 1,2,...,m\right\}. \tag{2}$$

If *m* is odd, in the *h*th cycle $(h = 1,2,...,n)$, let $X^{(r)}_{i(1:m)h}$ be the lowest ranked unit of the *i*th sample $(i = 1,2,...,k)$, $X^{(r)}_{i\left(\frac{m+1}{2}:m\right)h}$ be the median of the *i*th sample $(i = k+1,\ k+2,...,2k)$, and $X^{(r)}_{i(m:m)h}$ be the largest ranked unit of the *i*th sample $(i = 2k+1,\ \ 2k+2,...,3k)$. So that $X^{(r)}_{1(1:m)h}, X^{(r)}_{2(1:m)h}, ..., X^{(r)}_{k(1:m)h}, X^{(r)}_{k+1\left(\frac{m+1}{2}:m\right)h}$, $X^{(r)}_{k+2\left(\frac{m+1}{2}:m\right)h}$, …, $X^{(r)}_{2k\left(\frac{m+1}{2}:m\right)h}$, $X^{(r)}_{2k+1(m:m)h}, X^{(r)}_{2k+2(m:m)h}, ..., X^{(r)}_{3k(m:m)h}$ denote the measured MBGRSSO. Note that, the first *k* units are iid, and the second *k* units are iid and the last *k* units are iid. However, all units are independent but not identically distributed. The estimator of the population median using MBGRSSO is given by

$$\hat{\eta}^{(r)}_{MBGRSSO} = \text{median}\left\{\begin{array}{ll} X^{(r)}_{i(1:m)h} & , i = 1,...,k; \\ X^{(r)}_{i\left(\frac{m+1}{2}:m\right)h} & , i = k+1,...,2k; \\ X^{(r)}_{i(m:m)h} & , i = 2k+1,...,3k \end{array}\right\}. \tag{3}$$

If *m* is even, let $X^{(r)}_{i(1:m)h}$ be the lowest ranked unit of the *i*th sample $(i = 1,2,...,k)$, let $\frac{1}{2}\left(X^{(r)}_{i\left(\frac{m}{2}:m\right)h} + X^{(r)}_{i\left(\frac{m+2}{2}:m\right)h}\right)$ be the median of the *i*th sample $(i = k+1,k+2,...,2k)$, and $X^{(r)}_{i(m:m)h}$ be the largest ranked unit of the *i*th sample $(i = 2k+1,\ \ 2k+2,...,3k)$. In this case $X^{(r)}_{1(1:m)h}, ..., X^{(r)}_{k(1:m)h}$, $\frac{1}{2}\left(X^{(r)}_{k+1\left(\frac{m}{2}:m\right)h} + X^{(r)}_{k+1\left(\frac{m+2}{2}:m\right)h}\right), \ldots, \frac{1}{2}\left(X^{(r)}_{2k\left(\frac{m}{2}:m\right)h} + X^{(r)}_{2k\left(\frac{m+2}{2}:m\right)h}\right), X^{(r)}_{2k+1(m:m)h}$, $..., X^{(r)}_{3k(m:m)h}$ denote the measured MBGRSSE. In the case of even sample size, the MBGRSSE estimator is given by:

$$\hat{\eta}_{MBGRSSE}^{(r)} = \text{median}\left\{\begin{array}{ll} X_{i(1:m)h}^{(r)} & , i=1,...,k; \\ \frac{1}{2}\left(X_{i\left(\frac{m}{2}:m\right)h}^{(r)} + X_{i\left(\frac{m+2}{2}:m\right)h}^{(r)} \right) & , i=k+1,...,2k; \\ X_{i(m:m)h}^{(r)} & , i=2k+1,...,3k \end{array}\right\}. \quad (4)$$

4. Simulation Study

In this section, to compare the efficiency of proposed estimators of the population median using MBGRSS method relative to SRS method, seven probability distribution functions were considered for the populations: uniform, normal, beta, logistic, exponential, gamma and weibull. We compare the average of 60,000 sample estimates using $k=1,2,3$ corresponding to the sample sizes $m=3,6,9$ respectively. Without loss of generality for simulation, assume the cycle is repeated once. If the distribution is symmetric, the efficiency of RSS and MGBRSS relative to SRS is defined as:

$$eff\left(\hat{\eta}_{SRS},\hat{\eta}_{RSS}\right)=\frac{\text{Var}\left(\hat{\eta}_{SRS}\right)}{\text{Var}\left(\hat{\eta}_{RSS}\right)} \text{ and } eff\left(\hat{\eta}_{SRS},\hat{\eta}_{MBGRSS}^{(r)}\right)=\frac{\text{Var}\left(\hat{\eta}_{SRS}\right)}{\text{Var}\left(\hat{\eta}_{MBGRSS}^{(r)}\right)}.$$

If the distribution is asymmetric, the efficiency is defined as:

$$eff\left(\hat{\eta}_{SRS},\hat{\eta}_{RSS}\right)=\frac{\text{MSE}\left(\hat{\eta}_{SRS}\right)}{\text{MSE}\left(\hat{\eta}_{RSS}\right)} \text{ and } eff\left(\hat{\eta}_{SRS},\hat{\eta}_{MBGRSS}^{(r)}\right)=\frac{\text{MSE}\left(\hat{\eta}_{SRS}\right)}{\text{MSE}\left(\hat{\eta}_{MBGRSS}^{(r)}\right)},$$

where the $\text{MSE}\left(\hat{\eta}_{MBGRSS}^{(r)}\right)$ is the mean square error of $\hat{\eta}_{MBGRSS}^{(r)}$, which is given by

$$\text{MSE}\left(\hat{\eta}_{MBGRSSj}^{(r)}\right)=\text{Var}\left(\hat{\eta}_{MBGRSSj}^{(r)}\right)+\left[\text{Bias}\left(\hat{\eta}_{MBGRSSj}^{(r)}\right)\right]^2,\ j=E,O.$$

Results of the efficiency and bias values are given in Tables 1-3 with $m=3,6,9$ for $r=1,2,3$.

Table 1. The efficiency for estimating the population median using RSS and MBGRSS for sample size $m = 3$ and $r = 1, 2, 3$

Distribution		RSS	MBGRSS		
			$r = 1$	$r = 2$	$r = 3$
Uniform (0,1)	*Eff*	1.454	1.454	1.974	3.795
Normal (0,1)	*Eff*	1.613	1.613	2.393	5.102
Beta (4,4)	*Eff*	1.582	1.582	2.303	4.797
Logistic (0,1)	*Eff*	1.718	1.718	2.507	5.523
Exponential (1)	*Eff*	1.787	1.787	2.734	6.212
	Bias	0.086	0.086	0.062	0.027
Gamma (2,1)	*Eff*	1.702	1.702	2.583	5.597
	Bias	0.087	0.087	0.060	0.029
Weibull (1,3)	*Eff*	1.815	1.815	2.726	6.383
	Bias	0.258	0.258	0.174	0.088

Table 2. The efficiency for estimating the population median using RSS and MBGRSS for sample size $m = 6$ and $r = 1, 2, 3$

Distribution		RSS	MBGRSS		
			$r = 1$	$r = 2$	$r = 3$
Uniform (0,1)	*Eff*	2.413	2.131	8.245	37.523
Normal (0,1)	*Eff*	2.730	2.352	9.370	43.084
Beta (4,4)	*Eff*	2.595	2.379	8.977	41.539
Logistic (0,1)	*Eff*	2.786	2.363	9.726	45.404
Exponential (1)	*Eff*	2.872	2.337	9.428	38.932
	Bias	0.054	0.074	0.042	0.032
Gamma (2,1)	*Eff*	2.807	2.447	9.349	41.001
	Bias	0.048	0.073	0.045	0.033
Weibull (1,3)	*Eff*	2.841	2.464	9.311	38.232
	Bias	0.147	0.218	0.130	0.096

Table 3. The efficiency for estimating the population median using RSS and MBGRSS for sample size $m = 9$ and $r = 1, 2, 3$

Distribution		RSS	MBGRSS		
			$r = 1$	$r = 2$	$r = 3$
Uniform (0,1)	*Eff*	2.981	2.343	11.621	71.212
Normal (0,1)	*Eff*	2.942	2.482	13.591	82.241
Beta (4,4)	*Eff*	2.772	2.331	13.230	80.107
Logistic (0,1)	*Eff*	2.882	2.479	13.830	85.209
Exponential (1)	*Eff*	3.002	2.478	14.924	90.870
	Bias	0.018	0.022	0.004	0.001
Gamma (2,1)	*Eff*	2.901	2.461	14.323	86.035
	Bias	0.019	0.022	0.004	0.000
Weibull (1,3)	*Eff*	3.046	2.503	14.767	91.905
	Bias	0.050	0.067	0.011	0.002

The results in Tables 1-3, indicate that:

1. Gain in efficiency is obtained using MBGRSS for estimating the population median. As an example, with $m = 9$ and $r = 3$ for estimating the median of the normal distribution, the efficiency of MBGRSSO is 82.241.
2. MBGRSS estimators are unbiased of the population median when the underlying distribution is symmetric about the population mean.
3. The efficiency of MBGRSS is increasing in r for specific value of the sample size. For example, for $m = 6$ and $r = 1,2,3$ the efficiency values are 2.131, 8.245 and 37.523 respectively for estimating the median of the uniform distribution.
4. The efficiency of MBGRSS estimators is increasing in the sample size for specific r. As an example, when the underlying distribution is exponential with parameter 1, for $r = 3$ and $m = 3,6,9$, the efficiency values are 6.212, 38.932 and 90.870 respectively.
5. When the underlying distribution is asymmetric, the small bias of MBGRSS is decreasing in r. For example, when the underlying distribution is gamma with parameters (2,1), for $m = 9$ and $r = 1,2,3$, the efficiency of MBGRSSO is 2.461, 14.323 and 86.035 with biases 0.022, 0.004 and 0.000, respectively.

5. An Application

In this section, we will investigate the efficiency of MBGRSS method for estimating the population median of 64 olive trees' yields in West Jordan. The data was actually collected via RSS in 1999. The chosen ranker visually judged the olives yield for each set size See Al-Omari (1999). Results are summarized in Tables 4–6.

Table 4. The efficiency of RSS with respect to SRS for estimating the median of olive yields with $m = 3, 6, 9$

	Bias		Efficiency
	SRS	RSS	
$m = 3$	1.26461	1.06925	1.44859
$m = 6$	1.06864	0.832194	2.27847
$m = 9$	0.812175	0.533728	3.79903

Table 5. The efficiency and bias values of MBGRSS with respect to SRS for estimating the median of olive yields with $m = 3$ for $r = 1, 2, 3, 4, 5$

	$r = 1$	$r = 2$	$r = 3$	$r = 4$	$r = 5$
Bias	1.06925	0.921435	0.657382	0.430258	0.266557
Efficiency	1.44859	2.06416	4.38985	11.9381	40.4151

Table 6. The efficiency and bias values of MBGRSS with respect to SRS for estimating the median of olive yields with $m = 6, 9$ for $r = 1, 2, 3$

	$r = 1$	$r = 2$	$r = 3$
	$m = 6$		
Bias	0.98927	0.764353	0.656441
Efficiency	2.0332	6.01133	13.6648
	$m = 9$		
Bias	0.0666892	0.208683	0.153532
Efficiency	2.50285	32.0339	97.2748

Let w_i, $i = 1, 2, ..., 64$ be the olive yield of the ith tree. The exact median η and variance σ^2 of the population are

$$\eta = 8.1,\ \sigma^2 = \frac{1}{64}\sum_{i=1}^{64}(w_i - \mu)^2 = 26.112\ \text{kg}^2 / \text{tree}.$$

Based on Tables 4 and 5, it can be noted that the MBGRSS is more efficient than the SRS and RSS methods based on the same number of measured units. Also, the efficiency of MBGRSS is increasing as the sample size increasing.

6. Conclusions

Based on the above results we can conclude that the properties of $\hat{\eta}^{(r)}_{MBGRSS}$ are:

1. If the underlying distribution is symmetric about the population mean μ, then
   a. $\hat{\eta}^{(r)}_{MBGRSS}$ is an unbiased estimator of the population median.
   b. $\text{Var}\left(\hat{\eta}^{(r)}_{MBGRSS}\right) < \text{Var}\left(\hat{\eta}_{SRS}\right)$, implies that MBGRSS is more efficient than SRS.
   c. MBGRSS is more efficient than RSS for $r > 1$, since $\text{Var}\left(\hat{\eta}^{(r)}_{MBGRSS}\right) < \text{Var}\left(\hat{\eta}_{RSS}\right)$ for $r > 1$.
2. If the underlying distribution is asymmetric about μ, then $\hat{\eta}^{(r)}_{MBGRSS}$ has a small bias, and the bias decreases in *r*.
3. The efficiency of MBGRSS estimators is increasing as the number of stage increasing.

REFERENCES

AL-OMARI, A.I. (1999). Multistage ranked set sampling, Master Thesis, Department of Statistics, Yarmouk University, Jordan

AL-OMARI, A.I. and JABER, K. (2008). Percentile double ranked set sampling, Journal of Mathematics and Statistics. 4 (1): 60–64.

AL-SALEH, M. F. and AL-OMARI, A. I. (2002). Multistage ranked set sampling, Journal of Statistical Planning and Inference. 102(2): 273–286.

DELL, T. R. and CLUTTER, J. L. (1972). Ranked set sampling theory with order statistic background. Biometrics, 28: 545–553.

JEMAIN, A. A. and AL-OMARI, A. I. (2006). Multistage median ranked set samples for estimating the population mean. Pakistan Journal of Statistics. 22(3): 195–207.

JEMAIN, A. A., AL-OMARI, A. I., and IBRAHIM, K. (2007). Multistage extreme ranked set sampling for estimating the population mean. Journal of Statistical Theory and Applications. 6(4): 456–471.

MCINTYRE, G. A. (1952). A method for unbiased selective sampling using ranked sets, Australian Journal of Agricultural Research. 3: 385–390.

MUTTLAK, H. A. (2003a). Investigating the use of quartile ranked set samples for estimating the population mean. Journal of Applied Mathematics and Computation. 146: 437–443.

MUTTLAK, H. A. (2003b) Modified ranked set sampling methods. Pakistan Journal of Statistics. 19(3): 315–323.

MUTTLAK, H. A. (1997). Median ranked set sampling, Journal of Applied Statistical sciences. 6(4): 245–255.

OZTURK, O. and DESHPANDE, J.V. (2006). Ranked-set sample nonparametric quantile confidence intervals. Journal of Statistical Planning and Inference. 136, 570–577.

SAMAWI, H, ABU-DAYYEH, W and AHMED, S. (1996). Extreme ranked set sampling, The Biometrical Journal. 30: 577–586.

TAKAHASI, K. and WAKIMOTO, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering, Annals of the Institute Statistical Mathematics. 20: 1–31.

ZHENG, G. and MODARRES, R. (2006). A robust estimate of the correlation coefficient for bivariate normal distribution using ranked set sampling. Journal of Statistical Planning and Inference. 136 (2006) 298–309

TIENSUWAN, M., SARIKAVANIJ, S. and SINHA, B.K. (2007). Nonnegative unbiased estimation of scale parameters and associated quantiles based on a ranked set sample. *Communications in Statistics-Simulation and Computation*, 36, 3–31.

TSENG, Y.L. and SHAO, S.W. (2007). Ranked-Set-Sample-Based Tests for Normal and Exponential Means. Communications in Statistics—Simulation and Computation, 36: 761–782.

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 235—250

DOUBLE STAGE SHRINKAGE TESTIMATION IN EXPONENTIAL TYPE-II CENSORED DATA

Gyan Prakash[1], **D. C. Singh**[2]

ABSTRACT

The present paper investigates the properties of the shrinkage testimators for mean and variance of an Exponential distribution in double stage samples, by using the cost function when Type – II censored data are available.

Key words: Type-II censored data; Shrinkage factor; Shrinkage testimator; Level of significance; Effective Interval; Cost function.

1. Introduction

The most widely used lifetime distribution is the Exponential distribution, having the probability density function for any random variable x

$$f(x;\lambda) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \quad ; \; x \geq 0, \; \lambda \geq 0. \tag{1.1}$$

Here, the parameter λ is called as the scale parameter, better known as the average life and λ^2 is called as the variance. The mean time to failure is λ and $\bar{x}$ is an unbiased estimate for the parameter λ. Further, the survival function is $S(x) = e^{-\lambda x}$ and failure rate is $\gamma = 1/\lambda$. When the failure rate is constant, the Exponential model has been found to be useful.

In life testing, fatigue failures and other kinds of destructive test situations, the observations usually occurred in ordered manner in such a way that weakest items failed first and then the second one and so on. Let $x_{1i} (i = 1, 2, ..., r_1)$ and $x_{2j} (j = 1, 2, ..., r_2)$ be the two independent random samples of size r_1 and r_2

[1] Department of Statistics, Harish Chandra P. G. College, Varanasi, U. P., India. E-Mail: ggyanji@yahoo.com
[2] Department of Statistics, Harish Chandra P. G. College, Varanasi, U. P., India.

respectively, drawn from an Exponential population with mean λ and variance λ^2. Epstien & Sobel (1953) proved that the total test time

$$T_1 = \sum_{i=1}^{r_1} x_{(1i)} + (n - r_1)\, x_{(1r_1)} \;\; ; n > r_1 \qquad (1.2)$$

is complete sufficient statistic for λ and r_1 is the uncensored sample size. The minimum variance unbiased (MVU) estimator of λ is $\frac{T_1}{r_1}$ and for the variance λ^2 is $\frac{T_1^2}{r_1(r_1+1)}$. Further, the minimum mean square error (MMSE) estimator for the mean and the variance are given respectively as

$$\frac{T_1}{r_1+1}$$

and

$$\frac{T_1^2}{(r_1+2)(r_1+3)}.$$

The unbiased pooled estimates for λ and λ^2 are respectively given as

$$T_M = \frac{T_1 + T_2}{r_1 + r_2}$$

and

$$T_V = \frac{T_1^2 + T_2^2}{r_1(r_1+1) + r_2(r_2+1)}.$$

It has been obtained that the estimator, by using the preliminary testimator, adaptive testimator, conditional testimator or shrinkage testimator, is superior to the existing estimator if the sample size and level of significance are small. Bancroft & Han (1977) and Han et al. (1988) described the inference based on conditional specification. Pandey (1983), Ebrahimi & Hosmane (1987), Pandey & Srivastava (1987) and Pandey (1997) described shrinkage estimation under Exponential data.

Many authors including Stein (1945), Weiss (1955) and Chapman (1960) have used two stage samples. These authors used the first sample to make inference on the unknown nuisance parameter and then used it in conjunction with the second to estimate the parameter of interest. Katti (1962) and Shah (1964) used a two stage estimate, but used the first stage to test the accuracy of

the prior information (guess estimate) and if it is accurate, used first stage sample to estimate the parameter, otherwise used it in conjunction with the second sample to estimate the parameter. Al–Bayyati & Arnold (1969, 70, 72) proposed the double stage shrinkage estimators for the parameter using the first stage sample to test the accuracy of the guess value and if it is accurate, they used shrinkage of the usual estimator towards guess value, otherwise used the first sample in conjunction with a new second sample to estimate the parameter. Further, Pandey (1979) introduced the preliminary test in constructing the acceptance region in double stage shrinkage estimation. Waiker et al. (1984) and Adke et al. (1987) proposed two stage shrinkage testimators depending upon the outcome of the preliminary test of hypothesis.

The proposed double stage technique is to obtain first a sample of size r_1 and compute the usual estimate for the parameter. If the usual estimate implies that our prior estimate or guess value of the parameter was reasonable, we stop the sampling and estimate the parameter by using the shrinkage estimator. Otherwise, we obtain r_2 additional observations and then use an improved estimate based on all $(r_1 + r_2)$ observations.

In the present paper, we study the performance of the shrinkage testimators in double stage samples under the invariant form of the LINEX loss function for the mean and variance of the Exponential distribution when Type – II censored data are available.

2. The proposed class of estimators

The proposed class of estimators for the parameters λ and λ^2 based on two samples of sizes r_1 and r_2 are given as respectively

$$P_M = l_1 T_M ;\ l_1 \in R^+ \tag{2.1}$$

and

$$P_V = l_2 T_V ;\ l_2 \in R^+. \tag{2.2}$$

The invariant form of the LINEX loss function for any parameter μ (Basu & Ebrahimi, 1992) is given as

$$L(\Delta) = e^{a\Delta} - a\Delta - 1 ;\ a \neq 0,\ \Delta = \left(\frac{\hat{\mu}}{\mu} - 1\right), \tag{2.3}$$

where 'a' is the shape parameter and $\hat{\mu}$ is an estimate of the parameter μ.

The LINEX loss function is convex and its shape is to be determined by the value of 'a' (the sign of 'a' reflects the direction of asymmetry, $a > 0\ (a < 0)$ if

over estimation is more (less) serious than under estimation) and its magnitude reflects the degree of asymmetry. The LINEX loss function has been found to be appropriate in the situations where overestimation is more serious than underestimation and vice versa. The LINEX loss criterion will change to mean square error criterion if $|a| \to 0$.

The value of constant l_1 for which the risk of P_M is minimum under the LINEX loss function (2.3) is

$$\hat{l}_1 = \frac{r_1 + r_2}{a}\left(1 - e^{-a/r_1+r_2+1}\right).$$

Similarly, the value of $l_2 = \hat{l}_2$ (say) for which the risk of P_V is minimum under the loss (2.3), is obtained by solving the given equality

$$e^{a}\left(\Gamma(r_1+1)\Gamma r_2 + \Gamma r_1 \Gamma(r_2+1)\right) = I\left(0,\infty,0,\infty,e^{a\Delta_V}\right),$$

$$\text{where } \Delta_V = \frac{l_2\left(x^2+y^2\right)}{r_1(r_1+1)+r_2(r_2+1)},$$

$$I(x_1,x_2,y_1,y_2,w) = \int_{x_1}^{x_2}\int_{y_1}^{y_2} \frac{(w)e^{-x}e^{-y}}{\Gamma r_1 \Gamma r_2}$$

$x^{r_1-1}\, y^{r_2-1}\, dx\, dy$ and w may be the function of x and y.

Hence, the improved class of estimators for the parameters λ and λ^2 are

$$\hat{P}_M = \hat{l}_1 T_M$$

and

$$\hat{P}_V = \hat{l}_2 T_V$$

with the risk

$$R\left(\hat{P}_M\right) = (1+r_1+r_2)\left(e^{-a/1+r_1+r_2}-1\right) + a$$

and

$$R\left(\hat{P}_V\right) = I\left(0,\infty,0,\infty,e^{a\Delta_{V0}}\right) - a\, I\left(0,\infty,0,\infty,\Delta_{V0}\right) - 1,$$

where $\Delta_{V0} = \dfrac{\hat{l}_2\left(x^2+y^2\right)}{r_1(r_1+1)+r_2(r_2+1)}$.

3. The proposed shrinkage estimators

The shrinkage estimator (Thompson, 1968) for the mean λ and variance λ^2 based on the sample of size r_1, when prior point estimate of the parameters is available, are proposed:

$$T_{SM} = \lambda_0 + k\,(T_1 - \lambda_0 r_1)\, r_1^{-1} \tag{3.1}$$

and

$$T_{SV} = \lambda_0^2 + k\left(\frac{T_1^2}{r_1(r_1+1)} - \lambda_0^2\right);\ 0 \le k \le 1. \tag{3.2}$$

A value of shrinkage factor k near to zero implies strong belief in the guess value (prior point value) and near to one implies strong belief in the sample values. The risks of these shrinkage estimator under the loss (2.3) are obtained as

$$R(T_{SM}) = e^{a(\delta-1)} e^{-ak\delta}\left(1 - a\,k\,r_1^{-1}\right)^{-r_1} - a(\delta-1)(1-k) - 1$$

and

$$R(T_{SV}) = e^{a(\delta^2-1)} G\left(0, \infty, e^{a\Delta_{V1}}\right) - a(\delta^2-1)(1-k) - 1,$$

where $\Delta_{V1} = k\left(\frac{x^2}{r_1(r_1+1)} - \delta^2\right)$, $G(x_1, x_2, u) = \int_{x_1}^{x_2} \frac{u}{\Gamma r_1} e^{-x} x^{r_1-1}\, dx$,

$\delta = \frac{\lambda_0}{\lambda}$ and u may be the function of x.

4. Shrinkage testimators for the mean and their properties

Several researchers have studied the performance of the shrinkage estimators and found that the shrinkage estimator performs better with respect to any usual estimator or improved estimator when the guess value of the parameter is approximately equal to the true value of the parameter and the sample size is small. This implies that we may test the hypothesis $H_0: \delta = 1$ against $H_1: \delta \neq 1$. A test statistic $(2T_1/\lambda) \sim \chi^2_{(2r_1)}$ is available for testing the hypothesis H_0 against H_1. The proposed double stage shrinkage testimator for the mean λ is defined as

$$\hat{\lambda}_1 = \begin{cases} T_{SM} & \text{if } (t_1 \le T_1 \le t_2) \\ \hat{P}_M & \text{else} \end{cases},$$

where $t_i = \frac{m_i \lambda_0}{2}; i = 1, 2$. Here m_1 and m_2 are the values of the lower and upper $100\frac{\alpha}{2}\%$ points of the chi – square distribution with $2r_1$ degrees of freedom.

The relative bias of the shrinkage testimator $\hat{\lambda}_1$ is obtained as

$$RB(\hat{\lambda}_1) = E(\hat{\lambda}_1)\lambda^{-1} - 1$$

$$= G(x_1, x_2, (\Delta_{M1} + \delta)) - I(x_1, x_2, 0, \infty, \Delta_{M0}) + \hat{l}_1 - 1, \qquad (4.1)$$

where $\Delta_{M1} = k(x - \delta r_1) r_1^{-1}$, $\Delta_{M0} = \hat{l}_1 (r_1 + r_2)^{-1}(x + y)$.

The risk under the LINEX loss (2.3) for $\hat{\lambda}_1$ is

$$R(\hat{\lambda}_1) = e^{-a(\delta-1)} G(x_1, x_2, e^{a\Delta_{M1}}) - a\, G(x_1, x_2, (\Delta_{M1} + \delta))$$
$$- e^{-a} I(x_1, x_2, 0, \infty, e^{a\Delta_{M0}}) + a\, I(x_1, x_2, 0, \infty, \Delta_{M0})$$
$$+ (1 + r_1 + r_2)(e^{-a/1+r_1+r_2} - 1) + a. \qquad (4.2)$$

Our problem is considered as a sequential estimation problem with stopping random variable R defined as

$$R = \begin{cases} r_1 & \text{if } t_1 \le T_1 \le t_2 \\ r_1 + r_2 & \text{otherwise} \end{cases}. \qquad (4.3)$$

Let us introduce a cost $c(>0)$ also, for each observation. Then the risk of $\hat{\lambda}_1$ and $\hat{P}_M$ are

$$\tilde{R}(\hat{\lambda}_1) = R(\hat{\lambda}_1) + c\,E(R)$$

and

$$\tilde{R}(\hat{P}_M) = R(\hat{P}_M) + c(r_1 + r_2).$$

Therefore, the relative efficiency of $\hat{\lambda}_1$ with respect to $\hat{P}_M$ is given by

$$RE(\hat{\lambda}_1, \hat{P}_M) = \tilde{R}(\hat{P}_M) / \tilde{R}(\hat{\lambda}_1).$$

The expressions of the relative bias and relative efficiency are the functions of r_1, r_2, δ, a, c, k and α (level of significance). For the selected values of $r_1, r_2 =$ 03, 05, 07; $\delta = 0.25(0.25)1.75$; $a = 0.25, 0.50, 1.00$; $k = 0.25, 0.50, 0.75$; $c = 0.50$, 05, 10, 50 and $\alpha = 0.01, 0.05, 0.20$, the relative biases (not presented here), and the relative efficiency have been calculated. However, the relative efficiencies are presented only for $r_1 = 03, c = 0.50, \alpha = 0.01$ in Table 01.

We observe that the relative biases are small and lie between -0.140 and 0.281. The biases are negative for $\delta < 1.00$ and positive otherwise. The value of biases decreases as a, k or α increases when $\delta \geq 1.00$. The biases decrease as r_2 increases in the interval $0.50 \leq \delta \leq 1.25$ and increases as r_1 increases for all considered values of δ.

The testimator $\hat{\lambda}_1$ performs well with respect to $\hat{P}_M$ in the effective interval $0.50 \leq \delta \leq 1.25$ and the effective interval decreases with increase in r_1 and attains maximum efficiency at the point $\delta = 1.00$. The efficiency increases as r_2 increases for all considered values of δ but it decreases as r_1 or 'a' increases except for $\delta = 1.00$. The efficiency also increases as per unit cost c increases in the interval $0.75 \leq \delta \leq 1.25$ for small k, r_1 and α, and decreases otherwise. However, the gain in efficiency is nominal. In addition, as level of significance α increases, the efficiency decreases.

It may interest one to obtain the value of the shrinkage factor k that minimizes the risk of $\hat{\lambda}_1$ (4.2). For this the minimum value of $k = \hat{k}_M$ (say) is obtained numerically by solving the given equality

$$e^{-a(\delta-1)}\, G\left(x_1, x_2, \left(\Delta_{M1}\, e^{a\Delta_{M1}}\right)\right) = G\left(x_1, x_2, \Delta_{M1}\right). \qquad (4.4)$$

Hence, the proposed double stage shrinkage testimator for the mean is

$$\hat{\lambda}_2 = \begin{cases} \lambda_0 + \hat{k}_M \left(T_1 - \lambda_0 r_1\right) r_1^{-1} & \text{if } \left(t_1 \leq T_1 \leq t_2\right) \\ \hat{P}_M & \text{else} \end{cases}.$$

Thus, the expression of the relative bias and risk under the LINEX loss are obtained as

Table 1. RE $\left(\hat{\lambda}_1,\ \hat{P}_M\right)$

$\alpha = 0.01, c = 0.50$									
$k = 0.25$			δ						
r_1	r_2	a	0.25	0.50	0.75	1.00	1.25	1.50	1.75
03	03	0.25	1.1204	1.9647	2.3878	4.5153	2.1638	2.1749	2.1641
		0.50	1.1154	1.9525	2.3786	4.5382	2.1561	2.1467	2.0846
		1.00	1.0995	1.9134	2.3639	4.6256	2.1493	2.0262	1.7861
	05	0.25	1.1558	2.2141	2.9266	5.8443	2.8566	2.8894	2.8800
		0.50	1.1510	2.1987	2.9209	5.8574	2.8509	2.8424	2.7647
		1.00	1.1350	2.1469	2.9019	5.9074	2.8226	2.6501	2.3386
	07	0.25	1.1783	2.3974	3.3983	7.1011	3.5378	3.6008	3.5955
		0.50	1.1739	2.3811	3.3887	7.1076	3.5255	3.5366	3.4458
		1.00	1.1592	2.3255	3.3549	7.1320	3.4711	3.2775	2.8963
$k = 0.50$									
03	03	0.25	1.1206	1.9654	2.3766	4.5081	2.1609	2.1767	2.1747
		0.50	1.1161	1.9553	2.3740	4.5097	2.1540	2.1527	2.1255
		1.00	1.1018	1.9231	2.3679	4.5117	2.1164	2.0407	1.9149
	05	0.25	1.1560	2.2148	2.9252	5.8353	2.8528	2.8917	2.8942
		0.50	1.1516	2.2014	2.9157	5.8216	2.8350	2.8504	2.8190
		1.00	1.1369	2.1563	2.8838	5.7654	2.7541	2.6690	2.5073
	07	0.25	1.1785	2.3980	3.3968	7.0905	3.5331	3.6036	3.6132
		0.50	1.1744	2.3836	3.3831	7.0652	3.5060	3.5466	3.5134
		1.00	1.1608	2.3343	3.3354	6.9649	3.3875	3.3009	3.1051
$k = 0.75$									
03	03	0.25	1.1208	1.9658	2.3741	4.4963	2.1541	2.1719	2.1749
		0.50	1.1168	1.9568	2.3644	4.4626	2.1264	2.1326	2.1249
		1.00	1.1040	1.9290	2.3338	4.3301	2.0063	1.9562	1.8990
	05	0.25	1.1561	2.2152	2.9224	5.8204	2.8439	2.8853	2.8944
		0.50	1.1521	2.2028	2.9048	5.7624	2.7989	2.8238	2.8181
		1.00	1.1387	2.1619	2.8452	5.5388	2.6117	2.5587	2.4864
	07	0.25	1.1786	2.3983	3.3937	7.0728	3.5222	3.5957	3.6134
		0.50	1.1749	2.3850	3.3713	6.9954	3.4618	3.5135	3.5123
		1.00	1.1623	2.3396	3.2936	6.6976	3.2135	3.1647	3.0793

$$\mathrm{RB}\left(\hat{\lambda}_2\right) = \mathrm{G}\left(\mathrm{x}_1, \mathrm{x}_2, \left(\Delta_{\mathrm{M2}} + \delta\right)\right) - \mathrm{I}\left(\mathrm{x}_1, \mathrm{x}_2, 0, \infty, \Delta_{\mathrm{M0}}\right) + \hat{l}_1 - 1$$

and

$$\mathrm{R}\left(\hat{\lambda}_2\right) = \mathrm{e}^{-\mathrm{a}(\delta-1)}\,\mathrm{G}\left(\mathrm{x}_1, \mathrm{x}_2, \mathrm{e}^{\mathrm{a}\Delta_{\mathrm{M2}}}\right) - \mathrm{a}\,\mathrm{G}\left(\mathrm{x}_1, \mathrm{x}_2, \left(\Delta_{\mathrm{M2}} + \delta\right)\right)$$
$$-\mathrm{e}^{-\mathrm{a}}\,\mathrm{I}\left(\mathrm{x}_1, \mathrm{x}_2, 0, \infty, \mathrm{e}^{\mathrm{a}\Delta_{\mathrm{M0}}}\right) + \mathrm{a}\,\mathrm{I}\left(\mathrm{x}_1, \mathrm{x}_2, 0, \infty, \Delta_{\mathrm{M0}}\right)$$
$$+ \left(1 + \mathrm{r}_1 + \mathrm{r}_2\right)\left(\mathrm{e}^{-\mathrm{a}/1+\mathrm{r}_1+\mathrm{r}_2} - 1\right) + \mathrm{a}\ ;\ \Delta_{\mathrm{M2}} = \hat{\mathrm{k}}_{\mathrm{M}}\left(\mathrm{x} - \delta\,\mathrm{r}_1\right)\mathrm{r}_1^{-1}.$$

The relative efficiency of $\hat{\lambda}_2$ with respect to $\hat{\mathrm{P}}_{\mathrm{M}}$ is defined as

$$\mathrm{RE}\left(\hat{\lambda}_2, \hat{\mathrm{P}}_{\mathrm{M}}\right) = \widetilde{\mathrm{R}}\left(\hat{\mathrm{P}}_{\mathrm{M}}\right) / \widetilde{\mathrm{R}}\left(\hat{\lambda}_2\right).$$

The expressions of relative bias and the relative efficiency are the functions of the r_1, r_2, δ, a, c and α. For the similar set of selected values, the relative bias (not presented here) and the relative efficiency (presented in Table 02 for $\mathrm{r}_1 = 03, \mathrm{c} = 0.50, \alpha = 0.01$ only) have been calculated.

The relative biases are small and lie between -0.141 and 0.320. The bias increases as r_1 increases for $\delta \le 0.75$. The shrinkage testimator $\hat{\lambda}_2$ is more efficient than $\hat{\mathrm{P}}_{\mathrm{M}}$ in the effective interval $0.50 \le \delta \le 1.50$. Other properties of the shrinkage testimator $\hat{\lambda}_2$ related to the bias and the relative efficiency are similar to $\hat{\lambda}_1$. Based on the gain in relative efficiency, the shrinkage testimator $\hat{\lambda}_2$ is preferred over $\hat{\lambda}_1$ in the interval $0.50 \le \delta \le 1.50$.

5. Shrinkage testimators for the variance and their properties

The proposed double stage shrinkage testimator for the variance λ^2 is defined as

$$\hat{\lambda}_3 = \begin{cases} \mathrm{T}_{\mathrm{SV}} & \text{if } \left(\mathrm{t}_1 \le \mathrm{T}_1 \le \mathrm{t}_2\right) \\ \hat{\mathrm{P}}_{\mathrm{V}} & \text{else} \end{cases}.$$

The expressions of the relative bias and risk under the LINEX loss for the proposed double stage shrinkage testimator $\hat{\lambda}_3$ are obtained as

$$\mathrm{RB}\left(\hat{\lambda}_3\right) = \mathrm{G}\left(\mathrm{x}_1, \mathrm{x}_2, \left(\Delta_{\mathrm{V1}} + \delta^2\right)\right) - \mathrm{I}\left(\mathrm{x}_1, \mathrm{x}_2, 0, \infty, \Delta_{\mathrm{V0}}\right) + \hat{l}_2 - 1$$

and

$$R\left(\hat{\lambda}_3\right) = e^{-a\left(\delta^2-1\right)} G\left(x_1, x_2, e^{a\Delta_{V1}}\right) - a\, G\left(x_1, x_2, \left(\Delta_{V1} + \delta^2\right)\right)$$
$$- e^{-a}\, I\left(x_1, x_2, 0, \infty, e^{a\Delta_{V0}}\right) + a\, I\left(x_1, x_2, 0, \infty, \Delta_{V0}\right)$$
$$+ e^{-a}\, I\left(0, \infty, 0, \infty, e^{a\Delta_{V0}}\right) + a\left(1 - \hat{l}_2\right) - 1. \qquad (5.1)$$

Now, the risks under the sequential estimation problem with per unit cost $c\,(>0)$ are given as

$$\tilde{R}\left(\hat{\lambda}_3\right) = R\left(\hat{\lambda}_3\right) + c\,E(R)$$

Table 2. $RE\left(\hat{\lambda}_2,\ \hat{P}_M\right)$

$\alpha = 0.01$ $c = 0.50$									
			δ						
r_1	r_2	a	0.25	0.50	0.75	1.00	1.25	1.50	1.75
03	03	0.25	1.1205	2.1883	2.6471	5.0275	2.4012	2.4002	2.1368
		0.50	1.1158	2.1741	2.6433	5.0465	2.3737	2.2894	1.9497
		1.00	1.1010	2.1287	2.6328	5.1163	2.2367	1.7896	1.1065
	05	0.25	1.1559	2.4662	3.2582	6.5074	3.1702	3.1887	2.8438
		0.50	1.1514	2.4484	3.2466	6.5139	3.1245	3.0317	2.5860
		1.00	1.1363	2.3889	3.2068	6.5351	2.9116	2.3412	1.4489
	07	0.25	1.1784	2.6702	3.7834	7.9068	3.9263	3.9738	3.5503
		0.50	1.1743	2.6516	3.7671	7.9045	3.8645	3.7726	3.2231
		1.00	1.1603	2.5879	3.7094	7.8913	3.5825	2.8966	1.7946
05	03	0.25	1.0207	1.7685	2.0997	4.0534	1.9123	1.8461	1.5947
		0.50	1.0135	1.7311	2.0923	4.1851	1.8601	1.5850	1.2338
		1.00	0.9957	1.6429	2.0857	4.6473	1.6676	1.3321	0.6149
	05	0.25	1.0286	1.8787	2.4058	4.8907	2.3563	2.2881	1.9790
		0.50	1.0207	1.8291	2.3705	4.9772	2.2553	1.9293	1.5023
		1.00	0.9989	1.6990	2.2756	5.2884	1.9201	1.3819	0.7053
	07	0.25	1.0342	1.9621	2.6690	5.6836	2.7951	2.7303	2.3647
		0.50	1.0264	1.9086	2.6164	5.7377	2.6512	2.2787	1.7757
		1.00	1.0038	1.7592	2.4635	5.9364	2.1874	1.4326	0.8035
07	03	0.25	0.9924	1.5275	1.8722	4.2518	1.6202	1.2577	0.7578
		0.50	0.9847	1.4370	1.8619	5.4694	1.4852	1.2272	0.4461
		1.00	0.9708	1.4037	1.8618	7.0728	1.3231	1.1383	0.2929
	05	0.25	0.9937	1.5376	1.9383	4.6493	1.7910	1.2734	0.8327
		0.50	0.9886	1.4961	1.8790	5.6253	1.4925	1.2422	0.4390
		1.00	0.9793	1.4093	1.8715	7.2772	1.3478	1.2267	0.2632
	07	0.25	0.9962	1.5576	2.0320	5.0534	1.9745	1.3006	0.9164
		0.50	0.9953	1.5392	1.8863	5.8119	1.5244	1.2954	0.4426
		1.00	0.9804	1.5370	1.8717	7.6954	1.3776	1.2417	0.2433

and

$$\widetilde{R}\left(\hat{P}_V\right) = R\left(\hat{P}_V\right) + c\,(r_1 + r_2)$$

The relative efficiencies of $\hat{\lambda}_3$ with respect to $\hat{P}_V$ is defined as

$$RE\left(\hat{\lambda}_3, \hat{P}_V\right) = \widetilde{R}\left(\hat{P}_V\right) / \widetilde{R}\left(\hat{\lambda}_3\right).$$

For the similar set of selected values as considered in previous section, the relative bias (not presented here) and the relative efficiency have been calculated. The relative efficiencies are presented only for $r_1 = 03, c = 0.50, \alpha = 0.01$ in the Table 03.

Table 3. Table 03 :: RE $\left(\hat{\lambda}_3,\ \hat{P}_V\right)$

$\alpha = 0.01, c = 0.50$									
$k = 0.25$			δ						
r_1	r_2	a	0.25	0.50	0.75	1.00	1.25	1.50	1.75
03	03	0.25	1.1210	1.9652	2.4789	4.5201	2.1606	2.1465	2.0756
		0.50	1.1178	1.9552	2.3829	4.5584	2.1516	2.0258	1.7409
		1.00	1.1109	1.9527	2.3787	4.9161	2.1198	1.6419	1.1344
	05	0.25	1.1563	2.2142	2.9269	5.8497	2.8522	2.8512	2.7616
		0.50	1.1529	2.1996	2.9230	5.8792	2.8307	2.6810	2.3069
		1.00	1.1423	2.1549	2.9148	5.9953	2.7237	2.0469	1.1672
	07	0.25	1.1787	2.3971	3.3979	7.1062	3.5317	3.5526	3.4470
		0.50	1.1755	2.3807	3.3880	7.1276	3.4983	3.3338	2.8731
		1.00	1.1652	2.3285	3.3566	7.2092	3.3401	2.5249	1.4411
$k = 0.50$									
03	03	0.25	1.1211	1.9657	2.4649	4.5103	2.1568	2.1577	2.1257
		0.50	1.1180	1.9569	2.3787	4.5181	2.1320	2.0599	1.9022
		1.00	1.1114	1.9384	2.3775	4.7376	2.0734	1.6366	1.0795
	05	0.25	1.1564	2.2146	2.9256	5.8374	2.8471	2.8660	2.8283
		0.50	1.1531	2.2013	2.9182	5.8287	2.8051	2.7260	2.5206
		1.00	1.1428	2.1604	2.8993	5.7827	2.5706	2.0403	1.3483
	07	0.25	1.1788	2.3975	3.3965	7.0916	3.5255	3.5711	3.5302
		0.50	1.1756	2.3823	3.3828	7.0680	3.4669	3.3897	3.1393
		1.00	1.1656	2.3337	3.3400	6.9600	3.1535	2.5168	1.6648
$k = 0.75$									
03	03	0.25	1.1211	1.9660	2.4356	4.4938	2.1448	2.1504	2.1384
		0.50	1.1182	1.9582	2.3753	4.4500	2.0763	2.0096	1.9066
		1.00	1.1120	1.9228	2.3704	4.4347	1.7768	1.2574	1.1751
	05	0.25	1.1564	2.2150	2.9231	5.8166	2.8315	2.8564	2.8451
		0.50	1.1532	2.2025	2.9087	5.7431	2.7324	2.6596	2.5265
		1.00	1.1433	2.1647	2.8668	5.4211	2.2054	1.5685	1.2809
	07	0.25	1.1788	2.3978	3.3937	7.0670	3.5063	3.5591	3.5512
		0.50	1.1757	2.3834	3.3726	6.9669	3.3777	3.3072	3.1466
		1.00	1.1660	2.3378	3.3051	6.5351	2.7080	1.9352	1.3876

We observed that the relative biases are negligibly small and lie between -0.104 and 0.170. The biases are positive for $\delta \geq 1.00$ and negative otherwise. The values of bias decrease (increase) as 'a' (r_1) increases in the interval $\delta \leq 1.50$. The biases also decrease as level of significance increases in the interval $\delta \geq 0.75$. The bias decreases when the shrinkage factor k increases in the range $\delta \geq 1.00$ for small $\alpha\,(=0.01)$ and in the range $0.25 \leq \delta \leq 1.75$ for large α.

The shrinkage testimator $\hat{\lambda}_3$ is more efficient than $\hat{P}_V$ in the interval $0.25 \le \delta \le 1.50$ and the effective interval decreases with the shrinkage factor k or per unit cost c increase. The efficiency attains maximum at the point $\delta = 1.00$. The efficiency increases (decreases) as r_2 (r_1) increases for all considered values of δ. The decreasing trend in the efficiency is also seen when α increases when $\delta \le 1.25$ and when 'a' increase for all δ (except for $\delta = 1.00$). Around $\delta = 1.00,$ the efficiency also decreases with the shrinkage factor k increase. The nominal gain in efficiency is recorded when per unit cost c increases for all considered values of parametric space when $0.75 \le \delta \le 1.25$.

The value of $k = \hat{k}_V$ (say) that minimizes the risk of $\hat{\lambda}_3$ (5.1) can be obtained numerically by solving the given equality

$$e^{-a\left(\delta^2-1\right)} G\left(x_1, x_2, \left(\Delta_{V1}\, e^{a\Delta_{V1}}\right)\right) = G\left(x_1, x_2, \Delta_{V1}\right). \qquad (5.2)$$

Hence, the improved double stage shrinkage testimator of the variance is

$$\hat{\lambda}_4 = \begin{cases} \lambda_0^2 + \hat{k}_V \dfrac{\left(T_1^2 - \lambda_0^2\, r_1(r_1+1)\right)}{r_1(r_1+1)} & \text{if } \left(t_1 \le T_1 \le t_2\right) \\ \hat{P}_V & \text{else} \end{cases}$$

with relative bias

$$RB\left(\hat{\lambda}_4\right) = G\left(x_1, x_2, \left(\Delta_{V2} + \delta^2\right)\right) - I\left(x_1, x_2, 0, \infty, \Delta_{V0}\right) + \hat{l}_2 - 1$$

and the risk

$$\begin{aligned} R\left(\hat{\lambda}_4\right) = {} & e^{-a\left(\delta^2-1\right)} G\left(x_1, x_2, e^{a\Delta_{V2}}\right) - a\, G\left(x_1, x_2, \left(\Delta_{V2} + \delta^2\right)\right) \\ & - e^{-a}\, I\left(x_1, x_2, 0, \infty, e^{a\Delta_{V0}}\right) + a\, I\left(x_1, x_2, 0, \infty, \Delta_{V0}\right) \\ & + e^{-a}\, I\left(0, \infty, 0, \infty, e^{a\Delta_{V0}}\right) + a\left(1 - \hat{l}_2\right) - 1, \end{aligned} \qquad (5.2)$$

where $\Delta_{V2} = \hat{k}_V \left(\dfrac{x^2}{r_1(r_1+1)} - \delta^2 \right)$.

The relative efficiency of $\hat{\lambda}_4$ with respect to $\hat{P}_V$ is

$$RE\left(\hat{\lambda}_4, \hat{P}_V\right) = \tilde{R}\left(\hat{P}_V\right) / \tilde{R}\left(\hat{\lambda}_4\right).$$

For the similar set of selected values, the relative bias (not presented here) and the relative efficiency (presented in Table 04 only for $r_1 = 03, c = 0.50, \alpha$

$= 0.01$) have been calculated. The relative biases are negligible small and lie between -0.105 and 0.178. The biases are positive for $\delta > 1.00$ and negative for $\delta < 1.00$. Other properties are similar to the double stage shrinkage testimator $\hat{\lambda}_3$.

The shrinkage testimator $\hat{\lambda}_4$ is more efficient than $\hat{P}_V$ for all considered values of the parametric space (Table 04). Other properties are similar to the double stage shrinkage testimator $\hat{\lambda}_3$ except when r_1 increases, the efficiency decreases when $\delta \leq 1.25$.

Table 4. $\mathrm{RE}\left(\hat{\lambda}_4, \hat{P}_V\right)$

$\alpha = 0.01$ $c = 0.50$									
			δ						
r_1	r_2	a	0.25	0.50	0.75	1.00	1.25	1.50	1.75
03	03	0.25	1.1210	1.9650	2.3783	5.5171	3.1376	2.5920	1.8252
		0.50	1.1179	1.9546	2.3775	5.5453	3.0334	2.4312	1.1778
		1.00	1.1112	1.9510	2.1597	5.8492	2.6652	1.0545	1.0000
	05	0.25	1.1564	2.2140	2.9253	6.8460	3.8220	2.7355	2.4286
		0.50	1.1530	2.1991	2.9169	6.8628	3.6762	2.4949	1.2356
		1.00	1.1426	2.1532	2.8935	6.9156	2.9809	1.0681	1.0000
	07	0.25	1.1788	2.3969	3.3961	8.1017	4.4947	3.4087	3.0316
		0.50	1.1756	2.3802	3.3814	8.1082	4.3088	2.5760	1.2936
		1.00	1.1654	2.3269	3.3337	8.1159	3.2380	1.0840	1.0000
05	03	0.25	1.0235	1.6007	1.9886	4.6037	2.7259	2.7178	1.6847
		0.50	1.0231	1.5986	1.9020	4.6222	2.7064	2.5983	1.3338
		1.00	1.0219	1.5932	1.9005	4.6997	2.5991	1.9100	1.0502
	05	0.25	1.0315	1.7027	2.1712	5.3684	3.1379	3.1413	2.1032
		0.50	1.0310	1.6990	2.1692	5.3823	3.1099	2.9883	1.6612
		1.00	1.0295	1.6876	2.1651	5.4393	2.9626	2.1225	1.0619
	07	0.25	1.0370	1.7784	2.4124	6.0896	3.5433	3.5634	2.5216
		0.50	1.0365	1.7741	2.4084	6.1011	3.5079	3.3779	1.9892
		1.00	1.0349	1.7603	2.3962	6.1476	3.1252	2.3370	1.0737
07	03	0.25	1.0046	1.4656	1.6869	4.2037	2.5430	2.5470	1.5368
		0.50	1.0045	1.4651	1.6785	4.2155	2.5360	2.4962	1.4250
		1.00	1.0043	1.4640	1.6263	4.2656	2.4998	2.2331	1.1683
	05	0.25	1.0063	1.5126	1.8455	4.7199	2.8346	2.8520	1.8425
		0.50	1.0063	1.5115	1.8451	4.7287	2.8240	2.7887	1.7058
		1.00	1.0060	1.5080	1.8444	4.7655	2.7720	2.4663	1.6096
	07	0.25	1.0076	1.5482	1.9886	5.2046	3.1214	2.9156	2.1482
		0.50	1.0075	1.5467	1.9871	5.2118	3.1080	2.8808	1.9870
		1.00	1.0072	1.5420	1.9828	5.2415	3.0431	2.4711	1.9398

Based on the magnitude of the relative efficiency, one may prefer the shrinkage testimator $\hat{\lambda}_4$ in the interval $1.00 \leq \delta \leq 1.50$ and $\hat{\lambda}_3$ otherwise.

REFERENCES

ADKE, S. R., WAIKAR, V. B. and SCHUURMANN, F. J. (1987). A two stage shrinkage testimator for the mean of an exponential distribution. Communication in Statistics – Theory and Methods, 16, 1821–1834.

AL-BAYYATI, H. A. and ARNOLD, J. C. (1969). Double stage shrunken estimator of the variance. Ind. Statist. Theory Meth. Assoc., 7, 175–184.

AL-BAYYATI, H. A. and ARNOLD, J. C. (1972). On double stage estimation in simple liner regression using prior knowledge. Technometrices, 14, 405–414.

AL-BAYYATI, H. A. and ARNOLD, J. C. (1970). On double stage estimation of mean using prior knowledge. Biometrics, 26, 787–800.

BANCROFT, T. A. and HAN, C. P. (1977). Inference based on conditional specification. A note and a bibliography. International Statistical Review, 15, 117–127.

BASU, A. P. and EBRAHIMI, N. (1991). Bayesian approach to life testing and reliability estimation using asymmetric loss function. Journal of Statistical Planning and Inferences, 29, 21–31.

CHAPMAN, D. G. (1960). Some two sample test. Annals of Mathematical Statistics, 21, 601–606.

EBRAHIMI, N. and HOSMANE, B. (1987). On shrinkage estimation of the Exponential parameter. Communications in Statistics –Theory and Methods, 16, 2623–2637.

EPSTEIN, B. and SOBEL, M. (1953). Life Testing. Journal of American Statistical Association, 48, 486–507.

HAN, C. P., RAO, C. V. and RAVICHANDRAN, J. (1988). Inference based on the condition specification: A second bibliography. Communications in Statistics –Theory and Methods, A–17, 1–21.

KATTI, S. K. (1962). Use of some apriori knowledge in the estimation of mean from double samples. Biometrics, 18, 139–147.

PANDEY, B. N. (1979). Double stage estimation of population variance. Annals of Institute of Statistics Mathematical, 31, 225–233.

PANDEY, B. N. (1983). Shrinkage estimation of the Exponential scale parameter. IEEE Transaction on Reliability, R–32, 203–205.

PANDEY, B. N. (1997). Testimator of the scale parameter of the Exponential distribution using LINEX loss function. Communication in Statistics – Theory and Methods, 26, 2191–2200.

PANDEY, B. N and SRIVASTAVA, R. (1987). A shrinkage testimator for scale parameter of an exponential distribution. Microelectron Reliability, 27 (16), 949–951.

SHAH, S. M. (1964). Use a prior knowledge in the estimation of a parameter from double samples. Journal of Indian Statistical Association, 2, 41–51.

STEIN, C. (1945). A two-stage sample test for a linear hypothesis whose power is independent of the variance. Annals of Mathematical Statistics, 16, 243–258.

THOMPSON, J. R. (1968) Some Shrunken techniques for estimating the Mean. Journal of the American Statistical Association, 63, 113–122.

WAIKAR, V. B., SCHUURMANN, F. J. and RAGHUNATHAN, T. E. (1984). On a two stage shrunken testimator of the mean of a normal distribution. Communications in Statistics – Theory and Methods, 13, 1901–1913.

WEISS, L. (1945). On confidence intervals of given length for the mean of a Normal distribution with unknown variance. Annals of Mathematical Statistics, 16, 348–352.

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 251—267

SPATIAL M-QUANTILE MODELS FOR SMALL AREA ESTIMATION

Nicola Salvati[1], Monica Pratesi[1], Nikos Tzavidis[2], Ray Chambers[3]

ABSTRACT

In small area estimation direct survey estimates that rely only on area-specific data can exhibit large sampling variability due to small sample sizes at the small area level. Efficient small area estimates can be constructed using explicit linking models that borrow information from related areas. The most popular class of models for this purpose are models that include random area effects. Estimation for these models typically assumes that the random area effects are uncorrelated. In many situations, however, it is reasonable to assume that the effects of neighbouring areas are correlated. Models that extend conventional random effects models to account for spatial correlation between the small areas have been recently proposed in literature. A new semi-parametric approach to small area estimation is based on the use of M-quantile models. Unlike traditional random effects models, M-quantile models do not depend on strong distributional assumptions and are robust to the presence of outliers. In its current form, however, the M-quantile approach to small area estimation does not allow for spatially correlated area effects. The aim of this paper is to extend the M-quantile approach to account for such spatial correlation between small areas.

Key words: Quantile regression, Robust models, Spatial correlation, Weighted least squares

1. Introduction

In small area estimation direct survey estimates that rely only on area-specific data can exhibit large sampling variability due to small sample sizes. In order to increase the efficiency of small area estimates, it is common practice to construct small area estimates using explicit linking models that borrow information from

[1] Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa, via Ridolfi, 10–56124 Pisa, salvati@ec.unipi.it, m.pratesi@ec.unipi.it.
[2] Centre for Census and Survey Research, University of Manchester, Manchester M13 9PL, UK, nikos.tzavidis@manchester.ac.uk.
[3] Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522, Australia, ray@uow.edu.au.

related areas. The most popular class of these are models that include random area effects to account for between area variation beyond that explained by the auxiliary variables in the model. Estimation for these models is typically carried out assuming that the random area effects are uncorrelated (Rao 2003). In many situations, however, it is reasonable to assume that the effects of neighbouring areas – where neighbourhood is often defined in terms of a contiguity criterion – are correlated with the correlation decaying to zero as between area distance increases (Petrucci *et al.* 2005). In such cases the assumption of spatial independence of the random area effects becomes questionable. The problem of accounting for spatial correlation between the small areas has been recently tackled by extending the model of Battese *et al.* (1988) using a Simultaneously Autoregressive (SAR) process (Salvati 2004; Petrucci and Salvati 2006).

Chambers and Tzavidis (2006) have proposed a new approach to small area estimation that is based on the use of M-quantile models. Unlike traditional random effects models, M-quantile models do not depend on strong distributional assumptions and are robust to the presence of outliers. In its current form, however, the M-quantile approach to small area estimation does not allow for spatially correlated area effects. The aim of this paper is to extend the M-quantile approach to account for spatial correlation between small areas.

The paper is organized as follows. In section 2 we review random effects models that allow for spatially correlated random effects. In section 3 we propose an extension of the M-quantile approach to account for spatial correlation between the small areas. In section 4 we demonstrate usefulness of this framework through Monte Carlo simulation studies. The main focus of the comparisons is between the Spatial M-quantile and M-quantile models. In section 5 we present an application of spatial M-quantile models for estimating the average and median production of olives at the level of Local Economy System (LES) in Tuscany. Finally, in section 6 we summarise our main findings.

2. Small Area Models with Spatially Correlated Random Effects

Let $\mathbf{x}_i$ be a known vector of p auxiliary variables for each population unit j in small area i and assume that information for the variable of interest y is available only on the sample. The target is to use these data to estimate various area specific quantities. The most popular models used for this purpose are mixed effects models, i.e. models with random area effects. A linear mixed effects model has the following form:

$$y_{ij} = \mathbf{x}_{ij}^{T}\boldsymbol{\beta} + z_{ij}u_i + \varepsilon_{ij},\ j = 1,\ldots,n_i,\ i = 1,\ldots,m \qquad (0.1)$$

where u_i denotes a random area effect that characterizes differences in the conditional distribution of y given $\mathbf{x}$ between the m small areas, z_{ij} is a constant whose value is known for all units in the population and ε_{ij} is the error term

associated with the *j*-th unit within the *i*-th area. Conventionally, u_i and ε_{ij} are assumed to be independent and normally distributed with mean zero and variances σ_u^2 and σ_ε^2 respectively (Battese *et al.* 1988).

Model (0.1) can be extended to allow for correlated area effects. Let the deviations **v** from the fixed part of the model $\mathbf{x}^T\boldsymbol{\beta}$ be the result of an autoregressive process with parameter ρ and proximity matrix **W** (Cressie 1993; Anselin 1992), i.e.

$$\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{u} \Rightarrow \mathbf{v} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u} \tag{0.2}$$

where **I** is $m \times m$ identity matrix. The matrix $(\mathbf{I} - \rho\mathbf{W})$ needs to be strictly positive definite to ensure the existence of $(\mathbf{I} - \rho\mathbf{W})^{-1}$. This happens if $\rho \in \left(\frac{1}{\min(\lambda_i)}, \frac{1}{\max(\lambda_i)} \right)$ where λ_i's are the eigenvalues of matrix **W.** Combining (0.1) and (0.2), with $\boldsymbol{\varepsilon}$ independent of **v**, the model with spatially correlated errors can be expressed as

$$\mathbf{y} = \mathbf{x}^T\boldsymbol{\beta} + \mathbf{Z}(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u} + \boldsymbol{\varepsilon}. \tag{0.3}$$

The error term **v** then has the $m \times m$ Simultaneously Autoregressive (SAR) dispersion matrix:

$$\mathbf{G} = \sigma_u^2 \left[(\mathbf{I} - \rho\mathbf{W}^T)(\mathbf{I} - \rho\mathbf{W}) \right]^{-1}. \tag{0.4}$$

The **W** matrix describes the neighbourhood structure of the small areas whereas ρ defines the strength of the spatial relationship among the random effects associated with neighbouring areas. For ease of interpretation, the general spatial weight matrix is defined in row standardized form in which case ρ is referred to as the spatial autocorrelation parameter (Banerjee *et al.* 2004). Under (0.3), the Spatial Best Linear Unbiased Predictor (Spatial BLUP) estimator of the small area parameters and its empirical version (SEBLUP) are obtained following Henderson (1975). The SEBLUP estimator of the mean for small area *i*, $\bar{y}_i$, is

$$\begin{aligned} \hat{\bar{y}}_i = {} & \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}} + \mathbf{b}_i^T\hat{\sigma}_u^2 \left[(\mathbf{I} - \hat{\rho}\mathbf{W}^T)(\mathbf{I} - \hat{\rho}\mathbf{W}) \right]^{-1} \mathbf{Z}^T \\ & \left\{ \hat{\sigma}_\varepsilon^2 \mathbf{I}_n + \mathbf{Z}\hat{\sigma}_u^2 \left[(\mathbf{I} - \hat{\rho}\mathbf{W}^T)(\mathbf{I} - \hat{\rho}\mathbf{W}) \right]^{-1} \mathbf{Z}^T \right\}^{-1} (\mathbf{y}_s - \mathbf{x}^T\hat{\boldsymbol{\beta}}) \end{aligned} \tag{0.5}$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x}\right)^{-1}\mathbf{x}^T\mathbf{V}^{-1}\mathbf{y}_s$, $\bar{\mathbf{x}}_i^T$ denotes a known area specific vector of population means for the auxiliary variables, $\mathbf{y}_s$ is a $n \times 1$ vector of the sampled

observations, $\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2, \hat{\rho}$ are asymptotically consistent estimators of the parameters obtained by Maximum Likelihood (ML) or Restricted ML (REML) method; $\mathbf{b}_i^T$ is a $1 \times m$ vector $(0,0 \ldots 0,1,\ldots 0)$ with value 1 in the *i-th* position.

The mean squared error (MSE) of SEBLUP and its estimator are obtained following the results of Kackar and Harville (1984), Prasad and Rao (1990) and Datta and Lahiri (2000). More specifically, the MSE estimator consists of three components called g_1, g_2 and g_3. These are due to the variability associated with the estimation of the random effects (g_1), the estimation of $\boldsymbol{\beta}$ (g_2) and the estimation of $\sigma_u^2, \sigma_\varepsilon^2, \rho$ (g_3). The corresponding equations are in Singh *et al.* (2005) and Pratesi and Salvati (2008). Note that due to the introduction of the additional parameter ρ, the component g_3 of the MSE is not the same as in the case of the traditional EBLUP estimator (Saei and Chambers 2003; Singh *et al.* 2005; Petrucci and Salvati 2006, Pratesi and Salvati 2008).

3. Spatial M-quantile Models for Small Area Estimation

Chambers and Tzavidis (2006) have proposed a new approach to small area estimation that is based on modelling the M-quantiles of the conditional distribution of the study variable (y) given the covariates (Breckling and Chambers, 1988). Unlike mixed effects models, which assume that the variability associated with the conditional distribution of y given $\mathbf{x}$ can be at least partially explained by a pre-specified hierarchical structure, such as the small areas of interest, M-quantile regression does not depend on a hierarchical structure. Instead, we characterise the conditional variability across the population of interest by the so-called M-quantile coefficients of the population units. The corresponding M-quantile coefficients, $\{q_j; j \in s\}$, of the units in the sample are then estimated using a grid-based interpolation procedure. In particular, a fine grid on the (0,1) interval is first defined and, using the sample data, M-quantile regression lines are fitted at each value q on this grid using an iteratively reweighted least squares procedure (see Chambers and Tzavidis 2006 for details). If a data point lies exactly on a fitted M-quantile regression line, then the estimated M-quantile coefficient of the corresponding sample unit is set equal to q. Otherwise, if a data point lies between two fitted M-quantile regression lines, then the estimated M-quantile coefficient of the corresponding sample unit is derived by linear interpolation.

If a hierarchical structure does explain part of the variability in the population data, we expect units within clusters defined by this hierarchy to have similar M-quantile coefficients. Let $\overline{q}_i$ denote the average M-quantile coefficient for the population units in area i. An estimate of $\overline{q}_i$ is obtained by the corresponding

average value of the sample M-quantile coefficients of units j in area i, i.e. $\bar{\hat{q}}_i = \sum_{j \in s_i} q_j$. An estimator of the corresponding small area mean, $\bar{\hat{y}}_i$ is then

$$\bar{\hat{y}}_i = \frac{1}{N_i}\left(\sum_{j \in s_i} y_j + \sum_{j \in r_i} \mathbf{x}_j^T \hat{\boldsymbol{\beta}}(\bar{\hat{q}}_i) \right) \tag{0.6}$$

where $\hat{\boldsymbol{\beta}}(\bar{\hat{q}}_i)$ denotes the slope coefficient of the fitted M-quantile regression line at $\bar{\hat{q}}_i$, s_i and r_i respectively denote the sampled and non-sampled units in area i and N_i is the number of population units in area i. Note that (0.6) is equivalent to predicting the unobserved value y_j for population unit $j \in r_i$ using $\mathbf{x}_j^T \hat{\boldsymbol{\beta}}(\bar{\hat{q}}_i)$.

In this paper we propose an extension to the above approach to account for spatial correlation between the small areas. In particular, assuming that the target population is made up of m small areas, we propose modelling the sample M-quantile coefficients using the model

$$\mathbf{logQ} = \mathbf{1}_s^T \boldsymbol{\beta} + \mathbf{Z}_s \left(\mathbf{I} - \rho \mathbf{W} \right)^{-1} \mathbf{u} + \boldsymbol{\varepsilon} \tag{0.7}$$

where $\mathbf{logQ}$ is the $n \times 1$ vector of the *logit* transformation of q_j, $\log\left[q_j / \left(1 - q_j\right) \right]$, $\mathbf{1}_s^T$ is vector of ones with dimension n and $\mathbf{Z}_s$ is the $n \times m$ incidence matrix for the random effects vector. In expression (0.7) we could have employed alternative link functions such as the *probit* link function. However, we expect that the choice of the link function will have little impact upon the small area estimates. Under model (0.7), the Spatial Empirical Best Predictor of $\bar{q}_i$ is

$$\bar{\hat{q}}_i = \frac{exp(\hat{p}_i)}{1 + exp(\hat{p}_i)} \tag{0.8}$$

where

$$\hat{p}_i = \hat{\beta}_0 + \mathbf{b}_i^T \hat{\sigma}_u^2 \mathbf{D}_s^{-1} \mathbf{Z}_s^T \left\{ \hat{\sigma}_\varepsilon^2 \mathbf{I}_n + \mathbf{Z}_s \hat{\sigma}_u^2 \mathbf{D}_s^{-1} \mathbf{Z}_s^T \right\}^{-1} \left(\hat{\mathbf{q}}_s - \mathbf{1}_s \hat{\beta}_0 \right), \tag{0.9}$$

the coefficient $\hat{\beta}_0$ is the estimated value of the intercept of the mixed model and $\mathbf{D}_s = \left[\left(\mathbf{I} - \hat{\rho} \mathbf{W}^T \right) \left(\mathbf{I} - \hat{\rho} \mathbf{W} \right) \right]$. Here, $\hat{\mathbf{q}}_s$ is the $n \times 1$ vector of estimated M-quantile coefficients for the sample units q_j. An M-quantile estimator of the mean for area i that accounts for spatial correlation is then given by (0.6), but with $\bar{\hat{q}}_i$ now given by (0.8). A drawback of this specification is that although we use

the M-quantile approach in order to avoid using a parametric model in small area estimation, we still use the parametric model (0.7) to account for spatial correlation in the M-quantile coefficients. Ideally, we would like to employ a non-parametric approach to account for spatial correlation in the M-quantile coefficients. However, developing a fully non-parametric approach is beyond the scope of this paper.

As Tzavidis and Chambers (2006) note, the M-quantile estimator (0.6) can be biased particularly when small areas contain outliers. These authors have therefore proposed a bias-adjusted M-quantile estimator of the mean that is based on representing this estimator as a functional of the small area distribution function. More specifically, it is straightforward to see that (0.6) is derived by appropriately integrating the empirical distribution function

$$\hat{F}_i(t) = \frac{1}{N_i}\left(\sum_{j \in s_i} I\left(y_{ij} \le t\right) + \sum_{j \in r_i} I\left(\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\hat{\bar{q}}_i) \le t\right)\right). \tag{0.10}$$

Instead of using the empirical distribution function, the proposal of Tzavidis and Chambers (2007) is based on using the Chambers-Dunstan (1986) -hereafter denoted by a subscript CD- estimator of the small area distribution function

$$\hat{F}_{CD,i}(t) = \frac{1}{N_i}\left(\sum_{j \in s_i} I\left(y_{ij} \le t\right) + \frac{1}{n_i}\sum_{j \in r_i}\sum_{k \in s_i} I\left\{\left[\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\hat{\bar{q}}_i) + \left(y_{ik} - \mathbf{x}_{ik}^T \hat{\boldsymbol{\beta}}(\hat{\bar{q}}_i)\right)\right] \le t\right\}\right). \tag{0.11}$$

The corresponding bias-adjusted mean estimator for small area *i* is then

$$\hat{\bar{y}}_i = \int t d\hat{F}_{CD,i}(t) = \frac{1}{N_i}\left(\sum_{j \in s_{n_i}} y_{ij} + \sum_{j \in r_i} \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\hat{\bar{q}}_i) + \frac{N_i - n_i}{n_i} \sum_{j \in s_{n_i}} \left(y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\hat{\bar{q}}_i)\right)\right). \tag{0.12}$$

Estimates of other quantiles of the distribution of *y* in small area *i* can be obtained by appropriately integrating (0.11) (see also Tzavidis and Chambers 2007).

Tzavidis and Chambers (2006) and Chambers *et al.* (2007) proposed an estimator of the mean squared error of (0.12). The main limitation of this estimator is that it does not account for the variability introduced in estimating the area specific *q*'s. Empirical evaluations (Tzavidis *et al.* 2006), however, indicate that this mean squared error estimator provides a good approximation to the true mean squared error. In this paper, however, we focus our attention on the performance of M-quantile estimator (0.12), obtained with and without spatial information in the presence or not of outliers.

4. Simulation Experiments

Monte Carlo simulation experiments were designed for assessing the performance of the spatial M-quantile model described in the previous section. In particular the aim of these simulation exercises is to examine the usefulness of this framework for capturing the spatial structure of the data used for small area estimation and to investigate the performance of the small area methods in the presence of outliers. We illustrate the performance of the standard M-quantile estimator (0.6) and the CD form of the M-quantile estimator (0.12), with $\hat{\bar{q}}_i$ determined by simple averaging over area *i*, and the corresponding Spatial M-quantile versions of (0.6) and (0.12) with $\hat{\bar{q}}_i$ determined by (0.8). For reasons of completeness, we also considered other widely used methods for small area estimation such as the EBLUP estimator, and the Spatial EBLUP estimator (SEBLUP).

Synthetic population data are generated for the small areas using a spatial nested error regression model with random area effects distributed according to a SAR dispersion matrix with fixed spatial autoregressive coefficient given by

$$y_{ij} = 5 + 2x_{ij} + v_i + e_{ij}$$

for $i = 1\text{K}\ m$ and $j = 1\text{K}\ N_i$, with $m = 42$ and $N_i = 100$. The values x_{ij} of the auxiliary variable were drawn from a uniform distribution between 0 and 10, the vector $\mathbf{v} = [v_1,...v_m]^T$ of the random area specific effects was generated from $MVN\left(\mathbf{0}, \sigma_u^2\left[\left(\mathbf{I} - \rho\mathbf{W}^T\right)\left(\mathbf{I} - \rho\mathbf{W}\right)\right]^{-1}\right)$ with $\sigma_u^2 = 1$. The SAR dispersion matrix was generated with ρ equal to $0, \pm 0.25, \pm 0.50, \pm 0.75$ and the neighbourhood structure (**W**) was defined by randomly assigning neighbours for each area as follows: The value 1 was assigned to if the value drawn from a uniform $[0,1]$ distribution was greater than 0.5, otherwise it was set to 0. The maximum number of neighbours for each area was 5, and the **W** matrix was standardized by row, i.e. the row elements summed to one. We can therefore refer to ρ as an autocorrelation parameter. The **W** matrix was kept fixed for all simulations.

The experiment is designed to investigate also the performance of the small area methods in the presence of outliers. For this reason the individual error e_{ij} were generated from a contamination distribution. This means that in ten areas, randomly selected, a $(1-\delta)$ proportion of the e_{ij}'s were generated from the underlying 'true' distribution $N(0,4)$ and the remaining δ proportion of the e_{ij} 's were generated from the contaminated distribution $N(0,100)$. The choice

$\delta = 0$ corresponds to uncontaminated data. For contaminated data we fixed $\delta = 0.1$.

We conducted a total of T = 200 independent simulations, consisting of generating population and sample data as described above. For each sample drawn of size 420 with $n_i = 10 \forall i$, the small area mean was estimated using (a) the EBLUP estimator, (b) the SEBLUP estimator, (c) the M-quantile estimator and (d) the Spatial M-quantile estimator.

For each estimator we computed the Average Absolute Relative Bias ($\overline{ARB}$), the Average Relative Root MSE ($\overline{RRMSE}$) and the Percentage Relative Bias $\overline{PRB}$ defined as follows:

$$\overline{ARB} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=1}^{T}\left|\hat{Y}_{it} / Y_i - 1\right|$$

$$\overline{RRMSE} = \frac{1}{m}\sum_{i=1}^{m}\frac{\left[\overline{MSE}(\hat{Y}_i)^{1/2}\right]}{Y_i}$$

$$\overline{PRB} = \frac{1}{m}\sum_{i=1}^{m}\frac{B_i^2}{\overline{MSE}(\hat{Y}_i)}$$

where

$$\overline{MSE}(\hat{Y}_i) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=1}^{T}\left(\hat{Y}_{it} - Y_i\right)^2 \text{ and } B_i^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{Y}_{it} - Y_i\right)^2.$$

On the T = 200 simulations the coverage rate (CR) of the nominal 95% confidence interval has been obtained by using the analytical MSE estimators recalled in sections 2 and 3 for each of the presented estimators.

Tables 1 and 2 summarise the results from the simulation study for $\rho = \pm 0.25, \pm 0.75$. Results obtained with other ρ values are available from the authors.

Table 1 allows us to examine the usefulness of Spatial M-quantile modelling for capturing the spatial structure of the data used for small area estimation. When spatial correlation is high and positive the SEBLUP and Spatial M-quantile outperform the other estimators in term of $\overline{RRMSE}$. The estimators are biased and the Spatial M-quantile is a good competitor of SEBLUP only when the CD correction is adopted. In the case of negative spatial correlation the results are confirmed, even if the gain of spatial modelling in terms of $\overline{RRMSE}$ and bias reduction slightly decreases. If the spatial correlation is negligible, the traditional estimators achieve results that are similar to those of the spatial models.

Table 1. Comparison of small area estimators, uncontaminated data

	Estimator	$\overline{ARB}$ (%)	$\overline{RRMSE}$ (%)	PRB(%)	CR(%)
$\rho = 0.75$	SEBLUP	0.83	7.85	19.09	95.4
	EBLUP	0.84	8.02	17.31	95.3
	M-quantile	1.21	8.37	32.95	93.3
	M-quantile CD	0.20	8.67	0.38	92.2
	Spatial M-quantile	1.12	7.85	16.16	94.0
	Spatial M-quantile CD	0.20	8.67	0.38	91.9
$\rho = 0.25$	SEBLUP	0.99	8.58	21.06	95.4
	EBLUP	0.97	8.52	20.38	95.2
	M-quantile	1.05	8.61	28.06	94.4
	M-quantile CD	0.21	9.53	0.37	92.3
	Spatial M-quantile	1.27	8.64	18.17	93.9
	Spatial M-quantile CD	0.21	9.53	0.37	92.1
$\rho = -0.25$	SEBLUP	1.05	8.81	20.53	95.4
	EBLUP	1.03	8.75	19.78	95.2
	M-quantile	1.08	8.85	27.60	94.4
	M-quantile CD	0.22	9.77	0.37	92.2
	Spatial M-quantile	1.29	8.89	17.89	93.9
	Spatial M-quantile CD	0.22	9.77	0.38	92.1
$\rho = -0.75$	SEBLUP	1.14	9.09	18.33	95.3
	EBLUP	1.14	9.17	17.02	95.1
	M-quantile	1.37	9.45	28.71	93.9
	M-quantile CD	0.23	9.97	0.38	92.3
	Spatial M-quantile	1.45	9.21	16.27	93.7
	Spatial M-quantile CD	0.23	9.97	0.38	92.1

The exam of Table 2 is useful to investigate the performance of the small area methods in the presence of outliers.

Generally, both for high and low values of $|\rho|$, the M-quantile estimators have smaller bias in comparison with EBLUP and SEBLUP. This is an expected result as the peculiarity of the M-quantile modelling is to correct the influence of outliers on bias. Particularly, the Spatial M-quantile estimators present the lowest value of $\overline{ARB}$. This seems to be due to the combination of the control of outliers with the advantage of considering the spatial structure of the data.

This ability is also responsible for the good performance of Spatial M-quantile in terms of $\overline{RRMSE}$. When the spatial correlation is high and positive, the estimator achieves the lowest level of $\overline{RRMSE}$ in Table 2.

The coverage rate of the empirical confidence interval is always close to its 95% nominal value for all the investigated estimators under each level of spatial correlation and contamination scenario.

Table 2. Comparison of small area estimators, contaminated data, thirty-two uncontaminated areas

	Estimator	$\overline{ARB}$ (%)	$\overline{RRMSE}$ (%)	PRB(%)	CR(%)	AwCIs
$\rho = 0.75$	SEBLUP	1.88	8.98	40.56	98.5	2.88
	EBLUP	2.51	902	7.87	98.8	2.92
	M-quantile	1.88	8.80	31.44	93.4	2.31
	M-quantile CD	-0.01	9.21	0.53	92.3	2.30
	Spatial M-quantile	1.20	8.50	31.95	94.0	2.31
	Spatial M-quantile CD	-0.02	9.21	0.53	92.6	2.30
$\rho = 0.25$	SEBLUP	2.03	9.82	44.66	976	2.81
	EBLUP	2.03	9.56	42.88	97.5	2.87
	M-quantile	1.13	9.01	28.06	94.3	2.29
	M-quantile CD	-0.03	9.92	0.53	92.4	2.29
	Spatial M-quantile	1.31	9.19	34.46	94.0	2.30
	Spatial M-quantile CD	-0.04	9.93	0.53	92.5	2.30
$\rho = -0.25$	SEBLUP	2.01	10.12	43.89	97.4	2.82
	EBLUP	1.87	9.83	42.24	97.0	2.77
	M-quantile	1.00	9.26	27.99	94.0	2.29
	M-quantile CD	-0.04	10.11	0.54	92.3	2.29
	Spatial M-quantile	1.27	9.41	32.88	95.1	2.30
	Spatial M-quantile CD	-0.04	10.11	0.54	92.9	2.30
$\rho = -0.75$	SEBLUP	1.87	10.14	42.22	97.9	2.87
	EBLUP	2.05	10.14	44.80	98.1	2.88
	M-quantile	1.27	9.91	29.60	92.6	2.30
	M-quantile CD	-0.03	10.30	0.54	92.5	2.30
	Spatial M-quantile	1.62	9.87	31.68	93.4	2.31
	Spatial M-quantile CD	-0.03	10.30	0.54	92.7	2.30

5. Application

In the context of Italian agricultural surveys it is often of interest to produce accurate estimates of the average or of the total of farm production at local geographical areas, such as municipalities. However, such estimates can be difficult to produce due to the sparsity of the available survey data at this level of geography. As a result, previous work has focused on producing estimates at

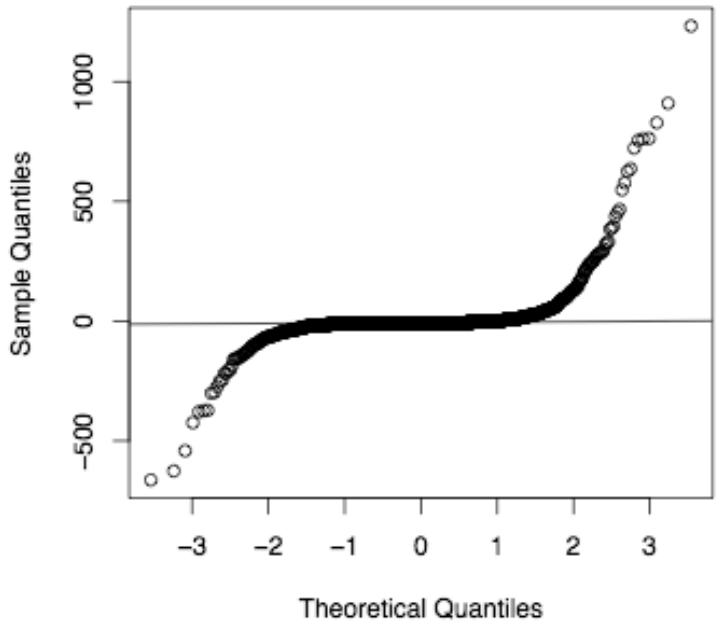
higher geographical levels such as Italian provinces (Benedetti *et al.* 2004). Accurate estimates at sub-regional level require either the enlargement of the sample size or the application of small area estimation methods.

In this application we employ data from the Farm Structure Survey (FSS - ISTAT 2003) that is carried out once every two years and collects information on farm land by type of cultivation, amount of animal production and structure and amount of farm employment from 55,030 farms. The target of inference is the average production of olives per farm in quintal units for each of the (small) areas making up the Local Economy System (LES) in the Tuscany region. However, as our exploratory analysis will show, the presence of outliers in the data suggests that it may be also useful to produce estimates of median olive production in each of the LES areas.

The Atlas of Coverage of the Tuscany Region maintained by the Geographical Information System of the Regione Toscana provided information on coordinates, surface area and positions of the small areas of interest (UTM system). The centroid of each area is the spatial reference for all the units residing in the same small area. The auxiliary variable we employ in our models is the surface area used for olive production.

Exploratory analysis was first used to test for the presence of spatial dependence in the data. Essential to this is the neighbourhood structure **W** that is defined as follows: the spatial weight, w_{ij}, is 1 if area *i* shares an edge with area *j* and 0 otherwise. For an easier interpretation, the general spatial weight matrix is defined in row standardized form, in which the row elements sum to one. In order to detect the spatial pattern (spatial association and spatial autocorrelation) of the average production of olives per farm, two standard global spatial statistics have been calculated: Moran's *I* and Geary's *C* (Cliff and Ord, 1981). The spatial dependence in the target variable is weak, but the value of Moran's *I* is statistically significant. This is consistent with the estimated value for Geary's *C*.

Figure 1. Normal probability plot of the linear regression model residuals



Using Restricted Maximum Likelihood estimation the value of spatial autoregressive coefficient, $\hat{\rho}$, is estimated to be 0.441 ($s.e. = 0.183$), which suggests a moderate spatial relationship. In addition, as part of our exploratory analysis we also used a regression model for investigating the relationship between the production of olives in quintal units and the surface area used for olive production. A normal probability plot of the model residuals shows a skewed distribution of the residuals and hence evidence of outlying observations (Figure 1). Given the spatial correlation in the data and the presence of outliers, instead of using the Skew-normal and Skew-t distribution (Genton, 2004) for dealing with possible anomalous features of the data, we decided to perform small area estimation using a small area model that appears robust to outliers estimation method. A model of this type is the proposed spatial M-quantile model.

Small area estimates of olive production per farm at LES level are therefore produced under this model using the Spatial M-quantile CD estimator. The choice of the Spatial M-quantile CD estimator is justified (i) because of the presence of outliers in the data; as Tzavidis and Chambers (2007) suggest, when outliers are present in the data, the M-quantile CD estimator of the small area average is more efficient than the corresponding M-quantile estimator, and (ii) because one of the targets of our analysis is to estimate the small area medians. In order to obtain consistent estimators of small area medians (and other quantiles), it is necessary to base these estimators on a consistent estimator of the small area distribution

such as the Chambers-Dunstan estimator. Finally, in order to complete our comparisons we also present small area mean estimates using the EBLUP and SEBLUP estimators.

The maps in Figures 2 and 3 depict small area model estimates. Figure 2 shows the predicted values of the (a) average (Figure 2a) and (b) median (Figure 2b) of olive production per farm for each of the 42 LES areas in the Tuscany region under the Spatial M-quantile model. Figure 3 presents the small area estimates of the average of olive production per farm under EBLUP (Figure 3a) and SEBLUP (Figure 3b) estimators. We can note that EBLUP and SEBLUP estimates are very similar and they differ from the estimates obtained under the Spatial M-quantile model. The spatial distribution of M-quantile-based estimates appears to be less variable than that obtained with the traditional EBLUP and SEBLUP approaches. At this point we should also mention that in two LES areas the EBLUP and SEBLUP estimates of the small area means were negative. This can happen when there are outliers in the data which invalidate the assumptions of the linear mixed model. For these two LES areas we therefore decided to replace the negative model-based estimates (EBLUP and SEBLUP) with the corresponding direct estimates. We should also mention that we did not encounter negative small area estimates when using the M-quantile model. This is explained by the robust estimation method employed for fitting the M-quantile models.

Figure 2. Small area estimates of the (a) average and (b) median olive production per farm in quintal units for each of the 42 LES in the Tuscany region under Spatial M-quantile model

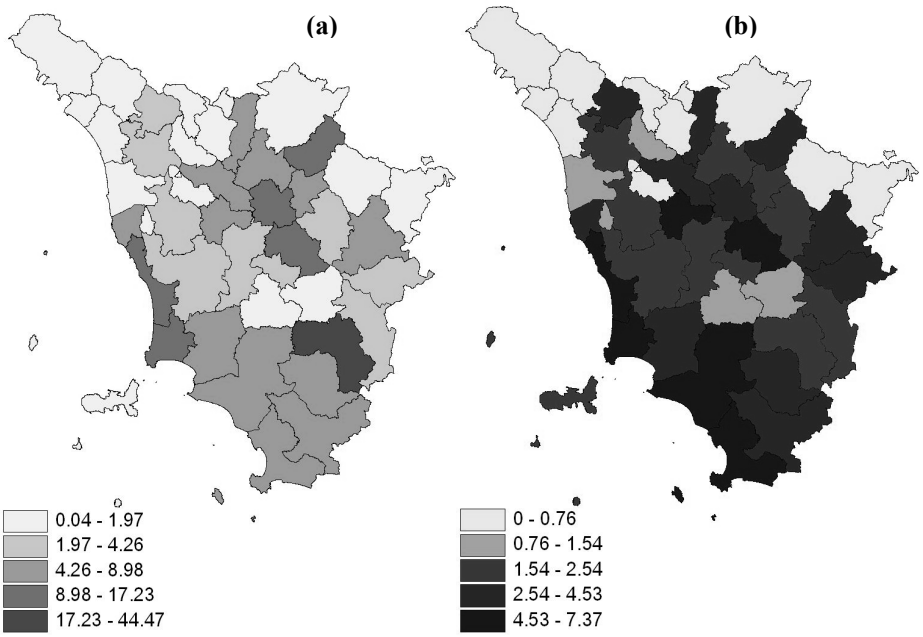
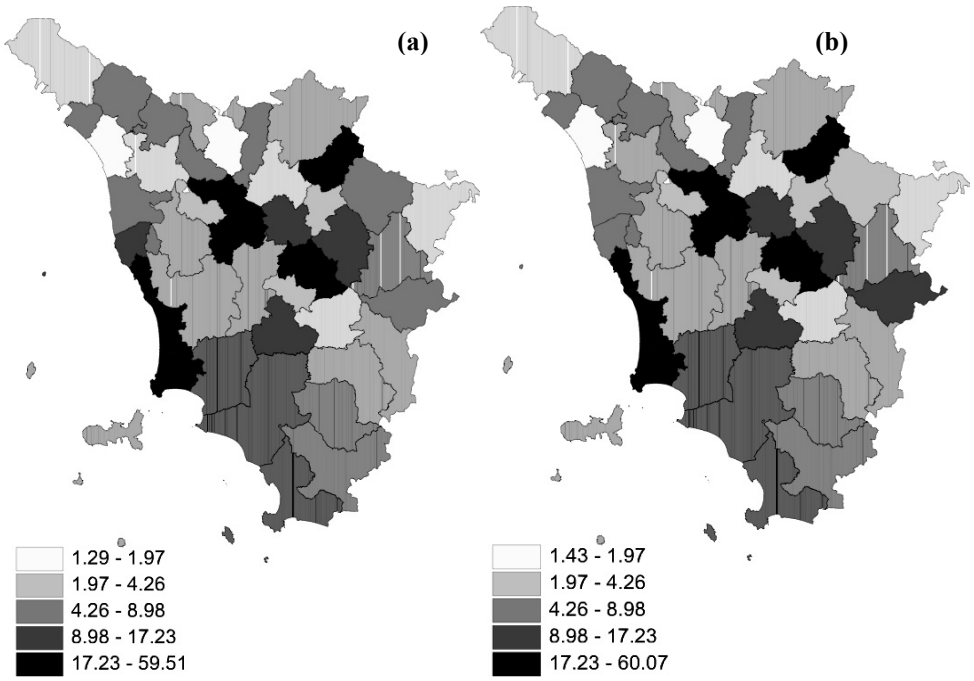


Figure 3. Small area estimates of the average olive production per farm in quintal units for each of the 42 LES in the Tuscany region under (a) EBLUP and (b) SEBLUP estimators.



Estimates of the small area median production of olives are also obtained under the Spatial M-quantile model. The median appears to be insensitive to the presence of few big farms that raise the average level of production. The spatial distribution of the median production also appears to be more homogenous than the corresponding spatial distribution of the small area means (see Figure 2). This emphasizes the importance of producing maps that represent not only the spatial distribution of the mean but also of other quantiles of the cumulative distribution function within each small area. The information contained in such maps is valuable both for agricultural policy interventions and for data users.

6. Conclusions

In this paper we propose an extension to the Chambers and Tzavidis (2006) small area M-quantile approach. The Spatial Empirical Best Predictor of $\overline{q}_i$ is obtained under a mixed model with spatially correlated random effects. Small area estimates are then obtained by fitting an M-quantile model at the average area specific M-quantile coefficient predicted under this parametric model. The proposed estimator of the small area mean, called Spatial M-quantile estimator, captures the spatial structure of the data and is robust in the presence of outliers.

Results from a simulation study show that this approach works well in comparison to the conventional M-quantile estimator. The main findings of our simulation study are: (i) when the data are not affected by outliers, in the case of strong spatial correlation the Spatial M-quantile estimator performs better than the M-quantile estimators; (ii) when the outliers contaminate the data, the properties of M-quantile estimators are confirmed, and in the case of high spatial correlation the Spatial M-quantile estimator emerges as the best choice in the small area mean estimation.

The empirical results confirm that the proposed semi-parametric approach offers one way of incorporating the spatial information in the M-quantile small area model preserving robustness to the presence of outliers.

A drawback of our approach is that we still need to specify a fully parametric model for the unit-specific M-quantile coefficients. We are currently investigating the use of non-parametric methods to incorporate spatial information into the M-quantile approach.

Acknowledgements

The work reported here has been developed under the support of the project PRIN *Metodologie di stima e problemi non campionari nelle indagini in campo agricolo-ambientale* awarded by the Italian Government to the Universities of Florence, Cassino, Pisa and Perugia. The authors are grateful to two anonymous Referees for their comments and suggestions that greatly improved the paper over earlier versions.

REFERENCES

ANSELIN, L. (1992) *Spatial Econometrics: Method and Models*, Kluwer Academic Publishers, Boston.

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

BENEDETTI, R., ESPA, G. and PIERSIMONI, F. (2004) Esperienze di stime provinciali indirette di variabili aziendali, *Atti del Convegno ISPA 2004*, 6 maggio 2004, Università degli Studi di Cassino, 65–81.

BNERJEE, S., CARLIN, B.P. and GELFAND, A.E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall, New York.

BRECKLING, J., CHAMBERS, R. (1988) M-quantiles, *Biometrika*, 75, 4, 761–771.

CHAMBERS, R., DUNSTAN, R. (1986) Estimating distribution function from survey data. *Biometrika*, 73, 597–604.

CHAMBERS, R. and TZAVIDIS, N. (2006) M-quantile Models for Small Area Estimation, *Biometrika*, 93, pp. 255–268.

CHAMBERS, R., CHANDRA, H. and TZAVIDIS, N. (2007) On robust mean squared error estimation for linear predictors for domains. *[Paper submitted for publication. A copy is available upon request]*.

CLIFF, A.D. and ORD, J.K. (1981) *Spatial Processes. Models & Applications*, Pion Limited, London.

CRESSIE, N. (1993) *Statistics for spatial data*, John Wiley & Sons, New York.

DATTA, G.S. and LAHIRI, P. (2000) A Unified Measure of Uncertainty of Estimates for Best Linear Unbiased Predictors in Small Area Estimation Problem, *Statistica Sinica,* 10, 613–627.

GENTON, M. G. (2004) *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Edited Volume, Chapman & Hall / CRC, Boca Raton, FL, 416 pp.

HENDERSON C. (1975) Best linear unbiased estimation and prediction under a selection model, *Biometrics*, *31*, 423–447.

KACKAR, R.N. and HARVILLE, D.A. (1984) Approximations for standard errors of estimators for fixed and randomeffects in mixed models, *Journal of the American Statistical Association*, 79, 853–862.

PETRUCCI, A. and SALVATI, N. (2005) "Small Area Estimation: the Spatial EBLUP at area and at unit level". Atti del Convegno "Metodi per l'integrazione di dati da più fonti", Roma.

PETRUCCI, A., PRATESI, M. and SALVATI, N. (2005) Geographic Information in Small Area Estimation: Small Area Models and Spatially Correlated Random Area Effects, *Statistics in Transition*, *7*, 3, 609–623.

PETRUCCI, A. and SALVATI, N. (2006) Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment, *Journal of Agricultural, Biological and Environmental Statistics*, *11*, *2*, 169–182.

PFEFFERMANN, D. (2002) Small Area Estimation - New Developments and Directions, *International Statistical Review*, *70*, *1*, 125–143.

PRASAD, N. and RAO, J. N. K. (1990), The estimation of the mean squared error of small-area estimators, *Journal of the American Statistical Association*, 85, 163–171.

PRATESI MONICA, SALVATI NICOLA (2008) Small Area Estimation: the EBLUP estimator based on spatially correlated random area effects, *Statistical Methods & Applications*, 17, 1, 114–131.

RAO, J.N.K. (2003) *Small area estimation,* John Wiley & Sons, New York.

SAEI, A. and CHAMBERS, R. (2003) Small Area Estimation Under Linear and Generalized Linear Model With Time and Area Effects, *Working Paper M03/15, Southampton Statistical Sciences Research Institute, University of Southampton.*

SALVATI, N. (2004) Small Area Estimation by Spatial Models: the Spatial Empirical Best Linear Unbiased Prediction (Spatial EBLUP), *Working Paper n 2004/04*, "G. Parenti" Department of Statistics, University of Florence.

SINGH, B.B., SHUKLA, G.K. and KUNDU, D. (2005) Spatio-Temporal Models in Small Area Estimation, *Survey Methodology*, *31*, *2*, 183–195.

TZAVIDIS, N. and CHAMBERS, R. (2006) Bias adjusted estimation for small areas with outlying values. *Southampton Statistical Sciences Research Institute*, *Working Paper M06/09*, Southampton.

TZAVIDIS, N. and CHAMBERS, R. (2007) Robust prediction of small area means and distributions. Submitted for publication.

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 269—275

A COMPOSITE ESTIMATOR FOR SMALL DOMAINS AND ITS SENSITIVITY INTERVAL FOR WEIGHTS α

G.C. Tikkiwal[1], Piyush Kant Rai[2]

ABSTRACT

The composite estimation methods, suggested in the literature, have major problems to deal with the estimation of optimum weights to combine synthetic and direct estimators. To take care of the absence of optimum weights, we have obtained the sensible interval of involved weights in the form of better performance interval of the weights with a view to retaining superiority for the different composite estimators i.e. for the direct ratio vs. synthetic ratio composite estimator and simple direct vs. synthetic ratio composite estimator.

Key words: Small area (domain) estimation, Synthetic and Composite estimation, Proportional inflation, TRS.

1. Introduction

Tikkiwal, G.C. and Ghiya (2000) define a generalized class of synthetic estimators, using auxiliary information. The proposed class of synthetic estimators gives consistent estimators if the corresponding synthetic assumptions are satisfied. The generalized class, among others includes the simple synthetic, ratio synthetic and product synthetic estimators. Further ,the generalized synthetic estimator is used to construct a generalized class of composite estimators by combining with generalized direct ratio estimators. Generalized class, among others include simple ratio, ratio synthetic, composite estimator taking combination of various direct and synthetic estimators. These authors, further, discuss the generalized class of synthetic and composite estimators under SRS and stratified random sampling schemes. It has been shown by Tikkiwal, G.C. and Ghiya (2004), Tikkiwal, G.C. and Pandey (2007) that, when auxiliary variable is closely related with variable under study, small area estimators based

[1] Department of Mathematics and Statistics, J.N.V.U, Jodhpur India.
[2] Department of Mathematics and Statistics, Banasthali University, Rajasthan India-304022, E-mail: raipiyush5@gmail.com, phone: +91-1438-228991, Fax: +91-1438-228649.

on auxiliary information perform better than those which do not use auxiliary information. Further, Tikkiwal, G.C. and Pandey (2007) discuss the generalized class of synthetic and composite estimators under Lahiri-Midzuno and systematic sampling schemes. The relative performance of these estimators is empirically assessed through a simulation study for the problem of crop acreage estimation for small domains. They recommend the use of composite estimator, which is a weighted sum of direct ratio and ratio synthetic estimators, for crop acreage estimation for small domains under the TRS (Timely Reporting Scheme), for such case where the synthetic assumption is satisfied. Most of the times such sampling schemes are only of theoretical interest, as in large-scale sample surveys mostly multistage sampling designs are used.

2. Composite Estimator and its sensitivity Interval for Weights α

Let us consider the case of composite estimator as direct ratio estimator shrinkage with ratio synthetic estimator, i.e. denote

$$\bar{y}_{c,a} = \alpha\, \bar{y}_{d,r,a} + (1-\alpha)\, \bar{y}_{syn,r,a} \tag{2.1}$$

Here α is the weight assign to them and "a" denote domain. Let us denote 'P' as the proportional inflation in the variance of $\bar{y}_{c,a}$ resulting from the use of some α other than α_{opt} i.e.

$$P = \frac{MSE(\bar{y}_{c,a}) - MSE_{opt}(\bar{y}_{c,a})}{MSE_{opt}(\bar{y}_{c,a})} \tag{2.2}$$

Now, to obtain P we have to compute MSE of composite estimator under α and α_{opt} . Thus, expression for P is given as

$$P = \frac{(\alpha^2 - \alpha_{opt}^2)MSE(\bar{y}_{d,r,a}) + \{(1-\alpha)^2 - (1-\alpha_{opt})^2\}MSE(\bar{y}_{syn,r,a})}{\alpha_{opt}^2 MSE(\bar{y}_{d,r,a}) + (1-\alpha_{opt})^2 MSE(\bar{y}_{syn,r,a})}$$

$$= \frac{\left(\frac{1-\alpha}{1-\alpha_{opt}}\right)^2 \left[\left(\frac{\alpha}{(1-\alpha)}\right)^2 MSE(\bar{y}_{d,r,a}) + MSE(\bar{y}_{syn,r,a})\right] - \left[\left(\frac{\alpha_{opt}}{(1-\alpha_{opt})}\right)^2 MSE(\bar{y}_{d,r,a}) + MSE(\bar{y}_{syn,r,a})\right]}{\left(\frac{\alpha_{opt}}{1-\alpha_{opt}}\right)^2 MSE(\bar{y}_{d,r,a}) + MSE(\bar{y}_{syn,r,a})}$$

$$= \left[\frac{1-\alpha}{1-\alpha_{opt}}\right]^2 P_1 \quad -1 \tag{2.3}$$

where P_1 is given as

$$P_1 = \frac{\left(\frac{\alpha}{1-\alpha}\right)^2 MSE(\bar{y}_{d,r,a}) + MSE(\bar{y}_{syn,r,a})}{\left(\frac{\alpha_{opt}}{1-\alpha_{opt}}\right)^2 MSE(\bar{y}_{d,r,a}) + MSE(\bar{y}_{syn,r,a})}$$

Now, equation (2.2) is positive and gives

$$P_1 \geq \left[\frac{1-\alpha_{opt}}{1-\alpha}\right]^2 \tag{2.4}$$

$$MSE(\bar{y}_{d,r,a}) = \left(\frac{N-n}{N\,n}\right)\left[\frac{N_a - 1}{N-1}\right] S^2_{EUa}$$

where

$$S^2_{EUa} = \frac{1}{N_a - 1}\sum_{U_a}(y_k - B_a x_k)^2$$

with

$$B_a = \frac{(\sum_{U_a} y_k)}{(\sum_{U_a} x_k)}$$

$$MSE(\bar{y}_{syn,r,a}) = \left(\frac{\bar{Y}}{\bar{X}}\bar{X}_a\right)^2\left[1 + \frac{N-n}{N\,n}\left\{3\,C_x^2 + C_y^2 - 4C_{xy}\right\}\right]$$

$$-2\,\bar{Y}_a\left(\frac{\bar{Y}}{\bar{X}}\bar{X}_a\right)\left[1 + \frac{N-n}{N\,n}\left\{\,C_x^2 - C_{xy}\right\}\right] + \bar{Y}_a^2$$

If synthetic assumption is satisfied then

$$MSE(\bar{y}_{syn,r,a}) = \frac{N-n}{N\,n}\left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2$$

Thus

$$P_1 = \frac{\left(\frac{\alpha}{1-\alpha}\right)^2 MSE(\bar{y}_{d,r,a}) + MSE(\bar{y}_{syn,r,a})}{\left(\frac{\alpha_{opt}}{1-\alpha_{opt}}\right)^2 MSE(\bar{y}_{d,r,a}) + MSE(\bar{y}_{syn,r,a})}$$

$$= \frac{\left(\frac{\alpha}{1-\alpha}\right)^2 \left(\frac{N_a-1}{N-1}\right) S_{EUa}^2 + \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2}{\left(\frac{\alpha_{opt}}{1-\alpha_{opt}}\right)^2 \left(\frac{N_a-1}{N-1}\right) S_{EUa}^2 + \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2}$$

From equation (2.4) we have

$$\frac{\left(\frac{\alpha}{1-\alpha}\right)^2 \left(\frac{N_a-1}{N-1}\right) S_{EUa}^2 + \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2}{\left(\frac{\alpha_{opt}}{1-\alpha_{opt}}\right)^2 \left(\frac{N_a-1}{N-1}\right) S_{EUa}^2 + \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2} \geq \frac{\left(1-\alpha_{opt}\right)^2}{\left(1-\alpha\right)^2}$$

so

$$\alpha^2 \left(\frac{N_a-1}{N-1}\right) S_{EUa}^2 + \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2 (1-\alpha)^2 \geq \alpha_{opt}^2 \left(\frac{N_a-1}{N-1}\right) S_{EUa}^2 + \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2 (1-\alpha_{opt})^2$$

$$\Rightarrow (\alpha^2 - \alpha_{opt}^2)\left(\frac{N_a-1}{N-1}\right) S_{EUa}^2 \geq \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2 \left\{(1-\alpha_{opt})^2 - (1-\alpha)^2\right\}$$

$$\Rightarrow (\alpha + \alpha_{opt})\left(\left(\frac{N_a-1}{N-1}\right) S_{EUa}^2 + \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2\right) \geq 2\left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2$$

$$\alpha \geq 2\frac{\left(C_x^2 + C_y^2 - C_{xy}\right)\bar{Y}_a^2}{\left(\frac{N_a - 1}{N-1}\right)S_{EUa}^2 + \left(C_x^2 + C_y^2 - C_{xy}\right)\bar{Y}_a^2} - \alpha_{opt}$$

or

$$\left(\alpha + \alpha_{opt}\right) \geq 2\frac{\left(C_x^2 + C_y^2 - C_{xy}\right)\bar{Y}_a^2}{\left(\frac{N_a - 1}{N-1}\right)S_{EUa}^2 + \left(C_x^2 + C_y^2 - C_{xy}\right)\bar{Y}_a^2} \tag{2.5}$$

Now, let us consider the case when we have composite estimator of the form simple direct and synthetic ratio, i.e.

$$\bar{y}_{c,a} = \alpha\, \bar{y}_{d,\,a} + (1-\alpha)\, \bar{y}_{syn,r,a} \tag{2.6}$$

Here

$$MSE(\bar{y}_{d,a}) = \frac{N-n}{N\,n}\left(\frac{S_a^2}{\phi_a}\right)$$

where $\phi_a = \dfrac{N_a}{N}$ and $S_a^2 = \dfrac{1}{N_a - 1}\sum_{i \in a}(y_i - \bar{Y}_a)^2$

Using (2.4) we have

$$\frac{\left(\frac{\alpha}{1-\alpha}\right)^2 \frac{S_a^2}{\phi_a} + \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2}{\left(\frac{\alpha_{opt}}{1-\alpha_{opt}}\right)^2 \frac{S_a^2}{\phi_a} + \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2} \geq \frac{\left(1-\alpha_{opt}\right)^2}{\left(1-\alpha\right)^2}$$

So

$$\alpha^2 \frac{S_a^2}{\phi_a} + \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2 (1-\alpha)^2 \geq \alpha_{opt}^2 \frac{S_a^2}{\phi_a} + \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2 (1-\alpha_{opt})^2$$

$$\Rightarrow (\alpha^2 - \alpha_{opt}^2)\frac{S_a^2}{\phi_a} \geq \left(C_x^2 + C_y^2 - 2\,C_{xy}\right)\bar{Y}_a^2 \left\{(1-\alpha_{opt})^2 - (1-\alpha)^2\right\}$$

$$\Rightarrow (\alpha+\alpha_{opt})\left(\frac{S_a^2}{\phi_a}+\left(C_x^2+C_y^2-2\,C_{xy}\right)\bar{Y}_a^2\right)\geq 2\left(C_x^2+C_y^2-2\,C_{xy}\right)\bar{Y}_a^2$$

$$\Rightarrow \alpha \geq 2\frac{\left(C_x^2+C_y^2-C_{xy}\right)\bar{Y}_a^2}{\frac{S_a^2}{\phi_a}+\left(C_x^2+C_y^2-C_{xy}\right)\bar{Y}_a^2}-\alpha_{opt}$$

or

$$\left(\alpha+\alpha_{opt}\right)\geq 2\frac{\left(C_x^2+C_y^2-C_{xy}\right)\bar{Y}_a^2}{\frac{S_a^2}{\phi_a}+\left(C_x^2+C_y^2-C_{xy}\right)\bar{Y}_a^2} \qquad (2.7)$$

Conclusion

In either cases when α takes values other than α_{opt}, from (2.5) and (2.7) we have the range of weights for the proposed composite estimators which provide P value positive i.e. $P=\dfrac{MSE(\bar{y}_{c,a})-MSE_{opt}(\bar{y}_{c,a})}{MSE_{opt}(\bar{y}_{c,a})}>0$

Acknowledgement

The authors are grateful to the University Grant Commission (UGC) New Delhi, India for providing support and facilitation to this research and development work.

REFERENCES

Agrawal, M.C. and Roy, D.C. (1997). Efficient Estimators for Small Domains. Jour. Ind. Soc. Ag. Statistics 52(3), 327–337.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An appraisal. Statistical Science, **9**, 55–93.

Hedayat, A.S. and Sinha, B.K. (1991). Design and Inference in Finite Population Sampling. John Wiley and Sons, New York.

Platek, R., Rao, J.N.K., Sarndal, C.E. and Singh, M.P. (1987). Small Area Statistics: An International Symposium. John Wiley and Sons, New York.

Purcell, N.J. and Kish, L. (1979). Estimation for small domain. Biometrics, **35**, 365–384.

Sarndal, C.E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag, New York.

Tikkiwal, B.D. and Tikkiwal, G.C. (1991). Sampling strategies in surveys. The role of the theory of T-classes and computers. Symposium Ind. Agri. Stat. Research Institute, New Delhi, 287–296.

Tikkiwal, G.C. and Gupta A.K. (1991). Estimation of population mean under successive sampling when various weights and regression coefficient are unknown. Biometrical Journal, **33**, 529–538.

Tikkiwal, G.C. and Ghiya, Alka (2000). A generalized class of synthetic estimators with application to crop acreage estimation for small domains. Biometrical Journal, 42 (7), 865–876.

Tikkiwal, G.C. and Ghiya, A. (2004). A generalized class of composite estimators with application of crop acreage estimation for small domains. Statistics in Transition, 6(5), 697–711.

Tikkiwal, G.C. and Pandey, K.K. (2007). On Some Aspects of Small Area Estimation Using Auxiliary Information. Ph.D. Thesis under supervision of Prof. G.C. Tikkiwal, Head of Department of Mathematics and Statistics J.N.V. University Jodhpur Rajasthan.

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 277—299

STATISTICAL MODELS FOR SOCIAL STRESS ANALYSIS

Luisa Canal[1], Walenty Ostasiewicz[2]

ABSTRACT

This paper presents the critical overview of the statistical models that could be used in the social stress analysis. Such analysis is considered to consist of the identification of the social stressors, and of the measurement of their potency to destroy the social harmony. Four main groups of methods are discussed: item response models, factorial models, latent classification, and paired comparison.

1. Introduction

In the 1970s the world entered a period of transition, whose length is difficult to predict. Almost all social and economical arrangements are crumbling. Neo-classical liberalism seems to dominate the economic thought in all countries. The influence of multiple factors of the globalization is visible in all aspects of human life. Such transition has positive effects as well as enormous negative impacts on the social life. Values of citizenships are decaying, societies are being based on individualistic egotism. Above all, social ties that are necessary for social cohesion are diminishing.

In addition to the negative influence of the globalization processes, there are various internal phenomena which contribute negatively to the harmony of social life. Those phenomena distress people, irritate them, and in a consequence can even destroy the existing social order. In this paper, we assume that any phenomenon, event, or condition which may have a destructive impact on the social life can be called *social stressor*.

The literature concerning the stress analysis is quite rich. However, it is to be found mainly within the psychology, and above all in the medical psychology. Furthermore, it concentrates on psychological distress rather than on the social one. Problems of social distress are scarcely documented.

The basic premise of this paper is that the prevention of fraying of the social fabric is a more appropriate policy than its mending. In order to prevent the

[1] Trient University.
[2] Wroclaw University of Economics.

fraying of the fabric of our societies, the underlying destructive forces have to be identified and then evaluated. This paper discusses the available statistical methods which might be useful in analyzing the social stressors. Those methods include the identification, the measuring and the evaluation of perils and risks to social cohesion.

Social stressor is any phenomenon, event or condition of living which irritates people thus deteriorating the social atmosphere of common living. It has a negative impact on social cohesion and social harmony.

The adjective "social" in the term "social stressor" is used to emphasize the fact that the investigated phenomenon produces distress which could be common for a large group of people. It means that the existence of some *common sense* or common feature characterizing a whole group of people is assumed. As this characteristic is not observed directly, it is called a latent variable. It will be denoted by the symbol Z, and it is assumed that it "drives", commands or controls people's reaction to stressful phenomena. For the lack of established terminology, a latent variable Z will be called *susceptibility*, endurance, resistance or patience. To keep the discussion general enough, we admit a number of aspects of the susceptibility. Therefore, the trait Z is considered as a d-dimensional variable $Z = (Z_1, Z_2, ..., Z_d)$.

As different people are endowed with different amounts of susceptibility, we will interpret the trait Z as a random variable. The cumulative distribution of it is denoted by $H(z) = H(z_1, z_2, ..., z_d)$.

For illustrative purposes, here are a few examples of stressors.

1. Almost legalized political corruption,
2. Cynicism of politicians,
3. Brutality in TV movies,
4. Immoral behaviour of higher officials.

All phenomena of that kind will be denoted by symbols $Y_1, Y_2, ..., Y_p$. They will be also called *items*.

In order to assess how dangerous those phenomena are in destroying the social cohesion, the survey methodology will be applied. The measurement of the strength of a stressor can be done by "observing" people's reaction. By a reaction we mean an answer to a question concerning undesired phenomena. Two kinds of questions and two broad approaches to analysis of collected responses are being discussed: categorical responses and comparative responses.

To collect data, items $Y_1, Y_2, ..., Y_p$ representing appropriate stressors are prepared, and then they are administered to a group of n people (respondents), who are treated as experts or judges.

In the case of categorical responses, the task of the subject is to endorse or reject the administered item. If, for example, we would like to know whether or not a non-controlled immigration is a risk for social cohesion, we could form the

following item: "non-controlled immigration is dangerous for social cohesion", and then ask respondents to confirm or to reject this assertion.

All responses to *p* items $Y_1, Y_2, ..., Y_p$. given by *n* respondents are presented in the form of the following matrix

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & & & \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

Symbol y_{ij} has the following meaning

$$y_{ij} = \begin{cases} 1, \text{if item } Y \text{ is endorsed by ith respondent} \\ 0, \text{if item } Y \text{ is rejected by ith respondent} \end{cases}$$

In order to avoid the confusion, the *i-th* row of this matrix

$$y_i = (y_{i1}, y_{i2}, ..., y_{ip})$$

will be denoted by $y^i = (y_1^i, y_2^i, ..., y_p^i)$.

Reaction (or response) obtained from *i-th* individual forms the vector (the *i-th* row of data matrix):

$$y^i = (y_1^i, y_2^i, ..., y_p^i)$$

In the case of a generic individual, a simplified notation can be used:

$$y = (y_1, y_2, ..., y_p).$$

For typographical reasons, instead of y_j^i , a symbol y_{ij}. will be used.

In the case of comparative responses, the subject's task is to decide which item from a pair (Y_j, Y_k) is more dangerous for social cohesion. The results of this survey form the following matrix:

$$\begin{bmatrix} n_{11} & n_{12} & \ldots & n_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ n_{p1} & n_{p2} & \cdots & n_{pp} \end{bmatrix}$$

Entry n_{jk} stands for the number of respondents who asserted that Y_j is at least as dangerous as Y_k. For convenience, we put $n_{jj} = n$, $j = 1,2,..., p.$

2. General model for categorical responses

There is a long tradition of investigating categorized data, particularly data obtained from the responses of people. Usually, the existence of some hypothetical variable "driving" these responses is assumed. The categorized responses are supposed to be determined by the individual's position in underlying space of latent characteristics, also called latent variables or latent traits. Usually, this space is a one-dimensional continuum.

The significance of the latent traits, on which much of the psychological theory is based, has been precisely explained by Birnbaum: "nowhere is there any necessary implication that traits exist in any physical or psychological sense. It is sufficient that person behaves as if he were in the possession of a certain amount of each of a number of relevant traits and that he behave as if these amounts substantially determined his behaviour" (see [14]).

All responses y_{ij} are supposed to be random outcomes. The response $y_{ij} \in \{0,1\}$ given by the *i-th* individual to the *j-th* item is interpreted as a realization of a Bernoullian random variable Y_{ij} with a distribution

$$P(Y_{ij} = 1), P(Y_{ij} = 0) = 1 - P(Y_{ij} = 1).$$

In the case of a generic individual, a simpler notation will be used:

$$P(Y_j = 1), P(Y_j = 0) = 1 - P(Y_j = 1).$$

The probability distribution of the responses to all the items is denoted as

$$f(y) = \mathrm{P}(\mathrm{Y}_1 = y_1, ...Y_p = y_p).$$

This distribution is supposed to depend on a random variable Z, and the conditional distribution is denoted by

$$f(y|z) = P(Y_1 = y_1, ..., Y_p = y_p) \big| Z_1 = z_1, ..., Z_p = z_p).$$

Using the rule that $P(A) = P(A|B) \cdot P(B)$ we can infer the following representation:

$$f(y) = \int f(y|z) dH(z).$$

This is the fundamental representation of the probability distribution of the observed data. It forms the basis for all theories commonly called latent variables models (see [6]). Unfortunately, in such general settings, it has no use in practice, as it can be neither falsified nor verified.

To make the settings more restrictive, three fundamental hypotheses (assumptions) are introduced (see [15,24]).

The first hypothesis.

The first restriction requires that people with a greater level of the susceptibility are more inclined to react to the stressful phenomena. In other words, more susceptible people can be easier provoked by stressors beyond their endurance. Using more technical language, this assumption can be formulated as follows:

Probability of an affirmative response to an item representing a stressful phenomenon is a non-decreasing function of the respondent's susceptibility.

This assumption is called monotonicity (M), and formally it is expressed as follows:

(M) $P(Y=1|Z=z)$ is a coordinate-wise non-decreasing function in Z.

The second hypothesis.

The second assumption, known as conditional independence or latent independence (LI) requires that reactions of one subject to different stressors are independent.

Formally it takes the following form:

$$\text{(LI)} \qquad P(Y_1=y_1,...,Y_p=y_p|Z_1=z_1,...,Z_d=z_d)=\prod_{j=1}^{p}P(Y_j=y_j|Z=z)$$

Using notations

$$f(y|z)=P(Y_1=y_1,...,Y_p=y_p)|Z_1=z_1,...,Z_p=z_p)$$

$$f_j(y_j|z)=P(Y_j=y_j|Z_1=z_1,...,Z_d=z_d)$$

this condition can be written down as follows:

$$f(y|z)=\prod_{j=1}^{p}f_j(y_j|z)\,.$$

Hence, the basic representation will take the form:

$$f(y)=\int\prod_{j=1}^{p}f_j(y_j|z)dH(z)\,.$$

To be consistent with the main stream of publications the following notation is introduced:

$$\pi_{ij}(z)=P(Y_{ij}=1|Z=z)=P(Y_{ij}=1|Z_1=z_1,Z_2=z_2,...,Z_d=z_d).$$

More compact form of it will also be used:

$$\pi_{ij}(z) = P(Y_{ij} = 1 | Z = z).$$

Omitting the index of the individual, it will take the form:

$$\pi_j(z) = P(Y_j = 1 / Z = z)$$

Taking into account that:

$$P(Y_j = 0 / Z = z) = 1 - \pi_j(z)$$

and that $y_j \in \{0,1\}$ one arrives at the distribution

$$f_j(y_j / z) = P(Y_j = y_j / Z = z) = \pi_j(z)^{y_j} (1 - \pi_j(z))^{1-y_j}.$$

As a result, the following representation is obtained:

$$f(y) = \int \prod_{j=1}^{p} \pi_j(z)^{y_j} (1 - \pi_j(z))^{1-y_j} \, dF(z)$$

The probability distribution function for observed (manifest) data for the *i-th* subject will take the following form:

$$f(y^i) = \int \prod_{j=1}^{p} \pi_{ij}(z)^{y_{ij}} (1 - \pi_{ij}(z))^{1-y_{ij}} \, dF(z).$$

The third hypothesis.

The third assumption requires that the susceptibility is a one-dimensional continuum. It is called unidimensionality (U), and its formal expression is very simple:

(U) d=1

When Z is unidimensional, probability $\pi_{ij}(z)$ treated as a function of z is usually called item characteristic function, or item characteristic curve (ICC).

3. General implications

The basic representation

$$f(y) = \int f(y|z) dH(z)$$

with the conditions (LI), (M) and (U) defines reasonable non-parametric model for the observed data. The three conditions: (LI), (M), and (U) are minimal assumptions necessary to obtain a falsifiable model for the analysis of the perceived stressors.

It means that from this model one can infer a number of properties for the observed data. The most important implications are briefly summarized below.

First of all, it is worthy to note that no two of three assumptions define a restrictive model. If any of these three conditions is completely relaxed, then the resulting "model" will fit any distribution of binary data (see[19]). For example, it is well known that the assumption of unidimensionality and local independence are independent in the sense that neither, either one or both assumptions, can hold for the above representation (see [29,30]).

What follows from this result is the suggestion for testing of the hypothesis.

For example, the evidence that follows from condition (LI) and the lack of fit is that $d \neq 1$. Condition $d = 1$ and the lack of fit might be considered as the evidence of non-local independence.

The necessary conditions for the representation $f(y)$ were found by P. W. Holland in 1981 (see[16]). By weakening the condition of local independence ([16,18]) to the form of local non-negative dependence (LND) it was possible to establish the necessary as well as the sufficient conditions.

The (LI) and (M) conditions imply that vector $(Y_1, Y_2, ..., Y_p)$ is *associated*, and this means that for all monotone item summaries expressed by non-decreasing functions g_1 and g_2 holds the inequality: $Cov(g_1(Y), (g_2(Y)) \geq 0$. Furthermore, more powerful result was obtained (see[19,28]):

$$Cov(g_1(Y_A), (g_2(Y_A) \mid h(Y_B) = z) \geq 0$$

where g_1 and g_2 are non-decreasing functions, h is any function and $\{Y_A, Y_B\}$ is a partition of the set $\{Y_1, Y_2, ..., Y_p\}$.

The interpretation of this important theorem is given in [19,20].

To detect the violation of (LI) condition, one can use the conditional covariance functions $Cov(Y_i, Y_j / Z = z)$. If conditions (LI) and (U) hold, then conditional covariance functions are identically equal to zero:

$$Cov(Y_i, Y_j / Z = z) = 0 \text{ for all z, and all pairs } i \text{ and } j.$$

In practice, however, condition (LI) never holds exactly. For this reason, in order to provide tests and measures of unidimensionality, the concept of *essential independence* with respect to the latent characteristic Z was introduced (see [29,30]). This concept requires that the average value of $\left|Cov(Y_i, Y_j / Z = z)\right|$ over all item pairs Y_i and Y_j should be small for all values z, instead of being equal to zero.

Using this concept, the concept of *the essential unidimensionality* of a set of items $Y_1, Y_2, ..., Y_p$ has been defined as the minimal dimensionality necessary to satisfy the assumption of the essential independence. The statistical procedure for

testing the null hypothesis of the essential unidimensionality has been proposed by Stout in 1987 (see[29]). This rather complicated procedure requires the construction of an estimate of $Cov(Y_i, Y_j \mid Z = z)$ at each value of Z. One way to obtain this estimate is by kernel smoothing. Almost all tests of fit that have been already proposed for non-parametric models of item responses are rather complicated and usually their asymptotic properties cannot be determined mathematically (see[28]).

One can distinguish the following three important particular cases of the general settings (see[17]):

1. Non-parametrical models.

   Probabilities $\pi_j(z)$ and $H(z)$ are both allowed to vary over all non-parametric classes of functions.

2. Semi-parametric models.

   In this class $\pi_j(z)$ have parametric form, and $H(z)$ is allowed to vary over all non-parametric classes of functions.

3. Parametric models.

   In this class, $\pi_j(z)$ and $H(z)$ have parametric forms (see[17]):

$$\pi_j(z) = \pi_0(a_j(z - b_j), c_j)$$

$$H(z) = H_0(z - \frac{\mu}{\sigma}, v)$$

where π_0 and H_0 are specified functions, and z is a location parameter (see[17]).

4. Simple logistic model

4.1. Model derivation

Let us assume the following conditions:

1) each respondent is characterized by latent parameter z_i and to each stressor a real number α_j characterizing its "capability" to provoke distress can be assigned,
2) functions $\pi_j(z) = \pi_j(y_j|z)$ are continuous and strictly monotone increasing for all latent trait parameters z,
3) $\lim_{z\to-\infty} \pi_j(z) = 0$, $\lim_{z\to\infty} \pi_j(z) = 1$
4) the "principle of local stochastic independence" is satisfied for all responses,

5) the row score $T_i = \sum_{j=1}^{p} Y_{ij}$ is sufficient statistics for parameter z_i.

It has been proved (see for example [1,14]) that if these conditions are satisfied, then π_j has the following form:

$$\pi_j(z_i) = P(Y_j = 1/Z = z_i) = \frac{\exp(z_i - \alpha_j)}{1 + \exp((z_i - \alpha_j)}$$

This model is called a simple logistic model, or Rasch model (see [27]).

It depends on $n+p$ parameters:

$$\alpha_1, \alpha_2, ..., \alpha_p, z_1, z_2, ..., z_n .$$

Parameters determining susceptibility of the respondents $z_1, z_2, ..., z_n$ are treated as nuisance parameters.

When estimating these parameters one can follow one of three possible ways:

1) nuisance parameters can be estimated simultaneously with the parameters determining the intensity of the stressors,
2) nuisance parameters can be eliminated by the conditioning,
3) nuisance parameters can be integrated out.

In the first case, joint maximum likelihood (JML) method can be used, in the second case – conditional maximum likelihood (CML) method, and in the third one - marginal maximum likelihood (MML) method.

4.2. Elementary method

The brief review of the above mentioned methods starts with the presentation of the simplest method. In [22] it is called *elementaren Schätzungen.*

Suppose that the Rasch model is valid. This means that the equality

$$\pi_{ij} = \frac{\exp(z_i - \alpha_j)}{1 + \exp((z_i - \alpha_j)}$$

$$\text{holds for all } i = 1, 2, ..., n, \; j = 1, 2, ..., p .$$

The equivalent form of these equations is as follows:

$$Log(\pi_{ij}) = z_i - \alpha_j$$

Replacing π_{ij} by the observed frequency $\frac{n_{ij}}{n}$, parameters z_i and α_j can be found by the solution of the following system of $n \cdot p$ equations:

$$\log\left(\frac{n_{ij}}{n}\right) = z_i - \alpha_j$$

The solution is following [22]:

$$\hat{z}_i = \bar{y}_{i\cdot} - 0{,}5$$
$$\hat{\alpha}_j = \bar{y} - \bar{y}_{\cdot j}$$

where

$$\bar{y}_{i\cdot} = \frac{1}{p}\sum_{j=1}^{p} y_{ij} \quad , \quad \bar{y}_{ij} = \frac{1}{n}\sum_{i=1}^{n} y_{ij}, \qquad \bar{y} = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij},$$

Theoretically, this way of estimation is not satisfactory since one cannot make any inferences. Practically, it is justified by its simplicity and, as reported in [22], the results are very close to those obtained by the sound methods.

4.3.Joint maximum likelihood

In order to use the JML let us observe that

$$f(y;z,\alpha) = \prod_{j=1}^{p} \pi_{ij}^{\,y_{ij}} (1-\pi_{ij})^{1-y_{ij}} = \prod_{j=1}^{p} \left[\frac{\exp(z_i - \alpha_j)}{1+\exp((z_i - \alpha_j)}\right]^{y_{ij}} \left[\frac{1}{1+\exp((z_i - \alpha_j)}\right]^{1-y_{ij}}$$

Hence, the joint maximum likelihood function is the following:

$$L(z_1, z_2, ..., z_n, \alpha_1, \alpha_2, ..., \alpha_p) = \prod_{j=1}^{n}\prod_{i=1}^{p} \pi_{ij}^{\,y_{ij}} \left[1-\pi_{ij}\right]^{1-y_{ij}} =$$

$$= \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{p} z_i y_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{p} \alpha_i y_{ij}\right)\left(\prod_{j=1}^{n}\prod_{i=1}^{p}\left(1+e^{z_i-\alpha_j}\right)\right)^{-1} =$$

$$= \exp\left(\sum_{i=1}^{n} z_i t_i - \sum_{j=1}^{p} \alpha_j q_j\right)\left(\prod_{i=1}^{n}\prod_{j=1}^{p}\left(1+\exp\left(z_i-\alpha_j\right)\right)\right)^{-1}$$

where

$$t_i = \sum_{j=1}^{p} y_{ij} \quad q_j = \sum_{i=1}^{n} y_{ij}$$

Estimators of the parameters $\alpha_1, \alpha_2, ..., \alpha_p$ and $z_1, z_2, ..., z_n$ are to be obtained by solving the equations:

$$\frac{\partial \ln L(z_1, z_2, ..., z_n, \alpha_1, \alpha_2, ..., \alpha_p)}{\partial z_i} = 0$$

$$\frac{\partial \ln L(z_1, z_2, ..., z_n, \alpha_1, \alpha_2, ..., \alpha_p)}{\partial \alpha_j} = 0$$

Identifability requires that either

$$\sum_{i=1}^{n} z_i = 0 \quad \text{or} \qquad \sum_{j=1}^{p} \alpha_j = 0 .$$

After simplification:

$$t_i = \sum_{j=1}^{p} \frac{\exp(z_i - \alpha_j)}{1 + \exp(z_i - \alpha_j)}, \quad i = 1,2,...,n$$

$$q_j = \sum_{i=1}^{n} \frac{\exp(z_i - \alpha_j)}{1 + \exp(z_i - \alpha_j)}, \quad j = 1,2,..., p$$

The solution of these equations can be obtained only numerically.

This solution has however some theoretical drawbacks (see[1, 3]).

The number of nuisance parameters grows with the sample size *n*, and item parameters are not consistent (see[14]).

4.4. Conditional maximum likelihood

The drawbacks of JML method can be avoided by using theoretically satisfactory method known as conditional maximum likelihood (CML) method. This method forms two separate likelihoods: one for the item parameters, and the second for the respondent parameters (see [6]).

To obtain the estimating equations, let us observe that

$$f(y) = f(y_{i1}, y_{i2}, ..., y_{ip}|) = f(y|t_i) g(t_i|z_i)$$

Hence, the likelihood function has the form:

$$L(z_1, z_2, ..., z_n, \alpha_1, \alpha_2, ..., \alpha_p) = \prod_{i=1}^{n} f(y_i|t_i) g(t_i|z_i) = \prod_{i=1}^{n} f(y_i|t_i) \prod_{i=1}^{n} g(t_i|z_i) = L_C \cdot L_M$$

Part L_C is used for the estimation of α parameters, and L_M part is used for the estimation of z parameters.

The required conditional distribution

$$f(y|t_i) = P(Y = y|T_i = t_i) = f(y_{i1}, y_{i2},..., y_{ip}|T_i = t_i)$$

is obtained by dividing *f(y)* by $g(t_i|z_i)$.

Let us observe that

$$f(y) = f(y_{i1}, y_{i2},..., y_{ip}) = \frac{\exp(z_i t_i - \sum_{j=1}^{p} \alpha_j y_{ij})}{\prod_{j=1}^{p}(1 + \exp((z_i - \alpha_j))},$$

then, by summation over all possible vectors $(y_{i1}, y_{i2},..., y_{ip})$ with the total $t_i = \sum y_{ij}$
one gets

$$g(t_i|z_i) = P(T_i = t_i) = P(\sum_{j=1}^{p} Y_{ij} = t_i) = \frac{\sum_{=t_i} \exp(z_i t_i - \sum_{j=1}^{p} \alpha_j y_{ij})}{\prod_{j=1}^{p}(1 + \exp((z_i - \alpha_j))}$$

where symbol $\sum_{=t_i}$ means that the summation is extended to all observations $(y_{i1}, y_{i2},..., y_{ip})$ such that $y_{i1} + y_{i2} + ... + y_{ip} = t_i$.

Hence, the conditional distribution $f(y|t)$ has the following form (see[1,5])

$$f(y|t_i) = f(y_{i1}, y_{i2},..., y_{ip}|T_i = t_i) = \frac{\exp(-\sum_{j=1}^{p} \alpha_j y_{ij})}{\gamma_{t_i}}$$

where $\gamma_{t_i} = \sum_{=t_i} \exp(-\sum_{j=1}^{p} \alpha_j y_{ij})$

The conditional likelihood L_C is independent of z_i (as it should be), and the estimates are determined by solving the equations (see [1,5]):

$$\frac{\partial L_c}{\partial \alpha_j} = 0.$$

For the estimation of z_i, the L_M part of the likelihood function is used.

Let us write down the function $g(t_i|z_i)$ in the equivalent form as follows (see [4]):

$$f(t_i|z_i) = \frac{\exp(z_i t_i)}{\prod_{j=1}^{p}(1+\exp((z_i-\alpha_j))} \cdot \sum_{=t_i} \exp(-\sum_{j=1}^{p}\alpha_j y_{ij})$$

Estimates of z_i are obtained by the maximization of the following likelihood function, known as marginal likelihood (see[4]):

$$L_M = L_M(\alpha_1, \alpha_2, ..., \alpha_p, z_1, z_2, ..., z_n) = \prod_{i=1}^{n} g(t_i|z_i).$$

System $\frac{\partial L_M}{\partial z_i} = 0$ is equivalent to (see [4])

$$t_i = \sum_{j=1}^{p} \frac{\exp(z_i - \alpha_j)}{1+\exp(z_i-\alpha_j)}$$

which can be solved for z_i replacing parameters α_j with their estimates $\hat{\alpha}_j$ obtained by conditional likelihood method.

4.5. Population likelihood

Suppose now that people's susceptibility to stressful phenomena is interpreted as a real valued random variable Z with a density distribution $h(z)$. Respondents are treated now as a random sample, and each z_i , $i = 1,2,...,n$ is interpreted as the realization of random variable Z_i, with distribution given by $h(z)$ or $H(z)$.

Furthermore, assume that $Z \sim N(\mu, \sigma^2)$. The problem is in the estimation of μ and σ^2. Parameters μ and σ^2 are estimated by the means of the so-called population likelihood function L_p. To derive this function let us observe that the probability density function of observed data can be expressed as follows (see [1,2,4]):

$$f(y) = \int f(y|t_i) \cdot g(t_i|z) dH(z) = \int f(y|t_i) \cdot g(t_i|z) h(z, \mu, \sigma^2) dz$$

hence, the likelihood function will have the form:

$$L(\alpha_1,\alpha_2,...,\alpha_p,\mu,\sigma^2)=\prod_{i=1}^{n} f(y_i|t_i)\prod_{i=1}^{n}\int g(t_i|z)h(z,\mu,\sigma^2)dz = L_C \cdot L_P$$

We see that it is a product of two functions L_C and L_P. For the estimation of parameters μ and σ^2 two stage procedure is applied.

First, parameters α_j , $j=1,2,...,p$ are estimated using the conditional likelihood function L_C . Then, they are treated as known, and they are used in the population likelihood function L_P for the estimation of μ and σ^2 .

It has been proven (see for example [3]) that observed score $T_i = Y_{i1} + Y_{i2} + ... + Y_{ip}$ is sufficient for (μ,σ^2). Statistics T_i takes on values $t=0,1,...,\mathrm{n}$. Let n_t be a number of response vectors with score t, i.e. n_t is a number of individuals with score t.

As respondents are sampled randomly, vector $(n_0,n_1,...,n_p)$ is to be interpreted as a realization of random vector $(N_0,N_1,...,N_p)$ which follows a multinomial distribution

$$M(n,g_0,g_1,...,g_p)$$

i.e.

$$P(N_0=n_0,N_1=n_1,...,N_p=n_p)=\frac{n!}{n_0!n_1!...n_p!}g_0^{n_0}g_1^{n_1}...g_p^{n_p}$$

Probabilities g_t are the following (see [3,4]):

$$g_t = g_t(\mu,\sigma^2)=\sum_{=t}\exp(-\sum_{j=1}^{p}\alpha_j y_{ij})\int \exp(z\cdot t)h(z,\mu,\sigma^2)\prod_{j=1}^{p}(1+\exp(z-\alpha_j))^{-1}dz$$

Population likelihood function is the following (see [1]):

$$L_p(\alpha_1,\alpha_2,...,\alpha_p,\mu,\sigma^2)=\prod_{t=0}^{p} g_t^{n_t}$$

Computational procedures for maximization of this function with the respect to μ and σ^2 are rather complicated (see [4]).

Statistics for testing goodness of fit is the following (see [3]):

$$z = 2\sum_{t=0}^{p} n_t[\ln n_t - \ln(ng_t(\hat{\mu},\hat{\sigma}^2))]$$

which is approximately χ^2 -distributed with *p-2* degrees of freedom for large *n*.

5. Feelings of morale classification

Some phenomena, particularly those of the political nature, might provoke drastically different reaction in different groups of people.

In the simplest case, the society under the investigation (respondents) could be divided into two classes. These classes could be called, for example, "content" and "malcontent", or "sensible" and "insensible". In such a dichotomized situation one can assume that the latent trait Z is a binary random variable with distribution $\eta = P(Z = 1) = p$ (respondent is content), $1-\eta = P(Z = 0) = p$ (respondent is malcontent).

Belonging to either of classes can be determined based on binary responses to items $Y_1, Y_2, ..., Y_p$.

Introducing notations

$$\pi_{j1} = P(Y_j = 1 \mid Z = 1) \quad , \qquad \pi_{j0} = P(Y_j = 1 \mid Z = 0) \quad ,$$

the basic representation for observed data

$$f(y) = \int g(y \mid z) dH(z)$$

will take the following form

$$f(y) = \eta \prod_{j=1}^{p} \pi_{j1}^{y_j} (1-\pi_{j1})^{1-y_j} + (1-\eta) \prod_{j=1}^{p} \pi_{j0}^{y_j} (1-\pi_{j0})^{1-y_j} \ .$$

In more general case we can assume that respondents belong to c classes indicated by numerals $1,2,...,c$.

Let us introduce notations:

$$w_k = P(W = k) \quad , \quad k = 1,2,...,c$$
$$\pi_{jk} = P(Y_j = 1 \mid Z = k)$$
$$g_j(y_j \mid z_k) = P(Y = y_j \mid Z = z_k) = \pi_{jk}^{y_j} (1-\pi_{jk})^{1-y_j} ,$$

then

$$f(y) = \sum_{k=1}^{c} w_k \prod_{j=1}^{p} g_j(y_j) = \sum_{k=1}^{c} w_k \cdot \prod_{j=1}^{p} \pi_{jk}^{y_j} (1-\pi_{jk})^{1-y_j} .$$

This probability distribution function is a finite mixture density and it depends on $pc + c$ parameters:

$$w_1, w_2, ..., w_c, \pi_{11}, \pi_{12}, ..., \pi_{1c}, ..., \pi_{p1}, \pi_{p2}, ..., \pi_{pc}.$$

The log-likelihood function is as follows (see [6,12,13]):

$$L = \sum_{i=1}^{n} \ln(\sum_{k=1}^{c} w_k \prod_{j=1}^{p} g_j(y_{ij} | z_k)),$$

where z_k stands for the event "$Z = k$."

Remembering that $\sum_{k=1}^{c} w_k = 1$, the estimates are derived from the equations (see [6,12]):

$$\hat{w}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{P}(k | y_k^i) \quad , \quad k = 1,2,...,c$$

$$\hat{\pi}_{jk} = \frac{1}{n \cdot \hat{w}_k} \sum_{i=1}^{n} y_{ij} \cdot \hat{P}(k | y^i) \quad , \quad j = 1,2,..., p, k = 1,2,..., c$$

Symbol $\hat{P}(k | y^i)$ is the estimate of probability that the individual with response vector $y^i = (y_1^i, y_2^i, ..., y_p^i) = (y_{i1,} y_{i2}, ..., y_{ip})$ belongs to the class *k*.

Although these equations have a simple form, posterior probability has a rather complicated form:

$$\hat{P}(k | y^i) = \frac{\hat{w}_k \prod_{j=1}^{p} \hat{\pi}_{jk}^{y_j^i} (1 - \hat{\pi}_{jk})^{1-y_j^i}}{\sum_{k=1}^{c} w_k \prod_{j=1}^{p} \hat{\pi}_{jk}^{y_j^i} (1 - \hat{\pi}_{jk})^{1-y_j^i}}.$$

For the solution the E-M algorithm has to be used (see [6]).

6. Factor analysis

Let us assume now that the reaction (response) to stressor Y_j is proportional to susceptibility Z and it is additively disturbed by a random error U_j.

Denoting the proportionality coefficient by α_j, we have $Y_j = \alpha_j Z + U_j$

If $Z \sim N(0,1)$ and $U_j \sim N(0, \sigma_j)$, and they are independent, then by introducing variable ε_j with a standard normal distribution, we can write

$$Y_j = \alpha_j Z + \sigma_j \varepsilon_j$$

This means that the response is a linear combination of susceptibility and random disturbances.

Without loosing generality, one can assume that Y_j is standardized by its standard deviation, i.e. $Var(Y_j) = 1$.

From this we have

$$Var(Y_j) = Var(\alpha_j Z) + Var(\sigma_j \varepsilon_j) = \alpha_j^2 + \sigma_j^2 = 1.$$

which implies that $\sigma_j^2 = 1 - \alpha_j^2$.

Hence, we have the following representation (see also [7,8]):

$$Y_j = \alpha_j Z + \sqrt{1 - \alpha_j^2} \cdot \varepsilon_j$$

The next crucial assumption is that we are not able to observe reaction to the stressor Y_j as it escalates. We are only able to note when the intensity of a stressor reaches a level when the respondent bursts out in anger (irritation).

Denoting this level by symbol γ_j we can define the new variable Y_j^* as follows (see [10]):

$$Y_j^* = \begin{cases} 1 & if \quad Y_j \geq \gamma_j \\ 0 & if \quad Y_j < \gamma_j \end{cases}$$

The probability that a subject with the susceptibility *z* bursts out, i.e. this subject endorses item Y_j is the following

$$P(Y_j^* = 1 \Big| z) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\gamma_j}^{\infty} \exp(-(x - \alpha_j z)^2 / 2\sigma^2) dx$$

Let

$$x_j = -\frac{\gamma_j - \alpha_j \cdot z}{\sigma}$$

then using Φ for cumulative distribution function of the standard normal variable, we have $P(Y_j^* = y_j \big| z) = \begin{cases} \Phi(x_j) & , \quad when \quad y_j = 1 \\ 1 - \Phi(x_j) & , \quad when \quad y_j = 0 \end{cases}$

The probability of the response vector $y = (y_1, y_2, ..., y_p)$ is given by

$$\prod_{j=1}^{p} P(Y_j^* = y_j | z)$$

Let us observe that all response vectors are divided into 2^p mutually exclusive categories (patterns). These categories are conveniently indexed by the decimal numbers:

$$k = 1 + \textit{binary number } y_{k1}y_{k2}...y_{kp} \quad , \quad k = 1,2,...,2^p$$

If P_k denotes the probability of *k-th* category, then we have (see [8])

$$P_k = \int_{-\infty}^{\infty} \prod_{j=1}^{p} P(Y_1^* = y_{k1}, Y_2^* = y_{k2}, ..., Y_p^* = y_{kp} | z) \varphi(z) dz$$

The definite integral in this formula cannot be expressed in the closed form, in [7] it is evaluated by Gauss-Hermite quadrature.

Remember that P_k determines the probability of the k-th category of response vector. This probability depends on threshold values $\gamma_1, \gamma_2, ...\gamma_p$ and correlations $\alpha_1, \alpha_2, ...\alpha_p$, which in turn are interpreted as the strength of stressors $Y_1, Y_2, ...Y_p$. The numbers N_k are multinomially distributed with the parameters n and P_k.

Therefore, the likelihood function

$$L = \frac{n!}{\prod_k n_k!} \prod_{k=1}^{2^p} P_k^{n_k}$$

is maximized with respect to α_j and γ_j using Newton-Raphson method.

Christofferson proposed simpler method of the generalized least squares using only first-order and second-order marginal proportions.

7. Stressor scaling by paired comparisons

Definition of models based on paired comparisons requires different assumptions (see[11]). First of all, one has to assume that each stressor can be located on a sensation scale. To determine this location, stressors are presented to individuals for judgement. They are presented in pairs, so that an individual can evaluate which stressor is more intensive.

Let us assume that stressors $Y_1, Y_2, ..., Y_p$ have intrinsic force to cause distress in the population, or to irritate people. As a result, a disruption of social

ties can occur. The strength of the negative impact of stressor Y_j on social cohesion is denoted as α_j, $j = 1,2,...,p$.

It is assumed that there is a probability space on which following random variables Y_{ij} are defined:

$$Y_{ij} = Y_i - Y_j .$$

and interpreted as the amount of predominance of Y_i over Y_j.

Quantities α_j are then treated as positional parameters of Y_j.

Assuming that ε_{ij} is a random variable having density $f(u)$ symmetric about 0, we arrive at the model

$$Y_{ij} = \alpha_i - \alpha_j + \varepsilon_{ij} .$$

Let π_{ij} denote the probability of the predominance of Y_i over Y_j:

$$\pi_{ij} = P(Y_i \prec Y_j) = P(Y_i - Y_j > 0) .$$

Because of symmetry of $f(u)$ the above formula will have the following form:

$$\pi_{ij} = P(Y_{ij} > 0) = P(\alpha_i - \alpha_j + \varepsilon_{ij} > 0)$$

$$= 1 - P(\alpha_i - \alpha_j + \varepsilon_{ij} \le 0) = F(\alpha_i - \alpha_j) = F(\delta_{ij})$$

where $F(u)$ is the cumulative distribution function (see[25,26]).

In the simplest case, the function of the uniform distribution over the interval $(-0{,}5 \;\; 0{,}5)$ can be used:

$$H(u) = \frac{1}{2} + u, \quad -0.5 \le u \le 0.5 .$$

However, either normal or logistic model is usually considered.

In the first case we have

$$\pi_{ij} = \frac{1}{\sqrt{2\pi}} \int_{-(\alpha_i - \alpha_j)}^{\infty} e^{-\frac{1}{2}z^2} dz$$

and in the second case we have (see[6,9]):

$$\pi_{ij} = \int_{-(\alpha_i - \alpha_j)}^{\infty} \frac{e^z}{(1+e^z)^2} dz = \frac{1}{1+e^{\alpha_j - \alpha_i}}$$

This model is known as Bradley-Terry model.

Positional parameters $\alpha_1, \alpha_2, ..., \alpha_p$ determine the scale values for stressors $Y_1, Y_2, ..., Y_p$.

With no loss of generality it may be assumed that

$$\sum_{i=1}^{p} \alpha_i = 0.$$

This equality guaranties the uniqueness of estimates $\hat{\alpha}_j$, $j = 1,2,..., p$.

For the estimation of these parameters paired comparisons are used.

If by n_{ij} we denote the number of individuals (respondents, experts, judges) who considered stressor Y_i more stressful than Y_j, then we can conclude that n_{ij} is a binomial variable

$$n_{ij} \sim B(n, \pi_{ij})$$

where n is the number of experts, and π_{ij} is the probability of "success".

It is not difficult to infer that

$$\alpha_j = \frac{1}{n} \sum_{k=1}^{n} \delta_{kj}$$

This equality suggests the method of estimation:

$$\hat{\alpha}_j = \frac{1}{n} \sum_{k=1}^{n} d_{kj}$$

where values of d_{kj} are to be obtained from the system of equations:

$$F(d_{ij}) = \frac{n_{ij}}{n}, \quad i,j = 1,2,...,p$$

The ratio $\frac{n_{ij}}{n}$ is the maximum likelihood estimate $\hat{\pi}_{ij}$ of π_{ij}.

If $n_{ij} = 0,$ than it is suggested to put $\hat{p}_{ij} = \frac{1}{2n}$, and in the case when $n_{ij} = 1$, to put $\hat{p}_{ij} = 1 - \frac{1}{2n}$.

Mosteller proved that estimates $\hat{\alpha}_j$ are least squares estimates (see[25,26]).

For testing the hypothesis that there is no difference between all stressors

$$\alpha_1 = \alpha_2 = ... = \alpha_p$$

one can use the following statistics (see[26]):

$$A_f = 4npf^2(0)\sum_{j=1}^{p}\hat{\alpha}_j^2$$

where the subscript *f* indicates that $\hat{\alpha}_j$ have been computed using the function $f(u)$.

Under the null hypothesis, A_f has an asymptotic chi square distribution with *p-1* degrees of freedom as number of sampled respondents tends to infinity.

8. Conclusions

From the discussion presented in this paper it follows that the statistical methods developed in different fields of psychology, education and bioassay can be easily adopted for modelling of the social phenomena. Particularly, the methods of item response theory can be directly used for social stressors analysis. Merely little changes in the interpretation of parameters are needed.

REFERENCES

1. ANDERSEN E.B. (1980): Discrete Statistical Models with Social Science Applications, North-Holland, Amsterdam.
2. ANDERSEN E.B., Latent trait models , Journal of econometrics, 22, 1983, 215–227
3. ANDERSEN E.B., Comparing latent distributions, Psychometrika, 45, 1980, 121–134
4. ANDERSEN E.B., MADSEN M., Estimating the parameters of the latent population distribution, *Psychometrika* 42, 1977, 357–374

5. ANDRICH D., Rasch models for measurement, SAGE University Paper, 1988
6. BARTHOLOMEW D.J., KNOTT M., Latent variable models and factor analysis, ARNOLD, London, 1999
7. BOCK R.D., AITKIN M. (1981): Marginal maximum likelihood estimation of item parameters: application of an EM algorithm, *Psychometrika* 46, pp. 443–459
8. BOCK R.D., LIEBERMAN M.(1970): Fitting a response model for n dichotomously scored items, *Psychometrika* 435(2), pp. 179–197.
9. BRUNK H. D. (1960), Mathematical models for ranking from paired comparisons, American Statistical Association Journal, 9, 503–21
10. CHRISTOFFERSON A. (1975): Factor analysis of dichotomised variables, *Psychometrika* 40(1), pp. 5–31.
11. DAVID H.A., The method of paired comparisons, Griffin, London, 1969
12. EVERITT B.S.(1984): An introduction to latent variable models, Chapman& Hall, London.
13. EVERITT B.S., HAND D.J.(1981): Finite mixture distributions, Chapman& Hall, London.
14. FISCHER G.H., MOLENAAR I.W. eds (1995): Rasch models: foundations, recent developments and applications, Springer-Verlag, New York.
15. HAMBLETON R.K, SWAMINATHAN H., ROGERS H.J. (1991): Fundamentals of itemresponse theory, Sage Publications, Newbury Park, CA.
16. HOLLAND P.W. (1981): When are item response models consistent with observe data?, *Psychometrika* 46, pp. 79–92.
17. HOLLAND P.W. (1990): On the sampling theory foundations of item response theory models, *Psychometrika 55,* pp. 577–601.
18. HOLLAND P.W., ROSENBAUM P.R. (1986): Conditional association and unidimensionality in monotone latent variable models, Annals of Statistics, vol. 14, pp. 1523–1543
19. JUNKER B.W. (1993): Conditional association, essential independence and monotone unidimensional item response models, *Annals of Statistics* 21, pp. 1359–1378.
20. JUNKER B.W., ELLIS J.L. (1997): A characterization of monotone unidimensional latent variable models, Annals of Statistics 25, pp. 1327–1343.

21. JUNKER B.W., SIJTSMA K., Nonparametric Item Response Theory in action, Applied Psychological Measurement,25, 2001, 211–220

22. KRAUTH J., Testkonstruktion und Testtheorie, BELTZ, 1995

23. LAZARSFELD P.F., HENRY N.W. (1968): Latent structure Analysis, Houghton-Mifflin, New York.

24. MOKKEN R.J. (1997): Nonparametric models for dichotomous responses, in W. van der Linden and R.K. Hambleton, eds, Handbook of Modern Item Response Theory, Springer-Verlag, New York pp. 351–367.

25. MOSTELLER F. (1951), Remarks on the method of paired comparisons, I, *Psychometrika* 16, pp. 3–9

26. NOETHER G., (1960), Remarks about a paired comparison model, *Psychometrika 25*, pp. 357–367

27. RASCH G. (1960): Probabilistic models for some intelligence and attainment tests, Pœdagogiske Institut, Copenhagen.

28. ROSENBAUM P. R.(1984), Testing the conditional independence and monotonicity assumptions of item response theory, *Psychometrika* , 49, 425–435.

29. STOUT W. (1987): A nonparametric approach for assessing latent trait unidimensionality, *Psychometrika* 52, pp. 589–617.

30. STOUT W. (1990): A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation, *Psychometrika* 55, pp. 293–325.

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 301—316

DEVELOPMENT OF THE REVISED WELL-BEING SCALE FOR CHINESE CITIZENS

Zhanjun Xing

ABSTRACT

This paper presents results of a second round of the studies in which a subjective well-being scale for Chinese citizens (SWBS-cc) has been developed in two consecutive steps. First, the scale composed of 54 items was applied to a sample of adult persons living in Shandong province, and a local norm was achieved. Next, a sample from a wider area in mainland of China was taken, and the psychometrics of SWBS-cc was examined. Based on that research,, the scale was revised and improved through factor analysis and logical analysis. The revised SWBS-cc included 40 items, and shown good psychometric properties. The standard norm sample was expanded to six metropolitan areas (capital cities) from the whole country, and the nationwide norm was achieved.

Key words: Subjective well-being, Psychological testing, Item analysis, norm.

1 Introduction

The subjective well-being (SWB) measures research appeared in the mid-20th century along with the research of SWB, and researchers from different traditions developed some widespread instruments (Bradburn & Caplovitz,1965; Kozma & Stones,1980; Ryff & Singer, 1996; Stones et.al., 1996; Diener et al., 1996; Cummins,1996; Luo,1998). In previous studies, we have developed a 54 items subjective well-being scale for Chinese citizens (SWBS-cc) based on these research. The SWBS-cc was developed through the view of experience. According to this view, subjective well-being is a positive or greatly satisfying state of existence that people experienced. This state of existence indicated if a person lived normally and at which degree a normal person achieved in material and spirit. The former indicated the people's state of physical and mental health. The later indicated the people's state of enjoyment and development, and it covers the main aspects of a human being's activity that he takes in a particular historical background (Zhanjun & Liqing, 2007). The regional norm of the scale has been established based on a sample from Shandong province (Zhanjun,2005).

A number of studies have showed that SWBS-cc has good measurement properties when applying to some special groups, such as the aged, youth, women, college student groups, post-graduate, teachers, rural residents and Christian group(Zhanjun,2003; Liqing,2004; Yu & Zhanjun, 2005;Huixiong & Limin,2005;Xinhua & Yu,2008; Yuanjiang & Zhanjun, 2009). However, there is still an obvious problem in these studies, that is mostly used samples from specific regions, thus, there is a big limitation to spread the scale to the nationwide. It is necessary to determine if the scale could be applied to the whole country and if it need to be revised. Moreover, the national norm of the revised scale also need to be established.

This paper consists of four relatively independent studies. Firstly the psychometric properties of SWBS-cc were examined in a broader area by sampling from another two capital cities. And then, as a clue, the outcomes of the psychometric properties examining were used to revise the scale to be more perfect. Finally, sampling from all the capital cities in Chinese mainland, we obtained the standard norm sample group to establish the national norm of revised SWBS-cc.

2 Study 1

2.1 Objective

Sampling in a broader area of Chinese mainland, we would test the psychometric properties of SWBS-cc(54 items) and examine the feasibility of its applying to further nationwide study. Study 1 was also designed to obtain clues for further revising of SWBS-cc through its psychometric properties examination.

2.2 Method

2.2.1 Participants

The participants were adult inhabitants from Shenyang (the capital city of Liaoning province) and Kunming (the capital city of Yunnan province) who lived there more than half a year and whose age were above 18 years (including 18 years). The sample size was 1244, and the demographic and social-economic characteristics could be summarized as following: Gender: male, 50.0%; female, 49.0%; missing,1.0%. Age: 18–34 years old, 42.4%; 35–54 years old, 47.6%; 55 years old and above,9.4%; missing,0.6%. Education: primary, 19.2%; secondary, 29.8%; higher, 49.5%; missing,1.5%.

2.2.2 Instruments

SWBS-cc. The scale consists of 54 items, and the respondent are required to respond to each item with 6-grade selection. The ten subscales are as following: experience of satisfaction and abundance, experience of mental health, experience

of growth and progress, experience of confidence towards society, experience of goal and personal value, experience of self-acceptance, experience of adaptation to interpersonal relation, experience of physical health, experience of psychological balance, experience of family atmosphere(Zhanjun,2005).

Personal Well-bing Index(PWI). The PWI scale contains seven items of satisfaction, each one corresponding to a quality of life domain as: standard of living, health, life achievement, personal relationships, personal safety, community-connectedness, and future security(Cummins,1996).

Satisfaction with life scale (SWLS). The scale was developed by ED Diener. It consists of five items, and the respondent are required to respond to each item with 7-grade selection response (Diener et al.,1985).

Singer-item self-report SWB scale (SISRSWBS). The scale contains only one question: "Overall, I am a happy people". The respondent are required to respond with a 7-grade likert-style selection.

2.3 Result

2.3.1 Reliability Analysis

Using the data of the sample from Shenyang and Kunming, we examined the internal consistency reliability of SWBS-cc and its ten subscales. The outcome showed that the full scale and its subscales have favorable reliability. The Cronbach's alpha coefficient of the full scale was 0.919, and the Cronbach's alpha coefficients of every subscale were also above 0.638 (see Table 1).

Table 1. The internal consistency reliability of ten subscales

Sub-scale1	Sub-scale2	Sub-scale3	Sub-scale4	Sub-scale5	Sub-scale6	Sub-scale7	Sub-scale8	Sub-scale9	Sub-scale10
0.851	0.797	0.829	0.814	0.820	0.723	0.740	0.736	0.683	0.638

**Data in the table were Cronbach's alpha coefficients.*

2.3.2 Criterion Validity Analysis

We used the pearson correlation coefficient between the scores on the full SWBS-cc (including the scores of its subscales) and the scores on PWI, SWLS and SISRSWBS as criterion index. The criterion validity of SWBS-cc applied to Shenyang and Kunming sample were 0.633, 0.400 and 0.520 respectively. All of the correlation coefficients were significant at 0.000 level. The SWBW-cc and its subscales showed favorable criterion validity (see Table 2).

Table 2. The criterion validity of SWBS-cc and its ten subscales

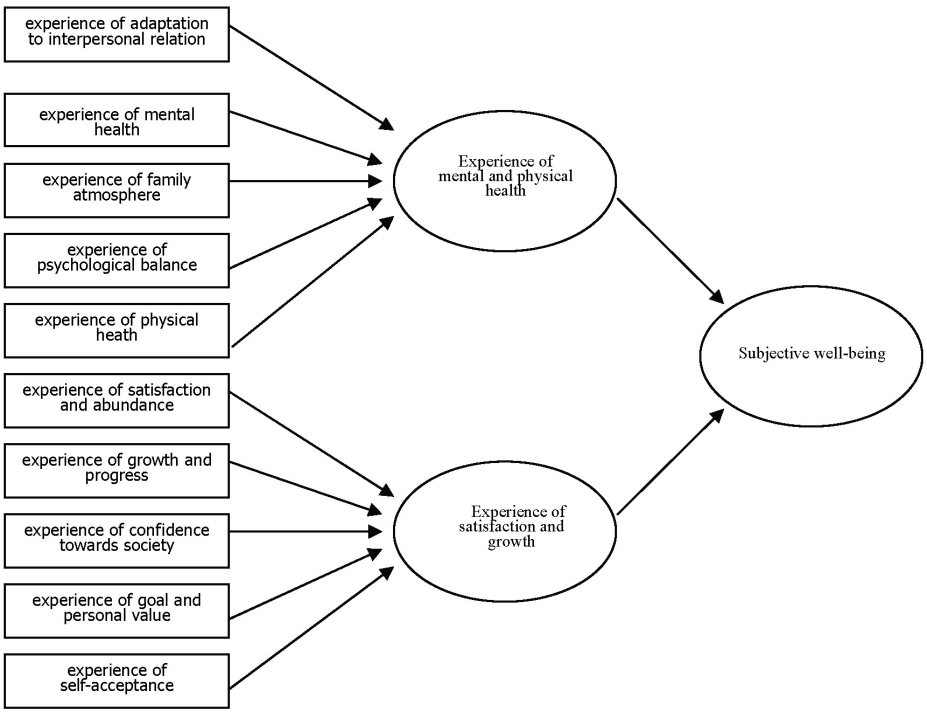
SWBS-cc	Sub-scale1	Sub-scale2	Sub-scale3	Sub-scale4	Sub-scale5	Sub-scale6	Sub-scale7	Sub-scale8	Sub-scale9	Sub-scale10
PWI	0.523	0.387	0.345	0.352	0.321	0.393	0.411	0.376	0.324	0.331
SWLS	0.594	0.148	0.242	0.118	0.158	0.202	0.131	0.334	0.032	0.144
SISRSWBS	0.402	0.243	0.300	0.383	0.250	0.340	0.275	0.343	0.251	0.449

**Data in the table were correlation coefficients.*

2.3.3 Structural Validity Analysis

Based on the former theory hypothesis and empirical exploration, we put forward a model of Chinese citizens' subjective well-being (see Chart 1).

Chart 1. The original structural model of Chinese people's subjective well-being



According to the model, the Structure of Chinese people's subjective well-being might be regard as two basic components. The first component was the experience of mental and psychological health(including experience of adaptation to interpersonal relation, experience of mental health, experience of family atmosphere, experience of psychological balance, and experience of physical health), and the second one was the experience of satisfaction and growth

(including experience of satisfaction and abundance, experience of growth and progress, experience of confidence towards society, experience of goal and personal value, experience of self-acceptance). The model's goodness-of-fit was testified through confirmatory factor analysis. It showed that the main goodness-of-fit index of the original model were unsatisfactory (see Table 3).

Table 3. The main goodness-of-fit index of the original model's testifying

x^2/df	GFI	AGFI	NFI	NNFI	CFI	IFI	RFI	RMR
20.9	0.88	0.82	0.77	0.73	0.77	0.78	0.72	0.13

2.3.4 Modification of SWB model and its Goodness-of-fit testing

According to the modification information provided by confirmatory factor analysis, the goodness-of-fit would be more satisfactory while the experience of goal and personal value adjusted to the first component and the experience of family atmosphere adjusted to the second component. In previous studies, secondary factor analysis showed that experience of family atmosphere and experience of goal and personal value both had heavy factor loading on the two principal components. Because the experience of goal and personal value has an important impact on a person's development experience, it is logically classified as an element of the second component. Actually, in a sense, the state of a person's experience of goal and personal value may be more affected to his basic living conditions, it is likely to affect his mental health. If a person lost his goal and no longer felt his own value, he is likely to face mental health problems. On the contrary, if a person has a realistic goal and aware of his own value clearly, he is more likely to be on the positive emotional adjustment and more easily adapts to the surrounding social environment, and maintains good mental health situation. As to the experience of family atmosphere, we classified it as the first component at first, mainly because this indicator is usually reflected the status of interpersonal adaptation in family environment. We assumed that a person's experience of the situation of his family atmosphere might much more affect his health. But in fact, through Chinese people's cultural value, family happiness has been regarded as an important index to evaluate one's life achivement. Thus, in Chinese cultural backgroud, if a person enjoyed his good experience of family atmosphere, it not only means one of his basic need has been satisfied, but also indicate that his self-potential has been fulfilled. Based on the above analysis, it is more convincing to classify this indicator as the second principal component, that is experience of satisfaction and growth. So the classification of the two indicators should be adjusted, and a modified model of Chinese people's subjective well-being is brought forward here(see Chart 2). Using the data collected from Shenyang and Kunming, the goodness-of-fit of the modified model was initially testified through confirmatory factor analysis. It showed that the main goodness-of-fit of the modified model were favorable (see Table 4).

Chart 2. The modified structural model of Chinese people's subjective well-being

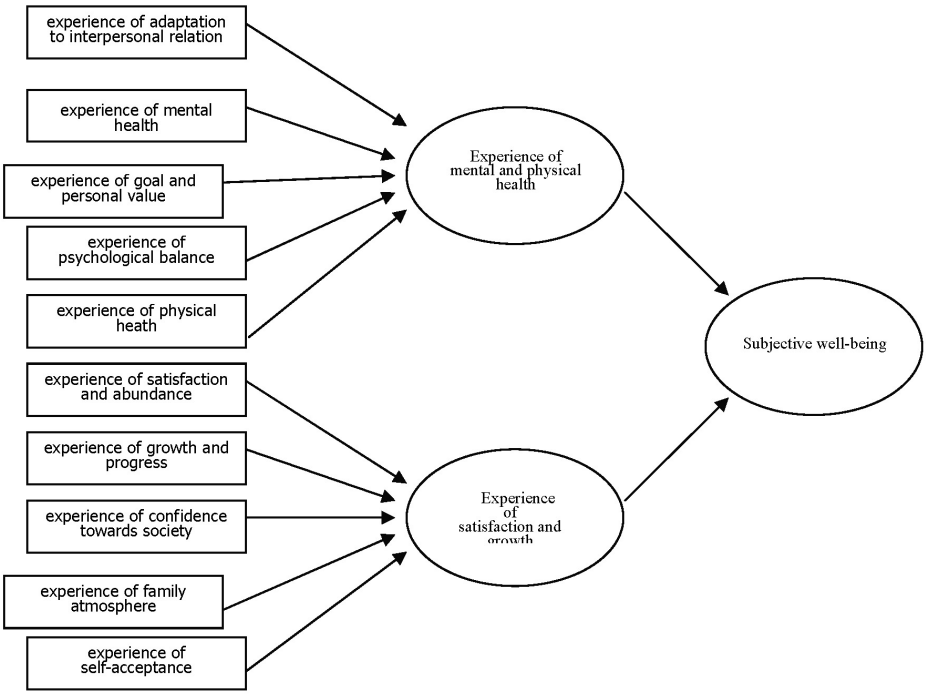


Table 4. The main goodness-of-fit index of the modified model's testifying

x^2/df	GFI	AGFI	NFI	NNFI	CFI	IFI	RFI	RMR
4.31	0.91	0.85	0.82	0.80	0.85	0.85	0.76	0.07

3 Study 2

3.1 Objective

Based on the result of the SWBS-cc applying to Chinese citizens in a broad area, this study was designed to form an ideal revised edition of SWBS-cc to satisfy the needs for national SWB research.

3.2 Method

3.2.1 Participants and instruments

The same as study 2.

3.2.2 The method of item analysis

3.2.2.1 Discrimination analysis

Item discrimination is an important index for item analysis. We examined the discrimination of each item in SWBS-cc by caculating the item-total correlation. The item selection criterions were as following:

1) Item scores and the full scale scores relate significantly;
2) Item scores and the subscale scores relate significantly;
3) The correlation coefficient between item scores and subscale scores is above 0.60;
4) The correlation coefficient between item scores and subscale scores is far lower than the correlation coefficient between item scores and other subscale scores.

3.2.2.2 Factor analysis and logical analysis

Exploratory factor analysis and logic analysis were used to further item analysis to ensure the revised SWBS has ideal structural validity. Items excluded criterions were as follows:

1) The number of extracted component would be decreased distinctly when the item was deleted;
2) The items whose factor loading in the principal component were less than 0.30;
3) Although an item influenced a certain principal component, but it was obviously different from other items influencing the same principal component ;
4) When several items influenced the same principal component, some of them were higher homogeneity or had containment relation;
5) Items originally belonged to a certain dimension, and now vested in another dimension.

3.2.2.3 The principle of optimization

In the process of the item selection according to the criterions listed above, we always insisted the principle of optimization. In order to keep balance on the number of items in each subscale, those sub-optimal items were deleted.

3.3 Result

3.3.1 Discrimination Analysis

In this research we used item-total correlation coefficient as the discrimination analysis index. Table 5 shows the correlation coefficient between item scores, full scale scores and total scores that item belongs to, in which Ai , Bi , Ci ⋯ ⋯ , Ji stand for item number in each subscale, corresponding to the ten subscales of SWBS-cc. In the table, we could find that all the correlation coefficients to be

examined reached statistically significant (two-tailed test). But In subscale experience of satisfaction and abundance, the correlation cofficient between item A9 and the subscale were below 0.60. In subscale experience of mental health, the correlation cofficients between item B2, B7, B9 and the subscale were also below 0.60. In subscale experience of self-acceptance, the correlation cofficients between item F4 and the subscale was below 0.60, and it was lower than its correlation cofficients with subscale experience of growth and progress. The items mentioned here need to be deleted according to the criterions, but the discrimination index number of B2 was 0.598, very close to 0.60, and in former studies it showed good psychometric properties, so it was reserved temporarily. After discrimination analysis, 50 items were reserved.

Table 5. Correlation coefficient between item scores and subscale scores

	Full-scale	Sub-scale1	Sub-scale2	Sub-scale3	Sub-scale4	Sub-scale5	Sub-scale6	Sub-scale7	Sub-scale8	Sub-scale9	Sub-scale10
A1	0.433	0.720									
A2	0.443	0.617									
A3	0.503	0.706									
A4	0.519	0.788									
A5	0.392	0.651									
A6	0.463	0.770									
A7	0.381	0.699									
A8	0.468	0.610									
A9	**0.376**	**0.509**									
B1	0.374		0.604								
B2	**0.388**		**0.598**								
B3	0.420		0.641								
B4	0.544		0.674								
B5	0.501		0.676								
B6	0.446		0.626								
B7	**0.398**		**0.538**								
B8	0.466		0.619								
B9	**0.416**		**0.577**								
C1	0.325			0.702							
C2	0.429			0.800							
C3	0.457			0.827							
C4	0.499			0.803							
C5	0.418			0.734							
D1	0.427				0.766						
D2	0.398				0.757						
D3	0.425				0.763						
D4	0.465				0.775						
D5	0.450				0.732						
E1	0.390					0.651					
E2	0.509					0.786					
E3	0.495					0.762					
E4	0.529					0.822					
E5	0.528					0.785					
F1	0.395						0.747				
F2	0.365						0.746				
F3	0.353						0.733				
F4	**0.416**				**0.583**		**0.486**				
F5	0.466						0.731				
G1	0.351							0.719			
G2	0.392							0.799			
G3	0.458							0.753			
G4	0.443							0.726			
H1	0.440								0.684		
H2	0.424								0.720		
H3	0.541								0.796		
H4	0.473								0.785		
I1	0.420									0.735	
I2	0.436									0.708	
I3	0.405									0.713	
I4	0.396									0.712	
J1	0.408										0.680
J2	0.423										0.685
J3	0.449										0.729
J4	0.392										0.688

3.3.2 Factor Analysis and logical analysis

By means of facotr analysis and logical analysis, the reserved 50 items of SWBS-cc were further analyzed. Data analysis showed that the KMO value was 0.917, Barlett spherical value of chi-square test was at significant level, indicating that factor analysis was suitable. With the factor analysis (principal component/varimax), 10 factors whose eigenvalue over 1 were extracted, and these factors could account for 56.369% of the total variance (Table 6).

In the former discussion about the method of item analysis, we have proposed five criterions for item selection by means of exploratory factor analysis and logical analysis. Item B6 and B8 were deleted according to criterion 5. Items A5 and C2 were deleted according to criterion 4. Items A4 and D2 were deleted according to criterion 3. Item A3 A8 B3 E1 were deleted to maintain balance on the number of items in each dimention. After the above item analysis, the revised SWBS-cc was made up of 40 items.

4 Study 3

4.1 Objective

Study 3 was designed to make a further examination on the psychometric properties of revised SWBS-cc(40 items). In this study, we selected another sample from six capital cities through random sampling technique. We examined the internal consistency reliability, criterion validity, and the construct validity of the scale.

4.2 Method

4.2.1 Participants

The participants were adult inhabitants from Beijiing, Shenyang, Xian, Guangzhou and Kunming. The valid sample size was 3090, and the demographic and social-economic characteristics could be summarized as following: Gender: male, 50.9%; female, 48.2%; missing,0.9%. Age: 18–34 years old, 49.5%; 35–54 years old, 39.7%; 55 years old and above, 10.7%; missing, 0.1%. Education: primary, 17.3%; secondary, 32.2%; higher, 48.2%; missing, 2.5%.

Table 6. Rotated component matrix of 50 items SWBS-cc

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9	Factor10
A5	**0.776**									
A2	0.770									
A1	0.732									
A7	0.689									
A4	**0.683**									
A6	0.637									
A3	**0.575**									
A8	**0.502**									
C3		0.789								
C2		**0.771**								
C4		0.726								
C5		0.686								
C1		0.641								
E4			0.782							
E2			0.729							
E5			0.713							
E3			0.704							
E1			**0.578**							
I1				0.657						
I3				0.636						
I2				0.592						
I4				0.579						
B8				**0.471**						
B6				**0.458**						
D2					**0.744**					
D3					0.699					
D1					0.681					
D4					0.670					
D5					0.504					
B1						0.692				
B2						0.684				
B5						0.629				
B4						0.616				
B3						**0.590**				
F3							0.716			
F2							0.679			
F1							0.627			
F5							0.605			
G2								0.798		
G3								0.694		
G1								0.688		
G4								0.627		
H4									0.654	
H3									0.617	
H2									0.512	
H1									0.509	
J4										0.717
J2										0.676
J3										0.478
J1										0.443

4.2.2 Instruments

SWBS-cc(revised). The scale consists of 40 items, and the respondent are required to respond to each item with 6-grade selection. See study 2.

PWI. See study 1.

SISRSWBS. See study 1.

4.3 Result

4.3.1 Reliability Analysis

Using the data of the sample from five capital cities, we examined the internal consistency reliability of the full scale (40 items) and its ten subscales. The outcome showed the full scale and its subscales had favorable reliability. The Cronbach's alpha coefficient of the full scale was 0.895, and the Cronbach's alpha coefficients of every subscale were also above 0.650 (see Table 7).

Table 7. The internal consistency reliability of ten subscales(revised)

Sub-scale1	Sub-scale2	Sub-scale3	Sub-scale4	Sub-scale5	Sub-scale6	Sub-scale7	Sub-scale8	Sub-scale9	Sub-scale10
0.769	0.689	0.743	0.738	0.805	0.650	0.689	0.689	0.725	0.668

**Data in the table were Cronbach's alpha coefficients.*

4.3.2 Criterion Validity Analysis

We used the pearson correlation coefficient between the scores on the full revised SWBS-cc (including the scores of its subscales) and the scores on PWI and SISRSWBS as criterion index. The criterion validity of revised SWBS-cc applied to the sample from five capital cities were 0.665 and 0.517 respectively. All of the correlation coefficients were significant at 0.000 level. The revised SWBW-cc and its subscales showed favorable criterion validity (see Table 8).

Table 8. The criterion validity of revised SWBS-cc and its ten subscales

SWBS-cc	Sub-scale1	Sub-scale2	Sub-scale3	Sub-scale4	Sub-scale5	Sub-scale6	Sub-scale7	Sub-scale8	Sub-scale9	Sub-scale10
PWI	0.343	0.342	0.377	0.389	0.434	0.397	0.462	0.463	0.408	0.406
SISRSWBS	0.305	0.173	0.339	0.452	0.256	0.370	0.294	0.288	0.255	0.417

**Data in the table were correlation coefficients.*

4.3.3 Structural Validity Analysis

Using the data collected from five capital cities, the goodness-of-fit of the modified structure model of Chinese people's SWB was further testified through confirmatory factor analysis. It showed that the main goodness-of-fit of the modified model were favorable (see Table 9).

Table 9. The main goodness-of-fit index of the modiied model's testifying

x^2/df	GFI	AGFI	NFI	NNFI	CFI	IFI	RFI	RMR
5.8	0.96	0.93	0.89	0.86	0.90	0.90	0.86	0.057

5 Study 4

5.1 Objective

This study was designed to establish the nationwide norm of Chinese capital cities through applying the revised SWBS-cc to a standard normal sample group.

5.2 Method

5.2.1 Sampling method

Multistage stratified sampling method was used to obtain the normal sample group. In the first stage, according to purposive sampling and based on the opinions of experts from related subjects, Beijing, Shenyang, Shanghai(which was replaced by Hangzhou later), Guangzhou, Xian and Kunming were selected. From the geographical point of view, the six cities are located in six geographic regions of Chinese Mainland, that is North China, North East, East China, South Central, North West, South West. The six geographic regions are in accordance with the six large administrative regions established early days of P.R.China, which have great impact on regional economic and social development for a certain period, and the impact has been even to this day. From the regional differences of economic and social development, three of six cities belong to the eastern area, and two cities belong to the central and western area. Although Shenyang is located in Northeast, its current degree of economic and social development is closer to the central city. From the historical and cultural point of view, the selected cities also have strong representation. Random sampling method was used to determine the sampling points in the second stage. In the third stage, according to the Fifth Census data, about 620 citizens whose age were above 18 years old before the survey were selected randomly from the sample points in each selected city.

5.2.2 Instruments

The living conditions of Chinese citizens questionnaire The questionnaire was made up of two parts. The first part was basic individual condition, such as birth, gender, education, and so on. The second part was the revise SWBS-cc and SISRSWBS. Here subjective well-being was replaced by living condition to reduce the face validity of the survey.

5.3 Result

5.3.1 Description of the nationwide normal sample group

The valid size of nationwide normal sample was 3710, and the demographic and social-economic characteristics could be summarized as following: Gender: male, 50.9%; female, 48.3%; missing,0.8%. Age: 18–34 years old, 49.5%; 35–54 years old, 39.7%; 55 years old and above, 10.7%; missing, 0.1%. Education: primary, 17.6%; secondary, 30.4%; higher, 49.6%; missing, 2.4%. Table 10 showed the occupational composition of the nationwide normal sample group.

Table 10. The occupational composition of the normal sample (N = 3710)

Career	Sample size	Percentage of the total sample
Individual labourer	427	11.5
The laid-off, Unemployed and Underemployed	274	7.4
Manufacturing	806 (retired 133)	21.7
Electricity, Gas and Water supplying	63	1.7
Construction	278 (retired 33)	7.5
Traffic, Transport, Storage and Post	234 (retired 32)	6.3
Wholesale and Retail trades	311 (retired 38)	8.4
Financial intermediation	81	2.2
Real Estate	62	1.7
Social Service	301	8.1
Health, Sports and Social welfare	126 (retired 5)	3.4
Education and Culture	260 (retired 53)	7.0
Science institute	171	4.6
Party and Government	196 (retired 35)	5.3
College students	120	3.0

5.3.2 Description of the nationwide norm of revised SWBS-cc

Through the nationwide normal sample, we obtained the norm of revised SWBS-cc in national capital cities. The mean score of the full scale was 61.2 (total score is 100), and the standard deviation is 12.07. Table 11 was the descriptive statistics of its ten subscales. Thus, The amendment of SWBS-cc in the whole country was finished. The revised SWBS-cc contained 40 items, and consisted of ten subscales in accordance with that of SWBS-cc.

Table 11. The descriptive statistics of revised SWBS-cc's subscales

Subscale	Mean scores	Standard deviation
Experience of satisfaction and abundance	4.820	2.405
Experience of mental health	5.566	2.082
Experience of growth and progress	6.154	2.074
Experience of confidence towards society	6.876	1.830
Experience of goal and personal value	6.274	2.350
Experience of self-acceptance	6.613	1.761
Experience of adaptation to interpersonal relation	5.678	2.223
Experience of physical health	5.750	2.099
Experience of psychological balance	6.653	2.453
Experience of family atmosphere	6.812	1.951

6 Conclusion

In the research, we developed the 40 items revised SWBS-cc using the samples from capital cities in Chinese mainland. In the course of scale amendment, discrimination analysis, factor analysis and logical analysis were used to item analysis. The revised SWBS-cc was tested to have good psychometric properties. The nationwide norm of revised SWBS-cc was achieved through the scale's applying to a national normal group. But there are still some limitation in this research. Firstly, the normal group was selected merely from six capital cities, some important factors such as the size of the city were not taken into account, and the number of selected cities was still too few. Secondly, the items used to form revised SWBS-cc were not large enough, which was limited in the SWBS-cc's 54 items, and in some dimensions there were too few items to be selected. Thirdly, As a result of funding constraints, the data collected in some cities were re-used in later studies, and this would have some impact on the result.

REFERENCE

BRADBURN, N.M. & CAPLOVITZ D.(1965), Reports on Happiness. A Pilot Study of Behavior Related to Mental Health. Chicago: Aldine Publishing Company, 9–23.

CUMMINS, R.A. (1996) The domains of life satisfaction: An attempt to order chaos. Social Indicators Research, 38, 303–332.

DIENER ED. EMMONS R.A. LARSEN R.J. et al. (1985), The satisfaction with life scale, Journal of personality assessment, 49 (1), 71–75.

DIENER ED. SUH EM. LUCAS R. et al.(1999).Subjective well–being: three decades of progress, Psychological Bulletin,125(2),276–302.

HUIXIONG, CHEN & LIMIN, WU (2005), Demonstrational Analysis of University Faculty in Zhejiang Province Based on the Investigation of Happiness and Miseries Resources. Journal of Higher Education,(8),14–18.

KOZMA A. & STONES M. J. (1980), The measurement of happiness: development of the memorial university Newfoundland scale of happiness, Journal of Gerontology, 35(6), 906–917.

LUO LU (1998), The meaning,measure,and correlates of happiness among Chinese people. Proceeding of the national science council part C: humanities and social sciences, 8, 115–137.

RYFF, C. D. & SINGER, B. H. (1996). Psychological well-being: Meaning, measurement, and implications for psychotherapy research. Psychotherapy and Psychosomatics, 65, 14–23.

STONES M. J. KOZMA. A. HIRDES J. et al. (1996), Short happiness and affect research protocol. Social Indicators Research,1(37),75–91.

XINHUA, XIAO & YU, DING (2008), An Investigation on the post-graduate's Subjective Well-being in Hunan Province. Academic Degrees & Graduate Education,(1),37–40.

YU, ZHANG & ZHANJUN, XING (2005), An Initial Study on College Student's Subjective Well-being. Youth & Juvenile Study,(1),7–9.

YUANJIANG, MIAO & ZHANJUN, XING (2009),The Application of the Brief Subjective Well-being Scale for Chinese Citizen in Christian Group. China Journal of Healthy Psychology, (6),678–680.

ZHANJUN, XING (2003), Research on the Subjective Well-being Scale for Chinese Citizens Applying to the Senior Citizens, Chinese Journal of Gerontology,(10),648–651.

ZHANJUN, XING (2005), The Measurement of Subjective Well-being. Beijing: People Press, 88–100.

ZHANJUN, XING (2007), An Initial Research on Assessment of Chinese Citizens' Subjective Well-being. Asian Social Science, (1),73–85.

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 317—335

SEARCHING FOR PERIODICITIES IN DATA SERIES

Nicolas Farmakis[1]

ABSTRACT

A new idea for searching and mining latent periodicities from labeled data (time series, aminoacid or DNA sequences, etc.) is proposed. This new method is based on systematic sampling procedures. A materializing algorithm is proposed and some supporting theoretical results are given. This method is very useful for biologists, environmental researchers, financial researchers and managers, etc., etc. Also some illustrating examples are given.

Key words: Systematic Sampling, Periodicity, Data Series, DNA sequence.

1. Introduction

Suppose that we have a series of labeled data of size N:

$$X_1, X_2, X_3,\ldots,X_N. \qquad (1.1)$$

The size of N runs from some decades to some millions or billions (e.g. banking or stock exchange data series or meteorological data series, etc.).

The values in (1.1) are arithmetic ones but they may represent elements of a symbolic sequence, e.g. aminoacid or DNA sequences, Pasquier, et al. (1998), Korotkov and Korotkova (1995).

The values of the series in (1.1) are taken to be random, i.e. we have a random variable X (rv X) with its N values in implementation. The implemented values of labeled data in (1.1) may represent the temperature values at 12h 00m (noon) of every day for 50 years, i.e. N=18250 more or less. Also (1.1) can represent the humidity in a meteorological station of a city or the noise level in a place in the city center. Obviously, there are many cases of labeled data series like the price of gross oil per barrel, in USA dollars, during the last 40 years, etc., etc.

[1] Aristotle University of Thessaloniki, Department of Mathematics, GR-54124 – Thessaloniki GREECE, farmakis@math.auth.gr.

Having to study on such a labeled series as the one in (1.1), we deal (among many other things) with periodicities of the data. The next definition is useful for the purposes of the present paper.

Definition1.1: A series of labeled data, like the one in (1.1), is called *periodic*, if there is an integer $T>1$ for which we have the equality

$$X_{i+T}=X_i,\ \forall\ i=1,2,3,\ldots,N-T \tag{1.2}$$

or more generally

$$X_{i+nT}=X_i,\ \forall\ i, n \text{ such that } i, i+nT\in\{1,2,3,\ldots,N\} \tag{1.3}$$

The integer T is the ***period*** of series (1.1). ∫

Definition (1.1) says that "in a periodic labeled data series we have only T different values for its element, even if the number N is in some cases a very big number, e.g. $N=10$ millions". Moreover, if we write the values of the series (1.1) in a table with T columns, we face the same value as we run on the j^{th} column, $j=1,2,3,\ldots,T$ of this table with the T columns. If the 1^{st} element of the j^{th} column is the X_j, then all the elements in this column are $X_{j+(m-1)\cdot T}=X_j,\ m=1,2,3,\ldots,[\frac{N}{T}]+b$, where b is 1 for the first T-1 (at most) columns and 0 for all the next columns. Also $[x]$ stands for the integer part of the real number x. Thus, the mean value of Xs in the j^{th} column drawn as a systematic sample (see Definition 1.2 below) is $\bar{x}_j = X_j$ and the sample variance is $\sigma_j^2 = 0,$ for all the values $j=1,2,3,\ldots,T$.

Remark 1.1: Due to many sources of errors and due to several kinds of data noise the relations $X_{j+mT}=X_j,\ m=1,2,3,\ldots,[\frac{N}{T}]+b$, $\bar{x}_j = X_j$ and $\sigma_j^2 = 0$, seem to be more or less ideal and it is better to substitute them by the more realistic ones in the every day applications of a researcher: $X_{j+mT}\approx X_j,\ m=1,2,3,\ldots,[\frac{N}{T}]+b$, $\bar{x}_j \approx X_j$ and $\sigma_j^2 \approx 0$.

Finally, if the row length of the table, i.e. the number of columns-samples, is $k \neq T$, then in every column we face at least two different values of data and so the variance grows up quickly. In any case is $\sigma_j^2 > 0$. **o**

®

We need the next definition for the systematic sample:

Definition1.2: Suppose we have a series of labeled data $X_1, X_2, X_3,\ldots,X_N$. We write these data with the hierarchy of labels in a table of k columns and n rows. If k is a divisor of N, then $n=\left[\frac{N}{k}\right]$ exactly. In other cases $n=\left[\frac{N}{k}\right]+1$ and the last row is not full completed. The last row contains $u=N-k\cdot n>0$ elements and its last $k-u$ places are empty. Every column of this table is called "*a systematic sample*" with step k. Obviously $k>1$. ∫

Corollary1.1: The element staying on the *(i,j)* place of the table of definition 1.2 is the $\{(i-1)\cdot k+j\}^{th}$ element of the data series in (1.1).

Proof: Obviously, before the i^{th} row of the table there are $(i-1)\cdot k$ elements belonging to the previous $i-1$ rows. Up to the element situated in the *(i,j)* place of the table we have j more elements. If we represent this element of the table by $Y_{i,j}$, we have $Y_{i,j}=X_{(i-1)\cdot k+j}$, and the proposition is proved. ◊

◊

2. Dealing with systematic samples of labeled data

In what follows we call the table in definition 1.2 a ***systematic table***. It is also more convenient to imagine that k is a divisor of N. All the results remain the same for the other case and we will make some specializations if there is any need. We suppose that we work on a labeled data series like the (1.1) with period $T>1$. The calculations will take place in the ideal (theoretic) field. In the sense of remark 1.1, no statistical error neither data noise is supposed.

In the case we have $k=T$ systematic samples of size n, i.e. $N=n\cdot k= n\cdot T$ is the size of the series in (1.1) and the number of elements of the systematic table. This also means that we have the next relations for every column $j=1,2,3,...,k$ of the table:

$$X_{j+m\cdot T}=X_j,\ m=1,2,3,\ldots,n-1,\quad \overline{x}_j(k)=X_j \quad \text{and} \quad \sigma_j^2(k)=0\,. \tag{2.1}$$

It is quite easy to note that the mean values of the samples are different and thus their variance is not zero, in general. It is given by

$$Var\overline{x}(T)=\sigma_{\overline{x}}^2(T)=E(\overline{x}(T)-\overline{X}(T))^2>0\,,\quad \overline{X}(T)=T^{-1}\cdot\sum_{j=1}^{T}\overline{x}_j(T) \tag{2.2}$$

Obviously, we have the same results with $k=\lambda\cdot T$ in (2.2). We are going to prove:

Theorem 2.1: For k=T we have

$$Var\overline{x}(T)=\sigma_{\overline{x}}^2(T)>Var(\overline{x}'(T))\geq 0 \tag{2.3}$$

$\overline{x}$ is the sample mean value before any permutation of elements between any two columns and $\overline{x}'$ is the sample mean after some permutations of elements between at least two columns of the systematic table.

Consequently

$$Var\overline{x}(T)=\sigma_{\overline{x}}^2(T)>Var(\overline{x}(k))=\sigma_{\overline{x}}^2(k)\geq 0\,,\ k\neq T\,,$$

i.e. finally it is $\max_k Var(\overline{x}(k))=Var(\overline{x}(T))$.

Proof: For $k=T$ and since all the elements of the j^{th} column are all equal we have

$$\bar{x}_j(T) = x_j \text{ , } j=1,2,3,...,T,$$

for every column-sample of the systematic table. This equality is destroyed for every other value of k.

Similarly, the equality is destroyed for every permutation of two at least elements belonging to different columns-samples.

Suppose now that, for any reason, the q^{th} sample exchanges an element with the r^{th} sample of the systematic table. This simply means that the mean values of the new systematic samples will be all the same with the previous except for the two above cases of order q and r. If we define the new sample means as $\bar{x}'_j(T)$, we have of course $\bar{x}'_j(T) = \bar{x}_j(T)$, $j \neq q, r$, and $\bar{x}'_t(T) \neq \bar{x}_t(T),\ t = q, r$. Thus, we arrive to the next two quantities for the variances of the sample means:

$$Var\bar{x}(T) = M + \frac{1}{k}\left\{\left(\bar{x}_q(T) - \bar{X}\right)^2 + \left(\bar{x}_r(T) - \bar{X}\right)^2\right\},\ \ \bar{X} = N^{-1} \cdot \sum_{i=1}^{N} X_i \qquad (2.4)$$

and

$$Var\bar{x}'(T) = M + \frac{1}{k}\left\{\left(\bar{x}'_q(T) - \bar{X}'\right)^2 + \left(\bar{x}'_r(T) - \bar{X}'\right)^2\right\},\ \ \bar{X}' = N^{-1} \cdot \sum_{i=1}^{N} X'_i \,. \qquad (2.5)$$

Obviously, in (2.4) , (2.5) it is $\bar{X}' = \bar{X}$ because the summation is over the same set of values via a different order only. Thus, we need to compare only the two next quantities:

$$k \cdot Q_1 = \left(\bar{x}_q(T) - \bar{X}\right)^2 + \left(\bar{x}_r(T) - \bar{X}\right)^2$$

and

$$k \cdot Q_2 = \left(\bar{x}'_q(T) - \bar{X}\right)^2 + \left(\bar{x}'_r(T) - \bar{X}\right)^2 .$$

If we prove that $Q_1 > Q_2$, then the theorem is also proved.

First of all, we note that

$$\bar{x}'_q = \frac{n-1}{n} \cdot \bar{x}_q + \frac{x_r}{n} = \frac{n-1}{n} \cdot \bar{x}_q + \frac{\bar{x}_r}{n}$$

$$\text{and } \bar{x}'_r = \frac{n-1}{n} \cdot \bar{x}_r + \frac{x_q}{n} = \frac{n-1}{n} \cdot \bar{x}_r + \frac{\bar{x}_q}{n} \qquad (2.6)$$

and thus it is

$$Q_2 = \left(\frac{n-1}{n} \cdot \bar{x}_q + \frac{\bar{x}_r}{n} - \bar{X} \right)^2 + \left(\frac{n-1}{n} \cdot \bar{x}_r + \frac{\bar{x}_q}{n} - \bar{X} \right)^2 = \ldots =$$

$$Q_1 - \frac{2 \cdot (n-1)}{n^2} \cdot \left(\bar{x}_q - \bar{x}_r \right)^2 < Q_1 \text{ q.d.e.}$$

Obviously, for *k=2,3,4,...,T-1* we have $\bar{x}_j(k) \neq x_j$ for some values *j=1,2,3,...k,* due to permutations. After this remark it is obvious that the next relation is valid

$$\max_k Var(\bar{x}(k)) = Var(\bar{x}(T)) . \quad \Diamond \qquad \Diamond$$

Now, a sharper version of the above theorem 2.1 is when we have exchanges of elements among more than two columns-samples of a systematic table with *k* columns and *n* rows. The biggest difference between Q_1 and Q_2 takes place when (evidently) we have exchanges among all the *k* columns. An element from, e.g. the m^{th} column, goes to the r^{th} one and one from the r^{th} goes to another, say to q^{th} one, and so on until all the columns give an element to another and every column receives an element from a column of the systematic table. The next theorem 2.2 is related to the range of the difference between Q_1 and Q_2:

Theorem 2.2: For every k=2,3,4,…,T-1 we have

$$Var\bar{x}(T) = \sigma^2_{\bar{x}}(T) > Var(\bar{x}'(T)) = \sigma^2_{\bar{x}'}(T) \geq 0 \tag{2.7}$$

Consequently

$$Var\bar{x}(T) = \sigma^2_{\bar{x}}(T) > Var(\bar{x}(k)) = \sigma^2_{\bar{x}}(k) \geq 0 \tag{2.8}$$

$$\text{i.e. } \max_k Var(\bar{x}(k)) = Var(\bar{x}(T)) .$$

Proof: For *k=T* and for every column-sample of the systematic table we have the $\bar{x}_j(T) = x_j$, *j=1,2,3,...,T,* because all the elements of the j^{th} column are all equal. For this case with the *k=T* movements we have the equalities:

$Var\bar{x}(T) = \frac{1}{T}\sum_{q=1}^{T}\left(\bar{x}_q - \bar{X}\right)^2$ and $Var\bar{x}'(T) = \frac{1}{T}\sum_{q=1}^{T}\left(\bar{x}'_q - \bar{X}\right)^2$ where it is

$$\bar{x}'_q = \frac{n-1}{n} \cdot \bar{x}_q + \frac{x_r}{n} = \frac{n-1}{n} \cdot \bar{x}_q + \frac{\bar{x}_r}{n}, \text{ with } q \neq r, \quad q, r = 1, 2, 3, \ldots, T \tag{2.9}$$

and the one

$$\bar{x}'_q - \bar{X} = \frac{(n-1) \cdot x_q + x_r - n \cdot \bar{X}}{n}, \text{ with } q \neq r, \quad q, r = 1, 2, 3, \ldots, T \tag{2.10}$$

From (2.9) and (2.10) we obtain

$$Q_2 = T \cdot Var(\overline{x}'(T)) = \sum_{q=1}^{T}\left(\overline{x}' - \overline{X}\right)^2 = \frac{1}{n^2}\cdot\sum_{\substack{q=1\\ q\neq j}}^{T}\left\{(n-1)\cdot x_q + x_j - n\cdot\overline{X}\right\}^2 = \ldots =$$

$$= \frac{1}{n^2}\cdot\left\{\left(n^2-2n+2\right)\cdot\sum_{q=1}^{T}x_q^2 + T\cdot n^2\cdot\overline{X}^2 + 2\cdot(n-1)\cdot\sum_{\substack{q=1\\ j\neq q}}^{T}x_j\cdot x_q - 2\cdot n^2\cdot\overline{X}\cdot\sum_{q=1}^{T}x_q\right\} = \ldots =$$

$$= Q_1 - \frac{n-1}{n^2}\cdot\sum_{\substack{q=1\\ j\neq q}}^{T}\left(x_q - x_j\right)^2 = T\cdot Var(\overline{x}(T)) - \frac{n-1}{n^2}\cdot\sum_{\substack{q=1\\ j\neq q}}^{T}\left(x_q - x_j\right)^2 \text{ and (2.7) is proved.}$$

Note that for a $k \neq T$ some exchanges of elements take place among the columns which now are a little different than the above assumed with $k=T$. The exchanges are more complicated than the described for the relation (2.7) but more chaotic than the above mentioned and they provide mean values with less variation than for those in (2.7). So we expect to arrive to (2.8) with freedom to say that the next equality holds:

$$\max_k Var(\overline{x}(k)) = Var(\overline{x}(T)). \qquad \Diamond \quad \Diamond$$

All that was mentioned in this paragraph could be more clear by some examples like that in the next paragraph.

3. Examples on systematic samples of labeled data

Some examples on systematic samples of size n from a Population **P** of labeled data are written in rows of size k=2,3,..., $\left[\frac{N}{2}\right]$. For reasons of space, and for simplicity reasons too, we use population with relatively small size in the present paper.

Example 3.1: We have some (say 50) values of a random variable X over the time (time series), the x_m, m=1,2,3,…,50. We can of course write them in a table with rows of size k=2,3,4,…,25. We will try first to write them in a table with rows of size:

(a) k=11, in Table 3.1

Table 3.1. In the last row we see the variances of the (every) column-sample data and in the previous one we see the mean values of the respective column-sample data

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th
2	100	40	30	50	1	200	40	40	40	0
180	40	50	40	2	300	40	50	50	2	300
30	30	40	1	250	40	50	40	0	200	40
45	45	3	400	80	40	40	2	320	40	50
40	1	250	40	40	50					
59.4	**43.2**	**76.6**	**102.2**	**84.4**	**86.2**	**82.5**	**33.0**	**102.5**	**70.5**	**97.5**
4822	**1299**	**9718**	**27969**	**9347**	**14637**	**6158**	**449**	**21492**	**7774**	**18692**

and afterwards with rows of size

(b) k=10, in Table 3.2

Table 3.2. In the last row we see the variances of the (every) column-sample data and in the previous one we see the mean values of the respective column-sample data

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
2	100	40	30	50	1	200	40	40	40
0	180	40	50	40	2	300	40	50	50
2	300	30	30	40	1	250	40	50	40
0	200	40	45	45	3	400	80	40	40
2	320	40	50	40	1	250	40	40	50
1.2	**220.0**	**38.0**	**41.0**	**43.0**	**1.6**	**280.0**	**48.0**	**44.0**	**44.0**
1.2	**8200**	**20**	**105**	**20**	**0.8**	**5750**	**320**	**30**	**30**

Its obvious that we have two (among the 24) ways of presentation of the same data x_m, m=1,2,3,…,50. The second way enables us to know a periodicity of the values of data, with a period T=5. Every graph of the five values $x_{5(t-1)+r}$, r=1,2,3,4,5 is giving the same (about) scheme for all t=1,2,3,…,10. The same icon is coming out from the mean values of the columns of data in pre-last row of Table 3.2, i.e. period T=5. Every column of this table is a systematic sample from the population of the 50 data, taken as 1 from 10; see in Cochran (1977), Farmakis (2002), ch. 4th, Thompson (1992) and Thompson (1997). The variance of the sample mean in Table 3.2 is $var\bar{x}_2 = 8007.23$ and the mean value of the sample variances is $E(s^2) = 1448$ with one unit accuracy.

From the other hand the presentation of Table 3.1 encrypts the above mentioned periodicity of the data series. There is not any discretional power even from the row of means in Table 3.1. Thus, the variance of the sample mean in

Table 3.1 is very low i.e. $var\overline{x}_1 = 481.68$ and a mean value of the sample variances $E(s^2) = 11123$, with one unit accuracy.

It becomes more obvious now that a presentation of the above data in a table with rows of size k=5 or k=15 or in general of $k=5t$, t=1,2,3,4,5 gives always the opportunity to "see" the period $T=5$ coming out. In the opposite, a size of rows $k=5t+p \geq 2$, p=1,2,3,4, t=0,1,2,3,4 keeps out of our optical field, the truth of the existence of a period T=5. For instance, in Table 3.3 with row size k=8. As in Table 3.1, we cannot see the period as we are checking the 50 values of data in Table 3.3. We cannot also see this period in the pre-last row of mean values of the columns (systematic samples). It seems to the observer now that the period T=5 is "near" (let us say so) the row size k=10=5·2 in the presentation of Table 3.2 and the same period is "far" from the row size k=11 of Table 3.1, even if the Euclidean distance between 10 and 11 is less than between 10 and 5.

From the pre-last row of Table 3.3, with the mean values of the suitable columns, we get the variance of the sample mean as $var\overline{x}_3 = 193.51$. It is very small in comparison with the $var\overline{x}_2 = 8007.23$ got from the last row of Table 3.2. We also note that the mean value of the sample variances is $E(s^2) = 10657$ (accuracy=1).

Table 3.3. In the last row we see the variances of the (every) column-sample data and in the previous one we see the mean values of the respective column-sample data

1st	2nd	3rd	4th	5th	6th	7th	8th
2	100	40	30	50	1	200	40
40	40	0	180	40	50	40	2
300	40	50	50	2	300	30	30
40	1	250	40	50	40	0	200
40	45	45	3	400	80	40	40
2	320	40	50	40	1	250	40
40	50						
66.29	**85.14**	**70.83**	**58.83**	**97.00**	**78.67**	**93.33**	**58.67**
10942	**11563**	**8024**	**3828**	**22350**	**12674**	**10867**	**5011**

Example 3.2: We have the same values of a random variable X over the time, the x_m, m=1,2,3,…,50, as in the previous example. We try to write them in a table with rows of size k=2,3,4,…,25 (in general it must be k=2,3,4,...,$\left[\frac{N}{2}\right]$, here it is N=50) and we found the results of the next Table 3.4, dealing with the same data:

Table 3.4.

k	2	3	4	5	6	7	8	9	10	11
$Var\bar{x}_k$	27	10	98	**7822**	151	290	194	130	**8007**	482
$E(s^2)$	9520	9718	9715	**1493**	10254	10349	10657	11095	**1448**	11123

Table 3.4. (continued)

k	12	13	14	15	16	17	18	19	20
$Var\bar{x}_k$	645	948	1395	**8386**	1403	2471	2333	2558	**8749**
$E(s^2)$	11032	10709	11700	**1534**	11580	10997	10780	12331	**1101**

Table 3.4. (continued)

k	21	22	23	24	25
$Var\bar{x}_k$	2849	3958	3839	3283	**8401**
$E(s^2)$	12306	10904	10547	12274	**1529**

From the point of view of $Var\bar{x}_k$, the values of row size $k=5t$, t=1,2,3,4,5 seem to be isolated among all the values of row size from 2 to 25. This result becomes sharper if we take as critical index the ratio $\frac{Var\bar{x}_k}{E(s^2)}$. This is obvious because the relatively small values of $Var\bar{x}_k$ are to be divided by the very big values of $E(s^2)$ and the (much) bigger values of $Var\bar{x}_k$ have to be divided by the relatively little values of $E(s^2)$.

From Table 3.4 we conclude that the values of row size equal to $k=5t$, $t=1,2,3,4,5$ correspond with very big values of $Var\bar{x}_k$. They are big values in comparison with the values of $Var\bar{x}_k$ for $k=5t+s$, $s=1,2,3,4$, so that they seem to be isolated among the 24 values from 2 to 25. This result means that we take that the period is:

$$T=5=min\ \{k:\ \text{with comparably big value of}\ Var\bar{x}_k\ \}.$$

Resuming the results of all the tables in example 3.1 and 3.2 we have some interesting conclusions:

1st) In order to affirm the existence of a periodicity among the data series (expressed by the random variable X) an observer must see the $Var\bar{x}_k$ for all the values of $k=2,3,4,\ldots, \left[\frac{N}{2}\right]$, where k is the step of the Systematic Sampling Procedure (SySP), i.e. the row size of the systematic table. The observer checks

the values of k with extremely big values of $Var\bar{x}_k$. This is because these sizes of $Var\bar{x}_k$ are the results of a (kind of) timbre between the values T of the period and k of the row size of the systematic table, when it is $k=\lambda \cdot T$, λ is an integer like 1, 2, 3,…

2nd) The quality of estimation for the mean value of data is the worst one when we choose in SySP such a sample size n as the period T is a divisor of $k=\left[\frac{N}{n}\right]$. Thus, a profit of the knowledge of the value of period T is to avoid confusion via those suitable values of n. The best would be to choose $k < \frac{T}{2}$ or at least $k<T$.

4. An algorithm mining periodicities

The previous paragraphs and their examples and conclusions lead directly to a four steps algorithm. This algorithm is based on the results until now, especially the remark that there is a timbre between T and k.

With the next algorithm we ask the computer to run for all values of k from 2 to $\left[\frac{N}{2}\right]$, in order to compare the values of variance when we have timbre with the others when we have not any timbre:

Algorithm 3.1:

We imagine all data of a series put in a table with row size k. Every column of the table is a systematic sample of size $n=\left[\frac{N}{\kappa}\right]$ or, for at most $k-1$ samples of size $n+1$. The random variable A represents the data with values a_j, $j=1,...,N$.

Step 1: For every $k=2,3,4,\ldots, \left[\frac{N}{2}\right]$ find the mean values $\bar{a}_j(k)$, $j=1,2,3,\ldots,k$ for all the suitable samples.

Step 2: For every $k=2,3,4,\ldots, \left[\frac{N}{2}\right]$ calculate the variance of the mean value estimator of Step 1, i.e. $Var\bar{a}(k)$.

Step 3: Check the values of k with extremely big values of the corresponding variance $Var\bar{a}(k)$, (write a list, e.g. as the two first rows of table 3.4).

Step 4: Find the value of $k_0=min\ \{k$: with a big suitable value of $Var\bar{a}(k)\}$.

Stop

End

If there is any (integer) period T for the checked data then it is $T= k_0$, in 100% of the cases. Some problem arises when T is not an integer. If (e.g.) it is $T=5.5$, then the algorithm will give the value $T=11$ as a period. In this case we get (probably) for $k=5$ and $k=6$ relatively big values of $Var\bar{a}(k)$. More complicated cases are faced as the quantity $T-[T]$ is different from 0 or 0.5, given that the symbol $[x]$ is the bigger integer which does not exceed the real x.

5. Sharper processes for mining periodicities

We are going to give a more sharp process for mining periodicities. This process will be based on some observations on tables 3.2, 3.4 and on 5.1 and 5.2. This leads us to some constitutive and very useful remarks afterwards. The basic remarks are:

Remark 5.1: Looking on tables 3.2 and 5.1 and more over on 3.4 we can state:

"If $k=\lambda \cdot T$, $\lambda=1,2,3,\ldots$ then we have the maximum of the variance of the mean values of the samples (i.e. max $Var\bar{a}(k)$) and also we have the minimum of the mean value of the sample variances (i.e. min $E\left(s^2(k)\right)$)". ®

Remark 5.2: Also looking on tables 3.2 and 5.1 and more over on 5.2 we can state:

"If $k=\lambda \cdot T$, $\lambda=1,2,3,\ldots$ then we have the maximum of the variance of the mean values of the samples (i.e. max $Var\bar{x}_k$) and also we have the minimum of the mean value of the sample coefficient of variation (i.e. min $E\left(CV(A(k))\right)$)". ®

The above remarks lead to two new indices for mining periodicities. These indices could be from 3.4 the $\frac{Var\bar{x}_k}{E\left(s^2\right)}$ and from 5.2 the $\frac{Var\bar{x}_k}{E(CVy_k)}$. We are going to prove the sharpness of them which is equivalent with the outgrowth of the gap between the relatively very big values of $Var\bar{x}_k$ (when T is an divisor of k) and the other values of $Var\bar{x}_k$ (when T is not an divisor of k) which are relatively very small.

Table 5.1.

1st	2nd	3rd	4th	5th
2	100	40	30	50
1	200	40	40	40
0	180	40	50	40
2	300	40	50	50
2	300	30	30	40
1	250	40	50	40
0	200	40	45	45
3	400	80	40	40
2	320	40	50	40
1	250	40	40	50
1.4	**250.0**	**43.0**	**42.5**	**43.5**
0.9	**7200**	**179**	**63**	**23**

In the last row we see the variances of the (every) column-sample data, accuracy=1 and in the previous one we see the mean values of the respective column-sample data with accuracy=0.1.

Table 5.2.

k	2	3	4	5	6	7	8	9	10	11
$Var\bar{x}_k$	27	10	98	**7822**	151	290	194	130	**8007**	482
$E(CVx_k)$	1.27	1.29	1.26	**0.33**	1.29	1.32	1.31	1.31	**0.35**	1.20

Table 5.2. (continued)

k	12	13	14	15	16	17	18	19	20
$Var\bar{x}_k$	645	948	1395	**8386**	1403	2471	2333	2558	**8749**
$E(CVx_k)$	1.14	1.10	1.08	**0.28**	0.93	1.03	1.02	1.04	**0.16**

Table 5.2. (continued)

k	21	22	23	24	25
$Var\bar{x}_k$	2849	3958	3839	3283	**8401**
$E(CVx_k)$	0.81	0.95	0.97	0.79	**0.32**

From the point of view of $Var\bar{x}_k$, the values of row size $k=5t$, $t=1,2,3,4,5$ seem to be isolated among all the values of row size from 2 to 25. This result becomes more sharp if we take as critical index the ratio $\frac{Var\bar{x}_k}{E(CVx_k)}$. This is obvious because the small values of $Var\bar{x}_k$ are divided by the very big values of $E(s^2)$ and the (much) bigger values of $E(CVx_k)$ have to be divided by the relatively small values of $E(CVx_k)$.

Theorem 5.1: If we have periodic labeled data, then the gap between the extremely big values of $Var\bar{x}_k$ and its other common values is smaller or equal to the gap of the corresponding values of the parameter $\frac{Var\bar{x}_k}{E(s^2)}$.

Proof: Obvious, because we divide the maximum of $Var\bar{x}_k$ (nominator) with the minimum of $E(s^2)$ (denominator). ◊

Theorem 5.2: If we have periodic labeled data, then the gap between the extremely big values of $\boldsymbol{Var\bar{x}_k}$ and its other common values is smaller or equal to the gap of the corresponding values of the parameter $\frac{\boldsymbol{Var\bar{x}_k}}{\boldsymbol{E(CVx_k)}}$.

Proof: Obvious, because we divide the maximum of $\boldsymbol{Var\bar{x}_k}$ (nominator) with the minimum of $\boldsymbol{E(CVx_k)}$ (denominator). ◊

6. The trivial case *T=2*

Suppose we have labeled data with period T=2.

This is equal to the supposition that the series of these data has the form a, b, a, b, a, b,….a, b,… where a, b are real numbers or that we have

$$X_{2t-1}=a \text{ and } X_{2t}=b,\ t=1,2,3,\ldots$$

The population of all these values has N elements. N could be odd or even.

All that was mentioned above lead to the idea of a series with the least period T=2, i.e. we deal with the trivial case of the labeled data series.

We are going to study this data series and to prove again that the variance of the sample mean values takes its maximum value when T is a divisor of k.

We have obviously the two cases for the population size:

$$N=2\cdot\lambda+1 \text{ and } N=2\cdot\lambda,\ \lambda=0,1,2,3,\ldots.$$

In order to be more precisely to the environment of the present problem we adopt the next symbolism for the population size N:

$N=T\cdot k\cdot\lambda+2\cdot\theta+1$ and $N=T\cdot k\cdot\lambda+2\cdot\theta$, θ=0,1,2,3,..., λ=1,2,3,… and k=2,3,4,…

Thus, the minimum value for N seems to be the $2\cdot T$ (here it is 4) but a well working value is much bigger, e.g. N=50 for even small T.

We are going to prove the next two theorems:

Theorem 6.1: Suppose we have the above mentioned series of labeled data

a, b, a, b, a, b, …

with population size

$N=T\cdot k\cdot\lambda+2\cdot\theta+1=2\cdot k\cdot\lambda+2\cdot\theta+1$, $\theta=0,1,2,\ldots,k-1$ and $\lambda=1,2,3,\ldots$ Thus, we can note that

$$\lambda=\frac{N-2\cdot\theta-1}{2\cdot\kappa},\ 2\cdot\lambda+1=\frac{N+\kappa-2\cdot\theta-1}{\kappa}$$

$$\text{and } \lambda+1=\frac{N+2\cdot k-2\cdot\theta-1}{2\cdot k}.$$

Then

(a) for $k=2\cdot\mu$, $\mu=1,2,3,\ldots$ we have $Var\bar{x}_k = \frac{(b-a)^2}{4}$ and

(b) *for* $k=2\cdot\mu+1$, $\mu=1,2,3,\ldots$ we have $Var\bar{x}_k < \frac{(b-a)^2}{4}$, sometimes

$Var\bar{x}_k < \frac{(b-a)^2}{16}$.

Proof: (a) Since we have only two values for X, the a and b, in the case with $k=2\cdot\mu$, $\mu=1,2,3,\ldots$ the systematic sampling table has in all its odd columns the value a and in all its even columns the value b. Thus, every odd column-sample gives a sample mean value equal to a and every even column gives a sample mean value equal to b. Since the number of columns is even, we have μ times sample mean the $\bar{x}_k = a$ and μ times sample mean the $\bar{x}_k = b$. Thus, it is $E(\bar{x}_k) = \frac{a+b}{2}$ and $Var\bar{x}_k = \frac{(b-a)^2}{4}$, q.d.e.

(b) Now, it is $k=2\cdot\mu+1$, $\mu=1,2,3,\ldots$

(b1) For $1+2\cdot\theta<k$ in the systematic sampling table we have λ pairs of rows, like

$$a\ b\ a\ b\ldots a$$

$$b\ a\ b\ a\ldots b$$

and a truncated row with $1+2\cdot\theta$ elements like

$$a\ b\ a\ldots a$$

After this, there are

$\theta+1$ columns-samples with mean value $\bar{x} = \frac{(\lambda+1)\cdot a+\lambda\cdot b}{2\cdot\lambda+1}$

θ columns-samples with mean value $\bar{x} = \frac{\lambda\cdot a+(\lambda+1)\cdot b}{2\cdot\lambda+1}$ and

$k-2\cdot\theta-1$ columns-samples with mean value $\bar{x} = \frac{a+b}{2}$.

Thus, the mean value of all the column mean values is given by the next quantity:

$$E(\bar{x}_k) = \frac{a\cdot(N+k-2\cdot\theta)+b\cdot(N+k-2\cdot\theta-2)}{2\cdot(N+k-2\cdot\theta-1)}.$$

The calculation of the variance of the ample mean values gives the

$$\boldsymbol{Var\bar{x}_k} = \frac{\{(2\cdot\theta+1)\cdot k-1\}\cdot(b-a)^2}{4\cdot(N+k-2\cdot\theta-1)^2} < \frac{(b-a)^2}{4}\cdot\frac{k^2}{N^2} < \frac{(b-a)^2}{16} \quad .$$

(b2) For $2\cdot\theta+1=k$ there are $\theta+1$ columns-samples with mean value $\bar{x} = \frac{(\lambda+1)\cdot a+\lambda\cdot b}{2\cdot\lambda+1}$ and θ columns-samples with mean value $\bar{x} = \frac{\lambda\cdot a+(\lambda+1)\cdot b}{2\cdot\lambda+1}$ and the mean value of these means is

$$E(\bar{x}_k) = \frac{(a+b)\cdot(N+k-2\cdot\theta-1)(\theta+1)}{2\cdot k\cdot(N+k-2\cdot\theta-1)}.$$

Thus, the variance of the sample mean values is

$$\boldsymbol{Var\bar{x}_k} = \frac{(b-a)^2\cdot(k+1)\cdot(k-1)}{4\cdot(N+k-2\cdot\theta-1)^2} < \frac{(b-a)^2\cdot k^2}{4\cdot N^2} < \frac{(b-a)^2}{16}.$$

(b3) For $k+1 \le 2\cdot\theta+1 \le 2\cdot k-1$ there are $2\cdot\theta+1-k$ columns-samples with mean value $\bar{x} = \frac{a+b}{2}$, $k-\theta$ with mean value $\bar{x} = \frac{(\lambda+1)\cdot a+\lambda\cdot b}{2\cdot\lambda+1}$ and $k-\theta-1$ with mean value $\bar{x} = \frac{\lambda\cdot a+(\lambda+1)\cdot b}{2\cdot\lambda+1}$.

Thus, the mean value of the sample means is

$$E(\bar{x}_k) = \frac{a\cdot(N-k-2\cdot\theta)+b\cdot(N+k-2\cdot\theta-2)}{2\cdot(N+k-2\cdot\theta-1)}$$

and also the $\boldsymbol{Var\bar{x}_k} = \frac{\{2\cdot k^2-(2\cdot\theta+1)\cdot k-1\}\cdot(b-a)^2\cdot}{4\cdot(N+k-2\cdot\theta-1)^2} < \frac{(b-a)^2}{4}\cdot\frac{k^2}{(N-k)^2} < \frac{(b-a)^2}{4}$.

Thus, the theorem is proved. ◊

Theorem 6.2: Suppose we have the above mentioned series of labeled data

$$a, b, a, b, a, b, \ldots$$

with population size $N=T\cdot k\cdot\lambda+2\cdot\theta=2\cdot k\cdot\lambda+2\cdot\theta$, $\theta=0,1,2,\ldots,2k-2$ and $\lambda=1,2,3,\ldots$

Then (a) for $k=2\cdot\mu$, $\mu=1,2,3,\ldots$ we have $\boldsymbol{Var\bar{x}_k} = \frac{(b-a)^2}{4}$

and (b) for $k=2\cdot\mu+1$, $\mu=1,2,3,\ldots$ we have $0 \le \boldsymbol{Var\bar{x}_k} < \frac{(b-a)^2}{4}$.

Proof: (a) Since we have only two values for X, the a and b, in the case with $k=2\cdot\mu$, $\mu=1,2,3,\ldots$ the systematic sampling table has in all its odd columns the

value *a* and in all its even columns the value *b*. Thus, every odd column (sample) gives as sample mean value the value *a* and every even column gives as sample mean value the *b*. Since the number of columns is even we have μ times sample mean the $\bar{x}_k = a$ and μ times sample mean the $\bar{x}_k = b$. Thus it is $E(\bar{x}_k) = \frac{a+b}{2}$ and (q.d.e.) $Var\bar{x}_k = \frac{(b-a)^2}{4}$, q.d.e.

(b) Now, it is k=2·μ+1, μ=1,2,3,…

(b1) For $\theta=0$ in the systematic sampling table we have λ pairs of rows, like

$$a\ b\ a\ b \ldots a$$

$$b\ a\ b\ a \ldots b$$

and for all the columns-samples the mean value is $\bar{x}_k = \frac{a+b}{2}$. Thus, the $Var\bar{x}_k = 0$.

(b2) For θ=1,2,3,…, $\frac{k-1}{2}$ there are θ columns-samples with mean value $\bar{x} = \frac{(\lambda+1)\cdot a + \lambda \cdot b}{2\cdot\lambda+1}$ and θ columns-samples with mean value $\bar{x} = \frac{\lambda \cdot a + (\lambda+1)\cdot b}{2\cdot\lambda+1}$ and $k-2\cdot\theta$ columns with mean value $\bar{x} = \frac{a+b}{2}$. Thus, the mean value of the sample means is $E(\bar{x}_k) = \frac{a+b}{2}$ and the

$$Var\bar{x}_k = \frac{\theta\cdot(b-a)^2\cdot k}{2\cdot(N+k-2\cdot\theta)^2} < \frac{(b-a)^2}{2}\cdot\frac{k^2}{2\cdot(N+k-2\cdot\theta)^2} < \frac{(b-a)^2}{16}.$$

Note that $N=2\cdot k\cdot\lambda+2\cdot\theta$ gives $\lambda = \frac{N-2\cdot\theta}{2\cdot k}$ and $2\cdot\lambda+1 = \frac{N+k-2\cdot\theta}{k}$. Also the inequality $2\cdot\theta \le k-1$ was adopted.

(b3) For $\theta = \frac{k+1}{2}, \frac{k+3}{2}, \ldots, k-1$, or $k+1 \le 2\cdot\theta \le 2\cdot k-2$, there are $2\cdot\theta-k$ columns-samples with mean value $\bar{x} = \frac{a+b}{2}$, $k-\theta$ with mean value $\bar{x} = \frac{\lambda\cdot a + (\lambda+1)\cdot b}{2\cdot\lambda+1}$ and $k-\theta$ with mean value $\bar{x} = \frac{(\lambda+1)\cdot a + \lambda\cdot b}{2\cdot\lambda+1}$. Thus, the

mean value of the sample means is $E(\bar{x}_k) = \frac{a+b}{2}$ and the

$$Var\bar{x}_k = \frac{(k-\theta)\cdot(b-a)^2\cdot k}{2\cdot(N+k-2\cdot\theta)^2} < \frac{(b-a)^2}{2}\cdot\frac{k\cdot(k-1)}{2\cdot(N+k-2\cdot\theta)^2} < \frac{(b-a)^2}{4}.$$

Note that $N=2\cdot k\cdot\lambda+2\cdot\theta$ gives $\lambda = \frac{N-2\cdot\theta}{2\cdot k}$ and $2\cdot\lambda+1 = \frac{N+k-2\cdot\theta}{k}$.

Thus, the theorem is proved. ◊ ◊

From a special point of view the case T=2 becomes more trivial if we adopt $b=1$ and $a=0$. The general form is $N = 2\cdot k\cdot\lambda + 2\cdot\theta + \upsilon$, where $\upsilon = 0,1$ and $\theta = 0,1,2,...,k-1$. We face two different cases dependent on k, if it is even or odd. The results are given in the next Table 6.1

Table 6.1.

$k \rightarrow$ / $\downarrow \upsilon$	$2\cdot\mu$	$2\cdot\mu+1$
0	$Var\bar{x}_k = \frac{1}{4}$	$0 \le Var\bar{x}_k < \frac{1}{4}$
1	$Var\bar{x}_k = \frac{1}{4}$	$0 \le Var\bar{x}_k < \frac{1}{4}$

The variance in the very trivial case $T=2, a=0, b=1$

7. Results and discussion

Our basic goal was to propose and support by theoretic tools an algorithm dealing with periodicities of labeled data.

The coming out spirit of the present article is that we can take help from systematic sampling with a *step* k=2, 3, 4, 5,…,$\left[\frac{N}{2}\right]$. When $k=\xi\cdot T$, ξ=1,2,3,… and T is the period of the labeled data, we have k samples from the systematic table and the mean values coming out from those samples have the maximum variance. The related variance, when it is $k \neq \xi\cdot T$ is not the maximum and it is usually extremely lower than the maximum one. For instance, T=2 means that data series has the form *a, b, a, b,…,a, b,…* and when it is k=2, 4, 6, 8, …, $2\cdot\xi$ we have very big values for the variance of the sample mean value. On the other hand, the variance in the case $k=2\cdot\lambda+1$ tends to zero and sometimes is exactly zero when we face even values of N.

Sometimes the maximum variance is extremely bigger than the other values. It is even one hundred times or more bigger, see Table 3.4. These extremely big values of the variance are a chance to use the proposed algorithm and its results. The systematic sampling algorithm becomes a tool to copy the motive of data via the timbre of period T and the number of systematic samples k. Moreover, if we want to see more clearly the periodicity and the value of the period T we can use the ratio indices

$$\frac{Var\bar{x}_k}{E\left(s^2\right)} \text{ and } \frac{Var\bar{x}_k}{E(CVy_k)} .$$

For each of these indices the difference between its maximum value and its other common values becomes more and more bigger than the appropriate difference between maximum and common values of the simple $Var\bar{x}_k$.

It is obvious that all the indices presented in this paper are very useful and applicable for many sectors of production and administration. For instance, the administration of a hospital wants to plan the supply of an antivirus vaccine. Their first duty is to see the number of events per months in the past (say) 10 years. If from the 120 values a period T=6 comes out, they have to be careful every six months with the picks in demand for the suitable vaccine.

An implication of the presented indices is their use by the researchers in every case of analysing data. Among many goals, one could be to see if we face any cyclic behaviour of the data separated or in parallel with a linear trend, etc., etc.

Announcement

I thank very much the referees for their suggestions and corrections.

REFERENCES

COCHRAN, W. (1977) "*Sampling Techniques*", John Wiley & Sons, New York, Toronto.

FARMAKIS, N. (2002) "*Introduction to Sampling*", Editing A & P Cristodoulidi, Thessaloniki, (in Greek).

KOROTKOV, E.V., KOROTKOVA, M.A. (1995) "Latent Periodicity of DNA Sequences from Some Human Gene Regions", *DNA Sequence-The Journal of Sequencing and Mapping*, **Vol. 5**, pp 353–358.

PASQUIER, C., PROBONAS, V., VARVAYANNIS, N., HAMODRAKAS, S.(1998) "A Web Server to Locate Periodicities in a Sequence", *Bioinformatics App. Note,* **Vol. 14**, No 8, pp 749–750.

THOMSON, M. E. (1997) "*Theory of Sample Surveys*", Chapman & Hall, London, New York.

THOMSON, S. K. (1992) "*Sampling*", John Wiley & Sons, New York, Toronto.

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 337—342

ONE HUNDREDTH BIRTHDAY OF STANISŁAW ULAM

Czesław Domański[1]

"Mathematics is a very concise way of formulating all rational thoughts"
S. Ulam

This year marks the centenary of birth of Stanisław Marcin Ulam, an extremely talented and creative mathematician, the author of original theories that affected several areas of the discipline and of its spectacular implementations. To commemorate Professor Ulam's 100th birthday and to pay tribute to his great achievements it has been decided by the organizers of the international conference MSA 2009 that a substantial part of the meeting will be devoted to the areas to which he contributed the most significantly – i. e. mathematical statistics, the Monte Carlo method and the use of computers in statistics.

Stanisław M. Ulam was born on April 13, 1909 in Lvov in the family of Polish Jews who had emigrated from Venice three generations before. Jozef Ulam – Stanisław's father, was a barrister. When the World War I broke out the Ulam family moved to Vienne.

After the war, in 1919 Stanisław started his education in the Polish secondary school in Lvov. At that time the theory of relativity by Einstein was being discussed in detail. Ulam resolved to understand the theory and its consequences which, to a large extent, he managed to do. During his school years he got interested and started to examine the problem of odd perfect numbers i.e. numbers which are the sum of all their factors except for the number itself.

In 1927 he enrolled at the Technical University of Lvov where he was attracted by the lectures on the theory of sets given by professor Kazimierz Kuratowski.

Professor Kuratowski also took interest in his new student and the two spent a lot of time together. Ulam solved the first of the problems in the theory of sets assigned to him by the professor, and by doing so he discovered independently the technique of graphs whose existence he had not been aware of. This allowed Kuratowski to confirm his deep belief in Ulam's talent.

[1] University of Lodz, Poland.

It is worth mentioning that Lvov was at the time the place where eminent professors such as Kazimierz Kuratowski, Stanisław Mazur and Stefan Banach lectured at Lvov University and Lvov Technical University. They introduced Ulam to the arcane of the mathematical way of thinking and discovering process. He used to spend hours with them in the Scottish Café poring over a sheet of paper with just one symbol or function on it. They stared at it as if it was a crystal ball, which allowed concentration, and exchanged opinions and suggestions on the problem. One of those sessions spent in the Scottish Café together with Banach and Mazur lasted for about 17 hours with just a few short breaks taken for meals.

Many years later, when he was writing about the Lvov School of Mathematics, Ulam characterized it as dealing with "the heart of the matters" which constitute mathematics. He also recalled his Lvov milieu and the years spent there with feelings of warmth and gratitude.

At the age of 18 Ulam wrote his first paper and published it in Fundamenta Mathematicae in 1929. He presented some of his findings at the congress of mathematicians in Vilnus in 1931. A year later he participated in the congress in Zurich which helped him to build a firm belief in the power of Polish mathematics and his personal ability to find new ways of solving problems.

In 1932 Stanislaw Ulam obtained a master's degree in the set theory, and in 1933 he defended his doctoral dissertation written under the supervision of Kazimierz Kuratowski at the General Faculty of Lvov Technical University and published it in the Polish periodical Fundamenta Mathematicae (Ulam1930, 1933).

His papers, which he published since he was 18, were often written in joint authorship with: K. Borsuk (1931, 1933), K. Kuratowski (1932, 1933), Z. Łomnicki (1934), S. Banach (1947), S. Mazur (1932); (see K. Kuratowski (1973), Mycielski (1990).

In 1935 Ulam went to Princeton in the USA at the invitation of one of the greatest mathematicians of the previous century John von Neumann. Between the years 1936–1940 he was a member of the "Society Fellows" at Harvard University where he became the lecturer in 1940. In the years 1941–1943 he was a professor at the University of Wisconsin in Madison. For over three decades (1944–1967) he worked for Los Alamos National Laboratory in New Mexico, where he arrived at the invitation of his friend John von Neumann and his co-workers. In Los Alamos he became a member of Edward Teller's team working on the "superbomb" project whose head was Robert Oppenheimer.

In the years 1965–1967 he was a professor at the University of Colorado in Boulder.

Having obtained his first problem in Los Alamos in 1944, Ulam worked together with Cornelis Everett and managed to prove that the implementation of the "superbomb" project was unfeasible.

It is worth mentioning here that Ulam and Everett did all their calculations with the use of slide rules and old-fashioned, manually operated office

calculators. A few months later, using one of the first electronic computers, von Neumann confirmed that Teller's project of the "superbomb" was impossible.

Another problem related to the bomb was solved by Ulam in cooperation with Fermi in 1950 and it was done also with the use of slide rules and office calculators. As the two projects proposed by Teller proved to be wrong, Ulam put forward in 1951 a new project of hydrogen bomb which he called "Teller-Ulam device". The above facts reveal that Ulam had some innovative ideas about the bomb.

He also proposed to apply fast computers for solving different problems with the use of random numbers and the Monte Carlo method. Ulam suggested that the probability of obtaining different results should be determined by a computer simulation. The Monte Carlo method allows to program computer in such a way that each step of a particular game is taken according to a well-known probability, and we can determine the accuracy of the final result depending on the number of games in a sample.

The works of Stanisław Ulam cover many areas of mathematics and the related fields e.g. theories of: graphs, numbers, sets, transformations, groups, ergodic and measure, as well as topology, geometric topology, algebra, the Monte Carlo method, theory of combinations, mathematical statistics, branching processes, computers and computing, biomathematics and astronomy (cf. Oktaba, 2000). Ulam was the first one to define the infinite game in the Scottish Book. Everett and Ulam dealt with processes of Galton-Watson type, and their works discuss probability related to cascades of elementary particles induced by impact of very energetic particles or by atom disintegration. Their basic findings concerned the size of population of different particles neurons and uranium nucleus and proportions between population sizes. They used the classic Frobenius-Perron theorem on power of matrix with non-negative terms.

The subject matter of Ulam's and Stein's work is connected with the classic work of Wolterra on periodicity of population sizes of various species of fish in ponds.

While experimenting with the computer E. Fermi, J. Pasta and S. Ulam discovered that a vibrating cord, whose classic equation was disturbed by a non-linear term almost returns to its initial position much earlier than it would follow from statistical considerations.

In the works of Neumann, Richtmyer and Ulam (1947) and Metropolis and Ulam (1949) we can find an application of computers for statistical examination of means with the use of random sampling.

Ulam was interested in the theory of games and genetic distances within species. He perceived the world from the naturalist's point of view and was of the opinion that works of mathematicians enable a closer examination of the world of nature. Physics, biology and mathematical problems were important as long as they were related to achievements of mathematicians.

Throughout his life he published 161 studies, a number of abstracts, research announcements and summaries.

In joint authorship with M. Kac (1968) he wrote a study for Encyclopaedia Britannica entitled "Mathematics and Logic: Retrospects and Prospect". (cf. Wiadomości Matematyczne, 1993).

One of the fathers of the theory of chaos and mathematical physicist–Mitchell Feigenbaum, who was half-Polish and just like Stanislaw Ulam and Marek Kac came from Krzemieniec, was also very much interested in the secrets of nature. The term "mathematical physics" is used here to mean applying and constructing mathematics within the context of physical reality.

While Kac was the pioneer of developing the mathematical theory of probability and its applications mostly in statistical physics, Ulam's name is closely connected with the Monte Carlo method of numerical simulations.

As it was mentioned above, Ulam was an extremely versatile man equally interested in mathematics, astronomy, physics or the theory of relativity. Nowadays, an ever-increasing specialization in the field of mathematics is observed. In the field of physics the problems are defined in a much more precise manner and experiments in physics lead to formulating numerous problems and theories. In the case of mathematics, experiments can be of purely mental or theoretical nature. These days mathematicians have a wonderful, new invention at their disposal–the computer. Since the very beginning of his work in Los Alamos Ulam was connected with computers. At some point he came to the conclusion that the Monte Carlo method was not a great intellectual achievement, yet it was a very useful tool.

It is worth mentioning that the book entitled "*The Scottish Book : Mathematics from the Scottish Café"* came out in print on the basis of the manuscript received from Hugo Steinhaus and translated by Ulam. It encompasses circa 190 problems in their original form written in "The Scottish Book" by Banach, Mazur, Ulam and others, and supplemented with a commentary on advances in solving them. Ulam held those extremely inspiring conversations with Stefan Banach, Kazimierz Kuratowski, Stanisław Mazur and Hugo Steinhaus in cafés of Lvov.

Stanisław Ulam was closely connected with Los Alamos National Laboratory for over 30 years of his life. In the period 1957–1967 he was one of the two scientific advisers of Morris Bradbury–the director of Los Alamos National Laboratory. He became a member of the scientific committee of NASA and US Air Force. He was one of those who proposed starting work on sending the man on the moon to J. Wiesnerow – the scientific adviser to President John F. Kennedy (cf. Kobos A.M., 1999).

Stanisław Marcin Ulam died of heart attack in Santa Fe on May 13, 1984. His wife Francoise Ulam buried his ashes in the Montmartre Cemetery in Paris.

BIBLIOGRAPHY

ULAM S., O teorii miary w ogólnej teorii mnogości, Ossolineum, Lwów 1933.

KOBOS A.M., (1999), Mędrzec większy niż życie, Zwoje (The Scrolls), 3, 16 http://www.zwoje-scrolls.com/zwoje16/text03p.htm.

KURATOWSKI K. (1973), Pół wieku matematyki polskiej (1920–1970), Wiedza Powszechna, Warszawa.

MYCIELSKI J. (1990), Stanisław Ulam (1909–1984), Wiadomości Matematyczne XXIX, s. 21–37.

OKTABA W. (2000), Probabiliści, statystycy, matematycy, ekonometrycy I biometrycy, Lubelskie Towarzystwo Naukowe, Lublin, s. 232–235

ULAM S. Remark on the generalized Bernsteins thcorem, Fundamenta Mathematicae 13 (1929), 281–283.

ULAM S. Concerning functions of sets, Fundamenta Mathematicae 14 (1929), 231–233.

ULAM S. ur Masstheorie in der allgemeinen Mengenlehre, Fundamenta Mathematicae 16 (1930), 140–150

ULAM S., KURATOWSKI K., (1932), Quelques proprietćs topologiąues dii produit combinatoire, Fundamenta Mathematicae 19, 247–251.

ULAM S., MAZUR S., (1932) Sur les transfarmations isom&lriąues d'espaces vectoriels normes, Comptes Rendus de l'Academie des Scienmces de Paris, 194, 946–948.

ULAM S. KURATOWSKI K., (1933), Sur un coefficient lie aux transformations continues d'ensembles, Fundamenta Mathematicae 20, 244–253

ULAM S. , BORSUK K., (1933), Uber gewisse Inuarianten der e-Ahbildungen, Math. Ann. 20, 311–318

ULAM S., Uber gewisse Zerlegungen von Mengen, Fundamenta Mathematicae 20 (1933), 221–223.

ULAM S. , ŁOMNICKI Z., (1934), Sur la theorie de la mesure dans les espaces combinatoires et son application au calcul des probabilites. I. Yariables independantes, Fundamenta Mathematicae 23, 237–278.

ULAM S., OXTOBY J.C., (1941), Measure-presenting homeomorphisms and metrical transitivity, Ann. of Math. (2) 42, 874–920.

ULAM S., METROPOLIS N., (1949), The Monte Carlo method, J. Amer. Statist. Assoc. 44, 335–341.

ULAM, S., KAC M., (1968), Mathematics and logic, retrospects and prospect, Britannica perspectives 1, 557–732; również: Mathematics and Logic, Praeger, N.Y. 1968;

Adventures of a Mathematician, Autobiography, Scribner's, N.Y. 1976, przedrukowane paperback 1977 I 1983

A collection of Mathematical problems, Interscience, N.Y. 1960, także Problems in Modern Mathematics, Science editions, 1964 (przekład rosyjski Niereszonnyje Matiematiczeskije Zadaczi, 1964)

Statistical methods in neutron diffusion, Work done by S. Ulam, J. von Neumann, report written by R.D. Richtmyer, J. von Neumann, LASL, LAMS-551 (1947) (odtajnione 1959), opublikowane w J. von Neumann, Collected Works, vol. V, 751–764, Pergamon Press, 1963, także Ann. Hist. Comput. 7(2) (1985), 148–155

Wiadomości Matematyczne (1993), Refleksje polskich mistrzów – wywiad ze Stanisławem Manickim i Markiem Kacem, przeprowadzony przez Mitchella Feigenbauma, s. 93–114

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 343—350

REPORTS

European Survey Research Association, ESRA 2009, Conference in Warsaw, 29.06–3.07.2009

The third Conference of the European Survey Research Association was held in Warsaw from the 29th of June until the 3rd of July 2009. It was organized in cooperation with the Institute of Philosophy and Sociology of the Polish Academy of Sciences and The Department of Law of the University of Warsaw. Nearly 600 conference participants, representing universities, research institutes, statistical agencies and opinion poll companies, came from 36 countries to give 369 talks. These talks were classified into 12 main conference topics including:

- Substantive Applications of Survey Methodology
- Sampling and Nonresponse
- Data Archive Teaching
- Design and Quality of Survey Questions
- Adding Data to Surveys
- Mode of Data Collection and Data Enhancement
- Social Indicators
- Data Analysis
- Qualitative Methods
- Comparative Research
- Fieldwork Monitoring
- Special Issues related to Survey Research

and were then further divided into specific subjects corresponding to 88 conference sessions. The huge range of themes and talks exceeds the scope of this report. Consequently, further relation would be limited out of necessity to *"Sampling and Nonresponse*" topics, which are the closest ones to the *Statistics in Transition* focus. This topic was divided into five specific subjects.

The first subject named "*New Challenges in Sampling*" incorporated sixteen talks.

- **Seppo Laaksonen** gave an introductory talk on the subject with the overview of challenges and nuisances in contemporary and future survey research from the sampling point of view, in relation to data collection modes, frame construction issues and nonresponse.
- **Jan Kordos** presented in a synthetic way the contribution of J. Neyman to the theory and practice of sampling, including his early works in the

1920's as well as his crucial paper of 1934. He stressed that in 1924, Neyman obtained his doctor's degree from the University of Warsaw, i.e. 85 years ago. Personal reflections of this cooperation were also mentioned. Next the impact of W. E. Deming's approach to the survey design on sample survey practice in Poland was discussed, with special emphasis on survey error decomposition.

- **Siegfried Gabler** and **Horst Stenger** discussed the role of systematic sampling in the context of two-stage sampling designs. They proposed to randomize the order of primary units to be selected systematically with subsequent drawing of secondary units via simple random sampling without replacement. The computation of design effect for the proposed sampling procedure was considered.
- **Matthias Ganninger** and **Siegfried Gabler** considered the estimation of the design effect in a cluster sampling scheme. They compared merits and drawbacks of design-based and model-based estimators for the design effect to finally provide recommendations that can guide the researcher's choice of an estimator.
- Another talk devoted to cluster sampling was given by **Artur Pokropek**, who considered the use of this sampling design in the context of large-scale educational research with application of multilevel modelling. Practical applications of this methodology were presented.
- **Janusz Wywiał** considered the estimation of a population mean using the sampling design dependent on the observed auxiliary variable. He proposed a sampling design that provides the selection probability of any particular sample proportionate to the order statistic of an auxiliary variable. The inclusion probabilities of the first and second order were evaluated. Then three quantile-based sampling strategies were compared with classical ones via computer simulation.
- **Paola Madalena Chiodini**, **Donata Marasini** and **Piero Quatto** presented a talk devoted to the use of so-called control sample known from epidemiological studies in the field of finite population sampling. They discussed a sampling plan to identify "effect" and "control" samples on which the computation of appropriate association measures may then be based. Finally, the choice of a proper association was discussed.
- The next talk by **Zerrin Aysan** and **Öztas Ayhan** was aimed at the study, definition and comparison of alternative sampling frames as potential bases for sample selection in internet surveys. A methodology for domain adjustment procedures eliminating the sample selection bias through appropriate corrections was debated.
- **Efi Markou, Nicolas Razafindratsima, Bernard the Cledat, Pernelle Issenhuth** and **Raphael Laurent** addressed another important problem of analysing results of the survey which is not based on probabilistic sampling. They discussed a practical example of the ELVIRE survey

conducted in 2006–2008 by the French National Demographic Institute, in order to investigate the languages used by the researchers working in the French public research Institutes and Universities. They discussed strategies implemented to obtain a database of sampling units as well as the execution of the survey itself and its results.

- **Giovanna Nicolini** and **Luciana Dalla Valle** dealt with related problem of self-selected sample resulting from voluntary participation of population units in the survey while the sampling frame is unavailable. They presented the application of Heckman's estimation method using the consumer satisfaction survey as an example.
- **Marc Christine** and **Sebastien Faivre** continued the discussion with their talk concerning the construction of annual sampling frames for national household surveys carried out by INSEE based on the lists of dwellings established by the new rotative Census that has been taking place in France since January 2004 and consists of five rotation groups surveyed in a five year cycle. The use of balancing technique by Deville and Tille to draw the sample was then discussed. A calibration procedure was then proposed to improve the sample so that the extent of underrepresentation and overrepresentation of some population subgroups is reduced.
- **Öztas Ayhan** and **Turgay Ünalan** discussed the methodology for the population projections permitting to update information for the future population representation. They used various mathematical models and cohort component projection models to make population projections. Various techniques of extrapolating existing information involving census data to the future were discussed. The problem of balancing domain estimates with the population estimate was also addressed.
- **Mónica Marti** and **Carmen Ródenas** dealt with the migration estimation based on the European Union Labour Force Survey. The problems associated with estimating migration flows were emphasized and their possible sources highlighted.
- **Andrey Veykher** considered the application of "independent" statistical indices to assess the quality of estimates of population characteristics obtained by sampling the adult population of St Petersburg. He discussed the application of various administrative data sources to measure the discrepancies between survey respondent and nonrespondent groups and to test the of hypotheses stating randomness of nonresponse, shedding light on the validity of survey estimates.
- The talk of **Volker Hüfken** was aimed at determination whether the exclusion of adults without mobile telephone access may bias estimates derived from political behaviour-related telephone surveys. This was verified using the data from the European Social Survey (ESS) 2006/07 with the use of logistic regression to compare the odds of political related behaviour for adults with fixed phone access to those for adults with

mobile telephone access and those without telephone access. The results of investigations were presented.

- The session ended with the presentation of the WEBSURVNET initiative carried out by **Pablo de Pedraza**, **Stephanie Steinmetz** and **Kea Tijdens**. This network aims to bring together survey methodologist, web surveys experts, social scientists and official survey institutes, synergize the knowledge of survey methodologists and web survey experts, develop guidelines for official bodies and statistical institutes, provide tools to take advance of technological changes, and foster the development of new cross-national research proposals.

The second main subject named "*Nonresponse*" incorporated five talks.

- **Elena von der Lippe**, **Patrick Schmich** and **Cornelia Lange** presented a methodological study on the use of advance letters sent to survey participants in order to reduce non-response rates. The report was based on an ongoing Health Interview Survey GEDA. The effect of advance letters was assessed by comparing groups of survey participants that were approached with or without advance letter.
- Another presentation authored by **Stefaan Demarest, Jean Tafforeau, Johan Van der Heyden, Lydia Gisle** and **Sabine Drieskens** was devoted to assessment of dependency of health interview survey response rates and the health status of the participants, which determines the scale of nonresponse error. The investigation was based on Belgian Health Interview Survey carried out in 2001. Their results indicate that only the presence of longstanding health problems positively affects the refusal rate but it does not affect the probability to be able to contact the household.
- **Margaret Ely** and **Richie Poulton** discussed the execution of the Dunedin Cohort Study and the use of the information on contact difficulty to assess the likely impact of attrition on the estimation of the prevalence of risk factors and associations of these risk factors with poor health outcomes. They have shown that efforts in retention can represent greater value for money than increases in sample size.
- **Michael Blohm, Achim Koch** and **Hanna Kaspar** present a study evaluating the impact of special interviewing techniques such as the deployment of a special group of interviewers, the personal briefing of interviewers, better payment of interviewers, and the use of respondent incentives and decreases in the respondent burden on response rates, with respect to the German General Social Survey (ALLBUS). They provided an overview on the comparison of distributions between the regular ALLBUS sample and an experimental intensive fieldwork sample.
- **Frode Berglund, Øyvin Kleven** and **Kristen Ringdal** examined the efficiency of follow-up efforts in the case European Social Survey. They compared estimates on selected variables from the response group and the

non-respondent group in order to elaborate on nonresponse errors in the main survey. On the background of empirical results, the value of follow-ups was discussed.

The third main subject named "*Selection Bias in Panel Research*" included seven talks.

- **Gerty Lensvelt-Mulders** provided introductory talk on the general problem of non-feasibility of probability sampling surveys due to non-availability of valid sampling frames, resulting in execution of most access panels on the basis of volunteer-opt-in samples. An overview of current streams of thinking on this problem was provided.
- **Rene Bekkers** analysed the civic duty weight factor. A procedure designed to correct self-selection of individuals with a higher sense of civic duty into survey respondent pools was presented. The procedure weighs observations from survey respondents based on self-reported behaviours for which population values are known. Weighted and unweighted estimates were then compared.
- **Stephanie Steinmetz, Kea Tijdens** and **Pablo de Pedraza** attempted to explore various statistical weighting procedures for volunteer web surveys and evaluate their effectiveness in adjusting biases arising from non-randomised sample selection. Poststratification and propensity score adjustment were compared. The efficiency of different weights was tested by comparing unweighted and weighted results from four different surveys. The sensitivity of the results to changes in the specification of the propensity score was also addressed.
- **Viviane Le Hay** analyzed the process of attrition of panel data by comparing four French electoral surveys between 1958 and 2007, conducted by phone and face to face over national (presidential and parliamentary) elections in France. She assessed the impact of elections types and survey design as well as the impact of the transformation of the general public in terms of level of education and generational renewal on panel attrition. Solutions for lessening attrition were proposed.
- **Elżbieta Getka-Wilczyńska** suggested to apply Poisson processes and basic methods of the reliability theory as a tool for interpretation, definition and analysis for some stochastic properties of Internet data collection process. The random size of uncontrolled sample is defined as an outcome of a counting process. The process of Internet data collection is considered a life test of the population surveyed. Internet survey events are interpreted as a lifetime, arrival, death of the element of the population. The basic characteristics of reliability of the length of the population lifetime are described, calculated and estimated by using the notions and methods of the reliability theory.
- **Marcel Das** discussed optimization in recruitment strategy for a newly established panel survey involving a probability sample of households

drawn from a population register by Statistics Netherlands and conducted via various data collection methods including the Internet. Experimental results aimed at finding the optimal combination of the contact mode, incentive amount, timing of the incentive, content of the information letter, and timing of the panel participation request were presented. The panel design was evaluated by comparing the composition of the recruited panel to population statistics, traditional face-to-face scientific studies and commercial access panels.

- **Miquelle Marchand** discussed the CentERdata LISS-panel, consisting of approximately 5000 households representative of the Dutch speaking population and combining a probability sample and traditional recruitment procedure with online interviewing. Reasons for not joining the panel were explored.

The fourth subject "*Strategies for Nonresponse Adjustments*" included six talks.

- **Joachim R. Frick, Markus M. Grabka** and **Olaf Groh-Samberg** discussed the problem of partial unit nonresponse. Using data on 24 waves of the German Socio-Economic Panel Study they evaluated four different strategies to deal with this phenomenon, described its incidence and time trends, analysed its selectivity, carried out a three-stage longitudinal imputation of missing income components at the individual level, and provided sensitivity analyses showing the variation in the results for income inequality and poverty using alternative imputation strategies.
- **Wojciech Gamrot** explored the phenomenon of respondent interaction and non-independent data missingness. Some simulation results concerning the properties of parameter estimates computed in such a situation were presented.
- **Oscar Breugelmans** investigated the influence of selective unit nonresponse on the analysis of annoyance levels and sleep disturbance due to aircraft noise exposure in the vicinity of Amsterdam Schiphol Airport. The nonresponse effects were assessed by comparing the results of three surveys carried out since 1996. Nonresponse correction methods were also discussed.
- Efi Markou, Bernard the Cledat, Nicolas Razafindratsima, Raphael Laurent **and** Pernelle Issenhuth **continued with the discussion of strategies for nonresponse bias reduction with special emphasis on follow-ups** and of its measurement.
- **Barbara Kowalczyk** and **Emilia Tomczyk** proposed analysis of selected properties of expectations expressed in business tendency surveys taking into account various nonresponse mechanisms. The empirical part of the presentation was based on business tendency surveys conducted by the Research Institute for Economic Development (RIED) of the Warsaw School of Economics.

- The presentation of **Michael Ruland** and **Britta Matthes** attempted to formulate general statements concerning selectivity by using nonresponse questionnaires regarding the non-attendance of foreigners, immigrants and persons with migration background, on the basis of the ALWA survey. It also attempted to explain which of the strategies used in the ALWA study have influenced the foreigners, immigrants and persons with migration background readiness to participate the most.
- **Bryce Weaver** examined the ways to get rid of normality assumption often made when the multiple imputation is used. He developed a technique of multiple imputation making no a priori assumptions on the form of the distribution, useful for the imputation of data when values are missing at random and designed to reproduce the cumulative distribution of any variable — continuous or dichotomous. A SAS implementation of the procedure was presented. An empirical comparison of the SAS PROC MI and the proposed procedure based on the Swiss Household Panel wave 8 (2007) was also provided.

The fifth subject *"Understanding Nonresponse and Attrition : Research from the UK Survey Design and Measurement Initiative"* included five talks.

- **Ian Plewis, Lisa Calderwood, Rebecca Taylor** and **Sosthenes Ketende** reported results from a randomised experiment that tested the efficacy of two different kinds of intervention aimed at converting refusals into productive cases in ongoing longitudinal studies. The intervention was applied in wave four of the Millennium Cohort Study, the fourth in the renowned series of UK birth cohort studies.
- **John Bynner, Harvey Goldstein** and **Gabriele Beissel-Durrant** reported work conducted in the project directed at reducing attrition in longitudinal surveys. It comprised an investigation of the correlates of nonresponse in longitudinal surveys through statistical modelling of attrition processes using data relating to attrition in established UK longitudinal surveys: British Household Panel Study (BHPS) 1958 and 1970 British birth cohort studies, Family and Children's Survey (FACS). The second stage was to formulate optimum strategy for encouraging continued participation as a basis for improved field strategy to reduce attrition. The third stage was to design and implement an exemplar field experiment to test the new strategy using a survey from a series of repeated cross-sectional ('Omnibus') surveys.
- **Gabriele Beissel-Durrant, Robert M. Groves, Laura Staetsky** and **Fiona Steele** investigated the influence of interviewers on unit-nonresponse across a number of UK government surveys. They focused on the effects of socio-demographic characteristics of the interviewer, interviewer experience, interviewing strategies and interviewer behaviours and attitudes. The interaction between household and interviewer characteristics was examined. Survey specific and survey independent

interviewer effects were investigated. The use of multilevel cross-classified models to analyse the effects of interviewers taking account of household level and area level characteristics was explored. Implications for survey practice were discussed.

- **Peter Lynn** and **Laura Fumagalli** reported on large-scale randomized experiments carried out to study possible methods of reducing panel attrition. The experiments addressed issues associated with ability to locate sample members and willingness of sample members to continue to participate. The experiments were carried out in 2008 between waves 17 and 18 of the British Household Panel Survey, involving a sample of around 12,500 persons. Relative costs of alternative procedures for tracking sample members an their impacts on nonresponse bias were considered.
- **Andrew Leicester** and **Zoë Oldfield**, considered the use of barcode scanners as a data acquisition tool in the expenditure surveys. Using data from market research company TNS on the food and grocery purchases of tens of thousands of households over 6 years, they explored implications of this mode of data collection in terms of response, attrition and representativeness by comparing the expenditure and demographic details of these data to existing data sources more traditionally used in social science research, including the EFS, the Census and the British Household Panel Survey.

Prepared by
Wojciech Gamrot
University of Economics, Katowice, Poland

STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 351—352

REPORT

The 6th Conference on Survey Sampling in Economic and Social Research, 21–22 September 2009, Katowice, Poland

The **6th Conference on Survey Sampling in Economic and Social Research** was held from 21st to 22nd September 2009 at the Faculty of Management of the University of Economics in Katowice (Poland). It was organized by the Department of Statistics at the University of Economics in Katowice in co-operation with the Department of Statistical Methods of Łódź University and Polish Statistical Association.

The Scientific Committee consisted of: Andrzej Barczak, Czesław Bracha, Czesław Domański, Nicholas T. Longford, Zdzisław Hellwig, Jan Kordos (chairman), Walenty Ostasiewicz, Jan Paradysz, Jan Steczkowski, Jacek Wesołowski, Janusz Wywiał.

Conference participants, representing universities, statistical agencies and opinion poll companies, came from 8 countries. At the conference were presented 3 invited lectures and 15 papers. The conference was organized to give an opportunity to present latest developments in survey sampling and related fields and to exchange experience on practical applications of survey sampling.

Topics discussed during the conference included:

- Estimation of population parameters based on complex samples
- Statistical inference based on incomplete data
- Small area estimation
- Sample size and cost optimization in survey sampling
- Sampling designs
- Statistical inference using auxiliary information
- Model-based estimation
- Longitudinal surveys
- Practical applications of survey sampling

The invited lectures were presented by:

a. Nicholas T. Longford (Pompeu Fabra University in Barcelona), *A house price index based on the potential outcomes framework*,
b. Malay Ghosh (University of Florida), *Benchmarked small area estimators*,
c. Partha Lahiri (University of Maryland), *Robust mean squared prediction error estimators of EBLUP of small area total under Fay-Herriot model.*

The list of the authors and titles of contributed papers is given below:

1. Paola Maddalena Chiodini, Rita Lima, Giancarlo Manzi, Bianca maria Martelli, Flavio Verrecchia, *Criticalities in Applying the Neyman's Optimality in Business Surveys: a Comparison of Selected Allocation Methods*.
2. Ewa Dziwok, *Yield curve estimation: a comparison of methods with an example of Polish data*.
3. Ryszard Gawlik, *Practical Problems in the Implementation of the Method of Sample Survey in the Field of Social Statistic Research*.
4. Elżbieta Getka-Wilczyńska, *Stochastic properties of the Internet sample.*
5. Alina Jędrzejczak, Jan Kubacki, *Estimation of Gini Coefficient for Regions from Polish Household Budget Survey Using Small Area Estimation Methods.*
6. Mauno Keto, Erkki Pahkinen, *On sample allocation for effective EBLUP estimation of small area totals*.
7. Jan Kordos, *Comparisons of Some Data Quality Issues in Statistical Publications in Poland in the Last Decade*.
8. Danute Krapavickaite, *Estimation of a total of a study variable having many zero values*.
9. Aleksandras Plikusas, *Calibrated estimators under different distance measures*.
10. Elżbieta Soszyńska, *Modelling the influence of human capital on economic growth-the role of samples, influential observations and outliers.*
11. Vladimira Hovorkova Valentova, *Treating Statistical Data Set with Missing Values*.
12. Jacek Wesołowski, *A Simulation Study of Gibbs Sampler for a Hierarchical Bayesian Model in Small Area Statistics*.
13. Janusz Wywiał, *Simulation analysis of accuracy estimation of population mean on the basis of strategy dependent on sampling design dependent on difference of order statistic of an auxiliary variable*.
14. Agnieszka Zięba, Jan Kordos, *Comparing three methods of standard error estimation for poverty measures*.
15. Tomasz Żądło, *On some pseudo-EBLUP in the case of modeling longitudinal profiles*.

Abstract of the presentations are available at http://web.ae.katowice.pl/metoda. The conference was sponsored by SPSS Poland. The 7th Conference on Survey Sampling in Economic and Social Research will take place in 2011.

Prepared by
Tomasz Żądło
Department of Statistics, University of Economics in Katowice