STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 213—216

FROM THE EDITOR

This issue contains seven articles on sampling and estimation methods, one article on small area estimation, three other articles on application of different statistical methods, and report on celebrating statistics devoted to 90th birthday of Professor Kazimnierz Zając from Cracow University of Economics.

There are following articles devoted to some problems of sampling and estimation methods and application of other statistical methods:

1. ***A Confidence Interval for ARPR - "at- risk- of- poverty rate"***(by Ryszard Zieliński form Polnad). The author starts with the European Commission Eurostat document (Doc. IPSE/65/04/EN page 11), where the "at-risk-of-poverty rate" *ARPR* is defined as the fraction of persons in a given population with the equivalised disposable income smaller than p=60 percent of the population median. In Zieliński (2006) the exact distribution of the natural estimator of ARPR was presented and it was proved that the natural estimator is practically unbiased though the distribution of the estimator heavily depends on the population distribution. In what follows he constructs a distribution-free confidence interval for *ARPR*. The interval appears to be highly conservative but it seems that the disadvantage can not be removed.
2. ***Modern Approach to Optimum Stratification: Review and Perspectives*** (by Marcin Kozak, Med Ram Verma from India and Andrzej Zieliński from Poland) The authors discuss a modern approach to stratification of a finite population. They present a general picture of univariate and multivariate stratification, and addressed issues such as strata geometry; an optimization function and constraints for it; dimensionality of stratification; approximate univariate stratification; the choice of an optimization method to perform stratification; initial parameters to be employed in optimization-based stratification; and other population and stratification attributes such as subdivision of a population into domains, domain-orientated approach, and a take-all stratum.
3. ***Innovative Role of Imputation Method/Analysis in Clinical Trial on Remnant Ablation in Differentiated Thyroid Cancer*** (by Prem Chandra, SN Dwivedi, CS Bal, Ajay Kumar, Arvind Pandey from India). This paper intended to emphasize the role of imputation analysis in the area of remnant ablation of differentiated thyroid cancer patients. In view of the fact that under the present data set the values in relation to some of the important covariates were found to be missing, multiple regression method was used to impute the missing covariate values of quantitative

variables. Whereas, discriminant function analysis was used to impute missing values related to categorical covariates. Comparative results i.e., results obtained before and after imputation analysis have been presented separately in detail. Regression models obtained after imputation analysis i.e., based on complete information revealed gained in validity and precision of the developed model. The present study has undoubtedly indicated the usefulness of imputation method in clinical trial regarding remnant ablation of differentiated thyroid cancer.

4. ***Effect of non-response on current occasion in search of good rotation patterns on successive occasions*** (by G. N. Singh and Kumari Priyanka from India). The present work is an attempt to study the effect of nonresponse at current occasion in search of good rotation patterns on successive occasions. Two difference type estimators have been proposed for estimating the population mean at current occasion in presence of nonresponse in two occasions successive (rotation) sampling. Detailed behavior of proposed estimators has been studied. Proposed estimators have been compared with the estimators for the same situations but in the absence of non-response.
5. ***On the Bias Reduction in Linear Variety of Alternative to Ratio-Cum-Product Estimator*** (by Rajesh Singh, Pankaj Chauhan and Nirmala Sawan from India). This paper proposes a method to reduce the bias appearing in the alternative to ratio-cum-product estimator for estimating finite population mean in sample surveys. The proposed technique is useful in survey sampling, when one auxiliary character has positive and high correlation with the study variable, whereas another auxiliary character has negative and high correlation with it. An empirical study is carried out to show the properties of the proposed estimator.
6. ***Estimating the Proportion of People Bearing a Sensitive Issue with an Option to Item Count Lists and Randomized Response*** (by Sanghamitra Pal from India). Randomized Response Techniques (RRT) pioneered by Warner (1965) and being developed quite rapidly ever since are well-known as devices for tackling sensitive issues. IRT is popular for its simplicity. So far IRT's need two independently drawn samples for data generation. A theory is developed for the estimation of the proportion of people bearing a sensitive characteristic in a given community on selecting samples with unequal probabilities.
7. ***Efficient Estimation Using Deep-Post Stratification Under Two Way R X R Set-Up*** (by Manish Trivedi and D. Shukla from India). This paper presents an estimation strategy for the population mean assuming a *r x r* deeply stratified population using a technique deep – post - stratification. The only known information is proportion of row and column size totals of deep – stratification. An efficient estimator is proposed and its optimum properties are examined. A relative comparison of efficiencies is

incorporated. It is concluded that fifty percent of the sum of rows – size – total properties and same of column – size – total proportions and same of column – size – total proportions provides an easy choice of unknown constant. An empirical study is made over obtained results showing a highly significant gain while using technique of deep – post – stratification. Approximate expression of mean square error (MSE) is Derived for this set-up.

One article in devoted to small area estimation:

8. ***Small Area Estimation for Spatially Correlated Populations — A Comparison of Direct and Indirect Model — Based Methods*** (by Hukum Chandra from the U.K., Nicola Salvati from Italy and Ray Chambers from Australia). In this paper the authors investigate small area estimation (SAE) based on linear models with spatially correlated small area effects where the neighbourhood structure is described by a contiguity matrix. Such models allow efficient use of spatial auxiliary information in SAE. In particular, they use simulation studies to compare the performances of model-based direct estimation (MBDE) and empirical best linear unbiased prediction (EBLUP) under such models. These simulations are based on theoretically generated populations as well as data obtained from two real populations (the ISTAT farm structure survey in Tuscany and the US Environmental Monitoring and Assessment Program survey).

There are following three articles devoted to application of statistical methods in different fields:

9. ***European Surveys in the Scope of Service Statistics*** (by Małgorzata Dytman and Agnieszka Matulska –Bachura from Poland). The subject matter of the article evolves around the harmonized methodology of surveys collecting data on the different aspects of services which have been undertaken within the European Statistical System. Poland actively participates in these works. In years 2003-2006 the Central Statistical Office of Poland took part in two pilot projects initiated by Eurostat: „Demand for services" and „Business services". In the article the following items of methodology are included: the subjective and objective scope of survey, the completion of survey, methodological processing and overall assessment of survey. In each of the mentioned items there are information on aspects of surveys conducted in countries participating in the given project as well as more detailed information related to survey realized by official statistics services
10. ***Changes of Employment Structure in Poland and EU Countries*** (by Anna Malina and Piotr Malina from Poland).The article presents the results of an analysis of Poland's employment structural changes as compared with the EU countries. The conducted research on Poland's similarity to the EU-15 countries and the countries which joined the Union in 2004, is based on specific structure similarity measures. Also,

the extent of Poland's and the EU's structural changes was assessed in 1980-2004. The analysis of the sector-based employment structure in Poland and the EU countries indicates the occurrence of similar trends: a decreasing share of employees in sector I (agriculture), in favour of sector III (service sector). The results of the research indicate that the major changes to Poland's employment structure occurred at the beginning of the economic transformation process, i.e. between 1990 and 1995.

11. ***Meta Analysis: What, Why and How*** (by Chandra Bhushan Tripathi, Prem Chandra, Neeraj Pandey and Nilanjan Roy from India). The paper describes the steps involved for carrying out meta-analysis. Brief overview of statistical methods, rationales for using these methods as well as formulae is discussed. The authors stress that meta-analysis was proposed more than 20 years ago as an innovative technique for pooling the results of a series of clinical studies. Meta-analysis has acquired a substantial following among both statisticians and clinicians. The technique was developed as a way to summarize the results of different research studies of related problems. A meta-analysis may be applied even when the studies are small and there is substantial variation in the specific issues studied, the research methods applied, the source and nature of the study subjects, and other factors that may have an important bearing on the findings. Biologists often consider meta analysis to be a simple way to summarize the existing knowledge and examine the actual strength of risk factor. The benefits or hazards that may not be detected in small studies can be found in meta-analysis that uses data from large scale studies. In this paper, the authors describe various methodological steps i.e., need of meta analysis, details about statistical methods used for the analysis, and its uses and limitations

The issue is concluded with the report on *CELEBRATING STATISTICS* - Conference in Honour of Professor Kazimierz ZAJĄC on the Occasion of his 90th Birthday, Cracow, Poland, 26th October 2006 (prepared by Józef Pociecha).

Jan Kordos,
The Editor

STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 217—222

A CONFIDENCE INTERVAL FOR *ARPR* - *"at- risk- of- poverty rate"*

Ryszard Zieliński[1]

ABSTRACT

In the European Commission Eurostat document Doc. IPSE/65/04/EN page 11, the "at-risk-of-poverty rate" *ARPR* is defined as the fraction of persons in a given population with the equivalised disposable income smaller than p=60 percent of the population median. In Zieliński (2006) the exact distribution of the natural estimator of ARPR was presented and it was proved that the natural estimator is practically unbiased though the distribution of the estimator heavily depends on the population distribution. In what follows we construct a distribution-free confidence interval for *ARPR*. The interval appears to be highly conservative but it seems that the disadvantage can not be removed.

Keywords: binomial distribution, confidence interval, non-observable successes.

1. Introduction

Let X be a positive random variable with an unknown continuous and strictly increasing on $[0,\infty)$ distribution function F with F(0)=0 and F(x)>0 for x>0. Let $X_1, X_2, ..., X_n$ be a random sample from F and let $X_{1:n}, X_{2:n}, \dots, X_{n:n}$ $(X_{1:n} \le X_{2:n} \le \dots \le X_{n:n})$ be order statistics from the sample. Given $q \in (0,1)$ and $\alpha \in (0,1)$, the problem is to construct a confidence interval for the fraction of the population below α·(qth quantile) of the distribution F, i.e. for:

$$\theta := F\left(\alpha F^{-1}(q)\right) = P\left\{X \le \alpha F^{-1}(q)\right\}. \tag{1}$$

In the original problem above $q = 0.5$ (the median) and α=0.6, so that *ARPR* = $F(0.6 \cdot F^{-1}(0.5)$.

[1] Institute of Mathematics Polish Academy of Sciences, 00-956 Warszawa, Poland, POB 21; e-mail: R.Zielinski@impan.gov.pl

A natural solution is to construct the standard confidence interval for the probability of success in the Bernoulli scheme with the number of successes

$$K = \#\{j : X_j \le \alpha \cdot F^{-1}(q)\}. \tag{2}$$

A problem however is that F is unknown so that F is not observable. Below we present a solution which gives us a conservative confidence interval but the surplus in the confidence level seems to be nor removable.

A construction of confidence intervals for the probability of success in Bernoulli scheme

Let S be a binomial random variable with the distribution

$$P\{S \le k\} = \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j}, \qquad k = 0,1,...,n.$$

Making use of the well known equality

$$\sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j} = B(n-k, k+1; 1-p), \qquad k = 0,1,\ldots,n,$$

where $B(\alpha,\beta;t)$ is the distribution function of a random variable distributed according to Beta distribution with parameters (α,β), at the point t, we extend the range of the random variable S to the whole interval $[-1,n]$ setting

$$P\{S \le x\} = B(n-x, x+1; 1-p), \qquad x \in [-1,n]. \tag{3}$$

For a fixed β, define the function $k_L(p)$, $p \in (0,1)$, by the formula

$$P\{S \le k_L(p)\} = \beta$$

and construct its inverse $k_L^{-1}(x)$, $x \in [-1,n]$. To this end take into account the following arguments:

$$\begin{aligned} k_L(p) = x \quad & \textit{iff} \quad P\{S \le x\} = \beta \\ & \textit{iff} \quad B(n-x, x+1; 1-p) = \beta \\ & \textit{iff} \quad B(x+1, n-x; p) = 1-\beta \\ & \textit{iff} \quad p = B^{-1}(x+1, n-x; 1-\beta) \end{aligned}$$

where $B^{-1}(\alpha,\beta;t)$ is the inverse function of $B(\alpha,\beta;t)$, or the quantile function of the beta distribution with parameter (α,β), at the point t.

It follows that

$$k_L^{-1}(x) = B^{-1}(x+1, n-x; 1-\beta).$$

Now

$$\beta = P\{S \le k_L(p)\} = P\{k^{-1}(S) \le p\} = P\{B^{-1}(S+1, n-S; 1-\beta) \le p\}$$

or

$$P\{p \le B^{-1}(S+1, n-S; 1-\beta)\} = 1-\beta.$$

It follows that

$$\left(0, B^{-1}(S+1, n-S; 1-\beta)\right)$$

is the one-sided confidence interval for p at the confidence level $1-\beta$.

Similarly, using the function $k_U(p)$ defined by the formula

$$P\{S \ge k_U(p)\} = \beta$$

and its inverse, we obtain the one-sided confidence interval for p at the confidence level $1-\beta$:

$$\left(B^{-1}(S, n-S+1; \beta), 1\right).$$

Combining the above results,

$$\left(B^{-1}(S, n-S+1; \frac{1-\gamma}{2}), \quad B^{-1}(S+1, n-S; \frac{1+\gamma}{2})\right)$$

gives us the two-sided confidence interval for p at the confidence level γ.

2. The confidence interval for $\theta = ARPR$

The statistic K defined in (2) has the binomial distribution

$$P\{K \le k\} = \sum_{j=0}^{k} \binom{n}{j} \theta^j (1-\theta)^{n-j}, \qquad k = 0,1,\ldots,n,$$

but due to the fact that F is unknown, it is not observable.

For a fixed small positive ε, let $\left(X_{k1:n}, X_{k2:n}\right)$ be a nonparametric confidence interval for $Med(F)$ at the confidence level $1-\varepsilon$:

$$P\{X_{k1:n} \le Med(F) \le X_{k2:n}\} \ge 1-\varepsilon .$$

Let C_ε denote the random event $\{X_{k1:n} \le Med(F) \le X_{k2:n}\}$. Let $K1$ be the number of observations $X_1, X_2, \ldots, X_n$ which are not greater than $\alpha \cdot X_{k1:n}$ and let $K2$ be the number of observations $X_1, X_2, \ldots, X_n$ which are not greater than $\alpha \cdot X_{k2:n}$. On the set C_ε we have $K1 \le K \le K2$. Both $K1$ and $K2$ are observable.

For every fixed t, the function $B^{-1}(x+1, n-x; t)$ is strictly increasing in $x \in [-1, n]$, and for every fixed x it is strictly increasing in t. It follows that on the set C_ε

$$B^{-1}(K1, n-K1+1; \frac{1-\gamma}{2}) \le B^{-1}(K, n-K+1; \frac{1-\gamma}{2}) \le B^{-1}(K+1, n-K, \frac{1+\gamma}{2}) \le B^{-1}(K2+1, n-K2; \frac{1+\gamma}{2}) \cdot$$

It follows that

$$\left(B^{-1}(K1, n-K1+1; \frac{1-\gamma}{2}),\ B^{-1}(K2+1, n-K2; \frac{1+\gamma}{2}) \right)$$

is a confidence interval for θ at a confidence level $\alpha \geq \gamma - \varepsilon$. The surplus in the confidence level $\alpha - \gamma - \varepsilon$ may be reduced if instead of the standard nonparametric confidence interval for the median $(X_{k1:n}, X_{k2:n})$ we take the best exact nonparametric confidence interval $(X_{I:n}, X_{J:n})$ with appropriate random indices I and J, such that $P\{X_{I:n} \leq Med(F) \leq X_{J:n}\} = 1-\varepsilon$ (Zieliński and Zieliński 2005).

3. Numerical examples

The actual confidence level α may be substantially higher than assumed $\gamma - \varepsilon$. For example, if $1-\varepsilon = 0.99$ and $\gamma = 0.9$, then the simulated (10,000 simulation runs) probability of coverage $\hat{\alpha}$ of the *ARPR* are presented in the Table below. The simulated length of the confidence interval is denoted by $\hat{\Delta}$.

Table. Simulated values of probability $\hat{\alpha}$ of covering and the length $\hat{\Delta}$ of the confidence interval

$F(x)$	$F(0.6 \cdot F^{-1}(0.5))$	$n=20$		$n=200$		$n=2000$	
		$\hat{\alpha}$	$\hat{\Delta}$	$\hat{\alpha}$	$\hat{\Delta}$	$\hat{\alpha}$	$\hat{\Delta}$
x^7	0.014	0.9972	0.171	0.9931	0.037	0.9587	0.010
x	0.3	0.9999	0.628	0.9999	0.218	1.0000	0.069
$x^{1/7}$	0.465	1.0000	0.805	1.0000	0.286	1.0000	0.091

There are at least two reasons why the actual confidence level α is higher than assumed $\gamma - \varepsilon$.

The first reason is a consequence of discreteness of the binomial distribution (see also Brown at al, 2001). For example, consider the one-sided confidence interval $(B^{-1}(K, n-K+1; \alpha), 1)$ at the confidence level $1-\alpha$ and let

$$\pi(\theta) = P\{B^{-1}(K, n-K+1; \alpha) < \theta\}$$

be the probability of covering the true probability of success θ. Now

$$P\{B^{-1}(K,n-K+1;\alpha)<\theta\} = P\{\alpha < B(K,n-K+1;\theta)\}$$
$$= P\{\alpha < \sum_{j=K}^{n}\binom{n}{j}\theta^{j}(1-\theta)^{n-j}\}$$
$$= P\{K \le \kappa(\theta,\beta)\}$$

where

$$\kappa(\theta,\alpha) = \max\{k: \quad \alpha < \sum_{j=k}^{n}\binom{n}{j}\theta^{j}(1-\theta)^{n-j}\}.$$

For example, for $=10$ and $\alpha = 0.1$, $\pi(\theta) \ge 0.9$ for all $\theta \in (0,1)$ (as assumed), $\pi(0.42) = 0.929$, $\pi(0.2) = 0.967$, and $\pi(\theta) = 1$ for $\theta \ge 0.8$.

The second reason is that in our problem $K = \#\{j : X_j \le \alpha \cdot F^{-1}(q)\}$ is not observable and what we observe are $K1$ and $K2$ which are known to satisfy $K1 \le K \le K2$ on a set of probability $1-\varepsilon$. To illustrate the consequences let $\pi_\delta(\theta)$ be the probability of covering the true θ if the confidence interval is constructed on the basis of the value $K-\delta$ instead of K itself; then

$$\pi_\delta(\theta) = P\{K-\delta \le \kappa(\theta,\alpha)\}.$$

For example, for $n = 100$, $\alpha = 0.1$, and $\theta = 0.5$ we have $\pi_0(\theta) = \pi(\theta) = 0.9033$, $\pi_1(\theta) = 0.9334$, $\pi_2(\theta) = 0.9557$, $\pi_3(\theta) = 0.9716$, $\pi_4(\theta) = 0.9824$, $\pi_5(\theta) = 0.9895$, and $\pi_\delta(\theta) = 1.0000$ for $\delta \ge 13$.

REFERENCES

BROWN, L.D., CAI, T.T., and DASGUPTA,A. (2001): Interval estimation for Binomial proportion. Statistical Science, Vol. 16, No.2, pp. 101—133

Eurostat Doc. IPSE/65/04/EN. Joint working party with candidate countries Statistics on income, poverty & social exclucion (IPSE) and EU/SILC (Statistics on income and living conditions). 19.4.2004

ZIELIŃSKI, R. and ZIELIŃSKI, W. (2006): Best exact nonparametric confidence intervals for quantiles. Statistics, Vol. 39, No.1, February 2005, pp. 67—71

ZIELIŃSKI, R. (2006): Exact distribution of the natural ARPR estimator in small samples from infinite populations. Statistics in Transition. Journal of the Polish Statistical Association, March 2006, Vol. 7, No.4, pp. 881—888

STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 223—250

MODERN APPROACH TO OPTIMUM STRATIFICATION: REVIEW AND PERSPECTIVES

Marcin Kozak[1], Med Ram Verma[2], Andrzej Zieliński[3]

Abstract

In the paper, we have discussed a modern approach to stratification of a finite population. We have presented a general picture of univariate and multivariate stratification, and addressed issues such as strata geometry; an optimization function and constraints for it; dimensionality of stratification; approximate univariate stratification; the choice of an optimization method to perform stratification; initial parameters to be employed in optimization-based stratification; and other population and stratification attributes such as subdivision of a population into domains, domain-orientated approach, and a take-all stratum.

Keywords: stratified sampling, sample allocation, multivariate surveys, auxiliary variable.

1. Introduction

Stratified sampling is one of the most often used sampling designs in sample surveys, for example conducted by the Central Statistical Office of Poland (Kursa and Lednicki, 2006), for several reasons. The most important one is that stratified sampling with random sampling without replacement (*SI*) within strata, i.e., *STSI* sampling, is simple to realize and provides efficient estimators of population parameters provided that the strata are well-constructed (Särndal et al., 1992, p. 100). When stratifying a population, two situations are possible. In the first one, a population is subdivided into strata prior to a study so a statistician has no

1 Department of Biometry, Warsaw Agricultural University, Nowoursynowska 159, 02-776 Warsaw, Poland, E-mail: m.kozak@omega.sggw.waw.pl
2 Division of Agricultural Economics & Statistics, ICAR Research Complex for N.E.H. Region, Umroi Road, Umiam (Barapani) Meghalaya, India, Pin 793 103, E-mail: mrverma19@yahoo.co.in
3 Department of Biometry, Warsaw Agricultural University, Nowoursynowska 159, 02-776 Warsaw, Poland

influence on this division at the design stage of a survey. Division of a population into administrative strata is the simplest example. In the second situation, a statistician divides a population into strata so as to optimize a survey (e.g., maximize the precision of estimation); the strata are to be constructed based on certain rules. In this paper, we discuss the second situation and take no notice of the first situation, which is much simpler to handle at the design stage of a survey as well as is less interesting from a stratification point of view.

Methodology of constructing strata is called stratification. In stratification, a population under study is divided into groups called strata based on available auxiliary information. Independent samples are drawn from each stratum according to a given sampling design. In this paper, we have assumed that an *SI* sample is to be drawn from each stratum; this is quite an often situation in survey practice—this is also the simplest stratified sample scheme. However, what will be said about this design usually refers also to stratified sampling with other than *SI* sampling designs within strata, but the corresponding formulae would have to be reformulated.

To stratify a population, knowledge of values of an auxiliary variable X (or, in a multivariate case, $\mathbf{X}=(X_1,...,X_k)^T$) for all elements of the population is required. If correlation between the auxiliary (X) and survey (Y) variables is high and a stratification method applied is effective, the stratification would be efficient (i.e., estimation would be precise). Otherwise, stratified sampling should not be used. (In fact, poor-quality auxiliary information should not be used at the design stage of a survey at all, irrespective of a sampling design/scheme to be used.) Stratification based on a qualitative auxiliary variable is easy to apply in practice. Nevertheless, stratification based on a quantitative auxiliary variable is not as easy and natural as stratification based on a qualitative variable: an efficient stratification method should be then used to construct the strata. In this paper, we consider construction of strata based on quantitative variable or variables.

Most surveys deal with multiple parameters for the estimation purpose. Two distinct cases can be pointed here out. In the first case, a multivariate stratification variable is used to optimize estimation of a parameter studied. In the second case, a multivariate stratification variable is used to optimize estimation of parameters of several survey variables. In fact, both these approaches do not differ from a practical point of view because in both of them stratification is to be performed in the same way. Univariate stratification methods are not efficient (and sometimes unlikely to be applied) in multivariate surveys so a multivariate stratification method should be applied.

In this paper, we present a picture of stratification from all the viewpoints mentioned. Namely, we discuss stratification as an optimization problem in which division of a population into strata is searched for so as to optimize some function of survey costs or precision of estimation, subject to given constraints. Furthermore, we discuss stratification under univariate and multivariate approaches; we also consider stratification of a population subdivided into

subpopulations (domains). In addition, we present several practical examples of stratification problems under univariate and multivariate approaches. Given the decision on which stratification method is to be applied, those approaches differ in aspects of population such as, e.g., division of a population into subpopulations (domains) or the type of an optimization function and its constraints to be taken into account.

The paper is organized as follows. In section 2, we provide basic ideas of stratification. In section 3, we discuss methods and approaches to univariate and multivariate stratification: approximate and optimization-based stratification is introduced and a general concept of multivariate stratification is presented. In sections 4, 5, 6, and 7, we present four different examples of univariate and multivariate stratification. Finally, we discuss the modern approach to stratification in section 8.

2. Stratification: basic ideas

2.1. Data modification technique

To start with, we present a data modification technique that is very useful in stratification. Kozak (2004b) proposed this technique for univariate stratification, and we will adapt it to multivariate stratification. Stratification applied for modified data is performed on a smaller dataset, which in turn results in less time consumption in computation.

Consider a population U consisting of N elements. A jth element of the population is represented by a vector $\mathbf{x}_j = \left(x_{j1},...,x_{jI}\right)^T$, j=1,…,N, of values of I auxiliary variables. The modification technique is as follows. Divide the population U into disjoint clusters; a particular cluster is to consist of population elements represented by the same vector $\mathbf{x}_j$. As we have N_c different vectors (distinct population elements), there are N_c clusters. Now change indices, by $\mathbf{x}_k$ denoting the vector corresponding to the kth cluster ($k = 1,...,N_c$). Based on these clusters, create a new frame, which, besides the vectors $\mathbf{x}_k$, includes weights w_k, w_k being the size of the kth cluster. Now we can consider a population U_c that comprises N_c clusters, each cluster being represented by the vector $\left(\mathbf{x}_k^T, w_k\right)^T = \left(\mathbf{x}_{k1},...,x_{kI}, w_k\right)^T$.

Consider a division of the population U_c into L strata U_{ch}, h=1,…,L, such that

$$U_{ch} \subset U_c;\ U_{ch} \cap U_{ch'} = \varnothing,\ h \neq h';\quad \bigcup_{h=1}^{H} U_{ch} = U_c \qquad (2.1)$$

Some basic characteristics of the population U can now be evaluated using the following weighted formulae:

$$N = \sum_{k=1}^{N_c} w_k \ , \ N_h = \sum_{k=1}^{N_{ch}} w_{hk} \quad , \ W_h = N_h / N$$

$$S_{ih}^2 = (N_h - 1)^{-1} \sum_{k=1}^{N_{ch}} w_{hk} \left(x_{ihk} - \overline{X}_{ih}\right)^2 \ , \ \overline{X}_{ih} = N_h^{-1} \sum_{k=1}^{N_{ch}} w_{hk} x_{ihk} \ , \quad (i = 1,...,I)$$

where w_{hk} is the weight of the kth cluster of the hth stratum; N_h is the number of elements of the population U that belong to the hth stratum; N_{ch} is the number of clusters that belong to the hth stratum; S_{ih}^2 and $\overline{X}_{ih}$ are the population variance and mean of the ith stratification variable in the hth stratum, respectively; and x_{ihk} is the X_i value for the kth cluster of the hth stratum.

2.2. Stratification

Stratification aims to subdivide the population U_c into L mutually disjoint groups U_{ch} (which fulfill the requirement given in Eq. (2.1)), called strata, in such a way so that a given objective function $f(\mathbf{a})$ is minimum under the two (possible) sets of constraints:

$$g_p(\mathbf{a}) = G_p, p = 1,...,P \tag{2.2}$$

$$h_q(\mathbf{a}) \le H_q, q = 1,...,Q \tag{2.3}$$

where $\mathbf{a}$ is the set of parameters that unequivocally defines the division (2.1), $g_p(\mathbf{a})$ and $h_q(\mathbf{a})$ are some functions of $\mathbf{a}$, and G_p and H_q are given constants. Thus, we consider a set of P equalities in Eq. (2.2) and a set of Q inequalities in Eq. (2.3) as the constraints.

A form of the objective function $f(\mathbf{a})$ and the constraints (2.2) and (2.3) depends on a stratification problem considered in a survey. Explicit formulae for $f(\mathbf{a})$ and the constraints (2.2) and (2.3) are given in subsequent sections, in which we consider particular stratification problems.

2.3. Special population and stratification attributes

Below we discuss some special attributes of a population and stratification. These attributes have a bearing on the form of an objective function and its constraints, and in general, they give an insight into the manner how the population should be stratified.

Dimensionality of stratification. Depending on a problem, stratification may be univariate or multivariate. Univariate stratification is orientated towards estimation of one population parameter; it is based on a univariate auxiliary variable. Multivariate stratification may be orientated towards estimation either one or several parameters; its multivariate character results from the fact that it is performed based on a multivariate stratification variable.

Definition of strata. As already introduced, **a** is a set of parameters that defines the division of a population into strata. In the case of a univariate population, it can be, for instance, a vector of X-based strata boundaries (which take values on the interval from min(X) to max(X)) or a vector of index-based strata boundaries (which take values on the interval from 1 to N). Strata boundaries cut an N-vector of values of the stratification variable X into L strata. The set **a** can also be represented by an N-vector of stratum indicators, the stratum indicator I_k being an integer on the interval from 1 to L (I_k indicates belonging of a kth cluster to an I_k-th stratum). The indicator-based approach is used in cluster-based stratification.

For instance, X-based strata boundaries should be understood as follows. Consider a sequence $a_{1,1} < ... < a_{L-1,1}$. The first stratum consists of clusters that satisfy the condition $x_k \le a_{1,1}$, $k = 1,...,N_c$. The last (Lth) stratum consists of clusters that satisfy the condition $x_k > a_{L-1,1}$. Each of the remaining strata consists of clusters that satisfy the condition $a_{h+1,1} < x_k \le a_{h,1}$ ($h = 2,...,L-1$).

In the multivariate case, construction of strata is usually not as natural as it is in the univariate case. Several manners of strata construction have been proposed. Mulvey (1983) listed stratification by cutpoints and by cluster analysis (for the latter, see also Heeler and Day (1975) and Mulvey and Crowder (1979)). Another approach to defining strata is via so-called L-rot-180 stratification geometry (Briggs and Duoba, 2000; Lednicki and Wieczorkowski, 2003). Under this peculiar stratification geometry, in the bivariate case, strata, the elements of which are points on a plane, have a form of the capital L rotated through 180 degrees. Such stratification can be generalized to more than two dimensions (Lednicki and Wieczorkowski, 2003). Then, the set **a** of strata boundaries can be presented as the following matrix:

$$\mathbf{a} = \begin{bmatrix} a_{1,1} & \dots & a_{1,I} \\ \dots & \dots & \dots \\ a_{L-1,1} & \dots & a_{L-1,I} \end{bmatrix} \tag{2.4}$$

a_{hi} being the particular stratum boundary that satisfies the condition

$$a_{h,i} < a_{h+1,i}, h = 1,\dots,L-2; i = 1,\dots,I$$

Strata are defined by the rows of the array **a** from Eq. (2.4). The first stratum consists of clusters that satisfy the conditions

$$x_{ki} \le a_{1,i}, k = 1,\dots,N_c;\ i = 1,\dots,I \tag{2.5}$$

The last, Lth stratum consists of clusters that satisfy the conditions

$$x_{ki} > a_{L-1,i},\ k = 1,\dots,N_c;\ i = 1,\dots,I \tag{2.6}$$

Each of the remaining strata is defined as follows. A hth (h=2,…,L–1) stratum consists of clusters that satisfy the following conditions at least for one i (i = 1,...,I):

$$a_{h-1,i} < x_{ki} \le a_{h,i} \text{ and } x_{ki'} \le a_{h,i'} \quad \text{for } i' = 1,\dots,I, i' \ne i, \tag{2.7}$$

$$k = 1,\dots,N_c;\ h = 2,\dots,L-1$$

An example of stratification via L-rot-180 geometry under the bivariate case with three strata constructed is presented in Figure 1.

In the case of stratification by cutpoints, strata are defined based on the same array **a** (2.4), but the definition of strata is done in a different manner, which is presented in Figure 2. In this approach, each of L strata defined by the L-180-rot geometry based on the array **a** is subdivided into a given number of sub-strata, viz., an hth stratum is subdivided into ($2h$ – 1) sub-strata; as a result, L^2 strata are obtained. (Note that the first stratum remains the same.)

While describing a particular multivariate stratification problem, we will not specify a type of strata definition to be applied. A general problem is to search for the optimum values of elements of a set **a**, irrespective of how the strata are defined: we let a statistician choose. A choice of strata geometry is also discussed in the last section of the paper.

Take-all-stratum approach Often, distribution of survey and stratification variables is positively skewed. In such a situation, population elements with the large value of a survey variable may strongly affect precision of estimation. Therefore, an efficient procedure to minimize the variation in the estimator of a parameter studied is to create a so-called "take-all" stratum; from such a stratum, all the elements are taken to a sample (Hidiroglou, 1986; Lavalée and Hidiloglou, 1988).

Other population attributes. A population may also be peculiarly divided or specified through some other attributes. For instance, it may be subdivided into subpopulations (domains). Stratification might be done independently within the domains then; a problem (optimization function and its constraints) would have to take into account this subdivision. (Note that in this paper we assume that the domain sizes are known prior to a study and that we do not consider small area estimation problem in which domains may be very small.) Other special features, as, e.g., stratification of a multi-stage population, may also be considered.

If multivariate stratification is to be performed, optimization criterion sometimes does not follow directly from the problem studied (Holmberg, 2002). For instance, this is the case when one aims to optimize estimation of a parameter studied (e.g., the population total) of several population characters with respect to a given sample size. Then, one has to choose a multivariate optimization criterion to be used in stratification; different criteria may provide different stratification.

Optimization problem. Stratification is performed so as to optimize some survey quantity/quantities. The two most common approaches are those in which (i) a total survey cost is minimized so as to achieve a desired level of precision of parameter(s) to be estimated, and (ii) a level of precision of estimation is optimized under a total survey cost assumed. In addition, a stratification problem can be defined in various ways, depending on a population and population/stratification attributes above-mentioned. For instance, consider a population subdivided into several domains; one may aim to minimize a total survey cost so as to achieve the same optimal level of precision of estimation in the domains.

Sample allocation. In stratification, sample—that of assumed size as well as that to be minimized—has to be allocated between strata. In univariate stratification, this usually is not an issue, but in multivariate stratification, it may be because sample allocation that is optimum for one characteristic may not be optimum for other characteristic(s). In such a situation, a compromise criterion is needed to find the allocation that is optimum from the multivariate point of view; such allocation is called the "compromise allocation." The compromise allocation is nearly never optimum for each variable from the univariate point of view. When the variables are highly correlated, the individual optimum allocations for them may differ relatively little; then, the sample allocation optimum from the multivariate point of view may be near-optimum for the characters from the univariate point of view. Many various multivariate allocation approaches have been proposed, including, for example, those by Greń (1964, 1966), Hartley (1965), Kokan and Khan (1967), Wywiał (1988), Bethel (1989), Kozak (2006a), and the like. Sample allocation applied to stratification has to be adapted to the function that is optimized in stratification, and it may be made in an approximate (e.g., Kozak, 2006b) or direct manner; the latter may require solving a numerical problem to find the optimum allocation (Kozak, 2006a).

An important issue in sample allocation is that the optimum allocations should be given in integers, not in real numbers. Analytical formulae do not take this into account so sample sizes from strata have to be rounded to integers, which may cause the results to be not optimum. This problem is omitted when sample allocation is posed as a numerical optimization problem, in which it is not a problem to look for the optimum sample sizes in the space of integer values higher than 1.

3. Stratification methods and approaches

3.1. Approximate stratification

Many approximate stratification methods/approaches have been proposed. Their specificity consists in taking no notice of the form of an optimization function and its constraints. They construct strata based on certain rules, which do not necessarily reflect aims defined by the stratification problem. Therefore, approximate methods do not optimize the stratification problem that is to be optimized—this makes strata constructed in such a manner approximate.

Since approximate stratification methods are commonly known, we will present neither their review nor detailed description; one can acquaint oneself with them in, e.g., Cochran (1977) and Särndal et al. (1992), or in many other text books on survey sampling. From among papers on approximate stratification, let us mention the classical ones: Dalenius (1950, 1952), Mahalanobis (1952), Dalenius and Hodges (1959), Ekman (1959), Sethi (1963), Schneeberger (1970), Singh (1971), or Thomsen (1976). Work on approximate stratification has not yet been finished. For instance, Hedlin (2000) attempted to improve Ekman's procedure. Gunning and Horgan (2004) proposed an alternative approach to approximate stratification based on a geometric progression and the assumption that a stratification variable is uniformly distributed within strata. Their approach aims to equalize values of the coefficient of variation of stratification variable within strata. Horgan (2006) presented a comparison of this approach with Dalenius and Hodges (1959), Ekman (1959), and Lavalée and Hidiroglou (1988) procedures; from her study it follows that stratification based on a geometric progression is the most efficient procedure among the procedures compared. However, Kozak and Verma (2006) obtained a different result in which stratification based on a geometric progression was less efficient than Lavalée and Hidiroglou's algorithm. The authors suggested that the stratification proposed by Gunning and Horgan (2004) might be successfully applied as a source of initial parameters in the optimization approach to stratification.

Kozak (2004c) studied three approximate stratification approaches, those of Mahalanobis (1952), Dalenius and Hodges (1959) and Ekman (1959), for their usefulness as a source of initial values in the optimization approach to optimum

stratification (see below). He detected that Mahalanobis' procedure was the best whereas Dalenius and Hodges' procedure was the worst among the three approaches studied.

Approximate stratification methods have some pros and cons. The most important pro is that they are usually simple to apply. Immediately the con follows: they are not optimum in the sense that the strata constructed are not optimum. The second pro is that they can be exploited to support, as initial strata boundaries, the optimization approach to stratification (Kozak, 2004c). However, the second con is that it may happen that approximate strata do not fulfill constraints of optimization problem. Without any doubt, approximate stratification approaches are useful as initial strata boundaries in stratification, but nowadays they should not be used as a means to provide the final strata (Kozak and Verma, 2006).

3.2. Optimization approach to stratification

As already mentioned, approximate stratification methods take into account neither optimization function nor its constraints. For this reason, it is likely to obtain strata that do not fulfill constraints of stratification; even if they fulfill the constraints, they would unlikely be optimum. Thus, optimization of stratification has been addressed by many authors—see, e.g., Mulvey and Crowder (1979), Mulvey (1983), Godfrey et al. (1984), Lavalleé and Hidiroglou (1988), Sweet and Sigman (1995), Niemiro (1999), Dorfman and Valiant (2000), Rivest (2002), Lednicki and Wieczorkowski (2003), or Kozak (2004b).

We call all these approaches the "optimization approaches to stratification" because they treat stratification as an optimization problem. Some of them do it more formally, by defining a problem (optimization function and its constraints), whereas others may ignore, for instance, some or even all constraints. It has, however, been proved that, in general, the optimization approach is much more efficient than the approximate approach. In fact, some optimization approaches to stratification are approximate because, for instance, they apply clustering procedures, and in such a way do not take into account an optimization function that is considered in a particular optimization problem (see Discussion section).

Kozak and Verma (2006) has recently proved that the optimization approach is more efficient than the approximate stratification, and disputed applying approximate stratification procedures in order to obtain final stratification points. In their opinion, state-of-the-art methodologies, that is, optimization approaches, should be applied to do it. In this paper, we do not refer to any particular optimization approach—we just claim that the optimization approach should be used to stratify a population. It is enough then to provide an optimization function to be minimized and sets of equality and inequality constraints for the function. Therefore, all optimization problems will be formulated as follows: find a set $\mathbf{a}$ (set of parameters defining strata) that minimizes an objective function $f(\mathbf{a})$

subject to specified equality $g_p(\mathbf{a})$, $p = 1,...,P$, and inequality $h_q(\mathbf{a})$, $q = 1,...,Q$, constraints. Which method will be used to perform stratification is the decision of a survey statistician designing a survey.

3.3. General note on multivariate stratification

In this subsection, we present a brief review of various methods of/approaches to multivariate stratification. Ghosh (1963) extended Dalenius's (1952) theory for univariate stratification to more than one variate, and developed the theory of optimum stratification with two characters under the proportional allocation. In a multi-character situation, strata have to be defined based on the joint variations in the characters under study. A problem of optimum stratification with multi-characters is thus a problem of determination of the optimum shapes (geometry) and sizes of strata. In this light, Samanta (1965) considered optimum stratification for more than one variate in the case of the proportional allocation by minimizing the generalized variance of estimators under the assumption that stratification variables are identical to survey variables. Sadasivan and Aggarwal (1978) considered bivariate stratification under the Neyman-optimum sample allocation. They extended the exact equations given by Dalenius (1950) to the bivariate case by taking study variables as the basis for stratification. They presented a set of equations giving the optimum, in the sense of minimizing the generalized variance, strata boundaries in two cases: (i) when ρ, the correlation coefficient between the variables, is constant within strata; and (ii) when ρ varies from stratum to stratum. Working along the same lines as Ekman (1959), they also developed approximate solutions to find stratification points under the assumption that there were numerous strata so the higher powers of stratum width could be omitted.

Gupta and Seth (1979) considered a problem of optimum stratification of a multivariate population based on one auxiliary character X under the proportional allocation. They showed that if the regression of each of Y_i (i=1,2,…p) on an auxiliary variable X was of the same form, stratification points evaluated based on X values yielded estimates that had the smallest concentration ellipsoid among all the possible stratification points based on X. (Note that this type of stratification may be seen as the univariate stratification, since determining of strata boundaries is to be done based on one auxiliary variable.)

Schneeberger and Pollot (1985) considered the problem of optimum stratification of two variates under the proportional and optimum allocation in the case of bivariate normal distribution; they took into account two stratification attributes, viz., a correlation coefficient ρ between the characters and a sampling fraction. Rizvi et al. (2000) developed the theory of optimum bivariate stratification under the proportional allocation. Rizvi et al. (2002) considered the problem of optimum bivariate stratification for two characters under the compromise allocation, and proposed a cumulative cube root rule for

determination of the optimum points of stratification. Verma et al. (2003) investigated the compromise allocation for determination of optimum strata boundaries for two sensitive variables.

Golder and Yeomans (1973), Heeler and Day (1975), and Yeomans and Golder (1975) investigated the use of cluster analysis in stratification. Following their concepts, Mulvey (1983) proposed to pose stratification as an optimization problem under the clustering approach; he achieved more efficient stratification than the former authors had. Nevertheless, note that Mulvey (1983) formulated the optimization function to be minimized as the sum of the multivariate within-strata variances. In addition, the constraints for the function he used did not insure that the constraints (2.2) and (2.3), considered in a particular optimization problem, would be fulfilled. Therefore, one cannot claim that such an approach is optimum in the sense of optimizing the problem posed as minimizing $f(\mathbf{a})$ subject to the constraints (2.2) and (2.3).

Recently, Briggs and Duoba (2000) presented a new approach to stratification of a bivariate population. In their approach, strata are constructed based on the L-rot-180 geometry discussed earlier. Lednicki and Wieczorkowski (2003) studied this approach and generalized the procedure to more than two dimensions. Such an approach may be treated as an optimization approach to multivariate stratification, since it aims to minimize $f(\mathbf{a})$ subject to the constraints (2.2) and (2.3). Later, the L-rot-180 geometry was applied by Kozak (2004a), who presented stratification under the compromise allocation in a problem of stratifying a multivariate population orientated towards minimizing the maximum coefficient of variation of estimators studied.

4. Stratification optimizing precision of a studied estimator with respect to fixed sample size

In this and the following sections, we use the information provided in earlier sections to present particular stratification problems. Each problem is specified in terms of the population and stratification attributes mentioned above, such as dimensionality, subdivision of the population, other population/stratification attributes, and finally, optimization function and constraints. As already mentioned, we will not set an optimization method and strata shape (in multivariate optimization) to be used—we let a statistician decide. Instead, we discuss this issue in order to help the statistician make the decision.

For the sake of simplicity, in the formulae in this and the following sections, we omit the index c regarding the population U_c even though we assume that stratification is to be done on a population being subjected to the modification technique given earlier.

In this section, we consider a classical stratification problem in which one aims to optimize precision (i.e., minimize the variance or coefficient of variation

of an estimator; in univariate studies these both approaches are equivalent) of the estimator of a population parameter for a study variable, with respect to a given sample size. Stratification is to be performed based on a univariate stratification (auxiliary) variable. Suppose that a parameter of interest is the population mean of a variable under study. (Note that optimum stratification for the population mean is also optimum for the population total.) We aim to construct L strata. Given sample size n, the problem can be posed as a nonlinear programming problem (Särndal et al, 1992). Under the Neyman-optimum sample allocation between strata, the objective function to be minimized is usually formulated as

$$f(\mathbf{a}) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h S_h\right)^2 - \sum_{h=1}^{L} N_h S_h^2 \tag{4.1}$$

where **a** is a set of parameters defining the subdivision of the population into strata, N_h is the size of the hth stratum, and S_h^2 is the population variance of the stratification variable restricted to the hth stratum.

In Eq. (4.1), the formula for the variance of the mean estimator under *STSI* sampling and under the Neyman-optimum sample allocation is presented. Särndal et al. (1992) proposed to ignore the component $\sum_{h=1}^{L} N_h S_h^2$ on the right hand side of Eq. (4.1), and to minimize the following simplified objective function:

$$f(\mathbf{a}) = \sum_{h=1}^{L} N_h S_h \tag{4.2}$$

Niemiro (1999) formulated the problem in the same way. Nevertheless, we do not see the need of ignoring the component $\sum_{h=1}^{L} N_h S_h^2$ in the optimization because it does not facilitate calculations in numerical optimization.

The formulation of the objective function made in Eq. (4.1) takes into account the Neyman-optimum sample allocation. A drawback of this approach is ignoring a set of restrictions related to sample sizes from strata, i.e.,

$$2\mathbf{I} \le \mathbf{n} \le \mathbf{N}, \tag{4.3}$$

where $\mathbf{n}=(n_1,\ldots,n_L)^T$, $\mathbf{N}=(N_1,\ldots,N_L)^T$, and **I** is the L-vector of ones; it also operates on real values (we mentioned this problem before). Ignoring this set of $2L$ inequality restrictions can cause sample size from some stratum/strata be (i) lower than 2, or (ii) larger than the corresponding stratum/strata size. These both situations make the allocation unacceptable; they are, however, quite common in survey practice, especially when a population under study is small or variation in a stratification variable is large. If the constraints (4.3) are not fulfilled, the sample allocation has to be changed. Note that the objective function (4.1) is not appropriate then because it assumes that the Neyman-optimum allocation is applied.

Therefore, let us consider an optimization problem in the appropriate form, in which the above-mentioned problems are not encountered. Our aim is to find the set **a** that minimizes the following objective function:

$$f(\mathbf{a}) = \sum_{h=1}^{L} W_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \tag{4.4}$$

where $W_h = N_h / N$. In (4.4),

$$n_h = n W_h S_h \left(\sum_{h=1}^{L} W_h S_h \right)^{-1}$$

is the Neyman-optimum sample size from the hth stratum, n being the assumed sample size . The constraints for the function (4.4) are

$$N_h \geq 2 \text{ ; } h=1,...,L \tag{4.5}$$

$$2 \leq n_h \leq N_h \text{ ; } h=1,...,L \tag{4.6}$$

$$\sum_{h=1}^{L} n_h = n \tag{4.7}$$

If any of the constraints (4.6) is not fulfilled (after rounding n_h to integers), the optimum components of the vector $\mathbf{n} = (n_1,...,n_L)^T$ are given by the solution of the following optimization problem:

Given **a**, minimize $f(\mathbf{n}) = \sum_{h=1}^{L} W_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right)$ (4.8)

subject to the constraints (4.6) and (4.7). (4.9)

In Eq. (4.8), **n** is the n-vector of integers satisfying the conditions given in Eqs (4.5)-(4.7). This comment applies also to all the allocation problems given hereafter.

Thus, it may happen that in one step of a stratification algorithm one will have to include another numerical sub-problem, that from Eqs (4.8)-(4.9).

Take-all-stratum approach

When one wants to construct a take-all stratum, the objective function (4.4) has to be changed to

$$f(\mathbf{a}) = \sum_{h=1}^{L-1} W_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \tag{4.10}$$

The constraints (4.5), (4.6), and (4.7) are to be redefined as follows:

$$N_h \geq 2 \text{ and } N_L \geq 1 \text{ , } h = 1,...,L-1 \tag{4.11}$$

$$2 \le n_h \le N_h,\ h = 1,\ldots,L-1 \tag{4.12}$$

$$N_L + \sum_{h=1}^{L-1} n_h = n \tag{4.13}$$

where N_L is the size of the take-all stratum (the Lth stratum).

Provided that the constraints (4.12) are fulfilled, the elements of the vector **n** are given by

$$n_h = (n - N_L) W_h S_h \left(\sum_{h=1}^{L-1} W_h S_h\right)^{-1},\ h = 1,\ldots,L-1;\ \ n_L = N_L$$

Otherwise, the following problem has to be solved numerically:

Given **a**, minimize $f(n_1,\ldots,n_{L-1}) = \sum_{h=1}^{L-1} W_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right)$ (4.14)

subject to the constraints (4.12) and (4.13). (4.15)

The restriction $N_L \ge 1$ from (4.11) is not required from a theoretical point of view, but it guarantees that the take-all stratum will not be empty. (Sometimes one may require that $N_L \ge A$, A being some integer greater than 1; a statistician is to decide about this value.)

In both optimization problems discussed in this section, one equality constraint and $3L$ inequality constraints are taken into account. Hence, a complexity of the problem increases with increasing L. Furthermore, in a particular step of the algorithm it may happen that one or more of the constraints related to sample sizes from the strata (i.e., (4.6) and/or (4.12)) are not fulfilled. Then, the optimum sample sizes would have to be obtained using a numerical method by solving a particular allocation problem provided earlier in this section; it would certainly result in more time consumption of execution of the algorithm.

In next sections, we provide formulae only for the case in which a take-all stratum is constructed, since the formulae for the without-take-all-stratum case are simpler and may be explicitly given based on the take-all-stratum formulae.

5. Stratification in domains minimizing overall sample size

Based on the paper by Rivest (2002), Lednicki and Wieczorkowski (2003) introduced a problem in which domains (subpopulations) of a population under study are stratified with the aim of minimizing an overall sample size (i.e., the sample size from the whole population) under restrictions on a level of precision of estimation in the domains. This problem can be presented as follows. We aim to find a set of subsets $\mathbf{a}_j$ ($\mathbf{a}_j$ defining the strata in the domain U_j, j=1,...,J) of a population U that minimizes the overall sample size subject to given values c_j of the coefficient of variation of a studied estimator in the domains.

Assuming that L_j denotes a number of strata to be constructed in the jth domain, the two-stage division of a population U means that

$$U = \bigcup_{j=1}^{J} U_j = \bigcup_{j=1}^{J} \bigcup_{h=1}^{L_j} U_{j,h} \text{ , and} \tag{5.1}$$

$$U_{j,h} \subset U_j \subset U,\ U_j \cap U_{j'} = \varnothing,\ U_{j,h} \cap U_{j,h'} = \varnothing$$

$$h, h'(h \neq h') = 1,\ldots,L_j,\ \ j, j'(j \neq j') = 1,\ldots,J$$

Let us define the optimization problem as follows: find sets $(\mathbf{a}_1,\ldots,\mathbf{a}_J)$ in the domains, $\mathbf{a}_j$ being related to the jth domain, that minimize the following objective function

$$f(\mathbf{a}_1,\ldots,\mathbf{a}_J) = \sum_{j=1}^{J} f(\mathbf{a}_j) \tag{5.2}$$

$$f(\mathbf{a}_j) = N_{jL_j} + \left(\sum_{h=1}^{L_j-1} W_{jh} S_{jh} \right)^2 \left(\overline{X}_j^2 c_j^2 + N_j^{-1} \sum_{h=1}^{L_j-1} W_{jh} S_{jh}^2 \right)^{-1} \tag{5.3}$$

where $f(\mathbf{a}_1,\ldots,\mathbf{a}_J)$ is the objective function; given $\mathbf{a}_j$, $f(\mathbf{a}_j)$ is the sample size required for the jth domain to obtain a level of precision of estimation equal to c_j; S_{jh}^2 is the population variance of the variable X_{jh}, X_{jh} being the variable X restricted to the hth stratum of the jth subpopulation; $\overline{X}_j$ is the population mean of X_j, X_j being the variable X restricted to the jth population; N_j is the size of the jth population; N_{jL_j} is the size of the take-all (L_j-th) stratum from the jth domain; and $W_{jh} = N_{jh} / N_j$.

The constraints for the functions (5.3) are

$$N_{jh} \geq 2\ ,\ j = 1,\ldots,J,\ \ h = 1,\ldots,L_j - 1\ ;\ N_{jL_j} \geq 1 \tag{5.4}$$

$$2 \leq n_{jh} \leq N_{jh}\ ,\ j = 1,\ldots,J,\ \ h = 1,\ldots,L_j - 1 \tag{5.5}$$

In evaluation of $f(\mathbf{a}_j)$ in Eq. (5.3), the Neyman-optimum formula is assumed. To check the constraints (5.5) for given j, the sample sizes n_{jh} $(h = 1,\ldots,L_j - 1)$ should be obtained via the Neyman-optimum allocation formula:

$$n_{jh} = \left[f\left(\mathbf{a}_j\right) - N_{jL}\right]\frac{S_{jh}W_{jh}}{\sum_{h=1}^{L_j-1} S_{jh}W_{jh}};\ h = 1,\ldots,L_j - 1;\ n_{L_j} = N_{L_j}$$

If any of the constraints from (5.5) is not fulfilled for a jth domain, the function (5.3) for this domain takes the form

$$f\left(\mathbf{a}_j\right) = N_{jL} + \sum_{h=1}^{L_j-1} n_{jh} \tag{5.6}$$

where integer sample sizes n_{jh} are determined by solving the following numerical problem:

Minimize $f\left(n_1,\ldots,n_{L_j-1}\right) = \sum_{h=1}^{L_j-1} n_h$

subject to constraints (5.4), (5.5), and $c_j = \overline{X}_j^{-1}\sqrt{\sum_{h=1}^{L_j-1} W_{jh}^2 S_{jh}^2\left(\frac{1}{n_{jh}} - \frac{1}{N_{jh}}\right)}$

The optimization problem (5.2), (5.4), and (5.5) may be treated as the sum of J sub-problems, each being the stratification problem in a particular domain. The final value of the objective function (5.2) is the sum of J values under the optimum solution in the sub-problems.

If one does not want to create the take-all strata in the domains, the formulae slightly change: summing is to be done from 1 to L_j, and the terms connected with N_{L_j} are omitted in all the formulae.

6. Multivariate stratification in domains minimizing overall sample size from population under domain-orientated approach

In this section, we consider stratification that is a multivariate generalization of the stratification discussed in the previous section. Consider the division (5.1) of a population under study U. Our objective is to find the set of strata boundaries within the domains that minimizes an overall sample size (i.e., the sum of sample sizes from the domains) with respect to assumed values c_{ij} of the coefficient of variation (cv) of the ith estimator in the jth domain. (Note that we can fix different levels of precision of estimation for different characters and domains.) In a jth domain, L_j strata are to be constructed. Let us consider I variables under study and assume that the population mean is the parameter of interest.

Let us formulate the optimization problem: Find the set of subsets $\mathbf{a}_j$, j=1,…,J, ($\mathbf{a}_j$ being a set of parameters defining the subdivision of the jth domain) that minimizes a sample size from the whole population with respect to fixed values c_{ij} of the coefficient of variation of estimators of the population mean of the variables in the domains. Therefore, the objective function may be written as

$$f(\mathbf{a}_1,...,\mathbf{a}_J) = \sum_{j=1}^{J} f(\mathbf{a}_j) \tag{6.1}$$

where $f(\mathbf{a}_1,...,\mathbf{a}_J)$ is the objective function and, given $\mathbf{a}_j$, $f(\mathbf{a}_j)$ is the minimum sample size from the jth domain that is required to obtain the given (c_{ij}) level of precision of estimation of the ith parameter in the jth domain. A procedure for evaluation of $f(\mathbf{a}_j)$ in the jth domain under the take-all-stratum approach is as follows. Given a set of strata boundaries $\mathbf{a}_j$, evaluate sample sizes for the variables Y_i, i=1,...,k, via the formula

$$f_i(\mathbf{a}_j) = N_{jL_j} + \left(\sum_{h=1}^{L_j-1} W_{jh} S_{ijh} \right)^2 \left(\overline{X}_{ij}^2 c_{ij}^2 + N_j^{-1} \sum_{h=1}^{L_j-1} W_{jh} S_{ijh}^2 \right)^{-1} \tag{6.2}$$

$$i = 1,...,I,\ j = 1,...,J$$

where $f_i(\mathbf{a}_j)$ is the minimal sample size for the ith variable from the jth domain that is required to obtain the value c_{ij} of the coefficient of variation of the estimator of the population mean of Y_i; S_{ijh}^2 is the variance of the variable X_{ijh}, X_{ijh} being the variable X_i restricted to the hth stratum of the jth domain; and $\overline{X}_{ij}$ is the population mean of X_{ij}. Next, based on the $f_i(\mathbf{a}_j)$ values, evaluate (Lednicki and Wieczorkowski, 2003)

$$f(\mathbf{a}_j) = N_{jL_j} + \sum_{h=1}^{L_j-1} n_{jh} \tag{6.3}$$

$$n_{jh} = \max_{i=1,...,I} \left\{ \left(n_j - N_{jL_j} \right) a_{ijh} \right\},\ n_j = \max_{i=1,...,I} \left\{ f_i(\mathbf{a}_j) \right\}$$

$$a_{ijh} = W_{jh} S_{ijh} \left(\sum_{h=1}^{L_j-1} W_{jh} S_{ijh} \right)^{-1}$$

Again, the function (6.3) assumes the Neyman-optimum sample allocation between strata. The constraints for the function (6.2) are

$$N_{jh} \geq 2,\ j = 1,...,J, h = 1,...,L_j - 1, \text{ and } N_{jL_j} \geq 1 \tag{6.4}$$

$$2 \leq n_{jh} \leq N_{jh},\ j = 1,...,J, h = 1,...,L_j - 1 \tag{6.5}$$

If any of the constraints (6.5) is not fulfilled for some j, n_{jh} from Eq. (6.3) should be determined by solving the following numerical problem (all ns are to be integers):

Given $\mathbf{a}_j$, minimize $f\left(n_1,...,n_{L_j-1}\right)=\sum_{h=1}^{L_j-1} n_{jh}$

subject to constraints (6.4), (6.5), and $c_{ij} \le \bar{X}_{ij}^{-1}\sqrt{\sum_{h=1}^{L_j-1} W_{jh}^2 S_{ijh}^2\left(\frac{1}{n_{jh}}-\frac{1}{N_{jh}}\right)}$ (6.6)

We have considered inequality in Eq. (6.6) because quite often it is impossible to achieve the exact assumed values of the coefficient of variation for many estimators; then, it is enough to set the upper bounds of their values, as we did in Eq. (6.6).

7. Domain-orientated multivariate stratification with respect to fixed sample sizes in domains

Consider the same population and multivariate stratification variable as in previous section. This time our objective is to design a survey under a fixed survey cost for each domain, so we have at our disposal a vector $\mathbf{n}=\left(n_1,...,n_j\right)^T$ of known sample sizes from the domains; such a situation is quite often encountered in survey practice. We aim to stratify the domains so that the maximum value of the coefficient of variation of considered estimators in the domains is minimal; such an approach aims to achieve as equal levels of precision of estimators in a particular domain as possible. The greatest "attention" is then paid to the estimators that have the worst precision; those estimators that are precise, for instance because of small variation in the corresponding variables, have less influence on stratification. Let us assume that a take-all stratum in each domain is to be constructed; as previously, formulae for the non-take-all-stratum approach easily follow from the formulae for the take-all-stratum approach so they are not given.

We could present, as a matter of fact, this stratification problem for a population that is not subdivided into domains. This is because stratification is to be performed independently in each domain, and results for one domain are not related to results for other domains. Therefore, the problem is not formulated as one general optimization function (as it was done previously), but as a set of J functions, each function being related to a particular domain. Nevertheless, because quite often populations are subdivided into domains, we have decided to retain the notation connected with domains of a population.

The problem for a jth domain is formulated as follows. Find a set $\mathbf{a}_j$ of parameters defining strata that minimizes the objective function

$$f(\mathbf{a}_j) = \max_{\substack{j=1,\dots,J,\\ i=1,\dots,I}} \{w_{ji} c_{ji}\} \tag{7.1}$$

$$c_{ji} = \frac{1}{\overline{X}_{ji}} \sqrt{\sum_{h=1}^{L_j-1} W_{jh}^2 S_{jih}^2 \left(\frac{1}{n_{jh}} - \frac{1}{N_{jh}} \right)}$$

where w_{ji} is the importance weight of the *i*th variable in the *j*th domain. Note that in the preceding stratification problems we did not have to include importance weights. Here it is important because sometimes one deals with some key variables, which are more important than other variables, and thus the precision of the estimators related to them is more important than that of other estimators. Conventionally, the weights w_{ji} should satisfy the condition

$$\sum_{i=1}^{I} w_{ji} = 1 \text{ for } j = 1,\dots,J$$

although it is not required from the theoretical viewpoint.

The weights should be chosen by experts designing a survey. Sometimes they might be evaluated in a different way. For example, adapting the approach by Skibicki and Wywiał (2003), we could use the weights

$$w_{ji} = S_{ji}^2 \left(\sum_{i'=1}^{I} S_{ji'}^2 \right)^{-1}$$

where S_{ji}^2 is the variance of the *i*th variable in the *j*th domain.

The constraints for the function (7.1) are

$$N_{jh} \geq 2\,,\ j = 1,\dots,J,\ \ h = 1,\dots,L_j - 1 \text{; and } N_{jL_j} \geq 1 \tag{7.2}$$

$$2 \leq n_{jh} \leq N_{jh}\,,\ j = 1,\dots,J,\ \ h = 1,\dots,L_j - 1 \tag{7.3}$$

$$\sum_{h=1}^{L_j-1} n_{jh} = n_j - N_{L_j} \tag{7.4}$$

Note again that to minimize the function (7.1) one needs to determine the optimum values of $\mathbf{a}_j$ and $\mathbf{n}_j = (n_{j1},\dots,n_{jL_j-1})^T$. Given $\mathbf{a}_j$, one is to determine the elements of the vector $\mathbf{n}_j$ that minimize the function (7.1) under the constraints (7.3) and (7.4). Because this problem cannot be solved directly using an analytical formula, based on Holmberg's (2002) results, Kozak (2004a, 2006b) proposed to use the following approximate formula

$$n_{jh} = \frac{n_j W_{jh} b_{jh}}{\sum_{g=1}^{L_j-1} W_{jg} b_{jg}} \tag{7.5}$$

$$b_{jh} = \sqrt{\sum_{i=1}^{I} d_{ji}^{-1} S_{jih}^2}$$

$$d_{ji} = \left(n_j - N_{L_j}\right)^{-1} \left(\sum_{h=1}^{L_j-1} W_{jh} S_{jih}\right)^2 - N_j^{-1} \sum_{h=1}^{L_j-1} W_{jh} S_{jih}^2$$

where d_{ji} is the variance of the mean's estimator for the *i*th variable in the *j*th domain under the univariate Neyman-optimum allocation for this variable. Such an approach is approximate because, given $\mathbf{a}_j$, it does not minimize the function (7.1) but the maximum loss of the efficiency of estimation under the multivariate sample allocation compared with the univariate allocation. To determine the optimum sample allocation, one may solve numerically the following problem:

Given $\mathbf{a}_j$, minimize $f\left(n_1,\ldots,n_{L_j-1}\right) = \max\limits_{\substack{j=1,\ldots,J,\\ i=1,\ldots,I}} \{w_{ji} c_{ji}\}$

subject to constraints (7.2), (7.3), and (7.4).

The above problem has to be solved also when the allocation (7.5) does not fulfill the constraints (7.3). A statistician should decide which sample allocation, the optimum (with sample allocation being the solution of the above problem) or approximate (with sample allocation (7.5)) one, should be chosen.

8. Discussion and conclusions

A population under study is stratified based on auxiliary character(s) that should be highly correlated with variable(s) under study. Efficiency of stratification depends on quality of auxiliary information and a stratification approach used. Consequently, if the auxiliary information is of good quality, the efficiency depends mainly on the stratification approach. If the information is not up-to-date, stratification may be far from optimum (it may even be less efficient than simple random sampling). Hence, it is very important to use stratification variable(s) of high quality.

The approach to optimum stratification presented in this paper is based on the use of numerical minimization of an optimization function with respect to certain constraints, the function and constraints being defined based on the stratification problem. The optimization problems we have considered ought to be solved using numerical methods. In the univariate case, optimization-based stratification is not troublesome since the optimization function and the constraints are not too complex. Nevertheless, the multivariate case may be troublesome, especially when there are many stratification variables; the choice of an optimization method

to be applied then is an important issue. For instance, the simplex method of Nelder and Mead (1965), which was used by Lednicki and Wieczorkowski (2003), may occur to be inefficient when a number of parameters searched for increases. In the multivariate stratification presented in section 7, there are k (L–1) parameters searched for in each domain. If one wants to construct seven strata for four variables without a take-all stratum, there are 24 parameters to be determined and 21 constraints to be fulfilled. Then, if a population size is large, classical optimization methods, as the simplex method, may be inefficient (Kozak, 2004b). However, in the case of univariate stratification of small populations, Nelder and Mead's (1965) method may be efficient because then it usually provides similar results as other stratification approaches (as, e.g., those by Lavalleé and Hidiroglou (1988) or a random search method (Niemiro,1999; Kozak, 2004b; Kozak and Verma, 2006)). Even for very large populations this method is usually efficient although a slight gain may be obtained by applying global optimization methods (Kozak, 2004c). The modification technique presented by Kozak (2004b) for univariate stratification and adapted in this paper to multivariate stratification makes the optimization works on smaller datasets. In such a way, the stratification is more efficient in terms of time consumption used for optimization of a given function.

Last but not least, an important issue to be addressed is shape (geometry) of strata. We have mentioned several possible shapes, such as L-rot-geometry, stratification by cutpoints and by cluster analysis. Other shapes may also be used; each one has some pros and cons. For instance, L-rot-geometry-shaped strata (Briggs and Duoba, 2000; Lednicki and Wieczorkowski, 2003) are easy to implement in some programming languages, as in R language (R Development Core Team, 2006), which was used, e.g., by Lednicki and Wieczorkowski (2003). There is no evidence, however, that this shape is the most efficient one; moreover, it is easy to imagine that there are more efficient strata shapes than that provided by the L-rot-geometry. On the other hand, in cutpoints-based stratification it may be difficult to obtain non-empty strata (see Figure 2: strata L5 and L9 have one item each). As a result, it may be troublesome to obtain the desirable number of non-empty strata, or to obtain efficient stratification under a desirable number of strata. Stratification by classical cluster analysis, on the other hand, is not efficient. It is so because cluster analysis has different aims than stratification has (more precisely, cluster analysis aims to optimize a different function under different constraints). Therefore, stratification by cluster analysis needs further work on to be adapted to stratification specificity.

It follows from this discussion what should be studied further. Two main directions of the investigation on multivariate stratification should be pointed out. First, the research should focus on determining the optimum shape of strata. Second, optimization method that would be efficient in stratification under the determined strata shape should be recognized; software that would handle this problem also ought to be chosen. It is possible that the optimum optimization

method for particular strata geometry is not optimum for other strata geometries. In addition, we are not able to claim that the same strata geometry is optimum for different optimization problems.

Which one of these two problems needs to be concentrated on more? This question is not easy to answer. As already mentioned, the perceptibly better results can be obtained by optimizing shape of strata and then by determining optimization method to be applied than by determining an optimization method to be applied for the L-rot-geometry. Nevertheless, these both problems are strictly connected, and the future work should give the solution of a combined problem, which is, "Which shape geometry and numerical method to optimize stratification under the geometry chosen should be applied?"

An additional problem is a choice of initial points to be applied in the optimization-based stratification. As already mentioned, this may have quite an impact on stratification efficiency, especially in multivariate stratification. In univariate stratification, approximate stratification method should be used (Kozak, 2004c; Kozak and Verma, 2006). In multivariate stratification, this issue is more complex and needs separate studies. One could apply a combined matrix of univariate approximate stratification points, but it is quite likely that such a matrix would not fulfill multivariate stratification constraints (especially under a large number of variables). Then, to optimize the optimization-based approach to stratification, research on optimum initial parameters in multivariate stratification should be conducted. The ideal initial points should fulfill the following requirements: (a) they should be easy to obtain (as approximate univariate stratification points are); (b) they should fulfill constraints of a particular stratification problem; and (c) they should provide optimum or near-optimum initial points for optimization, which means that the use of them should provide the optimum stratification.

REFERENCES

BETHEL, J (1989): Sample allocation in multivariate surveys, *Survey Methodology*, 15, 47—57.

BRIGGS, J., DUOBA, V. (2000): STRAT2D: Optimal bivariate stratification system. Statistics New Zealand (http://www.stats.gov.nz).

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

DALENIUS, T. (1950): The problem of optimum stratification, *Skandinavisk Aktuarietidskrift*, 33, 203—213.

DALENIUS, T. (1952): The problem of optimum stratification in a special type of design, *Skandinavisk Aktuarietidskrift*, 35, 61—70.

DALENIUS, T., HODGES, J.L. (1959): Minimum variance stratification, *Journal of the American Statistical Association*, 54, 88—101.

DORFMAN, A.H., VALIANT, R. (2000): Stratification by size revisited, *Journal of Official Statistics*, 16, 139—154.

EKMAN, G. (1959): An Approximation useful in univariate stratification, *Annals of Mathematical Statistics*, 30, 219—229.

GHOSH, S.P. (1963): Optimum stratification with two characters, *Annals of Mathematical Statistics*, 34, 866—872.

GODFREY, J., ROSHWALB, A., WRIGHT, R.L. (1984): Model-based stratification in inventory cost estimation, *Journal of Business and Economic Statistics*, 2, 1—9.

GOLDER, P.A., YEOMANS, K.A. (1973): Use of cluster analysis for stratification, *Applied Statistics*, 22, 213—219.

GUNNING, P., HORGAN, J.M. (2004): A simple algorithm for stratifying skewed populations, *Survey Methodology*, 30(2), 159—166.

GREŃ, J. (1964): Some Methods of the Sample Allocation in Stratified Multi-Parameter Sampling, *Przegląd Statystyczny*, 11, 361—369, (in Polish).

GREŃ, J. (1966): Some Application of the Non-Linear Programming in Sampling Method, *Przegląd Statystyczny*, 13, 203—217, (in Polish).

GUPTA, P.C., SETH, G.R. (1979): On stratification in sampling investigation involving more than one character, *Journal of Indian Society of Agricultural Statistics*, 31(2), 1—15.

HARTLEY, H. O. (1965): Multiple Purpose Optimum Allocation in Stratified Sampling. *Proceedings of the American Statistical Association, Social Statistics Section*, 258—261.

HEDLIN, D. (2000): A procedure for stratification by an extended Ekman Rule, *Journal of Official Statistics*, 16 (1), 15—29.

HEELER, R.M., DAY, G.S. (1975): Supplementary note on the use of cluster analysis for stratification, *Applied Statistics*, 24, 342—344.

HIDIROGLOU, M. (1986): The construction of a self-representing stratum of large units in survey design, *The American Statistician*, 40, 27—31.

HOLMBERG, A. (2002): A multiparameter perspective on the choice of sampling design in surveys, *Statistics in Transition*, 5, 969—994.

HORGAN, J.M. (2006): Stratification of skewed populations: a review, *International Statistical Review*, 74(1), 67—76.

KOKAN, A. R., KHAN, S. (1967): Optimum allocation in multivariate surveys: an analytical solution, *Journal of the Royal Statistical Society*, B 29, 115—125.

KOZAK, M. (2004a): New approach towards stratification of multidimensional population, *Wiadomości Statystyczne*, 12, 12—20 (in Polish).

KOZAK, M. (2004b): Optimal stratification using random search method in agricultural surveys, *Statistics in Transition*, 6(5), 797—806.

KOZAK, M. (2004c): Optimal stratification with the auxiliary variable, *Wiadomości Statystyczne*, 8, 29—34 (in Polish).

KOZAK, M. (2006a): Multivariate sample allocation: application of random search method. *Statistics in Transition*, 7(4), 889—900.

KOZAK, M. (2006b): On sample allocation in multivariate surveys, *Communications in Statistics: Simulation and Computation*, 35(4), 901—910.

KOZAK, M. and VERMA, M.R. (2006): Geometric versus optimization approach to stratification: a comparison of efficiency, *Survey Methodology*, 32(2), 157—163.

KURSA, L., LEDNICKI, B. (2006): The agricultural sample surveys in Poland in transition period, *Statistics in Transition*, 7, 981—1008.

LAVALLÉE, P. and HIDIROGLOU, M. (1988): On the stratification of skewed populations, *Survey Methodology*, 14, 30—43.

MAHALANOBIS, P.C. (1952): Some aspects of the design of sample surveys, *Sankhya*, 12, 1—7.

MULVEY, J.M. (1983): Multivariate stratified sampling by optimization, *Management Science*, 29, 715—724.

MULVEY, J.M. and CROWDER, H.P. (1979): Cluster analysis: an application of Lagrangian relaxation, *Management Science*, 25, 329—340.

NELDER, J.A. and MEAD, R. (1965): A Simplex Method for Function Minimization, *Computer Journal*, 7, 308—313.

NIEMIRO, W. (1999): Optimal Construction of Strata Using Random Search Method, *Wiadomości Statystyczne*, 10, 1—9 (in Polish).

R DEVELOPMENT CORE TEAM, (2006): R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria; URL http://www.R-project.org.

RIVEST, L.P. (2002): A Generalization of Lavallee and Hidiroglou algorithm for stratification in Business Surveys. *Survey Methodology*, 28, 207—214 (http://www.mat.ulaval.ca/pages/lpr/).

RIZVI, S.E.H., GUPTA, J.P., BHARGAVA, M. (2002): Optimum stratification based on auxiliary variable for compromise allocation, *Metron*, LX (3—4), 201—215.

RIZVI, S.E.H., GUPTA, J.P., SINGH, R. (2000): Approximately optimum stratification for two study variables using auxiliary information, *Journal of Indian Society of Agricultural Statistics*, 53(3), 287—298.

SADASIVAN, G., AGGARWAL, R. (1978): Optimum points of stratification in bivariate populations, *Sankhya* C, 40, 84—97.

SAMANTA, M. (1965): A Note on the problem of optimum stratification of a bivariate population in stratified random sampling, *Annals of Institute of Statistical Mathematics*, 17, 363—375.

SÄRNDAL, C.E., SWENSSON, B., WRETMAN, J. (1992): *Model assisted survey sampling*, Springer-Verlag.

SCHNEEBERGER, H. (1970): Optimierung in der stichprobentheorie durch schichtung, *Statistische Hefte*, 11(4), 242—253.

SCHNEEBERGER, H., POLLOT, J.P. (1985): Optimum stratification with two variates, *Statistische Hefte*, 26, 97—113.

SETHI, V.K. (1963): A Note on the optimum Stratification of Populations for Estimating the Population Means, *Australian Journal of Statistics*, 5, 20—33.

SINGH, R. (1971): Approximately optimal stratification of the auxiliary variable, *Journal of the American Statistical Association*, 66, 829—834.

SKIBICKI, M., WYWIAŁ, J. (2003): On Simultaneous Stratification of Population and Optimal Sample Allocation in Strata. In: Wywial J (Ed.) Metoda reprezentacyjna w badaniach społecznych (Survey Sampling in Social Science), Wydawnictwo Akademii Ekonomicznej, Katowice, pp. 157—168.

SWEET, E.M., SIGMAN, R.S. (1995): Evaluation of Model-Assisted Procedures for Stratifying Skewed Populations using Auxiliary Data, U.S. Bureau of the Census, (http://www.census.gov/srd/papers/pdf/sm95-22.pdf).

THOMPSEN, I. (1976): A Comparison of Approximately Optimal Stratification Given Proportional Allocation with other Methods of Stratification and Allocation, *Metrika*, 23, 15—25.

VERMA, M.R., JOOREL, J.P. SINGH, AGNIHOTRI, R.K. (2003): Approximately optimum stratification for two sensitive variables in the case of compromise method of allocation, *Sri Lankan Journal of Applied Statistics*, 4, 1—13.

WYWIAŁ, J. (1988): Minimizing the Spectral Radius of Means Vector from Sample Variance-Covariance Matrix Sample Allocation between Strata,

Prace Naukowe Akademii Ekonomicznej we Wrocławiu, 404, 223—235, (in Polish).

YEOMANS, K.A., GOLDER, P.A. (1975): Further Observations on the Stratification of Birmingham Wards by Clustering: A Riposte, *Applied Statistics*, 24, 345—346.

Figure 1. Stratification via the L-rot-180 geometry; three strata are constructed based on the 2×2 matrix **a**.

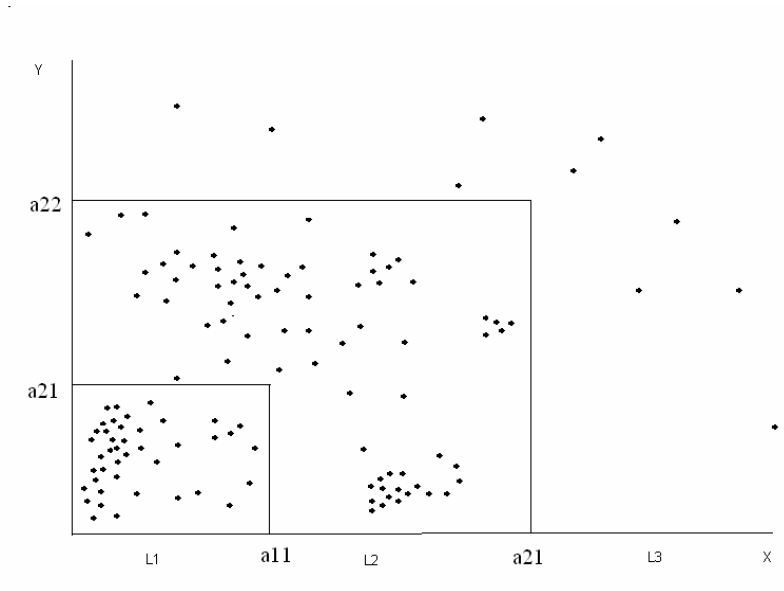
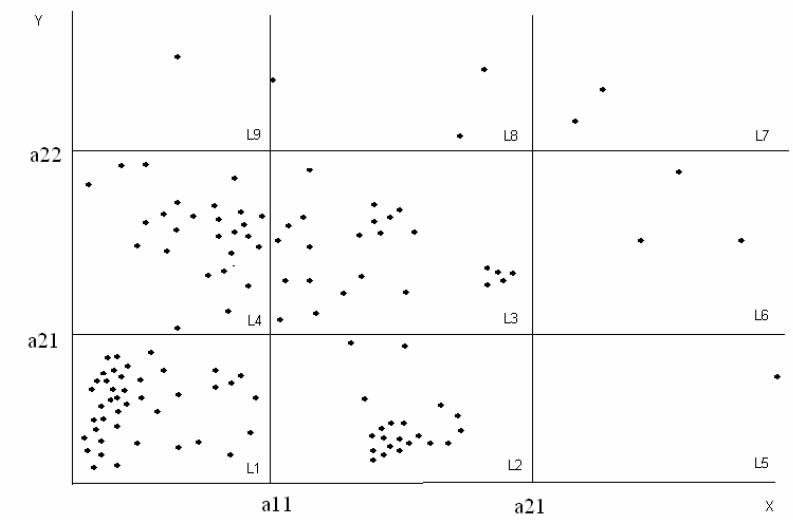


Figure 2. Stratification via the cut-points; nine strata are constructed based on the 2×2 matrix **a** (the same as in Fig. 1).



STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 251—272

INNOVATIVE ROLE OF IMPUTATION METHOD/ANALYSIS IN CLINICAL TRIAL ON REMNANT ABLATION IN DIFFERENTIATED THYROID CANCER*

Prem Chandra[1], SN Dwivedi[1], CS Bal[2], Ajay Kumar[2], Arvind Pandey[3]

ABSTRACT

Background: Prognostic models formalize the multivariate relationships between multiple patient characteristics and outcome and can be useful tools to aid clinical decision-making. The construction of a prognostic model ideally requires a large database with complete information on all potential prognostic factors. However, often some cases have missing covariate data, which may introduce bias and lead to misleading conclusions if handled inappropriately. Therefore it becomes essential to impute those missing values and then develop appropriate models based on complete information's. This paper intended to emphasize the role of imputation analysis in the area of remnant ablation of differentiated thyroid cancer patients. **Methods:** Between July '95 and January '02, a total of 565 patients with differentiated thyroid cancer, who fulfilled the inclusion criteria were recruited under this study. In view of the fact that under the present data set the values in relation to some of the important covariates were found to be missing, multiple regression method was used to impute the missing covariate values of quantitative variables. Whereas, discriminant function analysis was used to impute missing values related to categorical covariates. **Results:** Comparative results i.e., results obtained before and after imputation analysis have been presented separately in detail. Regression models obtained after imputation analysis i.e., based on complete information revealed gained in validity and precision of the developed model. It was found that model obtained after imputation provides not only gain in statistical significance but also have it's clinically relevance. **Conclusion**: The present study has

---

* Part of this work was presented at the 27th Annual Meeting of the Society for Clinical Trials, at Orlando, Florida (USA), from 21-24 May, 2006.
Department of [1]Biostatistics and [2] Nuclear Medicine, All India Institute of Medical Sciences, New Delhi, India;
[3] Director, National Institute of Medical Statistics, New Delhi, India.
Email: prem_s77@rediffmail.com

undoubtedly indicated the usefulness of imputation method in clinical trial regarding remnant ablation of differentiated thyroid cancer. Results obtained after imputation analysis had not only statistical significance but also had clinical relevance.

Keywords: Differentiated thyroid cancer, remnant ablation, prognostic model, imputation analysis

1. Introduction

Prognostic models formalize the multivariate relationships between multiple patient characteristics and outcome and can be useful tools to aid clinical decision-making. The construction of a prognostic model ideally requires a large database with complete information on all potential prognostic factors. However, often some cases have missing covariate data, which may introduce bias and lead to misleading conclusions if handled inappropriately.

Missing data arise for a multiplicity of reasons, and indeed are so common that one should be highly suspicious of any clinical trial report that claims to have no missing data. More commonly a patient fails to attend a scheduled study visit, or a blood sample goes astray or is inadequately refrigerated. Sometimes, patients retrospectively withdraw their consent to participate in the trial. Equally, since clinical trials protocols tend to evolve over time, a new baseline covariate may be added to the record form, or a simplified quality of life questionnaire replaces a version that was unacceptable to the patients. Of course, such problems should be identified during pilot studies, but time constraints are such that it is not always possible to conduct adequate pilot tests. If there are changes in data collection during the course of a trial, then relevant data may not be recorded for the early patients.

There are distinctly different kinds of missing data that arise in clinical trials, and there is no single approach that can cope with this variety of missing data. Before attempting to discuss solution to the handling of missing data, it is helpful to classify different types of “missingness”. Rubin (1976) gives a useful taxonomy, and this terminology for the different mechanism that can generate missing data is widely accepted. The first class comprises data that are “missing completely at random” (MCAR). Here, the presence or absence of an observation is completely unrelated to the value that might have been observed. For example, if new question is added to record form, then the data for early patients are expected to be MCAR. However, even in this simple situation, the missing data could fail to be MCAR if there is some underlying secular trend in the prognosis of the condition being studied. Similarly, if a laboratory test is not performed because equipment is out of order, then the data are most likely MCAR. The second class comprises data that are “missing at random” (MAR). In this class, the fact that an observation is missing, after conditioning on the observed data,

provided no further information. Murray and Findlay (1988) give an example of a study of hypertension, in which patients were withdrawn from the study if, at a study assessment, their diastolic blood pressure (DBP) exceeded 110 mmHg. Thus, the fact that an individual's DBP was not recorded at the eight- week of assessment provided no additional information to the observation that, at the four week assessment, the DBP exceeded the threshold of 110 mmHg. The final class comprises "nonignorable missing data". The most familiar example here is censored data, where it would be invalid to base inferences on a likelihood function that incorporate data only form the uncensored observation.

Determining the appropriate analytic approach in the presence of incomplete observation is a major question for data analysts. The development of statistical methods to address missing data has been an active area of research in recent decades (Rubin 1976; Little and Rubin 1987; Laird 1988; Ibrahim 1990). There are many strategies available for handling missing data (Little and Rubin, 1987; Schafer and Graham, 2002). These include the simple deletion approaches of complete case analysis, where only the cases with complete data for all collected variables are analysed, available case analysis, where the cases with complete data for the variables in the fitted model are analysed utilising the largest possible data set, and variable omission (Vach, 1997), where the incomplete variable is excluded from the model. Other techniques, utilising all cases, include analysing the missing data as a separate category (Greenland and Finkle, 1995), single imputation (Little and Rubin, 1987), in which a single value is substituted for each missing value, and multiple imputation (Rubin, 1987; Schafer, 1997), where more than one independently completed data sets are obtained.

Imputation, the practice of 'filling in' missing data with plausible values, is an attractive approach to analyzing incomplete data. It apparently solves the missing-data problem at the beginning of the analysis. However, a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests, as documented by Little and Rubin (1987) and others. The question of how to obtain valid inferences from imputed data was addressed by Rubin's (1987) book on multiple imputation (MI). Single imputation (SI) is a common way to deal with missing data. Each missing value is replaced by an imputed value, for example by interpolation. Single imputation refers to filling in a missing value with a single replacement value. There are two general approaches: arbitrary methods and regression-based imputation. Some researchers use arbitrary methods to impute missing data. Some of these including using (a) the mean of all observed values for all people, (b) the mean observed value for the same person in other time periods, (c) the mean of the previous and following values for the person, if they exist, or (d) the most recent observed value for the person. This latter method, known as last-observation-carried-forward (LOCF), is quite common. Other arbitrary methods can be created as well. In Regression-based imputation, the analyst estimates a regression model in which the dependent variable has missing values for some observations. In the

second step, the estimated regression coefficients are used to predict (impute) missing values of that variable. The proper regression model depends on the form of the dependent variable. However, the complexity of regression models used in imputation analysis should be carefully thought through by clinical trial practitioners, because the method assumes that the missing data process can be fully captured by the regression model employed on observed values. This assumption is called missing at random (MAR). MAR essentially says that the cause of the missing data may be dependent on observed data (such as data of previous visits) but must be independent of the missing value that would have been observed. However, discriminant Analysis is used to predict the missing values in case of categorical variable like type of gland.

Each strategy for handling missing data has an underlying assumption regarding the missing data mechanism (Little and Rubin, 1987), that is, the reasons for the occurrence of the missing covariate data, which if not satisfied could result in biased parameter estimates. For example, the commonly used complete case analysis assumes that the missingness in the covariates is not associated with the outcome (Vach and Blettner, 1998). Most single imputation and multiple imputation approaches assume that the missingness is related to the observed data but does not depend on the unobserved value itself.

To the best of our knowledge, there is no study available on missing data analysis in the area of remnant ablation of differentiated thyroid cancer (DTC) patients. Therefore, this study was intended to emphasize the importance of imputation methods (using multiple regression and discriminant analysis approach) in clinical trial on remnant ablation of differentiated thyroid cancer patients. Also, the comparative analysis (i.e., analysis before and after imputation) have been carried out and presented in detail.

2. Study Design

All India Institute of Medical Sciences, New Delhi, India, is a tertiary care teaching hospital serving about half-a-billion population of Northern India. The Department of Nuclear Medicine has been running a specialty thyroid cancer clinic for the last 30 years. Currently, 250-260 new differentiated thyroid cancer (DTC) patients are being treated annually. Between July ’95 and January ’02, a total of 565 patients with differentiated thyroid cancer, who fulfilled the inclusion criteria (patients having disease confirmed to be limited to the thyroid only by clinical, radiological, per operative and post-surgical ^{131}I scintigraphic examination and having no evidence of extrathyroidal or distant metastases at the time of presentation) were clinically treated with eight different doses of radioiodine; starting at 15 mCi and increasing the activity in increments of 5 mCi until 50 mCi. However, 56 patients were excluded for various reasons. Nineteen patients in whom nodal/distant metastases (nine nodal and ten pulmonary/skeletal metastases) were revealed in post therapy ^{131}I whole-body scans (WBSs) were excluded from the study and another 37 patients were lost to follow-up. Therefore,

finally data related to 509 patients were used in the analysis. Distribution of the patients (509) according to availability of the data for the different variables is given in table-1.

Table 1. Distribution of the patients (509) according to availability of the data for the different variables

Variable	Available Data	Missing cases (in %)
Age	509	0
Sex	509	0
Type of Surgery	509	0
Histopathology	509	0
First Dose of ^{131}I	509	0
Outcome after 1st Dose	509	0
Final Out Come	509	0
Post surgical radioiodine uptake	472	37 (7.2 %)
Duration of Illness	433	76 (14.9%)
Interval between Surgery & Therapy	446	63 (12.4%)
Radiation Absorbed Dose	472	37 (7.2%)
Type of Gland	451	58 (11.4%)
Tumor size	230	279 (54.8%)

3. Methods used for Imputation Analysis

Taking into account the merits/demerits of various imputation techniques, the regression approach (for quantitative dependent variables) and discriminant function analysis approach (for qualitative dependent variables) were used in the present analysis.

a) Multiple Regression Analysis (for quantitative missing covariate values)

Multiple stepwise regression method was used to impute the missing values of quantitative variables in various sub-categories of patients, such as (i) male ablated, (ii) female ablated, (iii) male non-ablated, and (iv) female non-ablated. A

series of regression equations considering each of the covariates with missing values as a dependent variable and covariates with no missing values as independent variables were generated. Based on the resulting model, the missing values were imputed for each variable separately (Little, 1992).

The following sets of regression equations (considering each subgroup) were used for imputation analysis, that is, to estimate the missing values of respective covariate.

Regression Equations:

Ablated Male

Tumor size = 1.179 + 0.0898 * age
RAIU1[a] = 18.592 – 12.042 * type of surgery + 0.587 * first dose of ^{131}I - 0.0953 * age
Radiation absorbed dose = 11840.516 + 4838.24 * first dose of ^{131}I - 325.693 * age
Duration of illness = 1.24 + 1.039 * age

Ablated Female

Tumor size = 2.988 + 0.03675 * age
RAIU1[a] = 13.915 – 5.474 * type of surgery - 0.0378 * follow-up time
Duration of illness = 2.944 + 1.439 * age - 21.342 * type of surgery
Radiation absorbed dose = 3041.086 + 4322.95 * first dose of ^{131}I

Non Ablated Male

Tumor size = -5.427 + 0.0338 * age
Durtn of illness = -8.692 + 1.261 * age + 8.071* first dose of ^{131}I - 41.287 * histopathology
Radiation absorbed dose = 18936.250 + 4924.874 * first dose of ^{131}I - 453.586 * age

Non Ablated Female

Tumor size = 1.119 + 0.0748 * age
RAIU1[a] = 20.223 – 5.017 * type of surgery - 5.847 * histopathology
Duration of illness = -18.578 + 0.803 * follow-up time - 1.024 * age
Radiation absorbed dose = 11433.625 + 3558.794 * first dose of ^{131}I

[a] radioiodine uptake (before therapy)

In case, where none of the factors were found to be significant (although these cases were very few), unconditional mean/median of respective classes were substituted to impute the missing values of those corresponding covariates.

b) Discriminant function analysis (for qualitative missing covariate values)

Discriminant function analysis (Altman, 1990; Pocock, 1983; Armitage, 2001) was used to determine which variables discriminate between two or more naturally occurring groups. Probably the most common application of discriminant function analysis is to include many measures in the study, in order to determine the ones that discriminate between groups. In another way, we want to build a "model" of how we can best predict to which group a case belongs. A discriminant function, also called a canonical root, is a latent variable which is created as a linear combination of independent variables, so that :

$$L = b_1 x_1 + b_2 x_2 + b_3 x_3 + ... + b_n x_n + c$$

Where, the b's are discriminant coefficients, the x's are discriminating variables, and c is a constant. This is analogous to multiple regression, but the b's are discriminant coefficients which maximise the distance between the means of the criterion (dependent) variable. To impute the missing covariates values which is of categorical type, discriminant function analysis was used (Little & Rubin, 2002; Little, 1992; Laurikkala et al, 1999). Since, some of the values corresponding to categorical variable as type of gland (solid thyroid nodule or multinomial goiter/thyroid swelling) were also found to be missing under the present study, the discriminant function analysis (DFA) was used in order to impute those missing values for ablated and non-ablated cases separately. The linear discriminant function equations were as follows:

Ablated Group

Type of gland (stn[a]) = - 7.911 + 0.236 *age + 1.114* first dose
Type of gland (mng[b]) = - 8.404 + 0.246 *age + 1.138* first dose

Non-Ablated Group

Type of gland (stn) = - 4.773 + 0.155 *age + 0.646* first dose
Type of gland (mng/ts) = - 5.497 + 0.177 *age + 0.628* first dose

[a] solid thyroid nodule; [b] multinodular goiter/thyroid swelling

4. Results

Characteristics of Data

The mean age of the patients was 37.5±12.7 years with female to male ratio of 2.6. Median duration of the illness before surgery was 18 months (mean= 37.7, SD= 54.6). Mean tumor size was 4.6 ± 2.4 cm ranging from 1 cm to 16 cm. The large tumor size was probably due to late referral to surgery. The initial surgical intervention was not uniform due to different surgical units operating upon them. The surgical procedure followed was near total thyroidectomy in 74% patients and subtotal- or hemithyroidectomy in the rest. The histopathological diagnosis was established in all patients and found that four hundred and ten (80.5%) patients had papillary thyroid carcinoma, and 99 (19.5%) had follicular thyroid carcinoma. The median interval between surgery and referral to Nuclear Medicine Department, AIIMS, for ^{131}I therapy was 2 months. The mean post-surgical RAIU was 9.1±7.2% and estimated mean radiation absorbed dose after the first dose of ^{131}I was 240±234 Gy.

It is evident from the table 2 that the values of all the baseline quantitative (age, duration of illness, tumor size, post surgical radioiodine uptake, interval between surgery and therapy, radiation absorbed dose to the tumor) and qualitative (gender, type of gland, type of surgery, histopathology) variables were almost comparable (except one variable i.e., radiation absorbed dose) across the eight different groups of radioiodine doses ($p>0.05$).

5. Missing Problem in the Data

It was observed that the values related to few of the important covariates like tumor size, duration of illness, post surgical radioiodine uptake, interval between surgery and therapy, radiation absorbed dose, and type of gland were missing in the data set used under this study. The proportion of missing values related to respective covariates has been given in Table 1. On account of missing covariates' values present in the data set, it becomes vital to estimate first those missing values using appropriate statistical methods/techniques and then developed the prognostic model based on complete data set. This may eventually enhance the model validity and precision as well. Hence, multiple regression method (for quantitative dependent variable) and discriminant function analysis (for qualitative dependent variable) were used to impute the missing values of the quantitative and qualitative covariates respectively (Little, 1992).

For the missing data analysis, the group of patients was considered in three subgroups as described below:

Subgroup 1: The number of patients for whom values of all the considered covariates were completely available.

Table 2. Distribution of Patients in Relation to Demographic and Clinical Profiles

Variables	15 mCi	20 mCi	25 mCi	30 mCi	35 mCi
Age (years)[1]	36.5±10.7 (36, 14-60) n = 47	37.7±11.5 (36, 8-70) n = 55	38.9±13.7 (38.5,12-73) n = 70	37.3±13.2 (37, 13-74) n = 73	35.3±10.8 (32, 12-63) n = 63
Sex[2] **Male** **Female**	 14 (29.8) 33 (70.2)	 16 (29.1) 39 (70.9)	 16 (22.9) 54 (77.1)	 16 (21.9) 57 (78.1)	 24 (38.1) 39 (61.9)
Duration of illness[1] **(months)**	26.2± 36.2 (12, 1-180) n = 28	39.5± 51.2 (19, 1-240) n = 36	42.3± 56.9 (16, 1-300) n = 63	41.8± 66.5 (23, 1-480) n = 66	26.6± 34.4 (14, 1-204) n = 57
Type of Gland[2] **STN** **MNG/TS**	 20 (64.5) 11 (35.5)	 26 (68.4) 12 (31.6)	 56 (83.6) 11 (16.4)	 49 (27.9) 19 (72.1)	 48 (81.4) 11 (18.6)
Tumor size[1] **(cm)**	4.5±1.6 (4.5, 2-7) n = 27	3.8±1.9 (3, 1.5-10) n = 21	5.3±2.8 (5, 2-15) n = 43	4.7±2.5 (4.5, 1-10) n = 33	3.6±1.5 (3.1, 1-7) n = 28
Type of Surgery[2] **NTT** **STT/HT**	 35 (74.4) 12 (25.6)	 42 (76.4) 13 (23.6)	 50 (71.4) 20 (28.6)	 54 (73.9) 19 (26.1)	 44 (69.8) 19 (30.2)
Histopathology[2] **Papillary** **Follicular**	 37 (78.7) 10 (21.3)	 48 (87.2) 7 (12.8)	 60 (85.7) 10 (14.3)	 57 (78.1) 16 (21.9)	 49 (77.8) 14 (22.2)
Interval between surgery and ^{131}I treatment[1] **(months)**	5.7±7.9 (3, 1-42) n = 29	5.7±7.2 (3, 1-39) n = 37	6.7±13.3 (3, 1-84) n = 66	6.4±14.6 (2, 1-108) n = 68	4.2±5.7 (2, 1-27) n = 58
Post-surgical RAIU[1] **(%)**	9.2±7.6 (9.2,1-26.3) n = 47	10.2±8.7 (6.3, 0.6-28.5) n = 55	8.1±5.6 (6.5,1.2-24) n = 70	8.3±7.8 (5, 0.45-21) n = 61	10.5±7.1 (9, 0.9-27) n = 49
RAD[1] **(Gy)**	93±74 (72, 15-255) n = 47	135±122 (102, 10-600) n = 55	173±120 (139, 29-705) n = 70	174±152 (123, 15-624) n = 61	269±188 (241, 31-875) n = 49

STN-solid thyroid nodule; MNG-multinodular goiter, TS-thyroid swelling; NTT-near total thyroidectomy,
HT-hemi thyroidectomy, STT-subtotal thyroidectomy; RAIU-radioiodine uptake;
RAD- Radiation absorbed dose.
[1]Parenthesis consists of median & range; [2]Parenthesis consists of percentage; n= Number of patients.

Table 2. Distribution of Patients Contd……

Variables	40 mCi	45 mCi	50 mCi	Total	p-value
Age (years)[1]	34.3±14.0 (31, 9-68) n = 60	38.3±13.6 (36.5, 7-65) n = 64	40.3±12.3 (40, 12-75) n = 77	37.5±12.7 (36, 7-75) n = 509	*0.167*
Sex[2] **Male** **Female**	 15 (25.0) 45 (75.0)	 18 (28.1) 46 (71.9)	 23 (29.9) 34 (70.1)	 142 (27.8) 367 (72.2)	*0.544*
Duration of illness[1] **(months)**	37.3±48.8 (24, 2-240) n = 57	44.3±68.9 (12, 1-300) n = 57	37.7±56.7 (18, 1-360) n = 69	37.7±54.6 (18, 1-480) n = 433	*0.578*
Type of Gland[2] **STN** **MNG/TS**	 44 (75.9) 14 (24.1)	 44 (77.2) 13 (22.8)	 48 (65.8) 25 (34.2)	 335 (74.2) 116 (25.8)	*0.183*
Tumor size[1] **(cm)**	4.9±3.3 (4, 2-16) n = 23	4.4±2.1 (4, 2-10) n = 25	4.6±1.8 (5, 2-10) n = 37	4.6±2.4 (4, 1-16) n = 230	*0.064*
Type of Surgery[2] **NTT** **STT/HT**	 41 (68.3) 19 (31.7)	 46 (71.9) 18 (28.1)	 55 (77.4) 22 (28.6)	 367 (72.1) 142 (27.9)	*0.972*
Histopathology[2] **Papillary** **Follicular**	 49 (81.7) 11 (18.3)	 52 (81.3) 12 (18.7)	 58 (75.3) 19 (24.7)	 410 (80.5) 99 (19.5)	*0.067*
Interval between surgery and ^{131}I treatment[1] **(months)**	3.5±6.6 (2, 1-47) n = 58	6.9±17.9 (2, 1-108) n = 58	6.3±12.5 (2, 1-72) n = 72	5.7±11.9 (2, 1-108) n = 446	*0.098*
Post-surgical RAIU[1] **(%)**	10.4±7.9 (8,0.4-26.2) n = 49	8.9±7.1 (7, 1.3-19) n = 64	8.1±6.7 (5.9,1-27) n = 77	9.1±7.2 (6.6, 0.4-28.5) n = 472	*0.485*
RAD[1] **(Gy)**	296±233 (238, 16-1000) n = 49	356±296 (264, 48-1260) n = 64	374±323 (280, 46-1350) n = 77	241±234 (170, 10-1350) n = 472	*0.001*

STN-solid thyroid nodule; MNG-multinomial goiter, TS-thyroid swelling; NTT-near total thyroidectomy, HT-hemi thyroidectomy, STT-subtotal thyroidectomy; RAIU-radioiodine uptake; RAD- Radiation absorbed dose.

[1] Parenthesis consists of median & range; [2]Parenthesis consists of percentage; n= Number of patients.

Subgroup 2: The number of patients for whom any or sub-group of covariates was missing and imputed values were worked out for the analysis, and

Subgroup 3: The complete group of patients that includes subgroup 1 and subgroup 2.

6. Results obtained after imputation analysis:

Characteristics of Subgroups of Data:

Table 3 (a and b) present the descriptive statistics for each group of patients i.e., subgroup-1, subgroup-2, and subgroup-3. It may be observed that, the mean values of the different covariates among the subgroups 3 (509 patients) were almost similar to the mean values observed for subgroup-1 (206 patients). Statistical test also revealed no significant difference among mean values of all the quantitative (tumor size, post surgical radioidine uptake, duration of illness, interval between surgery and therapy, radiation absorbed dose), and qualitative (type of gland) covariates between subgroup-1 and subgroup-3 ($p>0.05$).

There is no significant difference in mean tumor size between patients of subgroup 1 and subgroup 3 (4.57± 2.35 vs. 4.65± 2.31; $p=0.670$). Mean post surgical radioiodine uptake is almost similar in these two subgroups (8.92± 7.23 vs. 8.96± 7.06; $p=0.935$). Similarly, no significant difference is observed among the mean values of interval between surgery and therapy (5.72± 11.94 vs. 5.69± 11.18; $p=0.964$), and radiation absorbed dose to the tumor (241.3± 234.2 vs. 241.7± 222.1; $p=0.647$) between subgroup 1 and subgroup 3, respectively. In addition, it was also observed that there is no significant difference among the values of different categories of type of gland in these two subgroups.

This above analysis was carried to assess the significant difference (if any), between the covariates values obtained before and after imputation analysis. The above analysis revealed that none of the covariates values differs significantly, which indirectly ensures the validity of the imputation methods used under this study.

Table 3 (a). Comparison of mean levels of different covariates between Subgroup 1 and Subgroup 3

Variable	*(Subgroup- 1)*			*(Subgroup- 2)*			*(Subgroup- 3)*			*p-value**
	N	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	
Tumor size	230	4.57	2.35	279	4.72	2.28	509	4.65	2.31	*0.670*
Post surgical RAIU	472	8.92	7.23	37	9.44	4.39	509	8.96	7.06	*0.935*
Duration of illness	433	37.39	54.16	76	51.05	20.72	509	39.43	50.82	*0.618*
Interval between Surgery & Therapy	446	5.72	11.94	63	5.71	2.72	509	5.69	11.18	*0.964*
Radiation absorbed Dose (Gy)	472	241.04	234.22	37	250.22	221.42	509	241.71	222.13	*0.647*

* p-value (comparison subgroup 1 vs. subgroup 3)

Table 3 (b). Association between type of gland and various subgroups

Variable	*Subgroup- 1*	*Subgroup- 2*	*Subgroup- 3*	*p-value**
Type of Gland				
Solid thyroid nodule	335 (74.2)	30 (56.9)	365 (72.3)	0.489
Multinodular goiter	116 (25.8)	25 (43.1)	141 (27.7)	
Total	451	58	509	

**p-value (comparison subgroup 1 vs. subgroup 3); Number in parenthesis consists of percentage*

The following sections deal with comparison of the average/proportions of various quantitative and qualitative variables between ablated and not ablated groups. The comparative results (results based on only available cases and available plus imputed cases) have been presented in order to assess the significant difference, if any.

6.1. Univariate analysis

Unpaired 't'/Wilcoxon Rank Sum test was used to assess the average significant difference of various quantitative variables, such as age, duration of

Table 4. Association of Various Quantitative Covariates (among the Available Cases and Cases obtained after imputation) with first dose outcome

Variable	N	Mean	S.D.	p-value
Age (yrs)				
Ablated	395	37.9	12.5	*0.154*
Non Ablated	114	35.9	13.2	
Tumor Size (cm)				
Ablated	191	4.53	2.38	*0.553*
Non Ablated	39	4.78	2.23	
Total	230			
Tumor Size[a]				
Ablated	395	4.44	1.75	*0.006*
Non Ablated	114	5.40	3.55	
Total	509			
Duration of Illness (months)				
Ablated	352	36.08	52.77	*0.247*
Non Ablated	81	44.80	61.72	
Total	433			
Duration of Illness[a]				
Ablated	395	37.02	49.69	*0.021*
Non Ablated	114	47.79	53.91	
Total	509			
Interval between Surgery& Therapy (months)				
Ablated	363	5.47	11.84	*0.342*
Non Ablated	83	6.85	12.38	
Total	446			
Interval between Surgery& Therapy[a]				
Ablated	395	5.49	11.35	*0.375*
Non Ablated	114	6.54	10.56	
Total	509			
Post surgical RAIU (%)				
Ablated	362	8.01	6.62	*0.001*
Non Ablated	110	11.92	8.29	
Total	472			
Post surgical RAIU[a]				
Ablated	395	8.06	6.44	*0.001*
Non Ablated	114	12.07	8.19	
Total	509			
Radiation absorbed dose (Gy)				
Ablated	362	255.06	226.09	*0.504*
Non Ablated	110	237.23	227.13	
Total	472			
Radiation absorbed Dose[a]				
Ablated	395	257.74	224.09	*0.452*
Non Ablated	114	239.66	221.65	
Total	509			

[a] After imputation analysis

Table 5. Association of various Qualitative Covariates With First Dose Outcome (Ablated/Not-Ablated)

Variables	Out come after First Dose of ^{131} I		p-value
	Ablated	Not Ablated	
Sex			
Male	102 (71.8)	40 (28.2)	(χ^2=3.78)
Female	293 (79.8)	74 (20.2)	p= 0.052
Total	395 (77.6)	114 (22.4)	
Histopathology			
Papillary	320 (78.0)	90 (22.0)	(χ^2=0.24)
Follicular	75 (75.8)	24 (24.2)	p= 0.624
Total	395 (77.6)	114 (22.4)	
Type of Surgery			
NTT	305 (83.1)	62 (16.9)	(χ^2=22.92)
STT/HT	90 (63.4)	52 (36.6)	p= 0.0001
Total	395 (77.6)	114 (22.4)	
Type of Gland			
STN	274 (81.8)	61 (18.2)	(χ^2=0.35)
MNG/TS	92 (79.3)	24 (20.7)	p=0.556
Total	366 (81.2)	85 (18.8)	
Type of Gland[a]			
STN	293 (79.6)	75 (20.4)	(χ^2=3.11)
MNG/TS	102 (72.3)	39 (27.7)	p= 0.078
Total	395 (77.6)	114 (22.4)	
First Dose of ^{131}I			
15 mCi	28 (59.6)	19 (40.4)	
20 mCi	35 (63.6)	20 (36.4)	
25 mCi	57 (81.4)	13 (18.6)	($\chi^2_{\text{linear trend}}$
30 mCi	61 (83.6)	12 (16.4)	=9.77
35 mCi	50 (79.4)	13 (20.6)	p= 0.001
40 mCi	47 (78.3)	13 (21.7)	
45 mCi	54 (84.4)	10 (15.6)	
50 mCi	63 (81.8)	14 (18.2)	
Total	395 (77.6)	114 (22.4)	

[a] After imputation; STN: Solitary thyroid nodule, MNG: Multinodular goiter, NTT: Near total thyroidectomy, STT: Subtotal thyroidectomy, HT: Hemithyroidectomy

disease, tumor size, interval between surgery and ^{131}I treatment, post-surgical RAIU and radiation absorbed dose etc. between ablated and non ablated groups. This analysis revealed that only post-surgical RAIU had an effect on the first dose outcome i.e., ablation. Mean post surgical RAIU in ablated and non ablated patients was 8.1±6.4% and 12.1±8.2%, respectively (p=0.0001). There was no statistical difference among the means in the rest of the variables, mean age being 37.9 ± 12.5 *vs.* 35.9 ± 13.2 yr (p = 0.154), mean tumor size 4.53 ± 2.38 *vs.*4.78 ± 2.23 cm (0.553), mean duration of illness (before surgery) 36.08 ± 52.77 *vs.* 44.8 ± 61.72 months (p = 0.247), interval between surgery and ^{131}I treatment 5.5 ± 11.4 *vs.* 6.5 ± 10.6 months (p = 0.342), and radiation-absorbed dose to the thyroid remnant 255 ± 226 *vs.* 238 ± 227 Gy (p = 0.452) in ablated and non-ablated groups, respectively.

Table 4 shows some significant changes in the results obtained after imputation analysis. Tumor size and duration of illness which were not having significant association with first dose outcome turned out to have significant association with first dose outcome i.e., ablation. Of course as per discussion with concerned clinicians these parameters were also having clinical significance/relevance in determining the outcome in prospective patients. Apart from these two covariates relationship of remaining covariates with first dose outcome did not have any significant change.

Chi-square test revealed that only type of surgery and first dose of radioiodine had influence over the remnant ablation; adequate surgery (NTT) having much higher ablation rate 307/67 (83.2%) than inadequate surgery (HT/STT) 88/142 (63.3%); similarly the ablation rates among the doses ranging from 25mCi to 50mCi was significantly different in comparison of ablation with 15mCi and 20 mCi doses (table. 5). Among all the categorical variables there was only one variable i.e., type of gland in which some of the values were missing. After imputing those missing values, no significant changes were observed (when compared with available cases only).

6.2. Univariate Logistic Regression Analysis

Considering each of the various independent covariates individually such as age, sex, histopathology, duration of illness, tumor size, post surgical radioiodine uptake, first dose of radioiodine, type of surgery, type of gland, interval between surgery & therapy and dependent variable as ablated/not ablated, univariate logistic regression analysis was carried out to assess the relationship of these covariates with ablation. As evident from the Table 6, analysis revealed that tumor size, type of surgery, post surgical radioiodine uptake, and first dose of radioiodine were significantly associated with ablation.

This analysis revealed that patients with adequate surgery i.e., NTT has 2.34 times higher chances of ablation as compare to those with inadequate surgery i.e., STT/HT [Unadjusted OR= 2.84; 95% CI: (1.83, 4.40)]. Similarly, the odds of

ablation among the patients presented with small tumor size was significantly higher in comparison to those patients with large tumor size (Unadjusted OR= 0.857; 95% CI (0.79, 0.94)], However, chances of ablation among the patients having post surgical uptake value ≤ 5 % [Unadjusted OR=2.28; 95% CI (1.73, 5.37)] and 5-10% [Unadjusted OR=2.52; 95% CI (1.59, 4.22)] were 2.28 times and 2.52 times higher in comparison to those having uptake value >10%. In addition, odds of ablation among the patients receiving doses >35mCi of radioiodine was 2.74 times higher in comparison to those receiving doses <25mCi [Unadjusted OR=2.74; 95% CI (1.61; 4.68)]. Similarly, chances of ablation was 2.73 times higher when the patients received doses ranging between 25-35mCi in comparison to those receiving doses <25mCi [Unadjusted OR=2.73; 95% CI (1.63; 4.66)].

Therefore, it can be said that ablation among patients who had adequate surgery, smaller tumor size and lower post surgical radioiodine uptake, had higher odds of ablation. First dose of radioiodine was also positively associated ablation. It was found that, if the administered dose was more than or equal to 25 mCi of radioiodine, odds of having ablation was approximately three times higher in comparison to those receiving dose less than 25 mCi of radioiodine. Patients age, duration of illness, type of gland, interval between surgery and therapy, and histopathology were not significantly associated with ablation. However, chances of ablation among females patients was slightly higher in comparison to ablation rate among males [OR= 1.55; 95% CI (0.99, 2.42)]. Similarly, the odds of ablation among the patients having STN was comparatively higher as compare to patients presented with MNG [OR= 1.49; 95% CI (0.65, 2.33)]. Though these covariates were not statistically significantly associated with ablation, nevertheless it may have clinical significance in determining ablation. The following Table 6 presents the results obtained under univariate logistic regression analysis.

Table 6. Effect of various quantitative and qualitative covariates on first dose out come (ablated/not ablated)

Variable	Coefficients (β)	S. E.	Unadjusted Odds ratio (OR)	95% CI of OR
Age	0.012	0.009	1.02	(0.99, 1.03)
Sex				
Female	0.440	0.227	1.55	(0.99, 2.42)
Male			1.00	
Duration of illness	-0.004	0.002	0.99	(0.99, 1.00)
Type of gland				
STN	0.401	0.228	1.49	(0.95, 2.33)
MNG/TS			1.00	
Tumor size	-0.154	0.043	**0.857**	**(0.79, 0.94)**
Histopathology				
Papillary	0.129	0.263	1.14	(0.68, 1.91)
Follicular			1.00	
Type of Surgery				
NTT	1.045	0.223	**2.84**	**(1.83, 4.40)**
STT/HT			1.00	
Post surgical Radioiodine Uptake				
≤ 5	1.115	0.289	**2.28**	**(1.73, 5.37)**
5-10	0.995	0.248	**2.52**	**(1.59, 4.22)**
>10			1.00	
First Dose of ^{131}I				
>35 mCi	1.007	0.273	**2.744**	**(1.61, 4.68)**
25-35 mCi	1.090	0.232	**2.737**	**(1.63, 4.66)**
<25 mCi			1.00	
Interval between surgery & therapy	-0.007	0.009	0.99	(0.98, 1.01)

Reference Category: Not ablated; STN: Solitary thyroid nodule, MNG: Multinodular goiter; NTT: Near total thyroidectomy, STT: Subtotal thyroidectomy, HT: Hemithyroidectomy

6.3. Multiple Logistic Regression Analysis

In addition, multiple logistic regression analysis was carried out to assess the adjusted effects of various quantitative and qualitative covariates on remnant ablation.

i. Multiple Logistic Regression Analysis- (Before imputation of missing values)

Multiple Logistic Regression Analysis (before imputation of missing values) revealed that only post surgical radioiodine uptake was significantly associated with outcome after adjusting other covariates such as age, sex, histopathology, duration of illness, tumor size, post surgical radioiodine uptake, first dose of radioiodine, type of surgery, type of gland, interval between surgery & therapy. Chance of ablation among the patients having uptake values < 5% five was approximately five times higher in comparison to those having uptake values more than 10% [Adjusted OR=4.90; 95% CI: (1.51, 15.88)]. Similarly, odds of ablation among the patients in whom uptake values ranges from 5-10% had approximately four times higher in as compare to patients having uptake values more than 10 % [Adjusted OR= 3.43; 95%CI:(1.52, 7.82)] (See Table 7).

ii. Multiple logistic regression analysis- (After imputation analysis)

When multiple stepwise logistic regression analysis was applied on complete data set, that is the data set obtained after imputation analysis (n = 509) it was observed that tumor size, type of surgery, post surgical RAIU, and first dose of ^{131}I were found to have significant effect on the remnant ablation. It is evident form the Table 8, that tumor size was significantly (inversely) associated with ablation, higher size of the tumor may lead to lowered ablation i.e., chances of ablation among the patients with smaller tumor size was significantly higher in comparison to those patients having larger tumor size [Adjusted OR=0.77; 95% CI (0.69, 0.84)]. Similarly, chances of remnant ablation was 2.6 times more among the patients with adequate surgery (NTT) in comparison to those with inadequate surgery (STT/HT) [Adjusted OR= 2.55; 95% CI: (1.56, 4.18)].

It was also observed that among the patients having uptake values <5% and 5-10%, odds of ablation were approximately 2.53 [Adjusted OR=2.53; 95% CI (1.37, 4.67)] and 2.25 [Adjusted OR=2.25; 95% CI (1.29, 3.92)] times higher in comparison to those having uptake values >10%. Similarly, among the patients receiving doses >35mCi and 25-35mCi, respective chances of ablation were approximately 3.2 [Adjusted OR=3.19.; 95% CI (1.77, 5.35)] and 3.3 [Adjusted OR=3.25; 95% CI (1.81, 5.82)] times higher in comparison to those receiving <25mCi of radioiodine doses. Also, the insignificant result of Hosmer-Lemshow

Table 7. Adjusted effect of various quantitative and qualitative covariates on remnant ablation: a multiple stepwise logistic regression analysis (Before imputing the missing values: n= 206)

-2 Log likelihood= 171.47
Model Chisquare= 16.29 (p=0.001)
Hosmer-Lemshow goodness of fit-χ2= 5.06 (p=0.75)

Variable	Estimated Coefficients (β)	S. E.	Adjusted Odds ratio (OR)	95% CI of OR
Post surgical Radioiodine Uptake				
≤ 5	-0.040	0.600	**4.90**	(1.51, 15.88)
5-10	1.234	0.420	**3.43**	(1.52, 7.82)
>10			1.00	
Interval b/w surgery & therapy	-0.040	0.021	0.961	(0.92,1.01)
Constant	0.942	0.313		

Reference category- Not ablated

Table 8. Adjusted effect of various quantitative and qualitative covariates on remnant ablation: a multiple stepwise logistic regression analysis (After imputing the missing values n = 509)

-2 Log likelihood= 462.23
Model Chi-square= 79.22 (p=0.0001)
Hosmer-Lemshow goodness of fit-χ^2= 6.14 (p=0.63)

Variable	Estimated Coefficients (β)	S. E.	Adjusted Odds ratio (OR)	95% CI of OR
Type of Surgery				
Adequate	0.936	0.253	**2.55**	(1.56, 4.18)
Inadequate			1.00	
Post surgical RAIU (%)				
≤ 5	0.928	0..313	**2.53**	(1.37, 4.67)
5-10	0.811	0..283	**2.25**	(1.29, 3.92)
>10			1.00	
First Dose of ^{131}I				
>35 mCi	1.162	0.300	**3.19**	(1.77, 5.35)
25-35 mCi	1.179	0.298	**3.25**	(1.81, 5.82)
<25 mCi			1.00	
Tumor Size (cm)	-0.260	0.052	**0.771**	(0.69, 0.84)
Age	0.035	0.001	**1.03**	(1.01, 1.05)

Reference Category: Not ablated.

test of goodness of fit (p=0.63) shows that this model for ablation is satisfactorily describing the data.

It was found that this model differs in comparison with the models obtained in Table 7 & 8 in terms of significant covariates present in the model. It was also observed that there was significant changes in the likelihood values as well model chi-square and Hosmer- Lemshow goodness of fit test statistic. Therefore, imputing the missing values in relation to various important covariates was an essential task that will ultimately add merits in drawing any definite conclusions from the respective developed models. .

7. Discussion, Summary and Conclusions:

The use of imputations has become more common during last years. It is due to the increase in item/unit non-response rates, and to the advanced methodology for imputations. Appropriate analysis of data including development of prognostic model play a crucial role in the clinical decision making process. Unfortunately, missing covariate data impede the construction of valid and reliable models, potentially introducing bias, if handled inappropriately. The various approaches used for handling the different types of missing covariate data under the various clinical trials studies have been described (e.g., Rubin, 1976; Little & Rubin, 1987; Ibrahim, 1990; Little & Rubin, 2002; Little, 1992; Burton & Altman, 2004).

Little (1992) presented an extensive review of regression analysis with missing values of the independent variables. They have concluded and suggested that more widely distributed software is needed that advances beyond complete case analysis, available case analysis, and naïve imputation methods. A study by Schemper & Heinze (1997) demonstrated the utility of imputation technique for missing covariate values before carrying out logistic regression or Cox regression model. In 1990, Schemper and Smith recommended a conditional probability imputation technique (PIT) for the analysis of treatment studies which can be easily applied using standard software and which has been demonstrated to outperform the complete case and omission of covariates strategies.

A recent review (Burton & Altman, 2004) was conducted based on 100 articles reporting multivariate survival analyses to assess potential prognostic factors, published within seven cancer journals in 2002. This review has highlighted deficiencies in the reporting of missing covariate data. Guidelines for presenting prognostic studies with missing covariate data are proposed, which if followed should clarify and standardise the reporting in future articles. Musil et al (2002) studied the theoretical and empirical information for the selection and application of approaches for handling missing data on a single variable. They did compare and contrast five approaches (listwise deletion, mean substitution, simple regression, regression with an error term, and the expectation maximization [EM] algorithm) for dealing with missing data, and compare the effects of each method

on descriptive statistics and correlation coefficients for the imputed data (n = 96) and the entire sample (n = 492) when imputed data are included. All methods had limitations, however findings suggest that mean substitution was the least effective and regression with an error term produced estimates closest to those of the original variables.

As, there is no study available on missing data analysis in the area of remnant ablation of differentiated thyroid cancer (DTC) patients, comparing the results of present study with other studies is not possible. However, the present study has undoubtedly indicated the usefulness of imputation method in clinical trial regarding remnant ablation of differentiated thyroid cancer. Results obtained after imputation analysis gained not only statistical significance but also had clinical relevance.

8. Acknowledgements

Authors gratefully acknowledge Prof. G.K. Rath, Head, Department of Radiotherapy, Prof. K. R. Sundaram, Ex. Head, Department of Biostatistics, Dr. A. K. Karak, Department of Pathology, AIIMS, New Delhi, Dr. Padam Singh, Ex. Additional Director General, ICMR, New Delhi, India, for their constructive suggestions and consistent support throughout the study. The authors would like to thank AIIMS and ICMR, New Delhi, India for providing the financial assistance to carry out this study.

REFERENCES

LITTLE RJA AND RUBIN DB (1987). Statistical Analysis with Missing Data, NY, Wiley.

SCHAFER, JL AND GRAHAM, JW (2002). Missing data: our view of the state of the art, Psychological Methods, 7(2), 147—177.

VACH W (1997). Some issues in estimating the effect of prognostic factors from incomplete covariate data, Stat Med, 16, 57—72.

GREENLAND S, FINKLE WD (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses, Am J Epidemiol, 142, 1255—1264.

RUBIN, D.B. Multiple Imputation for Nonresponse Surveys (1987). John Wiley & Sons, Inc.

SCHAFER J.L. (1997). Analysis of Incomplete Multivariate Data. New York, Chapman & Hall.

VACH M, BLETTNER M (1998). Missing data in epidemiological studies. In Encyclopedia of Biostatistics, Armitage P, Colton T (eds) pp 2641—2654. New York: John Wiley and Sons.

RUBIN, D.B. (1976). Inference and missing data. Biometrika, 63, 581—592.

IBRAHIM J (1990). Incomplete data in generalized linear models, Journal of the American Statistical Association, 85, 765—769.

LITTLE RJA AND RUBIN DB (2002). Statistical Analysis with Missing Data, 2nd edition, New York, John Wiley.

MURRAY GD, FINDLAY JG (1988). Correcting for the bias caused by drop-outs in hypertension trials, Statistics in Medicine, 7, 941—946.

LITTLE RJA (1992). Regression with missing X's: a review. Journal of the American Statistical Association, 87, 1227—1237.

ALTMAN DG (1990). Practical Statistics for Medical Research. CRC Press..

POCOCK, S.J. (1983). Clinical Trials: A Practical Approach. Wiley.

ARMITAGE P, BERRY G. & MATTHEWS JNS (2001). Statistical Methods in Medical Research, 4th Edn, Blackwell Science Ltd.

LAURIKKALA J, KENTALA E, JUHOLA M, PYYKKO I (1999). Treatment of missing values with imputation for the analysis of otologic data, Stud Health Technol Inform., 68, 428—31.

BURTON A, ALTMAN DG (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines: Review, Br J Cancer, 5, 91(1), 4—8.

SCHEMPER M, HEINZE G (1997). Probability imputation revisited for prognostic factor studies, Stat Med., 16(1—3), 73—80.

SCHEMPER M, SMITH TL (1990). Efficient evaluation of treatment effects in the presence of missing covariate values. Stat Med., 9(7), 777—84.

HOSMER, DW AND LEMESHOW S (1980). A goodness-of –fit test for the multiple logistic regression model, Communication in Statistics, A10, 1043—1069.

HOSMER, DW AND LEMESHOW S (1989). Applied Logistic Regression. John Wiley and Sons, Inc.

STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 273—292

EFFECT OF NON-RESPONSE ON CURRENT OCCASION IN SEARCH OF GOOD ROTATION PATTERNS ON SUCCESSIVE OCCASIONS

G. N. Singh and Kumari Priyanka[1]

ABSTRACT

The present work is an attempt to study the effect of non-response at current occasion in search of good rotation patterns on successive occasions. Two difference type estimators have been proposed for estimating the population mean at current occasion in presence of non-response in two occasions successive (rotation) sampling. Detailed behavior of proposed estimators has been studied. Proposed estimators have been compared with the estimators for the same situations but in the absence of non-response. Empirical studies have been carried out to demonstrate the performance of the proposed estimators.

Keywords: Non-response, successive sampling, difference type estimator, bias, variance, optimum replacement policy.

1. Introduction

If the value of study character of a finite population is subject to change (dynamic) over time, a survey carried out on a single occasion will provide information about the characteristics of the surveyed population for the given occasion only and can not give any information on the nature or the rate of change of the characteristic over different occasions and the average value of the characteristic over all occasions or most recent occasion. To meet these requirements, rotation (successive) sampling provides a strong tool for generating the reliable estimates at different occasions.

Theory of rotation (successive) sampling appears to have started with the work of Jessen (1942). He was pioneer to utilize the entire information collected in the previous investigations (occasions). Further the theory of rotation

[1] Department of Applied Mathematics, Indian School of Mines University, Dhanbad-826 004, India. E. mail: gnsingh_ism@yahoo.com and priyanka_ismd@yahoo.co.in

(successive) sampling was extended by Patterson (1950), Rao and Graham (1964), Gupta (1979), Das (1982), Chaturvedi and Tripathi (1983) and many others. Sen (1971) developed estimators for the population mean on the current occasion using information on two auxiliary variables available on previous occasion. Further, Sen (1972, 1973) extended his work for p auxiliary variates. Singh et al. (1991) and Singh and Singh (2001) used the auxiliary information on current occasion for estimating the current population mean in two occasions successive sampling. Singh (2003) extended the work of Singh and Singh (2001) for h-occasions successive sampling. Feng and Zou (1997) and Biradar and Singh (2001) used the auxiliary information on both the occasions for estimating the current mean in successive sampling.

In many situations, information on an auxiliary variate may be readily available on the first as well as on the second occasions, for example, tonnage (or seat capacity) of each vehicle or ship is known in survey sampling of transportation, number of beds in different hospitals may be known in hospital surveys, number of polluting industries are known in environmental survey, nature of employment status, educational status, food availability & medical aids of a locality are well known in advance for estimating the various demographic parameters in demographic surveys. Many other situations in biological (life) sciences could be explored to show the benefits of the present study. Utilizing auxiliary information on both the occasions Singh (2005), Singh and Priyanka (2006 a) and Singh and Priyanka (2006 b) have proposed chain type ratio, difference and regression type estimators for estimating the population mean at current (second) occasion in two occasions successive sampling.

It is common experience in sample surveys that data cannot always be collected from all the units selected in the sample. For example, the selected families may not be at home at the first attempt and some may refuse to co-operate with the interviewer even if contacted. This is particularly true in mail surveys in which questionnaires are mailed to the sampled respondents who are requested to send back their returns by some deadline. As many respondents do not reply, available sample of returns is incomplete. The resulting incompleteness is called non-response and is sometimes so large that can completely vitiate the results.

Hansen and Hurwitz (1946) suggested a technique of handling non-response in mail surveys. These surveys have the advantage that the data can be collected relatively inexpensively. However, non-response is a common problem with mail surveys. Cochran (1977) and Fabian and Hyunshik (2000) extended the Hansen and Hurwitz technique to the case when besides the information on character under study, information is also available on auxiliary character. More recently Choudhary et al. (2004) used the Hansen and Hurwitz technique for estimation of population mean on current occasion in the context of sampling on two occasions.

The objective of the present work is to study the effect of non-response at current occasion in two occasions successive (rotation) sampling. In two

occasions successive sampling, a portion of sample is matched from the previous occasion and it is assumed that whole units respond at first occasion. So, we may think that as they are familiar with the questionnaire at first occasion, therefore, they may not have any hesitation in responding at the second occasion for the units in the matched portion of the sample. At the current occasion a sample is drawn afresh from the remaining units, so there may be possibility of non-response at current occasion. Motivated with the above points and using Hansen and Hurwitz (1946) technique, estimators are proposed to study the effect of non-response at current occasion in two occasions successive (rotation) sampling. In this work two different types of difference estimators T and Δ have been proposed for estimating the population mean of study character at current (second) occasion in two occasions rotation (successive) sampling. The estimator T has been proposed without utilizing any additional auxiliary information while the estimator Δ has been constructed with the utilization of all types of auxiliary information (including additional auxiliary information) for estimating the population mean of study character at the current occasion. The proposed estimators are compared with the estimator defined for the same situation but without non-response. Results have been shown in various tables and suitable recommendations are made.

2. Notations

Let $U = (U_1, U_2, - - -, U_N)$ be the finite population of N units, which has been sampled over two occasions. The character under study be denoted by x (y) on the first (second) occasions respectively. It is assumed that information on an auxiliary variable z (with known population mean), is available on both the occasions. A simple random sample (without replacement) of n units is taken on the first occasion. A random sub sample of $m = n\lambda$ units is retained (matched) for use on the second occasion. We assume that there is non- response at the current occasion, so that the population can be divided into two classes, those who will respond at the first attempt and those who will not. Let the sizes of these two classes be N_1 and N_2 respectively. Now, at the current occasion a simple random sample (without replacement) of $u = (n-m) = n\mu$ units is drawn afresh from the remaining (N-n) units of the population so that the sample size on the second occasion is also n. λ and μ ($\lambda + \mu = 1$) are the fractions of matched and fresh samples respectively at the second (current) occasion. We assume that in the unmatched portion of the sample on the two occasions u_1 units respond and u_2 units do not respond. Let u_{2h} denote the size of sub sample drawn from the non-response class in the unmatched portion of the sample on the current occasion. The following notations are considered for the further use:

$\bar{X}, \bar{Y}, \bar{Z}$: The population mean of the variables x, y and z respectively.

$\bar{x}_n, \bar{z}_n, \bar{y}_m, \bar{x}_m, \bar{z}_m, \bar{y}_{u1}, \bar{y}_{u2h}, \bar{z}_u$: The sample means of the respective variates of the sample sizes shown in suffices.

$\rho_{yx}, \rho_{xz}, \rho_{yz}$: The correlation coefficients between the variables shown in suffices.

S_x^2, S_y^2, S_z^2: The population mean squares of x, y and z respectively.

$W\left(=\frac{N_2}{N}\right)$: The proportion of non- responding units in the population

3. Formulation of Estimator T

To estimate the population mean $\bar{Y}$ on the second occasion, two independent estimators are suggested without utilizing the information on auxiliary character z. One is a Hansen and Hurwitz (1946) estimator based on sample of size u (= nμ) drawn afresh on the second occasion is given by

$$T_1 = \bar{y}_u^* = \frac{u_1\bar{y}_{u1} + u_2\bar{y}_{u2h}}{u} \tag{1}$$

Second estimator is difference type estimator based on the sample of size m (= nλ) common with both the occasions and is defined as

$$T_2 = \bar{y}_m + \beta_{yx}\left(\bar{x}_n - \bar{x}_m\right) \tag{2}$$

where β_{yx} is the known population regression coefficient between the variates shown in suffix.

Considering a convex linear combination of the estimators T_1 and T_2, we have the final estimator of $\bar{Y}$ as

$$T = \varphi T_1 + \left(1 - \varphi\right)T_2 \tag{3}$$

where φ is an unknown constant to be determined under certain criterion.

Remark 3.1: It is obvious that for estimating the mean on each occasion, the estimator T_1 is suitable, which implies that more belief on T_1 could be shown by choosing φ as 1, while for estimating the change from one occasion to the next, the estimator T_2 could be more useful by choosing φ as 0. For asserting both the problems simultaneously, the suitable (optimum) choice of φ is required.

4. Bias and Variance of T

Theorem 4.1: T is an unbiased estimator of $\bar{Y}$.

Proof: Since, $T_1 = \bar{y}_u^*$

$$E(T_1) = E(\bar{y}_u^*) = E\left[E(\bar{y}_u^*)|u_1, u_2\right] = E\left[E\left(\frac{u_1\bar{y}_{u1} + u_2\bar{y}_{u2h}}{u}\right)|u_1, u_2\right]$$

$$= \frac{1}{u} E\left[u_1 E(\bar{y}_{u1}|u_1, u_2) + u_2 E(\bar{y}_{u2h}|u_1, u_2)\right] = \frac{1}{u} E\left[u_1\bar{y}_{u1} + u_2\bar{y}_{u2}\right]$$

$$= E\left[\bar{y}_u\right] = \bar{Y},$$

therefore T_1 is unbiased for $\bar{Y}$. T_2 being a difference type estimator is also unbiased for $\bar{Y}$. The final estimator T is a convex linear combination of T_1 and T_2, therefore, T is also an unbiased estimator of $\bar{Y}$.

Theorem 4.2: Variance of T is obtained as

$$V(T) = \varphi^2 V(T_1) + (1-\varphi)^2 V(T_2) \tag{4}$$

where $$V(T_1) = \left[\left(\frac{1}{u} - \frac{1}{N}\right) + \frac{(f-1)N_2}{Nu}\right]S_y^2 \tag{5}$$

and $$V(T_2) = \left[\left(\frac{1}{m} - \frac{1}{n}\right)(1-\rho_{yx}^2) + \left(\frac{1}{n} - \frac{1}{N}\right)\right]S_y^2 \tag{6}$$

where $f = \dfrac{u_2}{u_{2h}}$.

Proof: Since, the samples are independent the variance of T (ignoring the covariance term) is given by

$$V(T) = E(T - \bar{Y})^2 = \varphi^2 V(T_1) + (1-\varphi)^2 V(T_2) \tag{7}$$

$$V(T_1) = V(\bar{y}_u^*) = V\left[E(\bar{y}_u^*|u_1, u_2)\right] + E\left[V(\bar{y}_u^*|u_1, u_2)\right]$$

$$= V\left(\bar{y}_u\right) + E\left[\frac{u_2}{u^2}(f-1)S_{y2}^2\right] = \left(\frac{1}{u} - \frac{1}{N}\right)S_y^2 + \frac{(f-1)}{u}S_{y2}^2\frac{N_2}{N}$$

where S_{y2}^2 is the population mean square of non response class at current occasion. Further we assume $S_y^2 = S_{y2}^2$, and hence the variance of T_1 is obtained as

$$V\left(T_1\right) = \left[\left(\frac{1}{u} - \frac{1}{N}\right) + \frac{(f-1)N_2}{Nu}\right]S_y^2 \tag{8}$$

Similarly, $V\left(T_2\right) = E\left(T_2 - \bar{Y}\right)^2 = E\left[\bar{y}_m + \beta_{yx}\left(\bar{x}_n - \bar{x}_m\right) - \bar{Y}\right]^2$

$$= \left[\left(\frac{1}{m} - \frac{1}{n}\right)\left(1 - \rho_{yx}^2\right) + \left(\frac{1}{n} - \frac{1}{N}\right)\right]S_y^2 \tag{9}$$

Using equations (8) and (9) in equation (7) we get the expression for V (T).

5. Minimum Variance of T

Since, variance of T in equation (4) is a function of unknown constant φ, therefore, it is minimized with respect to φ and subsequently the optimum value of φ is obtained as

$$\varphi_{opt.} = \frac{V\left(T_2\right)}{V\left(T_1\right) + V\left(T_2\right)} \tag{10}$$

and substituting the value of $\varphi_{opt.}$ in equation (9), we get the optimum variance of T as

$$V\left(T\right)_{opt.} = \frac{V\left(T_1\right).V\left(T_2\right)}{V(T_1)+V(T_2)} \tag{11}$$

Further substituting the values from equations (5) and (6) in equation (11), the simplified value of $V(T)_{opt.}$ are shown in theorem (5.1).

Theorem 5.1: The $V(T)_{opt.}$ is derived as

$$V\left(T\right)_{opt.} = \frac{\left[A_1\mu^2 + A_2\mu + A_3\right]}{[A_4\mu^2 + A_5\mu + A]} \tag{12}$$

where $A_1 = \left(\frac{\rho_{yx}^2}{N} - \frac{n}{N^2}\right)$, $A_2 = \left(\frac{n}{N^2} - \frac{A}{n}\rho_{yx}^2 + \frac{A-1}{N}\right)$, $A_3 = A\left(\frac{1}{n} - \frac{1}{N}\right)$, $A_4 = \left(\frac{2n}{N} - \rho_{yx}^2\right)$, $A_5 = \left(1 - A - \frac{2n}{N}\right)$, $A = 1+(f-1)W$, $W = \frac{N_2}{N}$ and $\mu\left(= \frac{u}{n}\right)$ is the fraction of fresh sample taken at second (current) occasion.

Again V $(T)_{opt.}$ derived in equation (12) is the function of μ. To estimate the population mean on each occasion the better choice of μ is 1 (case of no matching) while for estimating the change from one occasion to the other, μ should be 0 (case of complete matching). But to devise the amicable strategy for both the problems simultaneously, intuition suggests that an optimum choice of μ is desired. Hence, in the next section of this problem, an optimum replacement policy for μ has been discussed.

6. Optimum Replacement Policy

To determine the optimum value of μ (fraction of sample to be taken afresh at second occasion) so that $\bar{Y}$ may be estimated with maximum precision, we minimize V $(T)_{opt.}$ in (12) with respect to μ and hence we get the optimum value of μ as

$$\hat{\mu} = \frac{-B_2 \pm \sqrt{B_2^2 - 4B_1B_3}}{2B_1} = \mu_0 \text{ (say)} \tag{13}$$

where $B_1 = A_1A_5 - A_2A_4$, $B_2 = 2AA_1 - 2A_3A_4$, $B_3 = AA_2 - A_3A_5$

The real value of $\hat{\mu}$ exists if $B_2^2 - 4B_1B_3 \geq 0$. For certain situation, there might be two values of $\hat{\mu}$ satisfying the above condition, hence to choose a value of $\hat{\mu}$, it should be remembered that $0 \leq \hat{\mu} \leq 1$. All other values of $\hat{\mu}$ are inadmissible. In case if both the values of $\hat{\mu}$ are admissible, we choose the minimum of these two as μ_0. Substituting the value of $\hat{\mu}$ from equation (13) in (12) we have

$$V(T)_{opt.^*} = \frac{\left[A_1\mu_0^2 + A_2\mu_0 + A_3\right]}{[A_4\mu_0^2 + A_5\mu_0 + A]} S_y^2 \tag{14}$$

where $V(T)_{opt.^*}$ is the optimum value of T with respect to both φ and μ.

7. Efficiency Comparison

The percent relative loss in efficiency of T with respect to the estimator under same circumstance but in absence of non-response have been obtained to judge the effect of non-response in successive sampling. Let $T^* = \psi T_1^* + (1 - \psi) T_2$ be the estimator for estimating $\bar{Y}$ where $T_1^* = \bar{y}_u$, T_2 is defined in equation (2) and ψ is an unknown constant to be determined under certain criterion. Since, T^* is an unbiased estimator of $\bar{Y}$ and based on two independent samples, hence covariance term vanishes, therefore, following on the line of Sukhatme et al. (1984) the optimum variance of T^* is given by

$$V(T^*)_{opt.^*} = \frac{\left[C_1\mu_1^2 + C_2\mu_1 + C_3\right]}{[C_4\mu_1^2 + C_5\mu_1 + 1]} S_y^2 \tag{15}$$

where $\mu_1 = \dfrac{-D_2 \pm \sqrt{D_2^2 - 4D_1D_3}}{2D_1}$, $C_1 = \left(\dfrac{\rho_{yx}^2}{N} - \dfrac{n}{N^2}\right)$, $C_2 = \left(\dfrac{n}{N^2} - \dfrac{\rho_{yx}^2}{n}\right)$,

$C_3 = \left(\dfrac{1}{n} - \dfrac{1}{N}\right)$, $C_4 = \left(\dfrac{2n}{N} - \rho_{yx}^2\right)$, $C_5 = -\dfrac{2n}{N}$,

$D_1 = C_1C_5 - C_2C_4$, $D_2 = 2C_1 - 2C_3C_4$ $D_3 = C_2 - C_3C_5$.

Remark 7.1: μ_1 being the optimum value of μ, so for certain situation there may be two values of μ_1 so while choosing μ_1 we must remember $0 \le \mu_1 \le 1$. However, if both the values of μ_1 are admissible, we choose the minimum of the two values as μ_1.

The percentage loss in precision of T^* with respect to T both at optimality condition is given by

$$L = \frac{V(T)_{opt.^*} - V(T^*)_{opt.^*}}{V(T)_{opt.^*}} \times 100 \tag{16}$$

Table 1. The percentage loss in precision of T over T^* under optimal conditions for different choices of W, ρ_{yx}, f, n and N. **(N = 3000, n = 300)**

f →		1.5			2.0			2.5		
W ↓	ρ_{yx} ↓	μ_0	μ_1	L	μ_0	μ_1	L	μ_0	μ_1	L
0.2	0.6	0.3318	0.5556	2.76	0.1265	0.5556	1.90	*	-	-
	0.7	0.4691	0.5834	3.79	0.3624	0.5834	5.31	0.2619	0.5834	5.31
	0.8	0.5607	0.6250	4.28	0.5002	0.6250	6.94	0.4430	0.6250	8.48
	0.9	0.6625	0.6964	4.59	0.6304	0.6964	7.94	0.5999	0.6964	10.42
0.4	0.6	0.1265	0.5556	1.90	*	-	-	*	-	-
	0.7	0.3624	0.5834	5.31	0.1669	0.5834	4.20	*	-	-
	0.8	0.5002	0.6250	6.94	0.3885	0.6250	9.21	0.2864	0.6250	8.99
	0.9	0.6304	0.6964	7.94	0.5708	0.6964	12.24	0.5161	0.6964	14.53
0.6	0.6	*	-	-	*	-	-	*	-	-
	0.7	0.2619	0.5834	5.31	*	-	-	*	-	-
	0.8	0.4430	0.6250	8.48	0.2864	0.6250	8.99	0.1470	0.6250	5.99
	0.9	0.5999	0.6964	10.42	0.5161	0.6964	14.53	0.4410	0.6964	15.83
0.8	0.6	*	-	-	*	-	-	*	-	-
	0.7	0.1669	0.5834	4.21	*	-	-	*	-	-
	0.8	0.3885	0.6250	9.21	0.1918	0.6250	7.26	0.0201	0.6250	0.98
	0.9	0.5708	0.6954	12.24	0.4652	0.6964	15.60	0.3724	0.6964	15.62

Note: * indicates μ_0 does not exist.

Table 2. The percentage relative loss in precision for different randomly chosen values of μ_0, μ_1, W, f and ρ_{yx}. **(N = 3000, n = 300)**

		$\mu_0 = 0.1, \mu_1 = 0.1$			$\mu_0 = 0.3, \mu_1 = 0.3$			$\mu_0 = 0.5, \mu_1 = 0.5$		
f →		1.5	2.0	2.5	1.5	2.0	2.5	1.5	2.0	2.5
W ↓	ρ_{yx} ↓	L	L	L	L	L	L	L	L	L
0.2	0.5	0.83	1.52	2.10	2.56	4.69	6.48	4.43	8.08	11.15
	0.6	0.82	1.50	2.07	2.49	4.55	6.29	4.26	7.78	10.74
	0.7	0.81	1.48	2.04	2.40	4.39	6.07	4.06	7.41	10.22
	0.8	0.79	1.45	2.02	2.30	4.20	5.81	3.79	6.93	9.56
	0.9	0.78	1.43	1.98	2.18	3.98	5.50	3.47	6.34	8.75
0.4	0.5	1.52	2.60	3.41	2.56	4.69	6.48	8.08	13.77	17.99
	0.6	1.50	2.57	3.37	2.49	4.55	6.29	7.78	13.26	17.33
	0.7	1.48	2.53	3.32	2.40	4.39	6.07	7.41	12.62	16.49
	0.8	1.45	2.49	3.27	2.30	4.20	5.81	6.93	11.81	15.43
	0.9	1.43	2.45	3.21	2.18	3.98	5.50	6.34	10.80	14.11
0.6	0.5	2.10	3.41	4.30	8.69	6.47	10.48	11.15	17.99	22.61
	0.6	2.07	3.37	4.25	8.51	6.29	10.18	10.74	17.33	21.79
	0.7	2.4	3.32	4.19	8.29	6.07	9.82	10.22	16.49	20.73
	0.8	2.01	3.27	4.13	8.04	5.80	9.39	9.56	15.43	19.39
	0.9	1.98	3.21	4.05	7.75	5.50	8.89	8.74	14.11	17.73
0.8	0.5	2.60	4.04	4.95	8.01	12.93	15.16	13.77	21.25	25.94
	0.6	2.57	3.99	4.89	7.77	12.04	14.73	13.26	20.46	24.98
	0.7	2.54	3.93	4.82	7.50	11.62	14.21	12.62	19.47	23.77
	0.8	2.49	3.87	4.75	7.18	11.11	13.59	11.81	18.23	22.25
	0.9	2.45	3.80	4.66	6.80	10.52	12.88	10.80	16.66	20.35

8. Discussion

A close perusal of tables 1 and 2 reveals that for all cases, loss in precision is there. The percent relative loss (L) increases as ρ_{yx} increases. Hence, if non-response is considered at current occasion in successive sampling, there is loss in precision. The loss in not much appreciable but certain amount of loss is always there when non-response is taken into account. From table 2 it is observed that relative loss in precision increases with increase in the values of f and μ for fixed value of ρ_{yx} and W. However, for fixed values of f and μ and for increasing values of ρ_{yx} and W, values of L decrease.

9. Formulation of Estimator Δ

The estimator T proposed in equation (3) utilizes the information only on character under study through the sample values over two occasions, but there may be many situations when besides the information on character under study, information on auxiliary character may also be readily available on both the occasions. Therefore, with this argument, we assume that information on auxiliary character is readily available on both the occasions and is not changing rapidly over time. Under this supposition we propose an estimator of population mean $\bar{Y}$ considering the problem of non-response only at current occasion. Hence, utilizing the information on auxiliary character z the structure of newly proposed estimator is considered as

$$\Delta = \varphi^* \Delta_1 + \left(1 - \varphi^*\right)\Delta_2 \tag{17}$$

where $\Delta_1 = \bar{y}'_u + \beta_{yz}\left(\bar{Z} - \bar{z}_u\right)$ and $\Delta_2 = \bar{y}^*_m + \beta_{yx}\left(\bar{x}^*_n - \bar{x}^*_m\right)$

where $\bar{y}'_u = \dfrac{u_1\bar{y}_{u1} + \bar{y}_{u2h}}{u}$, $\bar{y}^*_m = \bar{y}_m + \beta_{yz}\left(\bar{Z} - \bar{z}_m\right)$, $\bar{x}^*_n = \bar{x}_n + \beta_{xz}\left(\bar{Z} - \bar{z}_n\right)$, $\bar{x}^*_m = \bar{x}_m + \beta_{xz}(\bar{Z} - \bar{z}_m)$ and φ^* is an unknown constant to be determined under certain criterion. The estimator Δ_1 is based on u units, which are drawn afresh at current occasion such that out of these u units, u_1 units respond and u_2 units do not respond. Let u_{2h} be the size of subsample drawn from the non-response class in the unmatched portion of the sample on current occasion and the estimator Δ_2 is based on m units, which are retained from previous occasion.

10. Bias and Variance of Δ

Theorem 10.1: Δ is an unbiased estimator of $\bar{Y}$.

Proof: Since, Δ_1 and Δ_2 are both difference type estimators and they are unbiased for $\bar{Y}$. The final estimator Δ is a convex linear combination of Δ_1 and Δ_2, therefore, Δ is also an unbiased estimator of $\bar{Y}$.

Theorem 10.2: Variance of Δ is obtained as

$$V(\Delta) = \varphi^{*2} V(\Delta_1) + \left(1 - \varphi^*\right)^2 V(\Delta_2) \tag{18}$$

where $V(\Delta_1) = \left(\frac{A^*}{u} - \frac{B}{N}\right) S_y^2$ (19)

and $V(\Delta_2) = \left[\left(\frac{1}{m} - \frac{1}{N}\right) B + \left(\frac{1}{m} - \frac{1}{n}\right) C\right] S_y^2$ (20)

where $A^* = 1 - \rho_{yz}^2 + f\,W - W$, $B = 1 - \rho_{yz}^2$

and $C = \rho_{yx}^2\left(\rho_{yz}^2 - 1\right) + 2\rho_{yx}\rho_{yz}\left(\rho_{yz} - \rho_{yx}\right)$

Proof: Since, the samples are independent the variance of Δ (ignoring the covariance term) is given by

$$V(\Delta) = E\left(\Delta - \bar{Y}\right)^2 = \varphi^{*2} V(\Delta_1) + \left(1 - \varphi^*\right)^2 V(\Delta_2) \tag{21}$$

$$V(\Delta_1) = V\left[\bar{y}'_u + \beta_{yz}\left(\bar{Z} - \bar{z}_u\right)\right]$$

$$= V\left[E\left(\bar{y}'_u + \beta_{yz}\left(\bar{Z} - \bar{z}_u\right) \middle| u_1, u_2\right)\right] + E\left[V\left(\bar{y}'_u + \beta_{yz}\left(\bar{Z} - \bar{z}_u\right) \middle| u_1, u_2\right)\right]$$

$$= V\left(\bar{y}_u + \beta_{yz}\left(\bar{Z} - \bar{z}\right)\right) + E\left[\frac{u_2}{u^2}(f-1)S_{y2}^2\right] = \left(\frac{1}{u} - \frac{1}{N}\right)\left(1 - \rho_{yz}^2\right)S_y^2 + \frac{(f-1)}{u} S_{y2}^2 \frac{N_2}{N}$$

where S_{y2}^2 is the population mean square of non response class at current occasion. Further we assume $S_y^2 = S_{y2}^2$, and hence the variance of Δ_1 is obtained as

$$V(\Delta_1) = \left[\left(\frac{A^*}{u} - \frac{B}{N}\right)\right] S_y^2 \tag{22}$$

where $A^* = 1 - \rho_{yz}^2 + f\,W - W$ and $B = 1 - \rho_{yz}^2$

Similarly, $V(\Delta_2) = E\left[\Delta_2 - \bar{Y}\right]^2$

$$= E\left[\left(\bar{y}_m^* - \bar{Y}\right)^2 + \beta_{yx}^2\left(\bar{x}_n^* - \bar{x}_m^*\right) + 2\beta_{yx}\left(\bar{y}_m^* - \bar{Y}\right)\left(\bar{x}_n^* - \bar{x}_m^*\right)\right]$$

Taking expectation we get, $V(\Delta_2)$ as

$$V(\Delta_2) = \left[\left(\frac{1}{m} - \frac{1}{N}\right)\left(1 - \rho_{yz}^2\right) + \left(\frac{1}{m} - \frac{1}{n}\right)\times\right.$$

$$\left.\left\{\left(1 - \rho_{xz}\right)^2\rho_{yx}^2 + 2\left(\rho_{xy}\rho_{yz}\rho_{xz} - \rho_{yx}^2\right)\right\}\right]S_y^2 \tag{23}$$

Further, we assume "$\rho_{xz} = \rho_{yz}$", this is an intuitive assumption, which has also been considered by Cochran (1977) and Feng and Zou (1997).

In the light of above assumptions equation (23) takes the following form, which is given as

$$V(\Delta_2) = \left[\frac{1}{m}\left(1 - \rho_{yz}^2\right) + \left(\frac{1}{m} - \frac{1}{n}\right)\left\{\rho_{yx}^2\left(\rho_{yz}^2 - 1\right) + 2\rho_{yz}\rho_{yx}\left(\rho_{yz} - \rho_{yx}\right)\right\}\right]S_y^2$$

$$V(\Delta_2) = \left[\left(\frac{1}{m} - \frac{1}{N}\right)B + \left(\frac{1}{m} - \frac{1}{n}\right)C\right]S_y^2 \tag{24}$$

where $B = 1 - \rho_{yz}^2$ and $C = \rho_{yx}^2\left(\rho_{yz}^2 - 1\right) + 2\rho_{yx}\rho_{yz}\left(\rho_{yz} - \rho_{yx}\right)$.

Now, substituting the values of $V(\Delta_1)$ and $V(\Delta_2)$ from equations (22) and (24) in equation (21), we get the $V(\Delta)$ as in equation (18).

11. Minimum Variance of Δ

Since, variance of Δ in equation (18) is a function of unknown constant φ^*, therefore, it is minimized with respect to φ^* and subsequently the optimum value of φ^* is obtained as

$$\varphi^*_{opt.} = \frac{V(\Delta_2)}{V(\Delta_1) + V(\Delta_2)} \tag{25}$$

and substituting the value of $\varphi^*_{opt.}$ in equation (25), we get the optimum variance of Δ as

$$V(\Delta)_{opt.} = \frac{V(\Delta_1).V(\Delta_2)}{V(\Delta_1)+V(\Delta_2)} \tag{26}$$

Further substituting the values from equations (19) and (20) in equation (26), the simplified value of $V(\Delta)_{opt.}$ is shown in theorem (11.1).

Theorem 11.1: The $V(\Delta)_{opt.}$ is derived as

$$V(\Delta)_{opt.} = \frac{d_1\mu^2 + d_2\mu + d_3}{d_4\mu^2 + d_5\mu + A^*} \tag{27}$$

where $d_1 = -\left(\frac{nB^2}{N^2} + \frac{BC}{N}\right)$, $d_2 = \left(\frac{nB^2}{N^2} + \frac{A^*B}{N} + \frac{A^*C}{n} - \frac{B^2}{N}\right)$, $d_3 = A^*B\left(\frac{1}{n} - \frac{1}{N}\right)$, $d_4 = \left(C + \frac{2Bn}{N}\right)$, $d_5 = \left(W - fW - \frac{2Bn}{N}\right)$ and $\mu\left(= \frac{u}{n}\right)$ is the fraction of fresh sample taken at second (current) occasion.

12. Optimum Replacement Policy

To determine the optimum value of μ (fraction of sample to be taken afresh at second occasion) so that $\bar{Y}$ may be estimated with maximum precision, we minimize $V(\Delta)_{opt.}$ in (27) with respect to μ and hence we get the optimum value of μ as

$$\hat{\mu} = \frac{-a_2 \pm \sqrt{a_2^2 - 4a_1a_3}}{2a_1} = \mu_0^* \text{ (say)} \tag{28}$$

where $a_1 = d_1d_5 - d_2d_4$, $a_2 = 2A^*d_1 - 2d_3d_4$ and $a_3 = A^*d_2 - d_3d_5$

Since, $a_2^2 - 4a_1a_3 \geq 0$, two values of $\hat{\mu}$ are possible, hence to choose a value of $\hat{\mu}$, it should be remembered that $0 \leq \hat{\mu} \leq 1$. All other values of $\hat{\mu}$ are inadmissible. In case if both the values of $\hat{\mu}$ are admissible, we choose the minimum of these two as μ_0^*. Substituting the value of $\hat{\mu}$ from equation (28) in (27) we have

$$V(\Delta)_{opt.} = \frac{d_1\mu_0^{*2} + d_2\mu_0^* + d_3}{d_4\mu_0^{*2} + d_5\mu_0^* + A^*} \tag{29}$$

where $V(\Delta)_{opt.^*}$ is the optimum value of Δ with respect to both φ^* and μ.

13. Efficiency Comparison

The percent relative loss in efficiency of estimator Δ with respect to Singh and Priyanka (2006 b) estimator Δ^* has been observed. The estimator Δ^* is defined under same circumstances as the estimator Δ, but in absence of non response. The structure of the estimator Δ^* is given as

$$\Delta^* = \psi^*\Delta_1^* + \left(1 - \psi^*\right)\Delta_2$$

where $\Delta_1^* = \bar{y}_u + \beta_{yz}\left(\bar{Z} - \bar{z}_u\right)$, Δ_2 is defined in the structure of the estimator Δ and ψ^* is an unknown constant to be determined under certain criterion. Since, Δ^* is an unbiased estimator of $\bar{Y}$, the optimum variance of Δ^* is given by

$$V\left(\Delta^*\right)_{opt.^*} = \frac{\left[t_1\mu_1^{*2} + t_2\mu_1^* + t_3\right]}{[t_4\mu_1^{*2} + t_5\mu_1^* + B]}S_y^2 \tag{30}$$

where $\mu_1^* = \dfrac{-s_2 \pm \sqrt{s_2^2 - 4s_1s_3}}{2s_1}$, $t_1 = -\left(\dfrac{n\,B^2}{N^2} + \dfrac{BC}{N}\right)$, $t_2 = \left(\dfrac{BC}{n} + \dfrac{nB^2}{N^2}\right)$,

$t_3 = B^2\left(\dfrac{1}{n} - \dfrac{1}{N}\right)$, $t_4 = \left(\dfrac{2nB}{N} + C\right)$, $t_5 = -\dfrac{2nB}{N}$,

$s_1 = t_1t_5 - t_2t_4$, $s_2 = 2Bt_1 - 2t_3t_4$ $s_3 = Bt_2 - t_3t_5$.

Remark 13.1: μ_1^* is the optimum value of μ, so for certain situation there may be two values of μ_1^* and while choosing μ_1^* we must remember that $0 \le \mu_1^* \le 1$. However, if both the values of μ_1^* are admissible, we choose the minimum of the two values as μ_1^*.

The percentage loss in precision of Δ with respect to Δ^* under their respective optimality conditions is given by

$$L^* = \frac{V(\Delta)_{opt.^*} - V(\Delta^*)_{opt.^*}}{V(\Delta)_{opt.^*}} \times 100 \qquad (31)$$

Table 3. The percentage loss in precision of Δ over Δ^* for different choices of W, ρ_{yx} .**(N = 3000, n = 300, f = 1.5)**

$\rho_{yx} \rightarrow$		0.5			0.7			0.9		
W ↓	ρ_{yz} ↓	μ_0^*	μ_1^*	L^*	μ_0^*	μ_1^*	L^*	μ_0^*	μ_1^*	L^*
0.2	0.3	0.2085	0.5473	3.52	0.5746	0.6368	4.97	*	-	-
	0.5	*	-	-	0.5631	0.6375	5.84	*	-	-
	0.7	0.9538	0.4970	12.81	0.3664	0.5834	6.95	*	-	-
	0.9	0.5911	0.3808	22.37	0.7401	0.4246	24.14	0.5358	0.6964	16.02
0.4	0.3	*	-	-	0.5163	0.6368	8.56	*	-	-
	0.5	*	-	-	0.4942	0.6375	9.79	*	-	-
	0.7	*	-	-	0.1742	0.5834	9.67	*	-	-
	0.9	0.7695	0.3808	37.82	*	-	-	0.4054	0.6964	21.66
0.6	0.3	*	-	-	0.4613	0.6368	11.19	*	-	-
	0.5	*	-	-	0.4298	0.6375	12.53	*	-	-
	0.7	*	-	-	0.0	0.5834	10.32	*	-	-
	0.9	0.9272	0.3808	49.24	*	-	-	0.2929	0.6964	24.08
0.8	0.3	*	-	-	0.4090	0.6368	13.16	*	-	-
	0.5	*	-	-	0.3691	0.3691	14.47	*	-	-
	0.7	*	-	-	*	-	-	*	-	-
	0.9	*	-	-	*	-	-	0.1925	0.6964	25.17

Note: * indicates μ_0^* does not exist.

Table 4. The percentage loss in precision of Δ over Δ^* for different choices of W, ρ_{yx} .**(N = 3000, n = 300, f = 2.5)**

ρ_{yx} →		0.5			0.7			0.9		
W ↓	ρ_{yz} ↓	μ_0^*	μ_1^*	L^*	μ_0^*	μ_1^*	L^*	μ_0^*	μ_1^*	L^*
0.2	0.3	*	-	-	0.4613	0.6368	11.20	*	-	-
	0.5	*	-	-	0.4298	0.6375	12.54	*	-	-
	0.7	*	-	-	0.2592	0.5834	10.33	*	-	-
	0.9	0.9272	0.3808	49.24	*	-	-	0.2929	0.6964	24.08
0.4	0.3	*	-	-	0.3115	0.6368	15.73	*	-	-
	0.5	*	-	-	0.2566	0.6375	16.79	*	-	-
	0.7	*	-	-	*	-	-	*	-	-
	0.9	*	-	-	*	-	-	0.0162	0.6964	25.78
0.6	0.3	*	-	-	0.1788	0.6368	17.60	*	-	-
	0.5	*	-	-	0.1051	0.6375	18.19	*	-	-
	0.7	*	-	-	*	-	-	*	-	-
	0.9	*	-	-	*	-	-	*	-	-
0.8	0.3	*	-	-	0.0585	0.6368	18.24	*	-	-
	0.5	*	-	-	*	-	-	*	-	-
	0.7	*	-	-	*	-	-	*	-	-
	0.9	*	-	-	*	-	-	*	-	-

Note: * indicates μ_0^* does not exist.

Table 5. The percentage relative loss in precision for different randomly chosen values of μ_0^*, μ_1^*, W, f, ρ_{yz} and ρ_{yx} . **(N = 3000, n = 300)**

		f =1.5, $\mu_0^* = 0.1, \mu_1^* = 0.1$			f =1.5, $\mu_0^* = 0.3, \mu_1^* = 0.3$			f =1.5, $\mu_0^* = 0.5, \mu_1^* = 0.5$		
ρ_{yx} →		0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
W ↓	ρ_{yz} ↓	L^*	L^*	L^*	L^*	L^*	L^*	L^*	L^*	L^*
0.2	0.5	1.07	1.02	0.96	5.75	4.82	2.99	5.72	4.82	2.99
	0.6	1.24	1.18	1.09	6.76	5.68	3.32	6.76	5.68	3.32
	0.7	1.53	1.45	1.33	8.61	7.28	4.08	8.61	7.29	4.08
	0.8	2.11	1.99	1.78	12.43	10.75	6.03	12.43	10.75	6.03
	0.9	3.74	3.48	2.95	23.21	21.04	12.99	23.21	21.04	13.00
0.4	0.5	1.92	1.83	1.71	10.19	8.58	5.33	10.19	8.58	5.37
	0.6	2.18	2.08	1.92	11.84	9.95	5.82	11.85	9.96	5.82
	0.7	2.64	2.50	2.28	14.70	12.44	6.97	14.70	12.45	6.97
	0.8	3.47	3.27	2.92	20.27	17.54	9.83	20.07	17.54	9.83
	0.9	5.54	5.17	4.39	34.22	31.02	19.16	34.22	31.02	19.16
0.6	0.5	2.60	2.48	2.32	13.77	11.61	7.20	13.77	11.61	7.20
	0.6	2.93	2.79	2.58	15.81	13.29	7.77	15.81	13.29	7.77
	0.7	3.46	3.28	2.99	19.24	16.29	9.13	19.24	16.29	9.13
	0.8	4.41	4.15	3.71	25.67	22.21	12.45	25.67	22.21	12.45
	0.9	6.62	6.16	5.23	40.65	36.84	22.76	40.65	36.84	22.76
0.8	0.5	3.16	3.02	2.82	16.71	14.08	8.74	16.71	14.08	8.74
	0.6	3.53	3.36	3.11	18.99	15.96	9.33	18.99	15.96	9.33
	0.7	4.11	3.89	3.55	22.75	19.26	10.79	22.76	19.26	10.80
	0.8	5.10	4.81	4.30	29.61	25.62	14.36	29.62	25.62	14.36
	0.9	7.32	6.82	5.79	44.86	40.66	25.12	44.86	40.67	25.11

14. Interpretation of Results

It may be seen from above tables that for all cases the relative percentage loss in precision is observed wherever the optimum value of μ exists when non-response is taken into account at current occasion. From table 5 it is observed that the loss in precision decreases with the increase in the values of ρ_{yx} whereas increases with the increase in the values of ρ_{yz}. For fixed values of f, μ_0^*, μ_1^*, ρ_{yz} and $\rho_{yx.}$ the values of L^* increases with the increase in the values of W. Similarly, for fixed values W, f, ρ_{yz} and ρ_{yx} , the loss in precision increases with the increase in the values of μ_0^* and μ_1^*, i.e., more the fraction of sample to be drawn at current occasion more loss in precision is observed due to non-response. From the tables it is clear that loss is observed due to the presence of non-response at current occasion, but the structure of estimator is such that the loss is not so appreciable. Hence, in the presence of non-response also the proposed estimator is performing well, so it may be recommended for further use.

Acknowledgement

Authors are grateful to the referees for their valuable and inspiring suggestions.

REFERENCES

BIRADAR, R. S. and SINGH, H. P. (2001): Successive sampling using auxiliary information on both occasions. *Cal. Stat. Assoc. Bull., 51,No. (203-204), 243-251.*

CHATURVEDI, D. K. and TRIPATHI, T. P. (1983): Estimation of population ratio on two occasions using multivariate auxiliary information. *Jour. Ind. Statist. Assoc., 21, 113—120.*

CHOUDHARY, R. *K. (2003):* Use of Hansen and Hurwitz technique in successive sampling. *M. Sc. Thesis submitted to P. G. School, I.A.R.I., New Delhi.*

CHOUDHARY, R. K., BATHAL, H. V. L. and SUD, U.C (2004): On Non-response in Sampling on Two Occasions. *J. Ind. Soc. Agril.Statist., 58(3) 331—343.*

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

DAS, A. K. (1982): Estimation of population ratio on two occasions. *Jour Ind. Soc. Ag Statist., 34, 1—9.*

ECKLER, A. *R.* (1955): Rotation Sampling. *Ann. Math. Statist., 664—685.*

FABIAN, C. O. and HYUNSHIK, L. (2000): Double Sampling for ratio and regression estimation with sub-sampling the non-respondents. *Survey Methodology, 26 (2), 183—188.*

FENG, S. and ZOU, G. (1997): Sample rotation method with auxiliary variable. *Commun. Statist. Theo-Meth., 26, 6, 1497—1509.*

HANSEN, M. H. and HURWITZ, W. N. (1946): The problem of the non-response in sample surveys. *J. Amer. Statist. Assoc., 41, 517—529.*

GUPTA, P. C. (1979): Sampling on two successive occasions. *Jour. Statist. Res., 13, 7—16.*

JESSEN, R. J. (1942): Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Road Bulletin No. 304, Ames, USA, 1—104.*

PATTERSON, H. D. (1950): Sampling on successive occasions with partial replacement of units. *Jour. Royal Statist. Assoc., Scr. B, 12, 241—255.*

RAJ, D. *(1979):* Sampling Theory. *Tata McGraw Hill, New Delhi, 152—162.*

RAO, J. N. K. and GRAHAM, J. E. (1964): Rotation design for sampling on repeated occasions. *Jour. Amer. Statist. Assoc., 59, 492—509.*

SEN, A. R. (1971): Successive sampling with two auxiliary variables. *Sankhya, Ser. B, 33, 371—378.*

SEN, A. R. (1972): Successive sampling with p $(p \geq 1)$ auxiliary variables. *Ann. Math. Statist., 43, 2031—2034.*

SEN, A. R. (1973): Theory and application of sampling on repeated occasions with several auxiliary variables. *Biometrics, 29, 381—385.*

SINGH, V. K., and SINGH, G. N. (1991): Chain-type regression estimators with two auxiliary variables under double sampling scheme. *Merton, 49, 279—289.*

SINGH, G. N. and SINGH, V. K. (2001): On the use of auxiliary information in successive sampling. *Jour. Ind. Soc. Agri. Statist., 54(1), 1—12.*

SINGH, G. N. (2003): Estimation of population mean using auxiliary information on recent occasion in h occasions successive sampling *Statistics in Transition., 6(4), 523—532.*

SINGH, G. N. (2005): On the use of chain-type ratio estimator in successive sampling. *Statistics in Transition, 7(1), 21—26.*

SINGH, G. N. and PRIYANKA, K. (2006 a): On the use of chain-type ratio to difference estimator in successive sampling. *IJAMAS, Vol. 5, No. S06, 41—49*

SINGH, G. N. and PRIYANKA, K. (2006 b): Search of good rotation patterns to improve the precision of the estimates at current occasion. *Submitted to Communications in Statistics (Theory and Methods)*

SUKHATME, P. V., SUKHATME, B. V., SUKHATME, S. and ASHOK, C. (1984): Sampling theory of surveys with applications, *Iowa State University Press, Iowa, USA, 3rd Revised Edition.*

TIKKIWAL, B. D. (1953): Theory of successive sapling. Unpublished Diploma thesis *submitted to ICAR, New Delhi.*

STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 293—300

ON THE BIAS REDUCTION IN LINEAR VARIETY OF ALTERNATIVE TO RATIO-CUM-PRODUCT ESTIMATOR

Rajesh Singh, Pankaj Chauhan and Nirmala Sawan[1]

ABSTRACT

This paper proposes a method to reduce the bias appearing in the alternative to ratio-cum-product estimator for estimating finite population mean in sample surveys. The proposed technique is useful in survey sampling, when one auxiliary character has positive and high correlation with the study variable, whereas another auxiliary character has negative and high correlation with it. An empirical study is carried out to show the properties of the proposed estimator.

Keywords: Auxiliary variable, alternative to ratio-cum-product estimator, bias, mean-squared error (MSE).

1. Introduction

Let U be a finite population consisting of N units. (U_1, U_2,...,U_N), where x is positively correlated with y and z is negatively correlated with y. We wish to estimate the population mean $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ of y, assuming that the population means ($\overline{X}, \overline{Z}$) of (x, z) are known. Assume that a simple random sample without replacement (SRSWOR) of size n is drawn from U.

The usual ratio and product estimator for $\overline{Y}$ are

$$\overline{y}_r = \overline{y}\left(\frac{\overline{X}}{\overline{x}}\right)$$

and

[1] School of Statistics, DAVV, Indore (M.P.), India. E-mail: rsingh.stat@dauniv.ac.in

$$\bar{y}_p = \bar{y}\left(\frac{\bar{z}}{\bar{Z}}\right),$$

respectively, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ are the sample means of y, x and z respectively.

Singh (1967) improved the ratio and product methods of estimation and suggested the ratio-cum-product estimator for $\bar{Y}$ as

$$\bar{y}_{rp} = \bar{y}\left(\frac{\bar{X}}{\bar{x}}\right)\left(\frac{\bar{z}}{\bar{Z}}\right).$$

Using transformation $x_i^* = \frac{\left(N\bar{X} - nx_i\right)}{\left(N-n\right)}$, (i=1,2,...,N) and $z_i^* = \frac{\left(N\bar{Z} - nz_i\right)}{\left(N-n\right)}$, (i=1,2,...,N).

Srivenkataramana (1980) and Bandyopadhyaya (1980) suggested a dual to ratio and product estimator as

$$t_1 = \bar{y}\left(\frac{\bar{x}^*}{\bar{X}}\right) \tag{1.1}$$

$$t_2 = \bar{y}\left(\frac{\bar{Z}}{\bar{z}^*}\right) \tag{1.2}$$

where $\bar{x}^* = \frac{\left(N\bar{X} - n\bar{x}\right)}{\left(N-n\right)}$, $\bar{z}^* = \frac{\left(N\bar{Z} - n\bar{z}\right)}{\left(N-n\right)}$.

Singh et.al.(2005) suggested a dual to the ratio-cum-product estimator as

$$t_3 = \bar{y}\left(\frac{\bar{x}^*}{\bar{X}}\right)\left(\frac{\bar{Z}}{\bar{z}^*}\right) \tag{1.3}$$

Estimator t_1, t_2 and t_3 are biased. In some application bias is disadvantageous. This led authors to suggest almost unbiased estimator of $\bar{Y}$.

2. Linear variety of estimators

Consider $t_0 = \bar{y}$ and t_1, t_2 and t_3 as defined in (1.1)-(1.3), respectively, such that $t_i \in H$, for i=0,1,2,3, where H denotes the set of all possible estimators of population mean $\bar{Y}$. The set H is a linear variety if

$$t = \sum_{i=0}^{3} h_i t_i \in H \tag{2.1}$$

$$\text{for } \sum_{i=0}^{3} h_i = 1 \text{ and } h_i \in R, \tag{2.2}$$

where h_i's (i=0,1,2,3) denote the real constants used for reducing the bias and R stands for the set of real numbers.

3. Mean-squared error (MSE)

To obtain the MSE of t, we write

$$\bar{y} = \bar{Y}(1+e_0),\ \bar{x} = \bar{X}(1+e_1),\ \bar{z} = \bar{Z}(1+e_2).$$

so that

$$E(e_0) = E(e_1) = E(e_2) = 0,$$

$$E(e_0^2) = \left(\frac{1-f}{n}\right)C_y^2,\ E(e_1^2) = \left(\frac{1-f}{n}\right)C_x^2, E(e_2^2) = \left(\frac{1-f}{n}\right)C_z^2,$$

$$E(e_0e_1) = \left(\frac{1-f}{n}\right)\rho_{yx}C_yC_x, E(e_0e_2) = \left(\frac{1-f}{n}\right)\rho_{yz}C_yC_z,$$

$$E(e_1e_2) = \left(\frac{1-f}{n}\right)\rho_{xz}C_xC_z.$$

where $C_y = \frac{S_y}{\bar{Y}}$, $C_x = \frac{S_x}{\bar{X}}$, $C_z = \frac{S_z}{\bar{Z}}$, $f = \frac{n}{N}$.

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

$$t = \bar{Y}(1+e_0)[h_0 + h_1(1-ge_1) + h_2(1-ge_2)^{-1} + h_3(1-ge_1)(1-ge_2)^{-1}] \tag{3.1}$$

where $g = \frac{n}{N-n}$.

Assuming that $|ge_2|<1$, so that $(1+\theta e_1)^x$ is expandable. Thus expanding the right hand side of the above expression (3.1) and retaining terms up to second power of e's, we have

$$t = \bar{Y}(1+e_0)[h_0 + h_1(1-ge_1) + h_2(1+ge_2+g^2e_2^2) + h_3(1-ge_1)(1+ge_2+g^2e_2^2)] \tag{3.2}$$

$$t \cong \overline{Y}[1 + e_0 - (h_1 + h_3)ge_1 + (h_2 + h_3)ge_2]$$

Let us choose

$$h_1 + h_3 = h \text{ (say, a constant)} \tag{3.3}$$

$$h_2 + h_3 = h^* \text{ (say, a constant)} \tag{3.4}$$

Putting values from (3.4), (3.5) into (3.3), we have

$$(t - \overline{Y}) = \overline{Y}[e_0 - hge_1 + h^* ge_2] \tag{3.5}$$

Squaring both sides of (3.6) and then taking expectations, we get the MSE of the estimator t, up to the first order of approximation, as

$$MSE(t) = \left(\frac{1-f}{n}\right)\overline{Y}^2[C_y^2 + h^2 g^2 C_x^2 + h^{*2} g^2 C_z^2 - 2hg\rho_{xy}C_y C_x -$$

$$2h^* hg^2 \rho_{xz} C_x C_z + 2h^* g\rho_{yz} C_y C_z] \tag{3.6}$$

Differentiating (3.7) with respect to h and h^* and equating to zero, we get

$$hgC_x - h^* g\rho_{xz} C_z = \rho_{xy} C_y \tag{3.7}$$

and

$$hg\rho_{xz} C_x - h^* gC_z = \rho_{yz} C_y \tag{3.8}$$

On solving these, we obtain

$$h = \frac{C_y(\rho_{yz}\rho_{xz} - \rho_{xy})}{gC_x(\rho_{xz}^2 - 1)} \tag{3.9}$$

and

$$h^* = \frac{C_y(\rho_{yz} - \rho_{xy}\rho_{xz})}{gC_z\left(\rho_{xz}^2 - 1\right)} \tag{3.10}$$

Substitution of (3.10) and (3.11) in (3.7), yields the minimum MSE of ‘t’ as

$$\min.MSE(t) = \left(\frac{1-f}{n}\right)S_y^2\left[1 - R_{y.xz}^2\right] \tag{3.11}$$

where $R_{0.12}^2 = \dfrac{\rho_{yx}^2 + \rho_{yz}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz}}{\left(1 - \rho_{xz}^2\right)}$

The expression (3.12) indicates that, to the first degree of approximation, the proposed linear variety of alternative to ratio-cum-product estimator attains the minimum MSE equal to that of the usual multiple regression estimator (based on two auxiliary characters).

4. Bias reduction process

From (2.2), (3.4), (3.5), (3.10) and (3.11), we have

$$\sum_{i=0}^{3} h_i = 1 \tag{4.1}$$

$$h_1 + h_2 = h \tag{4.2}$$

$$\text{and } h_2 + h_3 = h^* \tag{4.3}$$

From (4.1), (4.2), and (4.3), we have four unknowns to be determined from only three equations. It is, therefore, not possible to find a unique value for the real constants h_i's $i = (0,1,2,3)$ to be used for bias reduction. To do so, let us define a new linear constraint as a function of biases in the estimators forming the linear variety as

$$\sum_{i=0}^{3} h_i B(t_i) = 0\,, \tag{4.4}$$

where $B(t_i)$ denotes the bias of the ith estimator t_i of population mean. The above system of equations can be rewritten as:

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ B(t_0) & B(t_1) & B(t_2) & B(t_3) \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ h_3 \end{bmatrix} = \begin{bmatrix} h \\ h^* \\ 1 \\ 0 \end{bmatrix} \tag{4.5}$$

It is well known that to the first degree of approximation, the biases of $t_i (i = 1,2,3)$ are given by –

$$B(t_1) = -\left(\frac{1-f}{n}\right)\overline{Y} g \rho_{xy} C_x C_y \tag{4.6}$$

$$B(t_2) = \left(\frac{1-f}{n}\right)\overline{Y} g \left[g^2 C_z^2 + \rho_{xz} C_x C_z\right] \tag{4.7}$$

$$B(t_3) = \left(\frac{1-f}{n}\right)\overline{Y}g\left[gC_z^2 - g\rho_{xz}C_xC_z + g\rho_{yz}C_yC_z - g\rho_{yx}C_yC_z\right] \quad (4.8)$$

Solving (4.5), we have

$$h_0 = \frac{-B(t_1) - B(t_2) + B(t_3)}{B(1)(h^* - 1) + B(2)(h-1) + B(3)(1-h-h^*)} \quad (4.9)$$

$$h_1 = \frac{-B(t_1) - B(t_2) + B(t_3)}{B(t_2)(h^* - h) + h^*B(t_3)} \quad (4.10)$$

$$h_2 = \frac{-B(t_1) - B(t_2) + B(t_3)}{B(t_1)(h - h^*) + h^*B(t_3)} \quad (4.11)$$

$$h_3 = \frac{-B(t_1) - B(t_2) + B(t_3)}{-h^*B(t_2) - hB(t_1)} \quad (4.12)$$

Using (4.2), (4.3), (4.6) - (4.8) in (4.9) - (4.12), we get the values of h_i's $(i = 0,1,2,3)$ which reduces the bias up to the first order of approximation at (2.1).

5. Empirical study

The data for the empirical study is taken from natural population data set considered by Singh (1969).

Population: (Source: Singh, 1969, p.375).

The data for all 61 blocks of Ahmedabad city, Ward No. 1 taken from 1951 population census are considered. The variables y, x and z are the number of females employed, females in services and educated females, respectively. The required parameters are

$\overline{Y}$ = 7.46, $\overline{X}$ = 5.31, $\overline{Z}$ = 179.00, C_y^2 = 0.5046, C_x^2 = 0.5757, C_z^2 = 0.0633,

ρ_{yx} = 0.7737, ρ_{yz} = -0.2070, ρ_{xz} = -0.0033.

N = 61, n = 20.

In table 5.1 the values of scalars h_i's $(i = 0,1,2,3)$ are listed.

Table 5.1. Values of scalars h_i's $(i = 0,1,2,3)$

Scalars	Population I
h_0	0.6332
h_1	0.2918
h_2	-0.4872
h_3	0.5550

Using these values of h_i's $(i = 0,1,2,3)$ given in table 5.1, one can reduce the bias to the order $o(n^{-1})$ respectively, in the estimator t at (2.1).

In table 5.2 percent relative efficiencies (PRE) of $\bar{y}$, t_1, t_2, t_3 and t (min MSE of t) are computed with respect to $\bar{y}$.

Table 5.2. PRE of different estimators of $\bar{Y}$ with respect to $\bar{y}$

Estimator	PRE (., $\bar{y}$)
$\bar{y}$	100
t_1	214.95
t_2	104.34
t_3	235.76
t	278.09

Table 5.2 clearly indicates that the suggested estimator t in its optimum case is better than the usual unbiased estimator $\bar{y}$, Srivenkataramana (1980) and Bandyopadhyaya (1980) estimator's t_1 and t_2 and Singh et.al (2005) estimator t_3.

REFERENCES

BANDYOPADHYAYA, S. (1980): Improved ratio and product estimators, *Sankhya* C 42, 45—49.

SINGH, M.P. (1967): Ratio-cum-product method of estimation. *Metrika* 12, 34—72.

SINGH, M.P. (1969): Comparison of some ratio-cum-product estimators. *Sankhya* B, 31, 375—378.

SINGH, H. P., SINGH, R., Espejo, M.R., Pined, D.M., Nadarajah, S. (2005): On the efficiency of a dual to ratio-cum-product estimator in sample surveys. *Mathematical Proceedings of the Royal Irish Academy,* 105 A(2), 51—56.

SRIVENKATARAMANA, T (1980): A dual to ratio estimator in sample surveys. *Biometrika* 67, 199—204.

STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 301—310

ESTIMATING THE PROPORTION OF PEOPLE BEARING A SENSITIVE ISSUE WITH AN OPTION TO ITEM COUNT LISTS AND RANDOMIZED RESPONSE

Sanghamitra Pal[1]

ABSTRACT

Randomized Response (RR) Techniques (RRT) pioneered by Warner (1965) and being developed quite rapidly ever since are well-known as devices for tackling sensitive issues. Item Response Techniques (IRT) of Miller (1985), further being developed by Chaudhuri and Christofides (2006) and others are emerging as alternatives to RRT's. IRT is popular for its simplicity. So far IRT's need two independently drawn samples for data generation. Here we present a technique that offers each sampled person an option either to follow a specified RR device or to respond according to an IRT as clarified to him/her. But the option exercised need not be divulged to the investigator. A theory is developed for the estimation of the proportion of people bearing a sensitive characteristic in a given community on selecting samples with unequal probabilities. Unbiased estimation for the variance of the estimators is also provided. Theoretical results are accompanied by a simulation study.

Keywords*:* Item count method, randomized response, sensitive character, unbiased estimation.

1. Introduction

A Randomized Response (RR) Technique (RRT) is a data collection procedure on sensitive issues protecting privacy to the respondents. RRT was originally formulated by Warner (1965) and its alternate versions have been suggested by various authors including Abul-Ela, Greenberg and Horvitz (1967); Greenberg, Abul-Ela, Simmons and Horvitz (1969); Chaudhuri and Mukerjee (1988); Kuk (1990); Mangat and Singh (1990); Chaudhuri (2001) among others. Another method of indirect questioning called Item Count Technique or Item

[1] Sanghamitra Pal, Applied Statistics Unit, Indian Statistical Institute, 203, B.T.Road, Kolkata-700108, India, email: mitra_pal@yahoo.com

Response Technique (IRT) was reported by Miller (1985), Droitcour et. al. (1991) among others. But this is not as popular despite its simplicity. Here respondents are given a list of behaviours in which the sensitive behaviour is imbedded in a list of non-sensitive behaviours. Respondents indicate the number of the behaviours that apply to them rather than answering questions on the actual bahaviours. Recently Chaudhuri and Christofides (2006) have given a theoretical foundation of the technique.

In practice we observe that respondent's reaction to the indirect questioning technique is mixed though IRT is somewhat easier to administer than the RR technique. Some sampled persons feel comfortable with RR techniques. But some respondents like IRT. Here we propose an optional method where each sampled person is given the option either to produce an RR or to report the number of items applicable to him/her as described in the Item Count method without divulging the option exercised. In section 2 and 3 we present the Item Count method and our RRT for a general unequal probability sampling design. The section 4 presents the proposed optional method along with a detailed estimation method. A simulation study along with numerical findings is presented in section 5. Finally we present the conclusions in section 6.

2. Item Count method (or, Item Response Technique)

Let $U = (1,...,i,...,N)$ denote a population of a known number of N units.

Let $\theta_i = 1$ if the ith person bears a sensitive characteristic A

=0, he/she bears A^C, the complement of A.

Our problem is to estimate $\theta_A = \frac{\sum_{i=1}^{N} \theta_i}{N}$. In practice often the values of θ_i's, $i \in U$ are non-ascertainable. In that situation we shall estimate the unknown θ_i's by an RR technique.

To estimate the unknown proportion θ_A, we may alternatively employ Item Count Technique. Following Chaudhuri and Christofides (2006), in the item count method two nearly identical lists of behaviours are developed of which G (>1) items or behaviours of both the lists are innocuous and exactly same. The (G+1) st item of the first list stands for either the stigmatizing feature A or any fresh innocuous item (say F) or both of them (i.e $A \cap F$).

In the other list the (G+1) st item stands for either the complement of the stigmatizing question (A^C) or the complement of the fresh item (say (F^C)) or both the complement of the stigmatizing and the complement of the fresh innocuous item (i.e $(A^C \cap F^C)$).

Let $F_i = 1$ if the ith person bears the fresh innocuous item F mentioned above,

= 0, if the ith person does not bear the item F

and $\theta_i F_i = 1$ if the ith person bears both the character A and the fresh innocuous item F as above,

= 0, if the ith person does not bear either of the character A and the fresh innocuous item as above.

Let two independent samples s_1 and s_2 be drawn from U with a common sampling design P. Respondents of the sample s_1 are asked to simply provide the number of behaviours, say $y_i (i = 1,2.,,,.N)$ which apply to them (without indicating the specific behaviours) among the (G+1) items of the first list. Every unit in s_2 is presented with the second list of (G+1) items. The number of behaviours applicable for an individual j of the second sample is x_j (say). Following Chaudhuri and Christofides (2006) and using the Horvitz-Thompson estimation (1952) method , the unknown proportion θ_A may be estimated by $\hat{\theta}_A(1) = \frac{1}{N}\sum_{i\in s_1}\frac{y_i}{\pi_i} - \frac{1}{N}\sum_{j\in s_2}\frac{x_j}{\pi_j} + 1 - \theta_F$ where $\theta_F = \frac{1}{N}\sum_{k=1}^{N} F_k$ (which is known) and π_i, π_j's are the first order inclusion probabilities..

Writing the second order inclusion probabilities for a pair of units i and i' $(i \neq i')$ as $\pi_{ii'}$ and recalling from Chaudhuri and Pal (2002) and using Horvitz-Thompson (1952) method of estimation, an unbiased estimator of $V(\hat{\theta}_A(1))$ is

$$v_P(\hat{\theta}_A(1)) = \sum_{i<i'\in s_1}\sum\left(\frac{\pi_i\pi_{i'} - \pi_{ii'}}{\pi_{ii'}}\right)\left(\frac{y_i}{\pi_i} - \frac{y_{i'}}{\pi_{i''}}\right)^2 + \sum_{i\in s_1}\frac{y_i^{\,2}}{\pi_i^{\,2}}\alpha_i$$

$$+ \sum_{j<j'\in s_2}\sum\left(\frac{\pi_j\pi_{j'} - \pi_{jj'}}{\pi_{jj'}}\right)\left(\frac{x_i}{\pi_j} - \frac{x_j}{\pi_{j'}}\right)^2 + \sum_{j\in s_2}\frac{x_j^{\,2}}{\pi_j^{\,2}}\alpha_j$$

where $\alpha_i = 1 + \frac{1}{\pi_i}\sum_{i'\neq i}\pi_{ii'} - \sum_{i=1}^{N}\pi_i$, $\alpha_j = 1 + \frac{1}{\pi_j}\sum_{j\neq j'}\pi_{jj'} - \sum_{j=1}^{N}\pi_j$. In this case estimated coefficient of variation may be calculated as CV_{est} (IC)= $\frac{v_P(\hat{\theta}_A(1))}{\hat{\theta}_A(1)}$.

3. Randomized response (RR) technique

To estimate the proportion of people bearing a sensitive character several RR techniques are there in the literature. But here to cover the above item count values, a special RR technique has been introduced. Let u be a random variable following a discrete uniform distribution (0, G) i.e. Prob $[u = k] = \frac{1}{(G+1)}, k = 0,1,....,G.$ Then the random variable u takes the values 0,1,2,….,G. As his/her RR response, the ith individual of the sample s_1 will report $I_i = \theta_i + u_i$ where u_i is the value of the random variable u. By E_R and V_R we shall denote the expectation and variance operator with respect to the RR device. Thus $E_R(I_i) = \theta_i + \alpha$ where α is the mean of the random variable u. Clearly $\alpha = G/2$.

We suppose that θ_i is not ascertainable for a person i in a sample. Here an RR may be procured as $r_i = I_i - \alpha$ such that $(i) E_R(r_i) = \theta_i$, (ii) $V_R(r_i) = V_i(RR)(>0)$, (iii) r_i's are independent over i in U and (iv) there exists v_i ascertainable from RR's such that $E_R(v_i(RR)) = V_i(RR) = \frac{G(G+2)}{12}, i \in U$. Writing E_P, V_P as expectation and variance operators in respect of the design P, let us write $E = E_P E_R = E_R E_P$ and $V = V_P E_R + E_P V_R = E_R V_P + V_R E_P$ where E and V are the overall expectation and the variance operators respectively. Here our proposed estimator is $\hat{\theta}_A(2) = \frac{1}{N} \sum_{i \in s_1} \frac{r_i}{\pi_i}$. It may be estimated by

$$v\ (\hat{\theta}_A(2)) = \sum_{i<j \in s_1} \sum (\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}})(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j})^2 + \sum_{i \in s_1} \frac{r_i^2}{\pi_i^2} \alpha_i + \sum_{i \in s_1} \frac{v_i(RR)}{\pi_i}.$$

The term α_i is defined in our Section 2. Estimated coefficient of variation here is CV_{est} (RR)= $\frac{v(\hat{\theta}_A(2))}{\hat{\theta}_A(2)}$.

4. Suggested Optional technique

Here each respondent selected in the sample has either of the following two options:

(i)He/she can provide the Randomized Response (RR)

Or, (ii) the respondent can report the number of items applicable to him/her using the Item Count Technique.

Each individual will give out his/her item count value without disclosing whether he/she is doing so or follow the instructions as described in the RR device. Two independent samples s_1 and s_2 have been drawn for our proposed Optional technique. It is supposed that i th respondent of the sample s_1 chooses to give out the randomized response (RR) I_i with probability $c_i (0 \le c_i \le 1)$ (which is unknown) or the item count value (y_i) with probability (1-c_i), $i \in U$. For his/her RR value ith respondent will report the value $I_i = \theta_i + u_i$ where u_i is the value of the random variable u . Two independent responses may be gathered from each individual.

Let the first response of the ith person of the sample s_1 denoted as z_{1i} be defined as

$z_{1i} = \theta_i + u_i = I_i$ with an unknown positive probability c_i

or = y_i , with the complementary probability (1- c_i), $i = 1,2,...,N$.

A second independent response from each respondent i of the sample s_1 is also requested either to apply the RR device or to apply the item count technique. Denoting the second independent response of a sampled person i as z_{2i} we may write

$z_{2i} = \theta_i + u_i' = I_i'$ with an unknown positive probability c_i

or = y_i , with the complementary probability (1- c_i), $i = 1,2,...,N$ where u_i' is the value of the random variables u with a fresh independent draw.

Writing E_R and V_R as the operators for expectation and variance with respect to any RR device we may write $E_R(z_{1i}) = (\theta_i + \alpha)c_i + (1 - c_i)y_i = E_R(z_{2i})$

Let, $z_i = \dfrac{z_{1i} + z_{2i}}{2}$ then

$$E_R(z_i) = (\theta_i + \alpha)c_i + (1 - c_i)y_i = E_R(z_{1i}) = E_R(z_{2i}).$$

Defining $v_i = \dfrac{(z_{1i} - z_{2i})^2}{4}$, we also may write

$$E_R(v_i) = V_R(z_i) = V_i = V_R(z_{1i}) = V_R(z_{2i}).$$

Each respondent selected in the sample s_2 has either of the following two options:

(i) To make our procedure convenient the respondent numbered i of the sample s_2 is being requested to report the value $1+u_i-F_i$ as he/she chooses RR method

Or, (ii) he/she will report the value x_i (of section 2) as obtained by item count method. A second independent response is also requested from ith respondent of the sample s_2. Denoting their two responses by z_{1i}' and z_{2i}', we may write

$$E_R(z_{1i}') = (1+\alpha-F_i)c_i + (1-c_i)x_i = E_R(z_{2i}') \ .$$

Let, $z_i' = \dfrac{z_{1i}' + z_{2i}'}{2}$, $v_i' = \dfrac{(z_{1i}' - z_{2i}')^2}{4}$ then

$$E_R(z_i') = E_R(z_{1i}') = E_R(z_{2i}')$$

and $E_R(v_i') = V_R(z_i') = V_i' = V_R(z_{1i}') = V_R(z_{2i}')$.

In practical situation G is only 4 or 5. So for both the responses z_{1i} and z_{2i} (or, z_{1i}' and z_{2i}') of a particular individual the RR values may be same.

Define $e_1 = \dfrac{1}{N}\sum_{i\in s}\dfrac{z_i}{\pi_i}$ and $e_2 = \dfrac{1}{N}\sum_{i\in s}\dfrac{z_i'}{\pi_i}$ Now,

$$E(e_1) = E_P E_R(e_1) = \frac{1}{N}\sum_{i=1}^{N}[(\theta_i+\alpha)c_i + (1-c_i)y_i]$$

$$E(e_2) = E_P E_R(e_2) = \frac{1}{N}\sum_{i=1}^{N}[(1+\alpha-F_i)c_i + (1-c_i)x_i].$$

Writing $A^C \cup F^C = (A\cap F)^C$ and using De-Morgan's law for the indicator function , we may write

$$E(e_1) - E(e_2) = \frac{1}{N}[(\theta_i+\alpha)c_i - (1+\alpha-F_i)c_i + (1-c_i)\{\theta_i + F_i - \theta_i F_i - (1-\theta_i F_i)\}]$$

$$= \frac{1}{N}[\sum\theta_i + \sum F_i - N] = \theta_A + \theta_F - 1$$

where $\theta_F = \dfrac{\sum_{i=1}^{N} F_i}{N}$ which is known. Hence our proposed estimator is

$\hat{\theta}_A(3) = e_1 - e_2 - \theta_F + 1$. The variance of the proposed estimator is

$$V(\hat{\theta}_A(3)) = V(e_1) + V(e_2) \quad \text{where}$$
$$V(e_1) = E_R V_P(e_1) + V_R E_P(e_1).$$
$$= E_R V_P(e_1) + \sum V_i$$
$$\text{and} \quad V(e_2) = E_R V_P(e_2) + \sum V_i'$$

Our proposed unbiased estimator of $V(\hat{\theta}_A(3))$ is

$$v(\hat{\theta}_A(3)) = v_P(e_1) + \sum_{i \in s_1} \frac{v_i}{\pi_i} + v_P(e_2) + \sum_{i \in s_1} \frac{v_i'}{\pi_i}$$

Recalling from Chaudhuri and Pal (2002), $v(e_1)$ and $v(e_2)$ may be written as

$$v_P(e_1) = \sum_{i<j \in s_1} \sum \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{z_i}{\pi_i} - \frac{z_j}{\pi_j}\right)^2 + \sum_{i \in s_1} \frac{z_i^2}{\pi_i^2} \alpha_i$$

and $v_P(e_2) = \sum_{i<j \in s_2} \sum \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{z_i'}{\pi_i} - \frac{z_j'}{\pi_j}\right)^2 + \sum_{i \in s_2} \frac{z_i'^2}{\pi_i^2} \alpha_i$

For this ORR situation the estimated CV is CV_{est} (Optional)= $\frac{v(\hat{\theta}_A(3))}{\hat{\theta}_A(3)}$.

5. Simulation study

The two methods, RR and the Item Count of asking indirect questions about sensitive behaviours had different strengths and weaknesses. In this paper we are adopting an ORR technique that combines aspects of both RR and the Item count method.

In order to examine the magnitude of gain in relative efficiency of the proposed ORR procedure over RR and the Item count method, an empirical study was undertaken. Addiction to the cocaine among students was the sensitive topic chosen for our empirical study. Two lists L_1 and L_2 have been prepared for the samples s_1 and s_2 respectively.

The list L_1 consists of the following questions:

1. Is your birthday is in June
2. Do you have any brother and sister
3. Rode a bicycle without helmet.

4. Monthly income of the family to which he/she belongs is greater than 40,000 in Indian rupees.
5. The last or (G+1) st question consists of either used cocaine regularly one or more times (A) or liked coffee very much (F) or both $A \cap F$.

The list L_2 consists of the following questions:

The first four questions 1), 2), 3) and 4) will be same as earlier. The last or (G+1) st question is never used cocaine (A^C) or did not like coffee (F^C) or did not take cocaine and did not like coffee. Each individual of the sample s_1 and s_2 is being presented the lists L_1 and L_2 respectively. Each of them is being asked the question "how many of the things in this list are applicable for you?"

We consider a population of N=109 students from a medical college of the city Calcutta with the known values l_i and the unknown values θ_i for $i = 1,2,....., N$. Here $\theta_i = 1$ if ith person (student) had habit of using cocaine in any form, one or more times and $\theta_i = 0$ else.

$l_i =$ the per capita expenditure in Indian Rupees incurred in the household to which i belongs –this is the size measure used in choosing a sample of the medical students. Here G is taken as 4. In this survey the respondents are asked in a face-to-face interview to "report the number of items you feel appropriate" among the five items.

5.1. Sampling Scheme

The problem addressed is to estimate $\theta_A = \dfrac{\sum_{i=1}^{N} \theta_i}{N}$. In order to estimate θ_A we employ the scheme of sample selection given by Hartley and Rao (1962) using the size measure as l_i. Here the units in a population are first randomly permuted. Then, from the permuted vector of labeled units two probability proportionate to size (PPS) circular systematic samples (CSS) are chosen in n_1 and n_2 draws respectively, with n_i's as the intended sample sizes, provided $n_1 p_i < 1$ and $n_2 p_i < 1 \forall i \in U$, $p_i = \dfrac{l_i}{\sum_{i=1}^{N} l_i}$.

Two samples s_1 and s_2 of the sizes $n_1 = 39$ and $n_1 = 31$ have been drawn from the population U. Each individual of each sample is being requested to employ procedures as mentioned in the above Item count method, RR method and

the Optional method. We add Table 1 below to present our simulation-based study of the estimated coefficient of variations (CV_{est}'s).

Table 1. Performance characteristics based on different methods

Methods	**100(CV_{est})**
Item count (IC) method	12.6
Randomized response (RR) method	38.4
Optional method	8.6

6. Conclusion

There are two existing estimation methods (Item Count and Randomized Response) to estimate θ_A for complex survey data. A new method has been proposed in the paper where one can give an option for Item Count method or Randomized Response method so that the respondent can choose the option according to his/her choice. There is a rich growth of literature on randomized response (RR) procedure. Any RR device may be used here for which the RR values cover the ranges of Item count values.

As far as the CV criterion is concerned, it is shown that the optional method gives better estimate than the other two existing methods.

Acknowledgement

The author is grateful to Dr. Arijit Chaudhuri of Indian Statistical Institute for some useful discussions.

REFERENCES

ABUL-ELA, ABDEL-LATIF A, GREENBERG BG , HORVITZ DG (1967). A multiproportions RR model. Jour. of the Amer. Statist. Assoc., 62, 990—1008.

CHAUDHURI, A. (2001). Using a randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. Jour. Stat. Plan. Inf., 94(1), 37—42.

CHAUDHURI, A., CHRISTOFIDES, T.C. (2006). Item count technique in estimating the proportion of people with a sensitive feature. Accepted for publication in Jour. Stat. Plan. Inf.

CHAUDHURI, A., MUKHERJEE, R. (1988). Randomized response: Theory and Techniques. Marcel Dekker, Inc. N. Y.

DROITCOUR, J., CASPAR, R.A., HUBBARD, M.L, PARSLEY, T.L., VISSCHER, W. EZZATI, T.M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In measurement errors in surveys (eds. P.P. Biemerm et al), New York: John Wiley & Sons, Inc.

GREENBERG, B.G., ABUL-Ela, ABDEL-LATIF, A., SIMMONS, W.R. & HORVITZ, D.G. (1969) The unrelated question RR model: theoretical framework. Jour. Amer. Statist. Assoc., 64, 520—539

HARTLEY, H.O., RAO, J.N.K. (1962). Sampling with unequal probabilities and without replacement. Ann. Math. Stat., 33, 350—374

HORVITZ, D.G., THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. Jour. of the Amer. Statist. Assoc., 47, 663—685.

KUK, A.Y.C (1990). Asking sensitive questions indirectly. Biometrika, 77, 436—438

MANGAT, N.S , SINGH, R. (1990). An alternative randomized response procedure. Biometrika. 77, 439—442

MILLER, J.D. (1985). The nominative technique: A new method of estimating heroin prevalence. NIDA research monograph 57, 104—124.

WARNER, S.L (1965). RR: a survey technique for eliminating evasive answer bias. Jour. Amer. Stat. Assoc. 60, 63-69.

STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 311—330

EFFICIENT ESTIMATION USING DEEP-POST STRATIFICATION UNDER TWO WAY R X R SET-UP

Manish Trivedi[1] and D. Shukla[2]

ABSTRACT

This paper presents an estimation strategy for the population mean assuming a r x r deeply stratified population using a technique deep – post - stratification. If it assumed here that size of each strata is unknown along with unknown frame. The only known information is proportion of row and column size totals of deep – stratification. An efficient estimator is proposed and its optimum properties are examined. A relative comparison of efficiencies is incorporated. It is concluded that fifty percent of the sum of rows – size – total properties and same of column – size – total proportions and same of column – size – total proportions provides an easy choice of unknown constant. An empirical study is made over obtained results showing a highly significant gain while using technique of deep – post – stratification. Approximate expression of mean square error (MSE) is Derived for this set-up.

Keywords: Post - stratifications, SRSWOR, Optimal, Deep –Stratification.

1. Introduction

When a population is assumed stratified according to the r levels of an attribute and every level is further stratified as per r levels of another attribute then it constituted a two way r x r deep – stratified set – up studied by Bryant (1955). For selecting a random sample, it is necessary to have knowledge of complete frames of each strata in a stratified sampling set-up. Suppose, if strata sizes are known but list of units as per each stratum are absent then estimation has a difficulty under deep stratified set-up. This motivates to utilize the post-stratification sampling technique for estimating the population mean. According

[1] Department of Applied Mathematics Birla Institute of Technology Mesra, Ranchi (JH) – 835 215, INDIA, E-mail : manish_trivedi1976@yahoo.com
[2] Department of Mathematics and Statistics Dr. Hari Singh Gaur University Sagar (M.P.) – 470003, INDIA, E-mail : diwakarshukla@rediffmail.com

to Sukhatme [(1984), pp.134] the post-stratification with a large sample size is almost as precise as stratified sampling with proportional allocation. Further, suppose sizes of all r x r strata are also not available along with frames, rather only pooled information about their row and columns size totals are given, then to take be the estimation the estimation aspect of population mean is considered in this paper by establishing a deep-post-stratification scheme. Some useful contributions on post-stratification are Holt and Smith (1979), Jagers, Oden, and Trulsson (1985) and Jagers (1986). Bryant, Hartley and Jessen (1960) have proposed a technique to estimate the population mean when the sample is not large enough to provide an allocation to each stratum of a two-way stratified set-up. Let stratum wise frames and sizes both are not available but their row-sizes and column-size proportions of totals are known. To utilize this information in the estimation aspect of a population mean is considered in this paper.

2. Notations

We assume herein a r x r deeply stratified population of size N. Let Y_{ijk} be the kth value of $(i, j)^{th}$ strata, having size N_{ij} of a variable Y under study (i = 1, 2……..r, j = 1, 2……..r and k = 1, 2…..N_{ij}). We draw a large sample n by SRSWOR and deep post-stratified into n_{ij} units $\left(\sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij} = n\right)$ such that n_{ij} comes from N_{ij} $\left(\sum_{i=1}^{r}\sum_{j=1}^{r} N_{ij} = N\right)$.

Further, let $\bar{Y}_{ij}$ is the mean and S_{ij}^2 is the population mean square of $(i,j)^{th}$ strata. Also, $\bar{Y}$ and S^2 are population mean and population mean square whereas $\bar{y}_{ij}$ and $\bar{y}$ are sample means based on n_{ij} and n units respectively. Moreover N_i $\left(= \sum_{j=1}^{r} N_{ij}\right)$, $N_{.j}$ $\left(= \sum_{i=1}^{r} N_{ij}\right)$, $n_{i.}$ $\left(= \sum_{j=1}^{r} n_{ij}\right)$ and $n_{.j}$ $\left(= \sum_{i=1}^{r} n_{ij}\right)$ are row and column –size – totals while $\bar{Y}_{i.}$, $\bar{Y}_{.j}$, $\bar{y}_{i.}$, $\bar{y}_{.j}$, are population and sample means based on them.

2.1. Some results

Let $W_{ij} = \left(\frac{N_{ij}}{N}\right)$, $W_{i.} = \left(\frac{N_{i.}}{N}\right)$, $W_{.j} = \left(\frac{N_{.j}}{N}\right)$, $p_{ij} = \left(\frac{n_{ij}}{n}\right)$, $p_{i.} = \left(\frac{n_{i.}}{n}\right)$,

and $p_{.j} = \left(\frac{n_{.j}}{n}\right)$. Assume the sample n is large enough, with respect to r x r stratification, in order to strongly support followings:

$$\left.\begin{aligned} p_{ij} &= W_{ij}\ (1 + \in_{ij}) \\ p_{i'j} &= W_{i'j}\ (1 + \in_{i'j}) \\ p_{ij'} &= W_{ij'}\ (1 + \in_{ij'}) \end{aligned}\right\} \qquad (2.1.1)$$

Where $E[\in_{ij}] = E[\in_{i'j}] = E[\in_{ij'}] = 0;\ i \neq i',\ j \neq j' = 1,2\ldots..r$

And $E\lfloor p_{ij} \rfloor = W_{ij}, \quad E[p_{i.}] = W_{i.,} \quad E\lfloor p_{.j} \rfloor = W_{.j}$ (2.1.2)

Moreover,

$$E\left[\varepsilon_{ij}^2\right] = \frac{V\left[p_{ij}\right]}{W_{ij}^2} = \left(\frac{1}{W_{ij}^2}\right)\left[\frac{(N-n)\,W_{ij}\,(1-W_{ij})}{(N-1)\,n}\right] \qquad (2.1.3)$$

$$E\left[\varepsilon_{i'j}^2\right] = \frac{V\left[p_{i'j}\right]}{W_{i'j}^2} = \left(\frac{1}{W_{i'j}^2}\right)\left[\frac{(N-n)\,W_{i'j}\,(1-W_{i'j})}{(N-1)\,n}\right] \qquad (2.1.4)$$

$$E\left[\varepsilon_{ij'}^2\right] = \frac{V\left[p_{ij'}\right]}{W_{ij'}^2} = \left(\frac{1}{W_{ij'}^2}\right)\left[\frac{(N-n)\,W_{ij'}\,(1-W_{ij'})}{(N-1)\,n}\right] \qquad (2.1.5)$$

$$E\left[\varepsilon_{i'j}\ \varepsilon_{ij'}\right] = \frac{Cov\left[p_{i'j}\ \ p_{ij'}\right]}{\left\{W_{i'j}\ W_{ij'}\right\}} = -\left\{\frac{1}{\left(W_{i'j}\ \ W_{ij'}\right)}\right\}\left[\frac{(N-n)\,W_{ij'}\,W_{i'j}}{(N-1)n}\right] \qquad (2.1.6)$$

$$E\left[\varepsilon_{i'j}\ \ \varepsilon_{ij}\right] = \frac{Cov\left[p_{i'j}\ \ p_{ij}\right]}{\left\{W_{i'j}\ W_{ij}\right\}} = -\left\{\frac{1}{\left(W_{i'j}\ \ W_{ij}\right)}\right\}\left[\frac{(N-n)\,W_{ij}\,W_{i'j}}{(N-1)n}\right] \qquad (2.1.7)$$

$$E\left[\varepsilon_{ij}\ \ \varepsilon_{ij'}\right] = \frac{Cov\left[p_{ij}\ \ p_{ij'}\right]}{\left\{W_{ij}\ W_{ij'}\right\}} = -\left\{\frac{1}{\left(W_{ij}\ \ W_{ij'}\right)}\right\}\left[\frac{(N-n)\,W_{ij'}\,W_{ij}}{(N-1)n}\right] \qquad (2.1.8)$$

$$V\left[p_{i.}\right] = \left[\frac{(N-n)}{(N-1)} \cdot \frac{W_{.i}\ (1-W_{.i})}{n}\right] \qquad (2.1.9)$$

$$V\left[p_{.j}\right] = \left[\frac{(N-n)}{(N-1)} \cdot \frac{W_{.j}\ (1-W_{.j})}{n}\right] \qquad (2.1.10)$$

$$Cov\left[p_{i.}\ \ p_{.j'}\right] = -\left[\frac{(N-n)W_{i.}\ W_{.j}}{(N-1)n}\right] \qquad (2.1.11)$$

Theorem 2.1

Using (2.1.1) and avoiding terms of higher order, an approximate results, for j ≠ j', is

(a) $$E\left[\frac{p_{ij'}}{p_{ij}}\right] = \frac{W_{ij'}}{W_{ij}}\left[1+\frac{Var(P_{ij})}{W_{ij}^2} - \frac{Cov\left(p_{ij}\ \ p_{ij'}\right)}{W_{ij}\,W_{ij'}}\right]$$

(b) $$E\left[\frac{p_{ij'}^2}{p_{ij}}\right] = \frac{W_{ij'}^2}{W_{ij}}\left[1+\frac{Var\left(p_{ij}\right)}{W_{ij}^2} + \frac{Var\left(p_{ij'}\right)}{W_{ij'}^2} - \frac{2Cov\left(p_{ij}\ p_{ij'}\right)}{W_{ij}\,W_{ij'}}\right]$$

(c) $$E\left[\frac{p_{ij'}\ p_{i'j}}{p_{ij}}\right] = \frac{W_{ij'}\,W_{i'j}}{W_{ij}}\left[1+\frac{Var\left(p_{ij}\right)}{W_{ij}^2} + \frac{Cov\left(p_{i'j}\ p_{ij'}\right)}{W_{i'j}\,W_{ij'}} - \frac{Cov\left(p_{ij}\ p_{i'j}\right)}{W_{ij}\,W_{i'j}} - \frac{Cov\left(p_{ij}\ p_{ij'}\right)}{W_{ij}\,W_{ij'}}\right]$$

Proof: (a) $$E\left[\frac{p_{ij'}}{p_{ij}}\right] = E\left[\frac{W_{ij'}\left(1+\varepsilon_{ij'}\right)}{W_{ij}\left(1+\varepsilon_{ij}\right)}\right]$$

$$= \frac{W_{ij'}}{W_{ij}}\,E\left[\left(1+\varepsilon_{ij'}\right)\left(1+\varepsilon_{ij}\right)^{-1}\right]$$

$$= \frac{W_{ij'}}{W_{ij}}\,E\left[\left(1+\varepsilon_{ij'}\right)\left(1-\varepsilon_{ij}+\varepsilon_{ij}^2\ .....\right)\right]$$

$$= \frac{W_{ij'}}{W_{ij}}\,E\left[1+\varepsilon_{ij'}-\varepsilon_{ij}-\varepsilon_{ij}\,\varepsilon_{ij'}+\varepsilon_{ij}^2+\varepsilon_{ij}^2\,\varepsilon_{ij'}\,.....\right]$$

On avoiding all higher order terms $\left[\left(\varepsilon_{ij}\right)^r\ \left(\varepsilon_{ij'}\right)^s\right]$ for (r + s) > 2, we have

$$= \frac{W_{ij'}}{W_{ij}}\left[1+E\left(\varepsilon_{ij}^2\right)-E\left(\varepsilon_{ij}\ \ \varepsilon_{ij'}\right)\right]$$

$$= \frac{W_{ij'}}{W_{ij}}\left[1+\frac{Var\left(p_{ij}\right)}{W_{ij}^2} - \frac{Cov\left(p_{ij}\ p_{ij'}\right)}{W_{ij}\,W_{ij'}}\right]$$

(b) $$E\left[\frac{p_{ij'}^2}{p_{ij}}\right] = E\left[\frac{W_{ij'}^2\left(1+\varepsilon_{ij'}\right)^2}{W_{ij}\left(1+\varepsilon_{ij}\right)}\right]$$

$$= \frac{W_{ij'}^{2}}{W_{ij}} E\left[\left(1+\varepsilon_{ij'}\right)^{2}\left(1+\varepsilon_{ij}\right)^{-1}\right]$$

$$= \frac{W_{ij'}^{2}}{W_{ij}} E\left[\left(1+\varepsilon_{ij'}^{2} + 2\varepsilon_{ij''}\right)\left(1-\varepsilon_{ij} + \varepsilon_{ij}^{2} \ldots\ldots\right)\right]$$

$$= \frac{W_{ij'}^{2}}{W_{ij}} E\left[1+\varepsilon_{ij'}^{2} + 2\varepsilon_{ij'} - \varepsilon_{ij} - 2\varepsilon_{ij}\,\varepsilon_{ij'} + \varepsilon_{ij}^{2} \ldots\ldots\right]$$

On avoiding all higher order terms $\left[\left(\varepsilon_{ij}\right)^{r}\left(\varepsilon_{ij'}\right)^{s}\right]$ for (r + s) > 2, we have

$$= \frac{W_{ij'}^{2}}{W_{ij}}\left[1+E\left(\varepsilon_{ij}^{2}\right)+E\left(\varepsilon_{ij'}^{2}\right)-2E\left(\varepsilon_{ij}\,\varepsilon_{ij'}\right)\right]$$

$$= \frac{W_{ij'}^{2}}{W_{ij}}\left[1+\frac{Var\left(p_{ij}\right)}{W_{ij}^{2}}+\frac{Var\left(p_{ij'}\right)}{W_{ij'}^{2}}-\frac{2Cov\left(p_{ij}\,p_{ij'}\right)}{W_{ij}\,W_{ij'}}\right]$$

$$\text{(c)}\quad E\left[\frac{p_{i'j}\,p_{ij'}}{p_{ij}}\right] = E\left[\frac{W_{i'j}\left(1+\varepsilon_{i'j}\right)W_{ij'}\left(1+\varepsilon_{ij'}\right)}{W_{ij}\left(1+\varepsilon_{ij}\right)}\right]$$

$$= \frac{W_{i'j}\,W_{ij'}}{W_{ij}} E\left[\left(1+\varepsilon_{i'j}\right)\left(1+\varepsilon_{ij'}\right)\left(1+\varepsilon_{ij}\right)^{-1}\right]$$

$$= \frac{W_{i'j}\,W_{ij'}}{W_{ij}} E\left[\left(1+\varepsilon_{i'j}+\varepsilon_{ij'}+\varepsilon_{i'j}\,\varepsilon_{ij'}\right)\left(1-\varepsilon_{ij}+\varepsilon_{ij}^{2} \ldots\ldots\right)\right]$$

$$= \frac{W_{i'j}\,W_{ij'}}{W_{ij}} E\left[1+\varepsilon_{i'j}+\varepsilon_{ij'}+\varepsilon_{i'j}\varepsilon_{ij'}-\varepsilon_{ij}-\varepsilon_{ij}\varepsilon_{i'j}-\varepsilon_{ij}\,\varepsilon_{ij'}+\varepsilon_{ij}^{2} \ldots\ldots\right]$$

On avoiding terms $\left[\left(\varepsilon_{ij}\right)^{r}\left(\varepsilon_{ij'}\right)^{s}\left(\varepsilon_{i'j}\right)^{t}\right]$ for (r + s + t) > 2, we have

$$= \frac{W_{i'j}\,W_{ij'}}{W_{ij}}\left[1+E\left(\varepsilon_{ij}^{2}\right)+E\left(\varepsilon_{i'j}\,\varepsilon_{ij'}\right)-E\left(\varepsilon_{ij}\,\varepsilon_{i'j}\right)-E\left(\varepsilon_{ij}\,\varepsilon_{ij'}\right)\right]$$

$$= \frac{W_{i'j}\,W_{ij'}}{W_{ij}}\left[1+\frac{Var\left(p_{ij}\right)}{W_{ij}^{2}}+\frac{Cov\left(p_{i'j}\,p_{ij'}\right)}{W_{i'j}\,W_{ij'}}-\frac{Cov\left(p_{ij}\,p_{i'j}\right)}{W_{ij}\,W_{i'j}}-\frac{Cov\left(p_{ij}\,p_{ij'}\right)}{W_{ij}\,W_{ij'}}\right]$$

2.2. Some denotations

$$A_{ij(j')} = E\left[\frac{p_{ij'}}{p_{ij}}\right], B_{ij(j')} = E\left[\frac{p^2_{ij'}}{p_{ij}}\right] \quad C_{ij(i',j')} = E\left[\frac{p_{i'j}\, p_{ij'}}{p_{ij}}\right]$$

$$D_{ij} = E\left[\frac{1}{n_{ij}}\right] = \frac{1}{nW_{ij}} + \frac{(N-n)}{(N-1)}\frac{(1-W_{ij})}{n^2 W_{ij}^2}$$

$$F_{i.} = E\left[\frac{p^2_{i.}}{N_{ij}}\right] = \left(\frac{1}{N_{ij}}\right)\left\{\frac{(N-n)}{(N-1)}\frac{W_{i.}(1-W_{i.})}{n} + W^2_{i.}\right\}$$

$$F_{.j} = E\left[\frac{p^2_{.j.}}{N_{ij}}\right] = \left(\frac{1}{N_{ij}}\right)\left\{\frac{(N-n)}{(N-1)}\frac{W_{.j}(1-W_{.j})}{n} + W^2_{.j}\right\}$$

$$F_{ij} = E\left[\frac{p_{i.}\, p_{.j}}{N_{ij}}\right] = \left(\frac{1}{N_{ij}}\right)\left\{Cov\left(p_{i.}\ p_{.j}\right) + E(p_{i.})\,E(p_{.j})\right\}$$

$$M_{i.} = \sum_{j=1}^{r}\overline{Y}_{ij}, \quad M_{.j} = \sum_{i=1}^{r}\overline{Y}_{ij},$$

3. An estimation strategy

To recall, underlying assumptions here are:

(a) A two-way set-up of r x r stratified population N exists;
(b) frame of N units is available (with respect to any non-stratifying variable) but not the size and frame of each stratum for the criteria of deep-stratification;
(c) sample size n is sufficiently large to have a good representation of r x r;
(d) Although W_{ij} is unknown but information about $W_{i.}$ And $W_{.j}$ available by some other sources.

To estimate $\overline{Y}$, a "Deep-Post-Stratified" estimator is

$$\overline{y}_{gdps} = \sum_{i=1}^{r}\sum_{j=1}^{r} W_{\alpha ij}\ \overline{y}_{ij} \tag{3.1}$$

Where, $W_{\alpha ij} = \left[(\alpha/2)\{p_{i.} + W_{i.}\} + ((1-\alpha)/2)\{p_{.j} + W_{.j}\}\right]$

Here α is a suitable chosen constant such that $0 \le \alpha \le 1$.

3.1. Motivation and justification

I. The usual post-stratified and sample mean estimators for a r x r set-up , with known W_{ij} are

$$\bar{y}_{ps} = \sum_{i=1}^{r}\sum_{j=1}^{r} W_{ij}\,\bar{y}_{ij} \tag{3.2.1}$$

$$\bar{y} = \left(1/n\right)\sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij}\,\bar{y}_{ij}$$

$$v\left(\bar{y}\right) = \left\{\left(1/n\right) - \left(1/N\right)\right\} S^2 \tag{3.2.2}$$

When information of $W_{i.}$ And $W_{.j}$ are available but information about W_{ij}'s are absent the usual estimator (3.2.1) fails to perform estimation and one has to choose (3.2.2).

II. The information of $W_{i.}$ and $W_{.j}$ are common to be known priory like a university has p_1 proportion of meritorious students, p_2 average students and p_3 below average student ($p_1 + p_2 + p_3 = 1$) whereas same institution may with a proportion of p_4 students from poor families, p_5 from middle class and p_6 from rich class ($p_4 + p_5 + p_6 = 1$). This does not provide details of N_{ij} in 3 x 3 set-up, so the usual estimator (3.2.1) prossess a severe limitation.

III. Consider a telephone directory of N phone-holders in a city. From past record, guess, pilot survey, or otherwise information available that 40% businessman, 30% serviceman, 20% private doctors and 10% others are telephone – owners, whereas 40% postgraduates, 30% graduates, 20% secondary – educated and 10% Metric-educated are phone-owners in the city. This information needs to be utilized.

IV. A proper and intelligent utilization of known $W_{i.}$ and $W_{.j}$ in the construction of estimator, has motivated to propose estimator $\bar{y}_{gdps}$.

V. A logic for $W_{\alpha ij}$ is to choose a fraction of size of row and column-totals both, in the construction of estimator which is decided by a suitable selection of the constant α. A contribution by Agarwal and Panda (1993) support the motivations for choosing $W_{\alpha ij}$ in the present form.

3.2. Special estimators

(I) At $\alpha = 1$, $W_{1ij} = (1/2)\,[p_{i.} + W_{i.}]$

$$\left(\bar{y}_{gdps}\right)_1 = \sum_i^r \sum_j^r W_{1ij}\, \bar{y}_{ij} \qquad (3.2.3)$$

(II) At $\alpha = 0$, $W_{0ij} = (1/2)\,[p_{.j} + W_{.j}]$

$$\left(\bar{y}_{gdps}\right)_0 = \sum_i^r \sum_j^r W_{0ij}\, \bar{y}_{ij} \qquad (3.2.4)$$

(III) At $\alpha = ½$, $W_{1/2ij} = (1/4)\,[\{p_{i.} + W_{i.}\} + \{p_{.j} + W_{.j}\}]$

$$\left(\bar{y}_{gdps}\right)_{1/2} = \sum_i^r \sum_j^r W_{1/2ij}\, \bar{y}_{ij} \qquad (3.2.5)$$

Here $\left(\bar{y}_{gdps}\right)_1$ is purely based on row-totals pf strata size, $\left(\bar{y}_{gdps}\right)_0$ on column-totals and $\left(\bar{y}_{gdps}\right)_{½}$ is on an average of these two.

Theorem 3.1

The estimator $\bar{y}_{gdps}$ is biased for $\bar{Y}$.

Proof: Denote E [(.)/n_{ij}] as a conditional expectation given n_{ij}.

$$E\left(\bar{y}_{gdps}\right) = E\left[E\left[\left(\bar{y}_{gdps}\right) \Big/ n_{ij}\right]\right]$$

$$= E\left[\left\{\sum_{i=1}^r \sum_{j=1}^r W_{\alpha ij}\, E\left(\bar{y}_{ij}\right)\right\} \Big/ n_{ij}\right]$$

$$= \sum_{i=1}^r \sum_{j=1}^r E\left(W_{\alpha ij}\right) \bar{Y}_{ij}$$

$$= \sum_{i=1}^r \sum_{j=1}^r \left[\alpha \left(N_{i.} \big/ N\right) + (1-\alpha)\left(N_{.j} \big/ N\right)\right] \bar{Y}_{ij}$$

$$= \bar{Y} + \left[\alpha V_1 + (1-\alpha) V_2\right] = \bar{Y} + \left[Bias\left(\bar{y}_{gdps}\right)\right]$$

Where, $V_1 = \sum_{i=1}^r \sum_{j=1}^r \sum_{j' \neq j}^r W_{ij'}\, \bar{Y}_{ij}$, $\quad V_2 = \sum_{i=1}^r \sum_{i' \neq i}^r \sum_{j=1}^r W_{i'j}\, \bar{Y}_{ij}$

Theorem 3.2

M.S.E. of the estimator $\bar{y}_{gdps}$ is

$$MSE\left(\bar{y}_{gdps}\right) = (1/4)\left[U_1 + \alpha^2\left(R_1 + R_2 + 4V_1^2\right) + (1-\alpha)^2\left(S_1 + S_2 + 4V_2^2\right) + 2\alpha(1-\alpha)\left(T_1 + T_2 + 4V_1V_2\right)\right]$$

Where, $U_1 = \left[\left((4r-1)/n\right) - \left(3(2r-1)/N\right)\right] \sum_{i=1}^{r}\sum_{j=1}^{r} W_{ij}\ S_{ij}^2$

$$R_1 = \sum_{i=1}^{r}\sum_{j\neq j'=1}^{r}\left(\frac{1}{n}\right)B_{ij(j')}\ S_{ij}^2 + \sum_{i=1}^{r}\sum_{j=1}^{r} W_{i.}^2\ D_{ij}\ S_{ij}^2 + \sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r}\left(\frac{2(r-2)}{n}\right)C_{ij(i'.j')}\ S_{ij}^2$$
$$+\left(\frac{2}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r} W_{i.}\ A_{ij(j')}\ S_{ij}^2 - \sum_{i=1}^{r}\sum_{j=1}^{r} F_{i.}\ S_{ij}^2 - \left(\frac{(2r-3)}{N}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r}\left\{\frac{W_{ij'}^2\ S_{ij}^2}{W_{ij}}\right\}$$
$$-\left(\frac{6(r-2)}{N}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r}\frac{W_{i'j}\,W_{ij'}}{W_{ij}}\ S_{ij}^2$$

$$R_2 = \left\{\frac{(N-n)}{(N-1)n}\right\}\left[\sum_{i=1}^{r} W_{i.}\left(1-W_{i.}\right)M_{i.}^2 - \sum_{i=1}^{r}\sum_{i'=1}^{r} M_{i.}M_{i'.}W_{i.}W_{i.'}\right]$$

$$S_1 = \sum_{i=1}^{r}\sum_{j\neq j'=1}^{r}\left(\frac{1}{n}\right)B_{ij(j')}\ S_{ij}^2 + \sum_{i=1}^{r}\sum_{j=1}^{r} W_{.j}^2\ D_{ij}\ S_{ij}^2 + \sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r}\left(\frac{2(r-2)}{n}\right)C_{ij(i',j')}\ S_{ij}^2$$
$$+\left(\frac{2}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r} W_{.j}\ A_{ij(j')}\ S_{ij}^2 - \sum_{i=1}^{r}\sum_{j=1}^{r} F_{.j}\ S_{ij}^2 - \left(\frac{(2r-3)}{N}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r}\left\{\frac{\left(W_{ij'}^2\ S_{ij}^2\right)}{W_{ij}}\right\}$$
$$-\left(\frac{6(r-2)}{N}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r}\left(\frac{W_{i'j}\ W_{ij'}}{W_{ij}}\right)S_{ij}^2$$

$$S_2 = \left\{\frac{(N-n)}{(N-1)n}\right\}\left[\sum_{j=1}^{r} W_{.j}\left(1-W_{.j}\right)M_{.j}^2 - \sum_{j=1}^{r}\sum_{j'=1}^{r} M_{.j}M_{.j'}W_{.j}W_{.J'}\right]$$

$$T_1 = \sum_{i=1}^{r}\sum_{j=1}^{r} W_{i.}\,W_{.j}\ D_{ij}\ S_{ij}^2 + \sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r}\left(\frac{1}{n}\right)C_{ij(i',j')}\ S_{ij}^2 - \sum_{i=1}^{r}\sum_{j=1}^{r} F_{ij}\ S_{ij}^2$$
$$+\left(\frac{1}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r}\left(W_{i.} + W_{.j}\right)A_{ij(j')}\ S_{ij}^2 - \left(\frac{3}{N}\right)\sum_{i=i'=1}^{r}\sum_{j\neq j'=1}^{r}\left(\frac{W_{i'j}\ W_{ij'}}{W_{ij}}\right)S_{ij}^2$$

$$T_2 = -\left\{\frac{(N-n)}{(N-1)n}\right\}\sum_{i=1}^{r}\sum_{j=1}^{r} W_{i.}W_{.j}M_{i.}M_{.j}$$

A detailed description of notations is given in page no. 6.

Proof: $MSE(\bar{y}_{gdps}) = V(\bar{y}_{gdps}) + [Bias(\bar{y}_{gdps})]^2$ (3.2.6)

$$V(\bar{y}_{gdps}) = E[V(\bar{y}_{gdps})/n_{ij}] + V[E(\bar{y}_{gdps})/n_{ij}] \quad (3.2.7)$$

where $V[(.)/n_{ij}]$ is a conditional variance for given n_{ij}.Now, on picking first component of (3.2.7).

$$E[V(\bar{y}_{gdps})/n_{ij}] = E\left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{n_{ij}}\right)W_{\alpha ij}^2 S_{ij}^2\right] - E\left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{N_{ij}}\right)W_{\alpha ij}^2 S_{ij}^2\right] \quad (3.2.8)$$

The term one of (3.2.8) is

$$E\left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{n_{ij}}\right)W_{\alpha ij}^2 S_{ij}^2\right] = E\left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{n_{ij}}\right)\left\{\left(\frac{\alpha^2}{4}\right)[p_{i.}+W_{i.}]^2\right.\right.$$

$$\left.\left. + \left(\frac{(1-\alpha)^2}{4}\right)(p_{.j}+W_{.j})^2 + \left\{\frac{\alpha(1-\alpha)}{2}\right\}(p_{i.}+W_{i.})(p_{.j}+W_{.j})\right\} S_{ij}^2\right]$$

Here, we have results

$$a_1 : E\left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{n_{ij}}\right)\{p_{i.}+W_{i.}\}^2 S_{ij}^2\right] = \left\{\frac{(4r-1)}{n}\right\}\sum_{i=1}^{r}\sum_{j=1}^{r} W_{ij}S_{ij}^2 + \left(\frac{1}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r} B_{ij(j')}S_{ij}^2$$

$$+ \left\{\frac{2(r-2)}{n}\right\}\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r} C_{ij(i',j')}S_{ij}^2 + \sum_{i=1}^{r}\sum_{j=1}^{r} W_{i.}^2 D_{ij}S_{ij}^2 + \left(\frac{2}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r} W_{i.}A_{ij(j')}S_{ij}^2$$

$$a_2 : E\left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{n_{ij}}\right)\{p_{.j}+W_{.j}\}^2 S_{ij}^2\right] = \left\{\frac{(4r-1)}{n}\right\}\sum_{i=1}^{r}\sum_{j=1}^{r} W_{ij}S_{ij}^2 + \left(\frac{1}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r} B_{ij(j')}S_{ij}^2$$

$$+ \left(\frac{2(r-2)}{n}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r} C_{ij(i',j')}S_{ij}^2 + \sum_{i=1}^{r}\sum_{j=1}^{r} W_{.j}^2 D_{ij}S_{ij}^2 + \left(\frac{2}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r} W_{.j}A_{ij(j')}S_{ij}^2$$

$$a_3 : E\left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{n_{ij}}\right)\left\{\left(p_{i.}+W_{i.}\right)\left(p_{.j}+W_{.j}\right)\right\}S_{ij}^2\right] = \left\{\frac{(4r-1)}{n}\right\}\sum_{i=1}^{r}\sum_{j\neq j=1}^{r} W_{ij}\,S_{ij}^2$$

$$+\sum_{i=1}^{r}\sum_{j=1}^{s} W_{i.}W_{.j}\,D_{ij}\,S_{ij}^2 + \left(\frac{1}{n}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r} C_{ij(i',j')}\,S_{ij}^2 + \left(\frac{1}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r}\left(W_{i.}+W_{.j}\right)A_{ij(j')}\,S_{ij}^2$$

While obtaining results in a_1, a_2, a_3 theorem 2.1 is used wherever required. The use of a_1, a_2, and a_3 along with α and other terms helps of obtain an expression:

$$E\left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{n_{ij}}\right)W_{\alpha ij}^2\,S_{ij}^2\right] = \left(\frac{\{4r-1\}}{4n}\right)\sum_{i=1}^{r}\sum_{j=1}^{r} W_{ij}\,S_{ij}^2 + \left(\frac{\alpha^2}{4}\right)\left[\left(\frac{1}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r} B_{ij(j')}\,S_{ij}^2\right.$$

$$\left.+\left(\frac{2(r-2)}{n}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r} C_{ij(i',j')}\,S_{ij}^2 + \sum_{i=1}^{r}\sum_{j=1}^{r} W_{i.}^2\,D_{ij}\,S_{ij}^2 + \left(\frac{2}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r} W_{i.}A_{ij(j')}\,S_{ij}^2\right]$$

$$+\left(\frac{(1-\alpha)^2}{4}\right)\left[\left(\frac{1}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r} B_{ij(j')}\,S_{ij}^2 + \left(\frac{2(r-2)}{n}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r} C_{ij(i',j')}\,S_{ij}^2 + \sum_{i=1}^{r}\sum_{j=1}^{r} W_{.j}^2\,D_{ij}\,S_{ij}^2\right.$$

$$+\left(\frac{2}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r} W_{.j}\,A_{ij(j')}\,S_{ij}^2 + \left(\frac{\alpha(1-\alpha)}{2}\right)\left[\left(\frac{1}{n}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r} C_{ij(i',j')}\,S_{ij}^2\right.$$

$$\left.\left.+\sum_{i=1}^{r}\sum_{j=1}^{r} W_{i.}W_{.j}\,D_{ij}\,S_{ij}^2 + \left(\frac{1}{n}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r}\left(W_{i.}+W_{.j}\right)A_{ij(j')}\,S_{ij}^2\right] \qquad \ldots\ldots(3.2.9)$$

We also have,

$$a_4 = E\left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{N_{ij}}\right)\left\{\left(p_{i.}+W_{i.}\right)\right\}^2\,S_{ij}^2\right] = \left(\frac{3(2r-1)}{N}\right)\sum_{i=1}^{r}\sum_{j=1}^{r} W_{ij}\,S_{ij}^2 + \sum_{i=1}^{r}\sum_{j=1}^{r} F_{i.}\,S_{ij}^2$$

$$+\left(\frac{2r-3}{N}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r}\frac{W_{ij'}^2}{W_{ij}}\,S_{ij}^2 + \left(\frac{6(r-2)}{N}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r}\frac{W_{i'j}\,W_{ij'}}{W_j}\,S_{ij}^2$$

$$a_5 = E\left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{N_{ij}}\right)\left\{\left(p_{.j} + W_{.j}\right)\right\}^2 S_{ij}^2\right] = \left(\frac{3(2r-1)}{N}\right)\sum_{i=1}^{r}\sum_{j=1}^{r} W_{ij}\, S_{ij}^2 + \sum_{i=1}^{r}\sum_{j=1}^{r} F_{.j}\, S_{ij}^2$$

$$+\left(\frac{2r-3}{N}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r}\frac{W_{ij'}^2}{W_{ij}}S_{ij}^2 + \left(\frac{6(r-2)}{N}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r}\frac{W_{i'j}\,W_{ij'}}{W_{ij}}\,S_{ij}^2$$

$$a_6 = \left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{N_{ij}}\right)\left\{\left(p_{i.} + W_{i.}\right)\left(p_{.j} + W_{.j}\right)\right\}S_{ij}^2\right] = \left(\frac{3(2r-1)}{N}\right)\sum_{i=1}^{r}\sum_{j=1}^{r} W_{ij}\, S_{ij}^2$$

$$+\sum_{i=1}^{r}\sum_{j=1}^{r} F_{ij}\, S_{ij}^2 + \left(\frac{3}{N}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r}\frac{W_{i'j}\,W_{ij'}}{W_{ij}}\,S_{ij}^2$$

Theorem 2.1 is also used to obtained a_4, a_5 and a_6 and, with the help of there, we get

$$E\left[\sum_{i=1}^{r}\sum_{j=1}^{r}\left(\frac{1}{N_{ij}}\right)W_{\alpha ij}^2\, S_{ij}^2\right] = \left(\frac{1(2r-1)}{4N}\right)\sum_{i=1}^{r}\sum_{j=1}^{r} W_{ij}\, S_{ij}^2 + \left(\frac{\alpha^2}{4}\right)\left[\sum_{i=1}^{r}\sum_{j=1}^{r} F_{i.}\, S_{ij}^2\right.$$

$$\left.+\left(\frac{2r-3}{N}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r}\frac{W_{ij'}^2}{W_{ij}}S_{ij}^2 + \left(\frac{6(r-2)}{N}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r}\frac{W_{i'j}\,W_{ij'}}{W_{ij}}\,S_{ij}^2\right] + \left(\frac{(1-\alpha)^2}{4}\right)\left[\sum_{i=1}^{r}\sum_{j=1}^{r} F_{.j}\, S_{ij}^2\right.$$

$$\left.+\left(\frac{2r-3}{N}\right)\sum_{i=1}^{r}\sum_{j\neq j'=1}^{r}\frac{W_{ij'}^2}{W_{ij}}S_{ij}^2 + \left(\frac{6(r-2)}{N}\right)\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r}\frac{W_{i'j}\,W_{ij'}}{W_{ij}}\,S_{ij}^2\right] + \left(\frac{(1-\alpha)}{2}\right)\left[\sum_{i=1}^{r}\sum_{j=1}^{r} F_{ij}\, S_{ij}^2\right.$$

$$\left.+\left(\frac{3}{N}\right)\left[\sum_{i\neq i'=1}^{r}\sum_{j\neq j'=1}^{r}\frac{W_{i'j}\,W_{ij'}}{W_{ij}}\,S_{ij}^2\right]\right] \qquad \ldots\ldots(3.2.10)$$

On adding (3.2.8) and (3.2.10), one gets

$$E\left[V\left(\bar{y}_{gdps}\right)/n_{ij}\right] = \left(\frac{1}{4}\right)U_1 + \left(\frac{\alpha^2}{4}\right)R_1 + \left\{\frac{(1-\alpha)^2}{4}\right\}S_1 + \left\{\frac{\alpha\,(1-\alpha)}{2}\right\}T_1 \quad (3.2.11)$$

The second component of (3.2.7)presents

$$V\left[E\left(\bar{y}_{gdps}\right)/n_{ij}\right] = V\left[\sum_{i=1}^{r}\sum_{j=1}^{r} W_{\alpha ij}\overline{Y}_{ij}\right]$$

$$= \left(\frac{\alpha^2}{4}\right)\left\{\sum_{i=1}^{r} W_{i.}\left(1-W_{i.}\right)M_{i.}^{2} - \sum_{i=1}^{r}\sum_{i'=1}^{r} M_{i.}M_{i'.}W_{i.}W_{i.'}\right\}$$

$$+\left(\frac{(1-\alpha)^2}{4}\right)\left\{\sum_{j=1}^{r} W_{.j}\left(1-W_{.j}\right)M_{.j}^{2} - \sum_{j=1}^{r}\sum_{j'=1}^{r} M_{.j}M_{.j'}W_{.j}W_{.J'}\right\}$$

$$+\left(\frac{\alpha(1-\alpha)}{4}\right)\left\{\sum_{i=1}^{r}\sum_{j=1}^{r} W_{i.}W_{.j}M_{i.}M_{.j}\right\} \qquad \text{.....(3.2.12)}$$

On addition of (3.2.11) and (3.2.12)along with Bias term completes the proof

4. M.S.E. of special estimators

$$\text{At } \alpha = 1,\ MSE\left[\left(\bar{y}_{gdps}\right)_1\right] = \left(\frac{1}{4}\right)\left\{U_1 + \left(R_1 + R_2 + 4V_1^2\right)\right\}$$

$$\text{At } \alpha = 0,\ MSE\left[\left(\bar{y}_{gdps}\right)_0\right] = \left(\frac{1}{4}\right)\left\{U_1 + \left(S_1 + S_2 + 4V_2^2\right)\right\}$$

$$\text{At } \alpha = \frac{1}{2},\ MSE\left[\left(\bar{y}_{gdps}\right)_{1/2}\right] = \left(\frac{1}{16}\right)\left\{4U_1 + \left(R_1 + R_2 + 4V_1^2\right) + \left(S_1 + S_2 + 4V_2^2\right) + 2\left(T_1 + T_2 + 4V_1V_2\right)\right\}$$

5. Optimum estimator

On differentiating MSE with respect to α and equating to zero, one can easily obtain

$$\alpha_{opt} = \left[\frac{\left(S_1 + S_2 + 4V_2^2\right) - \left(T_1 + T_2 + 4V_1V_2\right)}{\left(R_1 + R_2 + 4V_1^2\right) + \left(S_1 + S_2 + 4V_2^2\right) - 2\left(T_1 + T_2 + 4V_1V_2\right)}\right]$$

Substituting α_{opt} an optimal estimator of $\bar{Y}$ is $\left(\bar{y}_{gdps}\right)_{opt}$ with optimum m.s.e.

$$MSE\left[\left(\bar{y}_{gdps}\right)_{opt}\right] = \left(\frac{1}{4}\right)\left[U_1 + \left\{\frac{\left(R_1 + R_2 + 4V_1^2\right)\left(S_1 + S_2 + 4V_2^2\right) - \left(T_1 + T_2 + 4V_1V_2\right)^2}{\left(R_1 + R_2 + 4V_1^2\right) + \left(S_1 + S_2 + 4V_2^2\right) - 2\left(T_1 + T_2 + 4V_1V_2\right)}\right\}\right]$$

6. Efficiency comparion

I. Estimator $\left(\bar{y}_{gdps}\right)_1$ will be efficient over $\left(\bar{y}_{gdps}\right)_0$

$$\text{if } \left(R_1 + R_2 + 4V_1^2\right) \leq \left(S_1 + S_2 + 4V_2^2\right)$$

II. Estimator $\left(\bar{y}_{gdps}\right)_1$ will be efficient over $\left(\bar{y}_{gdps}\right)_{1/2}$

$$\text{if } \left(R_1 + R_2 + 4V_1^2\right) \leq \left(\frac{1}{3}\right)\left[\left(S_1 + S_2 + 4V_2^2\right) + 2\ \left(T_1 + T_2 + 4V_1V_2\right)\right]$$

III. Estimator $\left(\bar{y}_{gdps}\right)_0$ will be efficient over $\left(\bar{y}_{gdps}\right)_{1/2}$

$$\text{if } \left(S_1 + S_2 + 4V_2^2\right) \leq \left(\frac{1}{3}\right)\left[\left(R_1 + R_2 + 4V_1^2\right) + 2\left(T_1 + T_2 + 4V_1V_2\right)\right]$$

7. Numerical illtrations

For providing a numerical support to expressions derived and results obtained, three populations of size N = 400, N = 650, and N = 490 are generated. Let random samples of size 160, 260, and 196 are drawn from these populations by SRSWOR respectively and post-stratified according to 2 x 2, to the first one and 3 x 3 to the next two. Calculation of some parameters of populations and samples are in tables given below:

Table 7.1. (for data set I)

B \ A		Attribute A: Low	Attribute A: High	Total
Attribute B	Low	$N_{11} = 90,\ n_{11} = 36$ $\overline{Y}_{11} = 71.72$ $W_{11} = 0.225$ $S_{11}^2 = 1713.79$	$N_{12} = 95,\ n_{12} = 38$ $\overline{Y}_{12} = 221.74$ $W_{12} = 0.2375$ $S_{12}^2 = 1872.94$	$N_{1.} = 185$ $n_1 = 74$ $\overline{Y}_{1.} = 113.86$ $w_{1.} = 0.462$
	High	$N_{21} = 105,\ n_{21} = 42$ $\overline{Y}_{21} = 378.54$ $W_{21} = 0.2625$ $S_{21}^2 = 1912.33$	$N_{22} = 110,\ n_{22} = 44$ $\overline{Y}_{22} = 431.9054$ $W_{22} = 0.11384$ $S_{22}^2 = 964.5512$	$N_{2.} = 215$ $n_{2.} = 86$ $\overline{Y}_{2.} = 455.811$ $W_{2.} = 0.5375$
Total		$N_{.1} = 195,\ n_{.1} = 78$ $\overline{Y}_{.1} = 236.93$ $W_{.1} = 0.4875$	$N_{.2} = 205,\ n_{.2} = 82$ $\overline{Y}_{.2} = 386.57$ $W_{.2} = 0.512$	$N = 400,$ $n = 160$ $\overline{Y} = 313.798$ $S^2 = 30962.781$

Table 7.2 (for data set II)

B \ A		Attribute A			Total
		Low	Medium	High	
Attribute B	Low	$N_{11} = 71,\ n_{11} = 28$ $\bar{Y}_{11} = 48.7464$ $W_{11} = 0.10923$ $S_{11}^2 = 857.187$	$N_{12} = 65,\ n_{12} = 26$ $\bar{Y}_{12} = 147.6923$ $W_{12} = 0.1$ $S_{12}^2 = 791.2476$	$N_{13} = 68,\ n_{13} = 27$ $\bar{Y}_{13} = 247.4264$ $W_{13} = 0.10461$ $S_{13}^2 = 876.9092$	$N_{1.} = 204,$ $n_{1.} = 81$ $\bar{Y}_{1.} = 146.499$ $W_{1.} = 0.3138$
	Medium	$N_{21} = 77,\ n_{21} = 31$ $\bar{Y}_{21} = 346.8831$ $W_{21} = 0.11846$ $S_{21}^2 = 866.6208$	$N_{22} = 74,\ n_{22} = 30$ $\bar{Y}_{22} = 431.9054$ $W_{22} = 0.11384$ $S_{22}^2 = 964.5512$	$N_{23} = 80,\ n_{23} = 32$ $\bar{Y}_{23} = 549.525$ $W_{23} = 0.12307$ $S_{23}^2 = 829.1359$	$N_{2.} = 231$ $n_{2.} = 93$ $\bar{Y}_{2.} = 444.29$ $W_{2.} = 0.3553$
	High	$N_{31} = 73,\ n_{31} = 29$ $\bar{Y}_{31} = 654.315$ $W_{31} = 0.1123$ $S_{31}^2 = 787.885$	$N_{32} = 70,\ n_{32} = 28$ $\bar{Y}_{32} = 737.957$ $W_{32} = 0.10769$ $S_{32}^2 = 1044.759$	$N_{33} = 72,\ n_{33} = 29$ $\bar{Y}_{33} = 846.597$ $W_{33} = 0.11076$ $S_{33}^2 = 780.469$	$N_{3.} = 215$ $n_{3.} = 86$ $\bar{Y}_{3.} = 745.939$ $W_{3.} = 0.3307$
Total		$N_{.1} = 221,\ n_{.1} = 88$ $\bar{Y}_{.1} = 352.6515$ $W_{.1} = 0.3399$	$N_{.2} = 209,\ n_{.2} = 84$ $\bar{Y}_{.2} = 446.01912$ $W_{.2} = 0.32153$	$N_{.3} = 220,\ n_{.3} = 88$ $\bar{Y}_{.3} = 553.3727$ $W_{.3} = 0.3384$	$N = 650$ $n = 260$ $\bar{Y} = 450.828$ $S^2 = 65530.35$

Table 7.3 (for data set III)

B \ A		Attribute A: Low	Attribute A: Medium	Attribute A: High	Total
Attribute B	Low	$N_{11} = 56,\ n_{11} = 22$ $\bar{Y}_{11} = 37.232$ $W_{11} = 0.1143$ $S_{11}^2 = 504.1068$	$N_{12} = 50,\ n_{12} = 20$ $\bar{Y}_{12} = 113.02$ $W_{12} = 0.102$ $S_{12}^2 = 434.947$	$N_{13} = 52,\ n_{13} = 21$ $\bar{Y}_{13} = 188.8846$ $W_{13} = 0.1061$ $S_{13}^2 = 505.163$	$N_{1.} = 158$ $n_{1.} = 63$ $\bar{Y}_{1.} = 111.1265$ $W_{1.} = 0.3224$
	Medium	$N_{21} = 48,\ n_{21} = 19$ $\bar{Y}_{21} = 267.3333$ $W_{21} = 0.09796$ $S_{21}^2 = 500.926$	$N_{22} = 62,\ n_{22} = 25$ $\bar{Y}_{22} = 321.258$ $W_{22} = 0.1265$ $S_{22}^2 = 768.06$	$N_{23} = 58,\ n_{23} = 23$ $\bar{Y}_{23} = 413.776$ $W_{23} = 0.11836$ $S_{23}^2 = 466.716$	$N_{2.} = 168$ $n_{2.} = 67$ $\bar{Y}_{2.} = 337.792$ $W_{2.} = 0.34282$
	High	$N_{31} = 54,\ n_{31} = 22$ $\bar{Y}_{31} = 483.037$ $W_{31} = 0.1102$ $S_{31}^2 = 564.56$	$N_{32} = 60,\ n_{32} = 24$ $\bar{Y}_{32} = 553.666$ $W_{32} = 0.12245$ $S_{32}^2 = 529.17$	$N_{33} = 50,\ n_{33} = 20$ $\bar{Y}_{33} = 625.000$ $W_{33} = 0.102$ $S_{33}^2 = 562.53$	$N_{3.} = 164$ $n_{3.} = 66$ $\bar{Y}_{3.} = 552.1583$ $W_{3.} = 0.33465$
Total		$N_{.1} = 158,\ n_{.1} = 63$ $\bar{Y}_{.1} = 259.4999$ $W_{.1} = 0.32246$	$N_{.2} = 172,\ n_{.2} = 69$ $\bar{Y}_{.2} = 341.796$ $W_{.2} = 0.35095$	$N_{.3} = 160,\ n_{.3} = 64$ $\bar{Y}_{.3} = 406.694$ $W_{.3} = 0.32646$	$N = 490$ $n = 196$ $\bar{Y} = 336.451$ $S^2 = 36084.05$

(a) Using parameters stated in the above tables, one can easily calculate:

Table 7.4

Estimators	DATA SET I	DATA SET II	DATA SET III
$\left[\left(\bar{y}_{gdps}\right)_1\right]$	MSE 81.2278 Bias 3.1003	MSE 118.45 Bias 9.012	MSE 89.29 Bias 6.7285
$\left[\left(\bar{y}_{gdps}\right)_0\right]$	MSE 78.087 Bias 6.906	MSE 99.2869 Bias 8.863	MSE 75.0276 Bias 6.5878
$\left[\left(\bar{y}_{gdps}\right)_{1/2}\right]$	MSE 9.8229 Bias 3.004	MSE 81.2278 Bias 3.1003	MSE 6.44 Bias 1.76
$\left[\left(\bar{y}_{gdps}\right)_{opt}\right]$ α_{opt}	MSE 9.8141 Bias 2.2463 0.4943	MSE 81.2278 Bias 3.1003 0.4773	MSE 5.70 Bias 1.655 0.4767
$V(\bar{y})$	116.11042	151.21785	110.4605
$V(\bar{y}_{ps})$	6.8987	2.06107	1.7247

(b) It seems that estimator $\left(\bar{y}_{gdps}\right)_{1/2}$ is more efficient than $\left(\bar{y}_{gdps}\right)_0$ and $\left(\bar{y}_{gdps}\right)_1$ both on these data sets. It is because the value $\alpha = (1/2)$ is very close to its optimum choice.

(c) The estimator $\left(\bar{y}_{gdps}\right)$ has made possible to estimate $\bar{Y}$ in a r x s set-up even without the prior knowledge of W_{ij} and frames. It has an effective utilization of row and column proportions $W_{i.}$ And $W_{.j}$

(d) The proposed estimator is found most efficient at optimal selection of α = 0.4943 for set – I, α = 0.4727 for set – II, α = 0.4767 for set – III. The gain in efficiency is almost double to the usual mean estimator.

(e) Although $\bar{y}_{ps}$ is most efficient in above data than all other estimators but it needs a knowledge of W_{ij} which is assumed absent. So, there is no need to compare the proposed estimators with $\bar{y}_{ps}$.

(f) The gain in efficiency of estimator to be compared.

Table 7.5

% Gain	DATA SET I	DATA SET II	DATA SET III
$\left[\left(\bar{y}_{gdps}\right)_1\right]$	30.0%	21.7%	19.1%
$\left[\left(\bar{y}_{gdps}\right)_0\right]$	32.7%	34.3%	32.0%
$\left[\left(\bar{y}_{gdps}\right)_{1/2}\right]$	91.5%	98.0%	95.0%
$\left[\left(\bar{y}_{gdps}\right)_{opt}\right]$	91.5%	98.1%	98.5%

(g) On the basis of data considered herein, one can think of choosing α to a value near to 0.5 which reveals that almost a fifty percent fraction of row-sum of size-proportions $[p_{i.} + W_{i.}]$ and rest fifty percent same from column generates an ideal, quick and easy choice of α which produces a high gain in efficiency. Thus, the proposed estimator is easy in selecting optimum value of α.

8. Acknowledgement

Authors are very much thankful to the referees for their valuable comments and suggestions. The proposed suggestions were very helpful to improve the quality of research manuscript.

REFERENCES

AGARWAL, M.C., PANDA, K.B. (1993): "An efficient estimator in post-stratification", Matron, Vol. 5, 3.4 pp. 179—187.

BRYANT,E.C. (1955): An Analysis of some two-way stratification, Metron, vol. 5,3—4, pp-179—187.

BRYANT, E.C., HARTLEY, H.O., JESSEN, R.J. (1960): Design and Estimation in two-way stratification, Jour. Amer. Stat. Assoc.,55,105—124.

HOLT, D., SMITH, T.M.F. (1979): "Post — stratification", J.R. Statistical Society, A, 142, 33—36.

JAGERS, P., ODEN, A., TRULSSON, L. (1985): "Post — stratification and ratio estimation", International Statistical Review, 53, 221—238.

JAGERS, P. (1986): "Post stratification against bias in sampling", International Statistical Review, 54, 159—167.

SMITH, T.M.F.(1991): Post-stratification, The Statistician , 40 , 315—323.

SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S. and ASOK, C. (1984) "Sampling Theory of Surveys with Applications", Iowa State University Press, Indian society of Agricultural Statistics Publication, New Delhi.

STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 331—350

SMALL AREA ESTIMATION FOR SPATIALLY CORRELATED POPULATIONS — A COMPARISON OF DIRECT AND INDIRECT MODEL — BASED METHODS

Hukum Chandra[1], Nicola Salvati[2] and Ray Chambers[3]

ABSTRACT

Linear mixed models underpin many small area estimation (SAE) methods. In this paper we investigate SAE based on linear models with spatially correlated small area effects where the neighbourhood structure is described by a contiguity matrix. Such models allow efficient use of spatial auxiliary information in SAE. In particular, we use simulation studies to compare the performances of model-based direct estimation (MBDE) and empirical best linear unbiased prediction (EBLUP) under such models. These simulations are based on theoretically generated populations as well as data obtained from two real populations (the ISTAT farm structure survey in Tuscany and the US Environmental Monitoring and Assessment Program survey). Our empirical results show only marginal gains when spatial dependence between areas is incorporated into the SAE model.

Keywords: Empirical Best Linear Unbiased Prediction, Spatial Models, Spatial EBLUP, Model-Based Direct Estimation

1. Introduction

Estimation of population characteristics for sub-national domains (or smaller regions) is an important objective for statistical surveys. In particular, geographically defined domains, e.g. regions, states, counties, wards and

[1] Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: hchandra@soton.ac.uk
[2] Dipartimento di Statistica e Matematica Applicata all'Economia, University of Pisa, Pisa, Italy, Email: salvati@ec.unipi.it
[3] Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW, 2522, Australia. Email: ray@uow.edu.au

metropolitan areas can be of interest. Estimates for these domains based on the usual design-based approach to survey sampling inference are typically referred to as direct estimates in the literature. However, sample sizes are typically small (or even zero) within the domains/areas of interest, leading to large sampling variability for these direct estimators. An alternative approach that is now widely used in small area estimation is the so-called indirect or model-based approach. This uses auxiliary information for the small areas of interest and has been characterised in the statistical literature as 'borrowing strength' from the relationship between the target variables and the auxiliary information. A flexible and popular way of borrowing strength is based on the use of linear mixed models with area specific random effects, with estimation and inferences typically carried out using empirical best linear unbiased prediction (EBLUP - see Rao, 2003). An alternative approach, discussed in Chandra and Chambers (2005), is based on the use of model-based direct estimation (MBDE) within the small areas. In this case an estimate for a small area of interest corresponds to a weighted linear combination of the sample data for that area, with weights based on a population level version of the linear mixed model. These weights 'borrow strength' via this model, which includes random area effects. Provided the assumed small area model is true, the EBLUP is asymptotically the most efficient estimator for a particular small area. In practice however the 'true' model for the data is unknown and the EBLUP can be inefficient under misspecification. In such circumstances, Chandra and Chambers (2005) note that MBDE offers an alternative to potentially unstable EBLUP. In particular, MBDE is easy to implement, produces sensible estimates when the sample data exhibit patterns of variability that are inconsistent with the assumed model (e.g. contain too many zeros) and generates robust MSE estimates.

Model-based methods of small area estimation (SAE) are often based on assuming a linear mixed model, with area-specific random effects to account for between area variation beyond that explained by auxiliary variables included in the fixed part of the model. Although it is customary to assume that these random area effects are independent, in practice most small area boundaries are arbitrary and there appears to be no good reason why population units just one side of such a boundary should not generally be correlated with population units just on the other side. In particular, it is often reasonable to assume that the effects of neighbouring areas (defined, for example, by a contiguity criterion) are correlated, with the correlation decaying to zero as the distance between these areas increases (Pratesi and Salvati, 2005). That is, small area models should allow for spatial correlation of area random effects. See Cressie (1991).

In this paper we consider linear unit level small area models (Battese *et al.*, 1988) and we extend MBDE and EBLUP for SAE to account for spatial correlation between the small areas. We then contrast the performance of these two approaches via empirical studies. Our aim in doing so is to explore how much efficiency is gained by incorporating spatial correlation into SAE. The paper is

organized as follows: In section 2 we review MBDE and EBLUP for SAE under random effect models with spatially independent area effects and discuss the extensions of these techniques to account for spatial dependence between the small areas. We define the resulting estimators for the small area means and their mean squared error estimators. In Section 3 we describe the design of our simulation studies and present empirical results. Besides using simulated population and sample data, we use two real data sets from the ISTAT farm structure survey (farm data) in Northern Tuscany and the Environmental Monitoring and Assessment Program (EMAP) survey of lakes in the north-east of the USA. Finally, in section 4 we provide some concluding remarks and identify further avenues of research.

2. Small Area Estimation under Linear Models with Random Area Effects

2.1. Models with Spatially Independent Effects

To begin, let y_{ij} denote the value of the variable of interest for the jth ($j = 1,\ldots,N_i$) unit in small area *i* ($i = 1,\ldots,m$) and let X_{ij} denote the vector of values of the *p* unit level auxiliary variables associated with this unit. We consider a nested error regression model of form

$$y_{ij} = X_{ij}^t \beta + u_i + e_{ij} \tag{1}$$

where β is a vector of *p* unknown fixed effects, u_i is the random area effect associated with small area *i*, assumed to have mean zero and variance σ_u^2, and e_{ij} is an individual level random error with mean zero and variance σ_e^2. The two error terms are assumed to be mutually independent, both across individuals as well as across areas. In addition, it is often assumed that they are normally distributed. In matrix notation, (1) is expressed as

$$Y_i = X_i \beta + u_i 1_{N_i} + e_i \tag{2}$$

where $Y_i = (y_{i1},....,y_{iN_i})^t$, $X_i = (X_{i1},...,X_{iN_i})^t$ is a $N_i \times p$ matrix and $e_i = (e_{i1},...,e_{iN_i})^t$. Here N_i is the number of population units in small area *i*. The covariance matrix of Y_i is $Var(Y_i) = V_i = \sigma_e^2 I_{N_i} + \sigma_u^2 1_{N_i} 1_{N_i}^t$, which depends on the vector $\theta = (\sigma_u^2, \sigma_e^2)^t$ of variance components of the model. Here 1_{N_i} is the unit vector of length N_i and I_{N_i} is the identity matrix of order N_i. Assuming (2)

holds, the population mean of Y in area i is $\bar{Y}_i = \bar{X}_i\beta + u_i + \bar{e}_i$, where $\bar{X}_i = N_i^{-1}\sum_{j=1}^{N_i} x_j$ is assumed known.

Grouping the area-specific models (2) over the population leads to the population level model

$$Y = X\beta + Zu + e \tag{3}$$

where $Y = (Y_1^t,\ldots,Y_m^t)^t$, $X = (X_1^t,\ldots,X_m^t)^t$, $Z = diag(Z_i = 1_{N_i}; 1 \le i \le m)$, $u = (u_1,\ldots,u_m)^t$ and $e = (e_1^t,\ldots,e_m^t)^t$. Since different areas are independent, the covariance matrix of Y has block diagonal structure given by $V = diag(V_i; 1 \le i \le m)$. We assume that X has full column rank p. In practice the variance components that define V are unknown and can be estimated from the sample data using methods described, for example, in Harville (1977). We denote these estimates by $\hat{\theta} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)^t$ and put a 'hat' on any quantity where these estimates are substituted for actual values. Thus $\hat{V} = diag(\hat{V}_i; 1 \le i \le m)$, with $\hat{V}_i = \hat{\sigma}_e^2 I_{N_i} + \hat{\sigma}_u^2 Z_i Z_i^t$.

Now consider the decomposition of Y, X, Z and V into sample and non-sample components so that X_s is the $n \times p$ matrix of sample values of the auxiliary variables, Z_s is the corresponding $n \times m$ matrix of sample components of Z and V_{ss} is the $n \times n$ covariance matrix associated with the n sample units that make up the $n \times 1$ sample vector Y_s. A subscript of r is used to denote corresponding quantities defined by the $N - n$ non-sample units, with V_{rs} denoting the $(N-n) \times n$ matrix defined by $Cov(Y_r, Y_s)$. In what follows we use 1_N, 1_s and 1_r to denote vectors of 1s of dimension N, n and $N - n$ respectively, with I_N, I_s and I_r denoting identity matrices of the same order. We use similar notation to denote restriction to small area level by introducing an extra subscript of i to denote the small area. For example, s_i corresponds to the set of n_i sample units in area i, r_i the corresponding set of $N_i - n_i$ non-sampled units, with associated variances and covariances $V_{iss} = \sigma_e^2 I_{is} + \sigma_u^2 Z_{is} Z_{is}^t$ and $V_{isr} = \sigma_u^2 Z_{is} Z_{ir}^t$.

Assuming (3) holds, the empirical best linear unbiased predictor (EBLUP) for the ith small area mean $\bar{Y}_i$ is

$$\hat{\bar{Y}}_{i,EBLUP} = f_i \bar{Y}_{is} + (1 - f_i)\left\{\bar{X}_{ir}^t \hat{\beta} + \hat{\gamma}_i(\bar{Y}_{is} - \bar{X}_{is}^t \hat{\beta})\right\} \tag{4}$$

where $\hat{\beta} = \left(\sum_i X_{is}^t \hat{V}_{iss}^{-1} X_{is}\right)^{-1} \left(\sum_i X_{is}^t \hat{V}_{iss}^{-1} Y_{is}\right)$ is the empirical best linear unbiased estimator (EBLUE) of β, $f_i = N_i^{-1} n_i$, $\hat{\gamma}_i = \hat{\sigma}_u^2 \left(\hat{\sigma}_u^2 + n_i^{-1} \hat{\sigma}_e^2\right)^{-1}$ is the shrinkage factor, $\bar{Y}_{is} = n_i^{-1} \sum_{s_i} y_j$ and $\bar{X}_{is} = n_i^{-1} \sum_{s_i} x_j$ are the sample means of *Y* and *X* for area *i*, while $\bar{X}_{ir} = (N_i - n_i)^{-1} (N_i \bar{X}_i - n_i \bar{X}_{is})$ is the corresponding mean of *X* for the $N_i - n_i$ non-sampled units in the area. An approximately unbiased estimator of the MSE of (4) under (3) is

$$M(\hat{\bar{Y}}_{i,EBLUP}) = (1 - f_i)^2 \left\{ g_{1i}(\hat{\theta}) + g_{2i}(\hat{\theta}) + 2 g_{3i}(\hat{\theta}) \right\} + N_i^{-1} (1 - f_i) \hat{\sigma}_e^2 \qquad (5)$$

where

$$g_{1i}(\hat{\theta}) = \hat{\sigma}_u^2 \left(1 - \hat{\sigma}_u^2 Z_{is}^t \hat{V}_{iss}^{-1} Z_{is}\right)$$

$$g_{2i}(\hat{\theta}) = \left(\bar{X}_{ir} - \hat{c}_i^t X_{is}\right)^t \left(\sum_i X_{is}^t \hat{V}_{iss}^{-1} X_{is}\right)^{-1} \left(\bar{X}_{ir} - \hat{c}_i^t X_{is}\right)$$

$$g_{3i}(\hat{\theta}) = tr \left\{ (\nabla \hat{c}_i) V_{iss} (\nabla \hat{c}_i)^t \hat{C}(\hat{\theta}) \right\}$$

with $\hat{c}_i = \hat{\sigma}_u^2 \hat{V}_{iss}^{-1} Z_{is}$, $\nabla \hat{c}_i = \partial \hat{c}_i / \partial \hat{\theta} = \left(\partial \hat{c}_i / \partial \hat{\sigma}_u^2, \partial \hat{c}_i / \partial \hat{\sigma}_e^2\right)^t$ and $\hat{C}(\hat{\theta})$ is the estimated asymptotic covariance matrix of $\hat{\theta}$ (i.e. the inverse of the observed information matrix for θ). For more details see Rao (2003, pp. 107-110).

Under the population level linear mixed model (3), the sample weights that define the EBLUP for the population total of *Y* are

$$w_{EBLUP} = (w_{j,EBLUP}) = 1_s + \hat{H}^t \left(X^t 1_N - X_s^t 1_s\right) + \left(I_s - \hat{H}^t X_s^t\right) \hat{V}_{ss}^{-1} \hat{V}_{sr} 1_r \qquad (6)$$

where $\hat{H} = \left(\sum_i X_{is}^t \hat{V}_{iss}^{-1} X_{is}\right)^{-1} \left(\sum_i X_{is}^t \hat{V}_{iss}^{-1}\right)$. See Royall (1976). The model-based direct estimator (MBDE, see Chambers and Chandra, 2006) of the ith small area mean is then defined as

$$\hat{\bar{Y}}_{i,MBD} = \sum_{j \in s_i} w_{j,EBLUP} y_j \Big/ \sum_{j \in s_i} w_{j,EBLUP} \,. \qquad (7)$$

A robust estimator (Chandra and Chambers, 2005; Royall and Cumberland, 1978) of the mean squared error of the MBDE (7) is

$$M(\hat{\bar{Y}}_{i,MBD}) = v(\hat{\bar{Y}}_{i,MBD}) + \left\{ b(\hat{\bar{Y}}_{i,MBD}) \right\}^2 \qquad (8)$$

where $v(\hat{\bar{Y}}_{i,MBD}) = \sum_{s_i} \lambda_j (y_j - x_j^t \hat{\beta})^2$, with $\lambda_j = N_i^{-2} \left\{ a_j^2 + (N_i - n_i)(n_i - 1)^{-1} \right\}$ and $a_j = \left(\sum_{s_i} w_k \right)^{-1} \left(N_i w_j - \sum_{s_i} w_k \right)$, is the estimate of the prediction variance of the MBDE, and $b(\hat{\bar{Y}}_{i,MBD}) = (\hat{\bar{X}}_{i,MBD} - \bar{X}_i)^t \hat{\beta}$ is the estimate of its prediction bias. Here $\hat{\bar{X}}_{i,MBD}$ denotes the weighted average of the sample values of the auxiliary variables in area *i* based on the EBLUP weights (6).

2.2. Models with Spatial Dependence

In order to take into account the correlation between neighbouring areas we consider the use of spatial models for random area effects (Cressie, 1991). In particular, we consider a linear regression model with spatial dependence in the error structure. In particular, we assume a Simultaneous Autoregressive (SAR) error process (Anselin, 1992), where the vector of random area effects $v = (v_i)$ satisfies

$$v = \rho W v + u \,. \tag{9}$$

Here ρ is a spatial autoregressive coefficient, W is a proximity matrix of order m and $u \sim N(0, \sigma_u^2 I_m)$. Since $v = (I - \rho W)^{-1} u$ with $E(u) = 0$ and $Var(u) = \sigma_u^2 I_m$, we have $E(v) = 0$ and $Var(v) = \sigma_u^2 [(I_m - \rho W)(I_m - \rho W^t)]^{-1} = G$. The W matrix describes how random effects from neighbouring areas are related, whereas ρ defines the strength of this spatial relationship. The simplest way to define W is as a contiguity matrix. That is, the elements of W take non-zero values only for those pairs of areas that are adjacent. Generally, for ease of interpretation, this matrix is defined in row-standardized form; in which case ρ is called the spatial autocorrelation parameter (Banerjee *et al.*, 2004). Formally, the element w_{jk} of a contiguity matrix takes the value 1 if area *j* shares an edge with area *k* and 0 otherwise. In row-standardised form this becomes

$$w_{jk} = \begin{cases} d_j^{-1} & \textit{if } j \textit{ and } k \textit{ are contiguous} \\ 0 & \textit{otherwise} \end{cases}$$

where d_j is the total number of areas that share an edge with area *j* (including area *j* itself). Contiguity is the simplest but not necessarily the best specification of a spatial interaction matrix. It may be more informative to express this interaction in a more detailed way, e.g. as some function of the length of shared border between neighbouring areas or as a function of the distance between

certain locations in each area. Furthermore, the concept of neighbours of a particular area can be defined not just in terms of contiguous areas, but also in terms of all areas within a certain radius of the area of interest. In the empirical evaluations reported later in this paper, however, we used simple contiguity (row-standardized) to define the spatial interaction between different areas.

In order to define the EBLUP in this situation, we replace (3) by a linear mixed model of form

$$Y = X\beta + Zv + e\,. \tag{10}$$

Here the vector v is an m-vector of spatially correlated area effects that satisfy the SAR model (9), with $Var(e) = \sigma_e^2 I_N$ and $Var(v) = G$. This model can then be rewritten as

$$Y = X\beta + Z(I - \rho W)^{-1}u + e\,. \tag{11}$$

It follows that the covariance matrix of Y is $Var(Y) = V = \sigma_e^2 I_N + ZGZ^t$. In practice, the vector of parameters $\theta = (\sigma_u^2, \sigma_e^2, \rho)^t$ is unknown. Replacing it with an asymptotically consistent estimator $\hat{\theta} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\rho})^t$, and assuming that (11) holds, the spatial-EBLUP (SEBLUP) for the ith small area mean $\bar{Y}_i$ is

$$\hat{\bar{Y}}_{i,SEBLUP} = f_i \bar{Y}_{is} + (1 - f_i)\left(\bar{X}_{ir}^t \hat{\beta}^s + m_i^t \hat{v}\right) \tag{12}$$

where $\hat{\beta}^s = \left(X_s^t \hat{V}_{ss}^{-1} X_s\right)^{-1}\left(X_s^t \hat{V}_{ss}^{-1} Y_s\right)$ is the empirical BLU estimator of β under (11), m_i is the m-vector $(0,0,0,\ldots,1,\ldots,0,0)^t$ with the 1 in the ith position and

$$\hat{v} = \hat{G} Z_s^t \hat{V}_{ss}^{-1}\left(Y_s - X_s \hat{\beta}^s\right).$$

Here $\hat{G} = \hat{\sigma}_u^2\left[(I_m - \hat{\rho}W)(I_m - \hat{\rho}W^t)\right]^{-1}$ and $\hat{V}_{ss} = \hat{\sigma}_e^2 I_n + Z_s \hat{\sigma}_u^2 [(I_m - \hat{\rho}W)(I_m - \hat{\rho}W^t)]^{-1} Z_s^t$. When all random effects are normally distributed, the parameter vector θ can be estimated via maximum likelihood (ML) as well as restricted maximum likelihood (REML) (Pratesi and Salvati, 2005; Singh *et al.*, 2005; Petrucci and Salvati 2006). Numerical approximations to either the ML or REML estimators $\hat{\sigma}_u^2, \hat{\sigma}_e^2$ and $\hat{\rho}$ can be obtained via a two-step procedure. At the first step, the Nelder-Mead algorithm (Nelder and Mead, 1965) is used to approximate these estimates. The second step then uses these approximations as starting values for a Fisher scoring algorithm. This is necessary because the log-likelihood function has multiple local maxima (Pratesi and Salvati, 2005). In empirical studies reported in Section 3 we carried

out parameter estimation via REML using the *lme* function in the *R* environment (Bates and Pinheiro, 1998).

Following the same approach as in Prasad and Rao (1990), an approximately unbiased estimator of the MSE of the SEBLUP (12) is given by

$$M(\hat{\bar{Y}}_{i,SEBLUP}) = (1-f_i)^2 \left\{ g_{1i}^{(s)}(\hat{\theta}) + g_{2i}^{(s)}(\hat{\theta}) + 2g_{3i}^{(s)}(\hat{\theta}) \right\} + N_i^{-1}(1-f_i)\hat{\sigma}_e^2 \quad (13)$$

where

$$g_{1i}^{(s)}(\hat{\theta}) = m_i^t \left(\hat{G} - \hat{G} Z_s^t \hat{V}_{ss}^{-1} \hat{G}_s \right) m_i$$

$$g_{2i}^{(s)}(\hat{\theta}) = \left(\bar{X}_{ir} - \hat{c}_i^t X_s \right)^t \left(X_s^t \hat{V}_{ss}^{-1} X_s \right)^{-1} \left(\bar{X}_{ir} - \hat{c}_i^t X_s \right)$$

$$g_{3i}^{s}(\hat{\theta}) = tr \left\{ (\nabla \hat{c}_i) V_{ss} (\nabla \hat{c}_i)^t \hat{C}(\hat{\theta}) \right\}$$

with $\hat{c}_i = \hat{V}_{ss}^{-1} Z_s \hat{G} m_i$ and $\nabla \hat{c}_i = \partial \hat{c}_i / \partial \hat{\theta} = \left(\partial \hat{c}_i / \partial \hat{\sigma}_u^2, \partial \hat{c}_i / \partial \hat{\sigma}_e^2, \partial \hat{c}_i / \partial \hat{\rho} \right)^t$. Here, after dropping 'hats' for the sake of clarity,

$$\begin{aligned} \frac{\partial c_i}{\partial \sigma_u^2} &= \frac{\partial V_{ss}^{-1} Z_s G}{\partial \sigma_u^2} m_i = \left\{ V_{ss}^{-1} Z_s \left(\frac{\partial G}{\partial \sigma_u^2} \right) + \left(\frac{\partial V_{ss}^{-1}}{\partial \sigma_u^2} \right) Z_s G \right\} m_i \\ &= \left\{ V_{ss}^{-1} Z_s \left(\frac{\partial G}{\partial \sigma_u^2} \right) - \left(V_{ss}^{-1} \frac{\partial V_{ss}}{\partial \sigma_u^2} V_{ss}^{-1} \right) Z_s G \right\} m_i \\ &= V_{ss}^{-1} Z_s D \left\{ I_s - Z_s^t V_{ss}^{-1} Z_s G \right\} m_i \end{aligned}$$

where

$$D = \frac{\partial G}{\partial \sigma_u^2} = \frac{\partial \sigma_u^2 [(I_m - \rho W)(I_m - \rho W^t)]^{-1}}{\partial \sigma_u^2} = [(I_m - \rho W)(I_m - \rho W^t)]^{-1}$$

and

$$\frac{\partial V_{ss}}{\partial \sigma_u^2} = \frac{\partial \left\{ \sigma_e^2 I_s + Z_s \sigma_u^2 [(I_m - \rho W)(I_m - \rho W^t)]^{-1} Z_s^t \right\}}{\partial \sigma_u^2} = Z_s D Z_s^t .$$

Similarly

$$\frac{\partial c_i}{\partial \sigma_e^2} = \frac{\partial V_{ss}^{-1} Z_s G}{\partial \sigma_e^2} m_i = V_{ss}^{-1} Z_s \left(\frac{\partial G}{\partial \sigma_e^2} \right) m_i + \left(\frac{\partial V_{ss}^{-1}}{\partial \sigma_e^2} \right) Z_s G m_i$$

$$= \left(-V_{ss}^{-1} \frac{\partial V_{ss}}{\partial \sigma_e^2} V_{ss}^{-1} \right) Z_s G m_i$$

$$= -\sigma_u^2 \left\{ V_{ss}^{-1} I_s V_{ss}^{-1} \right\} Z_s D m_i$$

since $G = \sigma_u^2 D$ and $\frac{\partial V_{ss}}{\partial \sigma_e^2} = I_s$. Finally

$$\frac{\partial c_i}{\partial \rho} = \frac{\partial V_{ss}^{-1} Z_s G}{\partial \rho} m_i = V_{ss}^{-1} Z_s \left(\frac{\partial G}{\partial \rho} \right) m_i + \left(\frac{\partial V_{ss}^{-1}}{\partial \rho} \right) Z_s G m_i$$

$$= \left\{ V_{ss}^{-1} Z_s A - V_{ss}^{-1} \left(\frac{\partial V_{ss}^{-1}}{\partial \rho} \right) V_{ss}^{-1} Z_s G \right\} m_i$$

$$= V_{ss}^{-1} Z_s A \left\{ I_s - Z_s^t V_{ss}^{-1} Z_s G \right\} m_i \ .$$

Here $\frac{\partial V_{ss}}{\partial \rho} = \frac{\partial \left\{ \sigma_e^2 I_s + Z_s G Z_s^t \right\}}{\partial \rho} = Z_s \left(\frac{\partial G}{\partial \rho} \right) Z_s^t = Z_s A Z_s^t$, with

$$A = \frac{\partial G}{\partial \rho} = \sigma_u^2 \left(\frac{\partial D}{\partial \rho} \right) = -\sigma_u^2 \left(D \frac{\partial D^{-1}}{\partial \rho} D \right) = -2\sigma_u^2 D \left(\rho W W^t - W \right) D$$

since $\frac{\partial D^{-1}}{\partial \rho} = \frac{\partial [(I - \rho W)(I - \rho W^t)]}{\partial \rho} = 2\left(\rho W W^t - W \right)$. We note that $\hat{C}(\hat{\theta}) = I^{-1}(\hat{\theta})$ is still the estimated asymptotic covariance matrix of $\hat{\theta}$ defined by the inverse of the information matrix $I(\hat{\theta})$ (Rao, 2003, pp. 107-110), with the $(i, j)^{th}$ element of $I(\hat{\theta})$ given by $\frac{1}{2} tr \left\{ P \left(\frac{\partial V_{ss}}{\partial \theta_i} \right) P \left(\frac{\partial V_{ss}}{\partial \theta_j} \right) \right\}$ with $P = V_{ss}^{-1} \left\{ I_s - X_s \left(X_s^t V_{ss}^{-1} X_s \right)^{-1} X_s^t V_{ss}^{-1} \right\}$.

Turning now to implementation of model-based direct estimation under (11) we note that the EBLUP sample weights (6) depend on the structure of the random area effects in the mixed model (3) only via the their sample and population covariance structure. Consequently, extension to more complex covariance structures requires only that $\hat{V}_{ss}^{-1}$ and $\hat{V}_{sr}$ be recomputed under these

more complex models. When (11) holds, the corresponding spatial EBLUP weights $w_{SEBLUP} = (w_{j,SEBLUP})$ are therefore still given by (6), but where now $\hat{V}_{ss}^{-1} = \left\{\hat{\sigma}_e^2 I_s + Z_s \hat{\sigma}_u^2 [(I_m - \hat{\rho}W)(I_m - \hat{\rho}W^t)]^{-1} Z_s^t\right\}^{-1}$ and $\hat{V}_{sr} = \hat{\sigma}_u^2 Z_s [(I_m - \hat{\rho}W)(I_m - \hat{\rho}W^t)]^{-1} Z_r^t$. The spatial-MBDE (denoted by SMBDE) of the ith small area mean $\bar{Y}_i$ and the corresponding estimator of its mean squared error are then given by (7) and (8) respectively, with the weights (6) used there replaced by the spatial EBLUP weights w_{SEBLUP} defined above.

3. Empirical Evaluations

In this section we use simulation to illustrate the performance of the four different methods of SAE discussed in the previous section. These are the EBLUP and MBDE under the linear mixed model (3) with spatially independent area effects (see section 2.1) and the SEBLUP and SMBDE under the linear mixed model (10) with spatially dependent area effects (see section 2.2). We computed three measures of estimation performance using the estimates generated in our simulations. These are the relative bias (RB) and the relative root mean squared error (RRMSE), both expressed as percentages, of estimates of the small area means and the coverage rate of nominal 95 per cent confidence intervals for these means (for more details see Chandra and Chambers, 2005).

We carried out two types of simulation studies. The first used real data and design-based simulation to evaluate the performance of these methods in the context of a real population and realistic sampling methods. The second used model-based simulation to generate artificial populations, from which samples were then taken. The sample data obtained in each case were then used to contrast the performance of different methods of small area estimation. The populations underpinning the design-based simulations were based on two different data sets:

(i) *The ISTAT farm structure survey*. This is a sample of 529 farms from the farm structure survey in Tuscany carried out by ISTAT. Here we used these sample farms to generate a population of N = 22977 farms by sampling with replacement from the original sample of 529 farms with probabilities proportional to their sample weights. We drew 1000 independent stratified random samples from this (fixed) population, with total sample size in each draw equal to the original sample size (529) and with the small areas of interest defined by the 23 Local Economy Systems (LESs) of the North Tuscany region. Sample sizes within these areas were fixed to be the same as in the original sample. Note that these varied from 4 to 48. Our aim was to estimate average olive production (quintals) in each LES using utilized olive surface (hectares) as the auxiliary variable. The results from this simulation are set out in Tables 1 and 2.

(ii) *The Environmental Monitoring and Assessment Program (EMAP) survey*. The data, on which this population was based, was provided by the Space-Time Aquatic Resources Modelling and Analysis Program (STARMAP) at Colorado State University. It consists of 551 measurements, taken between 1991 and 1996, from a sample of 349 of the 21,026 lakes located in the north-eastern United States. Here we define lakes grouped by 6-digit Hydrologic Unit Code (HUC) as our small areas of interest. Since three HUCS had sample sizes of one, these were combined with adjacent HUCS, leading to a total of 23 small areas. Sample sizes in these 23 areas varied from 2 to 45. A (fixed) population of size N = 21028 was then defined by sampling N times with replacement from the sample of 349 units, with probability proportional to a unit's sample weight. A total of 1000 independent stratified random samples of the same size as the original sample were selected from this simulated population, with HUC sample sizes fixed to be the same as in the original sample. The survey variable Y in this case was the Acid Neutralizing Capacity (ANC) of a lake - an indicator of the acidification risk of water bodies in water resource surveys - with elevation of the lake as the auxiliary variable X. Results from this simulation are set out in Tables 3 and 4.

In our model-based simulations we again used the data from the EMAP survey, but this time based the population model underlying our simulations on variance components obtained by fitting a linear mixed effects model to these data. In particular, we generated a population of size $N = 21028$, with the same small area (HUC) population sizes as before. We used a sample size $n = 349$ and constrained the small area sample sizes to be the same as in the EMAP survey. These population and sample sizes were kept fixed in all our simulations. The model used to generate the population corresponded to a nested error regression model with random area effects for neighbouring areas distributed according to a SAR spatial correlation structure. This was of the form

$$y_{ij} = 1000 - 3x_{ij} + v_i + e_{ij}$$

where the x_{ij} values were generated from the uniform distribution on [10, 700], $v = (v_i) = (I_m - \rho W)^{-1} u$ was an m-vector of spatially correlated area effects with $u = (u_i)$ an m-vector of independent realisations from $N(0, \sigma_u^2)$ and the e_{ij} were individual error terms distributed as $N(0, \sigma_e^2)$. Using estimates derived from the linear mixed model fitted to the original EMAP survey data, we put $\sigma_u^2 = 265000$ and $\sigma_e^2 = 125000$, with intra area effect, $\gamma = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2) = 0.68$. The row standardised SAR neighbourhood structure matrix W used in the simulations was kept fixed and corresponded to contiguous HUCs in the EMAP survey data set. Population data for four values of ρ (0.05,

0.25, 0.50 and 0.75) were generated and simple random samples selected from each small area, with a total of 1000 combinations independently simulated. The results from this simulation are set out in Table 5.

Table 1 shows the relative bias and relative root mean squared error for small area estimates calculated using the four different methods of small area estimation (EBLUP, MBDE, SEBLUP and SMBDE) based on repeated sampling from the simulated Northern Tuscany population. Corresponding coverage rates for nominal 95% intervals for area means generated by these four approaches are set out in Table 2. Tables 3 and 4 show analogous results for repeated sampling from the simulated EMAP population. Note that the estimated value of the spatial autocorrelation parameter ρ in the original ISTAT farm survey data was quite small ($\hat{\rho} = 0.025$), while for the EMAP survey data this estimate was considerably larger ($\hat{\rho} = 0.50$). Table 5 shows the average values of relative bias and relative root mean squared error, both expressed in percentage terms, and average coverage rates for the different methods generated under the model based simulations. All averages in Table 5 are over the 23 small areas of interest.

The results set out in Table 1 show that both EBLUP and SEBLUP are very unstable in a few small areas (e.g. regions 3, 6 and 14), due mainly to there being little or no variability in the variable of interest in these areas. In such situations, the SEBLUP seems to perform worse than the EBLUP. In contrast, the MBDE and SMBDE methods appear unaffected by such behaviour. Since the average values of performance measures are influenced by outlying estimates, we compare different methods using the median values of their area-specific performance measures. From Table 1 we see that the median relative bias of MBDE is smaller than that of EBLUP. In contrast, the median relative root mean squared error of EBLUP is smaller than that of MBDE. The median relative bias and median relative root mean squared error of SEBLUP is marginally smaller than that of EBLUP. However, these values are almost same for MBDE and SMBDE. Table 2 shows that average coverage rates increase when estimation methods are based on a spatial model (SEBLUP and SMBDE).

The results in Table 3 show that in region 1 (with sample size 2) all methods are very unstable, while in regions 2 and 3 (both with samples of size 3) EBLUP and SEBLUP are unstable. As noted earlier, EBLUP in these regions is affected by lack of variability in the data whereas MBDE is influenced by the presence of outlying values (see Chambers and Chandra, 2006). Although the estimated spatial autocorrelation is relatively higher for the EMAP data compared to the Northern Tuscany data, the simulation results for the EMAP data (Tables 3 and 4) are similar to those for the Northern Tuscany data (Tables 1 and 2). In both cases we see that the overall gain from introduced spatial dependence into small area estimation is rather small.

Finally, in Table 5 we show the performance of the different methods when population (and sample) data follow the assumed model. Here, we considered

four different values ($\rho = 0.05, 0.25, 0.50, 0.75$) for the spatial autocorrelation parameter ρ and a *W* matrix that characterises the neighbourhood structure of the small areas in terms of the contiguity characteristics of the sampled lakes in the EMAP data. As in Chambers and Chandra (2006) we note that when the assumed model is correct, estimation via EBLUP dominates estimation via MBD. These results also show that in this case the gain in small area estimation from taking account of the spatial correlation of random effects remains marginal for the MBD estimator for all values of ρ and only improves the performance of the EBLUP for large values of this parameter.

4. Concluding Remarks

This paper presents results from an initial exploration of the use of unit level models with spatially correlated area effects in small area estimation. In particular, we show how the EBLUP and MBD methods of estimation can be adapted for this situation. However, our empirical results, based both on real data as well as on simulated data under the spatial model, indicate that the gains from inclusion of spatial structure in small area estimation do not appear to be large. This is especially true for model-based direct estimation based on this structure (SMBDE), where the extra spatial information seems to have very little impact on the distribution of the SEBLUP weights that characterise this method of estimation.

There are many issues that still need to be explored in the context of using unit level models with spatially distributed area effects in small area estimation. The most important of these is identification of situations where inclusion of spatial information does have an impact, and the most appropriate way of then including this spatial information in the small area modelling process. An important practical issue in this regard relates to the computational burden in fitting spatial models to survey data. With the large data sets common in survey applications it can be extremely difficult to fit spatial models without access to high-end computational facilities. Although spatial information is becoming increasingly available in environmental, epidemiological and economic applications, there has been comparatively little work carried out on how to efficiently use this information. A further issue relates to the link between the survey data and the spatial information. In this paper we have assumed that all areas have sample units. In many situations this is not true, with survey data available only from a sample of areas. However, we often have spatial information for all areas. Saei and Chambers (2005) have explored the use of this spatial information in order to efficiently estimate the characteristics of the so-called ‘out of sample’ areas. Finally, we note that the spatial models considered in this paper have been based on neighbourhoods defined by contiguous areas. It is

easy to see that this is just one way of introducing spatial dependence between area effects, and several other options remain to be investigated.

REFERENCES

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

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

BANERJEE, S., CARLIN, B. and GELFAND, A. (2004). *Hierarchical Modelling and Analysis for Spatial Data*. Chapman and Hall, New York.

BATES, D.M. and PINHEIRO, J.C. (1998). Computational Methods for Multilevel Models. http://franz.stat.wisc.edu/pub/NLME/.

CHAMBERS, R.L. and CHANDRA, H. (2006). Improved Direct Estimators for Small Areas. Submitted for publication.

CHANDRA, H. and CHAMBERS, R.L. (2005). Comparing EBLUP and C-EBLUP for Small Area Estimation. *Statistics in Transition*, **7**, 637—648.

CRESSIE, N. (1991). Small-area prediction of undercount using the general linear model. *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality, Ottawa: Statistics Canada* pp. 93—105.

HARVILLE, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, **72**, 320—338.

NELDER, J. and MEAD, R. (1965). A Simplex Method for Function Minimization, *Computer Journal*, **7**, 308—313.

PRASAD, N.G.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.

PETRUCCI, A. and SALVATI, N. (2004). Small Area Estimation Considering Spatially Correlated Errors: the Unit Level Random Effects Model, *Working Paper 2004/10*, Dipartimento di Statistica"G. Parenti", Firenze.

PRATESI, M. and SALVATI, N. (2005). Small Area Estimation: the EBLUP Estimator with Autoregressive Random Area Effects, *Report No. 261*, Department of Statistics and Mathematics, University of Pisa. To appear *in Statistical Methods & Applications*.

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

RAO, J.N.K. (2003). *Small Area Estimation*. New York: Wiley.

ROYALL, R.M. (1976). The Linear Least-Squares Prediction Approach to Two-Stage Sampling. *Journal of the American Statistical Association*, **71**, 657—664.

ROYALL, R.M. and CUMBERLAND, W.G. (1978). Variance Estimation in Finite Population Sampling. *Journal of the American Statistical Association*,**73**, 351—358.

SAEI, A. and CHAMBERS, R. (2005). Out of Sample Estimation for Small Areas using Area Level Data. *Methodology Working Papers, M05/11*, Southampton Statistical Sciences Research Institute. (http://eprints.soton.ac.uk/14327)

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

Table 1. Relative Bias and Relative Root Mean Squared Errors generated by design based simulations using Northern Tuscany data. Regions are arranged in order of increasing population size.

Regions	Relative Bias (%)				Relative Root Mean Squared Error (%)			
	EBLUP	SEBLUP	MBDE	SMBDE	EBLUP	SEBLUP	MBDE	SMBDE
1	4.11	10.39	-3.03	-3.45	13.67	17.41	48.61	47.63
2	32.18	70.11	0.31	0.31	45.35	90.33	31.16	31.16
3	278.07	415.79	1.34	1.34	302.70	463.39	20.76	20.76
4	64.39	93.43	0.01	0.01	69.06	112.79	38.50	38.50
5	-4.90	-7.72	-1.23	-1.25	13.28	14.83	22.53	22.54
6	70.18	71.76	-9.47	-11.35	93.05	93.50	34.92	34.08
7	36.94	44.49	-0.13	-0.17	51.99	57.17	21.31	21.29
8	9.08	1.79	-1.43	-1.84	21.91	22.40	36.36	36.22
9	24.60	24.53	-0.73	-1.22	31.20	31.42	37.50	37.25
10	41.54	42.31	-2.13	-2.66	60.12	60.31	47.03	46.75
11	11.07	12.94	-0.42	-0.53	17.00	18.34	14.27	14.19
12	-3.90	-9.87	0.18	0.17	13.10	16.94	19.05	19.05
13	-15.11	-14.54	-0.47	-0.67	28.68	28.49	35.80	35.66
14	116.59	140.97	1.14	1.13	128.10	152.73	32.85	32.84
15	5.25	1.26	0.49	0.35	11.52	13.60	32.49	32.27
16	-8.17	-7.73	-1.73	-2.47	21.48	21.53	25.43	25.09
17	8.53	7.52	0.60	0.52	18.01	17.58	23.87	23.80
18	7.46	8.38	-1.59	-1.95	19.63	19.84	23.51	23.35
19	5.29	9.43	0.69	0.50	15.36	17.85	31.53	31.44
20	9.81	6.74	-0.02	-0.03	21.61	20.80	16.48	16.48
21	-5.89	-10.25	0.76	0.61	30.53	33.84	42.26	42.23
22	-9.14	-9.67	0.31	0.15	17.71	17.82	22.26	22.27
23	-11.42	-12.33	0.49	0.46	20.62	21.02	23.52	23.53
Mean	28.98	38.68	-0.70	-0.96	46.33	59.30	29.65	29.49
Median	8.53	8.38	-0.02	-0.03	21.61	21.53	31.16	31.16

Table 2. Coverage rates generated by design based simulations using Northern Tuscany data. Intervals are defined by the small area mean estimate plus or minus twice their corresponding estimated root mean squared error. Regions are arranged in order of increasing population size.

Regions	Coverage rates			
	EBLUP	SEBLUP	MBDE	SMBDE
1	1.00	1.00	1.00	1.00
2	1.00	1.00	0.99	0.99
3	1.00	1.00	1.00	1.00
4	1.00	1.00	0.97	0.97
5	1.00	1.00	0.95	0.95
6	0.50	0.51	1.00	1.00
7	1.00	1.00	1.00	1.00
8	1.00	1.00	0.97	0.97
9	0.98	0.99	0.98	0.98
10	0.99	1.00	0.97	0.97
11	1.00	1.00	0.99	0.99
12	1.00	1.00	0.94	0.94
13	0.65	0.68	0.91	0.91
14	1.00	1.00	1.00	1.00
15	1.00	1.00	0.99	0.99
16	0.49	0.50	0.91	0.91
17	1.00	1.00	0.97	0.98
18	0.98	0.98	0.98	0.98
19	1.00	1.00	0.97	0.98
20	1.00	1.00	1.00	1.00
21	0.98	0.98	0.63	0.63
22	0.91	0.93	0.93	0.93
23	0.99	0.99	0.92	0.92
Mean	0.93	0.94	0.95	0.96

Table 3. Relative Bias and Relative Root Mean Squared Errors generated by design based simulations using EMAP data. Regions are arranged in order of increasing population size.

Regions	Relative Bias (%)				Relative Root Mean Squared Error (%)			
	EBLUP	SEBLUP	MBDE	SMBDE	EBLUP	SEBLUP	MBDE	SMBDE
1	173.18	190.85	-8.85	-8.72	224.28	236.92	263.32	263.46
2	-17.18	-96.23	1.52	1.60	160.02	189.91	35.26	35.34
3	1044.89	1166.54	0.14	0.14	1056.36	1188.52	37.61	37.61
4	19.86	20.07	0.00	0.00	24.48	24.79	9.61	9.61
5	-12.08	-12.28	-0.39	-0.39	14.52	14.70	11.87	11.87
6	3.24	24.29	-0.32	-0.31	64.38	69.67	31.95	31.94
7	-0.28	-2.50	1.56	1.62	42.42	42.11	51.34	51.35
8	5.49	4.73	0.28	0.27	29.70	28.93	33.24	33.24
9	-8.07	-8.49	-0.08	-0.07	17.09	17.16	16.56	16.56
10	0.81	-0.25	0.94	1.01	22.54	22.34	27.23	27.24
11	7.19	4.51	0.51	0.57	35.00	34.51	30.32	30.38
12	-4.40	-5.07	-0.70	-0.66	21.48	21.48	25.31	25.31
13	-3.44	-0.23	0.87	0.87	32.10	31.77	34.46	34.47
14	3.07	4.10	0.29	0.48	23.90	24.18	22.51	22.51
15	-1.20	-1.09	0.30	0.33	11.65	11.59	12.48	12.47
16	22.12	25.14	-0.17	-0.18	49.42	50.92	38.06	38.06
17	4.15	2.53	0.22	0.25	16.44	16.22	11.21	11.22
18	-0.86	-3.29	0.36	0.36	10.28	10.83	7.25	7.25
19	0.38	1.51	1.93	2.12	15.50	15.57	20.47	20.52
20	-1.30	-2.10	-0.57	-0.58	17.36	17.43	16.83	16.83
21	2.15	1.49	1.17	1.27	12.25	12.16	14.96	14.98
22	4.85	2.90	-0.02	-0.03	15.82	15.29	13.83	13.83
23	-0.41	0.49	0.73	0.88	12.64	12.64	15.23	15.27
Mean	54.01	57.29	12.00	0.04	83.90	91.72	33.95	33.97
Median	0.81	1.49	0.28	0.27	22.54	22.34	22.51	22.51

Table 4. Coverage rates generated by design based simulations using EMAP data. Intervals are defined by the small area mean estimate plus or minus twice their corresponding estimated root mean squared error. Regions are arranged in order of increasing population size.

Regions	Coverage rates			
	EBLUP	SEBLUP	MBDE	SMBDE
1	1.00	1.00	0.98	0.99
2	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00
7	0.83	0.82	0.86	0.86
8	0.98	0.98	0.62	0.64
9	0.60	0.60	1.00	1.00
10	0.92	0.92	0.95	0.96
11	1.00	1.00	1.00	1.00
12	0.77	0.77	0.96	0.96
13	0.74	0.76	0.72	0.70
14	1.00	1.00	1.00	1.00
15	0.75	0.75	0.98	0.98
16	1.00	1.00	1.00	1.00
17	1.00	1.00	1.00	1.00
18	1.00	1.00	1.00	1.00
19	1.00	1.00	0.96	0.97
20	0.78	0.77	0.97	0.97
21	1.00	1.00	1.00	1.00
22	1.00	1.00	1.00	1.00
23	0.85	0.85	0.96	0.95
Mean	0.92	0.92	0.95	0.96

Table 5. Average Relative Bias (ARB,%), average Relative Root Mean Squared Error (ARRMSE, %) and average Coverage Rate (ACR) generated by model-based simulations. All averages are over the 23 small areas of interest.

Criteria	Methods	ρ			
		0.05	0.25	0.50	0.75
ARB, %	EBLUP	-8.22	-4.96	-25.56	45.84
	SEBLUP	-4.51	-5.13	-20.35	41.09
	MBDE	-35.25	5.02	9.80	69.16
	SMBDE	-35.12	5.03	9.64	68.74
ARRMSE,%	EBLUP	448.91	305.26	258.18	932.95
	SEBLUP	451.18	305.31	256.38	911.47
	MBDE	921.80	622.98	531.09	1924.81
	SMBDE	921.71	622.97	531.13	1922.11
ACR	EBLUP	0.95	0.94	0.95	0.96
	SEBLUP	0.95	0.94	0.95	0.95
	MBDE	0.98	0.98	0.98	0.98
	SMBDE	0.98	0.98	0.98	0.98

STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 351—368

EUROPEAN SURVEYS IN THE SCOPE OF SERVICE STATISTICS

Małgorzata Dytman and Agnieszka Matulska-Bachura[1]

ABSTRACT

The subject matter of the article evolves around the harmonized methodology of surveys collecting data on the different aspects of services which have been undertaken within the European Statistical System. Poland actively participates in these works. In years 2003—2006 the Central Statistical Office of Poland took part in two pilot projects initiated by Eurostat: „Demand for services" and „Business services". In the article the following items of methodology are included: the subjective and objective scope of survey, the completion of survey, methodological processing and overall assessment of survey. In each of the mentioned items there is information on aspects of surveys conducted in countries participating in the given project as well as more detailed information related to survey realized by official statistics services.

Keywords: service statistics, business services, survey methodology, data collection, Eurostat.

1. Introduction

Nowadays the service activities play the leading role in generating the value added and demand for labour. In 2003 in the developed countries the service sector produced about ¾ of value added of total economy. The largest share was achieved in Luxembourg — 83.1% and in the USA — 75.6%. Belgium, the United Kingdom and France recorded a little lower share — about 74.0%. As regards the creation of workplaces the service sector reached the largest share in Luxembourg — 77.2% and the Netherlands — 76.7% of total number of persons employed. In the United Kingdom that share amounted to 75.5%, in Sweden — 75.2%, and in the USA — 75.6%. Similar tendencies are being observed in Polish economy. In 1995 Polish enterprises running the service activities created 56.9%

[1] Services Statistics Division, Central Statistical Office of Poland, Warsaw, Poland. E-mail: DytmanM@stat.gov.pl

of value added while in 2003 that share jumped to the level of 66%. At the same time the share of people employed in the service activities raised from 43% to above 55% of number of persons employed in whole economy.

In the view of globalisation process the value added or numbers of persons employed are not the variables which sufficiently describe the phenomena taking place in the service sector. Due to the significant increase in demand for information on the service sector the need for the development of statistics concerning that sector has alluded. Therefore, within the framework of the European Statistical System the pilot studies, which one of the main objectives was to elaborate the harmonized methodology of surveys collecting information on the selected areas of services activities, were launched. In years 2003—2006 Poland took active participation in two pilot studies initiated by the Eurostat:"Demand for services" and"Business services". The main objective of these studies was to elaborate the harmonized methodology of surveys collecting information on the selected areas of services activities.

2. The pilot study „Demand for services"

The development of the service sector is driven by the growth in demand for services placed by households as well as economy. Therefore, in order to understand better the relationship between the service sector and other sectors of the economy within the works of European statisticians the efforts this aimed at uunderstanding reasons for growth of the services. To which extent is the growth caused by an externalisation of the production of services that previously have been produced in-house?

Moreover, hitherto the surveys have not been intended to collect detailed information of international trade in services. The pilot survey was expected to collect information which would enable to ddescribe the **internationalisation** of services and shed some light on the malfunctioning of the internal market for services.

The study was also to test the **feasibility of enlarging the statistical coverage of the services sector**. The purpose was to obtain the breakdown of variable "Total purchases of goods and services" into a "goods" and a "services" part and to further disaggregate the purchases of services into different types of services.

The project "Demand for services" was launched by a Eurostat grant call in March 2003. Surveys were carried out according to the methodology elaborated within the works of Eurostat's Task Force "Demand for services" in 9 Member States: Denmark, Finland, Germany, Greece, Latvia, Lithuania, Poland, Slovenia and Sweden. It enabled to describe the size of demand of companies for the selected types of services and provided information on the residence of the main service provider (country/abroad). Moreover, identifying the type and importance of barriers to the free movement of services between the EU countries as well as

the plans of enterprises for its purchase outside company were also the crucial items of study.

2.1. Population frame, sample and response rate

The survey aimed at covering the following economic sectors of NACE Rev.1.1:

- Manufacturing (by NACE Section D);
- Construction (by NACE Section F);
- Wholesale and retail trade, repair of motor vehicles, motorcycles and personal and households goods (by NACE Section G);
- Hotels and restaurants (by NACE Section H);
- Transport, storage and communication (by NACE Section I);
- Real estate, renting and business activities (by NACE Section K without division 70);
- Other community, social and personal service activities (by NACE division 90 and 92.1 + 92.2 of Section O);

The survey comprised enterprises with the number of persons employed 50 and more. However, most national statistical institutes decided to include also enterprises with the number of persons employed 20—49. The results of survey were presented at the level of the selected groupings by NACE broken down by size classes according to the number of persons employed for enterprises with the number of persons employed 50—249 and enterprises with the number of persons employed 250 and more. In Poland the examination was conducted as a full survey and comprised all units with the number of persons employed 50 and more.

In most countries the survey was based on a sample. The population frame was either the business register or the structural business statistics data. The stratified random sampling has been applied in the majority of countries, whereas two countries, Poland and Slovenia, have included all units from the frame and carried out a census.

In total nearly 40 000 enterprises received the questionnaire of survey and approximately 25 400 have responded, corresponding to the response rate of 64%. The response rate varied from 32% in Denmark and Finland as the lowest to 95% in Latvia. It was related with the fact whether the survey was conducted on voluntary or mandatory basis and whether reminder procedures are restricted.

The reminder procedures included the use of telephone reminder and mail reminders. The number of reminders varies from 1 to 3. This may be the reflection of different standard procedures, resources available, or the policy concerning the use of reminders in voluntary surveys. Germany has used another way of tackling the response burden: for each enterprise selected, two ‘back-up’ enterprises were pointed out where possible. If the first did not react, the questionnaire was send to the next enterprise etc.

In Poland the population frame was the business register. In the final card index of survey there were nearly 15 000 units and the response rate was amounted for 84%. When the deadline for submitting the Euro – U questionnaire expired the regional offices applied to units to encourage them to participate in the survey by filling in the form. In this case as well as in any voluntary pilot projects it was done in a manner of a request rather than typical official reminder. Finally the regional offices enterprises agreed to take part in the survey. Thanks to that the response rate was quite high in this survey.

2.2. The scope of information collected within the study

The objective scope of survey covered total amount of purchased services (expressed in PLN) and its structure by the following selected groups of services (according to PKWiU):

I. **Transport and logistic services** *(60—63; 64.1):*: land transport; transport via pipelines; water transport; air transport; cargo handling, storage and warehousing; other supporting and auxiliary transport services; services of travel agencies; post and courier services

II. **ICT services** *(72, 64.2)*: hardware and software consultancy; customized software**;** data processing and database services; web hosting; maintenance and repair services; other computer related services; telecommunication services (packaged software and hardware purchases are excluded)

III. **Marketing and sales-related services** *(51.1; 74.13, 74.4, 74.84,15):* intermediation in wholesale (commissions); market research and public opinion polling; advertising; organization of exhibitions, fairs and congresses.

IV. **Professional and business services** *(74.1 — excl. 74.13):* legal services, incl. patent advice; accounting**,** book-keeping and auditing services; tax consultancy; business and management consultancy; holdings (except market research and public opinion polling).

V. **Human resources related services** *(74.5; 80; 85.1):* labour recruitment and provision of personnel; education and trainings; human health services.

VI. **Financial intermediation services** except insurance and pension funding *(65—67):* financial intermediation; financial leasing; other financial intermediation; insurance and pension funding, administration of financial markets; security broking and fund management; services auxiliary to insurance and pension funding; services auxiliary to financial intermediation n.e.c..

VII. **Real estate, renting and operational leasing** *(70—71 — excl. 70.3):* real estate services with own property; buying and selling of own real estate; renting or leasing of automobiles; renting of other transport

equipment; renting of machinery and equipment without operator; renting of personal and household goods (excl. financial leasing and management of real estate on a fee or contract basic).

VIII. **Research and development services** *(73)*: — Research and experimental development on any kind of science.

IX. **Architectural, engineering and related technical consultancy services** *(74.2 & 74.3):* architectural design services for buildings and other structures; urban planning services; scientific and technical consultancy services related to engineering issues; composition and purity testing and analysis services; testing and analysis services of physical properties; testing and analysis services of integrated mechanical and electrical systems; technical automobile inspection services; other technical inspection services.

X. **Auxiliary services** (50.2; 52.7; 55.5; 70.3; 74.6; 74.7; 74.82; 74.83): maintenance and repair of motor vehicles; road assistance; repair of personal and household goods; canteens and catering; real estate services on a fee or contract basic; investigation and security services; industrial cleaning; packaging services; secretarial and translation services.

XI. **Royalties and license fees**: concessions, licenses and franchising, patents, industrial design, trademarks, copyright incl. film and music rights, excluding software. (excl. legal and patent advice, incl. other personnel related services)

XII. **Other services:** (renovation and construction services; repairs of the machinery and equipment, etc.).

as well as the costs related with using concessions, patents, trade marks, author rights, etc.

Moreover, supplementary qualitative information on the ways of service purchase (inside enterprise, within the group of enterprises, outside enterprise); types of contracts (current EU countries, outside EU countries); reasons for purchasing services abroad and plans of enterprises for purchasing the services were collected.

2.3. Questionnaire

The survey was conducted basing on the questionnaire which was developed in cooperation between Eurostat and the project leadership. The proposed questionnaire was tested in a pilot survey. It led to corrections and changes reflected in the final questionnaire. In most participating countries the questionnaire was adopted to national layout. However, these changes have not influenced the comparability.

In Poland the survey was conducted by using the Euro-U questionnaire “Report on purchase of external services in 2003”, which was designed on the

basis of the questionnaire proposed by the project task force. Before conducting the pilot study the questionnaire was tested. The test survey, which was carried out in March 2004, was based on two questionnaires: U-p and U-p/Euro. One of them was elaborated by a team of Polish experts, whereas, another one was the strict translation of the Eurostat draft questionnaire.

The test turned out to be very helpful stage of preparatory works because it enabled to verify information whether it is possible to fill in the questionnaire, correctness of records and descriptions, answering time. The reporting units could also include notes concerning the structure of the questionnaires, the way of understanding the explanations and also general remarks to the survey. Thanks to the test survey and remarks from enterprises additional knowledge on how the services of reporting units were entered into book-keepings accounts was gained. The responding units came across a lot of difficulties when fulfilling information on the costs of purchased services as the expenditures on the purchase of services from external enterprises are recorded in various fields of the profit and loss statement. While designing the questionnaire we had to take into account the ways of financial reporting required by Polish law on accounting.

2.4. The completion of the survey

In most countries the survey was voluntary. In all countries the survey was conducted as a postal survey, and only in Latvia electronic questionnaires were available for the respondents.

In Poland the „Demand for services" project was realized by official statistics services. Substantial and co-ordination works was carried out in the Central Statistical Office (CSO) with participation of statistical offices (whose main task was to carry out a field survey, send questionnaires and collect them, check, validate, contact with respondents, participate in the substantial and co-ordination works etc.) and other CSO experts who provided survey organisation and computer services for the project needs (incl. works related to processing and generating results tables).

In Poland the pilot survey was voluntary for the respondents. The regional offices sent out questionnaires to the reporting units by mail according to the previously received card index of survey and guidelines for conducting the survey. The survey with attached explanations was sent to the respondents on 31 May 2004 with the deadline for 25 June 2004.

2.5. Methodological processing

Various methods were used for correcting inconsistencies and item non-response:

- Item non-response corrected via telephone contact
- Via electronic processing
- Electronic processing combined with manual treatment
- Via electronic processing and telephone contact.
- Imputation for item non-response by different imputation methods

The method used for raising figures has been similar and the Horwitz-Thompson estimator has been used in almost all countries.

In Poland the software for entering, validating and processing of the collected data was prepared by the Central Statistical Computing Center (COIS). During the process of entering data the control of logical and book-keeping correctness as well as the completeness of survey was conducted. If there was no data in the particular part of questionnaire statistical offices tried to contact the respondents in order to clarify any incompatibilities. When there was the lack of the more detailed breakdown of quantitative data the enterprises were asked to prepare estimated data. All problems connected with individual data were explained directly with the responding units.

The next step of processing data was to create the control tables for voivodships' datasets and then compare the individual survey data with the data obtained during structural business surveys. After clarifying all arisen divergences datasets were accepted and merged into one national set. On its base the national control tables were elaborated.

The sets of individual data were aggregated into one national dataset. Taking into consideration a high response rate as well as the fact of using appropriate unit selection (full population of the units with the number of persons employed 49 and more) and the aim of survey (defining the structure of service purchase, area of the services; barriers in purchasing them as well as enterprises' plans in the filed of generation of the services) it was decided that grossing up of the collected data is not necessary.

2.6. Overall assessment of the survey — conclusions:

1. Generally speaking, the survey has been conducted successfully.
2. The coordinators find that a disaggregation of the SBS variable "Total purchases of goods and services" into a "goods" and a "services" into a "goods" and a "services" part is feasible, and that a further breakdown into different types of services could be possible with further investigation in service categories.
3. The test survey, as a preparatory stage for the "Demand for services" project, was unusually useful.

4. Generally, the questionnaire worked relatively well. The most common problem related to the questionnaire concerned the definitions of the services categories. It seems that even though an explanation had been developed, the enterprises found it difficult to relate their accounts and everyday production or purchase of services to the defined categories.
5. One of the main problems with the survey was that parts of the questionnaire were filled in by different people, e.g. the questions on use of external service providers may have been filled in by persons with knowledge of purchase strategies, whereas the questions on actual purchase of services and investments were typically filled in by persons with knowledge of accounts and therefore the results were not always consistent
6. Some countries stated that a higher response rate could have been reached if the qualitative part of the questionnaire had been separated from the quantitative part.
7. Almost all national statistical offices have had problems concerning the response burden.
8. Voluntary surveys should be kept as simple as possible
9. The most common problem related to the questionnaire concerns the definitions of the categories.

3. The pilot study „Business services"

Among the service activities the business services related activities[1] (by NACE division 72 and 74) develops very dynamically. Therefore, in 1999 Eurostat launched a development project "Methodological development and harmonised data collection for business services". It includes the number of pilot studies in the scope of business services and its main purpose was to improve the statistical coverage of the business services sector. A proposed questionnaire was developed and tested in a pilot survey in NACE 72. After an evaluation of the pilot survey a full scale survey in NACE 72, 74.12, 74.13, 74.14, 74.20 and 74.40 was launched. The work of the project has advanced to a stage where methodological guidelines have been developed and tested. The survey has now been carried out in three different years (2001, 2003 and 2004). During the years the questionnaire has been developed and changed in order to improve it. The coverage of division 74 was further extended to also cover NACE 74.11, 74.30,

[1] The project "Business services" includes the following types of service activities: computer and related activities, legal activities, accounting, book-keeping and auditing activities; tax consultancy; market research and public opinion polling activities; business and management consultancy activities; architectural and engineering services and related consultancy; technical testing inspection and analysis services; advertising; labour recruitment and provision of personnel activities

and 74.50 in the second year. Since the beginning of the project new participating countries have joined the project.

For the year 2004 the coverage of the questionnaire remained unchanged from that of 2003. The variables covered only presented marginal changes. On the other hand the 2004 survey included a series of qualitative questions relating to cross-border trade which are completely new.

Poland joined the project at its second edition. The implementation of the project required pre-assessment of the current works in the field of business services statistics with special focus on the scope of the available data as well as the organisation of the statistical surveys. As a result two pilots studied for the reference years 2003 and 2004 were carried out. Due to the large population of enterprises running the business services related activities in Poland they were surveyed within two separate studies.

3.1. Population frame, sample and response rate

The survey for the reference year 2004 aimed at covering the following economic activities of NACE rev. 1.1:

- 72.10 Hardware consultancy
- 72.21 Publishing of software
- 72.22 Other software consultancy and supply
- 72.30 Data processing
- 72.40 Database activities
- 72.50 Maintenance and repair of office, accounting and computing machinery
- 72.60 Other computer related activities
- 74.11 Legal activities
- 74.12 Accounting, book-keeping and auditing activities; tax consultancy
- 74.13 Market research and public opinion polling activities
- 74.14 Business and management consultancy activities
- 74.201 Architectural activities
- 74.202 Engineering activities and related technical consultancy
- 74.30 Technical testing and analysis
- 74.40 Advertising services
- 74.50 Labour recruitment and provision of personnel

The survey comprised the enterprises of all size classes. The results of survey were presented broken down by size classes according to the number of persons employed as follows for quantitative data (turnover data): enterprises with the number of persons employed 9 and less, enterprises with the number of persons employed between 10—49, enterprises with the number of persons employed between 50—249, enterprises with the number of persons employed 250 and more and for qualitative data (cross-border trade data): enterprises with the

number of persons employed 49 and less and enterprises with the number of persons employed 50 and more.

In Poland the above mentioned services activities were surveyed within two studies carried out for two consecutive reference years 2003 and 2004. The project for the reporting year 2003 covered the following services activities:

- computer and related activities (NACE 72),
- accounting, book-keeping and auditing activities; business and management consultancy activities (NACE 74.12+74.14),
- architectural and engineering activities (NACE 74.20),
- advertising (NACE 74.40)

The next edition of "Business services" project for the reporting year 2004 covered the enterprises running the following activities:

- legal activities (NACE 74.11),
- market research and public opinion polling (NACE 74.13),
- technical testing and analysis (NACE 74.3),
- labour recruitment and provision of personnel (NACE 74.5)

In Poland there is the lack of more detailed breakdown for the size class of enterprises with the number of persons employed 49 and more into enterprises with the number of persons employed between 50 and 249 and those with the number of persons employed 250 and more.

The population frame in the different participating countries was the business register or the structural business statistics frame work. Some countries have used stratified sample, while other have used the probability proportional to size sampling. In total nearly 81 000 enterprises have received a business services questionnaire, and approximately 31 500 have responded, corresponding to a response rate of 40 per cent (un-weighted). From approximately 20 per cent in Poland, as the lowest, to approximately 95 per cent in Denmark as an absolute maximum.

The procedures used for improving response rates where telephone reminders or mail reminders. The number of reminders procedures varies from 1 (in Finland, Germany and Slovakia) to 4 (in Latvia and Slovenia). This may be the reflection of different standard procedures, resources available, or the policy concerning the use of reminders in voluntary surveys.

In Poland the surveys were conducted as representative surveys. In the case of survey for the reference year 2003 at the planning stage it was considered to cover all of the large units (with the number of persons employed 49 and more) and 25% sample of the medium (with the number of persons employed 10—49) and small (with the number of persons employed 9 and less) units. However, due to the fact that in the test survey responded approx. 1 000 reporting units, the final sample covered all enterprises with the number of persons employed 10 and more and sample of enterprises with the number of persons employed 9 and less. The survey for the reference year 2004 covered all enterprises with number of persons

employed 10 and more and 45 % sample selected from the enterprises with number of persons employed 9 and less.

The base for defining the subjective scope of survey was the population frame for annual surveys on structural business statistics for the surveyed reference years 2003 and 2004 taking into account the selected classes and groups by NACE:

- survey for the reference year 2003: 7210, 7220, 7230, 7240, 7250, 7260, 7412, 7414, 7420, 7440;
- survey for the reference year 2004: 74.11; 74.13; 74.30; 74.50.

The frame was broken down by employment size classes: large enterprises (>49), medium enterprises (10—49) and small enterprises (<10).

The algorithm of sample allocation in small enterprises with number of persons employed 9 and less for both surveys was similar. Algorithm of the fixed precision allocation from Lednicki and Wieczorkowski (2003) was used. This algorithm allows division of the sample with given sample size to subsamples in such a way that expected precisions for the surveyed variable estimator (mean or total) should be optimal and equal in all subpopulations. In this real application it was also assumed that all coefficients of variations were equal, because there was no historical information available from the previous surveys. Sample allocation obtained from such algorithm for each class of NACE was finally divided proportionally into geographical regions (voivodships). Using this method the sample for enterprises with the number of persons employed 9 and less amounted to:

- survey for the reference year 2003 — 41 428 units
- survey for the reference year 2004 — 27 999 units

According to the sample allocation into strata the sample was drawn, independently in strata by simple random sampling without replacement using the SAS SURVEYSELECT procedure.

Table 1. Number of enterprises in population frame and sample of survey„Business services" for the reference years 2003 and 2004

Specification			Size classes by the number of persons employed					
Kind of activity	Symbol by NACE	year	1—9		10—49	50+	Total	
			population frame	sample			population frame	sample
Computer and related activities	72	2003	38 387	18 172	658	73	39 118	18 903
Legal activities	74.11	2004	17 939	8 632	187	14	18 140	8 833
Accounting, book-keeping and auditing activities	74.12	2003	30 856	5 846	396	14	31 266	6 256
Market research and public opinion polling	74.13	2004	6 797	4 825	63	14	6 874	4 902
Business and management consultancy activities	74.14	2003	26 383	5 666	459	26	26 868	6 151
Architectural and engineering activities	74.20	2003	60 634	6440	979	64	61 677	7483
Technical testing and analysis	74.30	2004	4 388	3 473	117	38	4 543	3 628
Advertising	74.40	2003	20 057	5 304	333	21	20 411	5 658
Labour recruitment and provision of personnel	74.50	2004	33 098	11 069	85	29	33 212	11 183

Source: REGON — Register of units GUS.

In the final samples (card index) there were 44 451 units for the reference year 2003 and 28 546 enterprises for the reference year 2004.

As far as the response rate is concerned it amounted to 28% in case of survey for the reference year 2003 and 20% in survey for the reference year 2004. The level of response rate was determined by the voluntary nature of survey and the major share (above 90%) of enterprises with the number of persons employed 9 and less in the population of surveyed enterprises. Moreover, the response rate depended on the employment size classes. In the group of enterprises with persons employed 50 and more, despite of the voluntary nature of the survey, in the survey for the reference year 2003 — 96% and in the survey for the reference year 2004 — 80% of enterprises responded. The smaller enterprise the lower response rate was recorded.

3.2. The scope of information collected within the study

Within the conducted surveys data on the value and structure of turnover broken down by:

- product;
- type of client: households and other non-commercial institutions, enterprises, public sector;
- residence of client: country, EU countries, non-EU countries

was collected.

Moreover, within the survey for the reporting year 2004, apart from the data on the structure of turnover broken down by product, type and residence of client, information on the types of cross-border trade, the reasons for its development and the perceived barriers was also collected.

3.3. Questionnaire

The survey was conducted basing on the draft questionnaire which was developed in cooperation between Eurostat, the Commission's DG-Enterprise and the project leadership as early as 1999. The proposed questionnaire was tested in a pilot survey during 2000 concerning enterprises in NACE 72, leading to corrections and changes. During the years the questionnaire has been developed and changed, and new variables have been included or some have been excluded. In the participating countries the questionnaires have been adopted to national layout. There were only small deviations from the one proposed by Eurostat.

The main change in the questionnaire concerning 2004 is that the qualitative questions concerning cross-border trade have been included. Almost all of the participating countries had more or less problems with this part of the survey, especially with the last question concerning the barriers met in cross-border trade.

There have also been problems with the definition of cross-border trade. In Sweden the definition has been interpreted as export, in Norway the definition was import and/or export of goods and/or services. In the United Kingdom for example, the majority of respondents were unsure about which countries were classed as cross-border. There was not any pilot survey on the cross-border trade questions. The questions were directly included in the business services questionnaire.

In Poland both surveys were conducted by using 8 different questionnaires designed separately for each of the surveyed activities:

1. Computer and related activities (by NACE 72) — „Report on turnover in computer and related services in 2003" — BS-Info;
2. Legal activities (by NACE 74.11) — „Report on turnover in legal activities in 2004" — BS-Praw;
3. Accounting, book-keeping and auditing activities; Business and management consultancy activities (by NACE 74.12+74.14) — „Report

on turnover in accounting, book-keeping and auditing services in 2003" — BS-Gosp;

4. Market research and public opinion polling (by NACE 74.13) — „Report on turnover in market research and public opinion polling in 2004" — BS-Ryn;
5. Architectural and engineering activities (by NACE 74.120) — „Report on turnover in architectural and engineering services in 2003" — BS-Arch;
6. Technical testing and analysis (by NACE 74.30) — „Report on turnover in technical testing and analysis in 2004" — BS-Tech;
7. Advertising (by NACE 74.40) — „Report on selling advertising services in 2003" — BS-Rek;
8. Labour recruitment and provision of personnel (by NACE 74.50) — „Report on selling labour recruitment and provision of personnel in 2004" — BS-Prac;

The test survey for the reference year 2003 aimed at obtaining information on the possibilities for filling in the questionnaires, correctness of the records and instructions. The final version of above specified questionnaires was designed basing on information collected during the test survey

As experiences gained during the realization of survey for the reference year 2003 were used in preparing the survey, therefore, it was not necessary to conduct the test survey for the reference year 2004.

3.4. The compilation of the survey

The survey was voluntary in four (Finland, Germany, Poland and Slovakia) of the 15 participating countries. In Sweden the qualitative part of the survey was voluntary, but the quantitative part was mandatory. In most of the countries the survey was carried out as a postal survey, but in Finland, Latvia, Norway, Spain and in Sweden the enterprises had the opportunity to choose an electronic questionnaire (by Internet). In Greece the private associations visited the enterprises and introduced the questionnaire.

In Poland the "Business services" project was conducted by official statistics services. Substantial and co-ordination works was carried out in the Central Statistical Office (Service Statistics Division) with participation of statistical offices and other CSO. The division of tasks was the same like in "Demand for services" project.

The questionnaires of survey were transferred to regional statistical offices, which then sent them out by regular post to reporting units according to the previously defined card index (sample) and guidelines for survey implementation. The questionnaires were accompanied by instructions and a cover letter signed by President of the CSO.

3.5. Methodological processing

The participating countries used various methods for correcting inconsistencies and item non-response, among others:

- Item non-response corrected via telephone contact
- Via electronic processing
- Electronic processing combined with manual treatment
- Via electronic processing and telephone contact
- Imputation for item non-response by different imputation methods

All participating countries more or less have problems with item non-response, especially the cross- border trade questions. If the questionnaire was not filled in properly, the enterprises in the participating countries were contacted and tried to correct the questionnaire or get the missing responses. This way of treatment on item non-response concerns mostly the turnover questions. However not so much effort was spent regarding the qualitative part of the questionnaire. If the data was still missing a range of imputation methods have been used. Some example of imputation methods are:

- In Finland item non-response in the questions on Barriers met in cross-border trade was treated using the option don't know/not applicable.
- In Spain questionnaires with item-non response is considered as not valid.
- In Slovenia item non-response was imputed using the mean imputation method and the hot-deck method. The latter means that every time the process of imputation of missing values is run, different donors donate their values in order to impute missing values.
- In Sweden the most common answer was used as imputation method if there were no tick mark for cross-border trade questions B, C and D.

Looking at the number of item non-response, the question block Barriers met in cross-border trade have the highest number of enterprises with no answer. As mentioned before most of the participating countries seem to have problem with this questions.

In Poland the logic and book-keeping control as well as completeness control of surveys were conducted during registration process of the collected data. When some part of the form was not completely filled in the personnel of the regional statistical offices contacted directly with enterprises in order to get missing data. Thanks to this, at the national level there were no missing answers or partly answered questions. Then, the individual datasets from the voivodships (regions) were merged into control tables. The comparative analyses of the individual data obtained during the surveys with the results from structural surveys of enterprises were conducted. After clarifying all divergences the datasets were approved and merged into one national compilation, and on its base the control tables for total Poland were compiled.

The obtained data was grossed up by number of units in the given population level and size class. The generalization was done with a co-operation with the

experts (mathematicians) in the scope of sample selection and generalization according to the following algorithm:

Original sampling weights were obtained by means of the SURVEYSELECT procedure and were simply ratios of the number of units in the frame to the number of units in the sample in the corresponding strata. These original weights were corrected by the so called activity coefficients according to formula:

$$f_A = \frac{1}{n}\left\{ RA_{01} + RA_{08}\frac{n - RA_{01}}{n - RA_{01} - RA_{09}} \right\}$$

where n means the number of units in the analyzed stratum, and RA_{code} is the number of units in the given stratum with the value of the *RA* field (symbol of the participation of the unit under examination in the survey e.g. *RA*=01 means that the unit participated in the survey, *RA*=08 means that the unit did not want to respond but was localized, *RA*=09 lack of contact and so on) equal *code.*

Corrected weights were denoted as W', where $W' = Wf_A$, and W denotes original weight. Weights calculated in this way then (after the correction) were modified according to the formula $W'' = W'\frac{\sum W'}{\sum_{RA=01} W'}$ (in every stratum) because of lack of the answer in strata.

Weights W'' were used to produce estimators for selected variables in defined subpopulations.

The corrected weights were also obtained separately for big and medium units using described algorithm, where original theoretical weights (equal to one) were modified in four NACE classes.

Based on the grossed-up results the national results tables were elaborated. The survey results were discussed together with the experts of structural business statistics and mathematicians. Comparing and matching data, for enterprises with number of persons employed 9 and less, from the pilot project with data from structural business statistics surveys it is important to highlight that both surveys (pilot project and SP3 “Reporting on business activities” survey) are representative and different methods of sampling and results estimation were used.

3.6. Overall assessment of survey

1. Generally speaking, the survey has been conducted successfully and the coordinators think it is feasible to collect this kind information, except the cross-border trade part of the survey where there has been a large number of item non-response.
2. It would also have been a good idea to test in a pilot study whether it is preferable to collect both the quantitative and qualitative data in the same questionnaire. Some participating countries mentioned that a higher response rate could have been reached, if the qualitative part of the questionnaire had been separated from the quantitative part. In Sweden this solution, as mentioned before, worked well.
3. One of the main problems with the survey was that different parts of the questionnaire were not always filled in by the same person, e.g. the questions of turnover may have been filled in by persons with knowledge of sales, whereas the questions on cross border trade were typically filled in by persons with knowledge of export and export strategies.
4. To get higher response rate the survey probably need to be mandatory, which was not the case in all participating countries carrying out this survey. In participating countries where the survey was voluntary the response rate was much lower than in countries where it was mandatory.
5. One of the main problems with the collection of turnover divided into different products is that the products are not known in advance by enterprises and that the product list is very detailed. Therefore the enterprises have problems in reporting these data since their accounting system do no match the product categories in the questionnaires. It is advisable that the enterprises should know the product categories well in advance in order to arrange their accounting system before the data collection.
6. Many participating countries also said that they have problem with break down of client by type of enterprise.

REFERENCES

GUS (2005), Final Report of Demand for services project, January 2005 (Grant Agreement 200344201025).

GUS (2005b), Final Report of Business Services project, June 2005 (Grant Agreement 200344201041).

GUS (2006), Final Methodological Report on the Survey on Business Services, June 2006 (Grant contract 2004.4420.1009).

EUROSTAT (2001 – 2003), Documents for the Working Group meeting "Business Services Statistics" 13-14 November 2001, Documents for the Working Group meeting "Business Services Statistics", 26-27 May 2003 , Minutes of the Business Services Task Force meeting, 16 December 2003.

EUROSTAT (2004) , Documents for the Task Force meeting Demand for Services (16 March 2004).

EUROSTAT (2004-2005), Documents for the Business Services Working Group meeting 21-22 October 2004, Minutes of the Business Services Working Group meeting 24-25 October 2005.

LEDNICKI, B., WIECZORKOWSKI, R (2003).Optimal Stratification and Sample Allocation between Subpopulations and Strata, *Statistics in Transition,* 2003, Vol.6, No.2 .

STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 369—382

CHANGES OF EMPLOYMENT STRUCTURE IN POLAND AND EU COUNTRIES

Anna Malina[1], **Piotr Malina**[2]

ABSTRACT

The article presents the results of an analysis of Poland's employment structural changes as compared with the EU countries. The conducted research on Poland's similarity to the EU-15 countries and the countries which joined the Union in 2004, is based on specific structure similarity measures. Also, the extent of Poland's and the EU's structural changes was assessed in 1980—2004. The analysis of the sector-based employment structure in Poland and the EU countries indicates the occurrence of similar trends: a decreasing share of employees in sector I (agriculture), in favour of sector III (service sector). Apart from the recorded favourable trends, there are considerable differences between Poland and EU countries. The share of farming-related employees in Poland is high (four times as high as average figures in EU-15), while the respective share in the service sector is low (approx. 50% of the total number of employees); the respective average level reached by EU-15 countries is 70%. The results of the research indicate that the major changes to Poland's employment structure occurred at the beginning of the economic transformation process, i.e. between 1990 and 1995.

Keywords: employment structure, employment rate, older workers, part-time work, classification, taxonomic distance.

1. Introduction

Structural changes occur at different levels and may relate to different economic areas. A number of experts believe that social and economic transformations at the national and regional level are mainly reflected in sector-, or industry-related employment structure. Employment structure in the economy's particular sectors constitutes a major criterion of assessing the country's economic

[1] Cracow University of Economics, Department of Statistics, 27 Rakowicka Street, 31-510 Krakow, Poland, e-mail: malinaa@ae.krakow.pl
[2] IBM, Poland, Armii Krajowej Street, Krakow.

advancement. On the basis of changes in the economies of many countries, the following thesis may be put forward: the country's economic advancement is coupled with a decreasing role of the farming sector in total employment in favour of the service industry [Głowacka, 1995; Głąbicka, 2000].

A superficial analysis of industry-based employment structure in Poland and EU member countries indicates that the major difference between Poland and West European countries lies in the fact that the transformation of this sector of Poland's economy is unavoidable, however difficult it may be from a social and economic point of view.

The paper focuses on the selected aspects of Poland's and the EU countries' employment structures in view of the systemic changes and European integration processes. The economic and social transformation initiated in 1989 led to changes in employment structure, affecting ownership status as well as the particular industries. A major shift of labour was recorded from the public to the private sector as well as from production to service areas.

The objective of this paper is to identify differences and assess structural changes in Poland's employment system after 1990, referring them to the EU countries. Issues related to the labour market, especially industry-related employment structure and unemployment rates, are certainly of vital importance. Apart from the national income per capita, they constitute one of the major criteria, of allocating EU structural funds.

2. Characteristics of employment structure in Poland and other EU countries

The EU employment resources (EU-25) amounted to nearly 377.5 million of people in 2004, including 194.5 million in employment. In the EU-15 the number of employees was at the level of 165.5 million in 2004, in the new member countries (EU-10) nearly 30 million, and in Poland — 13.8 million. It should be noted that Poland's labour resources (according to the *Eurostat's* data*)* account for about 46% of the total labour resources of the countries admitted to the UE in 2004 and represent approximately 9% of EU –15.[1]

The adjustment of the labour market in Poland to EU standards — just like in the case of the other countries which became fully-fledged EU member states in May 2004 — is a strategic objective to be achieved by all the new member states. The flexibility of the labour market is of vital significance not only in the particular member states but also in the entire European Union. An analysis of employment structure plays a special role in any examination of the labour market.

[1] Source: EU Labour Force Survey: „*Statistics in Focus*", 9/ 2005 (European Communities, Eurostat).

Employment structure may be considered from the point of view of different criteria. The basic classification criteria include the following: gender, age, education, qualifications, work experience, and employment status. In the era of systemic changes a major role is played by the employment structure from the point of view of industries and forms of ownership.

Table 1 presents basic indicators of employment structure in the EU countries as well as in Poland in 2004 .

Table 1. Employment rates in Poland and EU countries in 2004 (%).

Name	Total (age:15—64)	Men (age:15—64)	Women (age:15—64)	Older workers (age:55—64)
EU-25	63.3	70.9	55.7	41.4
EU-15	64.7	72.7	56.8	42.8
EU-10	60.4	67.3	53.1	37.1
Poland	***51.7***	***57.2***	***46.2***	***26.2***

Source: EU Labour Force Survey: Population and Social Conditions, *Statistics in Focus*, nr 9/2005 (*Eurostat).*

Table 1 indicates that in the EU-15, 64.7% of working age population (15—64 years) have jobs. The respective employment rate in new member states (*EU-10*) is at the level of 60.4%, while in Poland it amounts to 51.7%. In EU-15 countries the highest employment rates (above 70%) are recorded in Denmark (75.7), the Netherlands (73.1), Sweden (72.1), and the UK (71.6). The lowest rates are recorded in Greece (59.4), Spain (59.8) and Italy (57.6). In the countries which joined the EU in 2004, the highest rates are recorded in Cyprus (68.9), Czech Republic (64.2) and Slovenia (65.3). In 2004, Poland's employment rate is the lowest among all EU member states (51.7%). Poland's rate is lower than EU-*15* average rate by 13%. In EU-15, the average employment rate in the male population is higher than that for female population by nearly 16% (15—64 years). In Poland the respective gap is much lower amounting to 11%.

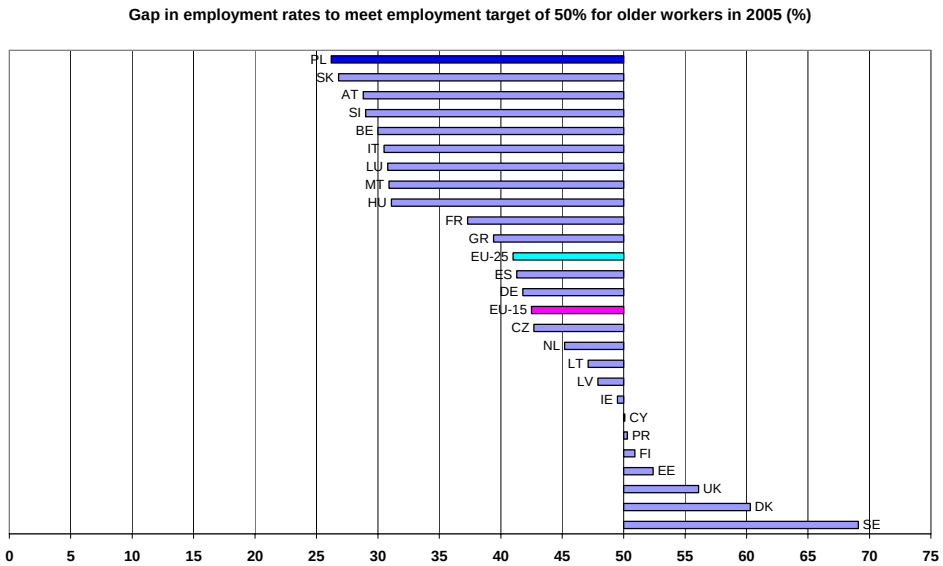
The employment of older workers[1]

The target set by the Stockholm Council is an employment rate of 50% for older workers, defined as the 55—64 year age group. In 2004, the employment rate of older workers in the EU-25 was 41%, a gap of nine percentage points. Only five countries within the EU-25 have employment rates above 50% for older workers, while a further two countries are close to the target rate. In the other EU member states, the employment rates of older workers are significantly below the

[1] The employment of older workers is calculated by dividing the number of people aged 55 to 64 years who are in employment, by the total population of the same age group.

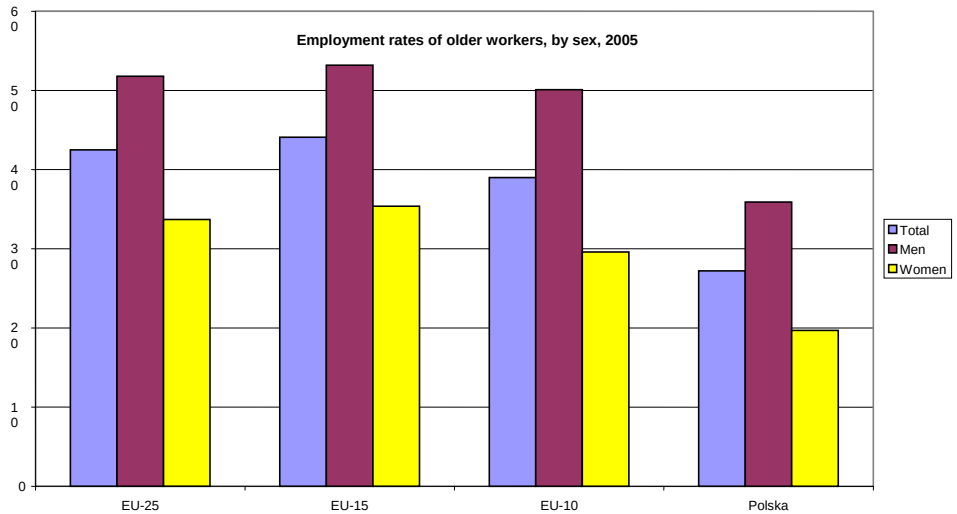
targeted figure. In Poland, the employment rate of older workers was 26%, lowest within the EU countries (EU-25), see fig. 1.

Figure 1. Gap in employment rates for older workers for EU countries to meet employment target of 50% for older workers.



Source: Author's own research on the basis of *Eurostat Data*, 2005.

Figure 2. Employment rates of older workers (age: 55—64) in Poland and EU group countries, in 2005.



Source: Author's own research on the basis of *Eurostat Data*, 2005

Female labour market participation

In 2005, the female employment rate for the EU-25 was 56.3%. The majority of countries are far behind the Lisbon target of 60% (fig. 3), although a few of them are above the target. For instance, Denmark and Sweden have female employment rates above 70%, and the Netherlands, the UK and Finland show rates above 65%.

Figure 3. Female employment rates in EU countries, 2004.

Source: Author's own research on the basis of *Eurostat Data*, 2005

Part-time employment

In EU-15, nearly 20% of the working population work part-time, with a large percentage of females. In the total, the proportion of working females, working part-time equals 33.5% (from 8% in Greece to 72.8 % in the Netherlands). The highest percentage of part-time employees is recorded in the Netherlands (46%), and the lowest in Slovakia — 2.4%. In Poland, part-time employees account for 10.8% of the working population (Table 2), and as compared with West European countries, the rates of part-time work are low (males: 4%; females: 6%).

Table 2. Part-time employment rates (age group 15—64) in Poland and EU countries (2004).

Name	Total	Men	Women	Share of total employment (%)
EU-25	10.9	4.5	17.2	18.2
EU-15	12.2	4.7	19.7	19.9
Poland	***5.1***	***4.1***	***6.0***	***10.8***
Max (Netherlands)	33.0	17.3	49.1	46.0
Min (Slovakia)	1.4	0.8	2.0	2.6

Source: EU Labour Force Survey: Population and Social Conditions, Statistic *in Focus,* nr 9/2005 and 6/2006, *Eurostat*.

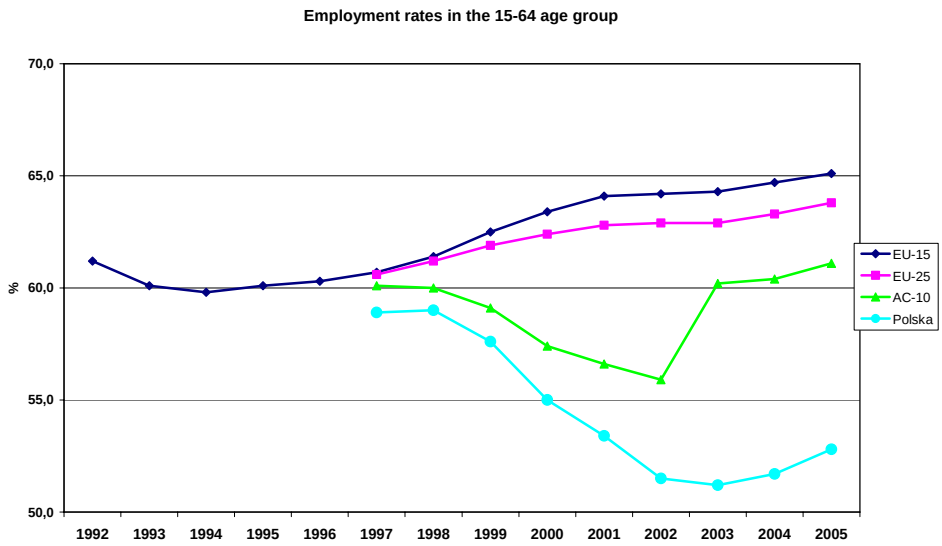
Table 3 shows changes in employment rates in selected EU countries and Poland. The general employment rates in EU-15 and the enlarged Union (EU-25) tend to growth during the investigated period, while in Poland and the countries which joined the EU in 2004, a reverse trend is recorded. This adverse trend changed as late as in 2003, in Poland in 2004 (see: Fig. 4).

Table 3. Employment rates (age 15—64) in Poland and EU countries in period 1997—2005 (%).

Name	1997	1998	2000	2002	2004	2005
EU-25	60.6	61.2	62.4	62.9	63.4	63.8
EU-15	60.7	61.4	63.4	64.2	64.7	65.1
EU-10	60.1	60.0	57.4	55.9	60.4	61.1
Poland	**58.9**	***59.0***	***55.0***	***51.5***	***51.7***	***52.8***

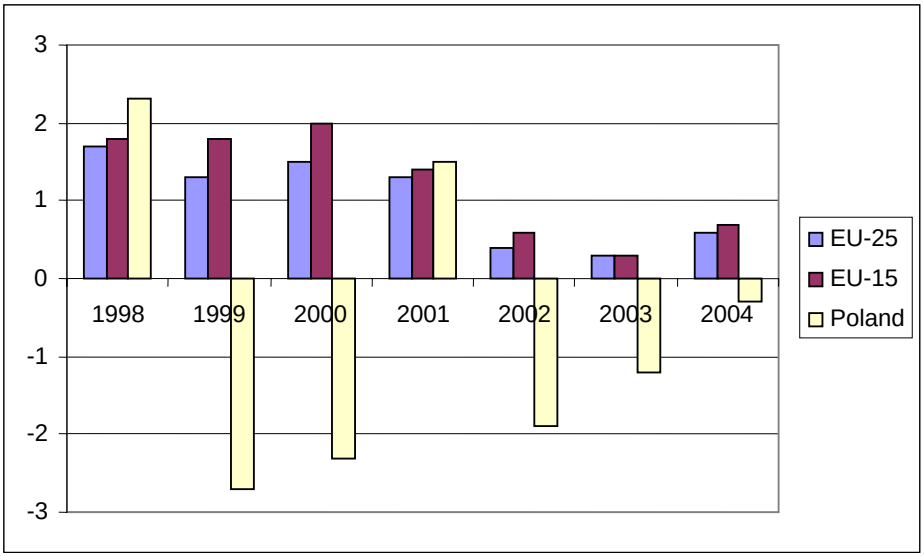
Source: Eurostat, 2005 and Employment in Europe 2005.

Figure 4. Employment rates (age 15—64) in Poland and EU countries in period 1997—2005 (%).



Annual percentage change in total employment in Poland and EU countries (EU-25, EU-15), in period 1998—2004 is presented in Fig. 5.

Figure 5. Annual percentage change in total employment in EU countries (EU-25, EU-15) and Poland in period 1998—2004.



Source: Author's own research.

Decline in the number of working population in the given period has been noticed in Poland since 1999. It was only in 2001 that Poland recorded a slight rise in the working population (by 1.5%). In this respect, Poland is not in a favourable position as compared with other EU countries.

3. Analysis of changes in employment structure across economy's sectors in Poland and EU countries

The subject matter of the comparative analysis of Poland and the EU countries is a sector-based employment structure. Employment structure includes three major sectors of the economy:

- sector I: employment in agriculture,
- sector II: employment in industry and construction,
- sector III: employment in services.

Table 4 presents data related to the share of employees in the particular sectors in Poland and EU countries in selected years of the period 1960—2004.

Table 4. Employment structure by sectors in Poland and EU countries in 1960-2004.

Year	European Union*						Poland		
	Arithmetic average			Standard deviation					
	Sector I	Sector II	Sector III	Sector I	Sector II	Sector III	Sector I	Sector II	Sector III
1960	16.45	43.45	41.33	7.39	3.44	6.97	48.0	28.9	23.1
1970	9.93	41.85	48.22	4.88	3.87	5.90	38.6	34.2	27.2
1973[a]	***9.82***	***39.22***	***50.95***	***6.31***	***5.28***	***6.15***	.	.	.
1980	7.45	34.12	58.42	4.56	4.53	5.23	30.2	37.9	31.9
1981[a]	***9.17***	***32.79***	***58.04***	***7.85***	***4.38***	***7.66***	.	.	.
1985	8.67	29.08	61.91	7.39	4.44	9.10	29.6	36.1	34.3
1986[a]	***10.17***	***28.55***	***61.27***	***7.57***	***4.36***	***8.04***	.	.	.
1990	8.12	29.05	61.77	5.61	4.34	7.22	27.6	35.5	36.9
1995[a]	6.95	28.29	64.78	4.71	3.53	5.57	27.8	30.1	42.2
2000	5.62	27.32	67.07	4.11	4.13	6.35	28.4	26.2	45.4
2002	5.17	26.54	68.27	3.90	4.15	6.38	27.6	26.2	46.2
2004[b]	***5.04***	***27.91***	***67.05***	***5.17***	***4.51***	***8.45***	20.1	28.5	51.4

*EU parameters are calculated on the basis of data referring to those countries which were EU members in the particular years.

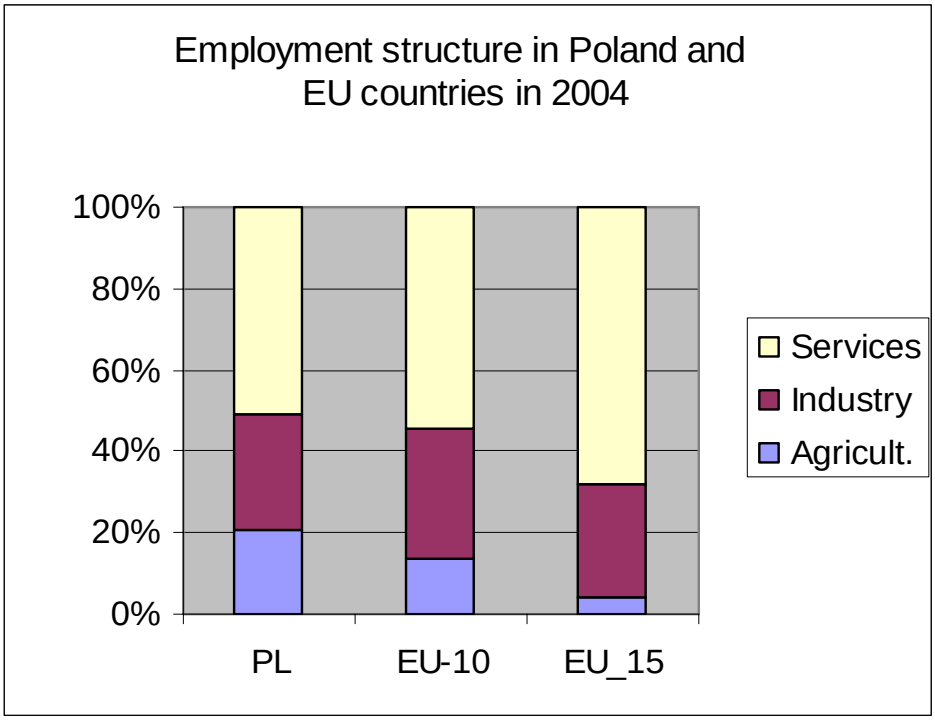
[a] Include a new member states: 1973 r. — Denmark, Ireland, United Kingdom; 1981 r. — Greece; 1986 r. — Spain, Portugal; 1995 r. — Austria, Finland, Sweden. [b] 2004 — EU-25.

Source: Author's own research (1980—1990 data were obtained from *International Statistics Yearbooks*, while 1995—2004 data — with reference to UE countries — from: *Labour Force Survey*, „Statistics in Focus" (*Eurostat*), and for Poland from *Statistical Yearbooks of the Republic of Poland,* 1996—2005 editions).

On the basis of descriptive parameters related to EU sector-based employment structure a distinct trend may be noted of a decrease in the average share in sector I (agriculture) as well as in sector II (industry and construction) in favour of sector III (services). The weakening of the trend occurred in 1985—1990, when new countries joined the Union — Greece in 1981, and Spain and Portugal in 1986. Similar trends of changes to employment sector structure are recorded in Poland. In the analyzed period Poland recorded a systematic increase in the number of employees in the service sector, which was coupled with a decrease in the number of people employed in agriculture and industry (see: Table 4).

Fig. 6 shows employment structure in Poland's basic economic sectors as compared with European Union countries (average figures for EU-15 countries and new European Union member states in 2004).

Figure 6. Employment structure in Poland, European Union countries (EU-15) and new member countries (EU-10) in 2004.



Source: Author's own research on the basis of *Eurostat* (*Labour Force Survey*, 2005).

The comparison of employment structure in the particular sectors of Poland's economy with average EU figures (EU-15) leads to the conclusion that apart from economic and structural changes, Poland differs considerably from EU member countries. The share of employment in sector I (agriculture) in Poland's economy

in 2004 (according to *Eurostat* data) was four times higher than in EU-15 countries, while the share of employees in the service sector amounted to 50%, while in EU countries it was at the average level of 70%. Also, considerable differences in employment structure are recorded between Poland and the countries which became new members in 2004 (in Fig. 6 these countries are referred to as EU-10).

On the basis of the conducted analysis it may be clearly inferred that Poland has the highest share of employees in sector I (agriculture) among all European Union countries, both EU-15 and new members (approx. 20% in 2004, according to *Eurostat*); in the service sector the level is close to average figures in new member countries (approx. 51%), while in the industry sector (sector II) it is lower than in the EU countries (Poland: 28%, EU-15: 26 %, and EU-10: 31%).

4. Assessment of the extent of employment structural changes in Poland and EU countries

Different types of measures may be applied in an analysis of structural changes at national or regional levels. The following measure (coefficient) was applied in the analysis[1]:

$$V_{t,t+\tau}^{(i)} = \left[\sum_{j=1}^{m} q_{j,t+\tau}^{i} \left(\frac{q_{j,t+\tau}^{i}}{q_{j,t}^{i}} - 1 \right)^2 \right]^{1/2} \quad (i = 1,...,n) \qquad (1)$$

where q_j^i - indicator of employment structure in economic *j*-th sector (share of employees in the particular economic sectors) in the *i*-th country.

When coefficient $V_{t,t+\tau}^{(i)}$ equals zero no changes occur in the structure of an *i*-th object (country) in the comparable periods *t* and *t*+τ. The more the value of the coefficient $V_{t,t+\tau}^{(i)}$ deviates from zero, the greater the structural changes in the comparable periods.

Table 5 shows the calculated values of coefficient (1) for Poland and selected countries which joined the EU in different years.

[1] This measure is based on Rutkowski's proposal [1981], see Malina [2004].

Table 5. Measures of the extent of structural changes in Poland and selected EU countries in period 1980—2004.

Years	Poland	EU	U.K.	Ireland	Greece	Spain	Portugal
1985/1980	0.053	0.104	***0.203***	***0.101***	0.053	0.092	0.134
1990/1985	0.058	0.018	0.018	0.097	***0.167***	***0.103***	0.098
1995/1990	***0.125***	***0.299***	0.030	0.052	0.092	0.067	***0.173***
2000/1995	0.084	0.053	0.059	0.106	0.088	0.073	0.075
2004 / 2000	0.064	0.046	0.038	0.056	0.039	0.045	0.039

Source: Author's own research.

The analysis of the data included in Table 5 indicates that EU employment structure changed considerably in 1995 as compared with 1990 (V =0.299), and in 1985 as compared with 1980. This may be due to the accession of new countries (Greece in 1981, Spain and Portugal in 1986). In those periods a structural equilibrium of the EU was, to some extent, disturbed (see: D. Strahl, 1996). In the subsequent years, however, the value of the coefficient recorded a systematic decrease, which implies the recovery of structural equilibrium.

Poland recorded major employment structural changes at the beginning of the economic transformation, in 1990—1995 ($V_{t,t+\tau}$ = 0.125). Major adjustment changes after the accession of new countries to the EU are usually recorded after the period of 5 years (e.g. Greece, Portugal — see: Table 5).

The conducted analysis indicates that the accession of new countries has a major impact on maintaining the stability of the system (considerable changes to employment structure parameters and the measures of the extent of structural changes). EU countries, being at a high, but varying level of economic advancement, are characterized by economic stability. The conducted analysis indicates that those countries are in the stage of relative economic equilibrium. It is the fear of the loss of that equilibrium that makes those countries set a high level of accession criteria. Therefore, countries applying for membership have to meet a number of legal, financial and economic requirements.

5. Classification of EU countries in terms of similarities of employment structure in economy's basic sectors

Ward's method was applied to classify EU countries in terms of similarities in their employment structures in basic industries. The classification was carried out for the enlarged EU (EU-25) on the basis of *Eurostat*'s data in 2004.

Table 6 presents EU classification results, showing the average shares of the working population in the particular industries (%) in selected groups of countries. The last column shows the value of GDP per capita for selected groups

of countries as a percentage of GDP in the entire EU (EU-25). Table 7 shows the calculated taxonomic distances between the selected groups of countries.

The classification of EU-25 clearly indicates that so called *old* and *new* members (with the exception of Malta and Cyprus) represent different types of the industry-based economic structure. The first group includes highly advanced nations, in which the service industry represents the highest proportion of the working population (74%), and where the proportion of farming-related jobs is very low (2.9%). The average income per capita in this group of countries exceeds EU levels by 33.5%. Among the countries which joined the EU in 2004, only Hungary and Estonia bear a substantial similarity to EU-15 in terms of employment structure. These countries are classified in the same group as Germany, Austria, Italy, Spain and Ireland. The other states representing so called post-Communist block countries of Central and Eastern Europe, along with Greece and Portugal, are divided into two groups (groups 3 and 4), marked by similar differences as compared with group one representing the most advanced nations. In group three the average income per capita represents about 70% of the income in EU-25, while in group four, including Poland, Greece, Lithuania and Latvia, only 56.6%. It should be noted that group three is marked by the dominance of industrial economy, while group four by farming (see: Table 6). The share of the service sector in those groups is relatively low (54% — 56 %).

Table 6. Average share of employed population in industries, and GDP per capita (EU25 =100) in EU selected groups of countries in 2004.

Group	Countries	Share of total employment (%)			GDP per capita (EU25=100)
		Agriculture Sector I	Industry Sector II	Services Sector III	
Gr.1	Belgium, Denmark, France, Netherlands, Luxembourg, Sweden, U.K., Cyprus	2.9	23.3	73.8	133.5
Gr.2	Germany, Italy, Austria, Finland, Spain, Ireland, Malta, Estonia, Hungary	5.2	30.6	64.2	108.6
Gr.3	Czech Republic, Slovakia, Slovenia, Portugal	8.4	37.6	54.0	70.2
Gr.4	Latvia, Lithuania, Greece, ***Poland***	17.3	26.1	56.6	56.3

Source: Author's own research.

Table 7. Taxonomic distances between selected group of countries.

Group	Gr. 1	Gr. 2	Gr.3	Gr. 4
Gr. 1	0.0000	0.0713	***0.1454***	***0.1406***
Gr. 2	0.0713	0.0000	0.0762	0.0870
Gr. 3	***0.1454***	0.0762	0.0000	0.0862
Gr. 4	***0.1406***	0.0870	0.0862	0.0000

Source: Author's own research.

The conducted research confirms that economic advancement is coupled with a considerable decline in the share of farming and a growing significance of services.

6. Conclusions

On the basis of the conducted analysis, it may be concluded that the share of the working population in industries I and II (farming and industry) is decreasing both in EU countries and in Poland, while the share of the service sector is increasing. Similar differences are recorded in the industry-based employment structure of the particular countries (excluding the periods of admitting new member states). The lowest diversification of EU employment structure is recorded in sector II, which is industry and construction, and the highest in the service sector.

The accession of Greece (1981), Spain and Portugal (1986), and CEE countries (2004) had a major influence on the basic parameters of employment structure (average value, standard deviation), while the admission of Austria, Finland and Sweden in 1995 did not lead to changes in EU employment structure parameters.

The analysis of EU industry-based employment structure indicates that new member states became increasingly *similar* to EU-founding countries in the period under investigation.

Apart from the similar tendencies of changes in the industry-based employment structure in Poland and other new member states, there are still considerable differences between Poland and EU-15. Poland has a high share of farming, which is four times as high as in EU-15. The conducted analysis indicates that the employment structure of Poland is most similar to that of Lithuania, Latvia and Greece. There is a considerable gap, however, between the aforementioned countries and EU-15.

REFERENCES

GŁOWACKA, J.E. (1995), Labour Market Trends in EU Countries (in Polish), „*Gospodarka Narodowa*", no 5.

GŁĄBICKA, K. (2000), Labour Market in EU Countries. State and Perspectives (in Polish), *Publishers „Difin"*, Warsaw.

KARPIŃSKI, A. (1996), Labour's Market Changes to 2010 (in Polish), "*Labour Marke*t", no 7.

KARPIŃSKI, A., PARADYSZ, S., ZIEMECKI, J. (1999), Changes of Economy Structure in Poland to 2010. Poland compared to EU Countries (in Polish), *Publishing House ELIPSA*, Warsaw.

MALINA, A. (2004), Multidimensional Analysis of the Spatial Differentiation of Poland's Economic Structure by Voivodship (in Polish), Cracow University of Economics, *„Science Pape*rs", special series*: Monografie*, no 162, Cracow.

RUTKOWSKI, J. (1981), Similarity of Structure and Structural Changes. Problems of Quantitative (in Polish) „*Wiadomości Statystyczne*", no 8.

STRAHL, D. (1985), Economic Structures Similarity (in Polish) „*Science Papers Economic Academy in Wrocław*", no 281, Wrocław.

STRAHL, D. (1996), Structural Equilibrium of Economic Object (in Polish), (in:) *Time Spatial Modelling and Predicting of Economic Phenomena,* (editor: A. Zeliaś), Cracow University of Economics Publishers, Cracow.

STRAHL, D., PANASIEWICZ Z. (1997), Long-term Patterns of Polish Structural Transition (in Polish), *International Compared Study*, IRiSS, Warsaw.

STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 383—398

META ANALYSIS: WHAT, WHY AND HOW

Chandra Bhushan Tripathi[1], Prem Chandra[2], Neeraj Pandey[1] and Nilanjan Roy[3]

ABSTRACT

Meta-analysis was proposed more than 20 years ago as an innovative technique for pooling the results of a series of clinical studies. Meta-analysis has acquired a substantial following among both statisticians and clinicians. The technique was developed as a way to summarize the results of different research studies of related problems. Meta-analysis may be applied even when the studies are small and there is substantial variation in the specific issues studied, the research methods applied, the source and nature of the study subjects, and other factors that may have an important bearing on the findings. A meta-analysis aims at gleaning more information from existing data by pooling the results of smaller studies and applying one or more statistical techniques. Biologists often consider meta analysis to be a simple way to summarize the existing knowledge and examine the actual strength of risk factor. The benefits or hazards that may not be detected in small studies can be found in meta-analysis that uses data from large scale studies. In this paper, we describe various methodological steps i.e., need of meta analysis, details about statistical methods used for the analysis, and its uses and limitations.

The paper describes the steps involved for carrying out meta-analysis. Brief overview of statistical methods, rationales for using these methods as well as formulae is discussed.

Meta-analysis can be very useful method for combining data across the studies, but it requires careful thought, planning and implementation.

Keywords: Meta-analysis, hazards, evidence-based medicine, fixed & random effect model.

[1] Department of Biostatistics, Institute of Human Behavior and Allied Sciences (IHBAS), Delhi.
[2] Department of Biostatistics, All India Institute of Medical Sciences (AIIMS), New Delhi.
[3] Department of Biotechnology, National Institute of Pharmaceutical Education & Research, (NIPER), Mohali, Chandigarh.

1. Introduction

Meta-analysis has acquired a substantial following among both statisticians and clinicians. The technique was developed as a way to summarize the results of different research studies of related problems. Meta-analysis may be applied even when the studies are small and there is substantial variation in the specific issues studied, the research methods applied, the source and nature of the subjects under study, and other factors that may have an important bearing on the findings. The core of meta-analysis is its systematic approach to the identification and abstracting of critical information from research reports. Doing meta-analysis correctly demands expertise in both the method and the substance and hence almost always requires collaboration between clinicians and an experienced statistician. The questions must be defined carefully to maximize the relevance of the reports to be included and to reduce uncertainties about procedures. The investigators must then try to find every relevant report by searching data bases, reviewing bibliographies, and asking widely about unpublished work.

Meta-analysis in clinical research is based on simple principles: systematically searching out, and, when possible, quantitatively combining the results of all studies that have addressed a similar research question. Principles of evidence-based methods to assess the effectiveness of heath care interventions and set policy are cited increasingly (Chalmers & Lau 1993, Basgett et al. 1997, Ohlsson 1994). There has been a growing international interest in the development of measures to ensure that public policy and practice decision making are based on better informed results of relevant and reliable research. This interest has been fuelled by evidence that some health and social interventions which have been commonly applied in the belief that they were doing good are actually harmful, that others are largely ineffective and thus wasteful of public resources. Furthermore, some interventions which reliable research shown to have significant benefits have been largely ignored. By bringing together the result of research in a systematic way, appraising its quality in the light of the question being asked, summarizing the results in an explicit way it is hope to achieve a greater sensitivity to the evidence by researchers, policy maker, practitioners and the public.

How can large amounts of independent quantitative information on the same question come together in coherent and meaningful manners? Many researchers have relied on this statistical technique to achieve such synthesis of evidence. Traditionally clinicians have relied for therapeutic guidance on traditional journal review articles and textbook chapters. Review articles and book chapters are generally tend to make selective appraisals of the evidence and usually do not provide a quantitative synthesis of the data (Mulrow 1987). While it can be applied to data of any sort, meta-analysis has been commonly used to combine results from different studies to draw conclusions. Meta-analysis has been defined as the ‘Statistical analysis of a large collection of analysis results from individual

studies for the purpose of integrating the findings' (Glass 1976). Although there has always been some controversy about its validity (Liberati 1995, Bailar 1997, Dickersin & Berlin 1992, Thompson & Pocock 1991, Chalmers 1991, Kassirer 1992), meta-analysis has become increasingly popular, as the number of studies with similar protocols has grown. By systematically combining studies, one attempts to overcome limits of size or scope of individual studies to obtain more reliable information about gene mutation malfunction or treatment effects.

In meta analysis the results of the various studies are tabulated compared, analyzed and discussed. Since it analyzes the results of the individual studies into new results a meta-analysis is more than a literature review. A meta-analysis also differs from a 'pooled data' analysis because the summary measures of the selected previous studies are combined for analysis not the results on individual subjects.

The extension of this research method in the pharmaceutical sciences has paralleled the explosion in the numbers of conducted randomized trials . Meta analysis has become an increasing need to provide an unbiased quantitative analysis in the era of evidence-based medicine. Meta-analysis has been used mostly for analysis of data from completed trails, but there is an increasing momentum for using the same principles of data analysis in the prospective design of clinical trails addressing a similar question (Laupacis et al. 1991, Henderson et al. 1995).

Meta-analysis has also been applied to epidemiologic studies to calculate better risk estimates of gene mutations and for evaluations of diagnostic tests to obtain more reliable estimates of test performance. In this paper, we focused, only on meta-analysis odds ratios.

2. Benefits & limitations

Meta analysis offers a rational and helpful way of dealing with a number of practical difficulties that beset anyone trying to make sense of effectiveness research. We perform meta-analyses because we often have a lot of information, from many studies, sometimes contradictory, and meta-analysis offers us a tool to help to integrate this. However, we should always remember that meta-analysis is only a tool, and it is simply one of many tools we use to help us to understand what a literature is trying to tell us, if anything. The benefits associated with meta-analysis and its possible limitations are enumerated as under:

2.1.Benefits:

(i) A clearer picture

Individual clinical trials may mean little, especially when they are small or medium sized. Small studies tend to be inconclusive, they may show no statistical difference between the treated and control groups, but on the other hand they may be unable to exclude the possibility of there being a sizeable effect (that is, they have low power). Aggregating studies in a systematic and unbiased way may allow a clearer picture to emerge. The question we are asking is whether, on average, a particular treatment confers significant benefits when used for specific patient groups. Meta analysis allows this aggregate picture to emerge.

(ii) Overcoming bias

The danger of unsystematic reviews is that there is plenty of scope for bias. Certain (perhaps favorable) reports may be preferred over those that show no benefit. Informal synthesis may be tainted by the prior beliefs of the reviewer. Meta analysis carried out on a rigorous systematic review can overcome these dangers offering an unbiased synthesis of the empirical data. Meta-analysis may reveal how heterogeneity among populations affects the effectiveness of medical interventions in different settings and in different patients (Loannidis & Lau 1999).

(iii) Precision

The precision with which the size of any effect can be estimated depends on (among other things) the number of patients studied. Meta analyses that draw on patients studied in many trials thus have more power to detect small but clinically significant effects, and can give more precise estimates of the size of any effects uncovered. This may be especially important when an investigator is looking for beneficial (or deleterious) effects in specific subgroups.

(iv) Transparency

It is not simply the case that meta-analyses can always exclude bias better than other forms of review. Their advantage also lies in the openness with which good meta-analyses reveal all the decisions that have been taken throughout the process of achieving the final aggregate effect sizes. Thus, good meta-analyses should allow readers to determine for themselves the reasonableness of the decisions taken and their likely impact on the final estimate of effect size. The rapidly increasing volume of research, often with discrepant findings, has led to an increased need for meta-analysis. The 225 published meta-analyses of randomized controlled trials in 1993, compared with 7 in 1980. At least 500 meta-analysis of randomized trails were published in 1996 and 1997(Chow 2000).

Finally when performed prospectively, a meta-analysis may allow advance planning in the collection and analysis of large-scale evidence.

2.2.Limitations:

Many criticisms have been imposed at meta-analysis (Thacker 1988, Goodman 1991). Meta-analysis cannot improve the quality or reporting of the original studies. Limitation may come from misapplications of the method, such as when study diversity is ignored or mishandled in the analysis or when the variability of patient populations, the quality of data, and the potential for underlying biases are not addressed. Publication bias is a major limitation of using only published studies. It has been accepted that research with statistically significant results is potentially more likely to be submitted, published or published more rapidly than work with null or non-significant results (Easterbook et al. 1991), which leads to a preponderance of biased results in the literature (Begg & Berlin 1989). This indicates for meta-analysis as combining only the identified published studies uncritically may leads to an incorrect, usually overoptimistic conclusion.

Publication bias must be detected by Begg and Mazumdar test (Begg 1994) and a funnel plot in which a plot of sample size versus treatment effect from individual studies in a meta-analysis should thus be shaped like a funnel if there is no publication bias (Light & Pillemar 1984). If the chance of publication is greater for studies with positive statistically significant result, or larger effect size estimates, or some other less defined mechanism, the shape of the funnel plot may become skewed. One analytic way to identify the impact of potential publication bias is the " fail-safe N" approach (Rosenthal 1979), which estimates the number of additional studies with null results required to reverse statistically significant findings of a meta-analysis. Although no satisfactory solution exists yet for correcting publication bias, research registries (Easterbook 1992, Dickersin 1994) and other reasonable means (following upon published abstract) can be used to track down unpublished studies.

3. Methodological steps involved while doing Meta Analysis:

- Decide on the topic.
- Decide on the hypothesis being tested.
- Review the literature for all studies which test that hypothesis. While this literature review may begin with a computerized search of the literature, such searches may miss important studies. Therefore other methods should be applied, such as careful study of the references in articles, examination of papers, abstracts, and presentations not published, and

other sources of unpublished (including government agencies, and rejected submissions). This needs to be done carefully to minimize bias.

- Evaluate each study carefully, to decide whether it is of sufficient quality to be worthy of inclusion, and whether it includes sufficient information to be included. This includes attention to endpoints, choice of the measure of effect size, and other information about quality. This task, too, needs to be done carefully to minimize bias.
- Create a database containing the information necessary for the analyses.
- Perform the meta-analysis
- Interpret the results.

4. Initial Requirements of Meta-analysis

(a) Study Identification

A difficult but crucial component is to identify the studies for meta-analysis. The study identification process should result in identification of all studies that could potentially be included in the meta-analysis. It should include published or unpublished studies, case clusters and observational epidemiologic studies. A search of computerized database (e.g., MEDLINE) may provide a reasonable start but is often insufficient because not all studies are included in such database. In particular, unpublished studies, studies published only as abstracts or published in journals which are not in database. Of course, identifying all such data may be impossible, but one should search through each identified report to find references to as yet unidentified reports (e.g. personal communication or unpublished data). One may also inquire among researchers in the topic area to acquire unpublished data on the topic. In the end, however, one's results will be vulnerable to bias due to systematic failures of other investigators to publish or report certain data (e.g. null results). Restricting one's analysis to published articles may only aggravate such bias.

(b) Data Extraction

Data to be extracted from individual studies for a meta-analysis include the effect estimate and an estimate of its variance and descriptive information about the study such as sample size, health outcome, exposure measure, design, time frame and control for confounding. The major issues that arise in connection with data extraction for meta-analysis of epidemiologic studies are introduction of bias in the selection of information and the sufficiency and quality of information available. Although the researchers agreed on the need to avoid bias, consensus was not reached on whether to use blinding where, individuals who extract data are blind to the journal, study location, study authors and study outcome, to minimize bias during data extraction.

5. Methods of analysis

In this section, the methods of meta-analysis will be discussed for only those study and results, which have been calculated from a two-by-two contingency table, with two groups and a binary outcome. A compendium of methods for combining continuous data as well as discrete data, can be found in The Handbook of Research Synthesis (Cooper & Hedges 1994) which is the most comprehensive book on statistical as well as non statistical matters pertaining to meta-analysis.

5.1. Assessment of Heterogeneity/Homogeneity

Once the studies selected and data have been abstracted from the studies for meta-analysis, effect sizes must be tested for homogeneity. In any meta-analysis, it is quiet obvious that point estimate of effect size will always differ to some degree. It is expected and is partly due to sampling error, which is present in every estimate. When effect sizes differ, but only due to sampling error, the effect estimates are considered to be homogenous, in other words difference between estimates are random variation and not due to systematic differences between studies (Sutton et al. 2000). A test of homogeneity is used to test whether there is evidence that the treatment effects are different across studies. Several slightly different formulae for general test are, for the most part, almost equivalent, being based on χ^2 (chi-square) or F-statistic (Dickersin & Berlin 1992). The statistics developed by Cochran (Cochran 1954), which is widely used, as follows:

$$Q = \sum_{i=1}^{k} w_i \left(T_i - \bar{T}\right)^2$$

Where k is the number of studies being combined, Ti is the treatment effect estimate in the ith study, $\bar{T} = \frac{\sum_{i=1}^{k} W_i T_i}{\sum_{i=1}^{k} W_i}$ is the weighted estimator of treatment effect, and w_i is the weight attached to that study usually calculated as the reciprocal of the variance (ν_i) of the outcome estimate from the ith study, in the meta-analysis. For example- if odds ratios (ad/bc) were used to combine the k studies in a meta-analysis, then each T_i could be calculated by taking the logarithm of odds ratios and equation $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$ could be used to calculate each corresponding ν_i (where a, b, c, d are frequency of 2x2 contingency table). A more convenient from of Q is

$$= \sum_{i=1}^{k} w_i T_i^{\,2} - \frac{\left(\sum_{i=1}^{k} w_i T_i\right)^2}{\sum_{i=1}^{k} w_i}$$

Q is approximately distributed as chi-square (χ^2) distribution with k-1 degree of freedom hence χ^2 tables can be used to obtain the corresponding p-value (White et al. 1979). Essentially, it tests whether it is reasonable to assume that all the studies to be combined are estimating a single underlying population parameter, and whether variation in study estimates is likely to be wholly random.

Variation between individual study estimates can be examined graphically too using forest plots. These plots are most common way for presentation of results of a meta-analysis, however, they are also useful at the exploratory stage of meta-analysis. Each study's effect size and their respective confidence intervals are plotted on one set of axes. For the purpose of reporting results, the combined estimate of effect size with its confidence interval is also plotted. Often the size of the plotting symbol used to mark the point estimate from each study is made proportional to the reciprocal of the variance of the estimate. Hence the more precise estimates are given the largest plotting symbols. From this plot, an idea about the variability of estimates from individual selected studies can be viewed.

After assessing the heterogeneity of effect size across the selected studies, the type of model should be decided for combining the effect size of all studies. Basically, there are two models have been employed to adjust appropriately the potential confounding effect of study.

(1) Fixed effect model

(2) Random Effect model

5.2. Fixed effect model

This assumes that all the studies in a meta-analysis are estimating the same underlying unknown true intervention effect. No heterogeneity between the study result is assumed and the only variability of treatment is within each study. Here we will discuss how to combine the odds ratios calculated from case-control studies.

(i) Mantel-Haenszel Method

This method was first described by Mantel-Haenszel (Mental & Haenszel 1959) for combining odds ration for stratified case-control studies. Later, Mantel (1963) reported the method that could be used for a wider class of problem. The combined estimate is calculated by

$$\bar{T}_{MH(OR)} = \frac{\sum_{i=1}^{k} a_i d_i / n_i}{\sum_{i=1}^{k} b_i c_i / n_i}$$

Where a_i, b_i, c_i, and d_i are the frequency of 2X 2 tables (four cells) for the i = 1,2,3,------K studies. Variance estimate for pooled odds ratios $T_{MH(OR)}$ can be calculated by the formula as

$$\nu_{MH(\ln(OR))} = \frac{\sum_{i=1}^{k} P_i R_i}{2(R_i)^2} + \frac{\sum_{i=1}^{k} (P_i S_i + Q_i R_i)}{2\left(\sum_{i=1}^{k} R_i\right)\left(\sum_{i=1}^{k} S_i\right)} + \frac{\sum_{i=1}^{k} Q_i S_i}{2\left(\sum_{i=1}^{k} S_i\right)^2}$$

Where $P_i = (a_i + d_i) / n_i$, $Q_i = (b_i + c_i) / n_i$, $R_i = a_i d_i / n_i$ and $S_i = b_i c_i / n_i$

Variance is required to calculate a confidence interval of pooled odds ratio. This formula calculates variance estimate for log of $T_{MH(OR)}$ and was derived by Robins et al. (1986). 100 (1-α) % confidence interval for summary odds ratio θ is given by

$$\text{Exp}\left[\ln\left(\overline{T}_{MH(OR)}\right) - Z_{\alpha/2}\left(\nu_{MH(OR)}\right)^{1/2}\right] \le \theta \le Exp\left[\ln\left(\overline{T}_{MH(OR)}\right) + Z_{\alpha/2}\left(\nu_{MH(OR)}\right)^{1/2}\right]$$

Where $Z_{\alpha/2}$ is the α/2 percentage point of a Standardized Normal Distribution

(ii) Peto's Method

This method was first described by Peto (1977) and more thoroughly by Yusuf et al. (1985). An advantage over the Mantel-Haenszel method is that it can be used when cell of 2x2 tables in individual studies are zero. Unfortunately, this method may produce serious under estimate (Fleiss 1994) when the odds ration is far from unity. For K-studies the combined estimate of odds ratio is given by

$$\bar{T}_{PETO(OR)} = \text{Exp}\left[\frac{\sum_{i=1}^{k} (O_i - Ei)}{\sum_{i=1}^{k} \nu_i}\right]$$

Where $\nu i = E_i [(n_i - n_{ti}) / n_i] [(n_i - d_i) / (n_i - 1)]$

O_i is the number of cases in exposed group

E_i is the expected number of cases in the exposed group and can be calculated by = $(n_{ti} / n_i) d_i$

Where n_i is the total number of study subjects in the study

n_{ti} is the total number of study subjects in the exposed group

d_i is the total number of cases (exposed + unexposed)

An estimate of approximate variance of the natural log of the estimated combined odds ratio is given by

$$\text{Var}\,(\ln \overline{T}_{PETO(OR)}) = \left(\sum_{i=1}^{k} \nu_i\right)$$

100 (1- α)% confidence interval is given by

$$\text{Exp}\left[\frac{\sum_{i=1}^{k}(O_i - E_i) \pm Z_{\alpha/2}\sqrt{\sum_{i=1}^{k}\nu_i}}{\sum_{i=1}^{k}\nu_i}\right]$$

(iii)General fixed effect model – the inverse variance weighted method

This method was first described by Birge (1932) and Cochran (1937) in the 1930s. In this method, each estimate is given a weight directly proportional to its precision (inversely proportional to its variance).

Let Ti (i=1, 2, ----k) be the observed effect size for the studies selected in meta-analysis, θi the underlying population effect size, with variance νi. For a fixed effect model, all population effect sizes are assumed equal (i.e. $\theta_1 = \theta_2 = \theta_3$ =--------$\theta_k = \theta$, where θ is the true common underlying effect size). The combined estimate is given by

$$\overline{T} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i}$$

Where $w_i = 1 / \nu_i$

For odds ratio νi (variance) = $1/ a_i + 1/ b_i + 1/c_i + 1/d_i$

Where a , b , c and d are the cell frequencies of 2×2 table.

Estimate of variance of the pooled estimate $\overline{T}$ is given by

$$\text{Var}\,(\overline{T}) = \frac{1}{\sum_{i=1}^{k} w_i}$$

100 (1- α)% confidence interval for the population effect θ is given by

$$\overline{T} - Z_{\alpha/2}\sqrt{1/\sum_{i=1}^{k} w_i} \le \theta \le \overline{T} + Z_{\alpha/2}\sqrt{1/\sum_{i=1}^{k} w_i}$$

5.2. Random Effect Model

In the fixed-effect model, which leads to inferences only about the studies assembled, the true treatment effect is assumed to be the same across all studies and the only variability of treatment effect considered is within each study (i.e. within-study variability). In the random-effect model, which leads to inference about all studies from a hypothetical population, each study can have a different effect that is randomly drawn from a normal distribution and is positioned about a central value; in this model, not only within-study variability but also variability of treatment affects across studies (between study–variability) is considered. Along with weighting each study by its sampling variability of treatment effect, which includes the number of events and sample size in each treatment group of a study, the random effect model also weights each study by an overall estimate of the variability of true treatment effects across studies. The random-effect model tends of give wider confidence intervals, as compared to fixed effect model, which makes the statistical significance of its results more conservative than those in the fixed effects model.

The standard random effects model applied in meta-analysis was described by Der Simonian and Laird (1986). The model assumes that the study specific effect sizes come from a random distribution of effect sizes with fixed mean and variance. Algebraically, the model can be expressed as

$$T_i = \theta_i + E_i$$

Where T_i is an estimate of effect size, θ_i is the true effect size in the ith study and e_i is the error with which T_i estimates θ_i and

$$\mathrm{Var}\,(T_i) = \tau_\theta^2 + \nu_i$$

where τ_θ^2 is the random effects variance and ν_i is the variance due to sampling error. If random effects variance is zero, the above model becomes fixed-effects model. The treatment point estimate for the mean treatment effect of all studies θ can then be computed by $\overline{T}_{.RND} = \dfrac{\sum_{i=1}^{k} w_i^* T_i}{\sum_{i=1}^{k} w_i^*}$

where w_i^* are adjusted weights for each of the studies and may now be calculated as

$$w_i^* = \frac{1}{\left[\left(1/w_i\right) + \hat{\tau}^2\right]},$$

where $\hat{\tau}^2$ denotes the between study variance of the studies effect sizes. The estimated component of variance due to inter study variation in effect size, $\hat{\tau}^2$, is calculated as

$$\hat{\tau}^2 = 0 \text{ if } Q \le k\text{-}1 \text{ and } \hat{\tau}^2 = [\,Q\text{-}(k\text{-}1)\,] / U \text{ if } Q > k\text{-}1$$

where Q is the heterogeneity test statistic defined earlier and

$$U = (k\text{-}1)\,(\,\overline{w} - s^2_w / k\overline{w}\,) \text{ and } s^2_W = \frac{1}{k-1}\left(\sum_{i=1}^{k} w_i - k\overline{w^2}\right) \;\&\; \overline{w} = \frac{\sum_{i=1}^{k} w_i}{k}$$

The variance of this estimate is simply given as

$$\text{Var}\,(\overline{T}_{.RND}) = 1\Big/\sum_{i=1}^{k} w_i^*$$

100 (1- α) % confidence interval can be calculated by

$$\overline{T}_{.RND} - Z_{\alpha/2}\Big/\sqrt{\sum_{i=1}^{k} w_i^*} \le \theta \le \overline{T}_{.RND} + Z_{\alpha/2}\Big/\sqrt{\sum_{i=1}^{k} w_i^*}$$

6. Area of application

Meta-analysis has been used extensively in human clinical trials and in the social sciences. Its use in epidemiology has been much more limited even though it can potentially enhance the value of epidemiologic data. Many applications of this technique appear in the literature. The results of meta-analyses of pharmaceutical have helped to determine optimal therapeutic use and to enhance the understanding of pharmaceuticals with respect to their efficacy, safety, bioequivalence and cost-effectiveness. Meta-analysis may be particularly applied when (Blair et al. 1995):

(i) there are many studies but no consensus.
(ii) refinement of the estimate of an effect is important
(iii) there is a need to increase the statistical power more than of single study
(iv) there are questions about generalizability of the results

(v) sources of heterogeneity are to be examined. For example, the cost effectiveness of enoxaparin was compared with that of low-dose warfarin in the prevention of deep-vein thrombosis (DVT) after total hip replacement (O'Brien et al. 1994). To know the proportion of DVT with enoxaparin and warfarin prophylaxis, the authors pooled the proportions from the available studies to arrive at the efficacy of the drugs for input into determining that the incremental cost-effectiveness of enoxaparin relative to warfarin was estimated as $ 29, 120 per life-year gained (in Canadian dollars)

REFERENCES

CHALMERS TC, LAU J. Meta-analysis stimulus for changes in clinical trials. Stat Methods Med Res. 1993;2:161—172

BADGETT RG, O' KEEFE M, HENDERSON MC. Using systematic reviews in clinical education. Ann Intern Med. 1997;126:886—891

OHLSSON A. Systematic reviews. SCAND J Clin Lab Invest Suppl. 1994;219:25—32

MULROW C. The medical review article: state of the science. Ann Intern Med. 1987;106:485—488

GLASS GV. Primary, secondary and meta-analysis of research. Educ Res.1976;5:3—8

LIBERATI A. Meta-analysis: statistical alchemy for the 21st century: discussion. A plea for a more balanced view of meta-analysis and systematic overviews of the effect of health care interventions. J clin Epidemiol. 1995;48:81—86

BAILAR JC 3rd. The promise and problems of meta-analysis. N Engl J Med. 1997;337:559—561

DICKERSIN K, BERLIN JA. Meta-analysis: state -of –the-science. Epidemiol Rev. 1992;14:154—176

THOMPSON SG, POCOCK SJ. Can meta-analyses be trusted? Lancet. 1991;338:1127—1130

CHALMERS TC. Problems induced by meta-analysis. Stat Med.1991;10:971—979

KASSIRER JP. Clinical trials and meta-analysis. What do they do for us. N Engl J Med. 1992; 327:273—274

LAUPACIS A, CONNOLLY SJ, GENT M, ROBERTS RS, CAIRNS J, JOYNER C. How should results from completed studies influence ongoing clinical trials? The CAFA study experience. Ann Intern Med. 1991; 115:818—822

HENDERSON WG, MORTIZ T, GOLDMAN S, COPELAND J, SETHI G. Use of cumulative meta-analysis in the design, monitoring and final analysis of a clinical trials: A case study. Controlled Clin Trials. 1995;16:331—341

IOANNIDIS JP, LAU J. Pooling research results: benefits and limitations of meta-analysis. Jt Comm J Qual Improve. 1999;25(9):462—469

CHOW S-C. Encyclopedia of Biopharmaceutical Statistics. Marcel Dekker, INC. New York. 2000: 307

THACKER SB. Meta-analysis: A Quantitative approach to research integration. JAMA. 1988; 259:1685—1689

GOODMAN SN. Have you ever seen a meta-analysis you didn't like? Ann Inern Med. 1991; 114:244—246

EASTERBROOK PJ BERLIN JA, GOPALAN R, MATTHEWS DR. Publication bias in Clinical research. Lancet. 1991; 337: 867—872

BEGG CB, BERLIN JA. Publication bias and dissemination of clinical research. J Natl Cancer Inst. 1989; 81:107—115

BEGG CB. In: Cooper H, Hedges LV (editors). The Handbook of Research Synthesis. New York: Russell Sage Foundation. 1994: 399—409

LIGHT RJ, PILLEMAR DB. Summing up: The science of Reviewing Research. Cambridge, MA: Harvard University Press.1984

ROSENTHAL R. The "file-drawer problem" and tolerance for null results. Psychol Bull. 1979; 86:638—641

EASTERBOOK PJ. Directory of registries of clinical trials. Stat Med. 1992; 11: 345—359

DICKERSIN K. In: Cooper H, Hedges LV (editors). The Handbook of Research Synthesis. New York: Russell Sage Foundation. 1994:71—83

COOPER H, HEDGES LV (editors). The Handbook of Research Synthesis. New York: Russell Sage Foundation. 1994

SUTTON AJ, ABRAMS KR, JONES DR, SHELDON TA, SONG F. Methods for Medical Research. John Wiley & Sons, Ltd.2000

COCHRAN WG. The combination of estimates from different experiments. Biometrics. 1954;10:101—129

WHITE J, YEATS A, SKIPWORTH G. Tables for statisticians. 3rd ed. Cheltenham: Stanley Thornes.1979

MANTEL N, HAENSZEL W. Statistical aspects of the analysis of data from retrospective studies of disease. J Nat Cancer Inst. 1959; 22:690—700

MANTEL N. Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel Procedure. J Am Statist Assoc. 1963; 58:690—700

ROBINS J, BRESLOW NE, GREENLAND S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. Biometrics. 1986; 42: 311—323

ROBINS J, GREENLAND S, BRESLOW NE. A general estimator for the variance of the Mantel-Haenszel odds ratio. Am J Epidemiol. 1986; 124: 719—723

PETO R, PIKE MC, ARMITAGE P, BRESLOW NE, COX DR, HOWARD SV, MANTEL N, MCPHERSON K, PETO J, SMITH PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II: analysis and examples. Br J Cancer. 1977; 35: 1—39

YUSUF S, PETO R LEWIS J, COLLINS R, SLEIGHT PETAL. Beta blockde during and after myocardial infarction : an overview of the randomized trials. Progress in Cardiovascular Diseases. 1985; 27:335—371

FLEISS JL. Measures of effect size for categorical data. In: Cooper H, Hedges LV (editors). The Handbook of Research Synthesis. New York: Russell Sage Foundation. 1994: 245—260

BIRGE RT. The Calculation of errors by the method of least squares. Phys Rev. 1932; 16: 1—32

COCHRAN WG. Problem arising in the analysis of a series of similar experiments. J Roy Statist Soc. 1937; 4 (supplement)102—118

DERSIMONIAN R, LAIRD N. Meta-analysis in clinical trials. Controlled Clin Trial. 1986; 7:177—188

BLAIR A, BURG J, FORAN J, GIBB H, GREENLAND S, MORRIS R, RAABE G, SAVITZ D, TETA J, WARTENBERG D WONG O, ZIMMERMAN R. Guidelines for environmental epidemiology. Regulatory Toxicology and Pharmacology. 1995; 22:189—197

O' BRIEN BJ, ANDERSON DR, GOEREE R. Cost-effectiveness of enoxaparin versus warfarin prophylaxis against deep-vein thrombosis after total hip replacement, Can. Med Assoc J. 1994; *150* :1083—1090.

STATISTICS IN TRANSITION-new series, August 2007
Vol. 8, No. 2, pp. 399—402

CELEBRATING STATISTICS

Conference in Honour of Professor Kazimierz ZAJĄC on the Occasion of his 90th Birthday, Cracow, Poland, 26th October 2006

On 26 October 2006 the Committee on Statistics and Econometrics of Polish Academy of Sciences solemnly celebrated Professor Kazimierz Zając's 90th birthday. More than sixty Polish statisticians and econometricians participated in the meeting. The ceremony was led by the honorary Chairman of the Committee Professor Wiesław Sadowski. At the beginning Professor Ryszard Borowiecki, the Rector of the Cracow University of Economics spoke of merits of the Nestor of Polish statisticians. He emphasised Professor Zając's services to the University wishing him long and scientifically active years in good health. Then the Dean of the Faculty of Management, Professor Józef Pociecha, presented Professor Zając's scientific path. Professor Kazimierz Zając in his speech emphasised the role of his master and predecessor at the Department of Statistics Professor Jerzy Fierich.

Professor Kazimierz Zając was born on September 20, 1916, in Krosno, Poland. In 1938 he started studies at the Academy of Commerce in Cracow, which were interrupted by The Second World War. During the German occupation he worked as a teacher in Krosno. He also organised underground education and aid for army officers in prisoner-of-war camps with the Polish Red Cross. He was also engaged in armed resistance as a soldier of the Home Army (Armia Krajowa). After the war he continued studies at the Academy of Commerce in Cracow and he defended his master thesis in 1947. Before graduation he got the post of assistant–volunteer in the Department of Economy, directed by Professor Jerzy Fierich. In 1947-1948 he took part in doctoral seminar led by professor Edward Lipiński at the Warsaw School of Economics. In 1950 together with Professor J. Fierich he joined the Department of Statistics of the University of Economics in Cracow. Four years later he started working as an assistant professor at the Department of Statistics. In 1958 he got a doctoral degree upon presenting a dissertation entitled *Analysis of Workers' Wages in Industry on the Basis of Statistical Research*. In 1963 on the basis of the work entitled *Econometric Methods of Household Budgets Study* he obtained a docent degree. Professor K. Zając became the head of the Department of Statistics in 1965 after Professor J. Fierich's death. In 1971 he became an associate professor of Economics and in 1976 a full professor. In 1969-1968 Professor Zając was the director of the Institute of Economic Theory. Although Professor retired in 1986, he has continued his academic activity after that date.

As the Head of the Department Professor for 21 years inspired a team of statisticians and econometricians and educated a large group of doctors, habilitated doctors and professors thus creating a widely recognised and distinguished Cracow school of statistics and econometrics.

He had also considerable achievements in the development of the University while he was twice the vice-dean (1954-1958, 1964-1968) and the dean in 1980-1986.

It should be mentioned that Professor also was giving his teaching at the Jagiellonian University, at the University of Science and Technology (AGH) and at the Pontifical Academy of Theology in Cracow.

Professor Kazimierz Zając is a modest man and he has never sought to obtain any positions. However a great scientific authority and confidence he gained were the reason for appointing him for many other prestigious offices, including the Polish Academy of Sciences, the Central Qualifying Commission for Scientific Degrees, and the Scientific Statistical Board of the Central Statistical Office.

To acknowledge Professor's scientific, didactic and organisational distinguished services a number of state, regional and social distinctions and Orders were conferred on him. The most important are:

Knight's Cross Order of Poland Reborn (1973), Commander's Cross Order of Poland Reborn (*Polonia Restituta)* (1985), Award of the National Education

Commission of Poland (1977). For his remarkable merits Professor Zając was given in 1995 the Honorary Doctorate by the Katowice University of Economics.

Professor's scientific life is characterised by versatility and diversity in his approach to current needs. Being a follower of Professor Fierich and Professor Lipiński he gained a thorough economic knowledge. It was in the fifties that he started dealing with nominal and real wages of industrial workers and related methodology of household budget survey. Taking into consideration the state of research in the field of wages and a complete lack of subject literature Professor's studies were pioneering. His habilitation dissertation gave methodological basis for research analysis of household budgets survey results, which was restarted by the Central Statistical Office in 1957.

Other areas of Professor's scientific interest activity include inter alia:

- research on production factors in the industry with the use of Cobb-Douglas function,
- statistical quality control,
- consumer demand,
- demographic research related to socio-economic factors
- historical demography,
- tourist infrastructure,
- taxonomic methods,
- social services,
- relation between demographic and economic development.

Although Professor retired 20 years ago he still publishes both as an author and editor. His scientific output includes 17 monographs, 227 journal papers and working papers, and 13 handbooks. His textbook entitled The Outline of Statistical Methods (Zarys metod statystycznych, PWE, Warszawa) has had five editions.

Professor is also a co-initiator of the Conferences of Statisticians, Econometricians and Mathematicians from the Universities of Economics in Southern Poland. It would be hard to count all those Polish statisticians and econometricians who got promoted directly or indirectly due to Professor. He supervised 38 doctoral theses and 22 habilitation dissertations. He prepared 80 reviews of doctoral theses, 60 reviews of habilitation dissertations and 68 reviews for a professorship. Professor is regarded to be an authority by a large number of Polish statisticians and econometricians, always willing to help with his advice and experience. He is unchangingly energetic, cheerful, open and tolerant.

Ad multos annos Professor!

Józef Pociecha, Cracow University of Economics, , Poland