FROM THE EDITOR

This issue contains nine article**s** on ***sampling methods***, two **other articles**, one ***book review*** and two ***reports***.

There are following nine articles on ***sampling methods***:

1. ***Practical Methods for Sampling Rare and Mobile Populations*** (by G. Katlon from USA. Many surveys focus on a rare population that constitutes only a small proportion of the total population. When a separate sampling frame is not available for the rare population, special cost-efficient sampling methods are needed. This paper reviews the use of disproportionate stratification, screening in the context of area sampling, and location sampling for sampling rare populations. Location sampling is also used for sampling mobile populations (such as international travelers), and its use for this type of population is discussed.

2. ***Estimation of the Population Total on the Current Occassion under Second Stage Unit Rotation Pattern*** (by B. Kowalczyk from Poland). The present paper gives the theory for the estimation of the population total on the current occasion under two-stage sampling design with a partial replacement of the second stage units. Under given rotation pattern we present composite estimator of the population total, which exploits information from the previous occasion. We derive the variance of the estimator and compare it with the variance of the usual estimator. It has been shown that the composite estimator is better than the usual one applied both for a sample selected independently of the previous period and a sample selected according to the given rotation pattern.

3. ***On the Use of Modified Randomization Device for Estimating the Prevalence of a Sensitive Attribute*** (by S. Singh, S. Horn, R. Singh, N.S. Mangat from USA, Australia, India and Canada respectively). The authors consider the use of a modified randomization device in the usual unrelated question randomized response model (UY-model). The proposed model based estimator is unbiased and fares better as compared to usual UY-model estimator when the proportion of population possessing the sensitive attribute is less than 0.5.

4. ***Estimation of Population Mean Using Auxiliary Information Recent Occasion in h Occasions Successive Sampling*** (G.N. Singh from India). A general theory for estimating the population mean in the presence of auxiliary information only on the h-th (recent) occasion has been developed. Difference type estimator has been proposed and compared empirically with the sample

mean estimator and another difference type estimator, which is the combination of the sample means of the matched and unmatched portion of the sample at the h-th (recent) occasion.

5. ***Estimation of a Finite Population Variance in The Presence of Auxiliary Information*** (P.A. Patel and R.D. Chaudhari from India). This article deals with a design-based estimation of the finite population variance. Various procedures for improving variance estimation with the aid of auxiliary information are discussed. A family of estimators that includes many estimators is suggested and its variance and variance estimators are derived. An empirical study has been carried out to study the behaviour of some estimators.

6. ***Second Order Properties of Some Estimators under Double Sampling*** (by M. I. Hossain, M. S. Rahman from Bangladesh, and M. S. Ahmed from Jordan). The authors have derived two estimators on the guild line of Srivastava (1967) and Walsh (1970) in case of double sampling. The comparisons among these estimators with traditional ratio and product estimators based on double sampling scheme have been presented. They have derived the second order biases and mean square errors of traditional ratio and product estimators and our estimators. Finally, numerical comparisons are given for relative comparisons.

7. ***Use of Known Correlation Coefficient in Estimating the Finite Population Mean*** (by H. P. Singh and R.Tailor from India). This paper deals with the problem of estimating the finite population means using known correlation coefficient. An estimator is suggested with its approximate bias and mean squared error. The regions of preference are obtained under which it is better than usual ratio estimator and sample mean. An empirical study is carried out to demonstrate the performance of the constructed estimator over usual ratio estimator and sample mean.

8. ***An Alternative Approach to the Estimation of Ratio in Two — Phase Sampling*** (by L. N. Sahoo, R. K. Sahoo, G. N. Singhand L. N. Upadhyaya from India). Using two-phase sampling mechanism, estimation of ratio of two unknown population means is considered when population mean of main auxiliary variable x is unknown, but that of an additional auxiliary variable z is known. Developing a new concept, we focus attention on the creation of some new estimators as well as a general class of estimators for our purpose.

9. ***A Note on the Estimation of Mean Using Auxiliary Information*** (by L. N. Upadhyaya and Housila P. Singh from India). In this paper the authors have suggested an estimator for the population mean $\overline{Y}$ of the study variate y by making use of auxiliary information on x and z, where x is positively correlated with y and z is negatively correlated with y. The bias and mean-squared error (MSE) of the proposed estimator has been obtained under the

large sample approximations. The proposed estimator has been compared with the usual unbiased estimator $\bar{y}$ and that of Srivastava (1965). Here we confine ourselves to SRSWOR sampling scheme.

The second part of this issue under the title **Other articles** contains two articles:

1) ***Tendencies in Restructuring the Agricultural Information System in Poland*** (by K. Mateńko from Poland). The system of agricultural information in Poland is created by institutions gathering information on agriculture by means of statistical surveys, as well as by institutions keeping various data bases and administrative registers. Mostly the CSO and the Ministry of Agriculture and Rural Development conduct agricultural surveys included in annual and multi-year programmes of official statistics. Changes taking place in the 90-ties aimed at adjusting agricultural statistics, as well as, the whole of Polish official statistics to international standards, including requirements of the European Union The building of statistics which would be coherent with the EU requirements shows the necessity to restructure the whole information system concerning agriculture. As a result of the talks conducted between Eurostat and Polish bodies there was created an Inter-sectoral Experts Team on Restructuring of the Agricultural Information System.

2) ***Is it Possible to Estimate Reliable Household Equivalence Scales?*** (by A. Szulc from Poland). Equivalence scales derived from the *Almost Ideal Demand System* are estimated using Polish microdata. The purpose of the study is to select an appropriate equivalence scale formula. The scales are defined assuming Barten scaling of prices and, alternatively, independence of base (IB). Three functional forms of equivalence scales are examined. The results prove strong sensitivity to the choice of formula. Moreover, the elasticities of scales with respect to some household attributes take unacceptable values if IB is imposed and the scales are defined by means of Cobb-Douglas function.

The part **Book Review** contains a review of the book: Sarjinder SINGH: *Advanced Sampling Theory with Applications***,** Kluwer Academic Publishers, 2003, 1194 pp. (prepared by J. Kordos).

There are also **two Reports:** (i) First Demographic Congress in Poland*:* "*Poland vs. Europe Demographic Processes at the Beginning of the 21st Century*" (prepared by E. Fratczak and Z. Strzelecki); (ii) *Conference on Survey Sampling in Economical and Social Surveys*, Łódź, Poland, November 3—5, 2003 (prepared by J. Wywiał).

Jan Kordos

The Editor

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 491—501

PRACTICAL METHODS FOR SAMPLING RARE AND MOBILE POPULATIONS

Graham Kalton[1]

ABSTRACT

Many surveys focus on a rare population that constitutes only a small proportion of the total population. When a separate sampling frame is not available for the rare population, special cost-efficient sampling methods are needed. This paper reviews the use of disproportionate stratification, screening in the context of area sampling, and location sampling for sampling rare populations. Location sampling is also used for sampling mobile populations (such as international travelers), and its use for this type of population is discussed.

Key words: Disproportionate Stratification, Screening with Area Sampling, Location Sampling

1. Introduction

Seymour Sudman, an internationally renowned survey methodologist who wrote widely on many aspects of survey methodology, died on May 2, 2000. This paper, which was presented in the Seymour Sudman Memorial Session at the Annual Meeting of the American Statistical Association in 2001, reviews an area of survey sampling in which Seymour made many important contributions. Seymour's work in survey sampling was firmly grounded in his practical experience. This experience is clearly to be seen in his papers on the frequently encountered and challenging problems involved in devising efficient and valid methods for sampling rare and mobile populations.

Many surveys focus on a subset of the total population, and that subset is often a small proportion of the total. Thus, for example, surveys may be concerned with minority populations, specific age/sex groups such as males aged

[1] Graham Kalton, Westat, 1650 Research Boulevard, Rockville, Maryland 20850.
E-mail: grahamkalton@westat.com. The paper appears in the 2001 Proceedings of the Section on Survey Research Methods, American Statistical Association (CD-ROM).

18 to 24, the disabled population, or persons with rare diseases. Sometimes a separate sampling frame with good coverage of the rare population is available, in which case the rare population can be sampled using standard methods. For example, a sample of births in 2001 was drawn from birth certificate records for the National Center for Education Statistics' Early Childhood Longitudinal Study — Birth Cohort. However, in most cases such frames are unavailable, and special sampling techniques are required.

A related topic is sampling mobile populations, such as international travelers, car passengers, visitors to museums or national parks, the homeless, voters at polling booths, hospital outpatients, and shoppers at a shopping mall. Sometimes methods for sampling mobile populations are used for sampling rare populations as discussed later.

Two main objectives in surveys of rare and mobile populations can be distinguished. One is simply to estimate the number of members of the rare population M and the prevalence of the rare population in the total population $P = M/N$, where N is the size of the total population. The second objective is to estimate characteristics of the rare or mobile population, such as the mean for some variable y, $\bar{Y} = \sum Y_i / M$. The proportion of the rare or mobile population with a given characteristic can also be represented by $\bar{Y}$, with $Y_i = 1$ if individual i has the characteristic and $Y_i = 0$ if not. This review is restricted to the second of these objectives.

A wide variety of methods has been used for sampling rare and mobile populations, including:

- Special lists;
- Multiple frames;
- Screening;
- Disproportionate stratification;
- Multiplicity sampling;
- Snowballing;
- Adaptive sampling;
- Multipurpose surveys;
- Location sampling;
- Cumulating cases over several surveys; and
- Sequential sampling.

Methods for sampling rare populations have been reviewed by Sudman and Kalton (1986), Sudman, Sirken and Cowan (1988), Kish (1965a, 1991), Kalton and Anderson (1986), and Kalton (1993a). The use of adaptive sampling for estimating the size of a rare population is not covered in these papers, being of more recent origin. It is described in Thompson and Seber (1996). Kalton (1991) reviews methods for sampling mobile populations. This paper cannot attempt a complete review of the methods. Instead, it will be confined to three widely used

techniques: disproportionate stratification, screening in the context of area sampling, and location sampling.

2. Disproportionate Stratification

Some rare populations are more heavily concentrated in certain parts of the population. When this concentration occurs, it can be advantageous to sample the parts with the heavier concentrations at high rates. Thus the various parts of the population are treated as strata, with higher sampling fractions being used in the strata with the greater concentrations. This procedure can be cost efficient since less screening is needed in the strata with higher concentrations to identify members of the rare population.

Assume, for simplicity, the following:

- The population mean of variable y in stratum h ($\bar{Y}_h$) is the same for all strata, i.e. $\bar{Y}_h = \bar{Y}$ for all h.
- The element variance of y in stratum h, S_h^2, is the same for all strata, so that $S_h^2 = S^2$.
- The cost of screening out a member of the non-rare population is the same for all strata, $c_{Sh} = c_S$.
- The cost of collecting data for a member of the rare population is the same for all strata, $c_{Rh} = c_R$.

Then, with simple random sampling within the strata, the optimum sampling fraction in stratum h is

$$f_h \propto \sqrt{\frac{P_h}{P_h(r-1)+1}} \tag{1}$$

where $P_h = M_h / N_h$ is the prevalence of the rare population in stratum h and $r = c_R/c_S$ (Kalton, 1993a). In most cases, r is greater than 1, often appreciably so. However, when the screening costs dominate, r may be approximately 1. With $r = 1$, the optimum sampling fraction reduces to $f_h \propto \sqrt{P_h}$. Consider, for example, two strata, with 64 percent of the members in one stratum and 4 percent in the other being members of the rare population. Then, with $r = 1$, the first stratum should be sampled at a rate 4 times as large as the second stratum. With $r = 4$, the first stratum should be sampled at a rate only 2.48 times larger than the second stratum.

The gain in precision from the use of disproportionate sampling with optimum sampling fractions over the use of proportionate stratification for the case when $r = 1$ is approximately

$$R = \frac{V(\bar{y}_{opt})}{V(\bar{y}_{prop})} = (\Sigma \sqrt{A_h W_h})^2 \qquad (2)$$

where A_h and W_h are the proportions of the rare and of the total population in stratum h. This formula clearly illustrates the need for the distributions of the rare population and of the total population to differ across the strata if a reduction in variance is to be obtained from a disproportionate allocation.

In the case of two strata, the above formula may be re-expressed in terms of the A_1 and P_1/P, where P_1 is the prevalence of the rare population in stratum 1 and P is its prevalence in the total population. Table 1 presents the value of R (in percent) for different values of A_1 (in percent) and the relative prevalence $RP = P_1/P$. The table shows that the gains from disproportionate stratification are modest (i.e., R is not much less than 1) unless two conditions both apply:

- The prevalence of the rare population in stratum 1 must be much higher than in the total population, i.e., RP must be much greater than 1.
- The proportion of the rare population in stratum 1 must be high, i.e., A_1 must be large.

A value of R of 80 percent is equivalent to a 25 percent variance reduction from the use of the optimum sampling fractions. The stepped line in the table divides the cell values into those with $R > 80$ percent from those with $R < 80$ percent. As can be seen from the table, even when the prevalence of the rare population in stratum 1 is 20 times higher than the average prevalence, a 25 percent or greater reduction in variance is not achieved unless stratum 1 contains more than 30 percent of the rare population. On the other hand, a 25 percent or greater reduction in variance can be achieved if the prevalence of the rare population in stratum 1 is only twice as great as the average prevalence provided that stratum 1 contains 90 percent or more of the rare population.

The reductions in variance that accrue when r is greater than 1 are less than when $r = 1$, in line with the lesser variation in the optimum sampling fractions that occurs in this case, as noted above. Table 2 presents comparable results to those in Table 1, but with $r = 7$. In this case the results depend on the overall prevalence level, which is here taken to be $P = 10$ percent. A comparison of the cell values shows that the R values for given RP and A_1 are appreciably larger in Table 2, and the stepped $R = 80$ line has shifted much further down in the right hand corner of the table. Reductions of variance of 25 percent or more ($R < 0.8$) occur only for values of A_1 of 80 percent or more and then only for sizeable values of RP.

These results show that disproportionate stratification can be useful but that major benefits from the use of this technique arise only when A_1 and RP are large. The benefits decline as the relative cost of data collection from a member of

the rare population to the cost of screening out a member of the rare population (r) increases.

Table 1. Values of R for two strata with r=1

RP	A_1 (Percent)									
	10	20	30	40	50	60	70	80	90	100
1	100	100	100	100	100	100	100	100	100	100
1.5	100	99	99	98	97	96	94	92	87	67
2	99	98	97	95	93	91	88	83	76	50
3	98	96	93	91	87	83	78	71	61	33
5	97	93	89	85	80	74	67	59	47	20
10	95	90	84	78	72	64	56	47	34	10
15	94	88	82	75	68	60	51	41	29	7
20	94	87	81	73	66	57	48	38	26	5

Table 2. Values of R for two strata with r=7, P=10%

RP	A_1 (Percent)									
	10	20	30	40	50	60	70	80	90	100
1	100	100	100	100	100	100	100	100	100	100
1.5	100	100	100	99	99	98	98	97	94	79
2	100	99	99	98	97	96	95	93	89	69
3	99	99	98	97	95	94	91	88	82	58
5	99	98	97	95	93	91	88	83	76	50
10	99	97	96	94	91	88	84	79	71	44
15	99	97	95	93	91	87	83	78	69	42
20	99	97	95	93	90	87	83	77	68	41

3. Screening with Area Sampling

Most national household surveys employ area sampling, and this is also generally so for national surveys of rare populations. Three situations may be usefully distinguished:

1. The rare population is evenly spread throughout the population.
2. The rare population is unevenly spread, with higher concentrations in some areas.
3. The rare population is unevenly spread, with many areas containing no members of the rare population.

These three situations are described in turn below. The section then concludes with a brief discussion of the noncoverage problem that often arises with screening.

3.1. Evenly-Spread Rare Populations

Consider first the estimation of a sample mean $\bar{y}$ from a simple two-stage sample with a equal-sized primary sampling units (PSUs) selected by simple random sampling and b individuals selected by simple random sampling within selected PSUs. Further assume a simple cost model of the form $C = aC_a + abc$, where C_a is the cost of including a PSU in the sample and c is the survey cost per selected individual. Then, from standard theory, the optimum value for b is

$$b_T = \sqrt{\frac{C_a}{c}\frac{1-\rho}{\rho}} \tag{3}$$

where ρ is the intra-class correlation of the *y*-values in the PSUs, and the subscript T denotes the fact that this result relates to an estimate of the mean for the total population (see, for instance, Kish, 1965b).

The above formula for the optimum b can also be applied for an evenly spread rare population (e.g., children aged 18-36 months in an immunization survey), but with the value of c changed to $c' = c_R + c_S(P^{-1} - 1)$, with c_R and c_S as defined earlier. This change applies because the cost of collecting data for a member of the rare population also includes the cost of screening out $(P^{-1} - 1)$ members of the non-rare population. Thus if the cost of survey data collection is the same for a member of the total population as for a member of the rare population excluding the screening cost (i.e., $c = c_R$), then the optimum value of b_R for the rare population will be smaller than b_T. In this situation, the relationship between b_R and b_T depends on the cost ratio c_R / c_S. If $c_S = 0$, $b_R = b_T$. If $c_R = c_S$, $c' = P^{-1}c_R$, and hence $b_R = \sqrt{P}b_T$. Thus

$$\sqrt{P}b_T < b_R < b_T \tag{4}$$

It should be noted that, although b_R will generally be less than b_T, the optimum screening sample size per PSU will still be large. For example, with $P = 10$ percent and $b_T = 20$, $6 < b_R < 20$ from the above equation. However, the PSU screening sample size needed to generate such a sample size for the rare population is $63 < n_T < 200$, where n_T is the screening sample size per PSU. Thus, the above discussion does not contradict the well-established advice to select large subsamples from selected clusters when sampling a rare population. However, the optimum subsample size is not as large as would occur by simply equating b_T and b_R.

3.2. Unevenly Spread Rare Populations

Some rare populations are more heavily concentrated in certain areas. For selecting a sample of such a population, disproportionate stratification may be employed, with the strata being defined geographically. As discussed in Section 2, such disproportionate stratification gives notable gains in precision only when the prevalence of the rare population is much higher than average in some geographical strata *and* when these strata contain a substantial proportion of the rare population. Also, as noted in Section 2, the relative cost of a full interview to a screening interview (r) affects the effectiveness of disproportionate stratification. The higher the value of r, other things being equal, the less the gain in precision.

Waksberg, Judkins and Massey (1997) provide an extensive and informative evaluation of the effectiveness of disproportionate geographic stratification for sampling racial minorities and the low income population, where the geographical areas were 1980 and 1990 Census blocks and block groups. Their findings indicate that disproportionate stratification is useful for sampling blacks and Hispanics for $r < 5$ or 10, and for other minorities for even larger values of r. However, the gains from disproportionate stratification for sampling the low income population are small because, although there are areas with high concentrations of the low income population, a high proportion of the low income population lives elsewhere. Waksberg *et al.* also point out that an assessment of the effects of disproportionate geographic stratification based on Census data needs to take into account the changes that will have occurred in the geographic distribution of the rare population between the time of the Census and the time of the survey.

3.3. Many Clusters Containing No Members of the Rare Population

There are some rare populations that go unrepresented in many geographic clusters. If the zero clusters can be identified in advance of the survey, they can simply be removed from the sampling frame. However, when they cannot be identified in advance, under standard designs the zero clusters are sampled and extensive, but unproductive, screening is conducted within them. Sudman (1972, 1985) has proposed a scheme to avoid this outcome. His scheme is based on the Mitofsky-Waksberg random digit dialing scheme for telephone surveys (Waksberg, 1978). The scheme involves the initial selection of one (or a few) elements in each sampled area. If the selected element is a member of the rare population, further screening is carried out until b more elements of the rare population are sampled. If the initially selected element is not a member of the rare population, no more screening is conducted in that area. There are, of course, fieldwork issues to be considered in applying this scheme. However, in some circumstances the scheme can be effective for very rare populations and when there are many zero clusters.

3.4. Noncoverage with Screening

Screening involves collecting data from the members of the initial sample in order to be able to classify them as members or nonmembers of the rare population. The identification of some rare populations requires only one or a few questions (e.g., children aged 18—36 months), but for other rare populations many questions may be needed (e.g., low-income white families with a male head under 25 and 2 or more children). Misclassification errors at the screener stage can give rise to serious levels of noncoverage (Sudman, 1972, 1976). Misclassifications of nonmembers as members (false positives) are usually corrected in the detailed interview that follows the screener, but misclassifications of members as nonmembers (false negatives) are not corrected and thus result in noncoverage. The risk of false negatives is heightened when the screener respondents can deduce the rare population of interest from the contents of the screener questions or from advanced material supplied to them, since they can then avoid the full interview through their choice of responses to the screener. Thus, designers of screening questionnaires generally attempt to avoid a transparent disclosure of the rare population. Even so, substantial levels of noncoverage of rare populations are widely encountered with screening, and particularly so when many questions need to be used to identify the rare population, and an incorrect answer to any one of them leads to a false negative outcome. As a typical example, Horrigan *et al.* (1999) report that only 75 percent of youths aged 12—23 years of age were located in the National Longitudinal Survey of Youth of 1997 (NSLY97); much of the loss was probably due to noncoverage.

Another issue is nonresponse at the screening phase. There is often a concern that screener nonresponse will be higher for the rare population than the total population. Thus, even a high overall screener response rate may mask a low response rate for members of the rare population.

4. Location Sampling

Location sampling refers to methods used to sample individuals who visit specific locations such as libraries, museums, shopping centers, and polling places. Sampling is usually conducted either as the visitors enter or as they leave a location. Two distinct units of analysis need to be distinguished—visits and visitors (Kalton, 1991). Location sampling can readily produce a probability sample of visits, with known selection probabilities, and hence visits are easily analyzed. Visits may be the appropriate unit of analysis for, say, a survey about satisfaction with visits to a museum. However, for many surveys using location sampling, the visitor is the appropriate unit of analysis. For example, the visitor is the appropriate unit of analysis in a survey of visitors to soup kitchens over a week to estimate the number of homeless, a survey of nomads visiting watering

holes to estimate the size of the nomadic population, or a survey of men who have sex with men (MSM) visiting gay bars to study the characteristics of the MSM population.

The use of the visitor as the unit of analysis is complicated by the fact that a visitor may make multiple visits during the survey's time frame. If a standard sample of visits is selected, the increased selection probabilities associated with multiple visits need to be taken into account in developing the survey weights. The problem lies in estimating the multiplicities, both because a sampled person may be unable to accurately recall past visits since the start of the survey's reference period and because he or she is unable to forecast visits to be made from the time of interview until the end of the reference period. As a result, the multiplicities may be based on simple reports about general frequency of visits.

An alternative solution to the multiplicity problem is to uniquely identify one of the visits with the visitor, treating the other visits as blanks, thereby avoiding the problem. The natural choice for the uniquely identified visit is the first one in the survey reference period: each sampled person is asked if the visit is his or her first since the start of the survey, is selected if the answer is "Yes", and is rejected if the answer is "No". From the fieldwork perspective, an unattractive feature of this procedure is that most visits near the start of the time period will be first visits, leading to interviews, whereas most near the end will not. To some extent, this problem can be addressed by sampling the time periods with probabilities proportional to appropriate size measures, but determining these measures is problematic.

The usual sample design for a location sample is a two-stage design (Kalton, 1991). Primary sampling units are constructed as combinations of locations (entrances or exits) and time segments when the location is open (e.g., a given Monday from 10 a.m. to 2 p.m.). The PSUs are sampled with probabilities proportional to size, with careful stratification by location and time. Then some form of systematic sample is employed to select visitors entering (or exiting) the location. Sudman (1980) outlines the application of this type of design for sampling visitors to a shopping center with several entrances and using half-hour time segments when the center is open.

Location sampling has been widely used for surveys of MSM concerning HIV risk and illness (Kalton, 1993b) with locations such as bars, dance clubs and street locations where MSM congregate. The Young Men's Survey conducted in 7 cities in 1994—1998 in 194 public locations is a major survey of this type (Valleroy *et al.*, 2000). It had a sample size of 3,492 MSM aged 15 to 22 years of age who consented to an interview and HIV testing. A complex weighting scheme was devised to address the multiplicity problem (MacKellar *et al.*, 1996).

5. Concluding Remarks

The sampling of rare and mobile populations often presents survey statisticians with major challenges. Although many methods have been devised for sampling these populations (only a few of which have been discussed here), finding cost-effective methods is frequently difficult and requires ingenuity. In a number of cases, a compromise needs to be made between scientific rigor and practicability. When this occurs, a careful assessment of likely biases and good judgment are required. Seymour Sudman's many valuable contributions to the subject exhibit this combination of ingenuity with a thoughtful balance of practicability and scientific rigor.

REFERENCES

HORRIGAN, M., MOORE, W., PEDLOW, S., AND WOLTER, K. (1999). Undercoverage in a large screening survey of youths. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 570—575.

KALTON, G. (1991). Sampling flows of mobile human populations. *Survey Methodology*, 17, 183—194.

KALTON, G. (1993a). *Sampling* Rare *and Elusive Populations*. New York: Department of Economic and Social Information and Policy Analysis Statistics Division, United Nations.

KALTON, G. (1993b). Sampling considerations in research on HIV risk and illness. In *Methodological Issues in AIDS Behavioral Research*, D.G. Ostrow and R.C. Kessler eds., pp. 53—74. New York: Plenum Press.

KALTON, G. AND ANDERSON, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, A*, 149, 65—82.

KISH, L. (1965a). Selection techniques for rare traits. *Genetics and the Epidemiology of Chronic Diseases*, Public Health Service Publication No. 1163.

KISH, L. (1965b). *Survey Sampling*. New York: Wiley.

KISH, L. (1991). Taxonomy of elusive populations. *Journal of Official Statistics*, 7, 339—347.

MACKELLAR, D., VALLEROY L., KARON, J., LEMP, G. AND JANSSEN, R. (1996). The Young Men's Survey: methods for estimating HIV seroprevalence and risk factors among young men who have sex with men. *Public Health Reports*, 111(Suppl. 1), 138—144.

SUDMAN, S. (1972). On sampling of very rare human populations. *Journal of the American Statistical Association,* 67, 335—339.

SUDMAN, S. (1976). *Applied Sampling.* New York: Academic Press.

SUDMAN, S. (1980). Improving the quality of shopping center sampling. *Journal of Marketing Research,*17, 423—431.

SUDMAN, S. (1985). Efficient screening methods for the sampling of geographically clustered special populations. *Journal of Marketing Research*, 22, 20—29.

SUDMAN, S. AND KALTON, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, 12, 401—429.

SUDMAN, S., SIRKEN, M.G. AND COWAN, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991—996.

THOMPSON, S.K. AND SEBER, G.A.F. (1996). *Adaptive Sampling*. New York: Wiley.

VALLEROY, L.A., MACKELLAR, D., KARON, J. *et al.* (2000). HIV prevalence and associated risks in young men who have sex with men. *Journal of the American Medical Association*, 284, 198—204.

WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40—46.

WAKSBERG, J., JUDKINS, D. AND MASSEY, J.T. (1997). Geographic-based oversampling in demographic surveys of the United States. *Survey Methodology*, 23, 61—71.

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 503—513

ESTIMATION OF THE POPULATION TOTAL ON THE CURRENT OCCASSION UNDER SECOND STAGE UNIT ROTATION PATTERN

Barbara Kowalczyk[1]

ABSTRACT

It is well known that it is desirable to use the information from the previous occasion to increase efficiency of the estimation on the current occasion. The present paper gives the theory for the estimation of the population total on the current occasion under two-stage sampling design with a partial replacement of the second stage units. Under given rotation pattern we present composite estimator of the population total, which exploits information from the previous occasion. We derive the variance of the estimator and compare it with the variance of the usual estimator. It has been shown that the composite estimator is better than the usual one applied both for a sample selected independently of the previous period and a sample selected according to the given rotation pattern.

***Key words*:** Sampling on Successive Occasions, Composite Estimation, Rotation Pattern, Partial Replacement of Units, Two-Stage Sampling Design.

1. Introduction

The problem of sampling on two successive occasions with a partial replacement of units was first considered by Jessen (1942). The main results for the mean estimation on more than two occasions are due to Patterson (1950). Patterson's theory was examined and extended by many authors e.g. Yates (1949), Eckler (1955), Rao and Graham (1964), Kordos (1967), Cochran (1977). Singh (1968) was the first to examine the theory of the population mean estimation on successive occasions for two-stage sampling design. He considered simple random sampling without replacement (SRSWOR) in both stages and assumed first stage unit rotation.

First stage unit rotation pattern looks as follows. On the first occasion a sample of m PSUs is drawn by the simple random method without replacement.

[1] Warsaw School of Economics, Warsaw, Poland, e-mail: bkowal@sgh.waw.pl

For the second occasion a subsample of mp $(0 \le p \le 1)$ PSUs is retained **along with all their SSUs** selected on the first occasion. The sample for the second occasion is supplemented by a sample of mq $(p+q=1)$ PSUs selected from the units not used on the first occasion.

The same sampling design but for the population ratio estimation on the second occasion was examined by Tripathi and Sinha (1976). Another two-stage sampling method was considered by Okafor (1987) and Okafor and Arnab (1987). The authors examined probability proportional to size with replacement PPSWR in the first stage and SRSWOR in the second stage and also assumed first stage unit rotation.

In the present paper, contrary to the previous authors, we examine the second stage unit rotation. This kind of rotation pattern seems to be very typical in applications. In many practical situations surveys conducted under two-stage sampling design are repeated in time in such a way that some second stage units that are examined on the current occasion have been also examined on the previous occasion. So it is desirable to use the information from the preceding occasion to improve the estimator on the current occasion [See also Duncan and Kalton (1987), Binder and Hidiroglou (1988)].

The present paper gives the theory for the population total estimation on the current occasion under two-stage sampling design with SRSWOR in both stages and rotation of units in the second stage.

2. Sampling design and rotation pattern

First consider a problem when the same population of N elements is sampled on two successive occasions using two-stage sampling design. Assume that the population does not change in composition over time (the same assumption was accepted in all theoretical papers mentioned in section 1, except for Duncan and Kalton (1987) which relates more to the practice of survey sampling). On the first occasion we apply usual two-stage sampling design with SRSWOR in both stages.

The corresponding notation is as follows. The subscripts used below vary respectively: k from 1 to M, j from 1 to N_k..

M — number of primary sampling units (PSUs) in the population,

N_k — number of second stage sampling units (SSUs) in the k-th PSU,

$$N = \sum_{k=1}^{M} N_k \ ,$$

m — number of PSUs in the sample,

n_k — number of SSUs to be selected from the k-th PSU,

q_k — fraction of the second stage sample from the k-th PSU to be rotated on the next occasion, $p_k = 1 - q_k$,

$Y_{(t)kj}$ — value of the variable Y under study on the *j*-th SSU in the *k*-th PSU on the *t*-th occasion, (2.1)

$\overline{Y}_{(t)k} = \frac{1}{N_k}\sum_{j=1}^{N_k} Y_{(t)kj}$ — population mean per SSU for the *k*-th PSU on the *t*-th occasion, 2.2)

$S_k^2(Y_{(t)}) = \frac{1}{N_k - 1}\sum_{j=1}^{N_k}(Y_{(t)kj} - \overline{Y}_{(t)k})^2$ — population variance per SSU for the *k*-th PSU on the *t*-th occasion, (2.3)

$$C_k(Y_{(t)}, Y_{(s)}) = \frac{1}{N_k - 1}\sum_{j=1}^{N_k}(Y_{(t)kj} - \overline{Y}_{(t)k})(Y_{(s)kj} - \overline{Y}_{(s)k}),$$

$$\rho_k(Y_{(t)}, Y_{(s)}) = \frac{C_k(Y_{(t)}, Y_{(s)})}{S_k(Y_{(t)})S_k(Y_{(s)})} \tag{2.4}$$

$Y_{(t)k} = N_k\overline{Y}_{(t)k}$, $Y_{(t)} = \sum_{k=1}^{M} Y_{(t)k}$ — total value on the *t*-th occasion for the *k*-th PSU and for the whole population respectively, (2.5)

$S_P^2(Y_{(t)}) = \frac{1}{M - 1}\sum_{k=1}^{M}\left(Y_{(t)k} - \frac{1}{M}Y_{(t)}\right)^2$ — population variance per PSU on the *t*-th occasion. (2.6)

On the first occasion we draw *m* PSUs according to SRSWOR. From the *g*-th, $g = 1,...,m$ selected PSU we draw the appropriate number of SSUs, i.e. n_g units.

For the second occasion we remain all PSUs selected on the first occasion. For each PSU in the sample we draw on the second occasion $n_g p_g$ SSUs according to SRSWOR from n_g units that have been examined on the first occasion. The remaining $n_g q_g$ SSUs $(\boldsymbol{p_g+q_g=1})$ we draw according to SRSWOR from the N_g-n_g SSUs that have not been examined on the first occasion.

The sample on both occasions consists of the same PSUs. The second stage sample for each PSU on the second occasion consists of two parts: units that have been also observed on the first occasion and new units that have not been observed on the first occasion.

The rotation pattern of SSUs can be illustrated as follows:

1-st occasion	xxxxxxx	xxxxx	
2-nd occasion		xxxxx	xxxxxxx
	$n_g q_g$ units	$n_g p_g$ units	$n_g q_g$ units

We use the following notation. The subscript g vary from 1 to m, t = 1,2.

$y_{(t)gi}$ − value of the variable Y on the SSU selected at the i-th second stage draw from the primary stage unit drawn at the g-th first stage draw on the t-th occasion (2.7)

$\bar{y}_{(t)g} = \frac{1}{n_g}\sum_{i=1}^{n_g} y_{(t)gi}$ — sample mean on the t-th occasion per SSU for the PSU drawn at the g-th first stage draw, (2.8)

We have

$$\bar{y}_{(t)g} = p_g \bar{y}_{(t)Mg} + q_g \bar{y}_{(t)Ug}, \tag{2.9}$$

where

$\bar{y}_{(t)Mg}$ — sample mean associated with the **matched** portion, i.e. with the $n_g p_g$ units observed on both occasion, (2.10)

$\bar{y}_{(t)Ug}$ — sample mean associated with the **unmatched** portion, i.e. with the $n_g q_g$ units observed only on one occasion. (2.11)

3. Proposed estimator

3.1. Theorem

Under introduced rotation pattern a statistic given by formula

$$y_{(2)} = \frac{M}{m}\sum_{g=1}^{m} N_g \left(Q_g \bar{y}_{(2)Ug} + (1-Q_g)\bar{y}^*_{(2)Mg}\right), \tag{3.1}$$

where

$$\bar{y}^*_{(2)Mg} = \bar{y}_{(2)Mg} + b_g(\bar{y}_{(1)g} - \bar{y}_{(1)Mg}), \tag{3.2}$$

$$Q_g = \frac{q_g[1-q_g\rho_g^2(Y_{(2)},Y_{(1)})]}{1-q_g^2\rho_g^2(Y_{(2)},Y_{(1)})}, \tag{3.3}$$

$$b_g = \frac{C_g(Y_{(2)},Y_{(1)})}{S_g^2(Y_{(1)})} \tag{3.4}$$

is an unbiased estimator of the population total on the second occasion $Y_{(2)}$. The variance of the estimator is given by:

$$D^2(y_{(2)}) = M^2\left(\frac{1}{m}-\frac{1}{M}\right)S_P^2(Y_{(2)}) +$$

$$\frac{M}{m}\sum_{k=1}^{M} N_k^2\left(\frac{1}{n_k}\frac{1-q_k\rho_k^2(Y_{(2)},Y_{(1)})}{1-q_k^2\rho_k^2(Y_{(2)},Y_{(1)})}-\frac{1}{N_k}\right)S_k^2(Y_{(2)}). \tag{3.5}$$

Proof:

First we will show that the estimator $y_{(2)}$ given by (3.1) is an unbiased estimator of the population total $Y_{(2)}$. From the general statistical theory it is well known that the expected value of a random variable can be expressed as the expected value of conditional expectations. So we have:

$$E(y_{(2)}) = E_1\left[E_2(y_{(2)} \mid I)\right] = \frac{M}{m} E_1\left[\sum_{g=1}^{m} N_g E_2(e_{(2)g} \mid I)\right],$$

where

$$e_{(2)g} = Q_g \bar{y}_{(2)Ug} + (1 - Q_g)\bar{y}^*_{(2)Mg}.$$

Symbol E_1 denotes the expected value over first stage sampling results, while $E_2(\cdot \mid I)$ denotes the expected value over second stage sampling results under condition that the first stage sample is realized. We have respectively:

$$E_2\left(Q_g \bar{y}_{(2)Ug} + (1 - Q_g)\bar{y}^*_{(2)Mg} \mid I\right) = Q_g \bar{Y}_{(2)g} + (1 - Q_g)\bar{Y}_{(2)g} = \bar{Y}_{(2)g}$$

and

$$E(y_{(2)}) = \frac{M}{m} E_1\left[\sum_{g=1}^{m} N_g \bar{Y}_{(2)g}\right] = \frac{M}{m}\left[\sum_{g=1}^{m} E_1(Y_{(2)g})\right] =$$

$$\frac{M}{m}\left[\sum_{g=1}^{m} \frac{1}{M} \sum_{k=1}^{M} Y_{(2)k}\right] = \sum_{k=1}^{M} Y_{(2)k} = Y_{(2)}.$$

Now we derive the formula for the variance of the estimator $y_{(2)}$. From the general statistical theory we have:

$$D^2(y_{(2)}) = D_1^2\left[E_2(y_{(2)} \mid I)\right] + E_1\left[D_2^2(y_{(2)} \mid I)\right].$$

The first component of the variance is of the form:

$$D_1^2\left[E_2(y_{(2)} \mid I)\right] = D_1^2\left[\frac{M}{m}\sum_{g=1}^{m} N_g E_2(e_{(2)g} \mid I)\right] = D_1^2\left[\frac{M}{m}\sum_{g=1}^{m} N_g \bar{Y}_{(2)g}\right] =$$

$$= M^2 D_1^2\left[\frac{1}{m}\sum_{g=1}^{m} Y_{(2)g}\right] = M^2\left(\frac{1}{m} - \frac{1}{M}\right) S_P^2(Y_{(2)}),$$

where $S_P^2(Y_{(t)})$ is defined in (3.6).

Now we compute the second component of the variance.

$$E_1\left[D_2^2(y_{(2)} \mid I)\right] = E_1\left[\frac{M^2}{m^2}\sum_{g=1}^{m} N_g^2 D_2^2(e_{(2)g} \mid I)\right].$$

Applying the theory that refers to simple random sampling we have:

$$D_2^2(y_{(2)} \mid I) = \left(\frac{1}{n_g} \frac{1-q_g \rho_g^2(Y_{(2)}, Y_{(1)})}{1-q_g^2 \rho_g^2(Y_{(2)}, Y_{(1)})} - \frac{1}{N_g} \right) S_g^2(Y_{(2)}).$$

[The appropriate formulas (but ignoring the finite population correction) can be found e.g. in Cochran (1977), pp. 346-347. Taking into consideration the fpc, the formulas can be found in Kowalczyk (2002), pp. 52-55].

Next we have:

$$E_1\left[D_2^2(y_{(2)} \mid I)\right] = \frac{M^2}{m^2} E_1\left[\sum_{g=1}^{m} N_g^2 \left(\frac{1}{n_g} \frac{1-q_g \rho_g^2(Y_{(2)}, Y_{(1)})}{1-q_g^2 \rho_g^2(Y_{(2)}, Y_{(1)})} - \frac{1}{N_g} \right) S_g^2(Y_{(2)}) \right] =$$

$$= \frac{M^2}{m^2} \left[\sum_{g=1}^{m} \frac{1}{M} \sum_{k=1}^{M} N_k^2 \left(\frac{1}{n_k} \frac{1-q_k \rho_k^2(Y_{(2)}, Y_{(1)})}{1-q_k^2 \rho_k^2(Y_{(2)}, Y_{(1)})} - \frac{1}{N_k} \right) S_k^2(Y_{(2)}) \right] =$$

$$= \frac{M}{m} \sum_{k=1}^{M} N_k^2 \left(\frac{1}{n_k} \frac{1-q_k \rho_k^2(Y_{(2)}, Y_{(1)})}{1-q_k^2 \rho_k^2(Y_{(2)}, Y_{(1)})} - \frac{1}{N_k} \right) S_k^2(Y_{(2)}).$$

Adding the two components we have the desired formula (3.5).

4. Analytical comparison of the estimators

In this section we will compare the proposed composite estimator $y_{(2)}$ of the population total on the second occasion with the usual estimator. We will study two cases.

The first case. On the second occasion a new sample is selected according to the standard two stage sampling design (i.e. independently of the first period). It is well known that the usual estimator of the population total under two stage sampling design (with SRSWOR in both stages) is of the form:

$$y_{(2)U} = \frac{M}{m} \sum_{g=1}^{m} \frac{N_g}{n_g} \sum_{i=1}^{n_g} y_{(2)gi} = \frac{M}{m} \sum_{g=1}^{m} N_g \bar{y}_{(2)g}. \tag{4.1}$$

The variance of this estimator is given by:

$$D^2(y_{(2)U}) = M^2 \left(\frac{1}{m} - \frac{1}{M} \right) S_P^2(Y_{(2)}) + \frac{M}{m} \sum_{k=1}^{M} N_k^2 \left(\frac{1}{n_k} - \frac{1}{N_k} \right) S_k^2(Y_{(2)}). \tag{4.2}$$

[see e.g. Särndal, Swensson, Wretman (1992), p.142].

The second case. A sample for the second occasion is selected according to the rotation pattern described in section 2. But this time for the estimation of the population total $Y_{(2)}$ the usual estimator of the form (4.1) is used (i.e. estimator that does not exploits information from the previous occasion). In this case the estimator given in (4.1) remains unbiased and its variation is still of the form (4.2). This results from the basic properties of the SRSWOR. Drawing in the second stage n_g elements (without replacement) directly from N_g elements is the same (in probability meaning) as drawing first n_g elements from N_g, and then $n_g p_g$ elements from the chosen n_g, and then $n_g q_g$ elements from the remaining N_g-n_g elements. Probability that a particular element is chosen remains the same $\frac{1}{N_g}$.

4.1. Corollary

The proposed estimator $y_{(2)}$ of the population total for the second occasion is better than the usual estimator $y_{(2)U}$, i.e. inequality holds:

$$D^2(y_{(2)}) \le D^2(y_{(2)U}). \tag{4.3}$$

Proof:
Using traditional differential calculus we have

$$\frac{1-q_k\rho_k^2(Y_{(2)},Y_{(1)})}{1-q_k^2\rho_k^2(Y_{(2)},Y_{(1)})} \le 1 \quad \text{for all k = 1,2,..., M,}$$

hence

$$\frac{M}{m}\sum_{k=1}^{M} N_k^2\left(\frac{1}{n_k}\frac{1-q_k\rho_k^2(Y_{(2)},Y_{(1)})}{1-q_k^2\rho_k^2(Y_{(2)},Y_{(1)})}-\frac{1}{N_k}\right)S_k^2(Y_{(2)}) \le$$

$$\frac{M}{m}\sum_{k=1}^{M} N_k^2\left(\frac{1}{n_k}-\frac{1}{N_k}\right)S_k^2(Y_{(2)})$$

which gives the desired inequality (4.3).

Relative percentage gain in efficiency is given by formula:

$$\frac{D^2(\bar{y}_{(2)U}) - D^2(\bar{y}_{(2)})}{D^2(\bar{y}_{(2)U})} \cdot 100\% = \frac{\sum_{k=1}^{M} N_k^2 \left(\frac{1}{n_k} \frac{p_k q_k \rho_k^2(Y_{(2)}, Y_{(1)})}{1 - q_k^2 \rho_k^2(Y_{(2)}, Y_{(1)})} \right) S_k^2(Y_{(2)})}{(M-m) S_P^2(Y_{(2)}) + \sum_{k=1}^{M} N_k^2 \left(\frac{1}{n_k} - \frac{1}{N_k} \right) S_k^2(Y_{(2)})} \cdot 100\% \tag{4.4}$$

5. Extension for more than two occasions

The results from the previous sections can be easily extended into more than two occasions under assumption that the population does not change in composition over time.

Now consider a problem when the same population is sampled on h successive occasions, $h \geq 2$. The rotation pattern is as follows. For all occasions PSUs remain the same, selected on the first occasion. For each PSU in the sample the second stage sample on the t-th occasion, t=2,3,...,h consists of two parts. The first part contains respectively $n_g p_g$ units drawn according to SRSWOR from units that have been examined on the (t-1)-th occasion but have not been examined on periods prior to (t-1), i.e. t-2, t-3,..., . The second part consists of $n_g q_g$ units drawn according to SRSWOR from elements that have not been examined so far. The rotation pattern implies that: $0{,}5 \leq q_g \leq 1$ [in case h = 2 (only two occasions) the condition above is not necessary]. We can repeat the procedure until the SSUs in selected PSUs are used up.

To distinguish between SSUs that have been examined on the previous occasion and SSUs that have passed on to examination on next occasion let us introduce notation:

$\bar{y}_{(t)Mg}$ — sample mean associated with the $n_g p_g$ units observed also on the (t-1)-th occasion,

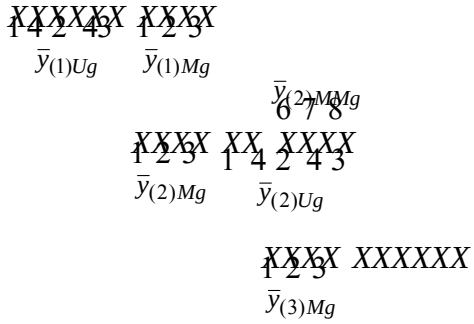
$\bar{y}_{(t)Ug}$ — sample mean associated with the $n_g q_g$ units observed for the first time on this occasion,

$\bar{y}_{(t)MMg}$ — sample mean associated with the $n_g p_g$ units observed also on the (t+1)-th occasion,

$\bar{y}_{(t)UUg}$ — sample mean associated with the $n_g q_g$ units not observed on the (t+1)-th occasion.

For the first occasion, to have the same notation as in previous sections, we denote: $\bar{y}_{(1)MMg} = \bar{y}_{(1)Mg}$, $\bar{y}_{(1)UUg} = \bar{y}_{(1)Ug}$.

As an example we have:



Under introduced rotation pattern a statistic given by formula

$$y_{(h)} = \frac{M}{m}\sum_{g=1}^{m} N_g e_{(h)g} \tag{5.1}$$

where

$$e_{(h)g} = Q_{(h)g}\bar{y}_{(h)Ug} + (1 - Q_{(h)g})\bar{y}^{*}_{(h)Mg}, \tag{5.2}$$

$$\bar{y}^{*}_{(h)Mg} = \bar{y}_{(h)Mg} + b_{(h)g}(e_{(h-1)g} - \bar{y}_{(h-1)MMg}), \tag{5.3}$$

$$b_{(h)g} = \frac{C_g(Y_{(h)}, Y_{(h-1)})}{S_g^2(Y_{(h-1)})}, \tag{5.4}$$

Q_h is given by recurrent formula:

$$1 - Q_{(h)g} = \frac{p_g}{1 - \rho_g^2(Y_{(h)}, Y_{(h-1)})(q_g - Q_{(h-1)g}\, p_g)}, \tag{5.5}$$

while

$$Q_{(1)g} = q_g, \quad e_{(1)g} = \bar{y}_{(1)g} \tag{5.6}$$

is an unbiased estimator of the population total on the current *h*-th occasion $Y_{(h)}$. The variance of the estimator is given by:

$$D^2(y_{(h)}) = M^2\left(\frac{1}{m} - \frac{1}{M}\right)S_P^2(Y_{(h)}) + \frac{M}{m}\sum_{k=1}^{M} N_k^2\left(\frac{Q_{(h)k}}{n_k q_k} - \frac{1}{N_k}\right)S_k^2(Y_{(h)}). \tag{5.7}$$

The proposed estimator is better than the usual $\bar{y}_{(h)U}$ estimator introduced in section 4, applied both for a sample selected independently of the previous period and a sample selected according to the given rotation pattern, i. e. inequality holds

$$D^2(y_{(h)}) \leq D^2(y_{(h)U}).$$

Proof of the above results is strictly analogous to the proof of theorem 3.1 and corollary 4.1. Results for simple random sampling are also used [see e.g. Patterson (1950), Cochran (1977), pp. 349—350 (ignoring the fpc and assuming additionally that the population variance is constant on successive occasions) and taking into consideration the fpc and without additional assumptions Kowalczyk (2002), pp. 42—53].

6. Concluding remarks

The procedure described in section 5, as mentioned above, can be repeated until SSUs in selected PSUs are used up. So it should be emphasized that in practice the given theory can be useful only for several occasions. Primary sampling units usually do not consist of a large number of second stage units so they can be relatively quickly exhausted.

In formulas for the composite estimator occur coefficients $\rho_g(Y_{(t)}, Y_{(s)})$, $b_{(t)g}$ which are assumed to be known (as in all theoretical papers mentioned in section 1). In practice however these quantities are usually unknown and ought to be estimated. Of course if the mentioned coefficients are replaced by their estimates then the received estimator will not be unbiased and its variance will not be of the form (5.7). However if the sample is large enough such proceeding should not change considerably stochastic properties of the estimator.

Also assumption that the population does not change in composition over time is very crucial. Sometimes (very seldom in fact) statisticians deal with populations that do not change in composition or (more often) changes in composition are insignificant. In some cases, even if the population does change over time, the changes are not included automatically in a sampling frame. Maintenance of a frame is cost and time expensive.

Changes in composition involve many serious problems in survey sampling practice. Theoretical results received under assumption that the population does not change in composition over time do not provide exact solutions for changing populations but indicate some procedure directions, which is also very important in practice.

All these problems connected with divergence between theory and practice are only mentioned here and need of course separate consideration.

REFERENCES

BINDER, D.A., HIDIROGLOU, M.A. (1988), Sampling in Time. P.R. Krishnaiah i R. Rao, eds., Handbook of Statistics, Vol. 6, PP. 187—211.

COCHRAN, W.G. (1977). Sampling Techniques. Third Edition. J. Wiley, New York-London.

DUNCAN, G.J. and KALTON, G. (1987). Issues of Design and Analysis of Surveys Across Time, Inernational Statistical Review 55.1, pp. 97—117.

ECKLER, A.R. (1955). Rotation Sampling. Annals of Mathematical Statistics 26, pp. 664—685.

JESSEN, R. (1942). Statistical investigation of a farm survey for obtaining farm facts. Iowa Agricultural Station Research Bulletin 304, pp. 54—59.

KORDOS, J. (1967). Metoda rotacyjna w badaniach reprezentacyjnych. (Rotation Metod in Sample Surveys) Przegląd Statystyczny 14.4, pp. 373—394.

KOWALCZYK, B. (2002). Sampling in Time (in Polish). Ph.D. Thesis, Warsaw School of Economics, Warsaw, Poland.

OKAFOR, F. and ARNAB, R. (1987). Some strategies of two-stage sampling for estimating population ratios over two occasions. Austral. J. Statist. 29, pp. 128—142.

OKAFOR, F.C. (1987). Comparison of Estimators of Population Total in Two-Stage Successive Sampling Using Auxiliary Information. Survey Methodology 13, pp. 109—121.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society, Series B 12, pp.241—255.

RAO, J. and GRAHAM, J. (1964). Rotation designs for sampling on repeated occasions. Journal of the American Statistical Association 50, pp.492—509.

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

SINGH, D. (1968). Estimates in successive sampling using multi-stage design. Journal of the American Statistical Association 63, pp.99—112.

TRIPATHI, T.P. i SINHA, S.K.P. (1976). Estimation of ratio on successive occasions. Paper presented at the „Symposium on Recent Developement in Survey Methodology" held at Indian Statistical Institute.

YATES, F. (1949). Sampling Methods for Censuses and Surveys. First Edition. Charles Griffin, London.

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 515—522

ON THE USE OF MODIFIED RANDOMIZATION DEVICE FOR ESTIMATING THE PREVALENCE OF A SENSITIVE ATTRIBUTE

Sarjinder Singh[1], Stephen Horn[2], Ravindra Singh[3], N.S. Mangat[4]

ABSTRACT

The present study considers the use of a modified randomization device in the usual unrelated question randomized response model (UY-model). The proposed model based estimator is unbiased and fares better as compared to usual UY-model estimator when the proportion of population possessing the sensitive attribute is less than 0.5.

Key words: Equal probabilities with replacement sampling; estimation of proportion; randomized response technique; unrelated question model.

1. Introduction

The unrelated question randomized response data-gathering device to procure trustworthy information on stigmatized characters was introduced by Horvitz *et al.* (1967). They claimed better cooperation from the respondents as compared to Warner's (1965) original model. Greenberg *et al*. (1969) while developing theoretical details for the Horvitz *et al.'s* model, considered both the situations where π_y, the proportion of innocuous character (say) Y is known and when it is not known. When π_y is known, we shall call this model as UY-model.

[1] Department of Statistics, St. Cloud State University, St. Cloud, MN 56301—4498,USA
E-mail: sarjinder@yahoo.com
[2] Dept. Family and Community Services, Box 7788, Canberra Mail Centre, ACT 2610, Australia
E-mail: stephen.horn@facs.gov.au
[3] Department of Mathematics and Statistics, Punjab Agricultural University, Ludhiana — 141 004, India.
[4] Morrison Scientific Inc., 815-706, 7th Ave. S.W. Calgary, AB T2P 0Z1, Canada.

In UY-model, each respondent in an equal probabilities with replacement sample of size n, is provided with a random device consisting of two statements: (i) "I belong to sensitive category A", and (ii) "I belong to category Y", represented with probabilities p_1 and $1-p_1$, respectively. The respondent selects one of these two statements randomly unobserved by the interviewer, and reports "yes" if he/she possesses the characteristic indicated by the selected statement and "no" otherwise. Assuming that the respondents speak complete truth, Greenberg *et al*. (1969) considered the following estimator of π (the proportion of population possessing sensitive attribute):

$$\hat{\pi} = \frac{\frac{n'}{n} - (1-p_1)\pi_y}{p_1}, \tag{1.1}$$

where π_y is the proportion of population possessing the character Y and n' is the number of observed "yes" answers. The estimator $\hat{\pi}$ is unbiased for π and has the variance

$$V(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{\pi(1-p_1)(1-2\pi_y)}{np_1} + \frac{\pi_y(1-p_1)\{1-(1-p_1)\pi_y\}}{np_1^2}. \tag{1.2}$$

The optimal p_1 and π_y are selected according to the following rule: The optimal $p_1>0$ is chosen close to 1. The value of π_y is chosen such that $|\pi_y - 0.5|$ is a maximum and π_y falls on the same side of 0.5 as π. If $\pi = 0.5$, the $|\pi_y - 0.5|$ could be maximum on either side. New developments in Warner (1965) randomized response model have been suggested by Chaudhuri and Mukherjee (1988), Ruiz-Espejo (1989), Mangat *et al*. (1993, 1995) and Tracy and Mangat (1995) else where.

In the present paper, we consider model which differs from the above discussed procedure in the sense that the randomization device used in the proposed procedure has three outcomes. In addition to the two outcomes in UY-model, we also have a third alternative which does not point to any character. In this case, the randomization device consisting of a deck of cards will carry two statements of the UY-model and some cards will be left blank and having no statement at all.

2. Proposed Model

As outlined in the preceding paragraph, the randomization device used in the proposed model carries three types of cards bearing statements: (i) "I belong to

sensitive category A", (ii) "I belong to category Y", and (iii) "Blank cards", with probabilities p_1, p_2 and p_3 such that $\sum_{i=1}^{3} p_i = 1$. In case the blank card is drawn by the respondent, he/she will report "no". The rest of the procedure remains as usual. The probability of "yes" answer will now be

$$\theta_1 = p_1\pi + p_2\pi_y . \tag{2.1}$$

Solving (2.1) for π, we get estimator of π as

$$\hat{\pi}_1 = \frac{\hat{\theta}_1 - p_2\pi_y}{p_1}, \tag{2.2}$$

where $\hat{\theta}_1$ is the proportion of "yes" answers obtained from the n sampled respondents.

Since the random variable $\hat{\theta}_1$ follows binomial distribution with parameters (n, θ_1) i.e. $\hat{\theta}_1 \cap B(n, \theta_1)$, we have the following theorem. The proof of the theorem being straightforward is omitted.

Theorem 2.1. The estimator $\hat{\pi}_1$ is unbiased for population proportion π.

In order to study the performance of the proposed estimator $\hat{\pi}_1$, we need its variance. This is obtained in Theorem 2.2.

Theorem 2.2. The variance of the estimator $\hat{\pi}_1$ is given by

$$V\left(\hat{\pi}_1\right) = \frac{\pi(1-\pi)}{n} + \frac{\pi\left(1 - p_1 - 2p_2\pi_y\right)}{np_1} + \frac{p_2\pi_y\left(1 - p_2\pi_y\right)}{np_1^2}. \tag{2.3}$$

Proof: Since $\hat{\theta}_1 \cap B(n, \theta_1)$, we have

$$V\left(\hat{\pi}_1\right) = \frac{\theta_1(1-\theta_1)}{np_1^2}. \tag{2.4}$$

On substituting the value of θ_1 from (2.1), the above expression for variance, after some algebraic simplifications, reduces to one given in (2.3). This completes the proof.

Obviously, (2.3) decreases as p_1 increases and will be minimum for p_1 close to 1. The optimal value of p_1 for two strategies under discussion will,

therefore, be the same. Using $p_1 > 0.5$, the variance $V\left(\hat{\pi}_1\right)$ values computed in Table 1 indicate that for $\pi \le 0.5$, one should choose π_y as close to 0 as it is possible. Thus for $\pi \le 0.5$, Greenberg *et al.* (1969) model and the proposed procedure have the same rules for choosing π_y. However for $\pi > 0.5$, the optimal choice of π_y is sometimes close to 0 and sometimes close to 1 depending on the choice of p_1, p_2 and $p_3 = 1 - p_1 - p_2$ and we are not able to arrive at a simple rule for choosing π_y for this case.

An unbiased estimator of variance $V\left(\hat{\pi}_1\right)$ can be obtained by replacing θ_1 by its estimate $\hat{\theta}_1$ and n by $n-1$ in (2.4). We thus have the following theorem.

Theorem 2.3. An unbiased estimator of variance $V\left(\hat{\pi}_1\right)$ is given by

$$v\left(\hat{\pi}_1\right) = \frac{\hat{\theta}_1\left(1-\hat{\theta}_1\right)}{(n-1)p_1^2}. \tag{2.5}$$

We now compare the efficiency of the proposed model based estimator $\hat{\pi}_1$ with the usual estimator $\hat{\pi}$ defined respectively in (1.1) and (2.2).

3. Efficiency Comparison

For the efficiency comparison, two cases i.e. when $(i)\ \pi \le 0.5$ and (ii) $\pi > 0.5$, are considered separately.

Case I. $\pi \le 0.5$: Here, optimal values for p_1 and π_y are the same for both models. Following this, we define relative efficiency of the proposed estimator $\hat{\pi}_1$ with respect to the usual UY-model based estimator $\hat{\pi}$ as

$$RE = V\left(\hat{\pi}\right) \Big/ V\left(\hat{\pi}_1\right).$$

The estimator $\hat{\pi}_1$ will be more efficient compared to $\hat{\pi}$ if

$$RE > 1.$$

Using (1.2) and (2.3), the above inequality becomes

$$\pi p_1(1-p_1)(1-2\pi_y)+\pi_y(1-p_1)\{1-(1-p_1)\pi_y\} > \pi p_1(1-p_1-2p_2\pi_y)+p_2\pi_y(1-p_2\pi_y)$$

or

$$\pi < \frac{1-\pi_y(1-p_1+p_2)}{2p_1}. \tag{3.1}$$

Obviously, the inequality (3.1) is likely to hold for smaller values of π. The different RE values computed in Table 2 for $\pi \le 0.5$ show that the inequality (3.1) will hold for all practicable values of p_1, p_2 and π_y $(\pi_y < 0.2)$. The proposed procedure is, therefore, recommended over the Greenberg *et al.* (1969) procedure for use in practice, when $\pi \le 0.5$.

Case II. $\pi > 0.5$: Using optimal values of π_y for two procedures, it has been empirically observed that chances of the proposed estimator $\hat{\pi}_1$ to be more efficient than the Greenberg *et al*. (1969) usual estimator $\hat{\pi}$ are rather remote. Therefore, when $\pi > 0.5$, one should use the procedure due to Greenberg *et al*. instead of the proposed procedure.

Acknowledgements

The authors are thankful to the Editor for accepting the paper for publication. We are very sorry to inform that Prof. Ravindra Singh passed away Feb. 4, 2003, which is a great loss to his erstwhile students and colleagues.

REFERENCES

CHAUDHURI, A. and MUKHERJEE, R. (1988). Randomized response Theory and Techniques. New York: Marcel Dekker.

GREENBERG, B. G., ABUL-ELA, A.L.A., SIMMONS, W.R. and HORVITZ, D.G. (1969). The unrelated question randomized response model: theoretical framework. *Jour. Amer. Statist. Assoc*., 64, 520—539.

HORVITZ, D.G., SHAH, B.V. and SIMMONS, W.R. (1967). The unrelated question randomized response model. *Proc. of Social Statistics Section, Amer. Statist. Assoc*., 65—72.

MANGAT, N.S., SINGH, S. and SINGH, R. (1993). On the use of a modified randomization device in randomized response inquiries. *Metron*, 211—216.

MANGAT, N.S., SINGH, S. and SINGH, R. (1995). On use of a modified randomization device in Warner's model. *Jour. Indian Soc. Statist. and Operations Research*, 16(1—4), 65—69.

RUIZ-ESPEZO, M. (1989). Estudio de Modelos de Muestreo Aplicado. Ph. D. Thesis, Universsidad Complutense de Madrid.

TRACY, D.S. and Mangat, N.S. (1996). Some developments in randomized response sampling during the last decade — a follow up of review by Chaudhury and Mukherjee. *Journal of Applied Statistical Science*, 4(2/3), 147—158.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Jour. Amer. Statist. Assoc*., 60, 63—69.

Table 1. Variance $V\left(\hat{\pi}_1\right)$ for different values of π, p_1, p_2 and π_y.

π	p_1	p_2	π_y				
			0.1	0.3	0.5	0.7	0.9
0.1	0.6	0.30	0.2275	0.3542	0.4608	0.5475	0.6142
		0.20	0.2044	0.2933	0.3733	0.4444	0.5067
		0.10	0.1808	0.2275	0.2719	0.3142	0.3542
	0.7	0.20	0.1671	0.2308	0.2879	0.3386	0.3827
		0.15	0.1587	0.2077	0.2530	0.2946	0.3326
		0.10	0.2742	0.3265	0.3722	0.4114	0.4441
	0.8	0.15	0.1343	0.1709	0.2046	0.2356	0.2637
		0.10	0.1279	0.1529	0.1767	0.1992	0.2204
0.5	0.6	0.30	0.6142	0.6608	0.6875	0.6942	0.6808
		0.20	0.6044	0.6400	0.6666	0.6844	0.6933
		0.10	0.5942	0.6142	0.6319	0.6475	0.6608
	0.7	0.20	0.4757	0.4936	0.5051	0.5100	0.5083
		0.15	0.4730	0.4877	0.4987	0.5060	0.5097
		0.10	0.4702	0.4808	0.4897	0.4971	0.5028
	0.8	0.15	0.3793	0.3858	0.3896	0.3905	0.3887
		0.10	0.3779	0.3829	0.3867	0.3892	0.3904
0.6	0.6	0.30	0.6608	0.6875	0.6941	0.6808	0.6475
		0.20	0.6544	0.6766	0.6900	0.6944	0.6999
		0.10	0.6475	0.6608	0.6719	0.6808	0.6875
	0.7	0.20	0.5028	0.5094	0.5094	0.5028	0.4897
		0.15	0.5015	0.5070	0.5101	0.5089	0.5040
		0.10	0.5002	0.5051	0.5083	0.5100	0.5100
	0.8	0.15	0.3906	0.3896	0.3858	0.3793	0.3699
		0.10	0.3904	0.3904	0.3892	0.3867	0.3829
0.7	0.6	0.30	0.6875	0.6942	0.6808	0.6475	0.5942
		0.20	0.6844	0.6933	0.6933	0.6844	0.6666
		0.10	0.6801	0.6875	0.6919	0.6942	0.6942
	0.7	0.20	0.5100	0.5051	0.4936	0.4757	0.4512
		0.15	0.5101	0.5077	0.5015	0.4917	0.4783
		0.10	0.5102	0.5093	0.5069	0.5028	0.4971
	0.8	0.15	0.3818	0.3733	0.3621	0.3480	0.3312
		0.10	0.3829	0.3779	0.3717	0.3642	0.3554
0.9	0.6	0.30	0.6808	0.6475	0.5941	0.5208	0.4275
		0.20	0.6844	0.6666	0.6399	0.6044	0.5599
		0.10	0.6875	0.6808	0.6719	0.6608	0.6475
	0.7	0.20	0.4642	0.4365	0.4022	0.3614	0.3141
		0.15	0.4672	0.4477	0.4244	0.3975	0.3668
		0.10	0.4702	0.4579	0.4440	0.4285	0.4114
	0.8	0.15	0.3043	0.2809	0.2546	0.2255	0.1937
		0.10	0.3079	0.2929	0.2767	0.2592	0.2405

Table 2. Per cent RE of estimator $\hat{\pi}_1$ with respect to $\hat{\pi}$.

π	p_1	p_2	π_y	
			0.1	0.2
0.1	0.6	0.30	109.9	114.0
		0.20	122.3	133.0
		0.10	138.3	163.6
	0.7	0.20	109.9	115.5
		0.15	115.7	125.7
		0.10	122.3	138.1
	0.8	0.15	104.7	107.9
		0.10	109.9	117.3
0.2	0.6	0.30	105.4	108.4
		0.20	111.6	119.1
		0.10	118.8	132.9
	0.7	0.20	105.0	108.4
		0.15	107.7	113.4
		0.10	110.7	119.7
	0.8	0.15	102.2	104.0
		0.10	104.6	108.4
0.3	0.6	0.30	103.4	105.5
		0.20	107.3	112.1
		0.10	111.5	120.3
	0.7	0.20	103.0	105.1
		0.15	104.6	108.1
		0.10	106.3	111.3
	0.8	0.15	101.3	102.3
		0.10	102.6	104.8
0.4	0.6	0.30	102.3	103.6
		0.20	104.8	107.9
		0.10	102.3	103.6
	0.7	0.20	101.9	103.1
		0.15	102.9	104.9
		0.10	103.9	106.9
	0.8	0.15	100.7	101.3
		0.10	101.5	102.7
0.5	0.6	0.30	101.5	102.3
		0.20	103.1	104.9
		0.10	100.7	101.2
	0.7	0.20	101.1	101.7
		0.15	101.7	102.7
		0.10	102.3	103.8
	0.8	0.15	100.3	100.5
		0.10	100.4	99.9

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 523—532

ESTIMATION OF POPULATION MEAN USING AUXILIARY INFORMATION ON RECENT OCCASION IN h OCCASIONS SUCCESSIVE SAMPLING

G. N. Singh[1]

ABSTRACT

A general theory for estimating the population mean in the presence of auxiliary information only on the h-th (recent) occasion has been developed. Difference type estimator has been proposed and compared empirically with the sample mean estimator and another difference type estimator, which is the combination of the sample means of the matched and unmatched portion of the sample at the h-th (recent) occasion.

Key word: Difference-type estimator, successive sampling, optimum estimator.

1. Introduction

Auxiliary information plays an important role in improving the survey estimates. In case of successive sampling, it has been seen that to utilize entire information collected in the previous investigations is very advantageous. Jessen (1942) was perhaps the first author who used the information collected on the previous occasion for improving the current estimate. This technique further extended by Yates (1949), Patterson (1950), Tikkiwal (1951), Eckler (1955), Rao and Graham (1964), Binder and Hidiroglou (1988), Kish (1998), McLaren and Steel (2000), Singh et.-al. (2001), Steel and McLaren (2002) and many others.

The method usually consists in retaining a fraction of the samples selected in the previous occasions in the repeated sampling enquiries and using the entire information available from such studies with the help of regression technique. Sen (1971, 73) generalized the theory of successive sampling over two occasions to provide the optimum estimate of the current average assuming that the information on p auxiliary variables with known population means and correlated with study variable are available from the previous occasion. Further, using the

[1] Department of Applied Mathematics,Indian School of Mines, Dhanbad-826 004. INDIA;
E. Mail: gnsingh_ism@yahoo.com

auxiliary information in successive sampling Singh et-al (1991), Feng and Zou (1997) and Biradar and Singh (2001) developed the estimation techniques in successive sampling. No doubt if the interval between successive surveys is small, the information gathered on auxiliary variable at the previous occasion may efficiently be utilized to develop precise estimators. But if the time lapsed between two successive surveys is large and population characteristics are changing rapidly over time, intuitively it will not be very much fruitful to utilize such auxiliary information collected on the previous occasion except the data on the study variable used in the matched portion of the sample. However, if any kind of auxiliary information is available at the recent occasion, it may be used through appropriate sampling design to improve the estimate at the current (recent) occasion. Influenced by this argument Singh and Singh (2001) proposed a difference type estimator for estimating the population mean on second occasion in two occasions successive sampling.

The aim of this work is to present a general study for estimating the population mean of the study variable (character) at the h-th (recent) occasion when the information on an auxiliary variable (character) is available only on the h-th (recent) occasion in the successive sampling of h occasions. The proposed estimator is compared (empirically) with those estimators when no such auxiliary information is used and recommendations are made.

2. The Proposed Estimator

Let $U = (U_1, U_2, -, -, -, U_N)$ be a finite population of size N, which is sampled over h occasions. Assume that the size of the population remains constants (unchanged) but values of units changes over occasions. When the data are available for all the previous (h-1) occasions, the estimation procedure for estimating the population mean of study character y is as follows. Let the sample size at the h-th occasion be n which consists the $n'_h = n\lambda_h$ units observed on the h-th occasion and common with the (h-1)-th occasion and $n''_h = n\mu_h$ units drawn afresh (not consisting the units drawn on (h-1)-th occasion) on the h-th occasion i. e. $n = n'_h + n''_h$. We further assume that the information on an auxiliary variable x, which is positively correlated to y is available only on the h-th occasion. The sampling scheme used on all the occasions is Simple Random Sampling Without Replacement (SRSWOR). In this work we use the following notations:

$\overline{Y}_h$: The population mean of the study variable y on the h-th occasion.

S^2_{hy} : The population mean square of y on the h-th occasion.

$\overline{y}'_h$: The sample mean based on n'_h units observed on the h-th occasion and common with the (h-1)-th occasion.

$\overline{y}''_h$: The sample mean based on n''_h units drawn afresh on the h-th occasion.

$\overline{X}_h$: The population mean of auxiliary variable x, which is available only on the h-th occasion.

S_{hx}^2 : The population mean square of x on the h-th occasion.

$\overline{x}'_h$: The sample mean of x based on n'_h units common with the (h-1)-th occasion and observed on the h-th occasion.

$\overline{x}''_h$: The sample mean of x based on n''_h units drawn afresh on the h-th occasion.

$\rho_{h,h-1}$: Correlation between the measurements on y variable of the same units on the h-th and (h-1)-th occasions.

ρ_{hxy} : Correlation between x and y on the h-th occasion.

For estimating $\overline{Y}_h$ based on successive sampling, two independent difference estimators can be proposed. First, based on sample of size n''_h drawn afresh on the h-th occasion and second, based on the sample of size n'_h common to h-th and (h-1)-th occasion. The proposed estimators are given as

$$\overline{y}''_{lh} = \overline{y}''_h + \beta_{hyx}(\overline{X}_h - \overline{x}''_h) \tag{1}$$

and

$$\overline{y}'_{lh} = \overline{y}'_h + \beta_{h,h-1}(\hat{\overline{Y}}_{h-1} - \overline{y}'_{h-1}) + \beta_{hyx}(\overline{X}_h - \overline{x}'_h) \tag{2}$$

where β_{hyx} is the population regression coefficient of y on x at the h-th occasion and $\beta_{h,h-1}$ stands for the regression coefficient of the variable y of the h-th occasion on the same variable of the (h-1)-th occasion. We assume that both β_{hyx} and $\beta_{h,h-1}$ are known and that the correlation between variates observed on two occasions more than one occasion apart is zero.

Combining the estimators defined in (1) and (2), we have the final estimator of $\overline{Y}_h$ as

$$\hat{\overline{Y}}_h = \varphi_h \overline{y}''_{lh} + (1-\varphi_h)\overline{y}'_{lh} \tag{3}$$

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

It is obvious that $\hat{\overline{Y}}_h$ is the best-weighted unbiased estimator of $\overline{Y}_h$ based on the data available up to and including h-th occasion.

3. Minimum Variance of $\hat{\bar{Y}}_h$

The estimators $\bar{y}''_{lh}$ and $\bar{y}'_{lh}$ are unbiased for the population mean $\bar{Y}_h$ of the study variable y at the h-th occasion. Therefore, the variance expression of $\hat{\bar{Y}}_h$ can be written as

$$V(\hat{\bar{Y}}_h) = \varphi_h^2 V(\bar{y}''_{lh}) + (1-\varphi_h)^2 V(\bar{y}'_{lh}) + 2\varphi_h(1-\varphi_h)\mathrm{Cov}(\bar{y}''_{lh}, \bar{y}'_{lh}) \quad (4)$$

As the auxiliary information available only on the h-th (recent) occasion and following Sukhatme et-al (1984) the variance and covariance terms in (4) can be derived as

$$V(\bar{y}''_{lh}) = \left(\frac{1}{n''_h} - \frac{1}{N}\right) S_{hy}^2 (1-\rho_{hxy}^2) \quad (5)$$

$$V(\bar{y}'_{lh}) = \left[\frac{(1-\rho_{h,h-1}^2 - \rho_{hxy}^2)}{n'_h} + \frac{\varphi_{h-1}\rho_{h,h-1}^2}{n''_{h-1}} - \frac{(1-\rho_{hxy}^2)}{N}\right] S_{hy}^2 \quad (6)$$

and

$$\mathrm{Cov}(\bar{y}''_{lh}, \bar{y}'_{lh}) = -\frac{S_{hy}^2}{N}(1-\rho_{hxy}^2) \quad (7)$$

Substituting the values of variances and covariance from (5), (6) and (7) in (4) we have the expression of the exact variance of the proposed estimator $\hat{\bar{Y}}_h$. Further, to minimize the variance of $\hat{\bar{Y}}_h$ we proceed as follows:

Define a function $f:(0,1)^2 \to (0,\infty)$, which represents the expression of the variance of $\hat{\bar{Y}}_h$ in (4) by

$$f(x,y) = s\left[a\left(\frac{x^2}{n(1-y)} - \frac{1}{N}\right) + \frac{(1-x)^2}{n}\left(\frac{b}{y} + c\right)\right], \quad (8)$$

where $s = S_{hy}^2$, $a = 1-\rho_{hxy}^2$, $b = a-\rho_{h,h-1}^2$, $c = g_{h-1}\rho_{h,h-1}^2$, $g_{h-1} = \frac{\varphi_{h-1}}{\mu_{h-1}}$ and $\mu_h = 1-\lambda_h$, the variables x, y are interpreted respectively, as $x = \varphi_h$, $y = \lambda_h$. Now we want to minimize f in $(0,1)^2$. Differentiating f with respect to x and y respectively and equating the derivatives to zero we get the following simple equations

$$\frac{ax}{1-x} = (1-y)\left(\frac{b}{y} + c\right), \tag{9}$$

and $$\frac{x}{1-x} = \sqrt{\frac{b}{a}}\,\frac{1-y}{y}. \tag{10}$$

The solution for y is found easily:

$$y = \frac{\sqrt{b}(a-b)}{(\sqrt{a}+\sqrt{b})c}. \tag{11}$$

Now, since

$$\frac{1-y}{x} = 1 + \frac{(\sqrt{a}-\sqrt{b})^2}{c}, \tag{12}$$

then one finds immediately that

$$\frac{1}{g_h} = \frac{1-y}{x} = 1 + \frac{r_h}{g_{h-1}}, \tag{13}$$

where

$$r_h = \frac{\sqrt{a}-\sqrt{b}}{\sqrt{a}+\sqrt{b}} = \frac{\sqrt{1-\rho_{hxy}^2} - \sqrt{1-\rho_{hxy}^2-\rho_{h,h-1}^2}}{\sqrt{1-\rho_{hxy}^2} + \sqrt{1-\rho_{hxy}^2-\rho_{h,h-1}^2}}. \tag{14}$$

Since $a \geq 0$, for $b \geq 0$ r_h is real. The values of b entirely depend on the values of correlations. Therefore, the pair of correlations $(\rho_{hxy}, \rho_{h,h-1})$ is said to be admissible, if $b \geq 0$ for this pair.

Further, iterating, one gets

$$g_h = \left(1 + \sum_{l=1}^{h}\prod_{k=l}^{h} r_k\right)^{-1}. \tag{15}$$

On the other hand plugging (9) into (8) yields

$$V(\hat{\bar{Y}}_h)_o = f(x,y) = \frac{as}{n}\left(g_h - \frac{n}{N}\right), \tag{16}$$

which with $g_h = \frac{x}{1-y}$ is the expression for the optimal value of the variance of $\hat{\bar{Y}}_h$. Further by solving (9) and (10) the optimal values of $\varphi_h (= x)$ could be obtained as

$$x = \frac{c-(\sqrt{a}-\sqrt{b})\sqrt{b}}{c+(\sqrt{a}-\sqrt{b})^2} \quad . \tag{17}$$

4. Efficiency Comparisons

We compare the proposed estimator $\hat{\bar{Y}}_h$ with the estimator $\hat{\bar{Y}}_h^*$ (when no auxiliary information is available on the h-th occasion) and $\bar{y}$ (sample mean estimator when no previous data are used). The $\hat{\bar{Y}}_h^*$ is defined as

$$\hat{\bar{Y}}_h^* = \varphi_h^* \bar{y}_h'' + (1-\varphi_h^*)\bar{y}_{lh}'^* \tag{18}$$

where

$$\bar{y}_{lh}'^* = \bar{y}_h' + \beta_{h,h-1}(\hat{\bar{Y}}_{h-1} - \bar{y}_{h-1}') .$$

It is clear that the minimum (optimum) variance of $\hat{\bar{Y}}_h^*$ could be obtained simply by substituting $\rho_{hxy} = 0$ in the previous results. Therefore, the expression for the optimum variance of $\hat{\bar{Y}}_h^*$ is obtained as

$$V(\hat{\bar{Y}}_h^*)_o = \frac{s}{n}\left[g_h^* - \frac{n}{N} \right], \tag{19}$$

where $g_h^* = \left(1 + \sum_{l=1}^{h} \prod_{k=l}^{h} r_k^* \right)^{-1}$ and $r_h^* = \frac{1-\sqrt{(1-\rho_{h,h-1}^2)}}{1+\sqrt{(1-\rho_{h,h-1}^2)}}$.

The variance of the sample mean estimator when no previous data are used, could be obtained as

$$V(\bar{y}) = \frac{s}{n}\left[1 - \frac{n}{N} \right]. \tag{20}$$

Therefore, the percent relative efficiencies (P.R.E.) of $\hat{\bar{Y}}_h$ with respect to $\hat{\bar{Y}}_h^*$ and $\bar{y}$ under optimum conditions could be expressed as

$$E_1 = \frac{V(\hat{\bar{Y}}_h^*)_o}{V(\hat{\bar{Y}}_h)_o} \text{ x } 100 \text{ and } E_2 = \frac{V(\bar{y})}{V(\hat{\bar{Y}}_h)_o} \text{ x } 100 .$$

Since correlation coefficients are independent of the occasions, therefore, for simplicity we assume that $\rho_{hxy} = \rho_x$ and $\rho_{h,h-1} = \rho$. To compare the proposed estimator $\hat{\bar{Y}}_h$ with the estimators $\hat{\bar{Y}}_h^*$ and $\bar{y}$ under the optimal conditions, the values of ρ_x considered as 0.4, 0.6 and 0.8 and the sampling fraction $f = \frac{n}{N}$ is taken as 0.4. For different suitably chosen values of ρ, the optimum values of φ_h and the percent relative efficiencies E_1 and E_2 are shown in Tables 1—3.

Table 1. Optimum values of φ_h and P.R.E. of $\hat{\bar{Y}}_h$ for different values of h and ρ and fixed values of ρ_x and f, where $\rho_x = 0.4$ and $f = 0.4$

Occasions h	ρ				
	0.2		0.4	0.6	0.8
2	φ_h	0.4940	0.4750	0.4400	0.3824
	E_1	119.45	121.04	125.93	151.85
	E_2	121.52	130.50	154.17	251.75
3	φ_h	0.4939	0.4737	0.4318	0.3462
	E_1	119.45	121.05	126.20	158.44
	E_2	121.52	130.53	154.83	269.34
4	φ_h	0.4939	0.4736	0.4307	0.3341
	E_1	119.45	121.05	126.24	161.18
	E_2	121.52	130.53	154.93	275.73
5	φ_h	0.4939	0.4736	0.4305	0.3301
	E_1	119.45	121.05	126.25	162.23
	E_2	121.52	130.53	154.94	277.98

From Tables 1—3, it is obvious that the proposed estimator $\hat{\bar{Y}}_h$ is preferable over the estimators $\hat{\bar{Y}}_h^*$ and $\bar{y}$ under optimal conditions. It could be observed that

for fixed value of ρ_x, $\hat{\bar{Y}}_h$ is performing gradually much better with the increase in the values of ρ. But it could be noticed that for fixed values of ρ_x and ρ performance of the proposed estimator is almost stable over the different occasions. Further for fixed value of ρ there is appreciable improvement in the performance of the estimator $\hat{\bar{Y}}_h$ with the increase in the values of ρ_x. It could be further noticed that the estimator $\hat{\bar{Y}}_h$ is enhancing its performance over $\hat{\bar{Y}}_h^*$ and $\bar{y}$ with the simultaneous increase in the values of the correlation coefficients ρ_x and ρ.

Table 2. Optimum values of φ_h and P.R.E. of $\hat{\bar{Y}}_h$ for different values of h and ρ and fixed values of ρ_x and f, where $\rho_x = 0.6$ and $f = 0.4$

Occasions h	ρ				
	0.2		0.4	0.6	0.8
2	φ_h	0.4921	0.4667	0.4182	0.3816
	E_1	157.84	164.52	190.02	243.06
	E_2	160.57	177.37	232.62	333.63
3	φ_h	0.4919	0.4643	0.4022	0.3449
	E_1	157.84	164.60	192.18	258.69
	E_2	160.57	177.48	235.79	357.73
4	φ_h	0.4919	0.4641	0.3989	0.3327
	E_1	157.84	164.60	192.66	264.77
	E_2	160.57	177.49	236.44	366.60
5	φ_h	0.4919	0.4641	0.3983	0.3284
	E_1	157.84	164.60	192.77	266.99
	E_2	160.57	177.49	236.58	369.74

Table 3. Optimum values of φ_h and P.R.E. of $\hat{\bar{Y}}_h$ for different values of h and ρ and fixed values of ρ_x and f, where $\rho_x = 0.8$ and $f = 0.4$

Occasions h	ρ				
	0.2		0.4	0.6	0.8
2	φ_h	0.4857	0.4667	0.4375	0.3950
	E_1	287.13	302.95	338.44	445.71
	E_2	292.10	315.33	364.89	505.96
3	φ_h	0.4853	0.4643	0.4286	0.3670
	E_1	287.14	303.12	340.09	462.99
	E_2	292.11	315.52	366.71	525.91
4	φ_h	0.4853	0.4641	0.4273	0.3592
	E_1	287.14	303.14	340.34	468.18
	E_2	292.11	315.53	366.98	531.81
5	φ_h	0.4853	0.4641	0.4271	0.3569
	E_1	287.14	303.14	340.37	469.68
	E_2	292.11	315.54	367.02	533.53

Acknowledgements

Author is thankful to the referees for their valuable comments for improving the original version of this paper.

REFERENCES

BINDER, D. A. and Hidiroglou (1988): "Sampling in Time" in Handbook of Statistics, Vol. 6, ed. by P. R. Krishaiah and C. R. Rao, Elsevion Science Publishers, B.V., 187—211.

BIRADAR, R. S. and SINGH, H. P. (2001): Successive sampling using auxiliary information on both the occasions. Cal. Stat. Assoc. Bull., 51(203—204), 243—251.

ECKLER, A. R. (1955): Rotation sampling. A. M. S., 26, 664—685.

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

JESSEN, R. J. (1942): Statistical investigations of a sample survey for obtaining farm facts. J. Agri. Exp. Stat. Res. Bull., 304.

KISH, L. (1998): Space/time variations and rolling samples. Journal of Official Statistics, 14, No. 2, 163—172.

McLAREN, C. H. and STEEL, D. G. (2000): The impact of different rotation patterns on the sampling variance of seasonally adjusted and trend estimates. Survey Methodology, 26, No. 2, 163—172.

PATTERSON, H. D. (1950): Sampling on successive occasions with partial replacement of units. J. Royal Statist. Soc., Ser. B, 12, 241—255.

RAO, J. N. K. and GRAHAM, J. E. (1964): Rotation designs for sampling on repeated occasions. J. 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. (1973): Some theory of sampling on successive occasions. Aust. Jour. of Statist., 15, 105—110.

SINGH, V. K., SINGH, G. N. and SHUKLA, D. (1991): An efficient family of ratio-cum-difference type estimators in successive sampling over two occasions. Jour. Sci. Res., 41 C, 149—159.

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

SINGH, A. C., KENNEDY, B. and WU, S. (2001): Regression composite estimation for the Canadian labour force survey with a rotating panel design. Survey Methodology, 27, No. 2, 33—44.

STEEL D. G. and McLAREN, C. H. (2002): In search of a good rotation pattern. Submitted to World Scientific.

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

TIKKIWAL, B. D. (1951): Theory of successive sampling, Unpublished Diploma Dissertation, Institute of Agricultural Research Statistics, New Delhi, India.

YATES, F. (1949): Sampling methods for census and surveys, Charles Griffin and Company Limited, London, First Edition.

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 533—542

ESTIMATION OF A FINITE POPULATION VARIANCE IN THE PRESENCE OF AUXILIARY INFORMATION

P.A. Patel[1] **and R.D. Chaudhari**[2]

ABSTRACT

This article deals with a design-based estimation of the finite population variance. Various procedures for improving variance estimation with the aid of auxiliary information are discussed. A family of estimators that includes many estimators is suggested and its variance and variance estimators are derived. An empirical study has been carried out to study the behavior of some estimators.

Key words: Auxiliary information, design-based, variance estimation

1. Introduction

The problem of estimating a finite population mean, total or distribution function in the presence of auxiliary information has been extensively discussed in survey sampling. Estimation of quadratic or other higher-order finite population functions is also important. For example, efficient estimators for finite population variances, covariances between two response variables, or variances of linear estimators are highly desirable. Shah and Patel (1996) presented several examples to illustrate why the estimation of population variances and covariances might be useful in their own right. Recently, Wu and Sitter (2001) and Sitter and Wu (2002) have extended model calibration and pseudo empirical likelihood method to obtain efficient estimators of quadratic and other second-order finite population functions.

In this article various procedures for improving variance estimation with the aid of auxiliary information are discussed and the estimators are compared empirically.

[1] Department of Statistics, Sardar Patel University, Vallabh Vidyanagar-388 120, Gujarat (India), E-mail: patelpraful_a@indiatimes.com

[2] Statistics Department, V.P. & R.P.T.P. Science College, Vallabh Vidyanagar-388 120, Gujarat (India), E-mail: rdchaudhari@indiatimes.com

We denote the finite survey population as U = {1,…, N}. Let y and x be the study variable and an auxiliary variable, respectively, with values y_i and x_i for unit i in the population. We want to estimate the population variance.

$$S_y^2 = \sum_{i,\ j\in U} \sum c_{ij} y_i y_j$$

where $c_{ij} = 1/N$ for $i = j$ and $c_{ij} = -1/N(N-1)$ for $i \neq j$, using y_i for the observed unit $i \in s$, where s is a sample drawn from U according to a given sampling design having the first four order inclusion probabilities $\pi_{ijkl} > 0$. We write $\sum_A y_i$, $\sum\sum_A y_i y_j$ and $\sum\sum'_A y_i y_j$ for $\sum_{i\in A} y_i$, $\sum_{i,\ j\in A} \sum y_i y_j$ and $\sum_{i\neq j\in A} \sum y_i y_j$ respectively.

The customary estimator of S_y^2 under SRSWOR is the sample variance $s_y^2 = \frac{1}{n}\sum_s y_i^2 - \frac{1}{n(n-1)}\sum\sum'_s y_i y_j$, whereas under varying probability sampling is the Horvitz-Thompson (HT) type estimator

$$Q_{HTY} = \sum\sum_s c_{ij} y_i y_j / \pi_{ij} = \sum\sum_s w_{ij} y_i y_j$$

where $w_{ij} = c_{ij}/\pi_{ij}$. Notice that the estimator Q_{HTY} does not make use of auxiliary population information at the estimation stage.

The variance and variance estimator of Q_{HTY} (Liu, 1984) are respectively given by

$$V(Q_{HTY}) = \sum'_U \Delta_{ijkl}(w_{ij} y_i y_j)(w_{kl} y_k y_l) = V_{yy} \text{ (say)} \quad (1.1)$$

and

$$\hat{V}(Q_{HTY}) = \sum'_s \overset{\vee}{\Delta}_{ijkl}(w_{ij} y_i y_j)(w_{kl} y_k y_l) = v_{yy} \text{ (say)} \quad (1.2)$$

where the single summation $\sum'_A$ is shorthand for $\sum_{i\in A}\sum_{j\in A}\sum_{k\in A}\sum_{l\in A}$,

$\Delta_{ijkl} = \pi_{ijkl} - \pi_{ij}\pi_{kl}$, and $\overset{\vee}{\Delta}_{ijkl} = \Delta_{ijkl}/\pi_{ijkl}$.

In case if π_i and π_{ij} are unrelated to the y_i and if the auxiliary information, x_i with known variance $S_x^2 = \sum\sum_U w_{ij} x_i x_j$ and closely related to

y_i is available then use the generalized ratio (GR)-type estimator (Shah and Patel, 1996),

$$Q_{GR} = \frac{Q_{HTY}}{Q_{HTX}} S_x^2$$

where $Q_{HTX} = \sum\sum_s w_{ij} x_i x_j$.

2. The Estimators

2.1. The HE π-Weighted Ratio-Type Estimator

The ordinary ratio estimator of the population mean is very precise when the population points (y_i, x_i) are tightly scattered around a straight line through the origin. Motivated by the use of ratio estimator for estimation of population mean we propose the following estimator of S_y^2.

The ratio-type adjustment to each of the terms of $S_y^2 = c_1 \sum_U y_i^2 + c_2 \sum\sum_U y_i y_j$ yields the following consistent estimator

$$Q_\pi = c_1 \left(\sum_U x_i^2\right) \frac{\sum_s y_i^2 / \pi_i}{\sum_s x_i^2 / \pi_i} + c_2 \left(\sum\sum{}'_U x_i x_j\right) \frac{\sum\sum{}'_s y_i y_j / \pi_{ij}}{\sum\sum{}'_s x_i x_j / \pi_{ij}}.$$

where $c_1 = 1/N$ and $c_2 = -1/N(N-1)$.

We call it a π-weighted ratio-type estimator.

2.2. The Optimal Estimator

Shah and Patel (1994, 1994-95) proposed the generalized difference (GD) — type estimator

$$Q_{GD} = Q_{HTY} + \gamma(S_x^2 - Q_{HTX})$$

where γ is a constant. The optimum value of γ can be obtained by minimizing the variance of Q_{GD} as

$$\gamma_O = V_{xy} / V_{xx}$$

where

$$V_{xy} = \mathrm{Cov}(Q_{HTY}, Q_{HTX}) = \sum_U \Delta_{ijkl} \left(w_{ij} x_i x_j\right) \left(w_{kl} y_k y_l\right),$$

$V_{xx} = V_{xy}$ when x = y, and $\Delta_{ijkl} = \pi_{ijkl} - \pi_{ij}\pi_{kl}$.

Inserting the usual design-unbiased estimators v_{xy} and v_{xx} of V_{xy} and V_{xx}, respectively, in γ_O, we obtain the optimal estimator as

$$Q_O = Q_{HTY} + \hat{\gamma}_O (S_x^2 - Q_{HTX})$$

with $\hat{\gamma}_O = v_{xy} / v_{xx}$.

Remark 1. A great disadvantage of the optimal estimator is that it requires fourth order inclusion probabilities. This is most likely possible within the simplest design such as stratified SRSWOR (WR) sampling.

2.3. The Calibration Estimator

Recently Singh et al. (1999) have proposed a high level calibration approach for Yates-Grundy variance estimation.

The calibration estimator of S_y^2 is given by

$$Q_C = \sum\sum_s w_{ij}^* y_i y_j$$

where $\left\{ w_{ij}^* \mid i, j \in s \right\}$ is a calibration weight system determined from the auxiliary information. Minimizing the chi-square-type distance measure

$$\sum\sum_s (w_{ij}^* - w_{ij})^2 q_{ij} / w_{ij}$$

subject to the calibration constraint

$$\sum\sum_s w_{ij} x_i x_j = S_x^2 ,$$

for fixed s, leads to the calibration weight

$$w_{ij}^* = w_{ij} \left\{ 1 + \frac{(S_x^2 - Q_{HTX}) x_i x_j / q_{ij}}{\sum\sum_s w_{ij} x_i^2 x_j^2 / q_{ij}} \right\},$$

where q_{ij} are known numbers, $Q_{HTX} = \sum\sum_s w_{ij} x_i x_j$ is the Horvitz-Thompson type estimator of S_x^2.

Therefore, the calibration estimator is

$$Q_C = Q_{HTY} + \hat{\gamma}_C (S_x^2 - Q_{HTX})$$

where $\hat{\gamma}_C = \left(\sum\sum_s w_{ij}x_i x_j y_i y_j / q_{ij}\right) \Big/ \left(\sum\sum_s w_{ij}x_i^2 x_j^2 / q_{ij}\right)$

A main drawback of the calibration estimator is that some of the weights can be negative for some populations. To overcome this difficulty the simplest way is to minimize a chi-squared type distance function subject to a calibration constraint with $w_{ij}^* \geq 0, \ \forall i, j \in s$.

3. Variance Estimation

A proposed family of estimators of S_y^2 is of the form

$$Q_G = \sum\sum_s g(s,ij) w_{ij} y_i y_j \tag{3.1}$$

where the g-weights $g(s,ij) = g(s,ji)$ are constants independent of the y-values and have the following properties.

Property 1. The g-weights yield perfect estimates when applied to the x_i values; i.e.

$$\sum_{i\in s} g(s,ii) x_i^2 / \pi_i = \sum_{i\in U} x_i^2$$

$$\sum_{i\neq j\in s}\sum g(s,ij) x_i x_j / \pi_{ij} = \sum_{i\neq j\in U}\sum x_i x_j .$$

Property 2. For any given $i, j, g(s,ij)$ is a random variable, the random elements being s, whose distribution is determined by the design, and for each $i, j \in U$, $g(s,ij)$ converges in design probability to unity under the mild restrictions on the first four order inclusion probabilities and on the limiting population moments of x_i (see, e.g., Shah and Patel, 1996).

It is generally not considered desirable to have g-weights that are far from unity. One reason is that large g-weights tend to increase the estimator variance. Negative g-weights are particularly undesirable, as they may lead to negative estimates for quantities that are intrinsically positive.

This family of estimators includes the ratio-type estimator (Q_{GR}), calibration estimator (Q_C), optimum estimator (Q_O) and π-weighted ratio-type estimator (Q_π).

Using the Taylor linearization technique (see, e.g., Wolter, 1985) the approximate variance and variance estimator of Q_G can be obtained as follows :
Consider the model

$$y_i = \beta x_i + \varepsilon_i$$

with the only assumption that y is well explained by the x.
Denote by E_i the 'census fit residual'

$$E_i = y_i - Bx_i \tag{3.2}$$

where $B = \left(\sum_U x_i y_i / q_i\right) \Big/ \left(\sum_U x_i^2 / q_i\right)$ the population regression coefficient.
Using (3.2), we can write the error of (3.1) as

$$Q_G - S_y^2 = \sum\sum_U \left(\frac{g(s,ij)\delta_{ij}}{\pi_{ij}} - 1\right) c_{ij} E_i E_j + 2B \sum\sum_U \left(\frac{g(s,ij)\delta_{ij}}{\pi_{ij}} - 1\right) c_{ij} x_i E_j \tag{3.3}$$

where $\delta_{ij} = 1$ if i & $j \in s$ and = 0, otherwise.

From (1.1) the variance of Q_G is given by

$$V(Q_G) = \sum_U \Delta_{ijkl} \left(a_{ij} a_{kl} + 4B^2 b_{ij} b_{kl} + 4B a_{ij} b_{kl}\right) \tag{3.4}$$

where

$$a_{ij} = w_{ij} E_i E_j = c_{ij} E_i E_j / \pi_{ij}\,,\ b_{ij} = w_{ij} x_i E_j = c_{ij} x_i E_j / \pi_{ij}\,.$$

A quick way to obtain (3.4) is to take $g(s,ij) = 1$ for all i, j in (3.3), a large sample approximation is justified by Property 2. The usual technique is to turn (3.4) into a variance estimator by replacing the unknown E_i by its sample-based counterpart e_{is}, given by

$$e_{is} = y_i - \hat{B} x_i$$

with

$$\hat{B} = \left(\sum_s x_i y_i / q_i \pi_i\right) \Big/ \left(\sum_s x_i^2 / q_i \pi_i\right),$$

that is, the simple variance estimator is

$$v_s(Q_G) = \sum_s^{'} \overset{\vee}{\Delta}_{ijkl} \left(a_{ij}(s) a_{kl}(s) + 4\hat{B}^2 b_{ij}(s) b_{kl}(s) + 4\hat{B} a_{ij}(s) b_{kl}(s)\right)$$

An alternative estimator of (3.4) is given by

$$v_g(Q_G) = \sum_s^{'} \overset{\vee}{\Delta}_{ijkl} \left(a_{ij}^*(s) a_{kl}^*(s) + 4\hat{B}^2 b_{ij}^*(s) b_{kl}^*(s) + 4\hat{B} a_{ij}^*(s) b_{kl}^*(s)\right)$$

where $a_{ij}(s) = w_{ij} e_{is} e_{js}$, $b_{ij}(s) = w_{ij} x_{is} e_{js}$

$a_{ij}^*(s) = w_{ij} g(s,ij) e_{is} e_{js}$, $b_{ij}^*(s) = w_{ij} g(s,ij) x_{is} e_{js}$

4. Empirical Study

The goal of this section is to compare empirically the efficiency of the estimators with that of the standard estimator Q_{HTY}. These estimators make use of the auxiliary information at the estimation stage. The HT-type estimator does not. The populations used in this simulation were listed in Table 1.

Table 1. Study Population

Popu-lation	N	Source	X	Y	ρ_{xy}
1	281	Särndal et al. (1992)	CS82: Number of conservative seats in municipal council	RMT × 10^{-4}: Revenue from the 1985 municipal taxation	0.657
2	338	Chambers et al. (1986)	Area assigned for sugarcane farms	Gross value of sugarcane	0.902
3	393	Valliant et al. (2000)	Number of beds	Number of patients discharged	0.910
4	301	Valliant et al. (2000)	Adult female population, 1960	Breast cancer mortality, 1950-69 (white female)	0.967
5	304	Valliant et al. (2000)	Number of households 1960	Population, excluding residents of group quarters, 1960	0.982
6	304	Valliant et al. (2000)	Number of households 1960	Population, excluding Residents of group quarters, 1960	0.998

A scatter plot of each of the populations 3 to 6 reveals that a linear model

$$y_i = \beta x_i + v(x_i)\varepsilon_i \qquad (4.1)$$

with $v(x_i) = x_i$, $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma^2$, might be appropriate and the relationship between y and x is strong. Population 2 seems to obey the model (4.1) with $v(x_i) = \sqrt{x_i}$.

From each population, listed in Table 1, a sample of size n=30 was drawn using Sunter's (1977a, 1977b) Sampling Scheme and (b) Poisson sampling (see Särndal et al., 1992). Under each scheme the variance estimators Q_{HTY}, Q_{GR}, Q_C and Q_π were computed. This process was repeated M=10,000 times. Each sample is replaced before the next is drawn.

The performance of the different variance estimators was measured and compared in terms of percentage relative bias (% RB) and relative efficiency (RE). These summary statistics over 10,000 samples were calculated for a particular variance estimator Q as

$$\%\,\mathrm{RB}\,(Q) = 100 \times \frac{\overline{Q} - S_y^2}{S_y^2}$$

$$\mathrm{RE}\,(Q) = \frac{\mathrm{MSE}(Q_{HTY})}{\mathrm{MSE}(Q)}$$

where $\overline{Q} = \frac{1}{M}\sum_{j=1}^{M} Q_{(j)}$, $\mathrm{MSE}(Q) = \frac{1}{M-1}\sum_{j=1}^{M}\left(Q_{(j)} - S_y^2\right)^2$.

Table 2 reports the RB and RE for estimators included in the simulation study. Table 3 reports the same results under Poisson Sampling.

It is clear from Tables 2 and 3 that the RB of Q_π is negligible, less than 1% (with one exception) and is more efficient than Q_C, which in turn more efficient than Q_{GR}. Under both the scheme Q_{GR} has performed better than the standard estimator Q_{HTY} for the first three populations under study.

Table 2. % RB and RE Under Sunter's Method

	Q_{HTY}	Q_{GR}	Q_C	Q_π
Population No.		**% Relative Bias**		
1	0.008	16.281	2.801	-0.036
2	-0.459	0.703	0.057	0.592
3	-1.488	-0.039	0.133	0.773
4	-0.385	-0.033	-0.065	-0.107
5	-0.309	0.162	0.066	-0.006
6	-0.358	-0.046	-0.016	0.113
		Relative Efficiency		
1	1.000	0.014	1.092	1.327
2	1.000	0.033	5.598	6.852
3	1.000	2.768	3.092	3.165
4	1.000	5.505	5.546	5.886
5	1.000	5.859	5.459	6.355
6	1.000	86.947	88.863	84.306

Table 3. % RB and RE Under Poisson Sampling

	Q_{HTY}	Q_{GR}	Q_C	Q_π
Population No.		**% Relative Bias**		
1	0.187	228.040	3.577	-0.341
2	5.278	28.364	-0.368	0.219
3	1.426	1.271	0.993	1.198
4	-0.042	0.074	-0.003	0.068
5	0.272	0.147	0.017	0.061
6	0.311	-0.004	0.047	0.115
		Relative Efficiency		
1	1.000	0.000	0.468	1.288
2	1.000	0.000	9.879	12.146
3	1.000	0.041	3.898	4.434
4	1.000	5.934	5.966	6.266
5	1.000	6.697	6.152	7.127
6	1.000	67.897	74.087	86.959

5. Conclusion

The absolute values of RBs for estimators Q_{HTY}, Q_{GR}, Q_C and Q_π are all within reasonable range (with a few exceptions). In terms of RE, Q_π and Q_C estimators outperformed the conventional HT-estimator Q_{HTY} under both the schemes. The ratio-type estimator Q_{GR} is better than Q_{HTY} for the last three populations. One possible reason for population 1 is than the correlation between y and x is moderate while for population 2 is that model variance of y_i is proportional to $x_i^{1/2}$. It is quite surprising that for population 3, Q_{GR} has performed better than Q_{HTY} under Sunter's scheme but this is reversed under Poisson scheme. If the population data are well described by a straight line passing through origin, the estimator Q_π normally will bring about a large variance reduction, as compared to the π-estimator, Q_{HTY}.

REFERENCES

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

LIU, T.P. (1984). Existence of a Sen-Yates-Grundy form of a dispersion. *Australian Journal of Statistics.* 26(1), 97—104.

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

SHAH, D. N. and PATEL, P. A. (1994—95). Uniformly minimum variance unbiased estimation for finite population variance. *Gujarat Statistical Review.*, (21—22), 17—24.

SHAH, D.N. and PATEL P. A. (1994) Optimum estimation for a finite population variance under certain super-population model. *Journal of Indian Statistical Association.*, 32(1), 21—28.

SHAH, D.N. and Patel, P.A. (1996). Asymptotic properties of a generalized regression-type predictor of a finite population variance in probability sampling. *Canadian Journal of Statistics.*, 24(3), 373—384.

SINGH, S, Horn, S. Chowdhury, S. and Yu. S. (1999). Calibration of the estimators of variance. *Australian And New Zealand Journal of Statistics.*, 41, 199—212.

SUNTER, A. B (1977a). Response burden, sample rotation, and classification renewal in economic surveys. *International Statistical Review.*, 45, 209—222.

SUNTER, A. B. (1977b). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics.* 26, 261—268.

SITTER and Wu (2002). Efficient Estimation of Quadratic Finite Population Functions in the Presence of Auxiliary Information. *Journal of American Statistical Association*, (To appear).

VALLIANT, R., DORFMAN, A. H. and ROYALL, R. M. (2000). *Finite Population Sampling and Inference*. John Wiley and Sons, Inc. New York.

WOLTER, K.M. (1985). *Introduction to Variance Estimation.* Springer Verlag, New York.

Wu and Sitter (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics.* Vol. 29 (7).

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 543—554

SECOND ORDER PROPERTIES OF SOME ESTIMATORS UNDER DOUBLE SAMPLING

M. I. Hossain[1], **M. S. Rahman and M. S. Ahmed**[2]

ABSTRACT

In this paper, we have derived two estimators on the guild line of Srivastava (1967) and Walsh (1970) in case of double sampling. The comparisons among these estimators with traditional ratio and product estimators based on double sampling scheme have been presented. We have derived the second order biases and mean square errors of traditional ratio and product estimators and our estimators. Finally, a numerical comparisons are given for relative comparisions.

***Key words and Phrases*:** Double sampling, auxiliary variables bias and mean square error.

1. Introduction

Srivastava (1967), Walsh (1970), Upadhyaya *et. al*. (1990, 1992), Srivastava *et. al*. (1990), Prasad *et. al*. (1992), Sahoo and Sahoo (1993) and Singh (1993) suggested several ratio and product estimators for a finite population mean using an auxiliary variable, which is available for the population. Here, we have proposed some estimators for the situations where the auxiliary variable is not available for population but available for large sample.

Suppose $U = \{U_1, U_2, ..., U_i, ..., U_N\}$ denotes a finite population of N distinct and identifiable units. Suppose X is an auxiliary variable available for estimating the population mean $\overline{Y}$ of a study variable Y. We assume that the auxiliary variable X is not known for the population but available for a large sample with less cost. In this situation, the method of double sampling is preferable to estimate $\overline{Y}$. Suppose the first phase sample S_1 (say) of fixed size n'

[1] Statistics Section, BARI, Gazipur, Bangladesh.
[2] Department of Statistics, Jahangirnagar University, Dhaka 1342, Bangladesh. Present address, Department of Statistics, Yarmouk University, Jordan.

is drawn from the population by using simple random sampling without replacement (SRSWOR) scheme and observed auxiliary variable *X*. Then a second phase sample S_2 (say) of size $n(n < n')$ is drawn from S_1 by SRSWOR to collect the information of study variable *Y*. Define, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{x}' = \frac{1}{n}\sum_{i=1}^{n} x_i'$. Then $\bar{x}'$ is an unbiased for $\bar{X}$ based on first phase sample, and $\bar{y}$ and $\bar{x}$ are the unbiased estimators of $\bar{Y}$ and $\bar{X}$ respectively, based on second phase sample. Further, $E_2(\bar{x}) = \bar{x}'$, $C_2(\bar{x}, \bar{x}') = 0$ and $C_2[\bar{x}, E_2(\bar{y})] = 0$

where E_2 and C_2 are conditional expectation and covariance respectively, given S_1.

2. Some Estimators in Double Sampling

For estimating the population mean of *Y*, the classical ratio estimator of the population mean $\bar{Y}$ is given by

$$t_{1d} = \bar{y}\frac{\bar{x}'}{\bar{x}} \tag{2.1}$$

where *d* stands for double sampling scheme.

The classical product type estimator is given by

$$t_{2d} = \bar{y}\frac{\bar{x}}{\bar{x}'} \tag{2.2}$$

On the guild of Srivastava (1967), we have considered a general estimator

$$t_{3d} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)^{\alpha} \tag{2.3}$$

where α is a suitably constant and obtained by minimizing mean square error (MSE). For α =1, t_{3d} is the same as traditional ratio estimator and for α = -1, it is traditional product estimator.

Similarly, we have considered another estimator from Walsh (1970)'s

$$t_{4d} = \bar{y}[\bar{x}'\{\theta\bar{x} + (1-\theta)\bar{x}'\}^{-1}] \tag{2.4}$$

where θ is the constant and suitably chosen by minimizing mean square error.

Define, $C_{ab} = \frac{\sigma_{ab}}{\bar{X}\bar{Y}}$, where $\sigma_{ab} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^a (y_i - \bar{Y})^b$ for *a*, *b* =0,1,2, 3.

The first order biases and mean square error (MSE) of the estimators t_{1d} and t_{2d} estimators are given respectively as

$$Bias(t_{1d}) = \bar{Y}\left(\frac{1}{n} - \frac{1}{n'}\right)(C_{20} - C_{11}) \tag{2.5}$$

$$Bias(t_{2d}) = \bar{Y}\left(\frac{1}{n} - \frac{1}{n'}\right)C_{11} \tag{2.6}$$

$$MSE(t_{1d}) = \bar{Y}^2\left[\left(\frac{1}{n} - \frac{1}{n'}\right)(C_{02} + C_{20} - 2C_{11}) + \left(\frac{1}{n'} - \frac{1}{N}\right)C_{02}\right] \tag{2.7}$$

and

$$MSE(t_{2d}) = \bar{Y}^2\left[\left(\frac{1}{n} - \frac{1}{n'}\right)(C_{02} + C_{20} + 2C_{11}) + \left(\frac{1}{n'} - \frac{1}{N}\right)C_{02}\right] \tag{2.8}$$

If ρ is the correlation coefficient between *Y* and *X* then , $\rho = \frac{C_{11}}{\sqrt{C_{20}C_{02}}}$

First order bias and mean square error (MSE) of the estimator t_{3d} are given respectively,

$$Bias(t_{3d}) = \frac{\bar{Y}}{2}\left(\frac{1}{n} - \frac{1}{n'}\right)(\alpha^2 C_{20} - 2\alpha C_{11}) \tag{2.9}$$

$$MSE(t_{3d}) = \bar{Y}^2\left[\left(\frac{1}{n} - \frac{1}{n'}\right)(C_{02} + \alpha^2 C_{20} - 2\alpha C_{11}) + \left(\frac{1}{n'} - \frac{1}{N}\right)C_{02}\right] \tag{2.10}$$

By minimizing $MSE(t_{3d})$, the optimum value of α, $\alpha_0 = \rho\sqrt{\frac{C_{02}}{C_{20}}}$ and the expression for the bias and MSE of t_{3d}, for the optimum value of α are given respectively by

$$Bias(t_{3d})_{opt} = \frac{\bar{Y}}{2}\left(\frac{1}{n} - \frac{1}{n'}\right)\rho^2\sqrt{C_{02}} \tag{2.11}$$

$$MSE(t_{3d})_{\min} = \bar{Y}^2\left[\left(\frac{1}{n} - \frac{1}{n'}\right)(1 - \rho^2)C_{02} + \left(\frac{1}{n'} - \frac{1}{N}\right)C_{02}\right] \tag{2.12}$$

The expression for the bias and MSE of t_{4d} to the first order of approximation are given respectively by

$$Bias(t_{4d}) = \bar{Y}\left(\frac{1}{n} - \frac{1}{n'}\right)\left(\theta^2 C_{20} - \theta C_{11}\right) \tag{2.13}$$

and

$$MSE(t_{4d}) = \bar{Y}^2\left[\left(\frac{1}{n} - \frac{1}{n'}\right)\left(C_{02} + \theta^2 C_{20} - 2\theta C_{11}\right) + \left(\frac{1}{n'} - \frac{1}{N}\right)C_{02}\right] \tag{2.14}$$

By minimizing $MSE(t_{4d})$, the optimum value of θ, $\theta_0 = \rho\sqrt{\frac{C_{02}}{C_{20}}}$ and the expression for the bias and MSE of t_{4d}, for the optimum value of θ are given respectively by

$$Bias(t_{4d})_{opt} = 0$$

and $MSE(t_{3d})_{\min} = MSE(t_{2d})_{\min}$

We observed that for the optimum case the estimator is unbiased but MSE of t_{3d} and t_{4d} are same. It is observed that the mean square errors of estimators t_{3d} and t_{4d} are always less than t_{1d} and t_{2d}. But the estimators t_{3d} and t_{4d} have same variance. To find the most efficient estimator among t_{3d} and t_{4d}, we have tried to find their second order biases and mean square errors.

Now we have discussed the bias and MSE of the given estimators up to the second order of approximation under simple random sampling without replacement (SRSWOR).

Define, $\delta_{\bar{y}} = \frac{\bar{y} - \bar{Y}}{\bar{Y}}$, $\delta_{\bar{x}} = \frac{\bar{x} - \bar{X}}{\bar{X}}$ and $\delta_{\bar{x}'} = \frac{\bar{x}' - \bar{X}}{\bar{X}}$ then

$$E(\delta_{\bar{x}}) = E(\delta_{\bar{x}'}) = E(\delta_{\bar{y}}) = 0$$

Before obtaining the bias and MSE of these estimators up to the second order approximation, we state the following lemmas.

Lemma 1

For SRSWOR at both the phases,

$V(\delta\bar{y}) = E\{(\delta\bar{y})^2\} = L_1 C_{02}$, $V(\delta\bar{x}) = E\{(\delta\bar{x})^2\} = L_1 C_{20}$,

$V(\delta\bar{x}') = E\{(\delta\bar{x}')^2\} = L_1' C_{20}$, $COV(\delta\bar{x}\delta\bar{y}) = E(\delta\bar{x}\delta\bar{y}) = L_1 C_{11}$ and

$COV(\delta\bar{x}'\delta\bar{y}) = E(\delta\bar{x}'\delta\bar{y}) = L_1' C_{11}$

where $L_1 = \dfrac{(N-n)}{n(N-1)}$ and $L_1' = \dfrac{(N-n')}{n(N-1)}$

Lemma 2

For SRSWOR at both the phases,

$E(\delta\bar{x}^2\delta\bar{y}) = L_2 C_{21}$, $E(\delta\bar{x}^3) = L_2 C_{30}$, $E(\delta\bar{x}'^2\delta\bar{y}) = L_2' C_{21}$

$E(\delta\bar{x}'^3) = L_2' C_{30}$

where $L_2 = \dfrac{(N-n)(N-2n)}{n^2(N-1)(N-2)}$ and $L_2' = \dfrac{(N-n')(N-2n')}{n'^2(N-1)(N-2)}$

Lemma 3

For SRSWOR at both the phases,

$E(\delta\bar{x}^3\delta\bar{y}) = L_3 C_{31} + 3L_4 C_{20} C_{11}$, $E(\delta\bar{x}^4) = L_3 C_{40} + 3L_4 C_{20}^2$,

$E(\delta\bar{x}'^3\delta\bar{y}) = L_3' C_{31} + 3L_4' C_{20} C_{11}$, $E(\delta\bar{x}'^4) = L_3' C_{40} + 3L_4' C_{20}^2$,

$E(\delta\bar{x}^2\delta\bar{y}^2) = L_3 C_{22} + L_4 (C_{20} C_{02} C_{11}^2)$ and

$E(\delta\bar{x}'^2\delta\bar{y}^2) = L_3' C_{22} + L_4' (C_{20} C_{02} C_{11}^2)$

where $L_3 = \dfrac{(N-n)(N^2+N-6nN+6n^2)}{n^3(N-1)(N-2)(N-3)}$,

$L_3' = \dfrac{(N-n')(N^2+N-6n'N+6n'^2)}{n'^3(N-1)(N-2)(N-3)}$, $L_4 = \dfrac{N(N-n)(N-n-1)(n-1)}{n^3(N-1)(N-2)(N-3)}$

and $L_4' = \dfrac{N(N-n')(N-n'-1)(n'-1)}{n'^3(N-1)(N-2)(N-3)}$

Proof of these lemmas and straight forward by using SRSWOR, see, Sukhatme *et al*.(1984).

Re-write t_{1d} as

$$t_{1d} = \bar{y}\frac{\bar{x}'}{\bar{x}} = \bar{Y}(1+\delta\bar{y})(1+\delta\bar{x}')(1+\delta\bar{x})^{-1}$$

$$\Rightarrow (t_{1d} - \bar{Y}) = \bar{Y}\left[\begin{array}{l}\delta\bar{y} + \delta\bar{x}' + \delta\bar{x}'\delta\bar{y} - \delta\bar{x} - \delta\bar{x}\delta\bar{y} - \delta\bar{x}\delta\bar{x}' + \delta\bar{x}\delta\bar{x}'\delta\bar{y} + \delta\bar{x}^2 + \delta\bar{x}^2\delta\bar{y} + \\ \delta\bar{x}^2\delta\bar{x}' + \delta\bar{x}^2\delta\bar{x}'\delta\bar{y} - \delta\bar{x}^3\delta\bar{y} - \delta\bar{x}^3\delta\bar{x}' + \delta\bar{x}^4 + ...\end{array}\right]$$

Taking expectation up to second order approximation, we have

$$E(t_{1d}-\bar{Y})=\bar{Y}\begin{bmatrix}E(\delta\bar{x}^2)-E(\delta\bar{x}\delta\bar{y})+E(\delta\bar{x}\delta\bar{y})-E(\delta\bar{x}'^2)+E(\delta\bar{x}^2\delta\bar{y})-E(\delta\bar{x}'^2\delta\bar{y})+\\ E(\delta\bar{x}'^3)-E(\delta\bar{x}^3)+E(\delta\bar{x}'^3\delta\bar{y})-E(\delta\bar{x}^3\delta\bar{y})-E(\delta\bar{x}^3)+E(\delta\bar{x}^4)\end{bmatrix}$$

Using Lemma1, Lemma 2 and Lemma 3, we get the following expression for the bias of t_{1d} up to the second order of approximation.

$$Bias_2(t_{1d})=\bar{Y}\begin{bmatrix}L_1C_{20}-L_1C_{11}+L_1'C_{11}-L_1'C_{20}+L_2C_{21}-L_2'C_{30}-L_2C_{30}+L_3'C_{31}-\\ 3L_4C_{20}C_{11}+L_3C_{11}+3L_4'C_{20}C_{11}-L_3'C_{40}-3L_4'C_{20}^2+L_3C_{40}+3L_4C_{20}^2\end{bmatrix}$$

$$=\bar{Y}\begin{bmatrix}L_1(C_{20}-C_{11})-L_1'(C_{20}-C_{11})+L_2(C_{21}-C_{30})-L_2'(C_{20}-C_{30})+\\ L_3(C_{40}-C_{31})-L_3'(C_{40}-C_{31})+3L_4(C_{20}^2-C_{20}C_{11})-3L_4'(C_{20}^2-C_{20}C_{11})\end{bmatrix}$$

Now, the MSE of t_{1d} is

$$E(t_{1d}-\bar{Y})^2=\bar{Y}^2E\begin{bmatrix}\delta\bar{y}^2+\delta\bar{x}^2+2\delta\bar{x}'\delta\bar{y}-\delta\bar{x}'-2\delta\bar{x}\delta\bar{y}+4\delta\bar{x}^2\delta\bar{y}-2\delta\bar{x}\delta\bar{y}^2 4\delta\bar{x}'^2\delta\bar{y}+2\delta\bar{x}'^2\delta\bar{y}^2\\ -2\delta\bar{x}^3+2\delta\bar{x}'^3-6\delta\bar{x}^3\delta\bar{y}+3\delta\bar{x}^2\delta\bar{y}^2-3\delta\bar{x}'^2\delta\bar{y}^2+3\delta\bar{x}^3-3\delta\bar{x}'^4+.....\end{bmatrix}$$

Using Lemma1, Lemma 2 and Lemma 3, the MSE of t_{1d} up to the second order of approximation is

$$MSE_2(t_{1d})=\bar{Y}^2\begin{bmatrix}L_1C_{02}+L_1C_{20}+L_1'C_{20}+2L_1'C_{20}-2L_1C_{21}-4L_2C_{21}-2L_2C_{12}-4L_2'C_{21}+\\ 2L_2'C_{12}-2L_2C_{30}+2L_2'C_{30}-6L_3C_{31}+18L_4'C_{20}C_{11}+3L_3C_{22}+3L_4(C_{20}C_{02}\\ +2C_{11}^2)-3L_3'C_{22}-3L_4'(C_{20}C_{02}+2C_{11}^2)+3L_3C_{40}+9L_4C_{40}-3L_3'C_{40}+9L_4'C_{20}^2\end{bmatrix}$$

$$=\bar{Y}^2\begin{bmatrix}L_1(C_{02}+C_{20}-C_{11})+L_1'(C_{20}-2C_{21})+2L_2(2C_{21}-C_{12}-C_{30})+\\ 2L_2'(2C_{21}-C_{12}-C_{30})+3L_3(C_{40}+C_{22}-2C_{31})-3L_3'(C_{40}+C_{22}-2C_{31})\\ +3L_4(3C_{20}^2+C_{20}C_{02}+2C_{11}^2-6C_{20}C_{11})-3L_4'(3C_{20}^2+C_{20}C_{02}+2C_{11}^2-6C_{20}C_{11})\end{bmatrix}$$

Again, rewrite t_{2d}

$$t_{2d}=\bar{y}\frac{\bar{x}}{\bar{x}'}=\bar{Y}(1+\delta\bar{y})(1+\delta\bar{x})(1+\delta\bar{x}')^{-1}$$

$$\Rightarrow(t_{2d}-\bar{Y})=\bar{Y}\begin{bmatrix}\delta\bar{y}+\delta\bar{x}-\delta\bar{x}'\delta\bar{y}-\delta\bar{x}'+\delta\bar{x}\delta\bar{y}+\delta\bar{x}\delta\bar{x}'-\delta\bar{x}\delta\bar{x}'\delta\bar{y}+\delta\bar{x}'^2+\delta\bar{x}'^2\delta\bar{y}+\\ \delta\bar{x}'^2\delta\bar{x}+\delta\bar{x}'^2\delta\bar{x}\delta\bar{y}-\delta\bar{x}'^3\delta\bar{y}-\delta\bar{x}'^3\delta\bar{x}+\delta\bar{x}'^4+.....\end{bmatrix}$$

Hence, the second order bias and mean square error up to second order approximation are given by

$$Bias_2(t_{2d}) = \bar{Y}\begin{bmatrix} L_1'C_{20} - L_1'C_{11} + L_1C_{11} - L_1C_{20} + L_2'C_{21} - -L_2'C_{30} - L_2C_{20} + L_2C_{30} - L_3'C_{31} + \\ 3L_4C_{20}C_{11} + L_3C_{31} - 3L_4'C_{20}C_{11} + L_3'C_{40} - 3L_4'C_{20}^2 - L_3C_{40} + 3L_4C_{20}^2 \end{bmatrix}$$

$$= \bar{Y}\begin{bmatrix} L_1'(C_{20} - C_{11}) - L_1(C_{20} - C_{11}) + L_2'(C_{21} - C_{30}) - L_2(C_{20} - C_{30}) + \\ L_3'(C_{40} - C_{31}) - L_3(C_{40} - C_{31}) + 3L_4'(C_{20}^2 - C_{20}C_{11}) - 3L_4(C_{20}^2 - C_{20}C_{11}) \end{bmatrix}$$

and

$$MSE_2(t_{2d}) = \bar{Y}^2\begin{bmatrix} L_1C_{02} - L_1C_{20} + L_1'C_{20} + 2L_1'C_{20} - 2L_1C_{21} - 4L_2C_{21} - 2L_2C_{12} - 4L_2'C_{21} + \\ 2L_2'C_{12} - 2L_2C_{30} + 2L_2'C_{30} - 6L_3C_{31} + 18L_4'C_{20}C_{11} + 3L_3C_{22} - 3L_4(C_{20}C_{02} \\ + 2C_{11}^2) - 3L_3'C_{22} + 3L_4'(C_{20}C_{02} + 2C_{11}^2) + 3L_3C_{40} + 9L_4C_{40} - 3L_3'C_{40} + 9L_4'C_{20}^2 \end{bmatrix}$$

$$= \bar{Y}^2\begin{bmatrix} L_1(C_{02} - C_{20} + C_{11}) + L_1'(C_{20} - 2C_{21}) - 2L_2(2C_{21} - C_{12} - C_{30}) + \\ 2L_2'(2C_{21} - C_{12} - C_{30}) - 3L_3(C_{40} + C_{22} - 2C_{31}) + 3L_3'(C_{40} + C_{22} - 2C_{31}) \\ - 3L_4(3C_{20}^2 + C_{20}C_{02} + 2C_{11}^2 - 6C_{20}C_{11}) + 3L_4'(3C_{20}^2 + C_{20}C_{02} + 2C_{11}^2 - 6C_{20}C_{11}) \end{bmatrix}$$

For estimator t_{3d}

$$t_{1d} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)^{\alpha} = \bar{Y}(1+\delta\bar{y})(1+\delta\bar{x}')^{\alpha}(1+\delta\bar{x})^{-\alpha}$$

The second order bias and MSE of t_{3d}

$$E(t_{1d} - \bar{Y}) = \bar{Y}\left[\alpha^2 E(\delta\bar{x}^2) - 2\alpha E(\delta\bar{x}\delta\bar{y}) + 2\alpha E(\delta\bar{x}\delta\bar{y}) - \alpha^2 E(\delta\bar{x}\delta\bar{x}')\right]$$

$$Bias_1(t_{3d}) = \frac{\bar{Y}}{2}(L_1\alpha^2 C_{20} - 2L_1\alpha C_{11} + 2L_1'\alpha C_{11} - L_1'\alpha^2 C_{20})$$

$$= \frac{\bar{Y}}{2}[L_1\alpha^2(C_{20} - C_{11}) + 2\alpha L_1'(C_{11} - C_{20})$$

$$Bias_2(t_{3d}) = \frac{\bar{Y}}{2}\begin{bmatrix} L_1\alpha^2 C_{20} - 2\alpha L_1 C_{11} + 2\alpha L_1' C_{11} - L_1'\alpha^2 C_{20} - L_2'\alpha^2 C_{21} + \alpha^2 L_2 C_{21} + \\ \frac{\alpha^3}{3}L_2' C_{30} - \frac{\alpha^3}{3}L_2 C_{30} + \frac{\alpha^3}{4}L_3' C_{31} - \alpha^3 L_4 C_{20}C_{11} + \frac{\alpha^4}{12}L_3 C_{40} + \\ \alpha^3 L_4' C_{20}C_{11} - \frac{\alpha^4}{12}L_3' C_{40} - \frac{\alpha^4}{4}L_4' C_{20}^2 + L_3 C_{40} + \frac{\alpha^4}{4}L_4 C_{20}^2 \end{bmatrix}$$

$$= \frac{\overline{Y}}{2}\begin{bmatrix} L_1(\alpha^2 C_{20} - 2\alpha C_{11}) + L_1'(\alpha^2 C_{20} - 2\alpha C_{11}) + L_2(\alpha^2 C_{21} - \frac{\alpha^3}{3} C_{30}) - \\ L_2'(\alpha^2 C_{21} - \frac{\alpha^3}{3} C_{30}) - L_3(\frac{\alpha^4}{12} C_{40} - \frac{\alpha^3}{3} C_{31}) - L_3'(\frac{\alpha^4}{12} C_{40} - \frac{\alpha^3}{3} C_{31}) \\ - 3L_4(\frac{\alpha^4}{12} C_{20}^2 - \frac{\alpha^3}{3} C_{20} C_{11}) - 3L_4'(\frac{\alpha^4}{12} C_{20}^2 - \frac{\alpha^3}{3} C_{20} C_{11}) \end{bmatrix}$$

$$MSE_2(t_{3d}) = \overline{Y}^2 \begin{bmatrix} L_1(C_{02} + \alpha^2 C_{20} - 2\alpha C_{11}) + L_1'(C_{20} - 2\alpha C_{21}) + L_2(3\alpha C_{21} - 2\alpha C_{12} - \alpha^3 C_{30}) - \\ L_2'(3\alpha C_{21} - 2\alpha C_{12} - \alpha^3 C_{30}) + L_3(\frac{7\alpha^4}{4} C_{40} + 2\alpha^2 C_{22} - \frac{7\alpha^3}{3} C_{31}) - \\ L_3'(\frac{7\alpha^4}{4} C_{40} + 2\alpha^2 C_{22} - \frac{7\alpha^3}{3} C_{31}) + L_4(\frac{21\alpha^4}{4} C_{20}^2 - \frac{21\alpha^3}{3} C_{20} C_{11} + 2\alpha^2 C_{20} C_{02} \\ + 4\alpha^2 C_{11}^2) - L_4'(\frac{21\alpha^4}{4} C_{20}^2 - \frac{21\alpha^3}{3} C_{20} C_{11} + 2\alpha^2 C_{20} C_{02} + 4\alpha^2 C_{11}^2) \end{bmatrix}$$

The optimum value of α is obtained by minimizing $MSE_2(t_{3d})$ and the solution for the determination of this optimum value is obtained by numerical iterations.

Similarly for the estimator t_{4d}

$$t_{4d} = \bar{y}[\bar{x}'\{\theta\bar{x} + (1-\theta)\bar{x}'\}^{-1}]$$

The bias and MSE of second order of this estimator are given by

$$Bias_2(t_{4d}) = \frac{\overline{Y}}{2}\begin{bmatrix} L_1(\theta^2 C_{20} - \theta C_{11}) - L_1'(\theta^2 C_{20} - \theta C_{11}) + L_2(\theta^2 C_{21} - \theta^3 C_{30}) - \\ L_2'(\theta^2 C_{21} - \theta^3 C_{30}) + L_3(\theta^4 C_{40} - \theta^3 C_{31}) - L_3'(\theta^4 C_{40} - \theta^3 C_{31}) \\ - 3L_4(\theta^4 C_{20}^2 - \theta^3 C_{20} C_{11}) - 3L_4'(\theta^4 C_{20}^2 - \theta^3 C_{20} C_{11}) \end{bmatrix}$$

$$MSE_2(t_{4d}) = \overline{Y}^2 \begin{bmatrix} L_1(C_{02} + \theta^2 C_{20} - 2\theta C_{11}) - L_1'(C_{20} - 2\theta C_{11}) + 2L_2(2\theta C_{21} - \theta C_{12} - \theta^3 C_{30}) - \\ 2L_2'(2\theta^2 C_{21} - \theta C_{12} - \theta^3 C_{30}) + 3L_3(\theta^4 C_{40} - \theta^2 C_{22} - 2\theta^3 C_{31}) - \\ 3L_3'(\theta^4 C_{40} - \theta^2 C_{22} - 2\theta^3 C_{31}) + 3L_4(3\theta^4 C_{20}^2 - \theta^2 C_{20} C_{11} - 6\theta^3 C_{20} C_{11} \\ + 2\theta^2 C_{11}^2) + 3L_4'(3\theta^4 C_{20}^2 - \theta^2 C_{20} C_{11} - 6\theta^3 C_{20} C_{02} + 2\theta^2 C_{11}^2) \end{bmatrix}$$

The optimum value of θ is obtained by minimizing $MSE_2(t_{4d})$ using numerical iteration techniques. We have considered the following numerical illustrations.

3. Numerical Illustrations for comparisons

For the two natural population data, we calculate the first and second order biases and the mean square errors of the above estimators. We also make a simulation study for 5000 repeated samples. In simulation study has been carried out by using "*Mathematica Software*" with same first and second phases sample sizes.

Data Set-1

The data for the empirical analysis are taken from 1981, Utter Pradesh District Census Handbook, Aligarh. The population consists of 340 villages under Koil police station, with Y=Number of agricultural labor in 1981 and X=Area of the villages (in acre) in 1981. The following values are obtained

$$\overline{Y} = 73.76765,\ \overline{X} = 2419.04,\ N = 340, n' = 120, n = 70,\ C_{02} = 0.7614,$$
$$C_{20} = 0.5557,\ C_{11} = 0.2667,\ C_{03} = 2.6942, C_{12} = 0.0747, C_{21} = 0.1589,$$
$$C_{30} = 0.7877, C_{13} = 0.1321,$$
$$C_{31} = 0.8851, C_{04} = 17.4275,\ C_{22} = 0.8424, C_{40} = 1.3051$$

Data Set-2

The data for the empirical analysis are taken from 1981, Utter Pradesh District Census Handbook, Aligarh. The population consist of 340 villages under Koil police station, with Y=Number of cultivators in the villages in 1981 and X=Area of the villages (in acre) in 1981. The following values are obtained

$$\overline{Y} = 141.1294,\ \overline{X} = 2419.04,\ N = 340, n' = 120, n = 70,\ C_{02} = 0.7614,$$
$$C_{20} = 0.5944,\ C_{11} = 0.2667,\ C_{03} = 2.6942, C_{12} = 0.4720, C_{21} = 0.4897,$$
$$C_{30} = 0.7877, C_{13} = 1.3923,$$
$$C_{31} = 1.5586, C_{04} = 17.4275,\ C_{22} = 1.2681, C_{40} = 2.8457$$

Table 3.1. The biases and mean square errors of the different estimators for the data set-1.

Estimator	Bias		MSE	
	First order	Second order	First order	Second order
t_{1d}	0.12726	0.12696	46.86763	48.35583
t_{2d}	0.11744	0.11863	82.52197	84.42031
t_{3d}	-0.02818	-0.02820	42.98526	43.08889
t_{4d}	0.00000	-0.00031	42.98526	43.03965

Table 3.2. The biases and mean square errors of the different estimators for the data set-2.

Estimator	Bias		MSE	
	First order	Second order	First order	Second order
t_{1d}	0.18292	0.18392	153.50681	155.00170
t_{2d}	0.31788	0.32924	332.95420	334.54630
t_{3d}	-0.10088	-0.10047	144.07733	144.67843
t_{4d}	0.00000	-0.00030	144.07733	144.39315

Table 3.3. The simulated biases and mean square errors of the different estimators for the data set-1.

Estimator	Simulated Bias	Simulated MSE
t_{1d}	0.009374	59.2557
t_{2d}	0.000762	120.0051
t_{3d}	0.0000386	38.00628
t_{4d}	0.0000129	37.7563

Table 3.4. The simulated biases and mean square errors of the different estimators for the data set-2.

Estimator	Simulated Bias	Simulated MSE
t_{1d}	0.0264813	268.24450
t_{2d}	0.985456	825.00245
t_{3d}	0.0000211	182.32742
t_{4d}	-0.0000013	179.30216

From Table 4.1 and Table 4.2 present the first and second order approximation the estimators t_{1d}, t_{2d}, t_{3d} and t_{4d} for two sets of data. The estimator t_{2d} is a product estimator and it's usually considered in case of negative correlation. So we have got the bias and mean square errors are quite large. For the classical ratio estimator t_{1d}, the bias and mean square errors are increased for second order. The estimators t_{3d} and t_{4d} have the same mean square error for first order but the mean square error of t_{4d} is less than t_{3d} for second order. So the second order mean square error of t_{3d} makes differ from t_{4d} for both data sets. Finally, the estimator t_{4d} shows better performance compare to other estimators. The simulation results in Table 4.3 and Table 4.4 almost exhibit similar performance.

REFERENCES

PRASAD, B., SINGH, R.S. and SINGH, H.P. (1992). Modified chain ratio estimator for finite population mean using two auxiliary variables in double sampling. *Statistical Series*, 249, 1—11.

SAHOO, J. and SAHOO, L.N. (1993). A class of estimator in two phase sampling using two auxiliary variables. *Jour. Ind. Stat. Assoc*., 31, 107—114.

SINGH, H.P. (1993). A chain ratio-cum-difference estimator using two auxiliary variates in double sampling. *Jour. Ravi. Univ*., B, 6, 79—83.

SRIVASTAVA, S.K. (1967). An estimator using auxiliary information in sample surveys. *Cal. Stat. Ass. Bull*. 15, 127—134.

SRIVASTAVA, RANI. S., KHARE, B. B. and SRIVASTAVA, S.K.. (1990). A generalised chain ratio estimator for mean of finite population. *Jour. Ind. Soc. Ag. Stat*. 42, 108—117.

SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S. and ASOK, C. (1984). *Sampling Theory of Surveys with Applications*. Indian Society of Agricultural Statistics, New Delhi.

UPADHYAYA, L..N., KUSHWAHA, K. S. and SINGH, H. P. (1990). A modified chain-type estimator in two-phase sampling using multi-auxiliary information. *Metron*, XLVIII, 381—393.

UPADHYAYA, L..N., DUBEY, S. P. and SINGH, H. P. (1992). A class of ratio-in-regression estimators using two auxiliary variables in double sampling. *Jour. Sci. Res*., 42, 127—134.

WALSH, J. B. (1970). Generalization of ratio estimate for population total. *Sankhya*. Ser. A. 32, 99—106.

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 555—560

USE OF KNOWN CORRELATION COEFFICIENT IN ESTIMATING THE FINITE POPULATION MEAN

Housila P. Singh and Rajesh Tailor[1]

ABSTRACT

This paper deals with the problem of estimating the finite population mean using known correlation coefficient. An estimator is suggested with its approximate bias and mean squared error. The regions of preference are obtained under which it is better than usual ratio estimator and sample mean. An empirical study is carried out to demonstrate the performance of the constructed estimator over usual ratio estimator and sample mean.

***Key words*:** Study variate, Auxiliary variate, Correlation coefficient, Finite Population mean, Bias, Mean squared error.

1. Introduction

Let $U=(U_1,U_2,...,U_N)$ be a finite population of N (given) units. Let y and x denote the study and the auxiliary variate taking real non-negative values y_i and x_i respectively on the unit $U_i (i=1,2,....,N)$. This non-negativity condition is met by almost all survey universes .Let y and x be positively correlated .Suppose that a simple random sample of n (<N) units, drawn without replacement, from the population U. Let $(\overline{X},\overline{Y})$ be the population means and $(\bar{x},\bar{y})$ sample means for respective variates. Then the commonly used ratio estimator for $\overline{Y}$ is defined by

$$\bar{y}_R=\bar{y}\left(\frac{\overline{X}}{\bar{x}}\right) \tag{1}$$

Here it is assumed that the population mean $\overline{X}$ of the auxiliary variate is known.

To the first degree of approximation, the bias and mean squared error(MSE) of $\bar{y}_R$ are respectively given by

[1] School of Studies in Statistics, Vikram University, Ujjain — 456 010, M.P., INDIA.

$$B(\bar{y}_R) = \left(\frac{1-f}{n}\right)\bar{Y}C_x^2(1-K) \quad , \tag{2}$$

$$MSE(\bar{y}_R) = \left(\frac{1-f}{n}\right)\bar{Y}^2\left[C_y^2 + C_x^2(1-2K)\right] \ , \tag{3}$$

where $C_y^2 = \frac{S_y^2}{\bar{Y}^2}, \quad C_x^2 = \frac{S_x^2}{\bar{X}^2}, \quad K = \rho\frac{C_y}{C_x}, \quad \rho = \frac{S_{xy}}{(S_x S_y)}$,

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

$S_y^2 = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ and f = n / N.

It is to be mentioned that the past association with experimental material might provide quite accurate value of the correlation coefficient (ρ) between y and x, for instance, see Sahai and Sahai (1995). On the other hand, in case y and x are same variable on two previous occasions ρ may be easily obtained, see, Singh and Singh (1984). Other relevant references in this context are Mohanty and Das (1971), Kulkarni (1978), Sisodia and Dwivedi (1981) and Upadhyaya and Singh (1999).

In this paper, we have suggested a ratio — type estimator exploiting the knowledge of correlation coefficient ρ and its properties are discussed.

2. Suggested Estimator

Let $z_i = x_i + \rho$ (i=1,2,…,N) so that $\bar{z} = \bar{x} + \rho$ is the sample mean of the transformed variate z and $\bar{Z} = \bar{X} + \rho$ is the corresponding population mean. We suggest a transformed ratio estimator for the population mean $\bar{Y}$ as

$$\hat{\bar{Y}} = \bar{y}(\bar{Z}/\bar{z})$$

$$= \bar{y}\left(\frac{\bar{X}+\rho}{\bar{x}+\rho}\right) \tag{4}$$

To obtain the bias and MSE of $\hat{\bar{Y}}$, we write

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

$\bar{x} = \bar{X}(1+e_1)$

such that

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

and $$E(e_0^2) = \frac{(1-f)}{n} C_y^2 \quad ,$$

$$E(e_1^2) = \frac{(1-f)}{n} C_x^2 \quad ,$$

$$E(e_0 e_1) = \frac{(1-f)}{n} \rho C_x C_y$$

$$= \frac{(1-f)}{n} K C_x^2 .$$

Expressing (4) in terms of e's , we have

$$\hat{\bar{Y}} = \bar{Y}(1+e_0)(1+\theta e_1)^{-1},$$

where $\theta = \dfrac{\bar{X}}{\bar{X}+\rho}$.

We assume that $|\theta e_1| < 1$ so that $(1+\theta e_1)^{-1}$ is expandable. Thus ,to the first degree of approximation , the bias and MSE of $\hat{\bar{Y}}$ are respectively, given by

$$B(\hat{\bar{Y}}) = \left(\frac{1-f}{n}\right) \bar{Y} C_x^2 \theta(\theta - K) \ , \tag{5}$$

$$MSE(\hat{\bar{Y}}) = \left(\frac{1-f}{n}\right) \bar{Y}^2 \left[C_y^2 + \theta C_x^2 (\theta - 2K)\right]. \tag{6}$$

Remark: When the two variates y and x are negatively correlated, the product type estimator

$$\bar{y}_p^* = \bar{y}\left(\frac{\bar{x}+\rho}{\bar{X}+\rho}\right) \tag{7}$$

can be used in practice quite effectively.

3. Efficiency Comparisons

It is well known under simple random sampling without replacement (SRSWOR) that

$$V(\bar{y}) = \frac{(1-f)}{n} S_y^2$$

$$= \frac{(1-f)}{n} \bar{Y}^2 C_y^2 \tag{8}$$

We have from (6) and (8) that $MSE(\hat{\bar{Y}}) < V(\bar{y})$ if

$$\rho > \frac{1}{2} \frac{C_x}{C_y} \frac{\bar{X}}{(\bar{X} + \rho)} \tag{9}$$

Further we note from (3) and (6) that the proposed estimator $\hat{\bar{Y}}$ would be better than the usual ratio estimator $\bar{y}_R$ if

$$\rho < \frac{1}{2} \frac{C_x}{C_y} \left[1 + \frac{\bar{X}}{(\bar{X} + \rho)} \right] \tag{10}$$

Combining (9) and (10) we obtain that the proposed estimator $\hat{\bar{Y}}$ is more efficient than the usual unbiased estimator $\bar{y}$ and the usual ratio estimator $\bar{y}_R$ if

$$\frac{1}{2} \frac{C_x}{C_y} \frac{\bar{X}}{(\bar{X} + \rho)} < \rho < \frac{1}{2} \frac{C_x}{C_y} \left[1 + \frac{\bar{X}}{(\bar{X} + \rho)} \right] \tag{11}$$

4. Empirical Study

For the purpose of illustration, we have considered three natural populations data.

Population-I- Source: Cochran(1977).

y: Estimated production in bushels of peach,

x: The number of peach trees in an orchard,

$\bar{X} = 44.5000$, $\bar{Y} = 56.4700$, $C_x = 1.4045$,

$C_y = 1.4177$, $\rho = 0.8870$, $K = 0.8954$.

Population-II- Source: Singh et al (1976).

y: Number of milch cows as enumerated in winter season 1957—8,

x: Number of milch cows as per live stock census 1956,

$\overline{X} = 40.0600$, $\overline{Y} = 33.2900$, $C_x = 0.5185$,

$C_y = 0.4944$, $\rho = 0.7200$, $K = 0.6865$.

Population-III- Source: Singh et al(1994).

y: Treating T – C (Total Serum) cholesterol (in mg/df)

x: Treating LDL – (low density lipoprotein cholesterol) (in mg/df)

$\overline{X} = 139.4700$ mg/df, $\overline{Y} = 210.9600$ mg/df, $C_x = 0.3886$,

$C_y = 0.2668$, $\rho = 0.9306$, $K = 0.6389$.

We have computed the percent relative efficiencies of $\overline{y}_R$ and $\hat{\overline{Y}}$ with respect to usual unbiased estimator $\overline{y}$ and displayed in Table 1.

Table 1. ercent relative efficiencies (PREs) of $\overline{y}_R$ and $\hat{\overline{Y}}$ with respect to $\overline{y}$.

Estimator	PREs(., $\overline{y}$)		
	Poulation		
	I	II	III
$\overline{y}$	100.000	100.000	100.000
$\overline{y}_R$	446.46	169.59	243.56
$\hat{\overline{Y}}$	453.84	173.08	249.68

Table 1 clearly indicates that the proposed estimator $\hat{\overline{X}}$ is more efficient than the usual unbiased estimator $\overline{y}$ and the ratio estimator $\overline{y}_R$. Thus the proposed estimator $\hat{\overline{Y}}$ is recommended for its use in practice.

Acknowledgement

Authors are grateful to the referee and the editor Professor J. Kordos for their valuable suggestions which improved the paper very much.

REFERENCES

COCHRAN ,W.G. (1977). Sampling Technique. 3rd edition,Wiley and Sons. New yark.

KULKARNI, S. P. (1978). A note on modified ratio estimator using transformation. Journal of Indian Society of Agricultural Statististics, 30(2), 125—128.

MOHANTY, S. and DAS, M.N. (1971). Use of transformation in sampling. Journal of Indian Society of Agricultural Statististics,23, 83—87.

SAHAI, A. and SAHAI, A. (1985). On efficient use of auxiliary information. Journal of Statistical Planning and Inference., 12, 203—212.

SINGH, D., SINGH,P. and KUMAR P. (1976). Hand Book on Sampling Methods. Indian Agricultural Statistics Research Institute, New Delhi, p.139.

SINGH, R. K. and SINGH G. (1984). A class of estimators for population variance using information on two auxiliary variates. Aligarh Journal of Statistics, 3&4, 43—49.

SINGH, V.K., SINGH, HARI P. and SINGH, HOUSILA P. (1994). Estimation of ratio and product of two finite population mean in two — phase sampling. Journal of Statistical Planning and Inference,41, 163—171.

SISODIA, B.V.S. and DWIVEDI, V.K. (1981). A modified ratio estimator using coefficient of variation of auxiliary variable. Journal of Indian Society of Agricultural Statistics, 33, 13—18.

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

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 561—570

AN ALTERNATIVE APPROACH TO THE ESTIMATION OF RATIO IN TWO — PHASE SAMPLING

L. N. Sahoo[1], R. K. Sahoo[1], G. N. Singh[2] and L. N. Upadhyaya[2]

ABSTRACT

Using two-phase sampling mechanism, estimation of ratio of two unknown population means is considered when population mean of main auxiliary variable x is unknown, but that of an additional auxiliary variable z is known. Developing a new concept, we focus attention on the creation of some new estimators as well as a general class of estimators for our purpose.

Key words: Auxiliary variable, bias, mean square error, two-phase sampling.

1. Introduction

Assume that the means $\overline{Y}$ and $\overline{U}$ of the variables y and u for a finite population Ω of N units are unknown and the parametric function that we want to estimate is the ratio of these means i.e. $R = \overline{Y} / \overline{U}$. In recent past, considerable attentions have been given to achieve acceptable precision in estimation of R incorporating prior knowledge on the mean $\overline{X}$ of an auxiliary variable x, starting with the work of Singh (1965, 1967). But, in absence of such an essential information, one seeks to estimate R, from a sample obtained through a two-phase selection. Allowing simple random sampling (WOR) in each phase, the two-phase sampling mechanism can be described as follows:

a. A first phase sample s' ($s' \subset \Omega$) of n' units is drawn from Ω to observe x only.
b. Given s', a second phase sample s ($s \subset s'$) of n units is drawn from s' to observe y only.

[1] Department of Statistics, Utkal University, Bhubaneswar-751 004, India.
[2] Department of Applied Mathematics, Indian School of Mines, Dhanbad — 826 004, India.E. mail: gnsingh_ism@yahoo.com

Two useful estimators in this context are $r_1 = r\frac{\bar{x}'}{\bar{x}}$ and $r_2 = r - b_1(\bar{x} - \bar{x}')$, where $r = \bar{y}/\bar{u}$, the conventional estimator of R; $\bar{y}$, $\bar{u}$, $\bar{x}$ are the sample means based on s, $\bar{x}'$ is the sample mean based on s′, b_1 is the consistent estimate of the composite parameter $B_1 = \frac{R}{\bar{X}}(\rho_{yx}\frac{C_y}{C_x} - \rho_{ux}\frac{C_u}{C_x})$ based on s, where ρ_{yx}, ρ_{ux} are the correlation coefficients and C_y, C_u, C_x are the respective coefficients of variation. However, a class of estimators covering r_1 and r_2 as its special cases can be defined as $r_h = h\left(r, \bar{x}, \bar{x}'\right)$, where h(., ., .) is a known function of r, $\bar{x}$ and $\bar{x}'$ satisfying certain regularity conditions. The asymptotic minimum mean square error (MSE) of the class is given by

$$\min M(r_h) = M(r) - (\phi - \phi')\, R^2\, C_x^2\, \Delta_1^2 \qquad (1.1)$$

where $\phi = \left(\frac{1}{n} - \frac{1}{N}\right)\phi' = \left(\frac{1}{n'} - \frac{1}{N}\right), M(r) = \phi R^2(C_y^2 - 2\rho_{yu}C_yC_u + C_u^2)$ is the approximate MSE of r and $\Delta_1 = B_1\bar{X}/R$. The estimator r_2 attains this minimum MSE.

In many situations even if $\bar{X}$ is unknown, the population mean $\bar{Z}$ of an additional auxiliary variable z closely related to x is known accurately. For example, suppose Ω is a population of villages in a district such that x and z equal to the cultivated area and geographical area of the village, respectively. To estimate yield rate of a crop, the value of $\bar{X}$ i.e. an average area per village under cultivation may not be available, but information on $\bar{Z}$ i.e. an average geographical area per village can be readily available from the district records.

In the above scenarios, the data available on s′ can be used to furnish a good estimate of $\bar{X}$ treating z as an auxiliary variable. As argued by Chand (1975) and Kiregyera (1980, 1984), substitution of such an estimate of $\bar{X}$ in place of $\bar{x}'$ will further increase precision of estimation under certain specific conditions. For example, when $\bar{x}'$ in r_1 is replaced by the ratio estimator $\frac{\bar{x}'}{\bar{z}'}\bar{Z}$, Srivastava et al (1989) proved that the resulting estimator $r_{11} = r\frac{\bar{x}'}{\bar{x}}\frac{\bar{Z}}{\bar{z}'}$ is more efficient than r_1 if

$$\Delta_2 = \left(\rho_{yz} \frac{C_y}{C_z} - \rho_{uz} \frac{C_u}{C_z} \right) > \frac{1}{2},$$

where $\bar{z}'$ is the sample mean of z based on $s'; \rho_{yz}, \rho_{uz}$ are the correlation coefficients between y, z and u, z; C_z is the coefficient of variation of z. Replacing $\bar{x}'$ in r_2 by a regression estimator $\bar{x}' - b'_{xz}(\bar{z}' - \bar{Z})$, where b'_{xz} is the sample regression coefficient of x on z based on s′, we see that the resulting estimator $r_{22} = r - b_1 \left[\bar{x} - \left\{ \bar{x}' - b'_{xz} \left(\bar{z}' - \bar{Z} \right) \right\} \right]$ is more efficient than r_2 if

$$\frac{\Delta_2}{\Delta_1} > \frac{1}{2} \rho_{xz} \frac{C_x}{C_z},$$

where ρ_{xz} is the correlation coefficient between x and z. In a similar manner one can create many estimators taking r_1 or r_2 as the base.

Srivastava et al (1988) proposed a generalized estimator

$$r_s = r \left(\frac{\bar{x}}{\bar{x}'} \right)^{\alpha_1} \left(\frac{\bar{z}'}{\bar{Z}} \right)^{\alpha_2},$$

α_1 and α_2 being suitably chosen constants. Singh et al (1994) developed a class of estimators defined by $r_p = p\left(r, \frac{\bar{x}}{\bar{x}'}, \frac{\bar{z}'}{\bar{Z}} \right)$, where p (., ., .) is a known function of $r, \frac{\bar{x}}{\bar{x}'}$ and $\frac{\bar{z}'}{\bar{Z}}$ satisfying certain regularity conditions. An analysis of the properties of r_s and r_p shows that both the classes (although r_s is a special case of r_p) possess the same asymptotic minimum MSE given by

$$\min M(r_s) = \min M(r_p) = M(r) - (\phi - \phi') R^2 \Delta_1^2 C_x^2 - \phi' R^2 \Delta_2^2 C_z^2 \qquad (1.2)$$

which is equal to the MSE of a regression — type estimator

$$r_{RG} = r - b_1 \left(\bar{x} - \bar{x}' \right) - b'_2 \left(\bar{z}' - \bar{Z} \right),$$

where b'_2 is the consistent estimate of $B_2 = \frac{R\Delta_2}{\bar{Z}}$ based on s′.

In this paper we consider an alternative approach to estimate R, with the same auxiliary information available, and also compose a class of estimators for this purpose.

2. The Alternative Approach

Motivated by Chand (1975) when an attempt has been made to create an estimator from r_1 or r_2, with replacement of $\bar{x}'$ by a better estimate of $\bar{X}$, one should think that $\bar{x}$ provides a less efficient estimate of $\bar{X}$ than $\bar{x}'$. Hence, one should hope for a better estimate of $\bar{X}$ than $\bar{x}$ using data on s, by taking advantage of correlation between x and z. Thus, there are many alternative ways to develop estimators on replacing $\bar{x}$ and $\bar{x}'$ simultaneously by some improved estimators of $\bar{X}$ taking z as an auxiliary variable. But, for simplicity we consider the creation of two estimators involving standard ratio and regression features only viz.

$$r_{11}^{*} = r\frac{\bar{x}'}{\bar{x}}\frac{\bar{z}}{\bar{z}'}\frac{\bar{Z}}{\bar{z}'}$$

replacing $\bar{x}$ and $\bar{x}'$ by $\frac{\bar{x}}{\bar{z}}\bar{z}'$ and $\frac{\bar{x}'}{\bar{z}'}\bar{Z}$ in r_1, and

$$r_{22}^{*} = r - b_1\left[\left\{\bar{x} - b_{xz}\left(\bar{z} - \bar{z}'\right)\right\} - \left\{\bar{x}' - b'_{xz}\left(\bar{z}' - \bar{Z}\right)\right\}\right]$$

replacing $\bar{x}$ and $\bar{x}'$ by $\bar{x} - b_{xz}\left(\bar{z} - \bar{z}'\right)$ and $\bar{x}' - b'_{xz}\left(\bar{z}' - \bar{Z}\right)$ in r_2, respectively, where b_{xz} the sample regression coefficient of x on z based on s.

Omitting details we now obtain the following useful results on the basis of the asymptotic MSE of different estimators:

(i) $M(r_{11}^{*}) < M(r_1)$ if $\frac{1}{2} < \Delta_2 < \frac{1}{2}\left(2\rho_{xz}\frac{C_x}{C_z} - 1\right)$ provided $\rho_{xz}\frac{C_x}{C_z} > 1$, and

$M(r_{11}^{*}) < M(r_{11})$ if $\Delta_2 < \frac{1}{2}\left(2\rho_{xz}\frac{C_x}{C_z} - 1\right)$.

Hence, when $\rho_{xz}\frac{C_x}{C_z} > 1,$, r_{11}^{*} is superior to both r_1 and r_{11} if

$$\frac{1}{2} < \Delta_2 < \frac{1}{2}\left(2\rho_{xz}\frac{C_x}{C_z} - 1\right). \tag{2.1}$$

(ii) $M(r_{22}^*) < M(r_2)$ if $\frac{\Delta_2}{\Delta_1} < \frac{1}{2}\rho_{xz}\frac{C_x}{C_z}$ and $\phi > 2\phi'$, and

$M(r_{22}^*) < M(r_{22})$ if $\frac{\Delta_2}{\Delta_1} < \frac{1}{2}\rho_{xz}\frac{C_x}{C_z}$.

Hence, r_{22}^* is superior to both r_2 and r_{22} even if r_{22} is inferior to r_2, when

$$\frac{\Delta_2}{\Delta_1} < \frac{1}{2}\rho_{xz}\frac{C_x}{C_z} \text{ and } \phi > 2\phi' \ . \tag{2.2}$$

Results in (2.1) and (2.2) clearly indicate that there are situations where our technique may be successfully employed for defining estimators with increased accuracy. These estimators are also simple to compute without any appreciable increase in cost as compared to the estimators developed in the line of Chand's approach. With this spirit we now develop a general class of estimators considering difference estimators $t_x = \bar{x} - d\left(\bar{z} - \bar{z}'\right)$ and $t_x' = \bar{x}' - d'\left(\bar{z}' - \bar{Z}\right)$ as alternatives to $\bar{x}$ and $\bar{x}'$ in r_h, where d and d′ are constants, in particular may be random variables converging in probability to D and D′ respectively with n, n′ $\to$ N. These difference estimators are easier to handle mathematically and also reduce to many well-known estimators when their coefficients are chosen in a right manner.

3. The Proposed Class of Estimators

Whatever be the samples s and s′ chosen, let (r, t_x, t_x') assumes values in a closed convex subspace R_3 of 3-dimensional real space containing the point $(R, \bar{X}, \bar{X})$. Suppose that $g(r, t_x, t_x')$ is a function of r, t_x and t_x' such that $g(R, \bar{X}, \bar{X}) = R$ and satisfying the following regularity conditions:

(i) $g(r, t_x, t_x')$ is continuous in R_3

(ii) The first and second order partial derivatives of $g(r, t_x, t_x')$ exist and are also continuous in R_3.

Then, the proposed general class of estimators of R is defined by

$t_g = g(r, t_x, t_x')$.

We note that r_h is a special case of t_g when $d = d' = 0$, and r_p is a special case of t_g when $d = \dfrac{\bar{x}\left(\bar{x}'/\bar{x} - 1\right)}{\bar{x}'\left(\bar{z}/\bar{z}' - \bar{x}/\bar{x}'\right)}$ and $d' = \dfrac{\bar{x}'/\bar{x} - \bar{z}'/\bar{Z}}{\bar{Z}\left(\bar{z}'/\bar{Z} - \bar{z}/\bar{Z}\right)}$. The estimators r_{11}^* and r_{22}^* considered in the earlier section and some generalized estimators like

$$t_{11} = r\frac{\bar{x}' - d'\left(\bar{z}' - \bar{Z}\right)}{\bar{x} - d\left(\bar{z} - \bar{z}'\right)}, \qquad t_{22} = r\text{-}$$

$$b_1\left[\left\{\bar{x} - d\left(\bar{z} - \bar{z}'\right)\right\} - \left\{\bar{x}' - d'\left(\bar{z}' - \bar{Z}\right)\right\}\right],$$

$$t_{33} = r\frac{\bar{x} - d\left(\bar{z} - \bar{z}'\right)}{\bar{x}' - d'\left(\bar{z}' - \bar{Z}\right)}$$ are particular cases of t_g.

We now analyse the properties of t_g in some depth by obtaining approximate expressions for its bias and mean square error. For this on expanding t_g around the point $(R, \bar{X}, \bar{X})$ by the first order Taylor's series, neglecting remainder term and noting that $g_{11} = 1$, $g_{12} = -g_{13}$, we get

$$t_g - R = (r - R) + g_{12}\left[(t_x - \bar{X}) - (t_x' - \bar{X})\right] \tag{3.1}$$

where g_{11}, g_{12} and g_{13} are the first order partial derivatives of $g(r, t_x, t_x')$ w.r.t. the corresponding arguments about $(R, \bar{X}, \bar{X})$. Using linear approximation $r - R \cong R\left(\delta\bar{y} - \delta\bar{u}\right)$, from (3.1) we have

$$t_g - R = R\left(\delta\bar{y} - \delta\bar{u}\right) + g_{12}\bar{X}\left[\delta\bar{x} - \delta\bar{x}' - \frac{\bar{Z}}{\bar{X}}\left(d\,\delta\bar{z} - d\,\delta\bar{z}' - d'\delta\bar{z}'\right)\right], \tag{3.2}$$

where $\delta\bar{y} = \dfrac{\bar{y} - \bar{Y}}{\bar{Y}}$, $\delta\bar{x}' = \dfrac{\bar{x}' - \bar{X}}{\bar{X}}$ etc.

If we continue to work with (3.2) in the usual way, after a considerable simplification we get,

$E(t_g) = R + O(n^{-1})$,

implying that the bias of t_g is of order n^{-1}, and the MSE of t_g to terms of order n^{-1} as

$$M(t_g) = M(r) + (\phi - \phi')\Big[\, g_{12}^2 \,\bar{X}^2 \left\{ C_x^2 - 2d\left(\frac{\bar{Z}}{\bar{X}}\right)\rho_{xz}\, C_x\, C_z + d^2\left(\frac{\bar{Z}}{\bar{X}}\right)^2 C_z^2 \right\}$$

$$+ 2g_{12}\,\bar{X}\, R\Delta_1\, C_x^2 - 2g_{12}\,\bar{X}\, d\left(\frac{\bar{Z}}{\bar{X}}\right) R\,\Delta_2\, C_z^2 \,\Big]$$

$$+ \phi'\Big[\, g_{12}^2\,\bar{X}^2\, d'^2\left(\frac{\bar{Z}}{\bar{X}}\right)^2 C_z^2 + 2g_{12}\,\bar{X}\, d'\left(\frac{\bar{Z}}{\bar{X}}\right) R\,\Delta_2\, C_z^2 \,\Big]. \qquad (3.3)$$

The MSE of tg in (3.3) is a function of g12, d and d′, and is minimized for

$$g_{12} = -\frac{\dfrac{R}{\bar{X}}\left(\Delta_1 - \rho_{xz}\dfrac{C_z}{C_x}\Delta_2\right)}{1-\rho_{xz}^2} = g_{12(opt)} \text{ (say)},\quad d = \left(\frac{\bar{X}}{\bar{Z}}\right)\frac{\rho_{xz}\dfrac{C_x}{C_z}\Delta_1 - \Delta_2}{\Delta_1 - \rho_{xz}\dfrac{C_z}{C_x}\Delta_2} = d_{opt} \text{ (say)}$$

$$\text{and } d' = \left(\frac{\bar{X}}{\bar{Z}}\right)\frac{\Delta_2\left(1-\rho_{xz}^2\right)}{\Delta_1 - \rho_{xz}\dfrac{C_z}{C_x}\Delta_2} = d'_{opt}\,(\text{say})$$

and the minimum MSE is given by

$$\min M(t_g) = M(r) - (\phi - \phi')\frac{R^2}{1-\rho_{xz}^2}\left[\Delta_1^2 C_x^2 - 2(\Delta_1 C_x)(\Delta_2 C_z)\rho_{xz} + \Delta_2^2 C_z^2\right] - \phi' R^2 \Delta_2^2 C_z^2 \qquad (3.4)$$

An estimator attaining this minimum MSE is a regression — type estimator of the form

$$r^* = r + g_{12(opt)}\left[\left\{\bar{x} - d_{opt}\left(\bar{z} - \bar{z}'\right)\right\} - \left\{\bar{x}' - d'_{opt}\left(\bar{z}' - \bar{Z}\right)\right\}\right].$$

It may be remarked here that the optimum values $g_{12(opt)}$, d_{opt} and d'_{opt} are usually unknown and thus need to be estimated by their consistent estimates computed from the sample data. However, the resulting estimator $\hat{r}^*$ (say) estimates R with certain amount of bias and to the first order of approximation $M(r^*) = M(\hat{r}^*)$.

As to comparison of the precision of t_g relative to r_h and r_p, we feel that one can not draw any meaningful conclusion by comparing all estimators belonging to two different classes. Since, an estimator has its own limitation and is suitable only for a particular situation in terms of the relationship between the variables

under consideration. However, for simplicity, if we accept minimum MSE as an intrinsic measure of precision of a class then from (1.1), (1.2) and (3.4) we see that

$$\min M(t_g) \leq \min M(r_h) \text{ or } \min M(r_p)$$

which implies that t_g represents a class of more precise estimators of R than that of r_h or r_p. This result also leads to a conclusion that the estimator $\hat{r}^*$ is asymptotically more efficient than r_2 and r_{RG}.

4. Numerical Illustrations

To study effectiveness of the suggested estimation technique numerically, we consider the data of 5 natural populations described in table 4.1. Relative efficiencies of various estimators, based on their asymptotic MSE expressions, are displayed in table 4.2.

Table 4.1. Description of Populations

Popu-lation. No.	Source	N	y	u	x	z
1.	Das (1982)	368 villages	No. of literate persons in 1951	No.of literate persons in 1961	No.of households in 1951	No.of households in 1961
2.	Fisher (1936)	50 Iris flowers (setosa)	Petal width	Sepal length	Petal length	Sepal width
3.	Fisher (1936)	50 Iris flowers (versicolor)	Sepal width	Petal length	Sepal length	Petal width
4.	Singh (1965)	61 blocks	Females employed	Females in services	Female population	Educated females
5.	Tripathi (1980)	225 households	Persons in services	Educated persons	Persons employed	Size of households.

Table 4.2. Relative efficiencies of different estimators w.r.t. r (in %)

Population	*Estimator*								
	R	r_1	r_{11}	r_{11}^*	r_2	r_{22}	r_{22}^*	r_{RG}	$\hat{r}^*$
1 (n′=100,n=40)	100.00	136.92	130.99	148.33	138.10	137.11	170.50	181.15	198.37
2 (n′=20,n=10)	100.00	105.84	105.14	108.34	105.95	106.24	115.20	118.55	127.31
3 (n′=20,n=10)	100.00	107.92	95.61	108.44	117.56	111.17	122.24	129.28	139.65
4 (n′=30,n=15)	100.00	96.58	81.48	99.86	101.09	98.85	103.28	109.10	121.90
5 (n′=90,n=30)	100.00	129.05	126.89	120.12	132.99	131.69	134.05	140.36	155.34

Findings of table 4.2 show that $\hat{r}*$ attains the maximum precision amongst all for the 5 populations whereas r_{RG} is the second best estimator as it is superior to the remaining estimators. An increase in precision of $\hat{r}*$ over r_{RG} also indicates that the class of estimators developed here i.e. t_g is definitely superior to r_s or r_p under minimum MSE criterion. For populations 1,2,3 and 4, inspite of poor performance of r_{11} over r_1, r_{11}^* shows to have better performance than both r_1 and r_{11}. In population 5, $\hat{r}_{11}^*$ is inferior to both r_1 and r_{11} because the favorable conditions in (2.1) do not fit the data. For all populations r_{22}^* is more efficient than its immediate competitors r_2 and r_{22} although r_{22} is less efficient than r_2 except for the second population. This numerical study (although has a limited scope) shows that even if the gain in efficiency of our suggested estimators over the conventional estimator r in some cases is either marginal or negligible, for some other cases the improvement is substantial. For example, in population 1 the gain in efficiency of $\hat{r}*$ over r is about 98%. However, the improvement depends on how well the favorable conditions for the suggested estimators fit to the data.

Acknowledgements

Authors are thankful to the referees for their valuable comments.

REFERENCES

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

DAS, A. K. (1982), Estimation of population ratio on two occasions. Jour. Ind. Soc. Agric. Stat., 34, 1—9.

FISHER, R. A. (1936), The use of multiple measurements in taxonomic problems. Ann. Eugenics, 7, 179—188.

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

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

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

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

SINGH, V. K., SINGH, Hari P., SINGH, HOUSILA P. and SHUKLA, D. (1994), A genral class of chain estimators for ratio and product of two means of a finite population. Comm. Stat. — Theo. Meth., 23, 1341—1355.

SRIVASTAVA, RANI S., KHARE, B. B. and SRIVASTAVA, S. R. (1988), On generalized chain estimator for ratio and product of two population means using auxiliary characters. Assam Stat. Rev., 2, 21—29.

SRIVASTAVA, RANI S., SRIVASTAVA, S. R. and KHARE, B. B. (1989), Chain ratio — type estimator for ratio of two population means using auxiliary characters. Comm. Stat. — Theo. Meth., 18, 3917—3926.

TRIPATHI, T. P. (1980), A general class of estimators for population ratio. Sankhya, C, 42, 63—75.

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 571—575

A NOTE ON THE ESTIMATION OF MEAN USING AUXILIARY INFORMATION

Lakshmi N. Upadhyaya[1] **and Housila P. Singh**[2]

ABSTRACT

In this paper we have suggested an estimator for the population mean $\overline{Y}$ of the study variate y by making use of auxiliary information on x and z, where x is positively correlated with y and z is negatively correlated with y. The bias and mean-squared error (MSE) of the proposed estimator has been obtained under the large sample approximations. The proposed estimator has been compared with the usual unbiased estimator $\overline{y}$ and that of Srivastava (1965). Here we confine ourselves to SRSWOR sampling scheme.

Key words: Auxiliary information, Mean-squared error, Large sample approximation, SRSWOR.

1. Introduction

Several authors including Ghosh (1947), Olkin (1958), Raj (1965), Srivastava (1965, 66), Rao and Adhvaryu (1975) have used multi-auxiliary-supplementary information in estimating the population mean $\overline{Y}$ and proposed a number of estimators and studied their properties.

Let the population consist of N units, y_i , x_i and z_i (i=1,2,…,n) denote the values of the i-th unit of the character y, x and z respectively. A simple random sample of size n is observed. Let $\overline{y}$, $\overline{x}$ and $\overline{z}$ denote the sample mean of y, x and z respectively. There are two well known estimates of the population mean $\overline{Y}$ which make use of ancillary information: one is the ratio estimate $\overline{y}\,(\overline{X}/\overline{x})$ where the two variables y and x have high positive correlation and the other is the product estimate $\overline{y}\,(\overline{z}/\overline{Z})$ where y and z have high negative correlation. It is assumed that the population mean $\overline{X}$ and $\overline{Z}$ of ancillary characters x and z are

[1] Department of Applied Mathematics, Indian School of Mines, Dhanbad-826 004. India.
[2] School of Studies in Statistics, Vikram University, Ujain-456 010. India.

known. In the above situation Srivastava (1965) has suggested the estimator of $\bar{Y}$ as:

$$\bar{y}_1 = \text{u.}\ \bar{y}\frac{\bar{X}}{\bar{x}} + \text{v.}\ \bar{y}\frac{\bar{z}}{\bar{Z}} \tag{1}$$

where u and v are the weights given to the two estimates (ratio and product) such that

u + v = 1.

Srivastava (1965) obtained the bias and MSE of the estimator $\bar{y}_1$.

2. Estimator and its properties

We propose as an estimator for the population mean $\bar{Y}$:

$$\bar{y}_2 = \bar{y}\left[u.\frac{\bar{X}}{\bar{x}} + \text{v}\frac{\bar{z}}{\bar{Z}}\right]^k \tag{2}$$

where u and v are the weights such that u + v = 1, and k is a constant to be chosen suitably.

The sample size n is chosen large enough so that we can assume that

$$\left|\frac{\bar{x}-\bar{X}}{\bar{X}}\right| < 1 \text{ and } \left|\frac{\bar{z}-\bar{Z}}{\bar{Z}}\right| < 1. \tag{3}$$

The bias and MSE of $\bar{y}_2$, to the terms of order $0(n^{-1})$, are given by

$$\text{B}(\bar{y}_2) = \left(\frac{\bar{Y}}{n}\right)\left[\, k(k-1)\left\{\text{u}^2 C_x^2 + \text{v}^2 C_z^2 - 2\,\text{u v}\,\rho_{xz} C_x C_z\right\}/2\right.$$

$$\left. - k\left\{(\rho_{yx} C_y C_x - C_x^2)\,\text{u} - \text{v}\,\rho_{yz} C_y C_z\right\}\,\right] \tag{4}$$

$$\text{M}(\bar{y}_2) = \left(\frac{\bar{Y}^2}{n}\right)\left[k^2\left\{\text{u}^2 C_x^2 + \text{v}^2 C_z^2 - 2\,\text{u v}\rho_{xz} C_x C_z\right\} - 2\right.$$

$$\left. k\left\{\text{u}\,\rho_{yx} C_y C_x - \text{v}\,\rho_{yz} C_y C_z\right\}\right] \tag{5}$$

where $C_x = \frac{\sigma_x}{\bar{X}}$, $C_y = \frac{\sigma_y}{\bar{Y}}$, $C_z = \frac{\sigma_z}{\bar{Z}}$,

$$\rho_{yx} = \frac{\text{Cov}(y,x)}{\sigma_y\,\sigma_x},\ \rho_{yz} = \frac{\text{Cov}(y,z)}{\sigma_y\,\sigma_z},\ \rho_{xz} = \frac{\text{Cov}(x,z)}{\sigma_x\,\sigma_z}.$$

Minimizing (5) subject to the condition that u + v = 1, we obtain the following normal equations

$$k = \frac{\mathrm{u\,S} - \rho_{yz}\, C_y\, C_z}{\mathrm{u}^2\,\mathrm{M} - 2\,\mathrm{u\,A} + C_z^2},$$

and $\quad \mathrm{u} = \dfrac{\mathrm{S} + \mathrm{A}\,k}{\mathrm{M}\,k} = 1 - \mathrm{v},$

where

$\mathrm{S} = \rho_{yx}\, C_y\, C_x + \rho_{yz}\, C_y\, C_z$,

$\mathrm{M} = C_x^2 + C_z^2 + 2\rho_{xz}\, C_x\, C_z$,

and $\quad \mathrm{A} = C_z^2 + \rho_{xz}\, C_x\, C_z$.

For the sake of simplicity, we may assume that

$C_x = C_z = C$

$\rho_{yx} = -\ \rho_{yz}$

so that

$$\mathrm{u} = \mathrm{v} = \frac{1}{2},$$

and

$$k = \frac{2\,\rho_{yx}\, C_y}{(1 - \rho_{xz})\, C}.$$

From (4), (5), (8) and (9), we obtain the minimum bias and MSE of $\bar{y}_2$ as

$$\mathrm{B}_0(\bar{y}_2) = \frac{\bar{Y}\,\rho_{yx}\, C_y}{2\,n\,(1 - \rho_{xz})}\left[(1 + \rho_{xz})\, C - 2\,\rho_{yx}\, C_y\right] \qquad (10)$$

$$\mathrm{M}_0(\bar{y}_2) = \left(\frac{\bar{Y}^2}{n}\right)\left[C_y^2 - \frac{4\,\rho_{yz}^2\, C_y^2}{(1 - \rho_{xz})^2\, C^2}\right]. \qquad (11)$$

The estimator $\bar{y}_2$ is unbiased when $C = \dfrac{2\,\rho_{yx}\, C_y}{(1 + \rho_{xz})}$.

It may be noted that when $\mathrm{u} = \mathrm{v} = \dfrac{1}{2}$, Srivastava (1965) obtained the minimum MSE of the estimator $\bar{y}_1$ as

$$\mathrm{M}_0(\bar{y}_1) = \left(\frac{\bar{Y}^2}{n}\right)\left[C_y^2 + \frac{\mathrm{C}^2}{2}(1 - \rho_{xz}) - 2\,\rho_{yx}\, C\, C_y\right] \qquad (12)$$

3. Comparison

The variance of the usual unbiased estimator $\bar{y}$ of the population mean $\bar{Y}$ is given by

$$\text{Var}(\bar{y}) = \frac{\bar{Y}^2 C_y^2}{n}. \tag{13}$$

From (11) and (13), we find

$$\text{Var}(\bar{y}) - \text{M}_0(\bar{y}_2) = \left(\frac{\bar{Y}^2}{n}\right)\left[\frac{4\rho_{yz}^2 C_y^2}{(1-\rho_{xz})^2 C^2}\right] \geq 0 \tag{14}$$

and from (11) and (12), we get

$$\text{M}_0(\bar{y}_1) - \text{M}_0(\bar{y}_2) = \frac{\bar{Y}^2}{2n(1-\rho_{xz})}\left[\{C(1-\rho_{xz}) - 2\rho_{yz} C_y\}^2\right] \geq 0 \tag{15}$$

Hence, the estimator $\bar{y}_2$ is more efficient than the usual unbiased estimator $\bar{y}$ and the estimator proposed by Srivastava (1965).

One can generalize the above estimator to the case of (m+n) auxiliary variables and can propose the following estimator for the population mean $\bar{Y}$ as

$$\bar{Y}_3 = \bar{y}\left[\sum_{i=1}^{m} u_i O_i + \sum_{j=1}^{n} v_j Q_j\right]^k \tag{16}$$

where k is a constant to be chosen suitably, $O_i = \frac{\bar{X}_i}{\bar{x}_i}$ (i=1,2,…, m), $Q_j = \frac{\bar{z}_j}{\bar{Z}_j}$ (j=1,2,…,n), $x_1, x_2, ..., x_m$ are the m auxiliary variables positively correlated with y and $z_1, z_2, ..., z_n$ are the n auxiliary variables negatively correlated with y. Further, it is assumed that the population means $\bar{X}_1$, $\bar{X}_2$, …, $\bar{X}_m$, $\bar{Z}_1$, $\bar{Z}_2$, …, $\bar{Z}_n$ are known.

REFERENCES

ADHVARYU, A., (1975): Combination of some estimators using supplementary information. Journal of Indian Society of Agricultural Statistics, Vol. 27, No. 2, 37—50.

GHOSH, B. (1958): Double sampling with many auxiliary variates. Calcutta Statistical Association Bulletin, 1, 91—93.

OLKIN, I. (1958): Multivariate ratio estimation for finite population. Biometrika, 45, 154—165.

RAJ, DES (1965): On a method of using multi-auxiliary information in sample surveys. Journal of American Statistical Association, 60, 270—277.

RAO, P.S.R.S. and MUDHOLKAR, G.S. (1967): Generalized multivariate estimator for the mean of finite population. Journal of American Statistical Association, 62, 1009—1012.

REDDY, V.N. (1974): On a transformed ratio method of estimation. Sankhya, Series B, 35, 59—70.

SRIVASTAVA, S.K. (1965): An estimate of the mean of a finite population using several auxiliary variables. Journal of Indian Statistical Association, Vol. 3, Nos. 2 & 3.

SRIVASTAVA, S.K. (1966): On ratio and linear regression methods of estimation with several auxiliary variables. Journal of Indian Statistical Association, Vol. 4, 66—72.

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 577—587

TENDENCIES IN RESTRUCTURING THE AGRICULTURAL INFORMATION SYSTEM IN POLAND

Krzysztof Mateńko[1]

ABSTRACT

The system of agricultural information in Poland is created by institutions gathering information on agriculture by means of statistical surveys and censuses, as well as by institutions keeping various databases and administrative registers. Agricultural surveys and censuses included in annual and multi-year programmes of official statistics are conducted mostly by the CSO and the Ministry of Agriculture and Rural Development.

Changes taking place in the 90-ties aimed at adjusting agricultural statistics, as well as, the whole of Polish official statistics to international standards, including requirements of the European Union. The building of statistics which would be coherent with the EU requirements shows the necessity to restructure the whole information system concerning agriculture. As a result of the talks conducted between Eurostat and Polish bodies there was created an Inter-sectoral Experts Team on Restructuring of the Agricultural Information System.

Key words: Agricultural information system, Census of agriculture, Agricultural sample survey, Official statistics, European Union.

1. Introduction

The system of agricultural information in Poland consists of several components, such as censuses of agriculture, agricultural sample surveys, registers and various databases. Required data are collected mostly by Central Statistical Office (CSO) with co-operation with the Ministry of Agriculture and Rural Development. Annual and multi-year programmes of statistical surveys are prepared by the Central Statistical Office (CSO) together with ministries, central government bodies, and local units of state and self-government administration.

[1] Central Statistical Office, Poland; E-mail: k.matenko@stat.gov.pl

Agricultural surveys and censuses included in annual and multi-year programmes of official statistics are conducted mostly by the CSO and the Ministry of Agriculture and Rural Development.

2. Transition Period

In the period preceding the transformation the following activities took place in the Polish agricultural statistics:

- annual national agricultural censuses up to the year 1988,
- sample surveys on individual agricultural holdings,
- full reporting from socialised units,
- expert estimates and monthly reports of voluntary agricultural and horticultural correspondents.

Changes taking place in the 90-ties aimed at adjusting agricultural statistics, as well as, the whole of Polish official statistics to international standards, including requirements of the European Union. The period was also characterised by reducing the reporting, and giving up annual national agricultural censuses. After the 1988 National Agricultural Census the Government decided that national agricultural censuses would be conducted every 5 years, and that, so called, June sample surveys (for the sample of 5%) on private agricultural holdings in the scope of land use, sown area and livestock, would be conducted in periods between the censuses (Kordos, Kursa, 1997; Wanke, 1997; Wanke, Zalewski, 1998). Due to some financial limitation the sample was reduced to 3%, and the 5-year period between the censuses were extended. A survey on agricultural holdings with the sample of about 10% was conducted in 1994, and full agricultural censuses were conducted in 1996 and 2002 (GUS, 2003; Wanke, 2003).

In the 90-ties the most important and biggest undertaking of the Polish agricultural statistics was conducting the National Agricultural Census in 1996 according to the FAO recommendations concerning 2000 the World Agricultural Census (FAO, 1995ab). The Census, to a great extent, was in compliance with the recommendations and requirements of the European Union. Apart from obtaining comprehensive information on actual conditions of agricultural holdings the census allowed to update the information base for surveys and statistical elaboration which was carried out after the Census. They constituted a new basis for estimates and updated sampling frames.

3. Harmonisation to EU requirements

The processes aiming at harmonisation to the European Union requirements caused a necessity for a considerable development of resources of statistical information on agriculture, as well as, for a change in the conception of statistical

surveys. The building of statistics which would be coherent with the EU requirements shows the necessity to restructure the whole information system concerning agriculture.

Economic aspects connected with obtaining information, its elaboration and dissemination also are becoming more and more important. In this context, it is very important to develop the co-operation between different organisations which participate in obtaining information on agriculture — in order to make the costs as small as possible through elimination of doubling the activities, or through ensuring the coherence of the used definitions and classifications. The close co-operation between the Central Statistical Office and the Ministry of Agriculture and Rural Development, (i.e. institutions leading in the official statistics and in the management of administrative data in this scope) play the biggest role in those activities.

As a result of the talks conducted between Eurostat, the Ministry of Agriculture and Rural Development, as well as, the Central Statistical Office there was created an *Inter-sectoral Experts Team on Restructuring of the Agricultural Information System*. To the basic tasks of the Team at the first stage of their activities belonged: cataloguing of the surveyed characteristic features in the system of agricultural information and determining the degree of adjustment of particular features to the requirements of the European Union. At a late stage, there was prepared a list of information gaps from the point of view of both national and European Union needs. Analyses were conducted in order to eliminate the doubling of the surveyed features.

Further work of the Team concentrates on determining the directions of activities undertaken for increasing the quality of the conducted surveys (both typically statistical and other specialist surveys) from the point of view of complete coherence of definitions, as well as, specialist subject-matter assessment and accuracy in the assessment at national and regional levels.The subsequent task is to indicate features which could be obtained from research and administrative systems created for the needs of the European Union which are to replace the hitherto existing administrative sources. It is also very important to determine a clear division of tasks between institutions which create the agricultural information system, considering the leading role of the Central Statistical Office in case of the system of statistical information.

4. Tasks of different institutions

The following lists have been elaborated within already conducted activities:

- a list of features covered by official statistics surveys conducted within the scope of the Programme of Surveys on Official Statistics which is implemented each year under the Regulation of the Prime Minister,
- a list of information obtained from the systems of market information kept by the Ministry of Agriculture and Rural Development,

- a list of information obtained and necessary to obtain by the Agricultural Market Agency,
- a list of features covered in newly created, by the Agency of Restructurization and Modernisation of Agriculture, administrative systems (IACS) for land utilisation, number of plots of land, and livestock number,
- a list of information covered by currently created FADN system.

The information on level of compliance to the EU requirements is included in elaborated lists together with the indication of particular UE legal act.

Lists of features included in surveys of official statistics contained in annual Programme of Statistical Surveys of Official Statistics concern following problems:

- number of agricultural holdings,
- type of user of agricultural holding,
- land area according to type of utilisation,
- area under forest and uncultivated area,
- sown and cultivated area of particular agricultural products,
- orchards area,
- crop area under particular fruit trees,
- crop area under particular fruit shrubs,
- yielding of particular crops,
- production of particular agricultural products,
- balances settlements of particular vegetable products (production, processing, external trade, stock, losses),
- livestock size of particular agricultural animals according to usability categories,
- production efficiency of particular agricultural animals,
- general number of animal production,
- number of animals destined to slaughter,
- hatching,
- covering of animals,
- balances of particular animal products (production, processing, external trade, stock, losses),
- procurement's size of particular agricultural products, animals and animal products,
- procurement prices and prices in marketplace,
- full set of structural features allowing to characterise agricultural holdings,
- means of agricultural production,
- agricultural economic accounts,
- agricultural forecasts etc.

Within the activities of official statistics the MARD (Ministry of Agriculture and Rural Development) conducts so called departmental surveys. These are

surveys which require specialist knowledge and contain sometimes administrative type of data as well. Their range is broad, various and concerns:

- realisation of investments in the scope of road building in the communal (gminas) rural areas and poviats,
- production and sale of dried fodder,
- veterinary activities,
- fishing economy,
- plant protection,
- livestock and animal production,
- ground utilisation — realisation of geodesic-installation work for agricultural purposes and realisation of uniting work,
- some economic aspects of environment protection,
- realisation of small storage reservoirs,
- resources and land use amendments, hazard and protection,
- agricultural and food products prices.

Administrative type of data are also gathered in The Ministry of Agriculture and Rural Development, including registers of entities, which are established on the basis of acts concerning particular areas of activity in agriculture and feeding economy.

Currently following registers are created in MARD:

- register of decisions on production, purification, contamination or dewatering, as well as, of ethyl alcohol and production of spirits,
- register of decisions on production or filling of the wine products,
- register of decisions on production of tobacco products.

Moreover in institutions subordinate to the Ministry of Agriculture and Rural Development various registers and databases are created.

6. The Agriculture and Food Quality Inspection

The Inspection keeps the following registers:

- register of entities reporting the carrying of agricultural and food economic activity,
- register of hops producers,
- register of processing plants designated for initial processing of the raw tobacco (including information on the size of procurement and the processed tobacco),
- register of ecological farms,
- register of ecological processing,
- register of units giving certificates declaring that a given farm, plant, etc. is conducted in compliance with the rules of ecological agriculture.

Besides, the inspection possesses the data on amount of import of agricultural and food products being subject to quality inspection (174 commodity

records with defined minimal level of import) and data on certificates concerning hops growing (detailed data are in possession of voivodships inspections, soon at voivodship level data on hop contracts will be registered).

The Main Inspection — Veterinary Inspection keeps the following data bases and registers:

- database of processing plants subject to veterinary control (meat, milk and fish sectors plants),
- register of entities undertaking or conducting activities relating to transport of livestock, animal production and animal feeding stuffs,
- register of processing plants that process animal by-products into bone meal and meat-and-bone meal (a separate register of plants producing meal for burning),
- database on production of feeds for ruminants on farms,
- database on imports of animals, products of animal origin and animal feeding stuffs,
- database on establishments certified for exporting goods into the European Union markets.

The considerable collection of information can be expected from databases under construction developed by The Agricultural Market Agency (AMA), and The Agency for Restructuring and Modernisation of Agriculture (ARMA) within the framework of the system IACS. Also work of research institutes, particularly The Institute of Agricultural and Food Economy (IAFE) is the valuable source of information. The importance of the system of agricultural holdings accountancy, established by IAFE, so called the Polish FADN is increasing permanently.

The Polish official statistics is fully harmonised with the EU requirements in the range of surveys on production of basic land crop and animal production.

7. Further development

The agricultural census conducted in 2002 allowed to extend and update the register of agricultural holdings, as well as, multi-purpose sample frame, which increased the surveys precision. Moreover, according to the European Union Regulation, a list of minor importance crops in Poland will be created.

Analysis of the system of plant production surveys indicates necessity for changes in the schedule of sample surveys and need for reducing the number of experts and voluntary agricultural correspondents, and simultaneously, increasing the role of sample surveys, also according to the European Union requirements as well.

In order to fully harmonise the horticulture statistics with the European Union requirements, it is planned, in the orchard sample survey (conducted every 5 years — according to the EU requirements) to include together with apple-tree variation survey, aubergine variation survey, and also to study the scale of

plantings and grubbing fruit trees and shrubs, by species. Taking into account the fact that peaches are cultivated in Poland on a small scale, contrary to the EU regulations, the survey on that species is not foreseen in Poland.

Some of the European Union Member States conduct vineyard area and yields statistics. It is only unprofessional grape growing in Poland; currently no surveys are conducted on production potential and production of that plant. In the 2002 Agricultural Census it was estimated, that the vineyard area in Poland covers at least 155 hectares total.

Basic Eurostat's requirements in the scope of animal production concern obligatory surveys of cattle stock, pig, sheep and goats. Those requirements do not determine methodological details, each Member State is free to use its own methods, but they contain basic rules, which must be accomplished.

The EU Member States are obliged to conduct particular surveys and to transmit results within the specified time limits. Sample surveys can be and should be completed by estimates. Those surveys might be partly, after consent of the Commission, substituted by administrative data (i. e. information from IACS).

Particular Member States might be entitled to limit number of surveys in case of small livestock population, or limit particular surveys only regarding to particular regions.

The survey should be conducted in a way, which would ensure grouping animals according to homogenous productively — utilising category, and presented results should be characterised by determined precision.

Population of pig, cattle, sheep and goats should be surveyed, in a way that would allow to determine their regional differentiation, at least once a year in the survey conducted in November or December. And that survey — at least every second year — should allow the presentation of the domestic livestock according to herd size. Certain Member States can -when agreed by the Commission- determine different reference deadline.

In Poland surveys on the livestock population are, since 1988, conducted according to the EU requirements. Using the results of the Agricultural Census 1996, frames to drawing the sample were actualised and developed, the frequency of the survey was changed from quarterly to trisemestral (pig population) or half-yearly (other livestock) and homogeneous productively - utilising categories were introduced.

The Polish official statistic is totally harmonised with the Eurostat requirements in the scope of survey and information about the size of livestock and herd structure of the basic livestock species: pig, cattle, sheep and goats.

It can be stated that statistical surveys of agricultural holdings, at an angle of conducted there animal production, require only current improving activities.

The subjects connected with slaughtering statistics, production prognosis and eggs for hatching and milk production statistics require further adjustment tasks in Poland. The tasks connected with the adjustment of the definition of milk products require special care. The tasks under compilation of balances of certain

agricultural products, including animal products, need also to be continued. Those aspects are also important because of the possibility of compound estimation both animal production and the market.

In the previous years agricultural censuses and June representative surveys were ones, which partly suited the surveys of agricultural holdings structure. Before 1996 they provided the data about the structure of agricultural holdings. They provided mostly the information on calculation of the agricultural production and conducting the analysis for the domestic needs.

Preparing the agricultural census 1996 the FAO recommendations included in the document „2000Agricultural Census" and some of the Eurostat requirements were taken into account. The co-operation with The Institute of Agricultural and Food Economy in the scope of typology of agricultural holdings was commenced. Thanks to that it became possible to transmit to Eurostat, for the first time, greater part of the data collection concerning the structure of agricultural holdings in Poland.

Further methodological work towards full adjustment of structure survey to the EU requirements was continued. These problems are also one of the most important elements of bilateral co-operation between Polish and German statisticians.

In June 2000 sample survey on land utilisation, sown area and livestock population was conducted. All possible to examine, according to the binding law, features required by EU were included. Unfortunately because of the financial restrictions it was necessary to diminish the sample from scheduled 10% to 3%. Such a sample size allowed generalising the results only for the country and voivodships.

The next stage of the tasks improving the surveys of the structure of agricultural holdings in Poland was to prepare and conduct in 2002 basic agricultural holding survey — general census. The census provided the whole range of information about the structure of agricultural holdings. This information will be available also in tabularised formulation compatible with the EUROFARM database.

The census will also allow — by using the SGM, calculated by The Institute of Agricultural and Food Economy — to calculate the economic values and to determine agricultural type for every agricultural holding. Census data will be the base for the creation of the frame for multi-purpose sample surveys conducted by the CSO, including the sample survey of the agricultural holding structure, which according to the EU calendar should be conducted in 2005 and 2007.

In the case of sample surveys of the structure the most important methodological task after the census 2002 will be designing the suitable sampling plan and determining optimum sample size for the 2005 survey.

Taking into a consideration the fact that many agricultural holdings don't conduct agricultural activity (according to PSR 1996 — about 10%) or conduct activity below so called economic threshold, it is necessary to settle the

agricultural holdings population, which should be included in the survey and to design such a sampling scheme, in which the agricultural holdings will be stratified not only according to agricultural land areas, but also according to economical type and size. The size of a sample must ensure the generalisation at least on the country, voivodship and county level (NTS1, NTS2, NTS3). In this range the co-operation with mathematicians and experts of sample method is necessary.

Very important element when introducing and improving the surveys of the agricultural holding structure is to hold the statistical register and the possibility of its actualisation.

As it was mentioned above the typology of the agricultural holdings is inseparably connected with the structure surveys. In this scope the CSO co-operates with The Institute of Agricultural and Food Economy. This co-operation will be tightened. The Institute will be responsible for the preparation and actualisation of the set of SGM for the certain regions of the country, agriculture crop and species and groups of livestock. It will be necessary to recognise, after conducting the survey 2002, what types of agricultural holdings are found in Poland and whether there is a need for specification, in Polish conditions, of certain level of the EU classification.

When introducing the surveys of the agricultural holding structures it will be necessary to prepare special training programme for the employees of statistical offices and interviewers.

The improving of the agricultural statistics in the economical scope is, above all, to implement (by MARD and IAFE) the accounting system of agricultural holdings (so called Polish FADN) and, with the use of obtained information, conduct the tasks connected with the agriculture economic accounts. There are still implementation works concerning not only accounts themselves, but also adjusting those works to the agriculture surveys calendar.

In the scope of the subject of price survey in agriculture, it will be necessary to replace the information about market prices obtained from the correspondents through the sample survey. The works aiming at the fully adjustment of the price survey in agriculture to the requirements of European Union were taken.

8. Concluding remarks

The balance operations of agricultural products are very important element of the information system about agriculture information concerning production and manufacture of agricultural products, foreign trade, consumption of feeds, loses in storage etc and concerning the „markets" of cereals, potatoes, sugar, rapes, vegetables and fruits and feeds are used in the balance completion.

They are prepared annually for certain cereals, oil plants and oils, sugar beets and sugar, potatoes and starch, fruits, vegetables, vines, pulses, feeds, meat, eggs, milk and its processed products, animal fats and honey in the producer's,

market's and consolidated balance sheet's approach. These sorts of completions are made for more than 100 agricultural products along with a much specified data calculations about the foreign trade of those products, with the possibility of checking them through the mirror collation. The compilations are conducted at the stage of prognostics and final operations.

The rules of calculations of the crop products balance within the Member States of European Union were devised by Eurostat in the suitable forms of draft books for certain groups of plants. Eurostat's requirements concerning the animal product balance are not totally determined.

The balance calculations are prepared according to the principles of Eurostat, and also in adjustment to the needs and calculations for the Common Agricultural Policy. In the calculations specialists from the CSO, the Ministry of Agriculture and Rural Development, The Institute of Agricultural and Food Economy, the Agricultural Market Agency and the National Vine Chamber are participating. Further methodological works to prepare feeds balance and animal production balances are conducted.

Because of the inter-institutional character of the information system about agriculture it seems necessary to stark working on preparation of the projects of suitable legal regulations, defining between all the task and responsibilities partition, in connection with updated survey calendar, compatible with the domestic and international requirements, especially with the EU ones. Legal regulations should also secure stability of the financial means, necessary for the full realisation of the commitments, determined among all under the UE law.

The restructurisation of the agricultural information system is a steady process, especially in the period of the socio — economic and politically — organisational changes. Its necessity results from the changing conditions, among which the most important are, as always, changes of the domestic and foreign recipients. New and greater informatics possibilities are an important element, which imposes the construction of the system which would be fully computerised.

The activities connected with the restructurisation of the information system about agriculture should then lead to:

- full protection of the necessary information covering of the national, EU, and international organisation needs,
- guarantying the homogeneity of definition and information classification,
- guarantying of the high quality of the matter information,
- minimisation of the costs of collecting the data,
- and also making the further development and improving of the system possible.

RREFERENCES

GUS (2003), Raport z wyników Powszechnego Spisu Rolnego 2002 (Report of Results of the 2002 Census of Agriculture), Główny Urząd Statystyczny, Warsaw.

FAO (1995a), Taking Agricultural Censuses — guidelines supplementing the programme for the 1980 World Census of Agriculture — FAO, Rome, 1995.

FAO (1995b), Programme for the World Census of Agriculture 2000, Statistical Development Series — FAO, Rome, 1995.

KORDOS, J. and KURSA, L. (1997), Agricultural Sample Surveys in Poland, *Statistics in Transition,* Vol.3, No.1, pp.75—108.

MATEŃKO, K. and WANKE, H. (1996), *Powszechny spis rolny 1996 (The Agricultural census 1996)*, Wiadomości Statystyczne No. 4, pp. 19—26.

WANKE, H. (1997), Censuses of Agriculture in Poland, *Statistics in Transition* Vol.3, No.1, pp. 57—74.

WANKE, H. (2003), The Censuses of Agriculture in Poland in 1996 And 2002, *Statistics in Transition* Vol.6, No.2, pp. 275—286.

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 589—611

IS IT POSSIBLE TO ESTIMATE RELIABLE HOUSEHOLD EQUIVALENCE SCALES?

Adam Szulc[1]

ABSTRACT

Equivalence scales derived from the Almost Ideal Demand System are estimated using Polish microdata. The purpose of the study is to select an appropriate equivalence scale formula. The scales are defined assuming Barten scaling of prices and, alternatively, independence of base (IB). Three functional forms of equivalence scales are examined. The results prove strong sensitivity to the choice of formula. Moreover, the elasticities of scales with respect to some household attributes take unacceptable values if IB is imposed and the scales are defined by means of Cobb-Douglas function.

Key words: equivalence scale identification, AIDS

1. Introduction

Equivalence scales may be generally defined as parameters intended to adjust households' costs of living with respect to their demographic compositions. They are, therefore, inevitable elements of welfare comparisons between households. The number of estimations of equivalence scales performed over two last decades is huge. Many of them use sophisticated econometric techniques and demand functions. On the other hand, in most of applied studies on income distribution or social policy addressed to wider audience (e.g. public institutions) the scales are defined in the simplest possible way, mainly by the OECD formula. The explanation of this phenomenon is straightforward. Most of estimates of household equivalence scales that capture simultaneously (at least) number of adults and children in the household cannot be accepted for being contradictory to the economic theory and/or common sense. The questions addressed here is: how the equivalence scales should be defined in order to reach acceptable numerical values? This issue is rarely explores in empirical studies. Most of the authors

[1] Warsaw School of Economics, Institute of Statistics and Demography.
Al. Niepodległości 162, 02-554 Warszawa, Poland, e-mail: aszulc@sgh.waw.pl

focus on checking functional forms of the demand system and/or sensitivity of the results to the method of estimation. Those two topics are not investigated in this paper.

It is well known that equivalence scales are identifiable only under explicit assumptions. Two of them are included in this research. First one is on the form in which demographic attributes are incorporated into the cost (or utility) function. Next assumption decides whether the scales are invariant to the utility level at which welfare comparisons are made (this invariance is hereafter referred to as independence of base or IB). Formally, the scales hold IB if and only if the cost function is separable in the utility level and the household attributes (see Lewbel, 1989 or Blackorby and Donaldson, 1993). Although this property is relatively frequently accepted in the economic literature, there is no rationale for assuming that the results of comparisons should be equal for, loosely speaking, "rich" and "poor" households. As claimed by Blundell and Lewbel (1991) imposing IB is sufficient for indetifiability of the equivalence scales. Incorporation of demographic attributes by means of Barten scaling of prices is alternative condition for indetifiability.

In this study the scales are estimated separately under both types of aforementioned assumptions. The first set of equivalence scales is obtained through estimation of the Almost Ideal Demand System (AIDS) in order to calculate Barten equivalence scales (introduced in sub-section 2.2), without imposing IB. Two next sets are derived from estimates of two AI demand systems with the cost function on which separability in the utility level and the attributes is imposed to ensure satisfying IB. At this stage of estimation two alternative specifications of Cobb-Douglas functional form of the scale are examined. First, equivalence scales take the form proposed by Blackorby and Donaldson (1993) and estimated by, among others, Garner and Johnson (1992), and Nelson (1993). Alternative form of Cobb Douglas function was proposed by Phipps (1997).

In this paper the scales are specified as combinations of selected parameters of an Almost Ideal Demand System (AIDS) with household attributes incorporated into the cost function. The estimates of these parameters are obtained by means of estimations of the complete demand system utilising Polish quarterly household data of the period 1992—1997. Estimation of the AIDS equivalence scales for Poland using data covering the so-called transition period (started after departure from the centrally planned economy in 1990) may be also considered a purpose in itself. Liquidation of shortages on the consumer market was one of obvious consequences of the transition. These shortages were intrinsic under the previous system, therefore assuming consumer utility maximisation could not be rational. Consequently, any estimation of neoclassical demand system using pre-transition data had to be biased. The previous published estimates of Polish equivalence scales were using partly pre-transition data (Szulc, 1992b, 2001) or were based on Engel curves (Betti, 1999, Ekkert-Jaffé et al, 2000, Keane and Prasad, 2002).

2. Specification of the demand system and equivalence scales

2.1. Almost Ideal Demand System

AIDS assumes the following cost function C for *k-th* household

$$\ln C(P,u) = a_0 + \sum_{i=1}^{n} a_i \ln p_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} \ln p_i \ln p_j + b_0 u \prod_{i=1}^{n} p_i^{d_i} \qquad (1)$$

The following restrictions are imposed on the parameters

$$\sum_{i=1}^{n} a_i = 1; \quad \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} a_{ij} = 0; \quad a_{ij} = a_{ji} \qquad for\ i, j = 1, 2, \ldots, n$$

where:
$\mathbf{P} = [p_1, p_2, \ldots, p_n]$ — vector of prices,
p_i — price of *i-th* (i=1,2,...,n) commodity (in empirical applications it is substituted by a price index number),
u — non-negative utility level.

Household attributes may be included into the cost function (1) by putting them into p_i (see section 2.2) or by adding a term including a set of household attributes (section 2.3). The latter representation is relevant when IB is imposed on equivalence scales. In both representations, differences between household preferences enter the cost function through the household attributes rather than through the parameters.

All scales calculated in the present research belong to the class defined by Deaton and Muellbauer (1980b). The equivalence scale, m_k, comparing cost of living for two households, *k-th* and *r-th*, is the following ratio

Definition

$$m_k = m(\mathbf{P}, u, \mathbf{A_k}, \mathbf{A_r}) = \frac{C(\mathbf{P}, u, \mathbf{A_k})}{C(\mathbf{P}, u, \mathbf{A_r})} \qquad (2)$$

2.2. Semi-exact formula (Barten scales)

The scales introduced in this subsection are based on Barten (1964) idea of scaling prices by means of functions of household attributes. The latter ones are incorporated into the cost function following the linear specification proposed by Muellbauer (1977)

$$\ln \mathbf{P_{it}} = \ln p_i + \sum_{l=1}^{m} e_{il} \ln A_{lt} \qquad (3)$$

where A_{lt} is *l-th* attribute (l=1,2,..,m) of *t-th* household (e. g. number of persons, number of children, age or location dummies), e_{il} is a parameter reflecting an impact of *l-th* household attribute on *i-th* commodity consumption. $\mathbf{P}_{ik}$ is usually referred to as an "effective price", as it indicates real price to be paid by a household characterised by the vector of attributes $\mathbf{A_t}$. The sets of household attributes do not need to be identical for each commodity *i* in the cost function. This property is utilised in estimation of the demand equations to relax impact of data multicolinearity (see section 4).

The Marshalian demand function for *i-th* commodity (i = 1,2,...,n) appears as

$$w_{it} = a_i + \sum_{j=1}^{m} a_{ij}(\ln p_j + \sum_{l=1}^{m} e_{jl} \ln A_{lt}) + d_i \ln(X/\mathbf{\Pi}) \tag{4}$$

where $\mathbf{\Pi}$ is a type of price vector defined by

$$\ln \mathbf{\Pi} = a_0 + \sum_{i=1}^{n} a_i \ln \mathbf{P_{it}} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \ln \mathbf{P_{it}} \ln \mathbf{P_{jt}} \tag{5}$$

In the above specification e_{jl} parameters are assumed to be identical for each *i-th* equation. Parameter a_0, which appears in estimation of eqn (4), cannot be identified from consumer patterns. Deaton and Muellbauer (1980a) suggest cardinalisation of the utility function by setting utility level at zero at a subsistence minimum level. After that, a_0 may be substituted out as logarithm of a subsistence minimum for the year selected as a base in calculation of price indices p_i and for a household with zero demographic attributes (in such a case all p_{ik}=0 and u=0 and a_0 is a sole non-zero value on the right-hand side of eqn 1). However, there is no reasonable ground for appointing the proxy for b_0, which should be put into eqn (2) to calculate the equivalence scale. Using the so-called theorem on "semi-exact" (SE) translog equivalence scales (by Szulc, 1992a; for a precise specification of the translog function following Diewert's (1976) concept see Appendix B) allows approximation of AIDS scales without knowledge of b_0 (the proof is outlined in Appendix B; the theorem was first published in Szulc (1992b) however the proof contains some errors).

Theorem

If:
r-th and *k-th* households minimise their translog cost functions
then

$$\ln \frac{C(\mathbf{P}, u^*, \mathbf{A_k})}{C(\mathbf{P}, u^*, \mathbf{A_r})} = \frac{1}{2} \sum_{l=1}^{m} \sum_{i=1}^{n} \left[e_{il}(w_{ik} + w_{ir}) \right] \ln \frac{A_{lk}}{A_{lr}} \tag{6}$$

where
w_{it} — *i-th* budget share of *t-th* household (t=k,r)

$u^* = (u_k u_r)^{1/2}$

When u is a geometric mean of utility levels reached by households under comparison, the left-hand side of (6) is equal to logarithm of the right-hand side of (2), i. e. to Deaton-Muellbauer equivalence scale. It depends on observable variables and parameters e_{il} (i=1,2, ... ,n; l=1,2, ... ,m) which have to be estimated. The theorem may be applied also to the AIDS cost function. Namely, if $ln(X/P^*)$, where P^* stands for any price index formula, is considered a proxy for u in eqn (4), then the latter takes a form of translog Hicksian budget share. Using such a proxy may be also rational when (Diewert's) translog demand system is estimated, as nonlinearities in the term containing utility make it practically inestimable in the exact form.

SE[1] Barten scales do not allow IB by the definition. Each pair of households is compared at the utility level being a geometric mean of utilities reached by them. It seems that such a choice is more justifiable than one utility level for all households, as indirectly assumed by imposing IB.

SE translog scales are attributed by at least two advantages:

1. Each scale may be attached individually to any single household, depending on its budget shares. This is consistent with the true that cost of living depends not only on the prices but also on the household consumption patterns.
2. The scales gain clear economic interpretation. It is easy to demonstrate why changes in budget shares may change a scale and in which way interaction between household attributes and consumption (being controlled by e_{il}) influences the cost of living. It becomes clear now why scales of low-income households, characterised by relatively high food shares, are more "steep" than scales of "rich" ones. Moreover, due to clear interpretation of parameters e_{il}, they may be inspected in a visual way in order to reject unrealistic estimates of the cost function. It is not possible for most of remaining parameters that are not utilised in calculation of the scales.

2.3. Utility independent formulas

To obtain equivalence scales defined by (2) and satisfying IB the cost function should be rewritten. It may be done in the following manner

[1] Scales were named "semi-exact" to refer their concept to exact index numbers introduced by Diewert (1976). Exact index numbers can be calculated from observable data alone (do not need any estimation work), while SE equivalence scales calculated from eqn (6) require knowledge of parameters e_{il} but not of "regular" parameters of C. This property is passed also by IB equivalence scales. Therefore, IB scales are also referred to by many of authors as "exact", although in that case the "exactness" is obtained by imposing separability on the cost function. To avoid confusion, the scales that are exact in the latter sense are referred to in this paper as "independent of base".

$$\ln C(\mathbf{P}, u, \mathbf{A_t}) = a_0 + \sum_{i=1}^{n} a_i \ln p_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} \ln p_i \ln p_j + b_0 u \prod_{i=1}^{n} p_i^{d_i} + \ln m(\mathbf{P}, \mathbf{A_t}) \quad (7)$$

where $m(\mathbf{P}, \mathbf{A_t})$ stands for an equivalence scale. Moreover, the standard restrictions (see eqn 1) are imposed on the parameters.

The first scale (comparing *k-th* household with the reference one) passing IB checked in this research is defined (after Blackorby and Donaldson, 1993) by the logarithmic equation

$$\ln m_{BD}(\mathbf{P}, \mathbf{A_k}) = e_1 \ln(N) + \sum_{i=1}^{n} \ln p_i \sum_{l=2}^{m} e_{il} A_{lk} \quad (8)$$

where N stands for the number of persons in the household and A_{lk} is *l-th* element of $\mathbf{A_k}$ (A_{1k} = N). The additional restriction ensures homogeneity of the demand system

$$\sum_{i=1}^{n} e_{il} = 0 \qquad \text{for } l = 2, \ldots, m$$

The functional (Cobb-Douglas) form appearing in (8) was previously applied by Garner and Johnson (1992), and Nelson (1993) who employed the AIDS cost function consistent with (7).

Parameter e_1 may be interpreted as the scale elasticity with respect to the household size. Parameters appearing in the second term on the right-hand side of (8) indicate relative needs of households of certain type. Positive signs reflect higher needs of households with positive attributes and vice versa.

Putting (8) into the cost function (7) yields budget shares

$$w_{it} = a_i + \sum_{j=1}^{n} a_{ij} \ln p_j + d_i \ln\left(\frac{X}{\Pi \cdot m_{BD}}\right) + \sum_{l=2}^{m} e_{il} A_{lt} \quad (9)$$

where Π is defined by eqn (5) with prices p_i replacing effective prices $\mathbf{P_{it}}$.

Phipps (1997) employed alternative formula (utilising translog indirect cost function) to estimate children scales with the number of adults fixed at two.

$$\ln m_P(\mathbf{P}, \mathbf{A_k}) = s_1 \ln \frac{(N_k + 2)}{2} + \sum_{l=2}^{m} s_l A_{lk} + N_k \sum_{i=1}^{n} g_i \ln p_i \quad (10)$$

where N_k stands for the number of kids and A_l represents l-th element of the vector of auxiliary household attributes. N_k in the first term on the right-hand side is transformed to allow use of logarithms. Homogeneity of the demand system is ensured by imposing

$$\sum_{i=1}^{n} g_i = 0$$

As scales estimated in the present study take into account also number of adults, Phipps' formula was modified to

$$\ln m_P(\mathbf{P}, \mathbf{A_k}) = s_1 \ln N + \sum_{l=2}^{m} s_l A_l + (N-1)\sum_{i=1}^{n} g_i \ln p_i \qquad (11)$$

Number of persons in the household (N) replaces the number of kids. N-1 in the third term on the right-hand side ensures zero value for household composed of one adult (selected in this study as a reference type).

Budget shares appear in the form

$$w_{it} = a_i + \sum_{j=1}^{m} a_{ij} \ln p_j + d_i \ln\left(\frac{X}{\Pi \cdot m_P}\right) + N g_i \qquad (12)$$

where Π is defined by eqn (5) with prices p_j replacing effective prices $\mathbf{P_{it}}$.

3. The data

Expenditure and demographic data came from the Household Budget Survey (HBS) being collected annually by the Central Statistical Office of Poland (CSO). In 1992 the sample was based on quarterly rotation and did not cover certain socio-economic groups (militants, self- employed not in the agriculture). Approximately 10,000 households were covered in 1992. Since 1993 data collection has been being based on monthly rotation and two new important groups were added: self-employed and social welfare recipients. More details on changes in Polish HBS may be found in Keane and Prasad, 2002, and Szulc, 2000. In spite of differences in the methods of data collection between 1992 and in remaining period, that year was included into the sample to increase price variation. Average yearly sample size in 1993—1997 was approximately 32,000 households. As the present study utilises quarterly data, totally 24 sub-periods are covered. Rural households were excluded from the sample, to relax the impact of natural consumption and that of spatial variation between prices, which is poorly covered by the official statistics. Large shifts in income distributions (see Keane and Prasad, 2002, and Szulc, 2000), changes in price relations due to institutional reforms (especially on the housing market) and alterations in consumer habits probably have had serious, harmful impact on quality of the estimation.

Price indices are published by the CSO on monthly, quarterly, and yearly basis. In general, 1992—1997 was characterised by relatively high, but systematically decreasing inflation rates: from 9% in the initial quarter to 3% in the concluding one

(see Table A1 in Appendix A). In the present study the first quarter of 1992 was selected as the base period.

4. Estimation of the AIDS with household attributes

In many empirical applications an AIDS is estimated in a linearised version. In such cases eqn (5) is substituted out by a certain price index number, usually Stone share weighted formula. The recent studies, however, (Buse, 1994, Buse and Chan, 2000, and Pashardes, 1993) spelled out serious bias and inconsistency of resulting estimators. Therefore, the original AIDS function (eqns 4 and 5) was employed in the present research.

Consumer expenditures were split into five groups:

1) food and non-alcoholic beverages,
2) clothing and footwear, healthcare, hygiene,
3) housing and energy (heating, electricity, natural gas etc.),
4) transportation (including fuel), telecommunication, educational and cultural expenses,
5) alcoholic beverages, tobacco, other expenses.

As a rule, grouping of commodities was based on Hicks theorem on exact aggregation. Groups are composed of commodities whose prices were moving possible closely in proportions. The only exception is group no. 3 composed of highly complementary commodities: housing and energy (dominant part of expenditures on energy is being spent on heating). One equation in the system has to be dropped in estimation. The last one (referring to group 5) seems to be the best choice, as it contains expenditures, which hardly conform to theoretical budget shares — alcohol and tobacco. Moreover, expenditures on these commodities are usually seriously underestimated in household surveys.

To increase household homogeneity, those with abnormally high shares of expenditures on any single commodity group (0.7 for food and 0.5 for remaining groups) were deleted. Also those with zero expenditures were removed from the sample. Finally, households with total expenditures below 40% of the median and above four times median were also dropped. The final sample size was approximately 76,500, i. e. about 90% of the initial urban sample.

Following household attributes were included into the cost function:

1) Number of persons in the household,
2) Proportion of children aged below 10,
3) Proportion of children aged between 10 and 15,
4) Dummy for household head's age below 30,
5) Dummy for household head's age 60 and over,
6) Dummy for household low expenditures (below 60% of median equivalent expenditure). In order to calculate equivalent expenditures previous estimates

of equivalence scales (see Szulc, 2000) were employed. It seems that the final estimates are quite robust to choice of the scales at that point.

Moreover, two seasonal dummies (for 1992—1993 and 1994—1995) were included into the cost function together with household attributes. They are intended to capture shifts in demographic scaling. A single, non-poor person aged between 30 and 60 years therefore constitutes the reference household. To ensure zero values of all attributes, observations should be collected from 1996—1997 period.

Parameter a_0 in the AIDS cost function is equal to logarithm of half of the so-called social minimum calculated by the Institute of Labour and Social Studies in Poland and corrected by Szulc (2000). In practice, the results are quite robust to a choice of this parameter, although Buse and Chan (2000) demonstrated that using exogenous a_0 results in minor bias in estimates of a_i. Since the latter estimates are not being used in calculation of scales, the impact of that bias seems to be negligible.

Estimation of the demand system represented by especially eqns (4) and (5) is complicated (see Carrascal, 2000) due to nonlinearities and imposed restrictions on the parameters, especially e_{il} (which are present in each equation). Moreover, data multicolinearity is very likely to appear. To relax its impact the vector of household attributes for 'other' commodities was restricted to the form of three dummies, indicating households of 2, 3 and 4 and more persons.

As parameters e_{j1} (i = 1, ... , 4) appearing in eqn (3), e_1 in (8) and s_1 in (11) represent impacts of the household sizes on consumption needs, the estimates should fit the interval [0,1]. Moreover, e_{11}, matched with food consumption should by the highest, while e_{13} attached to housing expenditures should be relatively low. As claimed by Carrascal (2000), in empirical applications those estimates may not fit theoretically plausible interval [0,1]. Therefore, in estimation reported by Carrascal it was necessary to use the transformation $sin^2(e_i)$ in estimated equation to "force" the results to comply with the expected bounds (the sole demographic attribute in that study was the household size). In the present estimation, however, all estimates reached reliable values without imposing additional restrictions. It was passed also for the estimates of e_1 in eqn (8) and s_1 in (11).

The results of the estimation depend strongly on applied econometric algorithms. As the equations are highly nonlinear in the parameters, precision of the iterative methods may play important role in obtaining rational results. Broyden-Fletcher-Goldfarb-Shanno algorithm (see Fletcher, 1980) available within full information maximum likelihood estimation (FIML) in TSP 4.4 was used to produce estimates of the parameters, standard errors and covariance matrices.

Tables 1, 2 and 3 report estimates of the cost functions' parameters obtained without (Tab. 1) and under (Tabs. 2 and 3) assumption of separability of household attributes and utility. These results may be interpreted as the estimates of the demand systems for the households of reference type (with household

attributes set at unity). Tables 4, 5 and 6 display estimates of parameters modelling interaction between consumption and household attributes. They are used furthermore in calculation of the equivalence scales. All parameters of two aforementioned categories were estimated within one demand system simultaneously.

Table 1. AIDS estimates without separability of household attributes and utility: Barten scaling of prices

	Parameter	Estimate	t-statistic
	Food		
a_{11}	(food)	0.04767	8.197
a_{12}	(clothing)	0.06882	32.168
a_{13}	(housing)	-0.04693	-13.262
a_{14}	(transportation)	-0.01780	-2.954
a_1	(intercept)	0.60286	258.985
d_1	(expenditure)	-0.14850	-111.297
	Clothing		
a_{22}	(clothing)	0.01617	7.081
a_{23}	(housing)	-0.02888	-12.014
a_{24}	(transportation)	-0.06508	-35.513
a_2	(intercept)	0.05510	39.172
d_2	(expenditure)	0.05678	63.024
	Housing		
a_{33}	(housing)	0.10425	23.234
a_{34}	(transportation)	-0.02400	-9.655
a_3	(intercept)	0.24942	118.072
d_3	(expenditure)	-0.01421	-10.984
	Transportation		
a_{44}	(transportation)	0.06188	12.029
a_4	(intercept)	0.02308	12.396
d_4	(expenditure)	0.10327	87.532

R^2 Food: 0.292 Clothing: 0.084 Housing: 0.047 Transportation: 0.214

All estimates of the reference household demand system obtained without separability assumption are statistically significant at 0.01 level. The signs of d_is, indicate that food and housing/energy are necessities, while 'clothing' (including also health and hygiene expenditures) and 'transportation' (including also educational and cultural expenditures) are luxuries. 'Other' goods (alcohol, tobacco and miscellaneous expenditures) are also found luxuries. Estimates obtained under separability assumption (see Tabs. 2 and 3) yield similar pictures. Parameters a_{ij} have usually the same signs under all methods, with a_{34} (the parameter modelling interaction between housing and transportation) estimate as a sole exception in this matter. It is negative without imposing separability and positive for both estimations under separability. One estimate in each estimation

under separability is significant at 0.08 level: a_{13} (interaction between food and housing) when Blackorby-Donaldson (BD) scales are employed and a_4 (intercept in 'transportation' equation) when Phipps scales are employed.

Table 2. AIDS estimates under separability of household attributes and utility: Blackorby-Donaldson equivalence scales

	Parameter	Estimate	t-statistic
	Food		
a_{11}	(food)	0.05658	3.764
a_{12}	(clothing)	0.13343	10.992
a_{13}	(housing)	-0.01767	-1.778
a_{14}	(transportation)	-0.15483	-11.734
a_1	(intercept)	0.60321	202.797
d_1	(expenditure)	-0.14705	-119.143
	Clothing		
a_{22}	(clothing)	0.21948	7.664
a_{23}	(housing)	-0.12878	-8.872
a_{24}	(transportation)	-0.28936	-15.835
a_2	(intercept)	0.05201	19.580
d_2	(expenditure)	0.05803	59.630
	Housing		
a_{33}	(housing)	0.12823	10.106
a_{34}	(transportation)	0.15327	13.378
a_3	(intercept)	0.23386	71.785
d_3	(expenditure)	-0.01526	-12.200
	Transportation		
a_{44}	(transportation)	0.42778	22.059
a_4	(intercept)	0.03289	10.716
d_4	(expenditure)	0.10415	94.164

R^2 Food: 0.283 Clothing: 0.085 Housing: 0.025 Transportation: 0.208

Table 3. AIDS estimates under separability of household attributes and utility: Phipps equivalence scales

	Parameter	Estimate	t-statistic
	Food		
a_{11}	(food)	0.05982	3.940
a_{12}	(clothing)	0.12314	10.458
a_{13}	(housing)	-0.03890	-4.076
a_{14}	(transportation)	-0.13998	-10.654
a_1	(intercept)	0.59773	217.402
d_1	(expenditure)	-0.14864	-142.172
	Clothing		
a_{22}	(clothing)	0.17753	7.441

a_{23}	(housing)	-0.15556	-11.184
a_{24}	(transportation)	-0.17664	-13.829
a_2	(intercept)	0.08520	49.450
d_2	(expenditure)	0.05380	72.445
	Housing		
a_{33}	(housing)	0.23966	23.763
a_{34}	(transportation)	0.08050	8.347
a_3	(intercept)	0.22227	135.985
d_3	(expenditure)	-0.01042	-12.160
	Transportation		
a_{44}	(transportation)	0.24899	17.244
a_4	(intercept)	0.00380	1.764
d_4	(expenditure)	0.10668	118.836

R^2 Food: 0.286 Clothing: 0.074 Housing: 0.037 Transportation: 0.202

Estimates of coefficients on household attributes are reported in Tables 4—6. Most of them are significant at 0.01 level. Again, more non-significant estimates (2 against 1) may be found for the estimations with imposed separability. Parameters representing economies of scale in the households (e_{i1} in the first estimation, e_1 in the second and s_1 in the third one) range between 0 and 1. Interpretations of other estimates differ between models, however the meanings of the signs are identical as far as first and second estimations are compared: positive sign means positive impact of the attribute on the household cost of living and vice versa. Apparently, for many parameters attached to the same attribute estimates suggest opposite impacts on household consumption. For example, in accord with the first model being poor reduces relative demand for food ($e_{18}<0$) and increases demand on 'clothing' ($e_{28} > 0$). Conclusions derived from the second model are reverse ($e_{17} > 0$ and $e_{27} < 0$, although the latter estimate is statistically significant at 0.1 level only). Also, presence of kids reduces 'clothing' and 'transportation' expenditures in accord with the first model (e_{22}, e_{23}, e_{42}, e_{43} are negative) but increases those expenditures in accord with the second model (e_{21}, e_{22}, e_{41}, e_{42} are positive). Finally, estimates on seasonal dummies rarely have the same signs. It is not surprising that aforementioned discrepancies lead to opposite conclusions on the equivalence scales (they are discussed in more details in succeeding section).

Estimates obtained with the use of Phipps scales cannot be compared directly with the previous ones. Instead, the estimates of s_l ($l = 1, ..., 8$) should be compared with the appropriate elasticities of SE and BD scales with respect to household attributes. This issue is discussed in succeeding section.

Table 4. Interaction between consumption and household attributes: estimation without separability, Barten scaling of prices

Attribute	Estimate	T-statistic
	Food	
e_{11} (size)	0.91987	103.001
e_{12} (kids1)	-0.18812	-6.699
e_{13} (kids2)	0.21974	6.572
e_{14} (seas1)	-0.04985	-5.455
e_{15} (seas2)	-0.05193	-6.973
e_{16} (age1)	-0.20559	-15.052
e_{17} (age2)	0.01244	1.111
e_{18} (poor)	-0.03056	-2.875
	Clothing	
e_{21} (size)	0.59263	59.697
e_{22} (kids1)	-0.65407	-15.609
e_{23} (kids2)	-0.83384	-18.226
e_{24} (seas1)	0.30352	20.987
e_{25} (seas2)	-0.07254	-7.804
e_{26} (age1)	0.12078	6.667
e_{27} (age2)	0.16296	10.929
e_{28} (poor)	0.09125	6.170
	Housing	
e_{31} (size)	0.42126	26.775
e_{32} (kids1)	-0.25132	-8.812
e_{33} (kids2)	-0.09249	2.6782
e_{34} (seas1)	0.12140	11.978
e_{35} (seas2)	0.03939	5.1916
e_{36} (age1)	-0.13182	-11.680
e_{37} (age2)	0.04837	4.621
e_{38} (poor)	-0.03604	-3.409
	Transportation	
e_{41} (size)	0.84698	89.241
e_{42} (kids1)	-0.59298	-16.490
e_{43} (kids2)	-0.13154	-3.106
e_{44} (seas1)	0.06885	5.864
e_{45} (seas2)	-0.13298	-14.514
e_{46} (age1)	-0.24645	-15.445
e_{47} (age2)	-0.23609	-20.406
e_{48} (poor)	0.06669	5.4245
	Other	
e_{51} (size2)	0.27806	16.666
e_{51} (size3)	1.02445	63.833
e_{51} (size4+)	1.32598	71.293

Table 5. Interaction between consumption and household attributes: estimation under separability, Blackorby-Donaldson equivalence scales

Attribute	Estimate	T-statistic
Whole consumption		
e_1 (size)	0.65106	108.190
Food price		
e_{11} (kids1)	-0.05098	-18.768
e_{12} (kids2)	-0.01739	-6.333
e_{13} (seas1)	0.03764	19.041
e_{14} (seas2)	-0.00368	-2.953
e_{15} (age1)	-0.01693	-12.981
e_{16} (age2)	0.01376	13.751
e_{17} (poor)	0.00765	5.953
Clothing price		
e_{21} (kids1)	0.04516	21.419
e_{22} (kids2)	0.01665	7.777
e_{23} (seas1)	0.00820	4.075
e_{24} (seas2)	0.00977	7.058
e_{25} (age1)	0.01442	13.816
e_{26} (age2)	0.01389	17.694
e_{27} (poor)	-0.00164	-1.640
Housing price		
e_{31} (kids1)	-0.03082	-12.212
e_{32} (kids2)	-0.03147	-12.240
e_{33} (seas1)	-0.02256	-10.877
e_{34} (seas2)	-0.00569	-4.802
e_{35} (age1)	0.00294	2.322
e_{36} (age2)	0.02004	21.166
e_{37} (poor)	-0.01319	-10.596
Transportation price		
e_{41} (kids1)	0.06157	25.420
e_{42} (kids2)	0.06043	24.627
e_{43} (seas1)	-0.04761	-20.566
e_{44} (seas2)	-0.01840	-14.011
e_{45} (age1)	-0.00122	-1.040
e_{46} (age2)	-0.03765	-41.883
e_{47} (poor)	0.00265	2.359

Note: estimates for 'miscellaneous' goods are obtained from the restriction:

$$e_{5l} = -(e_{1l} + e_{2l} + e_{3l} + e_{4l}),\ l = 1, \ldots, 7$$

Table 6. Interaction between consumption and household attributes: estimation under separability, Phipps equivalence scales

Attribute	Estimate	T-statistic
Whole consumption		
s_1 (size)	0.64726	44.022
s_2 (kids1)	-0.49752	-29.833
s_3 (kids2)	-0.30338	-18.553
s_4 (seas1)	0.22439	18.390
s_5 (seas2)	-0.00845	-1.113
s_6 (age1)	-0.08958	-11.787
s_7 (age2)	0.15268	24.583
s_8 (poor)	0.01088	1.532
Food price		
g_1 (size)	0.00872	10.567
Clothing price		
g_2 (size)	-0.00106	-3.004
Housing price		
g_3 (size)	-0.01328	-49.990
Transportation price		
g_4 (size)	0.00735	11.927

Note: estimates for 'miscellaneous' goods are obtained from the restriction:

$$g_5 = -(g_1 + g_2 + g_3 + g_4)$$

Legend to Tables 4, 5 and 6:

size — number of persons in the household

size2 — 2 person household (dummy)

size3 — 3 person household (dummy)

size4+ — at least 4 person household (dummy)

kids1 — share of children aged from 1 to 9

kids2 — share of children aged from 10 to 15

seas1 — data from 1992-1993

seas2 — data from 1994-1995

age1 — household head aged below 30 (dummy)

age2 — household head aged 60 or more (dummy)

poor — household equivalent expenditures below 60% of the median (dummy)

5. Equivalence scales: empirical results

All types of equivalence scales for selected categories of households, calculated for 1997 are displayed in Table 7. Due to including seasonal dummies (and also actual budget shares in SE scales, and actual prices in BD and Phipps scales), all scales are subject to change in time. As noted before, SE and IB equivalence scales are dissimilar. The previous ones are steeper, albeit disparities in economies of scale are not large. Differences between types of family (under fixed household size) are more substantial. It may be surprising that largest disproportion arose between BD and Phipps scales, i. e. those with IB imposed. Differences between scales with respect to a type of family are negligible in the case of BD scales. On the other hand, they are abnormally high when Phipps scales are considered. Similar differences between BD and Phipps scales might be derived from results by Garner and Johnson (1992), and Phipps (1997), in spite of using different demand systems and data sets. Results for SE scales lie between above-mentioned extremes.

Huge discrepancies between scales with IB imposed may be easily explained by analysing their specifications. Elasticities of BD scales with respect to household attributes other than number of persons may be calculated from

$$\frac{\partial \ln m_{BD}}{\partial \ln A_l} = \sum_{i=1}^{n} e_{il} \ln p_i A_{lk} \tag{13}$$

Since condition $\sum e_{il} = 0$ (ensuring homogeneity of the demand system) is not embed into the estimation but rather imposed on the estimates afterward, the magnitude of elasticity obtained by means of (13) is very low. In Phipps' formula elasticity for any attribute but number of persons (or children in the original study) is equal to $S_l A_l$. For SE translog formula (m_{SE}) elasticity with respect to any household attribute A_l may be calculated as follows

$$\frac{\partial \ln m_{SE}}{\partial \ln A_l} = \frac{1}{2} \sum_{i=1}^{n} \left[e_{il} (w_{ik} + w_{ir}) \right] \tag{14}$$

Therefore, it may be interpreted as an arithmetic mean of parameters e_{il} weighted by budget shares.

Table 7. Mean equivalence scales for selected households in 1997

Category of household					Type of scale		
No.of adults	Kids 1—9	Kids 10—15	Head 30—	Head 60+	SE	BD	Phipps
			Non-poor				
1	0	0	0	0	1.000	1.000	1.000
2	0	0	0	0	1.635	1.570	1.564
3	0	0	0	0	2.279	2.045	2.030
4	0	0	0	0	2.843	2.454	2.441
2	1	0	0	0	2.024	2.040	1.719
2	0	1	0	0	2.206	2.037	1.834
2	1	1	0	0	2.568	2.475	2.006
2	0	0	1	0	1.428	1.574	1.430
2	0	0	0	1	1.641	1.605	1.821
			Poor				
1	0	0	0	0	1.019	1.019	1.011
2	1	1	0	0	2.621	2.502	2.020

Differences between BD and Phipps scales are not only numerical but also lead to opposite economic conclusions. In accord with the first method small children (from 1 to 9 years of age) are cheaper to keep than older ones. The second method indicates that younger children are more expensive to keep. Differences arise also in the case of impact of poverty and age. As results from Phipps scales, poor households face higher than average costs of living and households headed by persons under 30 are cheaper than "mean-age" households are. BD scales lead to exactly reverse conclusions. Both methods prove higher costs of living faced by households headed by persons aged at least 60. On the other hand, there is a perfect consistency between Phipps and SE scales in the following conclusions: i/ small children are cheaper to keep than older ones are, ii/ households headed by persons below 30 are cheaper than those headed by persons between 30 and 60 are, while the latter ones are cheaper than those headed by persons at 60 or older, iii/ poor households face higher cost of living than non-poor do, iv/ the lowest scales are reported for 1994-1995, while the highest for 1992-1993.

6. Conclusions

The results depend strongly on specification of the scales, provided fixed (Almost Ideal) demand system. The scales with imposed independence of base are more flat with respect to the household size, as compared to the scales estimated without that restriction. The differences, however, are not very large in that case. Specification of equivalence scales, provided the demand system form, has more serious impact on the results. Using Cobb-Douglas function in the form proposed by Blackorby and Donaldson (1993) seems to underrate elasticity of scale with

respect to household attributes other than number of persons. On the other hand, employing the Cobb-Douglas function proposed by Phipps (1998) results in overestimation of those elasticities. Apart from some reservation that should be attributed to IB property, the scales of that type seek a proper specification. Cobb-Douglas formula does not seem to be a good choice for that purpose. Semi-exact equivalence scales, assuming Barten scaling of prices, result in much more reliable values.

The numerical results obtained for Polish households vary substantially from those calculated for such countries like Canada, the UK, or the US. Polish scales are much steeper in household sizes than the latter ones. This property may be easily explained by means of the budget share dependent (semi-exact translog) scales. Lower economies of scale result mainly from higher food ratios observed for Polish households. Nevertheless, some "universal" properties of equivalence scales are confirmed for two types of them ("semi-exact", that does not assume independence of base and Phipps formula, assuming this property): households with children are cheaper to keep and small children are cheaper than older ones. However, only semi-exact formula assuming Barten scaling of prices is able to produce acceptable estimates when wider set of demographic attributes is considered.

Acknowledgements

This study was sponsored by the Polish State Committee for Scientific Research (KBN). I would like to thank Dan Slesnick for his advice during my previous attempts at estimation of equivalence scales and Stanisław Maciej Kot for comments on the earlier draft of this paper. All remaining errors are the sole responsibility of mine.

REFERENCES

BETTI, G. (1999), A Non-parametric Approach to Equivalence Scales, *Statistics in Transition*, vol. 4, 181—194.

BARTEN, A. (1964) Family Composition, Prices, and Expenditure Patterns, in P. Hart, G. Mills and J. K. Whitaker (eds.), *Econometric Analysis for National Economic Planning*, Butterworth, London.

BLUNDELL, R. AND LEWBEL, A. (1991), The Information Content of Equivalence Scales, *Journal of Econometrics*, vol. 50, 49—68.

BUSE, A. (1994), Evaluating the Linearized Almost Ideal Demand System, *American Journal of Agricultural Economics*, vol. 76, 781—793.

BUSE, A. AND CHAN, W. H (2000), Invariance, Price Indices and Estimation in Almost Ideal Demand Systems, *Empirical Economics*, vol. 25, 519—539.

BLACKORBY, C. AND DONALDSON, D. (1993), Adult Equivalence Scales and the Economic Implementation of Interpersonal Comparisons of Well-Being, *Social Choice and Welfare*, vol. 10, 335—361.

CARRASCAL, U. (2000), AIDS Extended by Means of Family Composition, *Applied Economics Letters*, vol. 7, 21—23.

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

DEATON, A. AND MUELLBAUER, J. (1980b), *Economics and Consumer Behavior*, Cambridge, Cambridge University Press.

DIEWERT, W. E. (1976), Exact and Superlative Index Numbers, *Journal of Econometrics*, vol. 4, 115—145.

EKERT-JAFFÉ, O., GARDES, F AND STARZEC, C. (2000), Estimating the Cost of Children in Poland Using Panel Data, *paper presented at 26th General Conference of the International Association for Research in Income and Wealth,* Cracow, Poland, August 27— September 2.

FLETCHER, R. (1980) *Practical Methods of Optimization, Volume I: Unconstrained Optimization*, John Wiley and Sons, New York.

GARNER, T. AND JOHNSON, D. S. (1992), Estimating Unique Equivalence Scales, *paper presented at the 22nd General Conference of the International Association for Research in Income and Wealth*, Flims, Switzerland, August 31— September 5.

KEANE, M. AND PRASAD, E. (2002), Inequality, Transfers and Growth: New Evidence from the Economic Transition in Poland, *Review of Economics and Staistics*, vol. LXXXIV, pp. 324—341.

LEWBEL, A. (1989), Household Equivalence Scales and Welfare Comparisons, *Journal of Public Economics*, vol. 52, 377—391.

MUELLBAUER, J. (1977), Testing the Barten Model of Household Composition Effects and the Cost of Children, *Economic Journal*, vol. 87, 460—487.

NELSON, J. (1993), Independent of Base Equivalence Scales Estimation Using United States Micro-Level Data, *Annales D'Economie at de Statistique*, vol. 29, 43—63.

PASHARDES, P. (1993), Bias in Estimating the Almost Ideal Demand System with the Stone Index Approximation, *Economic Journal*, vol. 103, 908—915.

PHIPPS, S. (1998), What Is the Income "Cost of Child"? Exact Equivalence Scales for Canadian Two-Parent Families, *Review of Economics and Statistics*, vol. 80, 157—164.

SZULC, A. (1992a), Quasi-Exact Equivalence Scales Estimation, *Przegląd Statystyczny*, vol. 39, 302—307.

SZULC, A. (1992b), "Quasi-exact Equivalence Scales for Poland", *Economic Notes*, vol. 21, 175—183.

SZULC, A. (2000), Economic Transition, Poverty and Inequality: Poland in the 1990s, *Statistics in Transition*, vol. 4, 997—1017.

SZULC, A. (2001), Almost Ideal Demand System: Estimates for Poland with Applications in Welfare Measurement, in: S. Brosig and M. Hartman, *Analysis of Food Consumption in Central and Eastern Europe: Relevance and Empirical Methods*, Wissenschaftsverlag Vauk Kiel KG, 2001, pp. 2—18.

APPENDIX A. Descriptive Statistics

Table A1. Cumulative quarterly price indices (1992, 1st quarter=100)

Year, quarter	Total	Food	Clothing	Housing	Transp.	Other
1992, I	100,00	100,00	100,00	100,00	100,00	100,00
II	109,20	108,50	108,47	112,82	110,98	105,42
III	117,28	117,18	116,55	121,11	118,17	111,41
IV	128,66	132,53	127,61	129,97	127,88	117,40
1993,I	140,75	142,74	141,14	149,91	140,46	130,62
II	150,04	150,44	152,58	163,87	150,80	137,64
III	157,69	153,15	164,25	178,82	160,51	144,25
IV	171,73	171,68	180,83	189,86	172,63	153,96
1994,I	185,47	187,13	195,57	209,23	183,90	166,56
II	198,26	199,11	211,25	229,74	193,05	178,46
III	211,15	210,66	225,34	250,41	204,50	190,54
IV	229,31	229,20	243,94	275,87	218,19	210,55
1995,I	247,65	250,05	260,90	306,66	231,99	219,99
II	262,26	266,81	277,38	319,99	241,39	237,44
III	266,20	256,94	290,87	334,59	251,28	251,91
IV	280,31	273,89	306,67	342,57	263,45	267,26
1996,I	299,09	293,34	319,83	369,38	276,26	290,52
II	314,34	311,53	331,66	383,81	289,42	310,34
III	320,31	309,97	341,27	391,53	298,87	325,35
IV	333,45	326,09	355,17	405,38	308,00	335,71
1997,I	350,12	337,83	366,45	436,21	321,13	354,60
II	360,97	348,97	377,73	446,31	330,39	370,27
III	367,11	346,88	388,54	457,87	341,50	378,27
IV	378,86	360,76	402,73	465,99	349,63	389,07

Table A2.. Mean budget shares: whole sample

	Food	Clothing	Housing	Transp.	Other
1992	0.399	0.129	0.208	0.166	0.098
1993	0.395	0.142	0.207	0.195	0.060
1994	0.380	0.149	0.227	0.186	0.058
1995	0.381	0.148	0.235	0.174	0.063
1996	0.370	0.149	0.208	0.210	0.063
1997	0.358	0.155	0.210	0.215	0.062
1992-97	0.373	0.149	0.216	0.199	0.062

Legend to Tables A1 and A2

1/ "Clothing": apparel, health, and hygiene
2/ "Housing": shelter, energy
3/ "Transp.": transportation and telecommunication, fuel, education, culture
4/ "Other": alcohol, tobacco, other expenses not included in 1/ — 4/

Table A3. Household attributes: mean values for urban households

	Size	Kids 1—9	Kids 10—15	Head30—	Head60+	Poor
1992	2.622	0.363	0.270	0.067	0.343	0.185
1993	3.055	0.402	0.358	0.083	0.221	0.249
1994	3.092	0.431	0.368	0.083	0.222	0.265
1995	3.046	0.385	0.366	0.170	0.203	0.265
1996	3.013	0.367	0.355	0.083	0.233	0.252
1997	3.051	0.375	0.351	0.091	0.227	0.256
1992-97	3.031	0.391	0.335	0.100	0.227	0.254

Legend

1/ Kids 1—9 — number of kids aged from 1 to 9
2/ Kids 10—15 — number of kids aged from 10 to 15
3/ Head30— — household head below 30
4/ Head60+ — household head 60 and more
5/ Poor — proportion of households below 60% of median equivalent expenditure

APPENDIX B Outline of the proof of theorem on semi-exact equivalence scales

Translog cost function with demographic variables is defined as follows

$$\ln C(P,u,A_t) = c_0 + \sum_{i=1}^{n} c_i \ln \mathbf{P}_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} \ln \mathbf{P}_i \ln \mathbf{P}_j + b \ln u + c(\ln u)^2 + \sum_{i=1}^{n} d_i (\ln u) \ln \mathbf{P}_i \quad \text{(B1)}$$

where $\mathbf{P_i}$ is defined by eqn (3). Hicksian budget share appear in the form

$$w_{it} = c_i + \sum_{j=1}^{m} c_{ij}(\ln p_j + \sum_{l=1}^{m} e_{il} \ln A_{lt}) + d_i \ln u \quad \text{(B2)}$$

The following lemma (see Diewert, 1976) is employed in the proof

Quadratic Approximation Lemma (QAL)

$f : R^n \to R^1$ is the following quadratic function

$$f(Z) = \alpha_0 + \sum_{i=1}^{n} \alpha_i z_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_{ij} z_i z_j$$

($\alpha_{ij} = \alpha_{ji}$ for i, j = 1, ... ,n), iff for any Z_1 and Z_0 holds

$$f(Z_1) - f(Z_0) = \sum_{i=1}^{n} [\frac{\partial f(Z_1)}{\partial z_i} + \frac{\partial f(Z_0)}{\partial z_i}](z_{i1} - z_{i0})$$

Substituting $\alpha_i = c_i + d_i(\ln u)$ allows applying QAL to the cost function defined by eqn (B1) and then obtaining

$$\ln C(P,u,A_k) - \ln C(P,u,A_r) = \frac{1}{2}\sum_{i=1}^{n} [\frac{\partial \ln C(P,u,A_k)}{\partial \ln p_i} + \frac{\partial \ln C(P,u,A_r)}{\partial \ln p_i}] \sum_{l=1}^{m} m_{il} (\ln A_{lk} - \ln A_{lr}) \quad \text{(B3)}$$

For $u_c = (u_k u_r)^{1/2}$ this equation may be rewritten, recalling eqn (B2)

$$\frac{\ln C(P,u_c,A_k)}{\ln C(P,u_c,A_r)} = \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{m} m_{il} [\frac{\partial \ln C(P,u_k,A_k)}{\partial \ln p_i} + \frac{\partial \ln C(P,u_r,A_r)}{\partial \ln p_i}] \ln \frac{A_{lk}}{A_{lr}}$$

Finally, recalling again (B2), one can arrive at (6).

QED

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 613—614

Book Review

Advanced Sampling Theory with Applications

by **Sarjinder Singh** , Kluwer Academic Publishers, 2003, 1194 pp.

Advanced Sampling Theory with Applications: How Michael "selected" Amy is a comprehensive exposé of basic and advanced sampling techniques along with their applications in the diverse fields of science and technology.

This book is a multi-purpose document. It can be used as a text by teachers, as a reference manual by researchers, and as a practical guide by statisticians. It covers 1165 references from different research journals through almost 1900 citations across 1194 pages, a large number of complete proofs of theorems, important results such as corollaries, and 324 unsolved exercises from several research papers. It includes 159 solved, data-based, real life numerical examples in disciplines such as Agriculture, Demography, Social Science, Applied Economics, Engineering, Medicine, and Survey Sampling. These solved examples are very useful for an understanding of the applications of advanced sampling theory in our daily life and in diverse fields of science. An additional 173 unsolved practical problems are given at the end of the chapters. University and college professors may find these useful when assigning exercises to students. Each exercise gives exposure to several complete research papers for researchers and students. The data-based problems show statisticians how to select a sample and obtain estimates of parameters from a given population by using different sampling strategies, systematic sampling, stratified sampling, cluster sampling, and multi-stage sampling. Derivations of calibration weights from the design weights under single phase and two-phase sampling have been provided for simple numerical examples. These examples will be useful to understand the meaning of benchmarks to improve the design weights. These examples also explain the background of well-known scientific computer packages like CALMAR, GES, SAS, STATA, and SUDAAN etc., used to generate calibration weights by most organizations in the public and private sectors. The ideas of hot deck, cold deck, mean method of imputation, ratio method of imputation, compromised imputation, and multiple imputations have been explained with very simple numerical examples. Simple examples are also provided to understand Jackknife variance estimation under single phase, two-phase [or random non-response by following Sitter (1997)] and multi-stage stratified designs. This book also covers, in a very simple and compact way, many new topics not yet available in any book on the international market. A few of these interesting topics are:

median estimation under single phase and two-phase sampling, difference between low level and higher level calibration approach, calibration weights and design weights, estimation of parametric functions, hidden gangs in finite populations, compromised imputation, variance estimation using distinct units, general class of estimators of population mean and variance, wider class of estimators of population mean and variance, power transformation estimators, estimators based on the mean of non-sampled units of the auxiliary character, ratio and regression type estimators for estimating finite population variance similar to proposed by Isaki (1982), unbiased estimators of mean and variance under Midzuno's scheme of sampling, usual and modified jackknife variance estimator, estimation of regression coefficient, concept of revised selection probabilities, multi-character surveys sampling, overlapping, adaptive, and post cluster sampling, new techniques in systematic sampling, successive sampling, small area estimation, continuous populations, and estimation of measurement errors.

Contents

Volume I: — Preface.
1. Basic Concepts and Mathematical Notation.
2. Simple Random Sampling.
3. Use of Auxiliary Information. Simple Random Sampling.
4. Use of Auxiliary Information. Probability Proportional to Size and With Replacement (PPSWR) Sampling.
5. Use of Auxiliary Information. Probability Proportional to Size and Without Replacement (PPSWOR) Sampling.
6. Use of Auxiliary Information. Multi-Phase Sampling.

Volume II:
7. Systematic Sampling.
8. Stratified and Post-Stratified Sampling.
9. Non-Overlapping, Overlapping, Post, and Adaptive Cluster Sampling.
10. Multi-Stage, Successive, and Re-Sampling Strategies.
11. Randomized Response Sampling. Tools for Social Surveys.
12. Non-Response and its Treatments.
13. Miscellaneous Topics.

Appendix. Tables. Populations. Bibliography. Author Index. Handy Subject Index. Additional Information.

Prepared by Jan Kordos

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 615—620

REPORTS

FROM THE FIRST DEMOGRAPHIC CONGRESS IN POLAND: *'POLAND VS. EUROPE. DEMOGRAPHIC PROCESSES AT THE BEGINING OF THE 21ST CENTURY'*

Duration time: 15 September 2002—25/26 November 2003

It was the first time a Demographic Congress had been organized in Poland after World War II. Its main title was '***Europe and Poland. Demographic Processes at the Beginning of the 21st Century'.***

The initiative to organize a debate on assessment and especially the prospects of demographic development of our country came from Polish Demographic Society, The Committee of Demographic Research at the Polish Academy of Sciences, Government Population Council and Central Statistical Office.

The concept to convene the Congress came from a deep conviction that demographic processes are of fundamental importance for outlining the long-term strategy of development for our country. Such factors as education, innovative capability, civilization standards, affluence of citizens will determine, what position Poland will take among the countries worldwide.

The Congress was organized under the auspices of Aleksander Kwaśniewski, the President of Poland.

The aims of the Congress were:

- Synthetic assessment of demographic processes in Poland — determining the present condition and prospects of Poland's future development;
- Assessment of socio-economic, regional and political conditions of demographic development of Poland, and especially the changes in the period of system transformation;
- Determining the demographic prospects of the country, their conditioning and socio-economic effects;
- Defining the character, methods and means of population policy in Poland at the turn of the century and Poland's membership in the EU;
- Evaluation of the output of Polish demography and establishing its needs — its further development, research and teaching methods.

The Congress was very special in form and it lasted over a year. Official opening of the Congress took place on 15 September 2001 and the individual seminars, conferences and other meetings were scheduled for subsequent months of years 2001 and 2002.

Complexity of problems under debate lay behind the decision to divide the Congress into four parts:

Part I — Inauguration Session. It took place 15th of September 2001 and was devoted to the demographic situation in Poland against the demographic problems of the modern world. The most important factors to shape demographic development were presented at the session. Other vital issues under discussion included: the role of the family, ethical meaning of demographic processes and their significance to social progress. Information on the National Census 2002 was also presented. Theses formulated at this sitting set grounds for organization of seminars devoted to different problems and regions of the country during the second and the third Parts of the Congress.

Part II was a cycle of thematic seminars, devoted to the most important factors of demographic development of Poland, inclusive of living conditions of families, health status of citizens, education, dwelling conditions, employment, ecology and the cultural factors. The key point of this part of the Congress was organization of all-Polish conferences in response to the most perplexing social and demographic problems, which are the most difficult current nation-wide dilemmas in Poland.

Sixteen national conferences were organized under the program of this Part of the Congress:

- "*School medicine — main problems and trends in system solution*s" (organizer: J. Szymborski; Polish Hygienic Society, Mother and Child Institute, National Centre of Support to Health Education (Ministry of National Education and Sport);
- "*Social problems of handicap persons*" (organizer: J. Chapligo-Sikorska in cooperation with Polish Governmental Council and Institute of Philosophy and Sociology (Polish Academy of Sciences);
- "*Procreative health in Poland — current status and perspectives of development*" (organizer: B. Balińska; United Nations Development Programme, United Nations Population Fund and Mother and Child Institute);
- "*Demographic processes and labour market in Poland*" (organizer: B. Balcerzak-Paradowska; Instate of Labour and Social Affairs, Committee of Labour and Social Policy Sciences (Polish Academy of Sciences), Committee of Demographic Sciences (Polish Academy of Sciences), Polish Demographic Society);
- „*III Congress of Polish Perinatal Medicine Society*" (organizer: Polish Perinatal Medicine Society in cooperation with Perinatal Section of Polish Gynaecological Society, Polish Neonatological Society, Polish

Ultrasonographic Society, European Research Group on Feeding in Perinatology);

- „*Cardiology and Oncology in Poland at the beginning of 21st century. Status, challenges and perspectives"* (organizer: Z. Religa and M. Nowacki; Institute of Cardiology, Institute of Oncology);
- „*The system of values and demographic processes" (*organizer: K. Slany; Jagiellonian University);
- „*Methodological Basis of demographic research*" (organizer: Committee of Demographic Sciences (Polish Academy of Sciences), Polish Demographic Society in cooperation with Central Statistical Office in Poland);
- „*Family and family policy"* (organizer: B. Balcerzak-Paradowska; Institute of Labour and Social Affairs);
- „*Family and household in big agglomerations*" (organizer: S. Wierzchosławski; Centre of Family Research Academy of Economics in Poznań);
- „*Social consequences of population ageing"* (organizer: L. Frąckiewicz and A. Rączaszek; Social Policy Division, Academy of Economics in Katowice);
- „*The position of Poland in Europe in solving the problems of prophylactics, early detection and treatment of the gynaecological cancer*" (organizer: J. Zieliński; Polish Gynaecological Oncology Society in cooperation with Warsaw Division of Polish Gynaecological Society);
- „*The process of population ageing. Needs and challenges*" (organizer: J.T. Kowaleski; demographic Division University of Łódź, Committee of Demographic Sciences (Polish Academy of Sciences);
- „*Early pregnancy losses*" (organizer: B. Chazan and K. Czajkowski; Mother and Child Institute);
- „*Family of the Wielkopolski Region*", Trends in changes of attitudes and procreative behaviours (organizer: S. Wierzchosławski; Centre of Family Research, Academy of Economics in Poznań);
- „*Polish family at the beginning of New Millennium*" (organizer: W. Majkowski, University of Cardinal Stefan Wyszyński).

Part III was ***a series of regional seminars devoted to demographic problems of Poland with regard to regions***. Problems were discussed which had been proposed during the first session of the Congress, (regional problems included), as well as some other topics particularly important from the point of view of individual regions of the country:

- "*Impact of European Integration on Structural Transformation of Regions of High Unemployment"* (organizer: E.Z. Zdrojewski; Technical University in Koszalin in cooperation with Polish Demographic Society and Polish Economic Society);
- "*Contemporary and perspective problems of demographic development in Northern Poland*" (organizer: D. Śledź; Central Statistical Office of Poland);

- "*Demographic transformation in Świętokrzyskie voivodship*" (organizer: J. Śledzianowski);
- "*Role of local governments in demographic development*" (organizer: J. Hrynkiwicz, Local Government in Skierniewice);
- "*Current demographic problems and chances for socio-economic development in voivodships South-East Poland*" (organizer: J. Witkowski and A. Włoch; Statistical Office in Kielce);
- "*Demographic processes in European Union – comparison with Poland*" (organizer: I. Sobczak; Polish Demographic Society);
- "*Handicapped person in local society"* (organizer: SIKOŃ Foundation for Disabled Children; 15—16 October 2002).

Part IV, ***summing up the developments of the Congress***, consisted in a two-day plenary closing session. Closing ceremony organized by Governmental Population Council, Committee of Demographic Sciences (Polish Academy of Sciences), Polish Demographic Society, Central Statistical Office of Poland and United Nations Development Programme took place in Warsaw, 25-26 November 2002.

Representatives of the highest state authorities had been invited to participate in the Congress: the Government, Speakers of the Parliament and the Senate, people from medical and research environments, trade unions, politicians and representatives of the Polish Roman Catholic Church.

The Honorary Committee of the First Demographic Congress consisted of: The Prime Minister, the Speakers of the Parliament and the Senate and H.E Cardinal Józef Glemp, the Primate of Poland.

During time of the I. Demographic Congress and after on the base of workshops, seminars have been published series of the I. Congress Proceedings, together 17 volumes, devoted to different aspects of demographic processes in Poland.

Prepared by
Ewa Fratczak, Zbigniew Strzelecki
The Governmental Population Council,
Warsaw School of Economics

At the end of the First Demographic Congress in Poland, its participants adopted the following declaration:

DECLARATION OF I DEMOGRAPHIC CONGRESS IN POLAND

„POLAND AND EUROPE , DEMOGRAPHIC PROCESSES AT THE BEGINNING OF THE 21ST CENTURY' (26 November 2002)

Participants of the I Demographic Congress in Poland acknowledging that: generation continuity and the position of our nation and the country in the world are the issues of fundamental importance to our society, the objectives of sustained development and the 'UN Millennium Development Goals' constitute the fundamental principles of our country's development, beneficial demographic changes have occurred over the recent years; infant, children and youth mortality have dropped, life expectancy has grown; however: at the beginning of the 21st century, Poland is facing some serious demographic problems: decline in births, still high mortality, life expectancy at a lower level than that observed in the EU countries, bad health status of population, difficulties encountered in the process of family formation and dissolution, immediate prospect of decrease in population and serious destabilisation of the effective intergenerational balance.

Participants of I Demographic Congress call upon: Polish society with the advice of the existing grave hazards to the biological potential of the Nation the government, and civic organizations, the Catholic Church and religious organizations with a postulate to take some action in the areas of social and population policies, which might arrest the unfavourable demographic trends.

We assume it most important to:

- support such systems of values which add to the ethos of family and marriage in Polish society,
- spread the conviction that the State and the citizens bear the responsibility for the demographic future,
- change the directions of economic and social policies so that they allow for demographic issues,
- realise economic programmes bound to ensure satisfactory pace of economic growth, give the young generation job opportunities, decent pays and housing,
- increase involvement of the government in activities related to children upbringing,
- improve the quality of human potential through providing suitable conditions for education at all levels and in all forms,
- realise a coordinated public health policy, based on the principles of the National Health Programme,

- improve security through overcoming the crime hazard, and especially eliminating its causes; radically improve the traffic and railway security,
- work out a long-range migration policy, compatible with the goals of demographic and socio-economic development,
- use the results from the National Census as data resource to create programmes of action at different levels of country's administration — for the benefit of population and family development,
- respect equal rights and human dignity of women and men, goals and principles included in the Charter of the United Nations and the Universal Declaration of Human Rights and other international legal acts on human rights,
- the demographic change necessitates a broad social debate on its essence and consequences, development of research and widespread demographic education.

Participants of the I Demographic Congress in Poland acknowledge that the Congress:

- has substantially contributed to understanding and assessment of the current demographic processes,
- contributed to significant expansion and integration of the circles which are occupied with demographic issues,
- created the proper background for dissemination of demographic knowledge and its significance in the country,
- attracted kind attention of the highest State authorities, most prominent of them being the President of the Republic of Poland, and it has convinced them of the need to include these issues in social and economic policies of the State.

Participants of the I Demographic Congress in Poland call upon the Congress organizers to make the best use of the rich scientific materials and the recommendations included in the Declaration as well as those submitted during the Congress debates.

Participants of the I Demographic Congress in Poland recommend creating suitable conditions for organising the next Demographic Congress three years from now.

President of the I. Demographic Congress
Zbigniew Strzelecki

STATISTICS IN TRANSITION, December 2003
Vol. 6, No. 4, pp. 621—625

Conference on Survey Sampling in Economical and Social Surveys

Łódź, Poland, November 3—5, 2003

The conference was organised by Department of Statistics of University of Economics in Katowice, Department of Statistical Methods of University of Łódź and Polish Statistical Association. It was the meeting of Polish statisticians who deal with survey sampling. The scope of the conference covered theory and application of survey sampling methods. The following topics were discussed.

The problems connected with **inference on characteristics of small areas** were considered in several papers. Jan Kordos in the paper *My involvement and troubles with small area statistics*, presented his engagement in small area statistics in Poland and abroad, and analysed recent progress in small areas design and estimation. He confronted it with the quality statistics components accepted by the Eurostat, such as *relevance, accuracy, timelines and punctuality, accessibility and clarity, comparability and coherence*. Different sources of information used in small area estimation were discussed. The author's conclusion was that special attention should be paid to the study design stage, so as to be able to take into greater consideration the user's demands as well as the quality of the obtained estimates.

Tomasz Jurkiewicz and Krzysztof Najman prepared the paper *Analysis of efficiency of modified synthetic estimator using bootstrap methods.* In this paper a two-stage estimation procedure was suggested. The first stage consists in applying various classification methods to identify the degree of similarity between the sample units from the investigated domain and sample units representing other domains. In the second stage, those domains with units similar to units from domain of interest are used to provide sample information with specially constructed weights. On the basis of them the modified synthetic estimator is constructed and its accuracy is analysed on the basis of a sample survey of Continuos Vocational Training conducted in Poland in 2003.

Tomasz Żądło prepared the paper *On synthetic regression estimator based on superpopulation approach*. In the paper properties of the synthetic regression estimator of domain total is considered. The proof of its ξ-unbiasedness for regression superpopulation model in strata is given. The BLU predictor is also presented for this model. The variances of both estimators are derived. The

problem of model misspecification is considered, too. The comparison of accuracy is supported by simulation study.

The following papers were devoted to estimation of population parameters on two or more occasions. In the paper Estimating the mean in rotation designs Jacek Wesołowski considered the problem of inference on population mean on more than two occasions under the classical exponential correlation model approach. On the basis of A. Szarkowski's simulation studies, developed during practical survey sampling research, the following estimator was proposed. The optimum estimator can be dependent on data observed in the current study and in the two preceding studies. For particular rotation sampling designs an analytical formula for an optimum mean estimator in its iterative form was found. The discussed algorithm was implemented in the SAS system. The obtained numerical results are better than the results for the classical estimator.

Barbara Kowalczyk in the paper *Classical and composite estimator of the dynamics of change* occupied with the problem of how to assess the dynamics of change ratio in two successive periods of time in the case of rotating surveys. The estimator being a quotient of two regression estimators, under specified rotation pattern design was analysed and compared with the classical estimator. The conditions for the proposed estimator to be better then the classical one were given. The presented theoretical results were also illustrated by simulation studies.

Krystyna Pruska in the paper *A representative study of small areas on the basis of time cross-section samples* occupied with repeated representative study. The paper contains some proposals on how to arrive at a conclusion in the case of joint samples from different periods of time. The cases of identical and non-identical distribution of the studied variables in these periods are considered. The aim of such a procedure is to look for the estimators characterised by a variance, which is small when compared to their expected value.

In the paper *The way of organising representative studies in the causative analysis* Anna Bartkowiak tried to present some problems connected with planning a causative study. One of the major problems of the causative analysis is how to determine the span of time after the occurrence of a cause. In the paper, several proposals of how to solve some questions connected with causal analysis were discussed.

The problems of **optimal evaluation of sample sizes and related topics** were analysed, too. Bronisław Lednicki and Robert Wieczorkowski wrote the paper *The problem of optimal stratification and sample allocation in sub-populations and strata.* In the paper, a method of defining optimal borders between strata through a simple stratified sampling, which investigates all the upper stratum (since the distribution of the characteristics under study often tends to be asymmetrically right), was presented. A solution of the problem of optimal stratification by means of numeral optimisation methods with the use of appropriately thought of multiple variable function was also proposed. This

approach was already implemented in the R system. The method presented here was used to solve the problem of allocating the samples in sub-populations with the same precision. The approach described was actually used in one survey carried out by the Central Statistical Office. An ex-post analysis of the efficiency of the solutions was carried out.

Elżbieta Gołata and Grażyna Dehnel, in their paper *The problems of optimal allocation of a sample in estimating small areas* presented some empirical results together with their comments, results of a research of sample allocating in estimating small domains. The several optimisation procedures were taking into account. The research was part of the EURAREA project.

The paper *On Optimal Sample Allocation in Multivariate Surveys Using Semidefinite Programming* by Marcin Skibicki analysed the problem of sample sizes determination in stratified sampling. The case of simultaneous estimation of many means was considered. Minimisation of the spectral radius of covariance matrix of mean estimators' vector subject to a restricted survey cost was considered. This is a non-linear program, but can be also cast as a samidefinite program and efficiently solved by one of the semidefinite programming methods, i.e. interior point method.

Problems connected with **parameter estimation in the case of presence of non-sampling errors** are still important. That is why we can still expect many contributions in this topic. In the paper *Coherence measures in control surveys* by Paweł J. Szabłowski a certain measure of coherence between primary census and control census was presented. Its properties were discussed. The measure was applied to analyse the results of control microcensus for National Population Census (2002) and National Agricultural Census (2002).

Beata Zmyślona in the paper *Bayesian methods of incomplete data analysis* considered usefulness of auxiliary information, which is not obtained from the sample, and a priori knowledge on reasons of nonresponse in the inference problems. Bayesian methods of incomplete data analysis were presented. The example of incomplete categorical data analysis were considered.

Wojciech Gamrot in the paper *On some ratio estimators in two-phase sampling for nonresponse* occupied with improving estimation accuracy of mean value on the basis of call-back technique. It was assumed that the nonresponse is stochastic. Some modification of the well-known estimator was proposed and analysed.

In the paper On reduction of non-response bias through weighting survey data by means of an auxiliary variable's distribution function Janusz Wywiał occupied with estimation of a population mean in the case when a non-response error is present. The individual response probability model was considered. A simple sample was partitioned into sub-samples. It was assumed that each sub-sample consists of at least one unit that responds. The partition of an entire sample into sub-samples leads to clustering a population into sub-populations with assistance of an auxiliary variable. Two conditional estimators of the

population mean were considered. They are functions of a linear combination of the means from the sub-samples. Variances of estimators were derived. It was shown that response bias in comparison to the bias of the common simple sample mean is better reduced by one of the proposed estimator.

Some **computer simulation techniques** are used with success to study properties of rather complex sampling strategies. In the paper *L-Moment-Based Estimators Sampling Properties for Samples from Finite Populations* by Jiří Vohlídal and Libor Malenovský results of Monte-Carlo study aimed at a comparison of the estimators of the selected variability, relative variability, skewness and kurtosis measures, which are based on classical moments, quantiles and L-moments were presented. According to obtained results, the L-moment-based characteristics can be suggested as a good alternative to the classical moment-based and/or quantile-based characteristics.

Martin Zelený in the paper T*he use of Monte Carlo simulation in studying geographic properties of sample design* considered possibilities of using simulation techniques in evaluation of geographic properties of household survey designs. Two-stage designs with variations in number of units sampled per PSU were analysed and empirical distributions of number of municipalities to be visited were developed. The performance of two-stage designs in reducing the number of municipalities to be covered was evaluated.

Some **empirical survey sampling research** is usually inspiration of new study areas. Problem of inference in survey sampling is connected with evaluation of distribution functions of a statistics. In the case of randomised approach only limit distributions can be taken into account. But classical central theorems deal with simple sample. In the paper *Some remarks on border theorem in the case of repealing the sample's independence* Czesław Domański and Tomasz Zielonka presented some central theorem under non-classical assumptions. They suggested how to use these theorems in practical research.

Jarosław Górniak and Jolanta Perek-Białas in the paper *Analysis of the data in complex samples in sociological research* discussed practical consequences of different complex sampling schemes in the analysis of the data from social research. If some complex samples are drawn, this imposes an appropriate estimation procedure. In this context usefulness of a new statistical package is analysed

In the paper *Representative method in practical financial auditing - The results of a survey of auditors* Bartłomiej Janusz presented the results of a survey of Polish auditors conducted in the years 2002 - 2003. It shows the auditors' opinions on statistics' usefulness at their work and on the need to conduct further statistical research. The paper demonstrated how the auditors shape the sample to be studied and how often they use statistical methods to achieve conclusions referring to the whole population on the basis of a sample. The obtained results were compared with the results of a survey of American auditing firms conducted in the State of New York in 1992.

Grzegorz Krzykowski in the report *EU accession referendum – Exit poll survey* presented the description of the survey related to national referendum on Republic of Poland European Union accession. The description consists of theories and technical details for such surveys. Some specific problems related to non-response phenomenon and their solutions were presented.

Prepared by Janusz Wywiał