ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 194, 2005

*Krystyna Pruska**

TESTS FOR RATIO OF TWO MEANS IN CASE OF SMALL AREAS

Abstract

The relations between characteristics for subpopulations and for the whole population are very important in small area investigations.

In the paper there are proposed testing procedures for verification of hypothesis which says that there is no difference between the ratio of small area mean and population mean for analysed variable and auxiliary variable. The properties of one considered procedure are investigated with the use of simulation methods.

Key words: small area, synthetic estimator.

I. INTRODUCTION

In small area statistics estimation of unknown parameters for subpopulation is considered as general problem. Different estimators are constructed and applied and their properties are investigated. The synthetic estimators are considered, too. They can be used when some assumptions are true. In this paper we consider possibility of verification whether these assumptions are fulfilled.

II. SYNTHETIC ESTIMATORS

In statistical literature different definitions of synthetic estimators are given (see: Dol, 1991; Bracha, 1996; Särndal et al., 1997; Kordos, 1999, Domański and Pruska, 2001). Generally, their construction is possible when some relations between parameters for subpopulation and population are constant.

* Professor, Chair of Statistical Methods, University of Łódź.

In the paper there are considered finite population only.

We will introduce the following notations:

Y – investigated variable,

X – auxiliary variable,

TY_P – total value of variable Y for population,

TX_P – total value of variable X for population,

$TY_{.g}$ – total value of variable Y for stratum g of population,

$TX_{.g}$ – total value of variable X for stratum g of population,

TY_{hg} – total value of variable Y for stratum g and small area h of population,

$TX_{hg.}$ – total value of variable X for stratum g and small area h of population,

$TY_{h.}$ – total value of variable Y for small area h of population,

where $g = 1, \ldots, G$; $h = 1, \ldots, H$ and G is number of strata in the population, H is number of small areas in the population.

If we assume (see Dol, 1991):

$$\frac{TY_{.g}}{TX_{.g}} = \frac{TY_{hg}}{TX_{hg}} \quad \text{for} \quad g = 1, \ldots, G \quad \text{and} \quad h = 1, \ldots, H \tag{1}$$

then we can consider the synthetic estimator of total value of variable Y for small area h of the following form:

$$\hat{T}Y_{h.} = \sum_{g=1}^{G} \hat{\beta}_g TX_{hg} \tag{2}$$

where $\hat{\beta}_g$ is estimator of value $\frac{TY_{.g}}{TX_{.g}}$. There are different forms of statistic $\hat{\beta}_g$ (see: Dol, 1991).

Estimator (2) can be used when assumptions (1) are fulfilled. In empirical investigations we ought to verify whether the conditions are true. We may apply the estimator of TX_{hg} instead TX_{hg} in formula (2).

III. FORMULATING OF HYPOTHESIS

We can consider verification of possibility of the use of synthetic estimator as the verification of a suitable statistical hypothesis. Synthetic estimator (2) is constructed for the population which is divided into strata. Assumptions

(1) can be verified on the basis of independent samples drawn from each strata. All hypotheses which have the form:

$$H_{0g}: \frac{TY_{.g}}{TX_{.g}} = \frac{TY_{hg}}{TX_{hg}} \tag{3}$$

for $g = 1, \ldots, G$ and for the fixed h from the set $\{1, \ldots, H\}$ against an alternate hypothesis:

$$H_{1g}: \sim H_{0g} \tag{4}$$

can be verified analogously as hypothesis:

$$H_0: \frac{TY_P}{TX_P} = \frac{TY_{MO}}{TX_{MO}} \tag{5}$$

against hypothesis:

$$H_1: \sim H_0, \tag{6}$$

where TY_{MO} and TX_{MO} are total values for variables Y and X for fixed small area.

In this paper we will consider the following equivalent form of hypothesis H_0:

$$H_0: \frac{\mu_{YP}}{\mu_{XP}} = \frac{\mu_{YMO}}{\mu_{XMO}} \tag{7}$$

where μ_{YP}, μ_{XP}, μ_{YMO}, μ_{XMO} are means for variables Y and X for population and for small area, respectively.

IV. TEST PROCEDURES

Test statistic for the verification of hypothesis (7) can be random variable:

$$Z = \frac{\overline{Y}_P}{\overline{X}_P} - \frac{\overline{Y}_{MO}}{\overline{X}_{MO}} \tag{8}$$

or its function where $\overline{Y}_P$, $\overline{X}_P$, $\overline{Y}_{MO}$, $\overline{X}_{MO}$ are sample means for variables Y and X for population and for small area, respectively.

Determining the distribution of statistic Z is difficult when we do not know the distribution of variables Y and X.

Now the test procedure for verification of hypothesis (7) will be proposed. In this procedure the conditional distributions (for fixed values of statistics $\overline{X}_{MO}$, $\overline{X}_{P-MO}$ or $\overline{Y}_{MO}$, $\overline{X}_{P-MO}$ respectively) of the following statistics are used:

$$U_Y = \frac{\overline{Y}_{MO} - \frac{\overline{X}_{MO}}{\overline{X}_{P-MO}} \cdot \overline{Y}_{P-MO}}{\sqrt{\frac{S^2_{YMO}}{NMO} + \frac{\overline{X}^2_{MO}}{\overline{X}^2_{P-MO}} \frac{S^2_{YP-MO}}{NPMO}}} \tag{9}$$

and

$$U_X = \frac{\overline{X}_{MO} - \frac{\overline{Y}_{MO}}{\overline{Y}_{P-MO}} \cdot \overline{X}_{P-MO}}{\sqrt{\frac{S^2_{XMO}}{NMO} + \frac{\overline{Y}^2_{MO}}{\overline{Y}^2_{P-MO}} \frac{S^2_{XP-MO}}{NPMO}}} \tag{10}$$

where

$\overline{Y}_{P-MO}$, $\overline{X}_{P-MO}$ are sample means for set which is difference between population and small area and for variables Y and X, respectively;

S^2_{XMO}, S^2_{XP-MO} are sample variances for variable X for small area and for set which is difference between population and small area, respectively;

S^2_{YMO}, S^2_{YP-MO} are sample variances for variable Y for small area and for set which is difference between population and small area, respectively;

NMO is the number of these elements of sample from population, which belong to small area;

$NPMO$ is the number of these elements of sample from population, which do not belong to small area.

The test algorithm is the following:

1. We draw independently NP-element sample from the whole population. The elements of the sample belonging to small area are the sample for the small area and the elements which do not belong to small area are the sample for set which is difference between population and small area, respectively.

2. We determine the value of the following statistics: $\overline{X}_{MO}$, $\overline{X}_{P-MO}$, S^2_{XMO}, S^2_{XP-MO}, $\overline{Y}_{MO}$, $\overline{Y}_{P-MO}$, S^2_{YMO}, S^2_{YP-MO}, U_X, U_Y.

3. We verify whether $|u_X| \geqslant 1{,}96$ or $|u_Y| \geqslant 1{,}96$ where u_X and u_Y are values of variables U_X, U_Y, respectively. If one inequality is not true then we reject hypothesis (7).

We can notice that:

$$P(|U_X| \geqslant 1.96 \text{ or } |U_Y| \geqslant 1.96) = P(|U_X| \geqslant 1.96) + P(|U_Y| \geqslant 1.96) - P(|U_X| \geqslant 1.96 \text{ and } |U_Y| \geqslant 1.96) \leqslant P(|U_X| \geqslant 1.96) + P(|U_Y| \geqslant 1.96) = 0.05 + 0.05 = 0.1,$$

when we consider the probability for suitable conditional distribution of statistics UX and UY which are asymptotic normally.

If we know total values of auxiliary variable for population and small area (i.e. we know TX_P and TX_{MO}) then we can use the following random variable as test statistic:

$$U = \frac{\overline{Y}_{MO} - \frac{TX_{MO}}{TX_{P-MO}} \cdot \overline{Y}_{P-MO}}{\sqrt{\frac{S^2_{YMO}}{NMO} + \frac{TX^2_{MO}}{TX^2_{P-MO}} \frac{S^2_{YP-MO}}{NPMO}}}. \tag{11}$$

Asymptotic distribution of statistic (11) is normal $N(0; 1)$. In this case rejection region of hypothesis (7) is determined in the classic way. We reject hypothesis when the value of statistic (11) calculated on the basis of sample belongs to the rejection region.

V. MONTE CARLO ANALYSIS OF PROPOSED TEST PROCEDURE PROPERTIES

Monte Carlo analysis deals with the first presented test procedure, it means the case when the total values of auxiliary variable for population and small area are unknown.

The aim of the conducted experiments was determining the number of cases in which the hypothesis (7) was rejected in 1000 repetitions for fixed distribution of population.

The experiments were conducted in the following way:

1. Creating the population consisting of 50 000 values of variable (Y, X) which are generated from fixed distribution or two fixed distributions.
2. Determining the small area which consists of 5000 elements.
3. Drawing n-element sample from population ($n = 2000, 2500$).

4. Verification of hypothesis (7) by means of test whose statistics are variables (9) and (10).

5. Conducting 1000 repetitions of stages 3. and 4.

6. Determining the number of cases in which the hypothesis (7) was rejected in 1000 repetitions.

The results of Monte Carlo experiments are presented in Tables 1. and 2.

Table 1. Number of cases of rejection of hypothesis (7) among 1000 experiments for the same distribution of (Y, X) in population and small area

No.	Distribution of variable (Y, X)	Size of sample from population	Minimal size of sample from small area	Average size of sample from small area	Maximal size of sample from small area	Number of rejection of H_0
1	$N(m, \Sigma)$, $m^T = [100; 20]$ $\Sigma = \begin{bmatrix} 100 & 36 \\ 36 & 16 \end{bmatrix}$	2000 2500	156 207	200 250	251 299	123 113
2	$N(m, \Sigma)$, $m^T = [10; 20]$ $\Sigma = \begin{bmatrix} 1 & 0{,}8 \\ 0{,}8 & 1 \end{bmatrix}$	2000 2500	156 207	200 250	251 299	136 140
3	$X \sim N(10; 1)$ $Y = [X]$	2000 2500	149 196	200 250	247 305	0 0
4	$X \sim N(60; 12)$ $Y = [X]$	2000 2500	149 196	200 250	247 305	0 0
5	$X \sim P_5$ $Y = X + Z$ $Z \sim N\left(\frac{5}{2}; \frac{\sqrt{5}}{10}\right)$	2000 2500	161 188	200 250	252 312	0 1
6	$X \sim P_{10}$ $Y = X + Z$ $Z \sim N\left(5; \frac{\sqrt{10}}{10}\right)$	2000 2500	158 209	200 250	253 293	0 0

Source: Author's calculations.

Table 2. Number of cases of rejection of hypothesis (7) among 1000 experiments for the different distributions of (Y, X) in population and small area

No.	Distribution of variable (Y, X) in small area	Distribution of variable (Y, X) out of small area	Size of sample from population	Minimal size of sample from small area	Average of sample from small area	Maximal size of sample form small area	Number of cases of H_0
1	$N(m, \Sigma)$, $m^T = [120; 25]$ $\Sigma = \begin{bmatrix} 100 & 36 \\ 36 & 16 \end{bmatrix}$	$N(m, \Sigma)$, $m^T = [100; 20]$ $\Sigma = \begin{bmatrix} 100 & 36 \\ 36 & 16 \end{bmatrix}$	2000 2500	156 207	200 250	251 299	120 113
2	$N(m, \Sigma)$, $m^T = [12; 21]$ $\Sigma = \begin{bmatrix} 1 & 0{,}8 \\ 0{,}8 & 1 \end{bmatrix}$	$N(m, \Sigma)$, $m^T = [10; 20]$ $\Sigma = \begin{bmatrix} 1 & 0{,}8 \\ 0{,}8 & 1 \end{bmatrix}$	2000 2500	156 207	200 250	251 299	1000 1000
3	$X = U + 2$ $Y = V + 1$ $U \sim N(10; 1)$ $V = [U]$	$X \sim N(10; 1)$ $Y = [X]$	2000 2500	149 196	200 250	247 305	1000 1000
4	$X = U + 2$ $Y = V + 1$ $U \sim N(60; 12)$ $V = [U]$	$X \sim N(60; 12)$ $Y = [X]$	2000 2500	149 196	200 250	247 305	0 0
5	$X = U + 2$ $Y = V + 1$ $U \sim P_5$ $V = U + Z$ $Z \sim N\left(\frac{5}{2}; \frac{\sqrt{5}}{10}\right)$	$X \sim P_5$ $Y = X + Z$ $Z \sim N\left(\frac{5}{2}; \frac{\sqrt{5}}{10}\right)$	2000 2500	161 188	200 250	252 312	1000 1000
6	$X = U + 2$ $Y = V + 1$ $U \sim P_{10}$ $V = U + Z$ $Z \sim N\left(\frac{5}{2}; \frac{\sqrt{5}}{10}\right)$	$X \sim P_{10}$ $Y = X + Z$ $Z \sim N\left(\frac{5}{2}; \frac{\sqrt{5}}{10}\right)$	2000 2500	158 209	200 250	253 293	1000 1000

Source: Author's calculations.

We can notice that the considered test procedure does not reject hypothesis (7) or rejects in a few cases among 1000 repetitions when hypothesis is true. If hypothesis (7) is not true and distributions of investigated variable in population and small area differ significantly then hypothesis (7) is rejected in all repetitions for a given case. If the distributions do not differ significantly then hypothesis (7) is rejected in some repetitions or in no repetitions. Such results are not typical for significance tests. The proposed test procedure can be modified and its properties ought to be investigated.

VI. FINAL REMARKS

Different methods for the verification of hypothesis about relations between total values for subpopulation and population can be constructed. Some propositions are presented in the paper, another one (using bootstrap method) is given in the Pruska's paper (2002). The problems, considered in the paper, are important in investigations of small area. It seems necessary to continue the conducted analyses.

REFERENCES

Bracha Cz. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, PWN, Warszawa.

Dol W. (1991), *Small Area Estimation. A Synthesis between Sampling Theory and Econometrics*, Wolters Noordhoff, Groningen.

Domański Cz., Pruska K. (2001), *Metody statystyki małych obszarów*, Wyd. Uniwersytetu Łódzkiego, Łódź.

Kordos J. (1999), Problemy estymacji danych dla małych obszarów, *Wiadomości Statystyczne*, **1**, 85–101.

Pruska K. (2002), Statystyczna weryfikacja możliwości zastosowania estymatorów syntetycznych w badaniach małych obszarów, referat wygłoszony w Łagowie na konferencji „Statystyka regionalna w jednoczącej się Europie" (2–5.09.2002).

Särndal C., Swensson B., Wretman J. (1997), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Krystyna Pruska

TESTY DLA STOSUNKU DWÓCH ŚREDNICH W PRZYPADKU MAŁYCH OBSZARÓW

Streszczenie

Relacje pomiędzy charakterystykami podpopulacji i całej populacji są bardzo ważne w badaniach małych obszarów.

W pracy tej zaproponowane są procedury testowe, służące do weryfikacji hipotezy o równości stosunku średniej dla małego obszaru do średniej z populacji dla analizowanej zmiennej i zmiennej pomocniczej. Własności jednej z zaproponowanych procedur badane są za pomocą metod symulacyjnych.