ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

*Wojciech Gamrot**

ON SOME ESTIMATOR OF FINITE POPULATION SKEWNESS UNDER NONRESPONSE

Abstract. One of the most popular measures for the assymetry of distribution is the coefficient of skewness computed by standarizing the third central moment about the mean. In this paper the well-known two-phase sampling procedure is applied to estimate the finite population skewness under nonresponse. An estimator of this parameter is constructed as a function of well-known unbiased estimators of population totals. The properties of proposed estimator are investigated by the simulation study. The data obtained from agricultural census in boroughs of the Dąbrowa Tarnowska district is used in simulations.

Key words: finite population skewness, incomplete data, estimation.

I. INTRODUCTION

Consider finite population U=$\{u_1,...,u_N\}$. Let X be some population characteristic taking fixed values $x_1,...,x_N$. This paper focuses on the estimation of population skewness coefficient given by formula:

$$\lambda_3 = \frac{M_3}{M_2^{3/2}}$$

where:

$$M_r = \frac{1}{N}\sum_{i\in U}(x_i - \overline{X})^r$$

and $\overline{X} = N^{-1}\sum_{i\in U} x_i$. In order to estimate this parameter a simple random sample s of size n is drawn without replacement from U. It is assumed, that nonresponse appears in the survey and as a result the sample splits into two subsets s_1 and s_2 of sizes n_1 and n_2 such that units from s_1 respond while units from

* Ph.D., Department of Statistics, University of Economics, Katowice.

s_2 do not. Data incompleteness is treated as a random phenomenon described by means of response distribution $q(s_1|s) = q(s_2|s) = q(s_1,s_2|s)$ determining individual response probabilities $\rho_{i|s} = \sum_{s_1 \ni i} q(s_1 \mid s)$ (see. Cassel et al. 1983). To compensate for nonrespondent underrepresentation another phase of the survey is carried out with simple subsample s' of size $n'=cn_2$ being drawn without replacement from s_2. It is assumed that all subsampled units respond when re-contacted.

II. ESTIMATION

Let us consider the population total of the h-th power of X defined by the following expression:

$$t_h = \sum_{i \in U} x_i^h$$

Let us also consider the double-sample-based statistic:

$$\hat{t}_h = \frac{N}{n}\left(\sum_{i \in s_1} x_i^h + \frac{1}{c}\sum_{i \in s'} x_i^h\right)$$

Särndal et al (1992) shows that it is unbiased for t_h irrespective of the underlying response distribution, and provides the formula for the variance of $\hat{t}_h$ as well as its unbiased estimator. In particular for $h = 0$ we have $t_h = N$ and $\hat{t}_0$ is an unbiased estimator of n.

In order to construct the estimator of skewness coefficient, let us express this parameter as the following function of population totals:

$$\lambda_3 = \frac{t_0^2 t_3 - 3t_0 t_1 t_2 + 2t_1^3}{(t_0 t_2 - t_1^2)^{3/2}}.$$

Now let us replace unknown totals with their respective unbiased estimators computed from the two-phase sample. Hence, we obtain the following estimator of λ_3:

$$\hat{\lambda}_{2F} = \frac{\hat{t}_0^2 \hat{t}_3 - 3\hat{t}_0 \hat{t}_1 \hat{t}_2 + 2\hat{t}_1^3}{(\hat{t}_0 \hat{t}_2 - \hat{t}_1^2)^{3/2}}.$$

A simulation study was carried out to assess the properties of the proposed estimator and to compare it with the single-phase based estimator:

$$\hat{\lambda}_{UN} = \frac{\hat{t}_{0*}^2 \hat{t}_{3*} - 3\hat{t}_{0*}\hat{t}_{1*}\hat{t}_{2*} + 2\hat{t}_{1*}^3}{(\hat{t}_{0*}\hat{t}_{2*} - \hat{t}_{1*}^2)^{3/2}}$$

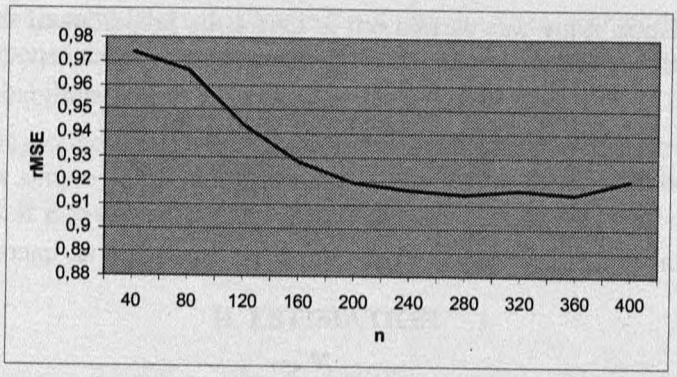
which is constructed by replacing each unknown total with an uncorrected estimator:

$$\hat{t}_{h*} = \frac{N}{n_1} \sum_{i \in s_1} x_i^h .$$

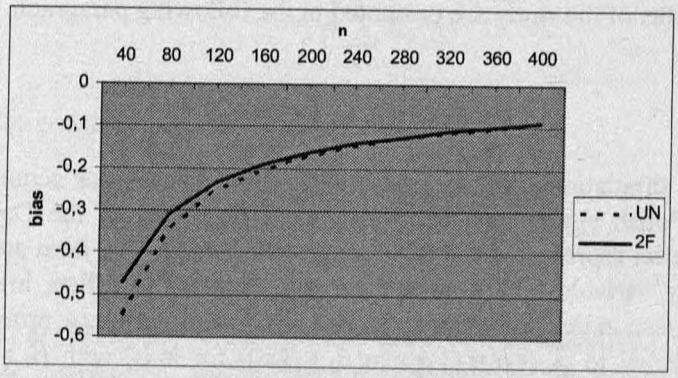
The results of the study are presented in the following paragraph.

III. SIMULATION RESULTS

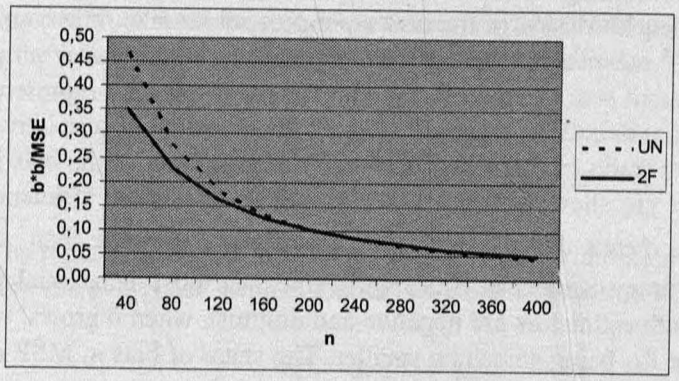
During simulations, the data describing 2420 households acquired during the 1996' Polish agricultural census in three boroughs of the Dąbrowa Tarnowska district represented the population under study. The farm area was used as the study variable. The nonresponse was assumed to follow logistic model with units responding independently and individual response probabilities respectively equal to $\rho_{i|s} = \rho_1 = (1 + \exp(\beta_0 + \beta_1 x_i))$ for $i \in U$, with β_0 and β_1 being arbitrarily chosen constants. Several sample – subsample pairs were repeatedly drawn from the population using the two-phase sampling design involving simple random sampling without replacement in both phases. The subsample size was always equal to 30% of the nonrespondent subset size. The empirical distribution of 10^6 estimates served as a basis for assessing estimator properties. The first experiment was carried out for $\beta_0 = 0$ and $\beta_1 = 0$ (this represents response probabilities unrelated to the characteristic under study). The relative efficiency of estimators (ratio of their MSE's), their bias and share of the bias in the mean square error are shown on pic. 1–3. The recorded relative efficiency is below unity which means that the estimator $\hat{\lambda}_{2F}$ is more accurate than $\hat{\lambda}_{UN}$, but the gain in accuracy is modest. The improvement is most substantial for $n > 200$. Biases of both estimators are negative and diminish when n grows, with the bias of estimator λ_{2F} being somehow smaller. The share of bias in MSE also quickly diminishes for both estimators.



Pic. 1. Relative efficiency for $\beta_0 = 0, \beta_1 = 0$

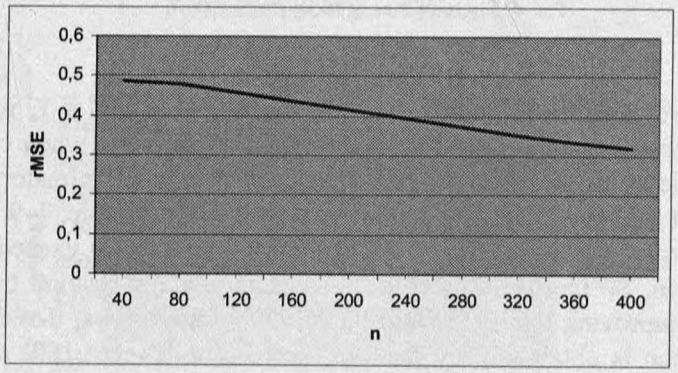


Pic. 2. Bias for $\beta_0 = 0, \beta_1 = 0$

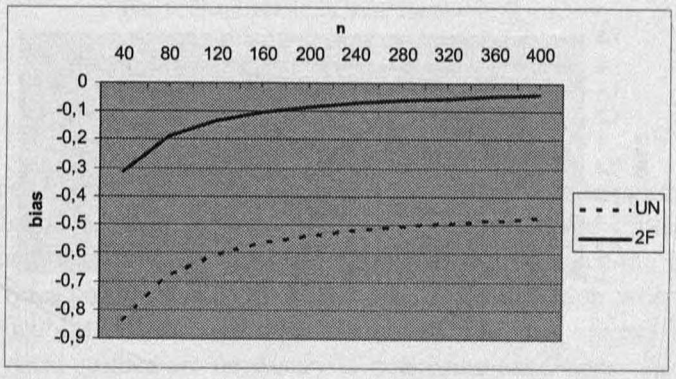


Pic. 3. Share of bias in MSE for $\beta_0 = 0, \beta_1 = 0$

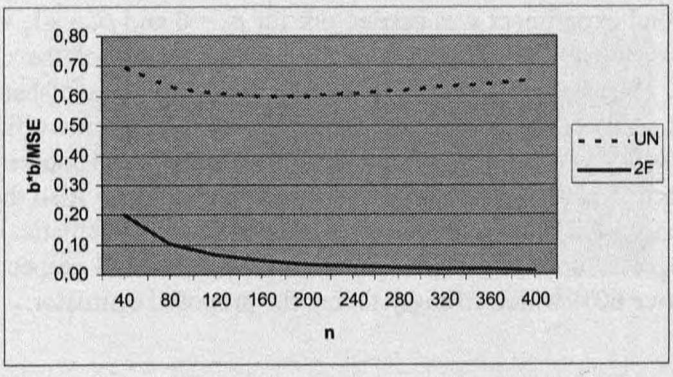
The second experiment was carried out for $\beta_0 = 0$ and $\beta_1 = -1$, which means that response probabilities increased with growing values of the characteristic under study. The relative efficiency of estimators, their bias and share of the bias in the mean square error are shown on pic. 4–6. Now relative efficiency takes values below 0.5 which indicates substantial gains in precision resulting from the use of second phase data, and it decreases when n grows. Also the bias of the two-phase-based estimator is significantly lower than the dramatically high bias of the uncorrected one. The shares of bias are at extremes – respectively below 20% and above 60% which strongly favors the proposed estimator.



Pic. 4. Relative efficiency for $\beta_0 = 0, \beta_1 = -1$

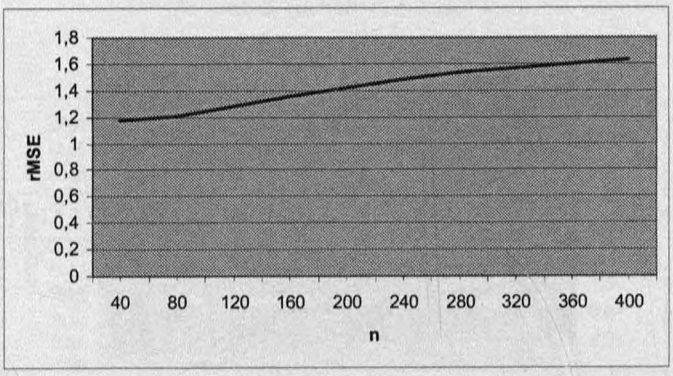


Pic. 5. Bias for $\beta_0 = 0, \beta_1 = -1$

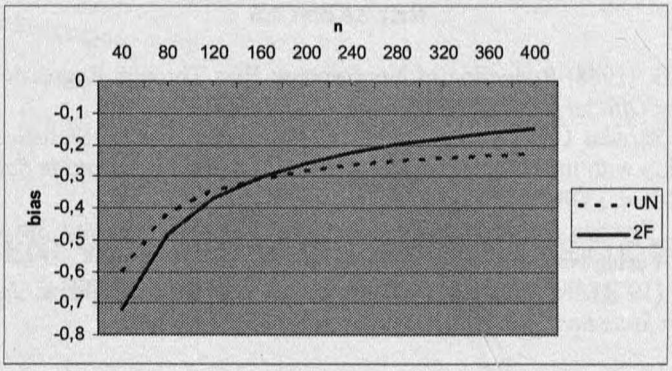


Pic. 6. Share of bias in MSE for $\beta_0 = 0, \beta_1 = -1$

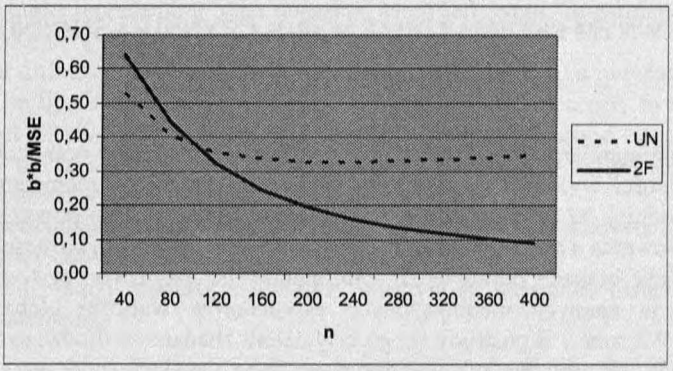
The third experiment was carried out for $\beta_0 = 0$ and $\beta_1 = 1$, which corresponds to the situation where response probabilities decrease when values of the characteristic under study grow. The relative efficiency of estimators, their bias and share of the bias in the mean square error are shown on pic. 7–9. The results are both surprising and disappointing: the relative efficiency exceeds unity and grows with n. So the use of second phase data actually distorts the estimates instead of improving them. The comparison of biases shows, that the effect of bias reduction is obtained only for quite large samples ($n > 160$) although for large samples the reduction is significant.



Pic. 7. Relative efficiency for $\beta_0 = 0, \beta_1 = 1$



Pic. 8. Bias for $\beta_0 = 0, \beta_1 = 1$



Pic. 9. Share of bias in MSE for $\beta_0 = 0, \beta_1 = 1$

IV. CONCLUSIONS

Presented simulation results indicate, that the two-phase sampling procedure has the potential to improve properties of estimates for the finite population skewness under nonresponse. It is particularly useful to reduce the nonresponse bias. However, special care should be taken in the situation when individual response probabilities are negatively correlated with the characteristic under study. Analytical studies are necessary to determine conditions where the two-phase sampling is preferable to other estimation methods..

REFERENCES

Bethlehem J.G. (1988) Reduction of Nonresponse Bias Through Regression Estimation *Journal of Official Statistics*, Vol. 4, No. 3, 251–160.

Cassel C.M. Särndal C.E. Wretman J.H. (1983), Some Uses of Statistical Models in Connection with the Nonresponse Problem [w:] *Incomplete Data in Sample Surveys* W.G. Madow, I.Olkin (red.), Academic Press New York.

Särndal C.E. Swensson B. Wretman J.H. (1992), *Model Assisted Survey Sampling* Springer Verlag New York.

Srinath K.P. (1971), Multiphase Sampling in Nonresponse Problems. *Journal of the American Statistical Association*, Vol. 66, No 335, 583–586.

Wojciech Gamrot

O PEWNYM ESTYMATORZE WSPÓŁCZYNNIKA SKOŚNOŚCI PRZY BRAKACH ODPOWIEDZI

Jedną z najpopularniejszych miar asymetrii rozkładu cechy w populacji jest współczynnik skośności wyznaczany poprzez standaryzację trzeciego momentu centralnego względem średniej. W niniejszej pracy rozważono wykorzystanie powszechnie znanej procedury losowania dwufazowego do szacowania współczynnika skośności w populacji skończonej przy brakach odpowiedzi. Zaproponowano estymator tego współczynnika będący funkcją znanych nieobciążonych estymatorów wartości globalnych cechy w populacji. Własności skonstruowanego estymatora zbadano w drodze symulacji komputerowych. W eksperymentach wykorzystano dane uzyskane podczas spisu rolnego w wybranych gminach powiatu Dąbrowa Tarnowska.