Stat Papers (2012) 53:887–894
DOI 10.1007/s00362-011-0392-3

REGULAR ARTICLE

Estimation of finite population kurtosis under two-phase sampling for nonresponse

Wojciech Gamrot

Received: 22 September 2010 / Revised: 21 June 2011 / Published online: 7 July 2011

Abstract In this paper an estimator of finite population kurtosis computed under the two-phase sampling for nonresponse is proposed. The formulas characterizing its asymptotic properties are derived using Taylor linearization technique for the general situation of arbitrary sampling designs in both phases and stochastic nonresponse represented by arbitrary response distribution. An important special case of simple random sampling without replacement and deterministic nonresponse is also considered.

Keywords Estimation · Kurtosis · Two-phase sampling · Non-response

Mathematics Subject Classification (2000) 62 · 62D05

1 Introduction

Several approaches have been proposed to deal with nonresponse bias including weighting adjustments (Bethlehem 1988; Ekholm and Laaksonen 1991), imputation methods (Rubin 1987; Schafer 1997; Meeden 2000), indirect estimation using auxiliary data (Rueda and González 2004; González et al. 2008) as well as randomized response techniques (Warner 1965; Chaudhuri 1987; Arnab 1998). Another procedure known as two-phase (or double) sampling relies on re-approaching some non-respondents in order to acquire a subsample of missing data which is needed to preserve unbiasedness. Possibilities of applying the two-phase sampling scheme to compensate for the nonresponse bias are thoroughly explored in the literature. This procedure is mostly considered in the context of estimating simple population parameters that

W. Gamrot (✉)
University of Economics in Katowice, Bogucicka 14, 40-226 Katowice, Poland
e-mail: wojciech.gamrot@ae.katowice.pl

may be expressed as linear combinations of population values such as the population total and population mean. Examples include recent papers of Okafor and Lee (2000), Sodipo and Obisesan (2007), Singh et al. (2010) as well as Singh and Kumar (2010). However, the two-phase sampling may also be adopted to estimate parameters which are defined as nonlinear functions of population values such as covariance, coefficient of variation or skewness. This paper focuses on constructing the estimators for finite population kurtosis on the basis of two-phase sample.

Let U be a finite population of size N. Let X be some fixed characteristic of population units taking values $x_1, \ldots, x_N$. Several population parameters may be defined including the r-th raw moment of X:

$$m_r = \frac{1}{N} \sum_{i \in U} x_i^r \tag{1}$$

and the r-th central moment of X about the mean:

$$M_r = \frac{1}{N} \sum_{i \in U} (x_i - m_1)^r \tag{2}$$

In particular this includes population total of the r-th power of X: $t_r = Nm_r$, population variance $S^2 = M_2$ and population standard deviation $S = M_2^{0.5}$. This paper focuses on the problem of estimating the dimensionless measure of distribution peakedness known as kurtosis and defined by the formula:

$$K = \frac{M_4}{M_2^2} \tag{3}$$

It is worth noting that sometimes the kurtosis is defined in a slightly different way, as K multiplied by some positive function of the sample size (Lütkepohl and Theilen 1991) or as a result of subtracting some constant from it, as in the case of *excess kurtosis* discussed by Zwillinger and Kokoska (2000). Anyway, the interpretation of so-defined kurtosis is the same: the higher value of K, the more leptokurtic is the distribution of X, and the lower K, the more platykurtic (flatter) is the distribution of X. The results presented in following paragraphs are easily applicable to the estimation of these quantities as well.

To estimate K the well-known sampling procedure attributed to Hansen and Hurwitz (1946) is adopted. In the first phase of the survey a random sample s of size n is drawn from U using some general sampling design $p(s)$ characterized by the set of inclusion probabilities of the first order $\pi_i = \sum_{s \ni i} p(s)$ and of the second order $\pi_{ij} = \sum_{s \ni i,j} p(s)$ for $i, j \in U$. Assume that some population units may fail to provide responses and corresponding values of X may remain unobserved. As a result the sample s splits into two subsets s_1 and s_2 of sizes n_1 and n_2, such that units from s_1 respond and units from s_2 do not. This may be described in terms of a probability distribution $q(s_1|s)$ known as *response distribution* (Cassel et al. 1983) determining individual response probabilities of the first order $\rho_{i|s} = \sum_{s_1 \ni i} q(s_1|s)$ and of the

second order $\rho_{ij|s} = \sum_{s_1 \ni i,j} q(s_1|s)$ for $i, j \in U$. To acquire knowledge about nonresponding units a second phase of the survey is then carried out. A subsample s' of size n' is drawn from s_2 according to another sampling design $p'(s'|s, s_2)$ characterized by another set of inclusion probabilities of the first order $\pi'_{i|s,s_2} = \sum_{s' \ni i} p'(s'|s, s_2)$ and second order $\pi'_{ij|s,s_2} = \sum_{s' \ni i,j} p'(s'|s, s_2)$. It is assumed that all units from s' respond in the second phase of the survey. As indicated by Lessler and Kalsbeek (1992) this assumption—although sometimes difficult to satisfy—is essential for the bias reduction.

2 Estimation of totals

We will now briefly review known results concerning the estimation of the population total. Let us consider the statistic:

$$\hat{t}_{r*} = \sum_{i \in s_1} \frac{x_i^r}{\pi_i} + \sum_{i \in s'} \frac{x_i^r}{\pi_i^*} \tag{4}$$

where

$$\pi_i^* = \pi_i \pi'_{i|s,s_2} \tag{5}$$

for $i \in U$. As indicated by Särndal et al. (1992) it is unbiased for $\hat{t}_r$ irrespective of underlying response distribution and its variance may be expressed as

$$V(\hat{t}_{r*}) = \sum_{i,j \in U} \frac{x_i^r x_j^r}{\pi_i \pi_j} \Delta_{ij} + E_{pq} \left(\sum_{i,j \in s_2} \frac{x_i^r x_j^r}{\pi_i^* \pi_j^*} \Delta'_{ij|s,s_2} \right) \tag{6}$$

where

$$\Delta_{ij} = \pi_{ij} - \pi_i \pi_j \tag{7}$$

$$\Delta'_{ij|s,s_2} = \pi'_{ij|s,s_2} - \pi'_{i|s,s_2} \pi'_{j|s,s_2} \tag{8}$$

while $E_{pq}(\cdot)$ represents the expectation computed with respect to the joint distribution of $p(s)$ and $q(s_1|s)$. This expectation operator may be eliminated from the variance formula by making additional assumptions on $q(s_1|s)$ and second-phase response probabilities. Särndal et al. (1992) also propose an assumption-independent estimator of this variance taking the form:

$$\widehat{V}(\hat{t}_{r*}) = \sum_{i,j \in s_1 \cup s'} \frac{x_i^r x_j^r}{\pi_i \pi_j \pi_{ij}^*} \Delta_{ij} + \sum_{i,j \in s'} \frac{x_i^r x_j^r}{\pi_i^* \pi_j^* \pi'_{ij|s,s_2}} \Delta'_{ij|s,s_2} \tag{9}$$

where

$$\pi_{ij}^{*} = \begin{cases} \pi_{ij}\pi'_{ij|s,s_2} & \text{for } i,j \in s_2 \\ \pi_{ij}\pi'_{i|s,s_2} & \text{for } i \in s_2,\ j \in s_1 \\ \pi_{ij}\pi'_{j|s,s_2} & \text{for } i \in s_1,\ j \in s_2 \\ \pi_{ij} & \text{for } i,j \in s_1 \end{cases} \tag{10}$$

The estimator $\widehat{V}(\hat{t}_{r*})$ is unbiased for $\hat{t}_{r*}$ irrespective of the response distribution.

3 Estimation of kurtosis

In order to estimate the population kurtosis K let us express this parameter as a function of population totals. This leads to the formula:

$$K = \frac{N^3 t_4 - 4N^2 t_3 t_1 + 6N t_2 t_1^2 - 3t_1^4}{(N t_2 - t_1^2)^2} \tag{11}$$

By replacing these unknown totals with respective unbiased two-phase based estimators we obtain the following estimator of K:

$$\widehat{K}_{*} = \frac{N^3 \hat{t}_{4*} - 4N^2 \hat{t}_{3*}\hat{t}_{1*} + 6N\hat{t}_{2*}\hat{t}_{1*}^2 - 3\hat{t}_{1*}^4}{(N\hat{t}_{2*} - \hat{t}_{1*}^2)^2} \tag{12}$$

Using Taylor linearization method we may express the approximate variance of $\widehat{K}_{*}$ in the form:

$$AV(\widehat{K}_{*}) = \sum_{i,j \in U} \frac{z_i z_j}{\pi_i \pi_j} \Delta_{ij} + E_{pq}\left(\sum_{i,j \in s_2} \frac{z_i z_j}{\pi_i^{*} \pi_j^{*}} \Delta'_{ij|s,s_2} \right) \tag{13}$$

where

$$z_i = \frac{1}{NS^6} \sum_{h=1}^{4} a_h x_i^h \tag{14}$$

and

$$a_1 = 4(m_4 m_1 - m_3 m_2 + 3m_2^2 m_1 - 3m_3 m_1^2) \tag{15}$$

$$a_2 = 2(4m_3 m_1 - 3m_2 m_1^2 - m^4) \tag{16}$$

$$a_3 = -4(m_1 m_2 - m_1^3) \tag{17}$$

$$a_4 = m_2 - m_1^2 \tag{18}$$

The same method leads to the derivation of its second-order approximate bias:

$$AB(\widehat{K}_*) = \sum_{i,j\in U} \frac{z_{ij}}{\pi_i \pi_j} \Delta_{ij} + E_{pq}\left(\sum_{i,j\in s_2} \frac{z_{ij}}{\pi_i^* \pi_j^*} \Delta'_{ij|s,s_2} \right) \tag{19}$$

where

$$z_{ij} = \frac{1}{N^2 S^8} \sum_{g=1}^{4} \sum_{h=1}^{min(g,2)} a_{gh}(x_i^g x_j^h + x_j^g x_i^h) \tag{20}$$

and

$$a_{11} = 2(m_4 m_2 + 3m_2^3 - 12m_3 m_2 m_1 + 5m_4 m_1^2 + 15m_2^2 m_1^2 - 12m_3 m_1^3) \tag{21}$$
$$a_{21} = 4(2m_3 m_2 - 3m_2^2 m_1 + 10m_3 m_1^2 - 6m_2 m_1^3 - 3m_4 m_1) \tag{22}$$
$$a_{22} = 3(2m_1^4 - 4m_3 m_1 + 2m_2 m_1^2 + m_4) \tag{23}$$
$$a_{31} = 4(3m_1^4 - 2m_2 m_1^2 - m_2^2) \tag{24}$$
$$a_{32} = 8(m_1 m_2 - m_1^3) \tag{25}$$
$$a_{41} = 4(m_1 m_2 - m_1^3) \tag{26}$$
$$a_{42} = -2(m_2 - m_1^2) \tag{27}$$

Moreover, the assumption-free variance estimator may be constructed as:

$$\widehat{V}(\widehat{K}_*) = \sum_{i,j\in s_1\cup s'} \frac{\hat{z}_i \hat{z}_j}{\pi_{ij}^* \pi_i \pi_j} \Delta_{ij} + \sum_{i,j\in s'} \frac{\hat{z}_i \hat{z}_j}{\pi_i \pi_j \pi'_{i|s,s_2} \pi'_{j|s,s_2} \pi'_{ij|s,s_2}} \Delta'_{ij|s,s_2} \tag{28}$$

where

$$\hat{z}_i = \frac{1}{N \hat{a}_4^3} \sum_{h=1}^{4} \hat{a}_h x_i^h \tag{29}$$

for $i \in U$ and $\hat{a}_1, \ldots, \hat{a}_1$ are respectively obtained from $a_1, \ldots, a_4$ by replacing each unknown population moment m_r in expressions (15)– (18) with its two-phase based unbiased estimator $\widehat{m}_r = \hat{t}_{r*}/N$. If all $\hat{z}_i$'s were exactly equal to corresponding z_i's, then the estimator (28) would be unbiased for $AV(\widehat{K}_*)$. In practice this will not hold exactly, but one may hope that the bias remains modest and tends to zero in large samples.

4 A special case

Let us assume that nonresponse is deterministic. The population is divided into two strata U_1 and U_2 of sizes N_1 and N_2 such that $\rho_{i|s} = 1$ for $i \in U_1$ and $\rho_{i|s} = 0$

otherwise. Moreover, assume that the simple random sampling without replacement is used in both phases of the survey with first-phase inclusion probabilities $\pi_i = n/N$, $\pi_{ij} = n(n-1)/(N(N-1))$ and second-phase inclusion probabilities $\pi'_{i|s,s_2} = n'/n_2$, $\pi'_{ij|s,s_2} = n'(n'-1)/(n_2(n_2-1))$ for $i \neq j \in U$ where $n' = c \cdot n_2$ and $0 < c < 1$ is a constant fixed in advance. The estimator $\hat{t}_{r*}$ of t_r may be expressed as:

$$\hat{t}_{r\circ} = \frac{N}{n}\left(\sum_{i\in s_1} x_i^r + \frac{1}{c}\sum_{i\in s'} x_i^r\right) \tag{30}$$

Consequently, the estimator $\widehat{K}_*$ of K takes the form:

$$\widehat{K}_\circ = \frac{N^3\hat{t}_{4\circ} - 4N^2\hat{t}_{3\circ}\hat{t}_{1\circ} + 6N\hat{t}_{2\circ}\hat{t}_{1\circ}^2 - 3\hat{t}_{1\circ}^4}{(N\hat{t}_{2\circ} - \hat{t}_{1\circ}^2)^2} \tag{31}$$

From (13) we obtain its approximate variance:

$$AV(\widehat{K}_\circ) = N^2\left(\frac{1-f}{n}S^2(z) + \frac{1-c}{c}\frac{W_2}{n}S_{U_2}^2(z)\right) \tag{32}$$

where $f = n/N$, $W_2 = N_2/N$ and

$$S^2(z) = \frac{1}{N-1}\sum_{i\in U}\left(z_i - \frac{1}{N}\sum_{j\in U} z_j\right)^2 \tag{33}$$

$$S_{U_2}^2(z) = \frac{1}{N_2-1}\sum_{i\in U_2}\left(z_i - \frac{1}{N_2}\sum_{j\in U_2} z_j\right)^2 \tag{34}$$

Hence, the approximate variance is a decreasing function of n. From (19) we also obtain the following second-order approximation of the bias:

$$AB(\widehat{K}_\circ) = \frac{1}{n}\cdot\frac{2}{N^2S^8}\cdot\sum_{g=1}^{4}\sum_{h=1}^{min(g,2)} a_{gh}c_{gh} \tag{35}$$

where constants a_{gh} are given by expressions (21)–(27) and

$$c_{gh} = (1-f)Cov(X^g, X^h) + W_2\frac{1-c}{c}Cov_{U_2}(X^g, X^h) \tag{36}$$

$$Cov(X^g, X^h) = \frac{1}{N-1}\sum_{i\in U}\left(x_i^g - \frac{1}{N}\sum_{j\in U} x_j^g\right)\left(x_i^h - \frac{1}{N}\sum_{j\in U} x_j^h\right) \tag{37}$$

$$Cov_{U_2}(X^g, X^h) = \frac{1}{N_2 - 1} \sum_{i \in U_2} \left(x_i^g - \frac{1}{N_2} \sum_{j \in U_2} x_j^g \right) \left(x_i^h - \frac{1}{N_2} \sum_{j \in U_2} x_j^h \right) \quad (38)$$

Hence, the approximate bias is also a decreasing function of n. The general variance estimator (28) takes now the form:

$$\widehat{V}(\widehat{K}_\circ) = A_0 \left(A_1 S_{s_1}^2(\hat{z}^\circ) + A_2 S_{s'}^2(\hat{z}^\circ) + A_3 \left(\overline{z}_{s_1}^\circ - \overline{z}_{s'}^\circ \right)^2 \right) \quad (39)$$

where

$$A_0 = \frac{N(N-1)}{n(n-1)} \quad (40)$$

$$A_1 = n_1 - 1 \quad (41)$$

$$A_2 = \frac{N(n_2 - 1) - cn_2(n-1) + n_1}{c(N-n)} \quad (42)$$

$$A_3 = \frac{n_1 n_2}{n} \quad (43)$$

while

$$\overline{z}_{s_1}^\circ = \frac{1}{n_1} \sum_{i \in s_1} \hat{z}_i^\circ \quad (44)$$

$$\overline{z}_{s'}^\circ = \frac{1}{n'} \sum_{i \in s'} \hat{z}_i^\circ \quad (45)$$

$$S_{s_1}^2(\hat{z}^\circ) = \frac{1}{n_1 - 1} \sum_{i \in s_1} \left(\hat{z}_i^\circ - \frac{1}{n_1} \sum_{j \in s_1} \hat{z}_j^\circ \right)^2 \quad (46)$$

$$S_{s'}^2(\hat{z}^\circ) = \frac{1}{n' - 1} \sum_{i \in s'} \left(\hat{z}_i^\circ - \frac{1}{n'} \sum_{j \in s'} \hat{z}_j^\circ \right)^2 \quad (47)$$

with

$$\hat{z}_i^\circ = \frac{1}{N \hat{a}_{4\circ}^3} \sum_{h=1}^{4} \hat{a}_{h\circ} x_i^h \quad (48)$$

and $\hat{a}_{h\circ}$'s are obtained from a_h's by replacing unknown raw moments $m_1, \ldots, m_4$ in expressions (15)–(18) with respective estimators: $\widehat{m}_{1\circ}, \ldots, \widehat{m}_{4\circ}$ where $\widehat{m}_{g\circ} = \hat{t}_{g\circ}/N$ for $g = 1, \ldots, 4$.

5 Conclusions

In this paper an estimator for the finite population kurtosis was proposed. It is constructed on the basis of a two-phase sample drawn in a nonresponse situation. A general double sampling procedure with arbitrary sampling designs in both phases was considered. The approximate bias and approximate variance formulas were derived for a stochastic nonresponse situation. They may be applied to assess the properties of the estimator in a wide range of situations. Finally, a special case of simple random sampling without replacement and deterministic nonresponse was considered. Presented results suggest that the proposed estimator is consistent at least in this special case.

Acknowledgment The work was supported by the grant No 1H02B 022 30 from the Ministry of Education and Science.

References

Arnab R (1998) Randomized response surveys: optimum estimation of a finite population total. Stat Pap39:405–408

Bethlehem JG (1988) Reduction of nonresponse bias through regression estimation. J Off Stat 4(3): 251–260

Cassel CM, Särndal CE, Wretman J (1983) Some uses of statistical models in connection with the nonresponse problem. In: Madow WG, Olkin I (eds) Incomplete data in sample surveys. Academic Press, New York

Chaudhuri A (1987) Randomized response surveys of a finite population: a unified approach with quantitative data. J Stat Plan Inference 15:157–165

Ekholm A, Laaksonen S (1991) Weighting via response modelling in the finnish household budget survey. J Off Stat 7(3):325–338

González S, Rueda M, Arcos A (2008) An improved estimator to analyse missing data. Stat Pap 49:791–796

Hansen MH, Hurwitz WN (1946) The problem of nonresponse in sample surveys. J Am Stat Soc 41: 517–529

Lessler JT, Kalsbeek WD (1992) Nonsampling error in surveys. Wiley, New York

Lütkepohl H, Theilen B (1991) Measures of multivariate skewness and kurtosis for tests of nonnormality. Stat Pap 32:179–193

Meeden G (2000) A decision theoretic approach to imputation in finite population sampling. J Am Stat Assoc 95:586–595

Okafor FC, Lee H (2000) Double sampling for ratio/regression estimation with subsampling for nonresponse. Surv Methodol 26(2):183–188

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York

Rueda M, González S (2004) Missing data and auxiliary information in surveys. Comput Stat 19:551–567

Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London

Singh HP, Kumar S (2010) Estimation of mean in presence of non-response using two phase sampling scheme. Stat Pap 51:559–582

Singh HP, Kumar S, Kozak M (2010) Improved estimation of finite-population mean using sub-sampling to deal with non-response in two-phase sampling scheme. Commun Stat Theory Methods 39:791–802

Sodipo AA, Obisesan KO (2007) Estimation of the population mean using difference cum ratio estimator with full response on the auxiliary character. Res J Appl Sci 2(6):769–772

Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc 60:63–69

Zwillinger D, Kokoska S (2000) Standard probability and statistics tables and formulae. Chapman & Hall/CRC, Boca Raton