STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 279—286

APPLICATION OF ORDER STATISTICS OF AUXILIARY VARIABLE TO ESTIMATION THE POPULATION MEAN

Janusz L. Wywiał[1]

ABSTRACT

Estimation of the population average in a finite population by means of sampling strategies dependent on an auxiliary variable highly correlated with a variable under study is considered. The sample is drawn with replacement on the basis of the probability distribution of an order statistic of the auxiliary variable. Observations of the variable under study are the values of the concomitant of the order statistic. The mean of the concomitant values is the estimator of a population mean of the variable under study. The expected value and the variance of the estimator are derived. The limit distributions of the considered estimators were considered. Finally, on the basis of simulation analysis, the accuracy of the estimator is considered.

Key words: Order statistic, sample quantile, auxiliary variable, sampling scheme, sampling design, concomitant.

1. Introduction and basic definitions

A fixed and finite population of size N denoted $U=(1,...,N)$ will be considered. The observation of a variable under study and an auxiliary variable are identifiable and denoted by y_i and x_i, $i=1,...,N$ respectively. The values of both variables are fixed. Our purpose is the estimation of the population mean $\bar{y}=\frac{1}{N}\sum_{k=1}^{N}y_k$. We assume that $x_i<x_{i+1}$, $i=1,...,N-1$. Let U_x be the set of all distinct values of the auxiliary variable.

Let Q be the set of all possible samples selected from the population U. The sampling design $P(s)$ has to fulfil the assumptions: $P(s)\geq 0$ for all $s\in Q$ and $\sum_{s\in Q}P(s)=1$.

[1] Katowice University of Economics, Department of Statistics, 14 Bogucicka, 40-226 Katowice, Poland. Email: janusz.wywial@ue.katowice.pl.

Let us consider simple random sampling design with replacement and with unequal probabilities of drawing population elements. The sample size is: $n > 2$. Let p_k be the probability of selecting the k-th population element in each draw. Let (X_i, Y_i) be a random variable whose values are the observations of an auxiliary variable and a variable under study, respectively, in the i-th draw. Thus, the probability function of the random variables (X_i, Y_i) is: $P(X_i = x_k, Y = y_k) = f(x_k) = p_k$ for $i = 1,...,n$ and $k = 1,...,N$. Thus, each random variable (X_i, Y_i) has the same probability function $f(x_k)$. Hence, the pairs of random variables $(X_i, Y_i),(X_2, X_2),...,(X_i, Y_i),...,(X_n, Y_n)$ are independent and they have the same distribution function defined by $P(X = x_k, Y = y_k) = f(x_k)$, $k = 1,...,N$. Thus, the sequence $\left((X_i, Y_i), i = 1,...,n\right)$ can be treated as data observed in the sample drawn with replacement from the population U with unequal probabilities $f(x_k)$, $k = 1,...,N$.

Let the sample $\left((X_i, Y_i), i = 1,...,n\right)$ ordered by the values of X_i be denoted by $\left((X_{r:n}, Y_{[r:n]}), r = 1,...,n\right)$ where $X_{r:n}$ is the r-th order statistic (so, $X_{r:n} \le X_{r+1:n}$, $r = 1,...,n-1$) and $X_{r:n}$ is concomitant of $X_{r:n}$, $r = 1,...,n$, see David and Nagaraja (2003), pp. 144. The distribution function of the r-th order statistic is as follows, see Arnold, Balakrishnan and Nagaraja (2008), pp. 42:

$$p_k(r:n) = P(X_{r:n} = x_k) = \sum_{i=0}^{r-1}\sum_{j=0}^{n-r} \frac{n!\left(F(x_{k-1})\right)^{r-1-i}\left(1 - F(x_k)\right)^{n-r-j}\left(f(x_k)\right)^{i+j+1}}{(r-1-i)!(n-r-j)!(j+j+1)!} \quad (1)$$

Let a k-th $k = 1,...,N$ population element be selected with replacement to the sample s of the size m in a single draw with the probability $p_k(r:n) = P(X_{r:n} = x_k)$. Hence, the sampling design of the defined sample is as follows:

$$P_{r:n}(s) = \prod_{k \in s} p_k(r:n)$$

Hence, the above sampling design is proportional to the product of the appropriate probabilities of the distribution function of r-th order statistic.

Moreover, let us note that another sampling scheme for selection of the sample is as follows. The simple sample of the size n is replicated m-times and in each of them the values x_k ($k = 1,...,N$) of the order statistic $X_{r:n}$ is observed.

The above defined sampling design $P_{r:n}(s)$ leads to selecting the sample s in which values $(x_{k_i}, y_{k_i}) k \in s$ are observations of the independent random pair $(X_{r:n}^{(k)}, Y_{[r:n]}^{(k)})$ where $X_{r:n}^{(k)}$ has the same distribution as the order statistic $X_{r:n}$ and $Y_{[r:n]}^{(k)}$ has the same distribution as the concomitant variable $Y_{[r:n]}$, $i = 1,...,z$.

Let us note that the proposed sampling design is similar to the sampling design considered by Wywiał (2) but the former one is proportional to the singular value of the order statistic of a positively valued auxiliary variable and it is drawn without replacement.

2. Estimation strategies

We are going to consider the basic properties of the strategy $\left(\bar{y}_s, P_{r:n}(s)\right)$ where $\bar{y}_s$ is the ordinary sample mean:

$$\bar{y}_s = \frac{1}{m}\sum_{i \in s} y_i = \frac{1}{m}\sum_{k=}^{m} Y_{[r:n]}^{(k)} \quad (2)$$

The expected value and the variance of the strategy $\left(\bar{y}_s, P_{r:n}(s)\right)$ are as follows:

$$E(\bar{y}_s, P_{r:n}(s)) = E(Y_{[r:n]}) \quad (3)$$

where

$$E(Y_{[r:n]}) = \sum_{k=1}^{N} y_k p_k(r:n) \quad (4)$$

$$V(\bar{y}_s, P_{r:n}(s)) = \frac{1}{m}\sum_{k=1}^{N} (y_k - E(Y_{[r:n]}))^2 p_k(r:n) \quad (5)$$

The unbiased estimator of the variance:

$$V_s(\bar{y}_s, P_{r:n}(s)) = \frac{1}{m-1}\sum_{k \in s} (y_k - \bar{y}_s)^2 \quad (6)$$

David and Nagaraja (2003), pp. 145, show that

$$E(Y_{[r:n]}) = \bar{y} + \frac{v_{xy}}{v_x}\left(E(X_{r:n}) - \bar{x}\right) \quad (7)$$

where

$$v_{xy} = \frac{1}{N}\sum_{k=1}^{N}(y_k - \bar{y})(y_k - \bar{y}) \quad , \quad v_x = \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x})^2 .$$

This straightforwardly leads to the following conclusion: when $E(X_{r:n}) = \bar{x}$, the estimation strategy $\left(\bar{y}_s, P_{r:n}(s)\right)$ is unbiased for the population mean. Thus, the parameters r and n should be assigned in such a way that $E(X_{r:n}) = \bar{x}$.

In the next section, the accuracy of the proposed strategy will be compared with the following ones. The first of them is the well known simple sample mean $\bar{y}_s = \sum_{k\in s} y_k$. The sampling design of the sample of the size n drawn without replacement is: $P_0(s) = \binom{N}{n}^{-1}$. The strategy $\left(\bar{y}_s, P_0(s)\right)$ is unbiased and its variance is $V(\bar{y}_S, P_0(s)) = \frac{N-n}{Nn} v_{*y}$, $v_{*y} = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{y})^2$.

Another well known strategy is called the mean from the sample of the fixed size n drawn with replacement. Each population element is drawn to the sample from a population with probability proportional to a value of the auxiliary variable. Let $p_k (k = 1,\dots,N, \sum_{k=1}^{N} p_k = 1)$ be the probability of selection of a k-th population element in a single draw. Using the simplified notation the multinomial sampling design explains the following expression, see e.g. Tillé (2006):

$$P_1(s) = \prod_{k\in s} p_k .$$

Usually the probabilities p_k, $k = 1,\dots,N$ are determined by the expression:

$$p_k = \frac{x_k}{\sum_{i=1}^{N} x_i}, \text{ for } x_k > 0, \; k = 1,\dots,N .$$

The unbiased estimation strategy is denoted by $\left(\hat{y}_s, P_1(s)\right)$ where the estimator is of the Hansen-Hurvitz (1943) type:

$$\hat{y}_s = \frac{1}{m}\sum_{k\in s}\frac{y_k}{p_k} \tag{8}$$

The variance of the strategy is:

$$V\left(\hat{y}_s, P_1(s)\right) = \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k}{Np_k} - \bar{y}\right)^2 p_k \tag{9}$$

The last estimator using the sample drawn according to the sampling design $P_{r:n}(s)$ is as follows:

$$\tilde{y}_s = \frac{1}{m}\sum_{k\in s}\frac{y_k}{p_k(r:n)} \tag{10}$$

$E(\tilde{y}_{m,z,s}) = \bar{y}$ and

$$V\left(\tilde{y}_s, P_{r:n}(s)\right) = \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k}{Np_k(r:n)} - \bar{y}\right)^2 p_k(r:n) \tag{11}$$

The construction of the strategy $\left(\bar{y}_s, P_{r:n}(s)\right)$ leads to the conclusion that its distribution converges to the normal distribution with parameters given by the expression (4) and (5) if $m \to \infty$. This conclusion results straightforwardly from the well known Lindeberg theorem, see e.g. Billingsley (1979).

On the basis of the same theorem it is easy to proof that under sufficiently large sample size m the strategies $\left(\hat{y}_s, P_1(s)\right) \sim N(\bar{y}, V(\hat{y}_s, P_1(s))$ and $\left(\tilde{y}_s, P_{r:n}(s)\right) \sim N(\bar{y}, V(\tilde{y}_s, P_{r:n}(s))$ where $V\left(\hat{y}_s, P_1(s)\right)$ and $V\left(\tilde{y}_s, P_{r:n}(s)\right)$ are given by the expressions (9) and (11), respectively.

3. Comparison of the estimators' accuracy

The accuracy of the considered strategies is based on the following relative accuracy coefficient:

$$e\left(t_s, P(s)\right) = \frac{V\left(t_s, P(s)\right)}{V\left(\bar{y}_s, P_0(s)\right)}$$

Thus, it is the ratio of the variance $V\left(t_s, P(s)\right)$ and the variance of the simple sample (drawn without replacement) mean. We use the following notation:

$$e_1 = e(\bar{y}_s, P_{r:n}(s)), \quad e_2 = e(\hat{y}_s, P_2(s)).$$

The simulation experiments were based on the procedure prepared by means of the program R. Firstly, according to a theoretical probability distribution, pseudo-values of the random variable (X,Y) have been generated. Two two-dimensional distribution functions have been considered. The first of them has been two dimensional normal distribution denoted by $N(100,0,1,1,\rho)$ where $E(X)=100$, $E(Y)=0$, $V(X)=V(Y)=1$ and the correlation coefficient $\rho=\rho(X,Y)$. The other distribution of the random variable (X,Y) was a two-dimensional exponential one where $X=Z+U$, $Y=Z+V$ and U, V, Z are independent, $Z \sim EXP(\alpha^{-1})$, $U \sim EXP(\beta^{-1})$, $V \sim EXP(\beta^{-1})$, $\alpha^2+\beta^2=1$, $\rho=\beta^2$. The considered sets of generated pseudo-values were of size $N=500$. The program has calculated the inclusion probabilities and finally the mean square error for different population for the considered strategies.

We assume that $m=n$. Thus, in Tables 1 and 2, the parameter r is the rank of the auxiliary variable order statistic for which $|E(X_{r:n})-\bar{x}|=minimum$.

The analysis of Table 1 lets us say that in the case of the normal distribution of the variables (X,Y), the strategy $(\bar{y}_s, P_{r:n}(s))$ is evidently better than $(\hat{y}_s, P_1(s))$. Moreover, the strategy is less accurate than the simple sample mean.

Table 1. The relative efficiency coefficients (%) of the strategies $(X,Y) \sim N(100,0,1,1,\rho)$.

	ρ:	0.5		0.8		0.95	
n	r	e_1	e_2	e_1	e_2	e_1	e_2
10	6	62	102	35	102	20	102
14	8	59	103	31	103	16	103
20	11	57	104	29	104	12	104
24	13	56	105	28	105	11	104
30	16	56	106	27	106	9	106
40	21	56	109	26	108	8	108
50	26	56	111	25	111	7	111

Source: The author's own calculations.

Table 2. The relative efficiency coefficients (%) of the strategies. The exponential distribution of (X,Y).

	ρ:	0.5		0.8		0.95	
n	r	e_1	e_2	e_1	e_2	e_1	e_2
10	2	62	138	38	54	20	15
14	3	67	140	40	55	21	15
20	4	58	141	38	55	16	15
24	5	62	142	39	56	17	16
30	5	58	144	39	57	15	16
40	7	62	147	40	58	16	16
50	7	63	151	41	59	15	17

Source: The author's own calculations.

The accuracy of the estimation in the case of a highly asymmetric two-dimensional exponential distribution of the variable under study and an auxiliary variable is considered in Table 2. When the correlation coefficient is high, $\rho = 0.95$, the accuracy of the strategies $(\hat{y}_s, P_1(s))$ and $(\bar{y}_s, P_{r:n}(s))$ is comparable. We infer that for a rather small correlation coefficient $\rho = 0.5$, the strategy $(\bar{y}_s, P_{r:n}(s))$ is evidently better than the strategy $(\hat{y}_s, P_1(s))$. In this case, the strategy $(\hat{y}_s, P_2(s))$ is less accurate than the simple sample mean. Finally, in the medium case, when $\rho = 0.8$, the strategy $(\bar{y}_s, P_1(s))$ is a little better than $(\hat{y}_s, P_1(s))$.

4. Conclusions

The proposed $(\bar{y}_s, P_{r:n}(s))$ strategy for the population mean (total) is based on sampling design dependent on order statistics of an auxiliary variable. The basic parameters of the strategy are derived. It was shown when it is unbiased. It is quite easy to show that the considered estimators has an approximately normal distribution when the sample size is sufficiently large. Thus, it is possible to construct confidence intervals for a population mean (total) of a variable under study as well as to test statistical hypotheses about those parameters.

The simulation analysis let us conclude that in the case of two-dimensional distribution of variable under study and the auxiliary one, the proposed estimation strategy $\left(\bar{y}_s, P_{r:n}(s)\right)$ is not worse than the strategy $\left(\hat{y}_s, P_1(s)\right)$.

Acknowledgements

The research was supported by the grant number N N111 434137 from the Ministry of Science and Higher Education. I am grateful to professor Malay Ghosh for his valuable comments.

REFERENCES

ARNOLD, B.C., BALAKRISHNAN N., NAGARAJA H.N. (2008). *A First Course in Order Statistics*. Society for Industrial and Applied Mathematics. Philadelphia.

BILLINGSLEY, P. (1979). *Probability and Measure*. John Wiley & Sons, New York-Chichester-Brisbane- Toronto.

DAVID, H.A., NAGARAJA, H.N. (2003). *Order statistics*. John Wiley & Sons.

HANSEN, M.H., HURVITZ, W.N. (1943). *On the theory of sampling from finite population. Annals of Mathematical Statistic*, 14, 333-362.

TILLE, Y. (2006). *Sampling algorithms*. Springer.

WYWIAŁ, J.L. (2008). Sampling design proportional to order statistic of auxiliary variable. *Statistical Papers* vol. 49, Nr. 2/April, pp. 277-289.