ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 8, 2004

Approximate design-based variance of functions of covariance matrix

IMBI TRAAT

ABSTRACT. Functions of a design-weighted estimator $\hat{S}$ of the finite population covariance matrix are considered. For these functions (determinant, Hotelling's T^2) the approximate (Taylor series based) variances are derived. For $\hat{S}$ also the exact dispersion matrix is derived. These are generalizations of the earlier results for independent identically distributed (*i.i.d.*) variables. A simulation study supports the derived formulae.

1. Introduction

Finite population estimation theory is mostly one-dimensional. It means that one study variable is considered at a time and subject of the estimation is its certain parameter (a number) like population mean, total of values, or variance. In this paper we consider many study variables at a time with their joint behavior being characterized by the finite population covariance matrix S. This is a multivariate parameter of the finite population.

A natural estimator for S in the finite population context is a design-weighted sample covariance matrix $\hat{S}$. For the self-weighting designs it reduces to the classical sample covariance matrix. Distribution of $\hat{S}$, and also of its functions, is created by the sampling design (so called design-based distribution). Distribution of $\hat{S}$ and of many of its functions is thoroughly studied only under one special sampling design – simple random sampling with replacement (SIR sampling). Corresponding sample satisfies the *i.i.d.* assumption of data which is the basic assumption in classical statistics while studying distributions of sample functions.

Received November 4, 2003.

2000 *Mathematics Subject Classification.* 62D05, 62F12, 62H10.

Key words and phrases. Finite population, sampling, covariance matrix, determinant, Hotelling's T^2, Taylor linearization.

The author is thankful for the financial support of Estonian Science Foundation Grant 5523 and of Target Financed Project 0181776s01.

In this paper we observe a general situation. Let the sample be drawn by any sampling design. We use the distributional approach by Traat (2000) allowing simultaneous consideration of with- and without-replacement (WR and WOR) sampling designs. Under these circumstances we are interested in the distributional characteristics of $\hat{S}$ and some of its functions. In Traat (2003) an approximate (Taylor-linearization based) dispersion matrix of $\hat{S}$ is derived. Here we derive the exact dispersion matrix of $\hat{S}$ being a generalization of the earlier *i.i.d.* result by Traat (1984).

Our main interest in this paper is turned to the important functions of $\hat{S}$ – determinant of $\hat{S}$ and Hotelling's T^2-statistic. They are valuable inferential tools in classical statistics – $|\hat{S}|$ estimates generalized variance of the set of variables, T^2 is used for testing hypothesis about the mean vector. These inferential procedures are based on the knowledge of distributions of these functions. The distributions are known only under *i.i.d.* assumption, in some very special cases exactly, otherwise asymptotically. In Kollo (1990) the asymptotic normal distribution of the above functions is established and corresponding asymptotic variances are presented. In this paper we derive these asymptotic variances under more general assumptions, valid for any sampling design. We show that in the special case of SIR sampling they coincide with the results by Kollo (1990). The simulation study confirms our results. It can be also seen that convergence to the asymptotic distributions is much slower in the complex sampling situation than in the *i.i.d.* case.

2. Finite population, sampling

Finite population is a matrix of fixed values

$$Y:\ p\times N,$$

involving measurements of p study variables on N units. Usually the case $p=1$ is considered in sampling literature. Inference is desired on Y, in practice, usually on totals and related quantities like means, proportions etc. In the multivariate setting all totals can be presented simultaneously as

$$t=Y\mathbf{1}:\ p\times 1,$$

where $\mathbf{1}:\ N\times 1$ is a vector of ones. Estimation of t can be straightforwardly built up on the estimation of its components which is well covered in the sampling theory literature. More complex multivariate parameters are almost not considered in the finite population sampling theory. The important representatives, observed in this paper, are the population covariance matrix

$$S=\frac{1}{N}(YY'-\frac{1}{N}Y\mathbf{1}\mathbf{1}'Y'):\ p\times p \tag{1}$$

and its functions. A convenient tool for describing sample selection from a finite population is the sampling or design vector (see Traat 2000):

$$I = (I_1, I_2, \dots, I_N)'.$$

It is a random vector with I_i showing the selection count of the population unit i ($I_i = 0$ meaning that i is not selected). Here EI_i is called the sampling expectation and $\pi_i = Pr(I_i > 0)$ is the inclusion probability of unit i. Distribution of I is the sampling design. It is a discrete multivariate distribution. Consideration of WOR and WR designs is unified in this way. The observed data can be conveniently expressed as (YI_d, I), where $I_d : N \times N$ is a diagonal matrix with the vector I on the diagonal. Multiplying with I_d extracts columns of Y which correspond to sampled units. Randomness which in the design-based approach comes in through I, is now explicitly given in data.

For the estimation purposes the expanded sampling vector is needed:

$$\breve{I} = (\breve{I}_1, \breve{I}_2, \dots, \breve{I}_N), \text{ where } \breve{I}_i = \frac{I_i}{EI_i}.$$

Note that

$$E\breve{I} = \mathbf{1} : \ N \times 1,$$

offering an especially simple form for unbiased estimation of t:

$$\hat{t} = Y\breve{I}. \tag{2}$$

The most important design characteristic in the estimation formulae is the design covariance matrix, more precisely, the covariance matrix of the expanded design vector:

$$\Delta = E(\breve{I} - \mathbf{1})(\breve{I} - \mathbf{1})' : \ N \times N. \tag{3}$$

The matrix Δ defines variance of $\hat{t}$ in (2):

$$V(\hat{t}) = Y\Delta Y', \tag{4}$$

The matrix Δ defines also approximate (Taylor linearization based) variance of nonlinear statistics, some of them considered later in this paper.

3. Estimator of finite population covariance matrix

Finite population covariance matrix (1) is an expression of totals

$$S = \frac{1}{N}(T - \frac{1}{N}\, t\, t'), \text{ where } T = YY', \ t = Y\mathbf{1}. \tag{5}$$

A consistent estimator of S is received by replacing totals by their unbiased estimators:

$$\hat{S} = \frac{1}{\hat{N}}(\hat{T} - \frac{1}{\hat{N}}\, \hat{t}\, \hat{t}'), \tag{6}$$

where

$$\hat{T} = Y\check{I}_d Y', \; \hat{t} = Y\check{I}, \; \hat{N} = \mathbf{1}'\check{I}. \tag{7}$$

The consistency of $\hat{S}$ holds both in the sense of infinite population (Särndal et al. 1992, p. 166-168) and in the sense of finite population, postulating that if the full population is sampled then $\hat{S} = S$. Note that in spite of unbiased components, $\hat{S}$ is biased for S, the bias matrix being

$$B = E\hat{S} - S = -\frac{1}{N^2} Y \Delta Y'. \tag{8}$$

We see that $\hat{S}$ underestimates S and the bias is $1/N^2$ times the covariance matrix of $\hat{t}$.

The approximate (Taylor linearization based) dispersion matrix of vec$\hat{S}$ is given in Traat (2003):

$$AV(\text{vec}\hat{S}) = U \Delta U', \tag{9}$$

where $U : \; p^2 \times N$ is such that its ith column is

$$U_i = \frac{1}{N}\;((y_i - \frac{t}{N}) \otimes (y_i - \frac{t}{N}) - \text{vec}S). \tag{10}$$

In the special case of the SIR design,

$$\Delta_{ii} = \frac{N-1}{n}, \; \Delta_{ij} = -\frac{1}{n},$$

we get from (9) the classical *i.i.d.* result (see e.g. Parring, 1979):

$$AV(\text{vec}\hat{S}) = \frac{1}{n}(M_4 - \text{vec } S \text{ vec}' \; S),$$

where

$$M_4 = \frac{1}{N}\sum_U (y_i - \frac{t}{N}) \otimes (y_i - \frac{t}{N})' \otimes (y_i - \frac{t}{N}) \otimes (y_i - \frac{t}{N})' \tag{11}$$

is the fourth central moment of the population. The one-dimensional special case of (9) is given in Särndal et al. (1992, p. 186-188).

Below we present the exact design-based dispersion matrix of $\hat{S}$. Looking at the form of S in (1) we immediately see a possible estimator

$$\hat{S} = \frac{1}{N} Y(\check{I}_d - \frac{1}{N}\check{I}\check{I}')Y'. \tag{12}$$

Apart from the formula (6) we do not use $\hat{N}$ here. The estimator is not linear in I but still considerably simple (with random part separated in the middle).

First, we are interested in the exact mean square error matrix of $\hat{S}$:

$$MSE(\text{vec}\hat{S}) = E[\text{vec}(\hat{S} - S)\text{vec}'(\hat{S} - S)].$$

Exploiting the expressions (5) and (12) together with the property $\mathrm{vec}(ABC) = (C' \otimes A)\mathrm{vec}B$ we have

$$\mathrm{vec}(\hat{S} - S) = \frac{1}{N}(Y \otimes Y)D,$$

where

$$D = [\mathrm{vec}(\breve{I}_d - \mathbf{1}_d) - \frac{1}{N}(\breve{I} \otimes \breve{I} - \mathbf{1} \otimes \mathbf{1})].$$

Now it is easy to get MSE-matrix of of $\hat{S}$:

$$MSE(vec\hat{S}) = \frac{1}{N^2}(Y \otimes Y)\ E(DD')\ (Y' \otimes Y').$$

With the help of MSE- and bias-matrices the exact dispersion matrix of $\hat{S}$ is:

$$V(vec\hat{S}) = MSE(vec\hat{S}) - \mathrm{vec}B\mathrm{vec}'B.$$

Its special case under *i.i.d.* sampling is given in Traat (1984). We see that both, the exact MSE- and dispersion matrices of $\hat{S}$, depend on the moments of the design vector $\breve{I}$ up to the 4th order. These moments are given if the sampling design is given. Nevertheless, they are often difficult to calculate. For smaller populations these moments can be estimated by repeated sampling of the frame.

4. Functions of covariance matrix

Our main aim is to present approximate variances of the following important functions of $\hat{S}$:

- $|\hat{S}|$ – determinant, generalized variance;
- $T^2 = \hat{\hat{Y}}'\hat{S}^{-1}\hat{\hat{Y}}$ – Hotelling's T^2-statistic, where $\hat{\hat{Y}}$ is given in (15).

We use general technique of Taylor linearization. Assume $f(Z): \ R^p \to R^1$ allows Taylor approximation (linear) around Z_0,

$$f(Z) \approx f(Z_0) + D(Z - Z_0),$$

where a row-vector

$$D = \frac{df}{dZ}|_{Z=Z_0}$$

is a matrix derivative (Magnus and Neudecker, 1999). Since $f(Z_0)$ and DZ_0 are constants, then the variability of $f(Z)$ is produced by the term DZ. Consequently, the approximate variance of $f(Z)$ is:

$$AV[f(Z)] = D\ AV(Z)\ D', \tag{13}$$

where $AV(Z)$ is the exact or approximate covariance matrix of Z, usually found from its Taylor expansion.

For functions considered in this paper $Z = \mathrm{vec}\hat{S}$ and $Z = (\hat{\hat{Y}}'\vdots\mathrm{vec}'\hat{S})'$. For the first case $AV(Z)$ is available in (9), for the second case we derive it.

It remains to find the first matrix derivatives of our functions of Z. These derivatives have been found earlier in the matrix literature. Putting the pieces together we present below the Taylor variances for the above listed functions of $\hat{S}$. They are valid for any sampling design. We point out that in the special case of SIR design they coincide with the classical *i.i.d.* results.

4.1. Taylor variance of determinant. Using the derivative (Magnus and Neudecker, 1999, p. 178),

$$D = \frac{d|\hat{S}|}{d\hat{S}}|_{\hat{S}=S} = |S|\mathrm{vec}'(S^{-1})$$

and the expression (9) we get from the general result (13):

$$AV(|\hat{S}|) = |S|^2\mathrm{vec}'(S^{-1})U\Delta U'\mathrm{vec}(S^{-1}). \tag{14}$$

By specifying Δ the result is expressed for different sampling designs. For example, for multinomial design (WR with unequal selection probabilities p_i) $\Delta_{ii} = \frac{1-p_i}{np_i}$, $\Delta_{ij} = -\frac{1}{n}$. For SIR design, $p_i = 1/N$, and we get from (14) the known classical result (Kollo, 1990),

$$AV_{SIR}(|\hat{S}|) = \frac{1}{n}|S|^2[\mathrm{vec}'(S^{-1})M_4\mathrm{vec}(S^{-1}) - p^2],$$

where M_4 is given by (11). For SI design $\Delta_{ii} = \frac{N-n}{n}$, $\Delta_{ij} = -\frac{N-n}{n(N-1)}$, the asymptotic variance of the determinant will become smaller:

$$AV_{SI}(|\hat{S}|) = \frac{N-n}{N-1}AV_{SIR}(|\hat{S}|).$$

4.2. Taylor variance of Hotelling's T^2. The Hotelling's $T^2 = \hat{\bar{Y}}'\hat{S}^{-1}\hat{\bar{Y}}$ is a function of the mean estimator

$$\hat{\bar{Y}} = \frac{\hat{t}}{\hat{N}} \tag{15}$$

and of $\hat{S}$. By Traat (2003) we have Taylor linearized vec$\hat{S}$ in the form

$$\mathrm{vec}\hat{S}_0 = U\check{I}, \tag{16}$$

where U is defined in (10). Taylor linearized $\hat{\bar{Y}}$ (without constant term) is

$$\hat{\bar{Y}}_0 = (\hat{t} - \bar{Y}\hat{N})/N, \tag{17}$$

or after replacing $\hat{t}$ and $\hat{N}$ with their expressions through $\check{I}$,

$$\hat{\bar{Y}}_0 = W\check{I}, \tag{18}$$

where $W : p \times N$ is a matrix with $(y_i - \bar{Y})/N$ in its ith column, $\bar{Y} = t/N$. Now it is easy to see that $V(\hat{\bar{Y}}_0) = W\Delta W'$ and $Cov(\hat{\bar{Y}}_0, \mathrm{vec}\hat{S}_0) = W\Delta U'$.

Finally, denoting $Z = (\hat{\bar{Y}}' \vdots \, \mathrm{vec}'\hat{S})' : \ 1 \times p^3$, we get its Taylor variance in the block form:

$$AV(Z) = \begin{bmatrix} W\Delta W' & \vdots & W\Delta U' \\ \cdots & \cdots & \cdots \\ U\Delta W' & \vdots & U\Delta U' \end{bmatrix}. \tag{19}$$

Using the expression of derivative (Kollo 1990):

$$D = \frac{d(\hat{\bar{Y}}'\hat{S}^{-1}\hat{\bar{Y}})}{dZ}|_{Z=Z_0} = (2\bar{Y}'S^{-1} \vdots \, \bar{Y}'S^{-1} \otimes \bar{Y}'S^{-1}), \tag{20}$$

where $Z_0 = (\bar{Y}' \vdots \, \mathrm{vec}'S)'$ we get from the general result (13) for $\bar{Y} \neq 0$ the approximate variance of Hotelling's T^2-statistic:

$$AV(T^2) = D \ AV(Z) \ D'. \tag{21}$$

For the special case of SIR design (19) reduces to the form:

$$AV(Z) = \frac{1}{n} \begin{bmatrix} S & \vdots & M_3' \\ \cdots & \cdots & \cdots \\ M_3 & \vdots & M_4 - \mathrm{vec}S \ \mathrm{vec}'S \end{bmatrix},$$

where

$$M_3 = \frac{1}{N} \sum_U (y_i - \frac{t}{N}) \otimes (y_i - \frac{t}{N})' \otimes (y_i - \frac{t}{N}).$$

Consequently, in this case $AV(T^2)$ reduces to the classical *i.i.d.* result given in Kollo (1990).

5. Simulation

The Taylor variances of determinant and Hotelling's T^2 derived in this paper are compared with corresponding simulated values.

The population $Y : \ p \times N$, where $p = 2$ and N=1000 is generated from the bivariate normal distribution. The population mean vector $\bar{Y}$ and the covariance matrix are:

$$\bar{Y} = (5.007, \ 4.934), \quad S = \begin{bmatrix} 0.942 & 0.399 \\ 0.399 & 1.002 \end{bmatrix},$$

where

$$|S| = 0.785 \text{ and } \bar{Y}'S^{-1}Y = 36.13.$$

Two different designs were used for drawing samples from Y:

- SIR design with $EI_i \equiv \frac{n}{1000}$, where n is the sample size.

- Multinomial design with
$$EI_i = b + (1 - bN/n)np_i, \tag{22}$$
where $b = 0.02$, $p_i = (x_i - \bar{x})^2 / \sum_{i=1}^{N} x_i$ and x_i is the auxiliary variable known for the whole population. With (22) we get highly variable sampling expectations, in our case ranging from 0.02 to 4.85, to check our formulae in most difficult situations. With $b = 0.02$ we protect us against nearly zero sampling expectations.
- Sample sizes $n = 100$, $n = 400$ were considered.

The mean and variance in tables are simulated values over 1000 repetitions, AV is Taylor linearized variance derived in this paper. In Tables 1 – 2 the auxiliary variable x_i is correlated with one of the study variables, in Tables 3 – 4 it is not.

Table 1. Mean and variance of $|\hat{S}|$ *, correlated auxiliary*

	SIR		Multinomial	
n	100	400	100	400
mean	0.751	0.779	1.180	1.000
variance	0.023	0.006	0.064	0.021
AV, (14)	0.025	0.006	0.037	0.021

Table 2. Mean and variance of T^2 *, correlated auxiliary*

	SIR		Multinomial	
n	100	400	100	400
mean	37.50	36.47	34.06	34.63
variance	36.22	6.94	32.82	13.74
AV, (19) – (21)	28.63	7.16	41.82	21.25

The mean in tables confirms the fact that both $|\hat{S}|$ and T^2 are biased for corresponding population values. The bias decreases when n increases. The bias is bigger and decreases more slowly for multinomial design. Comparing variances we see that AV approximates well the true (i.e. simulated) variance in classical SIR case. In multinomial case with highly variable selection probabilities p_i the approximation is not so good, but it becomes better with increasing n. The approximation works better for $|\hat{S}|$ than for T^2.

Below we consider an auxiliary variable uncorrelated with study variables. We assume study variables to be uncorrelated with each other too. Our finite population is now

$$\bar{Y} = (4.954,\ 4.941), \quad S = \begin{bmatrix} 1.048 & -0.053 \\ -0.053 & 1.001 \end{bmatrix},$$

where

$$|S| = 1.046 \text{ and } \bar{Y}'S^{-1}Y = 50.42.$$

With these changes we expect to increase the variability of considered statistics.

Table 3. Mean and variance of $|\hat{S}|$, *uncorrelated auxiliary*

	SIR		Multinomial	
n	100	400	100	400
mean	1.011	1.036	0.930	0.970
variance	0.041	0.011	0.068	0.034
AV, (14)	0.042	0.011	0.103	0.061

Table 4. Mean and variance of T^2, *uncorrelated auxiliary*

	SIR		Multinomial	
n	100	400	100	400
mean	52.46	50.89	54.45	52.76
variance	53.04	12.42	99.71	45.81
AV, (19) – (21)	48.68	12.17	113.68	65.32

The same tendencies as described above are visible in Tables 3–4 too. Here the variability of statistics is bigger. The AV captures the increasing variability. With increasing n the approximation becomes very good under SIR design, whereas under multinomial design it is moderate.

Conclusions

In this paper we considered a consistent estimator $\hat{S}$ of the finite population covariance matrix. We derived the exact design-based covariance matrix of vec$\hat{S}$. We considered also the functions of $\hat{S}$ such as determinant $|\hat{S}|$ and Hotelling's T^2-statistic. With Taylor linearization we derived their approximate design-based variances. The results are more general than one can find in literature. The classical *i.i.d.* results follow from ours under one special sampling design – simple random sampling with replacement. The derived formulae were supported by the simulation study. The approximation works very well for the designs with equal EI_i. For the designs with highly variable EI_i the approximation follows the true variance, though not in the best possible way.

References

Kollo, T. (1990). Investigation of the convergence of functions of sample means and covariances to limit distributions. In: *Probability Theory and Mathematical Statistics.*

Proceedings of the Fifth Vilnius Conference. Vol. 1., Eds: Grigelionis, B. et al., Mokslas/VSP, Vilnius/Utrecht, 638–646.

Magnus, J. R. and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Economics.* Wiley, Chichester.

Parring, A.-M. (1979). Calculation of asymptotic characteristics of sample functions. *Tartu Riikl. Ül. Toimetised* **492**, 86–90. (In Russian)

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

Traat, I. (1984). Moments of sample covariance matrix. *Trudy Vychisl. Tsentra Tartu. Gos. Univ.* **51**, 108–125. (In Russian)

Traat, I. (2000). Sampling design as a multivariate distribution. In: *Multivariate Statistics. Proceedings of the 6th Tartu Conference*, Eds: Kollo, T. et al., VSP/TEV, Vilnius/Utrecht, 195–208.

Traat, I. (2003). On the estimation of finite population covariance matrix. *Statistics in Transition* **6**, 67–83.

Institute of Mathematical Statistics, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia
E-mail address: imbi@ut.ee