ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 17, Number 1, June 2013
Available online at www.math.ut.ee/acta/

Influence of informative sampling on dependence between variables

Julia Aru

Abstract. In the case of informative sampling the sampling scheme explicitly or implicitly depends on the response variables. As a result, neither the sample distribution of response variables, nor the covariance matrix reflects the corresponding population counterparts. In this paper, a relationship between multivariate sample and population distributions is used. Based on this, the influence of the informative sampling on the covariance matrix is investigated. It is shown that with inclusion probabilities in a multiplicative form with respect to study variables, the independence between variables is preserved in the sample. Further, it is shown that with inclusion probabilities exponentially depending on the study variables, the multivariate exponential family is invariant under sampling. The sample distribution belongs to the same family as the population distribution but with different parameters. The relationship between parameters is given. The multinomial and multivariate normal distributions are examined in more detail and the parameters of their sample distributions are derived explicitly. The effect of the informative sampling on the respective covariance matrices and correlations is analysed and illustrated in the examples.

1. Introduction

Many statistical methods are based on the dependence between variables. In the multivariate case, the most important dependence characteristic to use is the covariance matrix of variables. Therefore, valid estimation of the covariance matrix is of utmost importance in data analysis. But with survey data the task is not trivial; sampling may disturb relationships between variables. In this article we focus on the multivariate distributions and their

Received May 15, 2012.
2010 *Mathematics Subject Classification.* 62D05.
Key words and phrases. Covariance matrix, inclusion probabilities, informative sampling, multinomial distribution, multivariate exponential family, multivariate normal distribution.
http://dx.doi.org/10.12697/ACUTM.2013.17.06

covariance matrices under non-ignorable or informative sampling. In the case of informative sampling the sampling scheme explicitly or implicitly depends on the study variable(s). As a result, the sample distribution of the study variables does not reflect the population distribution and does not approximate it after increasing the size of a sample either. The traditional sample estimates are biased for the population parameters.

We use an analytical form of sample probability density function as proposed by Pfeffermann and Sverchkov (1999). So far this approach was used for one-dimentional response variable in the modelling context. We use their approach to derive sample probability density functions for multivariate distributions. We concentrate on the exponential family. We assume inclusion probabilities to depend on the study variables. Having chosen an exponential function for this relationship we show that the multivariate sample probability density function again belongs to the exponential family but with different parameters. Knowledge of the sample distribution makes classical statistical inference possible with survey data. In particular, we give explicitly the sample probability density functions for the multinomial and multivariate normal distributions. We present covariance matrices of the obtained sample distributions and analyse differences between population and theoretical sample covariances.

It should be mentioned that the theoretical sample covariance matrix can be estimated from a sample in the classical way by a traditional sample covariance matrix. The established relationship between the theoretical sample and population covariance matrices now allows to produce an estimate for the population covariance matrix.

In sample surveys, the population covariance matrix can also be estimated in other ways, for example by a design-weighted sample covariance matrix (Traat, 2003). These other ways do not lend themselves to analytical comparisons of theoretical sample and population covariances. With our approach the effect of the informative sampling can be analysed from the derived expressions.

2. Preliminaries

Let $U = \{1, 2, \ldots, i, \ldots, N\}$ define the finite population of size N. The vector of study variables at object i is denoted by $\mathbf{y}_i = (y_i^1, y_i^2, \ldots, y_i^k)'$, where k is the number of study variables. Let $f_p(\mathbf{y}_i)$ be the probability density function (pdf) of the study variables in the population, either discrete or continuous. The vector of parameters indexing f_p is denoted by $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$.

The sample from a population is denoted by s and consists of n objects from U, selected according to some sample selection scheme with inclusion probabilities $\pi_i = P(i \in s)$. In what follows we consider single stage sampling

with inclusion probabilities π_i. In practice, probabilities often depend on the values of study variable(s), the values of the auxiliary variables and, possibly, values of design variables used for sample selection. For simplicity we consider the case without any auxiliary or design variables, but all results are also valid for conditional distributions given with the auxiliary or design variables. Sample pdf f_s of study variables is characterized by additional condition of object i being included in the sample s:

$$f_s(\mathbf{y}_i) = f_p(\mathbf{y}_i, i \in s).$$

As shown by Pfeffermann and Sverchkov (1999), by applying Bayes theorem to the sample pdf we get the following relationship between sample and population pdfs ($E_p()$ denotes expectation):

$$f_s(\mathbf{y}_i) = \frac{E_p(\pi_i|\mathbf{y}_i) f_p(\mathbf{y}_i)}{E_p(\pi_i)}. \tag{1}$$

Expression (1) defines the relationship between population and sample distributions, so that if π_i depends on $\mathbf{y}_i$, then $E_p(\pi_i|\mathbf{y}_i) \neq E_p(\pi_i)$ and $f_p(\mathbf{y}_i) \neq f_s(\mathbf{y}_i)$. In this case the population distribution differs from the sample distribution and the sample design is informative. Note that $E_p(\pi_i)$ is the normalizing constant in (1), i.e.,

$$f_s(\mathbf{y}_i) \propto E_p(\pi_i|\mathbf{y}_i) f_p(\mathbf{y}_i). \tag{2}$$

Relationship (1) or (2) is the basis for the following sections.

3. Independence in the population

Consider the case when variables $y^1, \ldots, y^k$ are independent in the population. Then the population pdf can be rewritten as the product of marginal distributions:

$$f_p(\mathbf{y}_i) = f_p(y_i^1) f_p(y_i^2) \ldots f_p(y_i^k). \tag{3}$$

Let the sample selection probabilities depend on the study variables and have expectations in multiplicative form,

$$E_p(\pi_i|\mathbf{y}_i) = E_p(\pi_i|y_i^1) E_p(\pi_i|y_i^2) \ldots E_p(\pi_i|y_i^k). \tag{4}$$

Then

$$E_p(\pi_i) = E_p(E_p(\pi_i|\mathbf{y}_i)) = E_p(E_p(\pi_i|y_i^1)) \ldots E_p(E_p(\pi_i|y_i^k)). \tag{5}$$

Substituting (3) – (5) into (1), we see that the sample pdf is again the product of marginal pdfs, i.e.,

$$f_s(\mathbf{y}_i) = f_s(y_i^1) f_s(y_i^2) \ldots f_s(y_i^k).$$

and so variables are independent in sample as well. So, even in the case of highly informative sampling, the independence between variables can be preserved if effects of different variables in the inclusion probabilities are multiplicative, like in (4).

Example 1. Consider the population of $N = 300$ objects and two study variables, y and z, which are independent and follow the standard normal distribution in the population. We take a sample of $n = 40$ objects from that population with two versions of inclusion probabilities which make sampling informative:

$$\pi_i = c_a(5 + y_i z_i + \epsilon_i),$$
$$\pi_i = c_b((3 + y_i)(3 + z_i) + \epsilon_i)$$

where ϵ_i is a random error following a uniform distribution $U(-0.15, 0.15)$ and c_a and c_b are normalizing constants. We use Pareto sampling as described by Traat et al. (2004).

Conditional population expectations of inclusion probabilities are then

$$E_p(\pi_i|y_i, z_i) = c_a(5 + y_i z_i), \tag{7a}$$
$$E_p(\pi_i|y_i, z_i) = c_b(3 + y_i)(3 + z_i). \tag{7b}$$

Graphs on Figure 1 show the population objects and objects selected into the sample in cases (7a) and (7b). In case (7a) effects of y and z were not

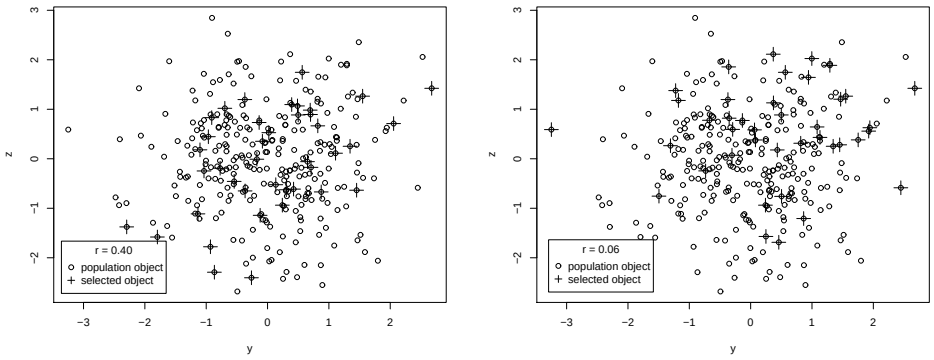

FIGURE 1. Population and selected objects, independence not preserved (on the left) and independence preserved (on the right)

multiplicative and objects with both positive or both negative values of y and z were selected into the sample, thus introducing correlation in the sampled values, $r = 0.40$. While in case (7b) independence was preserved, $r = 0.06$, the sampling is still informative: sample distributions of y and z are different from population distributions (sample mean of y is 0.17, sample mean of z is 0.41).

4. Population distribution in multivariate exponential family

An example of exponential family in the univariate case with informative sampling was examined in Pfeffermann et al. (1998). Presented here is a generalisation for the multivariate case. In general, due to informative sampling, a multivariate distribution in the population has another form in the sample. Also dependence characteristics, such as covariance, change. In some cases the sample covariance depending on the population parameters can be derived analytically.

Let the population distribution belong to the multivariate exponential family (Lehman, Casella 1998),

$$f_p(\mathbf{y}|\boldsymbol{\eta}) = h(\mathbf{y}) \exp\left(\sum_{j=1}^{m} \eta_j T_j(\mathbf{y}) - B(\boldsymbol{\eta})\right),$$

where $\boldsymbol{\eta} = (\eta_j)$ is an m-dimensional vector of canonical parameters taking values in the parameter space $\boldsymbol{\Psi} \subset \mathbb{R}^m$, $h(\cdot) : \mathbb{R}^k \to \mathbb{R}$ and $T_j(\cdot) : \mathbb{R}^k \to \mathbb{R}$ are functions of $\mathbf{y}$, $B(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is a function of $\boldsymbol{\eta}$.

If the inclusion probabilities also have an exponential form

$$E_p(\pi|\mathbf{y}) \propto \exp\left(\sum_{j=1}^{m} p_j T_j(\mathbf{y})\right),$$

where p_j are some constants, then, according to (1), the sample distribution belongs to the same family as the population distribution but with different parameters, $\eta_j^* = (\eta_j + p_j), j = 1, \ldots, m$, provided $\boldsymbol{\eta}^*$ lies in $\boldsymbol{\Psi}$,

$$\begin{aligned} f_s(\mathbf{y}|\boldsymbol{\eta}) &= \frac{f_p(\mathbf{y}|\boldsymbol{\eta}) E_p(\pi|\mathbf{y})}{E_p(\pi)} \\ &\propto h(\mathbf{y}) \exp\left(\sum_{j=1}^{m} \eta_j T_j(\mathbf{y}) - B(\boldsymbol{\eta})\right) \exp\left(\sum_{j=1}^{m} p_j T_j(\mathbf{y})\right) \\ &\propto h(\mathbf{y}) \exp\left(\sum_{j=1}^{m} (\eta_j + p_j) T_j(\mathbf{y}) - B(\boldsymbol{\eta})\right). \end{aligned}$$

5. Multinomial population distribution

Consider for example the k-dimensional multinomial distribution $\mathbf{y} = (y^1, y^2, \ldots, y^k)$ with number of trials l and probabilities $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, i.e.,

$$f_p(\mathbf{y}|l, \boldsymbol{\theta}) = \frac{l!}{y^1! \ldots y^k!} \exp\left(\sum_{i=1}^{k} y^i \log \theta_i\right),$$

where $\sum_{i=1}^{k} \theta_i = 1$ and $\sum_{i=1}^{k} y^i = l$. The vector of canonical parameters is $\boldsymbol{\eta} = (\log\theta_1, \ldots, \log\theta_k)$, with parameter space $\boldsymbol{\Psi} \subset \mathbb{R}^k$ such that $\sum_{i=1}^{k} \exp(\eta_i) = \sum_{i=1}^{k} \theta_i = 1$, and $T_j(\mathbf{y}) = y^j$.

Let the population inclusion probabilities have expectations

$$E_p(\pi|\mathbf{y}) \propto \exp\left(\sum_{i=1}^{k} p_i y^i\right),$$

then the sample distribution of the vector $\mathbf{y}$ is again multinomial with canonical parameters $\boldsymbol{\eta}^* = (\log\theta_1 + p_1, \ldots, \log\theta_k + p_k)$, provided $\boldsymbol{\eta}^*$ lies in $\boldsymbol{\Psi}$. The vector of probabilities for the sample distribution is thus $\boldsymbol{\theta}^* = (\theta_1 e^{p_1}, \ldots, \theta_k e^{p_k})$. By defining a vector $\mathbf{e} = (e^{p_1}, \ldots, e^{p_k})$, we can write $\boldsymbol{\theta}^*$ as an elementwise product $\boldsymbol{\theta}^* = \boldsymbol{\theta} \cdot \mathbf{e}$. The sample covariance matrix thus becomes

$$D_s\mathbf{y} = l(diag(\boldsymbol{\theta} \cdot \mathbf{e}) - (\boldsymbol{\theta} \cdot \mathbf{e})(\boldsymbol{\theta} \cdot \mathbf{e})').$$

The correlation coefficients between the variables y^i and y^j in the population and the sample are, respectively,

$$\rho_p(y^i, y^j) = -\sqrt{\frac{\theta_i\theta_j}{(1-\theta_i)(1-\theta_j)}} \text{ and } \rho_s(y^i, y^j) = -\sqrt{\frac{\theta_i e^{p_i}\theta_j e^{p_j}}{(1-\theta_i e^{p_i})(1-\theta_j e^{p_j})}}.$$

So, if p_i and p_j are positive, i.e., objects with larger y^i and y^j tend to be selected more often, then the negative correlation between y^i and y^j in the sample is stronger than that in the population.

6. Multivariate normal population distribution

The multivariate normal distribution belongs to the exponential family. Its density function equals

$$f_p(\mathbf{y}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right),$$

where $\boldsymbol{\mu}$ is the vector of expectations and $\boldsymbol{\Sigma}$ is the covariance matrix. We will derive its sample distribution explicitly.

Suppose that the inclusion probabilities are again in exponential form, but we now present them in matrix form

$$E_p(\pi|\mathbf{y}) \propto \exp\left(\mathbf{y}'\mathbf{A}\mathbf{y} + \mathbf{b}'\mathbf{y}\right), \tag{8}$$

where $\mathbf{A}$ is a $(k \times k)$ symmetric matrix and $\mathbf{b}$ is a $(k \times 1)$ vector. After applying (1) we see, given that the matrix $\left(\boldsymbol{\Sigma}^{-1} - 2\mathbf{A}\right)^{-1}$ is positive definite, the sample distribution is in this case again normal with the vector of expectations $\boldsymbol{\lambda}$ and the covariance matrix $\boldsymbol{\Omega}$ having the forms

$$\boldsymbol{\lambda} = \left(\boldsymbol{\Sigma}^{-1} - 2\mathbf{A}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{b}\right), \tag{9}$$

$$\boldsymbol{\Omega} = \left(\boldsymbol{\Sigma}^{-1} - 2\mathbf{A}\right)^{-1}. \tag{10}$$

From the above expressions we can make some conclusions on the relationship between the population and the sample covariance matrices in the case of a normal distribution:

- The sample covariance matrix is different from the population covariance matrix only if the matrix $\mathbf{A}$ is different from the matrix of zeros, that is, if the expectation of inclusion probabilities depends on the squares and products of study variables. If $\mathbf{A} = 0$, then the mean of the distribution changes but not the structure of dependencies.
- If variables are independent in the population, i.e., $\boldsymbol{\Sigma}$ is diagonal, then independence is preserved in the sample iff $\mathbf{A}$ is also diagonal (including $\mathbf{A} = 0$).
- With the choices of $\mathbf{A}$, the structure of dependencies between variables can drastically change: dependent variables can become independent, and vice versa, the sign of the covariance can change, etc.

So, with a normal population distribution and inclusion probabilities having the form (8), parameters of the sample distribution (including covariance matrix) can be calculated analytically using expressions (9) and (10). The same relationships can be used in an opposite way to derive population parameters from the sample parameters. It is important to note that the sample parameters can be estimated in the classical way from a sample by the sample mean and the sample covariance matrix.

Example 2. To illustrate how different the population and sample correlations may be, let us consider a population of $N = 10000$ objects with two study variables, y and z. Let the study variables follow a standard normal distribution with correlation coefficient r. We take a sample of 1000 objects with exponential inclusion probabilities of the form (8) with $\mathbf{A} = \begin{pmatrix} -1 & 1/2 \\ 1/2 & -1 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 1 & 1 \end{pmatrix}'$, and use Pareto sampling. Table 1 shows the population correlation r, the theoretical sample correlation $\tilde{r}$ calculated with the help of relationship (10) and the empirical sample correlation $\hat{r}$,

$$\hat{r} = \frac{\sum_{i\in s}[(y_i - \bar{y})(z_i - \bar{z})]}{\sqrt{\sum_{i\in s}(y_i - \bar{y})^2}\sqrt{\sum_{i\in s}(z_i - \bar{z})^2}},$$

averaged over 1000 repetitions. The theory states, that $\hat{r}$ is a consistent estimator for $\tilde{r}$, but not for r. We see that both strength and direction of the correlation may drastically change in the sample as compared to the population, negative correlation can become positive and independent variables

TABLE 1. Population and sample correlation coefficients

r	$\tilde{r}$	$\hat{r}$
-1	-1	-1
-0.8	-0.26	-0.25
-0.6	0.02	0.01
-0.4	0.16	0.17
-0.2	0.26	0.27
0	0.33	0.33
0.2	0.40	0.39
0.4	0.46	0.46
0.6	0.54	0.53
0.8	0.67	0.68
1	1	1

in the population can become correlated in the sample. We also see that $\hat{r}$ estimates $\tilde{r}$ very well.

7. Summary

In this paper we considered the multivariate exponential family. We showed that this family is closed under informative sampling if the inclusion probabilities are of exponential form. We derived sample distributions for the multinomial and multivariate normal distributions, and presented formulas of the respective covariance matrices. We showed that informative sampling can drastically change the dependence structure in the sample as compared to that in the population; sampling can even make dependent variables independent and vice versa. However, special cases exist where informative sampling does not influence covariances between variables. For example, independence between variables is preserved if the inclusion probabilities depend on the study variables in a multiplicative way.

Acknowledgements

The research was partially supported by Estonian Science Foundation Grant 8789. The author also thanks the supervisor Imbi Traat for her very helpful comments and suggestions, as well as the anonymous referee for several valuable remarks concerning the text of this paper.

References

Aru, J. (2008), *Influence of informative sampling on covariance between variables*, in: Baltic-Nordic Workshop on Survey Sampling Theory and Methodology, Kuressaare, 25–29 August 2008. Tallinn: Statistics Estonia, 2008, pp. 50–55.

Lehman, E. L., and Casella, G. (1998), *Theory of Point Estimation*, Second Edition, Springer, New York.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998), *Parametric distributions of complex survey data under informative probability sampling*, Statist. Sinica **8**, 1087–1114.

Pfeffermann, D., and Sverchkov, M. (1999), *Parametric and semi-parametric estimation of regression models fitted to survey data*, Indian J. Statist. Ser. B, Special Issue on Sample Surveys, **61**, 166–186.

Traat, I. (2003), *On the estimation of finite population covariance matrix*, Statistics in Transition, **6**, 67–82.

Traat, I., Bondesson, L., and Meister, K. (2004), *Sampling design and sample selection through distribution theory*, J. Statist. Plann. Inference, **123**, 395–413.

Statistics Estonia, 15174 Tallinn, Estonia
E-mail address: julia.aru@stat.ee