Communications in Statistics—Theory and Methods, 36: 1607–1610, 2007

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920601125813

Taylor & Francis
Taylor & Francis Group

A Note on Sample Size Determination for the Estimation of the Mean Vector of a Multivariate Population

WEN CUI[1] AND FEIQI ZHU[2]

[1]Department of Computer Information Systems & Quantitative Methods, McCoy College of Business Administration, Texas State University – San Marcos, San Marcos, Texas, USA
[2]Systems and Technology Group, IBM Corp., Austin, Texas, USA

In this article, we present a straightforward Bonferroni approach for determining sample size for estimating the mean vector of a multivariate population under two scenarios: (1) *a pre-specified overall confidence level is desired; and* (2) *a pre-specified confidence level needs to be guaranteed for each individual variable. It is demonstrated that correlation between variables helps reduce the sample size. The formula to calculate the reduced sample size is derived. A binormal example is presented to illustrate the effect of correlation on sample size reduction for various values of the correlation coefficient.*

Keywords Bonferroni Inequality; Correlation; Sample size determination; Sample size reduction.

Mathematics Subject Classification Primary 62H12; Secondary 62D05.

1. Introduction

Sample size determination is an important step in survey planning. For estimating the mean of a univariate population, methods have been proposed and many of them have been standard in practice. For example, the most common formula is

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{d^2}, \tag{1}$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$th percentile of the standard normal distribution, σ^2 is the population variance, and d is a pre-specified allowable margin of error. In addition,

Received August 12, 2006; Accepted September 15, 2006
Address correspondence to Wen Cui, Department of Computer Information Systems & Quantitative Methods, McCoy College of Business Administration, Texas State University – San Marcos, 601 University Drive, San Marcos, TX 78666, USA; E-mail: jcui@txstate.edu

practical methods such as two-stage methods (Cohran, 1977; Gillett, 1989) based on Stein (1945) are also available for cases in which the population variance σ^2 is unknown.

For survey studies involving estimation of the mean vector of a multivariate population, a general approach is to choose the maximum of individual sample sizes computed from (1) or from variants of (1) (Ballenger and McCune, 1990; Cohran, 1977). These methods were suggested for the scenario in which individual confidence levels and margin of errors need to be secured. In this article, we go one step further to consider the scenario in which the overall confidence levels and margin of errors need to be secured. We present a Bonferroni approach to determine the sample size and show that the maximum of the individual sample sizes guarantees the desired confidence levels in both scenarios. In addition, we demonstrate that correlation between variables helps reduce the sample size. A general formula is derived for calculating the reduced sample size and an illustrative example is then presented for a binormal case.

2. Bonferroni Approach

Let $X_1, X_2, \ldots, X_p$ be p dependent variables and X_i be from a population with unknown mean μ_i and known variance σ_i^2. Let A_i denote the confidence interval of μ_i, that is, $A_i = [\overline{X}_i - d_i, \overline{X}_i + d_i]$, d_i be the margin of error for estimating μ_i, and $P(A_i)$ denote the corresponding confidence level. Therefore, $P(\bigcap_{i=1}^{p} A_i)$ is the overall confidence level.

In the scenario that a pre-specified overall confidence level needs to be guaranteed, set the overall confidence level at $(1-\alpha)100\%$, i.e., $P(\bigcap_{i=1}^{p} A_i) = 1-\alpha$. Denote the individual confidence level as $P(A_1) = P(A_2) = \cdots = P(A_p) = 1-\alpha'$. By Bonferroni Inequality, $P(\bigcap_{i=1}^{p} A_i) \geq \sum_{i=1}^{p} P(A_i) - (p-1)$. Thus, $1-\alpha \geq \sum_{i=1}^{p}(1-\alpha') - (p-1) = (p - p\alpha') - (p-1)$. This implies $\alpha' \geq \frac{\alpha}{p}$. Then, from (1), each individual sample size n_i is

$$n_i = \frac{z_{\alpha'/2}^2 \sigma_i^2}{d_i^2}, \quad \text{where } \alpha' = \frac{\alpha}{p}, \ i = 1, 2, \ldots, p. \tag{2}$$

We can choose the overall sample size to be the maximum of the set of individual sample sizes calculated from (2). That is, $n_{\text{overall}} = n_{\max} = \max\{n_1, n_2, \ldots, n_p\}$. Since $n_i \leq n_{\max}$ for each $i = 1, 2, \ldots, p$, therefore, $P_{\text{actual}}(A_i) \geq 1-\alpha'$. By Bonferroni Inequality,

$$P\left(\bigcap_{i=1}^{p} A_i\right) \geq \sum_{i=1}^{p} P_{\text{actual}}(A_i) - (p-1) \geq p(1-\alpha') - (p-1) = 1 - p\alpha' = 1-\alpha. \tag{3}$$

The result in (3) shows that the maximum of the individual sample sizes computed from (2) guarantees that the overall confidence level is at least $(1-\alpha)100\%$. However, while the Bonferroni approach is simple and straightforward, it can lead to an undesirably large sample size when p is large because α' can be very small in that case.

In the scenario that the confidence level for each individual variable needs to be guaranteed, let the individual confidence level $P(A_1) = P(A_2) = \cdots = P(A_p) = 1-\alpha$.

Individual sample sizes, n_i for $i = 1, 2, \ldots, p$, can be obtained similarly from (1). Again, we choose the overall sample size to be the maximum of the p individual sample sizes, i.e., $n_{\text{overall}} = n_{\max} = \max\{n_1, n_2, \ldots, n_p\}$. It can be easily seen that the actual individual confidence level is guaranteed to be at least $(1 - \alpha)100\%$, because of each $n_i \leq n_{\max}$, $i = 1, 2, \ldots, p$. By Bonferroni Inequality, the overall confidence level is

$$P\left(\bigcap_{i=1}^{p} A_i\right) \geq \sum_{i=1}^{p} P_{\text{actual}}(A_i) - (p-1) \geq p(1-\alpha) - (p-1) = 1 - p\alpha, \tag{4}$$

When p is small, a fairly good overall confidence level can be achieved while each individual confidence level is guaranteed. However, if p is large, the overall confidence level is not really meaningful, because $1 - p\alpha$ may become zero or negative in this case. For example, when α is chosen to be 0.05 and p is 20, the only thing we know from (4) is that the overall confidence level is at least 0%. One other drawback of the Bonferroni approaches is that it does not take into consideration of the correlation and, therefore, can lead to large sample size which may result in an excessive cost on surveys. Next we discuss how to incorporate the correlation and how the correlation helps reduce sample size.

3. Sample Size Reduction

We consider a multinormal case in which variables are correlated with each other and one variable is of primary interests of researchers. Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)' \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a p-dimensional multinormal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$ and a positive definite covariance matrix $\boldsymbol{\Sigma}$. Suppose, without loss of generality, X_1 is the primary variable. Let $\mathbf{X} = \begin{bmatrix} X_1 \\ \mathbf{X_{(2)}} \end{bmatrix}$, $\mu = \begin{bmatrix} \mu_1 \\ \boldsymbol{\mu}_{(2)} \end{bmatrix}$, and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma_{22}} \end{bmatrix}$. It is known (Anderson, 1984) that the marginal distribution of X_1 is $N(\mu_1, \sigma_1^2)$, and the conditional distribution of X_1 given $\mathbf{X}_{(2)} = (X_2, X_3, \ldots, X_p)'$ is the multivariate normal distribution, $MN(\mu_1 + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}(\mathbf{X_{(2)}} - \boldsymbol{\mu_{(2)}}), \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\boldsymbol{\Sigma_{12}'})$. Now, suppose that the sample size for X_1 is n_1 and the sample size for X_1 given $\mathbf{X_{(2)}}$ is $n_{1|(2)}$. Then for the given confidence level of $(1 - \alpha)100\%$ and the allowable margin of error d, by (1), we have

$$n_1 = z_{\alpha/2}^2 \sigma_1^2 / d^2 \tag{5}$$

$$n_{1|(2)} = z_{\alpha/2}^2 (\sigma_1^2 - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\boldsymbol{\Sigma_{12}'}) / d^2. \tag{6}$$

Thus, the reduced sample size is

$$n_{1|(2)} = n_1\left(1 - \frac{\boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\boldsymbol{\Sigma_{12}'}}{\sigma_1^2}\right). \tag{7}$$

For illustration purpose, we look at a binormal case, in which $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$, where ρ is the correlation coefficient between X_1 and X_2. Thus, the sample size for X_1 given X_2 is $n_{1|2} = n_1(1 - \rho^2)$, which says the sample size for the primary variable can be reduced by a multiple of ρ^2 in the binormal case. In Table 1, we give the magnitude of the reduction for various values of ρ.

Table 1
Reduction of sample size for X_1 given ρ between X_1 and X_2.
Suppose $n_1 = 100$

$\lvert\rho\rvert$	0.1	0.25	0.5	0.75	0.9
Reduction	1%	6.3%	25%	56.3%	81%
$n_{1\lvert 2}$	99	94	75	44	19

It can be seen from Table 1 that tremendous reduction can be obtained even when the correlation is moderate or moderately high: the reduction is 25% when the correlation coefficient is ± 0.25, and the reduction is over 50% when the correlation is ± 0.75.

In (7), we assume that the covariance matrix is known for simplicity. When it is unknown, estimators may be obtained from a pilot sample. One simple and largely accepted estimator is the unbiased estimator, $\widehat{\mathbf{\Sigma}} = \mathbf{S}/(\mathbf{n_p} - \mathbf{1})$, where n_p is the size of the pilot sample and $\mathbf{S} = (\mathbf{X} - \overline{\mathbf{X}}1')(\mathbf{X} - \overline{\mathbf{X}}1')'$, $\mathbf{1}$ is a $n_p \times 1$ vector of ones. However, when n_p is small, it can be statistically inefficient in the sense that improved estimators can be found. Moreover, if $n_p < p$, $\widehat{\Sigma}$ will be singular and it is impossible to compute the inverse of the covariance estimator. Actually, many improved estimators (mainly under decision theoretic perspective) have been proposed. For example, the minimax estimator proposed by James and Stein (1961) was shown to be uniformly better than $\widehat{\Sigma} = S/(n_p - 1)$ in terms of Stein's loss. The estimator is of the form $\widehat{\Sigma}^{JS} = KDK'$, where $D = diag(d_1, \ldots, d_p)$, $d_i = 1/(n_p + p - 2i + 1)$ $(i = 1, \ldots, p)$, K is a lower triangular matrix with positive diagonal elements and $KK' = S$. Compared with $\widehat{\Sigma} = S/(n_p - 1)$, this estimator is not computationally straightforward, even though the triangular decomposition of S is not a problem given the powerful computers we have today. Actually, many other minimax estimators improved upon James and Stein's estimator have been developed later (for example, see Dey and Srinivasan, 1985; Stein, 1975). These estimators usually offer better efficiency than $\widehat{\Sigma} = S/(n_p - 1)$, but normally involve more computational efforts and more complicate structures.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. 2nd ed. New York: John Wiley & Sons.

Ballenger, J. K., McCune, S. K. (1990). A procedure for determining sample size for multiple population parameter studies. *J. Market. Educ.* 12(Fall):30–33.

Cohran, W. G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.

Dey, D. K., Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* 13:1581–1591.

Gillett, R. (1989). Confidence interval construction by Stein's method: a practical and economical approach to sample size determination. *J. Market. Res.* 26(May):237–240.

James, W., Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Mathemat. Statist. Probab.* 1:361–380.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Mathemat. Statist.* 16(3):243–258.

Stein, C. (1975). Estimation of a covariance matrix. Rietz Lecture, 39th Annual Meeting IMS, Atlanta, Georgia.