A TWO-STAGE PROCEDURE FOR FIXED-SIZE CONFIDENCE REGION OF A CONDITIONAL MEAN

Nakao, Hiroyuki
Graduate School of Mathematics, Kyushu University

Hyakutake, Hiroto
Graduate School of Mathematics, Kyushu University

Kanda, Takashi
Department of Environmental Design, Hiroshima Institute of Technology

https://doi.org/10.5109/13483

出版情報：Bulletin of informatics and cybernetics. 31 (1), pp.101-108, 1999-03. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 31, No. 1, 1999

A TWO-STAGE PROCEDURE FOR FIXED-SIZE CONFIDENCE REGION OF A CONDITIONAL MEAN

By

Hiroyuki NAKAO*, Hiroto HYAKUTAKE*
and
Takashi KANDA†

Abstract

A two-stage procedure for constructing fixed-size confidence region of a conditional mean in multivariate normal distribution is proposed. A distribution of a statistic, which is appeared in the two-stage procedure, is approximated by an F-distribution. The accuracy of the approximation is examined by simulation. A numerical example is also given.

Key Words and Phrases: conditional mean, confidence region, multivariate normal distribution, simulation, two-stage procedure.

1. Introduction

Let $\boldsymbol{x}_i$, $(i = 1, 2, \cdots)$ be independent and identically distributed random vectors having the p-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, that is $N_p(\boldsymbol{\mu}, \Sigma)$. The covariance matrix Σ is a positive definite matrix. Let $\boldsymbol{x}_i$ be partitioned as $(\boldsymbol{x}_{i1}', \boldsymbol{x}_{i2}')'$, $(r+s)\times 1$, and $\boldsymbol{\mu}_j$ and Σ_{jk} $(j, k = 1, 2)$ be the corresponding partitions of $\boldsymbol{\mu}$ and Σ, respectively. The conditional mean of $\boldsymbol{x}_{02}$ given $\boldsymbol{x}_{01}$ is $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_{01} - \boldsymbol{\mu}_1)$. $\Sigma_{21}\Sigma_{11}^{-1}$ is a regression matrix. In the usual linear regression problem, Chatterjee (1962) developed Stein's (1945) two-stage procedure for the fixed-size confidence region of the regression parameters (see Ghosh, Mukhopadhyay and Sen, 1997). Healy (1956) gave a multivariate two-stage procedure for fixed-size confidence regions of $\boldsymbol{\mu}$.

The problem is to determine the sample size satisfying

$$P\{(\hat{\boldsymbol{\mu}}_{2.1} - \boldsymbol{\mu}_{2.1})'(\hat{\boldsymbol{\mu}}_{2.1} - \boldsymbol{\mu}_{2.1}) \leq d^2\} \geq 1 - \alpha, \tag{1}$$

where $d > 0$ and α $(0 < \alpha < 1)$ are given and $\hat{\boldsymbol{\mu}}_{2.1}$ is an estimate of $\boldsymbol{\mu}_{2.1}$. Let $\bar{\boldsymbol{x}}_n' = (\bar{\boldsymbol{x}}_{1,n}'\bar{\boldsymbol{x}}_{2,n}')$ be the usual sample mean with sample size n. If Σ were known, it is easily seen that

* Graduate School of Mathematics, Kyushu University, Ropponmatsu, Fukuoka 810-8560, Japan

† Department of Environmental Design, Hiroshima Institute of Technology, Hiroshima 731-5193, Japan

$$
\begin{aligned}
& P\{(\hat{\mu}_{2.1} - \mu_{2.1})'(\hat{\mu}_{2.1} - \mu_{2.1}) \leq d^2\} \\
\geq\ & P\{n(\hat{\mu}_{2.1} - \mu_{2.1})'\Sigma_{22.1}^{-1}(\hat{\mu}_{2.1} - \mu_{2.1}) \leq nd^2/\lambda\} \\
=\ & G_s(nd^2/\lambda),
\end{aligned}
$$

where G_s is the cumulative distribution function (c.d.f.) of the chi-square distribution with s degrees of freedom (d.f.), λ is the maximum characteristic root (ch.r.) of $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, and $\hat{\mu}_{2.1} = \bar{\boldsymbol{x}}_{2,n} + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_{01} - \bar{\boldsymbol{x}}_{1,n})$. Hence, if the sample size n is chosen such that

$$
n \geq n^* = u\lambda/d^2, \tag{2}
$$

where u is the solution of the equation $G_s(u) = 1 - \alpha$, then (1) is satisfied.

When Σ is unknown and the sample size n is fixed, Σ is estimated by

$$
S_n = \begin{pmatrix} S_{11,n} & S_{12,n} \\ S_{21,n} & S_{22,n} \end{pmatrix} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)'
$$

and the statistic

$$
t_n = \frac{n-s}{(n-r-1)s}\frac{(\hat{\mu}_{2.1,n} - \mu_{2.1})'S_{22.1,n}^{-1}(\hat{\mu}_{2.1,n} - \mu_{2.1})}{1/n + Q_n/(n-1)}, \tag{3}
$$

is used for constructing the confidence region of the conditional mean, where $Q_n = (\boldsymbol{x}_{01} - \bar{\boldsymbol{x}}_{1,n})'S_{11,n}^{-1}(\boldsymbol{x}_{01} - \bar{\boldsymbol{x}}_{1,n})$. But, there is no fixed sample procedure satisfying (1), when Σ is unknown. So, we need at least two-stage procedure. Since the size of the confidence region depends on $S_{22.1,n}$ and Q_n, it will be impossible to give a two-stage procedure which is asymptotically efficient. In this paper, we propose a two-stage procedure to satisfy (1) approximately. The two-stage procedure and its property are given in Section 2. In Section 3, the accuracy of approximation in the procedure is examined by simulation. In Section 4, we give an example by using the data of Potthoff and Roy (1964).

2. Two-stage procedure

Since Σ is unknown, we give a two-stage procedure and its asymptotic efficiency. In the two-stage procedure, take m observations as the initial sample and compute S_m. We define the total sample size N as

$$
N = \max\{m,\ [c\ell/d^2] + 1\} + 1, \tag{4}
$$

where $[b]$ denotes the greatest integer less than b, ℓ is the maximum ch.r. of $S_{22.1,m} = S_{22,m} - S_{21,m}S_{11,m}^{-1}S_{12,m}$,

$$c = \frac{(m-r-1)s}{m-s} f\{1 + (\boldsymbol{x}_{01} - \bar{\boldsymbol{x}}_{1,m})' S_{11,m}^{-1} (\boldsymbol{x}_{01} - \bar{\boldsymbol{x}}_{1,m})\}, \tag{5}$$

and f is the upper $100(1-\alpha)\%$ point of the F-distribution with $(s, m-p)$ d.f. Next take $N-m$ additional observations and compute

$$\hat{\boldsymbol{\mu}}_{2.1,N} = \bar{\boldsymbol{x}}_{2,N} + S_{21,N} S_{11,N}^{-1} (\boldsymbol{x}_{01} - \bar{\boldsymbol{x}}_{1,N}) \tag{6}$$

based on N observations. By (4), it follows that

$$\begin{aligned} & P\{(\hat{\boldsymbol{\mu}}_{2.1,N} - \boldsymbol{\mu}_{2.1})'(\hat{\boldsymbol{\mu}}_{2.1,N} - \boldsymbol{\mu}_{2.1}) \le d^2\} \\ \ge\ & P\{(N-1)(\hat{\boldsymbol{\mu}}_{2.1,N} - \boldsymbol{\mu}_{2.1})'(\hat{\boldsymbol{\mu}}_{2.1,N} - \boldsymbol{\mu}_{2.1}) \le c\ell\} \\ \ge\ & P\{\frac{m-s}{(m-r-1)s} \frac{(\hat{\boldsymbol{\mu}}_{2.1,N} - \boldsymbol{\mu}_{2.1})' S_{22.1,m}^{-1} (\hat{\boldsymbol{\mu}}_{2.1,N} - \boldsymbol{\mu}_{2.1})}{1/N + Q_m/(N-1)} \le f\}. \end{aligned} \tag{7}$$

If $\boldsymbol{\mu}_{2.1}$ is estimated by m (fixed sample size) observations instead of N in (6), then the statistic t_m is distributed as the F-distribution with $(s, m-p)$ d.f. exactly (see e.g. Siotani, Hayakawa and Fujikoshi, 1985). If $\bar{\boldsymbol{x}}_{1,N}$, $S_{11,N}$ and $S_{22.1,N}$ were fixed, the estimate $\hat{\boldsymbol{\mu}}_{2.1,N}$ in (6) is conditionally distributed as s-variate normal with mean $\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_{01} - \bar{\boldsymbol{x}}_{1,N})$ and covariance matrix $\{1/N + Q/(N-1)\}\Sigma_{22.1}$. $(m-1)S_{22.1}$ is distributed as Wishart with covariance matrix $\Sigma_{22.1}$ and $(m-r)$ d.f. Hence the distribution of the statistic

$$\frac{m-s}{(m-r-1)s} \frac{(\hat{\boldsymbol{\mu}}_{2.1,N} - \boldsymbol{\mu}_{2.1})' S_{22.1,m}^{-1} (\hat{\boldsymbol{\mu}}_{2.1,N} - \boldsymbol{\mu}_{2.1})}{1/N + Q_m/(N-1)} \tag{8}$$

in (7) may be approximated by the F-distribution. That is, the right-hand side of (7) is approximated by $1-\alpha$, then (1) is satisfied.

Takada (1988) showed that the usual multivariate two-stage procedure is asymptotically efficient when the initial sample size satisfies appropriate condition. We examine the asymptotic efficiency when

$$m \to \infty \ \text{ and } \ d^2 m \to 0 \ \text{ as } \ d \to 0. \tag{9}$$

It follows that $\lim_{d\to 0} sf = u$ and $\lim_{d\to 0} \ell = \lambda$ under (9) and

$$\frac{c\ell/d^2 + 1}{n^*} \le \frac{N}{n^*} \le \frac{c\ell/d^2 + 1 + m}{n^*}$$

by (4). Hence we have the asymptotic efficiency as

$$\lim_{d\to 0}\frac{E(N)}{n^*} = 1 + (\boldsymbol{x}_{01} - \boldsymbol{\mu}_1)'\Sigma_{11}^{-1}(\boldsymbol{x}_{01} - \boldsymbol{\mu}_1). \tag{10}$$

The procedure (4) is asymptotically efficient if $\boldsymbol{x}_{01} = \boldsymbol{\mu}_1$.

3. Accuracy of approximation

The distribution of the statistic in (8) appeared in the two-stage procedure is approximated by the F-distribution. In order to examine the accuracy of the approximation, a simulation experiment was conducted. In the simulation, the coverage probability (1) is also estimated. The simulation results are based on 10000 replications and are given in Table 1. The table lists the proportion of the statistic in (8) which is less than f. The values in the parentheses are the proportion of

$$(\hat{\boldsymbol{\mu}}_{2.1} - \boldsymbol{\mu}_{2.1})'(\hat{\boldsymbol{\mu}}_{2.1} - \boldsymbol{\mu}_{2.1}) \le d^2.$$

The confidence coefficient is chosen to be $1 - \alpha = 0.95$. It is chosen that $p = 2,\ 3$, $d = 0.4,\ 0.6,\ 0.8$ and $\boldsymbol{\mu} = \mathbf{0}$. For $p = 2$ $(r = s = 1)$, we choose $m = 10,\ 15,\ 25$, $\boldsymbol{x}_{01} = 0.2,\ 0.5,\ 1.0$ and $\Sigma = \Sigma_j$ $(j = 1, 2, 3)$, where

$$\Sigma_1 = \begin{pmatrix} 2.0 & 0.5 \\ 0.5 & 2.0 \end{pmatrix},\quad \Sigma_2 = \begin{pmatrix} 2.0 & 1.0 \\ 1.0 & 2.0 \end{pmatrix},\quad \Sigma_3 = \begin{pmatrix} 1.6 & 1.0 \\ 1.0 & 2.0 \end{pmatrix}.$$

For $p = 3$ $(r = 2, s = 1)$, we choose $\boldsymbol{x}_{01}' = (0.2, 0.2),\ (0.2, 0.5),\ (0.5, 0.5),\ (0.5, 1.0)$, $m = 15,\ 25$ and $\Sigma = \Sigma_j$ $(j = 4, 5)$, where

$$\Sigma_4 = \begin{pmatrix} 2.0 & 1.0 & 1.0 \\ 1.0 & 2.0 & 1.0 \\ 1.0 & 1.0 & 2.0 \end{pmatrix},\quad \Sigma_5 = \begin{pmatrix} 1.5 & 1.0 & 0.6 \\ 1.0 & 2.0 & 1.0 \\ 0.6 & 1.0 & 1.8 \end{pmatrix}.$$

Table 1. Accuracy of approximation

Σ_1

$\boldsymbol{x}_{01}$	d	$m = 10$	$m = 15$	$m = 25$
0.2	0.4	.9364 (.9406)	.9404 (.9455)	.9424 (.9475)
	0.6	.9302 (.9444)	.9352 (.9542)	.9409 (.9766)
	0.8	.9290 (.9645)	.9363 (.9800)	.9405 (.9966)
0.5	0.4	.9327 (.9384)	.9378 (.9431)	.9413 (.9461)
	0.6	.9305 (.9448)	.9360 (.9524)	.9414 (.9695)
	0.8	.9283 (.9591)	.9368 (.9766)	.9403 (.9954)
1.0	0.4	.9272 (.9301)	.9366 (.9401)	.9427 (.9453)
	0.6	.9261 (.9359)	.9391 (.9481)	.9435 (.9573)
	0.8	.9271 (.9484)	.9376 (.9653)	.9431 (.9859)

Table 1. Accuracy of approximation (continued)

Σ_2

x_{01}	d	$m = 10$	$m = 15$	$m = 25$
0.2	0.4	.9321 (.9378)	.9372 (.9445)	.9410 (.9487)
	0.6	.9290 (.9511)	.9360 (.9627)	.9403 (.9871)
	0.8	.9292 (.9734)	.9352 (.9894)	.9405 (.9995)
0.5	0.4	.9322 (.9380)	.9363 (.9432)	.9413 (.9480)
	0.6	.9268 (.9458)	.9362 (.9600)	.9407 (.9837)
	0.8	.9271 (.9688)	.9356 (.9855)	.9404 (.9987)
1.0	0.4	.9288 (.9343)	.9366 (.9404)	.9429 (.9464)
	0.6	.9274 (.9402)	.9393 (.9521)	.9423 (.9669)
	0.8	.9292 (.9595)	.9359 (.9750)	.9422 (.9928)

Σ_3

x_{01}	d	$m = 10$	$m = 15$	$m = 25$
0.2	0.4	.9299 (.9373)	.9402 (.9472)	.9419 (.9526)
	0.6	.9287 (.9525)	.9365 (.9651)	.9405 (.9897)
	0.8	.9279 (.9766)	.9357 (.9926)	.9408 (.9999)
0.5	0.4	.9304 (.9377)	.9377 (.9425)	.9418 (.9492)
	0.6	.9297 (.9493)	.9367 (.9627)	.9402 (.9861)
	0.8	.9277 (.9782)	.9365 (.9889)	.9397 (.9989)
1.0	0.4	.9286 (.9337)	.9359 (.9395)	.9434 (.9465)
	0.6	.9289 (.9423)	.9383 (.9514)	.9421 (.9675)
	0.8	.9297 (.9605)	.9366 (.9757)	.9420 (.9927)

Σ_4

x'_{01}	d	$m = 15$	$m = 25$
(0.2, 0.2)	0.4	.8775 (.8879)	.9212 (.9334)
	0.6	.9020 (.9437)	.9286 (.9833)
	0.8	.9062 (.9806)	.9292 (.9978)
(0.2, 0.5)	0.4	.8477 (.8564)	.9094 (.9177)
	0.6	.8927 (.9290)	.9278 (.9762)
	0.8	.9062 (.9744)	.9290 (.9959)
(0.5, 0.5)	0.4	.8062 (.8145)	.8960 (.9045)
	0.6	.8802 (.9165)	.9264 (.9745)
	0.8	.9019 (.9705)	.9278 (.9958)
(0.5, 1.0)	0.4	.7235 (.7300)	.8467 (.8518)
	0.6	.8387 (.8617)	.9186 (.9552)
	0.8	.8857 (.9407)	.9253 (.9870)

Table 1. Accuracy of approximation (continued)

Σ_5

x'_{01}	d	$m = 15$	$m = 25$
(0.2, 0.2)	0.4	.8878 (.8988)	.9246 (.9362)
	0.6	.9032 (.9470)	.9297 (.9847)
	0.8	.9070 (.9825)	.9302 (.9981)
(0.2, 0.5)	0.4	.8438 (.8530)	.9117 (.9213)
	0.6	.8932 (.9295)	.9285 (.9782)
	0.8	.9073 (.9761)	.9294 (.9965)
(0.5, 0.5)	0.4	.8282 (.8374)	.9051 (.9147)
	0.6	.8864 (.9257)	.9269 (.9761)
	0.8	.9021 (.9731)	.9273 (.9963)
(0.5, 1.0)	0.4	.7237 (.7311)	.8487 (.8550)
	0.6	.8394 (.8645)	.9183 (.9571)
	0.8	.8871 (.9426)	.9247 (.9883)

Table 1 shows that all values are smaller than 0.95, namely the approximation of the distribution of the statistic (8) to the F-distribution is not good. But, when d and m are large, the table shows that (1) is satisfied. For example, if $d = 0.8$ and $m = 25$, all values in the parentheses are greater than 0.985. The coverage probability (1) is increasing in d and m and is decreasing in the distance of $\mathbf{0}$ (mean μ_1) and $\boldsymbol{x}_{01}$. This finding suggests that we may use the procedure if d and m are large and the distance is small. Note that the asymptotic efficiency in (10) is increasing in the Mahalanobis distance

$$(\boldsymbol{x}_{01} - \mu_1)'\Sigma_{11}^{-1}(\boldsymbol{x}_{01} - \mu_1).$$

Thus when the distance is small, then the both of asymptotic efficiency and accuracy of approximation are good, and the proposed method might be useful. The approximation for $p = 2$ is better than that for $p = 3$.

4. Example

We give a numerical example by using the data ($p = 4$) of Potthoff and Roy (1964), which are the dental measurement of the distance (mm) from the center of the pituitary to the pterygomaxillary fissure. x_i has been measured at age $2i + 6$ ($i = 1, 2, 3, 4$). Let $r = 3$, $s = 1$ and the first stage sample size m is selected as 20. The values of x_i's are given in Table 2.

Table 2. First stage samples

	x_1	x_2	x_3	x_4		x_1	x_2	x_3	x_4
1	21.0	20.0	21.5	23.0	11	24.5	25.0	28.0	28.0
2	21.0	21.5	24.0	25.5	12	26.0	25.0	29.0	31.0
3	20.5	24.0	24.5	26.0	13	21.5	22.5	23.0	26.5
4	23.5	24.5	25.0	26.5	14	23.0	22.5	24.0	27.5
5	21.5	23.0	22.5	23.5	15	25.5	27.5	26.5	27.0
6	20.0	21.0	21.0	22.5	16	20.0	23.5	22.5	26.0
7	21.5	22.5	23.0	25.0	17	24.5	25.5	27.0	28.5
8	23.0	23.0	23.5	24.0	18	22.0	22.0	24.5	26.5
9	20.0	21.0	22.0	21.5	19	24.0	21.5	24.5	25.5
10	16.5	19.0	19.0	19.5	20	23.0	20.5	31.0	26.0

The unbiased covariance matrix is

$$S_{20} = \begin{pmatrix} 5.102 & 3.388 & 5.184 & 4.845 \\ & 4.329 & 3.118 & 3.914 \\ & & 8.221 & 5.863 \\ & & & 6.802 \end{pmatrix}.$$

Let us take $d = 1.5$, $1 - \alpha = 0.95$, and

$$\boldsymbol{x}_{01} = (27.5,\ 28.0,\ 31.0)'$$

After computation, $\ell = S_{22.1,m} = 1.588$ and $c = 33.26$. Then the total sample size are computed as $N = 25$. The additional observations are taken and those values are given in Table 3.

Table 3. Second stage samples

	x_1	x_2	x_3	x_4
21	23.0	23.0	23.5	25.0
22	21.5	23.5	24.0	28.0
23	17.0	24.5	26.0	29.5
24	22.5	25.5	25.5	26.0
25	23.0	24.5	26.0	30.0

Based on 25 samples, $\bar{\boldsymbol{x}}_{25} = (21.98,\ 23.04,\ 24.44,\ 25.92)'$ and

$$S_{25} = \begin{pmatrix} 5.198 & 2.449 & 3.822 & 3.134 \\ & 3.936 & 2.794 & 3.722 \\ & & 6.819 & 5.193 \\ & & & 6.993 \end{pmatrix}.$$

Hence the estimator (6) is computed as $\hat{\mu}_{2.1,N} = 32.09$, and the confidence region (interval) is obtained as

$$30.59 \leq \mu_{2.1} \leq 33.59.$$

In fact, x_{04} is measured as 31.5, which is included in the region.

References

Chatterjee, S.K. (1962), Sequential inference of Stein type for a class of multivariate regression problem. *Ann. Math. Statist.*, **33**, 1039–1064.

Ghosh, M., Mukhopadhyay, N. and Sen, P.K. (1997), *Sequential Estimation*, Wiley, New York.

Healy, W.C.Jr. (1956), Two-sample procedures in simultaneous estimation. *Ann. Math. Statist.*, **27**, 687–702.

Potthoff, R.F. and Roy, S.N. (1964), A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.

Siotani, M., and Hayakawa, T. and Fujikoshi, Y. (1985), *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*, American Sciences Press, Columbus, Ohio.

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

Takada, Y. (1988), Two-stage procedures for a multivariate normal distribution. *Kumamoto J. Math.*, **1**, 1–8.

Received October 23, 1997
Revised November 2, 1998