This work was supported by the National Institutes of Health, Institute of General Medical Sciences, Grant No. GM-12868-04

A CENTRAL TOLERANCE REGION FOR THE MULTIVARIATE NORMAL DISTRIBUTION II

by

David G. Kleinbaum and S. John [1]

University of North Carolina
Institute of Statistics Mimeo Series No. 620

January 1969

---

[1] Now at the Australian National University, Canberra

A CENTRAL TOLERANCE REGION FOR THE MULTIVARIATE NORMAL DISTRIBUTION II

1. Introduction

In a recent paper of the same title, John (1968) considered the problem of determining from a random sample of size N from a p-variate normal distribution a region which with probability β includes the region

$$R = \{ x: \ (x - \mu)' \ \Sigma^{-1}(x - \mu) \leq \chi^2_p \ (1 - \alpha) \}$$

Here μ and Σ are respectively the mean vector and covariance matrix of the distribution considered and $\chi^2_p \ (1 - \alpha)$ is the number exceeded by a chi square of p degrees of freedom with probability $1 - \alpha$. The solution to the problem when μ and Σ are unknown is the region

$$R_{\bar{x},s} = \{x: \ (x - \bar{x})' \ S^{-1} \ (x - \bar{x}) \leq K_{N,n} \} ,$$

where $\bar{x}$ is the vector of sample means, S is the sample covariance matrix and $K_{N,n}$ is the number exceeded by

$$U = [\{ n \chi^2_p \ (1 - \alpha) \ / \ t_p \}^{\frac{1}{2}} + \{(x - \mu)' \ S^{-1}(x - \mu)\}^{\frac{1}{2}}]^2$$

with probability $1 - \beta$. Here t_p is a random variable distributed independently of $\bar{x}$ as the smallest root of a standard p x p Wishart matrix of $n = N - 1$ degrees of freedom. Since it is difficult to determine $K_{N,n}$ exactly, two approximations were given:

$$(1) \quad \hat{K}^{(1)}_{N,n} = \left\{ [n \chi^2_p (1 - \alpha)/t_p(0.5) \right\}^{\frac{1}{2}} + \{ N(n-p+1) \}^{-\frac{1}{2}} (np)^{\frac{1}{2}} \left\{ F_{p,n-p+1}(1-\beta) \right\}^{\frac{1}{2}}]^2$$

and

$$(2)\quad K^{(2)}_{N,n} = [\{ n\,\chi^2_p(1-\alpha)/(n-p-1)\}^{\frac{1}{2}} + \{N(n-p+1)\}^{-\frac{1}{2}}(np)^{\frac{1}{2}}\{ F_{p,n-p+1}(1-\beta)\}^{\frac{1}{2}}]^2$$

where $F_{p,n-p+1}(1-\beta)$ is the number exceeded with probability $1-\beta$ by a random variable having the F-distribution with p and n-p+1 degrees of freedom and $t_p(\beta)$ is the number which is exceeded by t_p with probability β. The corresponding regions approximating $R_{\bar{x},s}$ will henceforth be denoted by adding the appropriate superscript to $\hat{R}_{\bar{x},s}$.

In this paper we show that the above approximations to $K_{N,n}$ are inadequate unless n is unusually large and we propose two new approximations which are accurate. These results were obtained primarily by computer simulation, and only the case p = 2 was considered. Also we have allowed n to be independent of N, so that S may be an estimate of Σ obtained independently of the N observations in our sample.

2. Evidence for the Inadequacy of the Earlier Approximation

Since $t_p(0.5)$ and n-p-1 both represented central values of the t_p distribution, it was assumed that adequacy or inadequacy of (2) implied the adequacy or inadequacy of (1) (and vice versa). For the computer simulation it was assumed without loss of generality that the true parameters were $\mu = 0$ and $\Sigma = I$. The region R was thus the circle

$$R = \{ x:\ x'x \leq \chi^2_2\ (1-\alpha) \}\ .$$

For specified values of α, β, N and n, 100 independent sample values of $\bar{x}$ and S were generated by the method outlined in Section 5, and the proportion PR of $\hat{R}^{(2)}_{\bar{x},s}$'s that contained R was computed. The proportion PR was then compared with β. A few examples of the results are given in Table 1. For each specified value of α, β and N it was found that a value of n graater than 300 was required in order for the proportion PR to be equal to β. Also for

all values of n less than 300, PR was smaller than β.

TABLE 1[3] Evidence of the Inadequacy of Approximation $\hat{K}^{(2)}_{N,n}$ for (β, α, N) = (.95, .95, 25)

n	24	75	120	300	500	600
PR	.35	.74	.81	.87	.92	.95

3. The New Approximation to $K_{N,n}$ and Evidence in Support of It

The following approximation was shown by simulation to be adequate for p = 2:

$$(3)\quad \hat{K}^{(3)}_{N,n} = [\{ n\ \chi^2_p(1-\alpha)/t_p(\beta)\}^{\frac{1}{2}} + \{N(n-p+1)\}^{-\frac{1}{2}}\ (np)^{\frac{1}{2}}\{F_{p,n-p+1}(1-\beta)\}^{\frac{1}{2}}]^2$$

(A table of values of $t_p(\beta)$ for p = 2 has been prepared by Kleinbaum and John (1969). This approximation was suggested by the fact that the approximation (3) is exact in the two special cases $n = \infty$, $N = \infty$. (John, 1968). The method of simulation was the same as that used for studying the earlier approximations except that 200 runs were made for each given set of (α, β, N, n) in order to obtain more accuracy. However, it was necessary to compute β for specified values of $t_2(\beta)$. This was accomplished using the following formula obtained by John (1963):

$$\beta = [1-F_{2n}(2t_2(\beta))] - [\Gamma(\tfrac{1}{2})/\Gamma(\tfrac{1}{2}n)]\left[\frac{t_2(\beta)}{2}\right]^{\frac{1}{2}}(n-1)e^{-\frac{1}{2}t_2(\beta)}\ [1-F_{n+1}(t_2(\beta))],$$

where $F_m(t) = \Pr\{\chi^2_m \le t\}$.

[3] PR = $\frac{\#\text{ of times } \hat{R}_{\bar{x},s} \text{ contains R (100 runs)}}{100}$

Table 2 presents some of the results obtained for various values of α, β, N and n. It was generally found that the approximation was best for small n but different values of N and α produced no specific trend with regard to the accuracy of the approximation.

4. A Fourth Approximation

For the computer simulation described in the previous section, PR exceeded β for all but four of the parameter sets (α, β, N and n) used; the four exceptions can be attributed to sampling error. Thus approximation (3) was experimentally "safe." Nevertheless, this is difficult to prove mathematically. This leads us to a fourth approximation which can be mathematically proved to be safe, although it is generally larger than the approximation (3). Here we say that an approximation $\hat{K}_{N,n}$ is safe if

$$\Pr\{ \hat{R}_{\bar{x},s} \supset R_{\bar{x},s} \} \geq \beta$$

It is easy to see that this is so if

$$\Pr\{ U \leq \hat{K}_{N,n} \} \geq \beta.$$

Approximation (4) is given by

$$\hat{K}^{(4)}_{N,n} = \left[\left\{ \frac{n\chi^2_p(1-\alpha)}{t_p(\beta_1)} \right\}^{\frac{1}{2}} + \{N(n-p+1)\}^{-\frac{1}{2}} (np)^{\frac{1}{2}} \{F_{p,n-p+1}(1-\beta_2)\}^{\frac{1}{2}} \right]^2$$

where $\beta_1 + \beta_2 - 1 \geq \beta$

The proof that (4) is safe is as follows:

$$\Pr\{U \leq \left[\left\{ \frac{n\chi^2_p(1-\alpha)}{t_p(\beta_1)} \right\}^{\frac{1}{2}} + \{N(n-p+1)\}^{-\frac{1}{2}} (np)^{\frac{1}{2}} \{F_{p,n-p+1}(1-\beta_2)\}^{\frac{1}{2}} \right]^2 \}$$

$$\geq \Pr\{t_p \geq t_p(\beta_1) \text{ and } (\bar{x}-\mu)' S^{-1} (\bar{x}-\mu) \leq \{N(n-p+1)\}^{-1} (np) F_{p,n-p+1}(1-\beta_2)\}$$

$$= 1 - \Pr\{t_p < t_p(\beta_1) \text{ or } (\bar{x}-\mu)' S^{-1} (\bar{x}-\mu) > \{N(n-p+1)\}^{-1} (np) F_{p,n-p+1}(1-\beta_2) \}$$

$$\geq 1 - \Pr\{t_p < t_p(\beta_1)\}$$

$$- \Pr \{(\bar{x} - \mu)' S^{-1} (\bar{x} - \mu) > \{N(n-p-1)\}^{-1}(np)F_{p,n-p+1}(1-\beta_2)\}$$

$$= 1 - (1-\beta_1) - (1-\beta_2)$$

$$= \beta_1 + \beta_2 - 1$$

$$\geq \beta, \text{ for and } 0 \leq \beta_i \leq 1,\ i = 1, 2, \text{ if } \beta_1 + \beta_2 - 1 \geq \beta$$

In particular we may choose β_1 and β_2 so that $K_{N,n}$ is minimum. In this case $\beta_1 + \beta_2 - 1 = \beta$.

5. Outline of Simulation Method

Values of $\bar{x}$ were generated using a standard computer procedure for generating independent N(0,1) variables and then dividing each of a resulting pair by $\sqrt{N}$. S was generated using a computer technique by Odell and Feiveson (1966) and specialized for our problem as follows:

(i) Generate 3 independent N(0,1) variates U_1, U_2 and U_3.

(ii) Use U_1 and U_2 to generate two independent chi square variates V_1 and V_2 using the Wilson-Hilferty approximation, where $V_1 \overset{\cdot}{\sim} \chi^2_{n-1}$, $V_2 \overset{\cdot}{\sim} \chi^2_{n-2}$ and

$$V_1 = n\{1 - 2/[9n] + U_1[2/9n]^{1/2}\}^3$$

$$V_2 = (n-1)\{1 - 2/[9(n-1)] + U_2[2/9(n-1)]^{1/2}\}^3$$

(iii) The variance covariance matrix S is given by

$S = ((s_{ij}))$ where

$s_{11} = V_1/n$, $s_{22} = [V_2 + U_3^2]/n$

$s_{12} = U_3\sqrt{V_1}/n = s_{21}$.

To determine whether or not $\hat{R}_{\bar{x},s}$ contained R, it was first checked whether the origin (0,0) was inside $\hat{R}_{\bar{x},s}$. If not, then $\hat{R}_{\bar{x},s} \not\supset R$. If so, it was then determined by solving a fourth degree polynomial whether the circle R

and the ellipse $\hat{R}_{\bar{x},s}$ had any points of intersection. This depended on whether there were any real roots of the equation. If so, then $\hat{R}_{\bar{x},s} \not\supset R$. If all the roots were imaginary, then $\hat{R}_{\bar{x},s} \supset R$ provided the point $(\sqrt{\chi_2^2 (1 - \alpha)}, 0)$ was within the ellipse $\hat{R}_{\bar{x},s}$. Conversly, if the roots were imaginary and $(\sqrt{\chi_2^2 (1 - \alpha)}, 0)$ was outside the ellipse, then $\hat{R}_{\bar{x},s} \not\supset R$.

TABLE 2[4]: Evidence in Support of Approximation $\hat{K}^{(3)}_{N,n}$

Input Parameters				
PR	β	α	N	n
.880	.879	.95	120	5
.895	.879	.50	120	5
.880	.879	.05	120	5
.885	.879	.95	11	5
.890	.879	.50	11	5
.885	.879	.95	5	5
.875	.879	.50	5	5
.900	.835	.95	120	61
.915	.835	.05	120	61
.785	.716	.95	120	61
.795	.716	.95	61	61
.815	.716	.95	25	61
.860	.716	.95	11	61
.835	.716	.95	5	61
.995	.974	.95	120	61
.990	.977	.95	120	25
.990	.981	.95	120	11
.970	.986	.95	120	5
.470	.518	.95	61	5
.545	.518	.95	120	5
.808	.777	.95	120	5
.805	.777	.95	61	5

[4] PR = $\frac{\#\text{ of times } \hat{\bar{R}}_{\bar{x},s} \text{ contains R (200 runs)}}{200}$

REFERENCES

1. John, S. (1963). A tolerance region for multivariate normal distributions. Sankhyá, Series A, Vol. 25, pp. 363-368.

2. John, S. (1968). A central tolerance region for the multivariate normal distribution. J. Roy Statist. Soc., Ser. B.

3. Kleinbaum, David G. and S. John (1969). A table of percentage points of the smallest latent root of a 2 x 2 Wishart matrix. (Submitted for publication).

4. Odell, P. L. and Feiveson, A. H. (1966). A numerical procedure to generate a covariance matrix. Journal of the American Statistical Association, Vol. 61, pp. 199-203.