A TOLERANCE REGION FOR MULTIVARIATE NORMAL DISTRIBUTIONS

By S. JOHN

Indian Statistical Institute

SUMMARY. In this paper is developed a method of determining from a random sample from a p-variate normal population a region regarding which it can be asserted that, with probability β, a proportion not less than α of the individuals in the population are contained in it. Solutions to some related problems are given in the final section.

1. Introduction

In most statistical populations, whether they be populations of income or of blood pressure or of tensile strength of metal castings, a preponderant majority of individuals are concentrated over a relatively narrow range. This enables us, in many situations, to act as if individuals falling outside such intervals did not exist. Theoretically, information concerning such regions is implicit in the probability distribution, though actual determination of them is often a matter of some mathematical difficulty, esp:cially when the distribution is not unidimensional. The problem of tolerance regions is that of determining from only a random sample from the population, a region, regarding which we can assert that, with probability β, a proportion not less than α of the individuals in the population are contained in it.

The earliest formulation of the problem of tolerance regions is that of Wilks (1941). Wilks discovered a simple method of determining non-parametric tolerance regions for univariate populations. The corresponding multivariate problem was solved by Wald (1943). In Wald (1942) can be found an asymptotic solution of the problem of tolerance regions for parametric families of multivariate distributions. Tukey (1947, 1948), Tukey and Scheffé (1945), Fraser (1951, 1953), Fraser and Wormleighton (1951), Murphy (1948) and Kemperman (1956) report later work on the problem of non-parametric tolerance regions.

Though a general (asymptotic) solution of the problem of tolerance regions for a parametric family of distributions was given by Wald (1942), specialisation of his solution to particular families of distributions does not generally lead to the best or the simplest solution possible. On the other hand, the non-parametric solution can be inefficient, as demonstrated by Wilks (1941), when applied to such special families. For these reasons, Wald and Wolfowitz (1946) worked out a separate solution for univariate normal distributions. Our purpose in this paper is to work out such a solution for multivariate normal distributions.

2. Notation

We shall denote by $g_x(\mu, \Sigma)$ the density function of the multivariate normal distribution of dimension p, having μ for mean vector and Σ for dispersion matrix.

For any region R in the sample space, we shall set

$$\nu(R;\ \mu, \Sigma) = \int \underset{R}{\cdots} \int\ g_x(\mu, \Sigma)\, dx_1 dx_2 \ldots dx_p. \qquad \ldots \quad (2.1)$$

The function $f(z;\ \lambda, m)$ is defined as follows :

$$f(z; \lambda, m) = \frac{e^{-\lambda}}{(\sqrt{2})^m} \sum_{0}^{\infty} \frac{(\frac{1}{2}\lambda)^j}{j!\,\Gamma(\frac{1}{2}m+j)} z^{\frac{1}{2}m+j-1}\, e^{-\frac{1}{2}z} \qquad \ldots \quad (2.2)$$

It is the density function of the non-central chi-square variable of noncentrality λ and degree of freedom m. Also, we shall set

$$F(z;\ \lambda,\ m) = \int_0^z f(u;\ \lambda,\ m)\, du. \qquad \ldots \quad (2.3)$$

The equation

$$F(P(\theta;\ \lambda,\ m); \lambda, m) = \theta \qquad \ldots \quad (2.4)$$

defines the function $P(\theta; \lambda, m)$.

3. Procedure for determining the tolerance region

Let $\bar{x}$ be the arithmetic mean of N observations of a random vector $x = (x_1, \ldots, x_p)$ distributed according to the density function $g_x(\mu, \Sigma)$. Let V be a realisation of an independent Wishart variable of n degrees of freedom having $n\Sigma$ for its expectation.*

Denote by R_k the region of all x-vectors satisfying the inequality

$$(x-\bar{x})V^{-1}\ (x-\bar{x})' \leqslant k. \qquad \ldots \quad (3.1)$$

Let

$$v_1 = P(\alpha;\ \tfrac{1}{2}\, N^{-1}\, p, p), \qquad \ldots \quad (3.2)$$

and

$$v_2 = P(1-\beta;\ 0, n\, p). \qquad \ldots \quad (3.3)$$

Set

$$K = (v_1/v_2)p. \qquad \ldots \quad (3.4)$$

Regarding the region R_K, we can make the following assertion :

$$\text{prob}\ (\nu(R_K;\ \mu, \Sigma) \geqslant \alpha) \approx \beta. \qquad \ldots \quad (3.5)$$

The difference between the two members of (3.5) is small provided n and N are at least moderately large.

The constant v_2 can be determined from the table of the percentage points of the chi-square distribution given by Fisher and Yates (1953). If N is large $P(\alpha; N^{-1}p, p) \approx P(\alpha; 0,\ p)$. Hence when N is large, v_1 also can be determined from these same tables. If (p/N) is not small enough, we have to resort to methods developed by Patnaik (1949) and Abdel-Aty (1954). The short tables which they give would be of help in determining v_2.

* The matrix of corrected sum of products calculated from a random sample of size $n+1$ satisfies our requirements.

4. Proof of equation (3.5)

Let M be any non-singular matrix. Let $R_{k,M}$ be the region of all x-vectors satisfying the inequality

$$(x-\bar{x}M)(M'\ V\ M)^{-1}\ (x-\bar{x}M)' \leqslant k. \qquad \ldots \quad (4.1)$$

Then,
$$\nu(R_k;\ \mu,\ \Sigma) = \nu(R_{k,M};\ \mu\ M,\ M'\Sigma\ M). \qquad \ldots \quad (4.2)$$

We now choose A so that $A'\ \Sigma\ A = I$ and $A'\ V\ A$ is a diagonal matrix. Let the diagonal elements of $A'V\ \ A$ be $t_i (i = 1, 2, \ldots, p)$. We can assume, without loss of generality, that $t_1 \leqslant t_2 \leqslant \ldots \leqslant t_p$. The region $R_{k,A}$ is then the region of all x-vectors satisfying the inequality

$$\sum_1^p (x_i-a_i)^2/t_i \leqslant k, \qquad \ldots \quad (4.3)$$

where a_i is the i-th component of the vector $\bar{x}\ A$.

Denote by R'_k the region of all x-vectors satisfying the inequality

$$\sum_1^p (x_i-a_i)^2 \leqslant k. \qquad \ldots \quad (4.4)$$

Set
$$k' = (v_1/v_2)pt_1,\ \ k'' = (v_1/v_2)\ \sum_1^p\ t_i \qquad \ldots \quad (4.5)$$

and
$$k''' = (v_1/v_2)pt_p. \qquad \ldots \quad (4.6)$$

Equation (4.2), together with (4.3) and (3.4), leads to the following :

$$\text{prob}\ \{\nu(R'_{k'};\ \mu\ A, I) \geqslant \alpha\} \leqslant \text{prob}\ \{\nu(R_K;\ \mu, \Sigma) \geqslant \alpha\} \leqslant \text{prob}\ \{\nu(R'_{k'''};\ \mu A,\ I) \geqslant \alpha\}. \qquad \ldots \quad (4.7)$$

Since $t_1 \leqslant (\Sigma\ t_i)/p \leqslant t_p$, we have also

$$\text{prob}\ \{\nu(R'_{k'};\ \mu A,\ I) \geqslant \alpha\} \leqslant \text{prob}\ \{\nu(R'_{k''};\ \mu A,\ I) \geqslant \alpha\} \leqslant \text{prob}\ \{\nu(R'_{k'''};\ \mu A,\ I) \geqslant \alpha\}. \qquad \ldots \quad (4.8)$$

It is easy to demonstrate that

$$\text{prob}\ \{\nu(R'_{k'''}\ ;\ \mu A,\ I) \geqslant \alpha\} - \text{prob}\ \{\nu(R'_{k'}\ ;\ \mu A,\ I) \geqslant \alpha\}$$

tends to zero as $n\to\infty$. Therefore, *a fortiori*,

$$|\text{prob}\ \{\nu(R'_{k''};\ \mu A,\ I) \geqslant \alpha\} - \text{prob}\ \{\nu(R_K;\ \mu,\ \Sigma) \geqslant \alpha\}|$$

tends to zero as $n\to\infty$. Hence, equation (3.5) will be established if we show that

$$\text{prob}\ \{\nu(R'_{k''};\ \mu A,\ I) \geqslant \alpha\} \approx \beta. \qquad \ldots \quad (4.9)$$

Now,
$$\nu(R'_k;\ \mu A,\ I) = F(k;\ w,\ p), \qquad \ldots \quad (4.10)$$

where
$$w = \tfrac{1}{2}(\bar{x}-\mu)\Sigma^{-1}(\bar{x}-\mu)'. \qquad \ldots \quad (4.11)$$

Therefore,

$$\begin{aligned}
&\text{prob}\{\nu(R'_{k^*};\ \mu A,\ I) \geqslant \alpha\} \\
&= \text{prob}\{F(k^*;\, w, p) \geqslant \alpha\}, \\
&= E_w\, \text{prob}\{v_1 v_2^{-1} v \geqslant P(\alpha;\, w, p)\,|\,w\},\ \text{where } v = \sum_1^p t_i, \\
&= E_w\, \text{prob}\{v \geqslant v_2 v_1^{-1}\, P(\alpha;\, w, p)\,|\,w\}, \\
&= 1 - E_w F(v_2 v_1^{-1}\, P(\alpha;\ w,\ p);\ 0,\ np), \\
&= 1 - E_w F(v_2 v_1^{-1}\, P(\alpha;\ \tfrac{1}{2}N^{-1}p,\, p);\ 0,\ np) \\
&\quad - E_w(w - \tfrac{1}{2}N^{-1}p)[(\partial/\partial\lambda)\, F(v_2 v_1^{-1} P(\alpha;\, \lambda,\, p);\, 0,\, np)]_{\lambda = \frac{1}{2}N^{-1}p} \\
&\quad - \tfrac{1}{2}E_w(w - \tfrac{1}{2}N^{-1}p)^2[\partial^2/\partial\lambda^2)F(v_2 v_1^{-1}\, P(\alpha;\, \lambda,\, p);\, 0,\, np)]_{\lambda = \gamma(w)}, \quad \ldots \quad (4.12)
\end{aligned}$$

by Taylor's theorem. Here $\gamma(w)$ is a function of w bounded by $\frac{1}{2}N^{-1}p$ and w.

Because of (3.2) and (3.3), the second term of the last member of (4.12) is $1-\beta$. The random variable $2Nw$ has the chi-square distribution with p degrees of freedom. From this it follows that the third term is zero. Finally, it is possible to prove that $(\partial^2/\partial\lambda^2)F(v_2 v_1^{-1}\ P(\alpha;\ \lambda,\ p);\ 0,\ np)$ is bounded. Further, $E(w-\frac{1}{2}N^{-1}p)^2 = \frac{1}{2}p/N^2$. Therefore, the absolute value of the fourth term is less than B/N^2, where B is some finite positive number. This proves that, if terms of order two in $(1/N)$ can be neglected, then

$$\text{prob}\{\nu(R'_{k^*};\, \mu A,\, I) \geqslant \alpha\} = \beta. \quad \ldots \quad (4.13)$$

5. Alternative procedures

Let $\xi(t_1, t_2, \ldots, t_p)$ be any 'average' of $t_1, t_2, \ldots, t_p$. Let v_3 be a number such that

$$\text{prob}\{\xi(t_1, t_2, \ldots, t_p) < v_3\} = \beta. \quad \ldots \quad (5.1)$$

Set

$$k^{**} = v_1/v_3. \quad \ldots \quad (5.2)$$

We can then prove, by arguments exactly similar to those employed earlier, that

$$\text{prob}\{\nu(R_{k^{**}};\ \mu,\ \Sigma) \geqslant \alpha\} \approx \beta. \quad \ldots \quad (5.3)$$

The procedure discussed in Section 3 corresponds to the choice of the arithmetic mean for ξ. This choice has the advantage that the exact value of v_3 can be determined quite easily using tables of the percentage points of the chi-square distribution. Some alternative choices for ξ are considered below.

(I) $$\xi(t_1, t_2, \ldots, t_p) = (t_1\, t_2, \ldots, t_p)^{1/p}.$$

Hoel (1937) shows that the density function of $\xi(t_1, t_2, \ldots, t_p)$ is approximately

$$c^{\frac{1}{2}p(n-p+1)}\{\Gamma(\tfrac{1}{2}p[(n-p+1]))\}^{-1}\ \xi^{\frac{1}{2}p(n-p+1)-1}e^{-c\xi}, \quad \ldots \quad (5.4)$$

where

$$c = \tfrac{1}{2}p[1 - \tfrac{1}{2}(p-1)(p-2)/n]. \quad \ldots \quad (5.5)$$

If $p = 1$ or 2, (5.4) is the exact density function of $\xi(t_1, t_2, \ldots, t_p)$. We can thus determine v_3 from tables of the chi-square distribution.

(II) $$\xi(t_1, t_2, \ldots, t_p) = p\Big/\Big(\sum_1^p t_i^{-1}\Big).$$

In this case, $p\xi$ is distributed approximately as a chi-square with $\{np-p(p+1)+2\}$ degrees of freedom.

6. Bounds for prob $\{\nu(R_k;\ \mu,\ \Sigma) \geqslant \alpha\}$

If we neglect terms of order two in $(1/N)$,

$$\text{prob}\ (t_1 > v_2/p) \leqslant \text{prob}\ \{\nu(R_k;\ \mu,\ \Sigma) > \alpha\} \leqslant \text{prob}\ (t_p > v_2/p). \qquad \ldots \quad (6.1)$$

These inequalities are just another version of inequalities (4.7), obtained by the application of equation (4.10).

We give below simple expressions for the distribution functions of t_1 and t_p, in the case $p = 2$. Pillai (1954) gives recurrence relations connecting distribution functions of different orders.

If $p = 2$, starting from the joint distribution of t_1 and t_2, which is given, for instance, by Fisher (1939), we can show that

$$\text{prob}\ (t_1 > t) = [1-F(2t;\ 0,\ 2n)]-[\Gamma(\tfrac{1}{2})/\Gamma(\tfrac{1}{2}n)](t/2)^{\frac{1}{2}(n-1)}\ e^{-\frac{1}{2}t}\cdot[1-F(t;\ 0,\ n+1)] \qquad \ldots \quad (6.2)$$

and that,

$$\text{prob}\ (t_2 > t) = [1-F(2t;\ 0,\ 2n)]+[\Gamma(\tfrac{1}{2})/\Gamma(\tfrac{1}{2}n)](t/2)^{\frac{1}{2}(n-1)}\ e^{-\frac{1}{2}t}\ F(t,\ 0,\ n+1) \qquad \ldots \quad (6.3)$$

From (3.3), (6.2) and (6.3) we see that both extreme members of inequalities (6.1) are very nearly equal to β even for moderately large values of n. Therefore, the middle member is more so.

7. Related problems and concluding remarks

In some situations we face a slightly different problem. Here $\Sigma = \sigma^2\Lambda$ where Λ is a known positive definite matrix and σ^2 an unknown positive number. An unbiased estimate s^2 of σ^2, independent of the estimate $\bar{x}$ of μ, is available. The quantity ns^2 is a realisation of a chi-square variable with n degrees of freedom. A situation of this kind arises, for example, if we want to determine a tolerance region for the distribution of estimates of regression parameters in a linear model. The procedure of Section 3 applies to this case also if we set

$$ns^2\Lambda = V, \qquad \ldots \quad (7.1)$$

and

$$v_2 = P(\alpha;\ 0,\ n). \qquad \ldots \quad (7.2)$$

A problem closely related to that which we have been discussing in Sections 1 to 6 is that of determining from a random sample from the population a (random) region R regarding which we can make the following assertion :

$$E\nu(R;\ \mu,\ \Sigma) = \alpha. \qquad \ldots \quad (7.3)$$

The region R_k of Section 3 will satisfy this requirement if we choose k so that $kN(n-p+1)/[p(N+1)]$ is the upper $100(1-\alpha)$ percent point of the F-distribution with p and $n-p+1$ degrees of freedom. Fraser and Guttman (1956) prove that among regions satisfying condition (7.3), R_k is, in many respects, best.

In the practical application of the procedure of Section 3, it would be convenient to have at hand a table of values of K for various values of N, n and p. Such a table we hope to make available at a later date.*

* Tables required in the univariate case are given by Weissberg and Beatty (1960).

References

Abdel-Aty, S. H. (1954): Approximate formula for the percentage points and the probability integral of the non-central χ^2 distribution. *Biometrika*, **41**, 538-540.

Fisher, R. A. (1939): The sampling distribution of some statistics obtained from non-linear equations. *Ann. Eugen.*, **9**, 238-249.

Fisher, R. A. and Yates, F. (1953): *Statistical Tables for Biological, Agricultural and Medical Research*, Fourth Edition, Oliver and Boyd, Edinburgh.

Fraser, D. A. S. (1951): Sequentially determined statistically equivalent blocks. *Ann. Math. Stat.*, **22**, 294-298.

——— (1953): Nonparametric tolerance regions. *Ann. Math. Stat.*, **25**, 44-55.

Fraser, D. A. S. and Guttman, Irwin (1956): Tolerance regions. *Ann. Math. Stat.*, **27**, 162-179.

Fraser, D. A. S. and Wormleighton, R. (1951): Nonparametric estimation, IV. *Ann. Math. Stat.*, **22**, 294-298.

Hoel, Paul, G. (1937): A significance test for component analysis. *Ann. Math. Stat.*, **8**, 149-158.

Kemperman, J. H. B. (1956): Generalized tolerance limits. *Ann. Math. Stat.*, **27**, 180-186.

Murphy, R. B. (1948): Nonparametric tolerance limits. *Ann. Math. Stat.*, **19**, 581-589.

Patnaik, P. B. (1949): The noncentral χ^2 and F distributions and their applications. *Biometrika*, **36**, 202-232.

Pillai, K. C. S. (1954): *On Some Distribution Problems in Multivariate Analysis*, Institute of Statistics Mimeograph Series No. 88, 34-39.

Scheffé, H. and Tukey, J. W. (1945): Nonparametric estimation, I. Validation of order statistics., *Ann. Math. Stat.*, **16**, 187-192.

Tukey, J. W. (1947): Nonparametric estimation, II. Statistically equivalent blocks and tolerance regions — the continuous case. *Ann. Math. Stat.*., **18**, 529-539.

——— (1948). Nonparametric estimation, III. Statistically equivalent blocks and multivariate tolerance regions — the discontinuous case. *Ann. Math. Stat.*, **19**, 30-39.

Wald, A. (1942): Setting tolerance limits when the sample size is large. *Ann. Math. Stat.*, **13**, 389-399.

——— (1943): An extension of Wilk's method of setting tolerance limits. *Ann. Math. Stat.*, **14**, 45-55.

Wald, A. and Wolfowitz, J. (1946): Tolerance limits for a normal distribution. *Ann. Math. Stat.*, **17**, 208-215.

Weissberg, Alfred and Beatty, Glen, H. (1960): Tables of tolerance-limit factors for normal distributions. *Technometrics*, **3**, 483-501.

Wilks, S. S. (1941): Determination of sample sizes for setting tolerance limits. *Ann. Math. Stat.*, **12**, 91-96.

——— (1942): Statistical prediction with special reference to the problem of tolerance limits. *Ann. Math. Stat.*, **13**, 400-409.

Paper received : March, 1961.