IN-65

153123

P.9

NASA Technical Paper 3304

March 1993

Improving the Chi-Squared Approximation for Bivariate Normal Tolerance Regions

Alan H. Feiveson

(NASA-TP-3304) IMPROVING THE CHI-SQUARED APPROXIMATION FOR BIVARIATE NORMAL TOLERANCE REGIONS (NASA) 9 p

N93-22670

Unclas

H1/65 0153123

NASA

NASA
Technical
Paper
3304

1993

Improving the Chi-Squared Approximation for Bivariate Normal Tolerance Regions

Alan H. Feiveson
Lyndon B. Johnson Space Center
Houston, Texas

NASA

National Aeronautics and
Space Administration

Office of Management

Scientific and Technical
Information Program

1.0 Introduction

Let $X_1,..,X_N$ be a sample of N observations taken from a bivariate normal distribution with unknown mean vector μ and covariance matrix Σ, and let $\overline{X}$ and S be the respective 2 x 1 sample mean vector and 2 x 2 sample covariance matrix calculated from the X_i. Given a desired containment probability β and a level of confidence γ, the problem addressed is to find a region about $\overline{X}$ that contains at least $100\beta\%$ of the X-distribution with probability γ.

When S^{-1} exists, an ellipse R about $\overline{X}$ for any positive number c may be defined by

$$R = R(\overline{X}, S, c) = \{x: (x-\overline{X})' S^{-1} (x-\overline{X}) \le c\} \tag{1-1}$$

For a future observation X distributed according to $N_2(\mu,\Sigma)$, the probability that $X \in R$ is given by

$$I(c) = \frac{1}{2\pi|\Sigma|^{1/2}} \int_R e^{-\frac{1}{2}(x-\overline{X})'\Sigma^{-1}(x-\overline{X})} dx \tag{1-2}$$

For each new sample of N observations, R, which depends on $\overline{X}$ and S, is a random region in 2-dimensional Euclidean space; thus for a fixed general c, $I(c)$ is a random variable taking values in $(0,1)$. In particular, we seek c^* (not depending on the unknown μ or Σ), such that

$$P\{I(c*) \ge \beta\} = \gamma \tag{1-3}$$

The corresponding ellipse $R(\overline{X},S,c^*)$, known as a *tolerance region*, solves the problem.

Based on the work of John (1963) for a general p-variate normal distribution, the following approximation for c^* is given by Chew (1966):

$$\tilde{c} = \frac{(N-1)\,p\mu'_\beta(p, \frac{p}{N})}{\mu_{1-\gamma}((N-1)p)} \tag{1-4}$$

where $u'_\beta(p,\lambda)$ is the β-percentage point of the noncentral chi-squared distribution[1] with p degrees of freedom and noncentrality parameter λ, and $u_{1-\gamma}(m)$ is the $(1-\gamma)$-percentage point of the central chi-squared distribution with m degrees of freedom. Equation (1-4) is easier to evaluate than more precise but complicated expressions, such as given by Siotani (1964). Chew states, "the approximation is good if $1/N^2$ is negligible"; however, it appears (see section 3) that for the bivariate normal case ($p = 2$), $\tilde{c}$ underestimates c^* by a factor $1 - A/N$ where A depends on β and γ.

For general values of p, (1-4) has stood the test of time; for example it was cited and used by Rode and Chincilli (1988) in their paper on transforming clinical laboratory measurements. When $p = 2$, however, it is feasible to significantly improve the approximation by direct calculation of $I(c)$ within a Monte-Carlo simulation of values of $\overline{X}$ and S (see section 2). Estimation of A by comparing the resulting more accurate estimates of c^* with $\tilde{c}$ makes it feasible to use a corrected form of (1-4) to obtain accurate easily computed tolerance regions.

---

[1] Chew defines the noncentrality parameter "in accordance with that in Wilks" (1962); i.e. a noncentral chi-squared random variable with m degrees of freedom and noncentrality parameter λ is distributed as $Z^2 + Y$, where $Z \sim N(\lambda^{1/2},1)$ and Y has a central chi-squared distribution with $m-1$ degrees of freedom.

2.0 Monte-Carlo Estimation of c*

For $X \sim N(\mu, \Sigma)$ and any level of confidence, it can be shown that as N becomes large, c^* approaches $c_0 = -2log(1-\beta)$. This is because c_0 satisfies

$$P\{(X-\mu)'\Sigma^{-1}(X-\mu) \le c_0\} = 1-e^{-\frac{c_0}{2}} = \beta \tag{2-1}$$

(e.g., see Cramér (1963)) and $\overline{X}$ and S converge in probability to μ and Σ as N increases. For finite N, the solution to (1-3) is $c^* = Kc_0$ for some $K > 1$.

Let $\Sigma^{1/2}$ be a "square-root" of Σ in the sense that $\Sigma^{1/2}(\Sigma^{1/2})' = \Sigma$. By making the transformation $y = \Sigma^{-1/2}(x-\mu)$ in (1-2), it can be shown that the solution to (1-3) is the same as when $\mu = 0$, $\Sigma = I$ and $\overline{X}$ and S are obtained from a sample of N observations from the N($0,I$)-distribution. As a result, it will be henceforth assumed that $\mu = 0$ and $\Sigma = I$.

For each combination of $N = 10, 40(5), 50$, $\beta = .90, .95, .99, .999$ and selected values of c in the range $c = Kc_0$ $(1 < K \le 7.5)$, 1000 realizations of $\overline{X}$ and S were randomly generated taking $\mu = 0$ and $\Sigma = I$. (This can be done without generation of the individual observations; see Odell and Feiveson (1966)). For each $\overline{X}$ and S, $I(c)$ was then calculated by numerical integration (see appendix).

With β fixed, $Q(c) = P\{I(c) \ge \beta\}$ is a monotonic increasing function of c, with c^* being the root $Q(c^*) = \gamma$. From the simulation, for each trial value of c, say c_i, the observed proportion of times, q_i, that $I(c_i)$ exceeds β, is an estimate of $Q(c_i)$. For each N and β, an interpolating quadratic function was fitted to the points (y_i, c_i), where $y_i = -log(1-q_i)$; $(.80 \le q_i \le .999)$, then set equal to $y_\gamma = -log(1-\gamma)$ to solve for $\hat{c}_\gamma$, the estimate of c^* for $\gamma = .90, .95$ and $.99$. As an example, a plot of y_i vs c_i along with the interpolating quadratic function is shown for $N = 10$ and $\beta = .99$ in figure 1. The three horizontal lines represent the values of y_γ which define $\hat{c}_\gamma$.

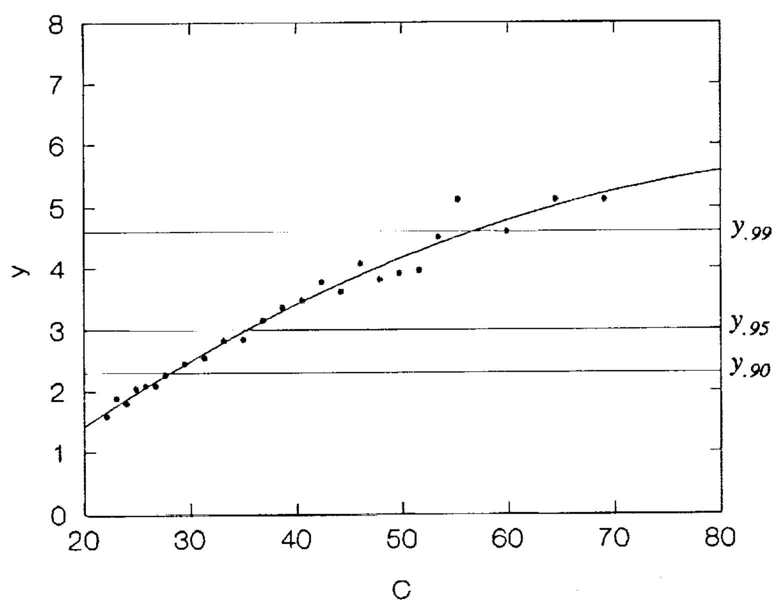

Figure 1. $y_i = -log(1-q_i)$ vs c_i for $N = 10$; $\beta = .99$

Originally, $\overline{X}$ and S were kept fixed as c was varied for given N and β; however, it was noticed that unnatural patterns in plots of y vs c would often result. Consequently, it was decided to avoid all dependence between results by regenerating $\overline{X}$ and S for each value of c, despite the extra computational effort.

3.0 Accuracy of Point Estimates

The major sources of error in $\hat{c}_\gamma$ are (1) the numerical integration used to compute $I(c)$ and (2) the process of fitting an interpolating polynomial to the q_i and solving for $\hat{c}$. A check against the first error was made by using a sufficiently small step size so that values of $I(c)$ resulted in 1 (to within 5 decimal places) as c was made arbitrarily large. The second error contains a random component induced by the binomial distribution of the q_i and some bias due to inverting the estimated y *vs* c relationship as well as possible model error (the true relationship may not be quadratic). Because the obtained fits were so tight (e.g., see fig. 1), the bias in $\hat{c}_\gamma$ was considered negligible.

To estimate the variance of the random error, a replicate of the entire simulation was made. Under the assumption of constant coefficient of variation, differences between the results were used to estimate a CV of about 1.5% for individual values of $\hat{c}_\gamma$. For the final smoothing described below, results of the two runs were averaged, further reducing the error CV by a factor of $\sqrt{2}$.

4.0 Final Results

Values of $\hat{c}_{.90}$, $\hat{c}_{.95}$, and $\hat{c}_{.99}$ for each N and β are shown in table 1 along with corresponding values of $\tilde{c}$ obtained from (1-4). A comparison reveals that the latter tend to be smaller by a factor of $1 - A/N$ where A depends on β and γ, thus suggesting smoothing the $\hat{c}_\gamma$ using $\tilde{c}$ as a concomitant variable. Given $\tilde{c}$, one may then better estimate c^* by

$$\hat{c}' = \tilde{c}\,[N/(N\text{-}A)]\ . \tag{4-1}$$

Using the data in table 1, estimated values of A for various β and γ were obtained by regression through the origin of $1 - \tilde{c}_\gamma/\hat{c}_\gamma$ against $1/N$. These values are shown in table 2. Although there were only eight values of N for each of the nine regressions, the fits were almost exact. Standard errors of A-estimates ranged from .04 to .22, corresponding to pertubations in $\hat{c}'$ between 0.4 and 2.3 percent for $N = 10$, and between 0.08 and 0.45 percent for $N = 50$. By contrast, errors in uncorrected $\tilde{c}$ are about 50% for $N = 10$ and 5 - 10% for $N = 50$.

Table 1. Values of $\hat{c}_\gamma$ and $\tilde{c}$
γ-Confidence Tolerance Ellipsoid of Content β;
Bivariate Normal Distribution

		$\gamma = .90$		$\gamma = .95$		$\gamma = .99$	
N	β	$\hat{c}_\gamma$	$\tilde{c}_\gamma$	$\hat{c}_\gamma$	$\tilde{c}_\gamma$	$\hat{c}_\gamma$	$\tilde{c}_\gamma$
10	.900	12.53	8.38	15.45	9.69	24.93	12.97
10	.950	17.07	10.89	21.40	12.60	34.69	16.86
10	.990	28.05	16.72	35.46	19.34	58.85	25.89
10	.999	45.42	25.06	57.25	28.99	91.78	38.81

Table 1 (Cont.), Values of $\hat{c}_\gamma$ and $\tilde{c}$
γ-Confidence Tolerance Ellipsoid of Content β;
Bivariate Normal Distribution

		$\gamma = .90$		$\gamma = .95$		$\gamma = .99$	
N	β	$\hat{c}_\gamma$	$\tilde{c}_\gamma$	$\hat{c}_\gamma$	$\tilde{c}_\gamma$	$\hat{c}_\gamma$	$\tilde{c}_\gamma$
15	.900	9.23	7.26	10.68	8.12	14.33	10.13
15	.950	12.47	9.43	14.67	10.56	21.01	13.17
15	.990	20.23	14.50	23.65	16.22	32.80	20.24
15	.999	31.75	21.71	37.90	24.29	52.91	30.31
20	.900	7.91	6.72	8.93	7.38	11.36	8.87
20	.950	10.44	8.74	11.78	9.60	15.09	11.54
20	.990	16.86	13.43	19.20	14.75	25.43	17.74
20	.999	26.57	20.15	30.21	22.14	40.58	26.63
25	.900	7.25	6.39	8.00	6.94	9.95	8.16
25	.950	9.66	8.32	10.68	9.03	13.07	10.61
25	.990	15.31	12.78	17.09	13.88	21.57	16.30
25	.999	24.19	19.19	27.09	20.85	34.43	24.49
30	.900	6.83	6.17	7.45	6.65	8.97	7.68
30	.950	9.04	8.03	9.89	8.65	11.91	9.99
30	.990	14.23	12.34	15.78	13.30	19.65	15.36
30	.999	22.26	18.49	24.58	19.92	30.19	23.01
35	.900	6.53	6.01	7.01	6.44	8.27	7.35
35	.950	8.58	7.82	9.32	8.38	11.11	9.56
35	.990	13.52	12.02	14.65	12.87	17.50	14.69
35	.999	21.33	18.02	23.33	19.29	27.52	22.01
40	.900	6.30	5.89	6.78	6.28	7.91	7.09
40	.950	8.36	7.66	8.98	8.16	10.40	9.23
40	.990	13.14	11.78	14.20	12.55	16.78	14.18
40	.999	20.31	17.67	21.97	18.83	26.04	21.27
50	.900	6.02	5.71	6.42	6.04	7.33	6.73
50	.950	7.90	7.43	8.40	7.86	9.70	8.75
50	.990	12.52	11.43	13.36	12.09	15.40	13.46
50	.999	19.03	17.15	20.41	18.13	23.75	20.19

Table 2. Values of A for Correcting
Non-central Chi-Squared Approximation for
Bivariate Normal Tolerance Regions.

		γ		
		0.90	0.95	0.99
β	0.900	3.153	3.543	4.553
	0.950	3.521	3.994	5.103
	0.990	4.093	4.606	5.800
	0.999	4.725	5.254	6.334

Correction is $\hat{c}' = \tilde{c}[N/(N-A)]$.

As an example, for $N = 10$, a 90%-tolerance region ($\gamma = .90$) that contains at least 99% of the population ($\beta = .99$) is found by first computing the chi-squared approximation (1-4), giving $\tilde{c} = 16.72$, and then correcting it with equation (4-1). Table 2 gives $A = 4.093$; hence $\hat{c}' = 16.72[10/(10 - 4.093)] = 28.31$. The desired tolerance region is the ellipse $\{x:(x-\overline{X})'S^{-1}(x-\overline{X}) \leq 28.31\}$.

5.0 Concluding Remarks

This paper has illustated how Monte-Carlo simulation, along with simple regression modelling, can be used to improve a theoretical approximation for a useful special case. The approximation is easily obtained if one has access (through software or tables) to the percentage points of the central and non-central chi-squared distributions. Correction to more accurate values for bivariate normal tolerance regions is readily accomplished for conventional values of β and γ using the appropriate value of A in table 2.

If one does not have a ready means of obtaining non-central chi-squared percentage points $u'_\beta(p,\lambda)$, an approximation given in Abramowitz and Stegun (1966) provides even greater simplification of computation with little loss of accuracy when $p = 2$. Abramowitz and Stegun give $u'_\beta(p,\lambda) \approx (1+b)u_\beta(p^*)$ where $p^* = a/(1+b)$, $a = p + \lambda$ and $b = \lambda/(p + \lambda)$. Here, $p = 2$ and $\lambda = 2/N$, hence $1 + b = (N+2)/(N+1)$ and $p^* = 2(N+1)^2/[N(N+2)] \approx 2$ so that for larger values of N, one may simply use

$$\begin{aligned} u'_\beta(p,\lambda) &\approx \frac{(N+2)}{(N+1)} u_\beta(2) \\ &\approx -2\frac{(N+2)}{(N+1)} \log(1-\beta) \end{aligned} \tag{5-1}$$

References

Abramowitz, M. and Stegun, I. A. (1966). *Handbook of Mathematical Functions,* National Bureau of Standards Applied Mathematics Series 55. Washington. D. C.: U. S. Government Printing Office.

Anderson, T.W. (1964). *An Introduction to Multivariate Statistical Analysis.* New York: John Wiley and Sons.

Chew, V. (1966). "Confidence, Prediction and Tolerance Regions for the Multivariate Normal Distribution," *Journal of the American Statistical Association, 61*, 605-617.

Cramér, H. (1963). *Mathematical Methods of Statistics.* Princeton University Press.

John, S. (1963). "A Tolerance Region for Multivariate Normal Distributions," *Sankhyā, 25*, 363-8.

Odell, P.L. and Feiveson, A. H. (1966). "A Numerical Procedure to Generate Sample Covariance Matrices," *Journal of the American Statistical Association, 61*, 199-203.

Rode, R. A. and Chinchilli, V. M. (1988). "The Use of Box-Cox Transformations in the Development of Multivariate Tolerance Regions with Applications to Clinical Chemistry," *American Statistician, 42*, 23-30.

Siotani, M. (1964). "Tolerance Regions for a Multivariate Normal Population," *Annals of the Institute of Statistical Mathematics, 16*, 135-53.

Wilks, S. S. (1964). *Mathematical Statistics.* New York: John Wiley and Sons.

Appendix

Obtaining $I(c)$ by Numerical Integration

Equation (1-2) may be rewritten

$$I(c) = \int_{x_{2L}}^{x_{2H}} f(x_2) \int_{g_L(x_2)}^{g_H(x_2)} f(x_1|x_2)\, dx_1 dx_2 \qquad \text{(A-1)}$$

where $f(x_2)$ is the density of x_2; i.e. N(0,1) and $f(x_1|x_2)$ is the density of the conditional distribution of x_1 given x_2, which is also standard normal, since $\Sigma = I$. The limits x_{2H} and x_{2L} are given by $\overline{X}_2 \pm (cS_{22})^{1/2}$ and for fixed x_2, the limits of x_1 are given by

$$g_H(x_2),\, g_L(x_2) = \frac{S_{21}(x_2 - \bar{X}_2) \pm \sqrt{|S|\,[cS_{22} - (x_2 - \bar{X}_2)^2]}}{S_{22}} \qquad \text{(A-2)}$$

where $S = (S_{ij})$. The inner integral in (A-1) is easily computed as $\Phi[(g_H(x_2)] - \Phi[(g_L(x_2)]$ where Φ is the standard normal cumulative distribution function for which good approximations are available.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)

2. REPORT DATE
March 1993

3. REPORT TYPE AND DATES COVERED
Technical Paper

4. TITLE AND SUBTITLE
Improving the Chi-Squared Approximation for Bivariate Normal Tolerance Regions

5. FUNDING NUMBERS

6. AUTHOR(S)
Alan H. Feiveson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Lyndon B. Johnson Space Center
Houston, Texas 77058

8. PERFORMING ORGANIZATION REPORT NUMBER
S-698

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, D.C. 20546-001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

NASA TP-3304

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified/Unlimited

Subject Category 65

12b. DISTRIBUTION CODE

13. ABSTRACT

Let X be a two-dimensional random variable distributed according to $N_2(\mu,\Sigma)$ and let $\overline{X}$ and S be the respective sample mean and covariance matrix calculated from N observations of X. Given a containment probability β and a level of confidence γ, we seek a number c, depending only on N, β and γ such that the ellipsoid $R = \{x: (x - \overline{X})'S^{-1}(x - \overline{X}) \leq c\}$ is a tolerance region of content β and level γ; i.e., R has probability γ of containing at least 100β percent of the distribution of X. Various approximations for c exist in the literature, but one of the simplest to compute — a multiple of the ratio of certain chi-squared percentage points — is badly biased for small N. For the bivariate normal case, most of the bias can be removed by simple adjustment using a factor A which depends on β and γ. This paper provides values of A for various β and γ so that the simple approximation for c can be made viable for any reasonable sample size. The methodology provides an illustrative example of how a combination of Monte-Carlo simulation and simple regression modelling can be used to improve an existing approximation.

14. SUBJECT TERMS
Approximation; Monte Carlo Method; Regression Analysis; Bivariate Analysis; Mathematical Models; Statistical Distributions; Tolerance Region

15. NUMBER OF PAGES
12

16. PRICE CODE
A03

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
Unclassified