The Philippine Statistician, 2001
Vol. 50 Nos. 1-4, pp. 55-62.

Effects of Sample Size and Skewness of Multivariate Populations to Confidence Ellipsoids for Population Means[1]

Millard R. Mamhot[2]

ABSTRACT

In this paper, samples of different sizes were drawn from bivariate distributions with wide-ranging measures of skewness. The Hotelling's T^2 statistic, $n(\bar{X}-\mu)S^{-1}(\bar{X}-\mu)$, was computed for each of the samples. It was found out that as the sample size n increases, $P\left(T^2 > \frac{(n-1)\,pF_{n,(n-p)}(\alpha)}{n-p}\right)$ decreases for a fixed α and p = 2. Also for given n, these probabilities tend to be proportional to the skewness of the distributions. These findings led to the feasibility of locating the appropriate sample size n so that $P\left(T^2 > \frac{(n-1)pF_{n,(n-p)}(\alpha)}{n-p}\right) = \alpha$ for a given α and for different measures of skewness. Simulation results are given and supported by a saddlepoint approximation to density functions and an Edgeworth expansion of multivariate normal distributions.

1. INTRODUCTION

The Central Limit Theorem states that if samples of size n are drawn from a population with mean μ and variance σ^2, then the sampling distribution of the sample mean $\bar{x}$ is approximately normally distributed with mean $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}}^2 = \sigma^2/n$, and that this approximation gets better as the sample size increases. A rule of thumb tells us that this approximation is good for as long as $n \geq 30$. However, some studies reveal that knowledge on the nature of the underlying distribution for the sample helps in determining the value of n such that the probability coverage be exactly $(1-\alpha)\times 100\%$. For instance, if the distribution of a random variable is U(0,1), then it is sufficient that n = 12, and for asymmetric distributions, such as the exponential distribution, n could be much greater than 30.

Boos and Oliver (2000) provided explicit Edgeworth expansions of normal density functions to illustrate the effects of both skewness and kurtosis of the underlying population on the accuracy of normal approximations. Sen et al. (1992) offered a discussion of the minimum sample size needed to ensures the validity of classical confidence intervals for means with platykurtic distributions. Chen (1995) mentioned how skewness may affect the accuracy of tests of hypothesis about means of normal populations using the classical t-tests. This paper investigates the same problem for multivariate populations.

[1] One of Two Winning Entries in the 2001 Student Paper Competition in Statistics organized by the Statistical Research and Training Center. This paper was presented during the Student Session of the Eighth National Convention on Statistics (organized by the National Statistical Coordination Board), held on October 1-2, 2001 at the Westin Philippine Plaza.

[2] Ph.D. Student, Mindanao Polytechnic State College, C.M. Recto Avenue, Lapasan, Cagayan de Oro City *9000*.
Email: *m.millard@eudoramail.com*

In a multivariate case, if $X_1, X_2, \ldots, X_n$ are random sample from a multivariate normal distribution with mean μ and covariance Σ, then the statistic, $(X - \mu)'\Sigma^{-1}(X - \mu)$ has a chi-square distribution with the corresponding probability coverage:

$$P\left(n(\bar{X}-\mu)'\Sigma^{-1}(\bar{X}-\mu) \le \chi_p^2(\alpha)\right) = 1-\alpha \quad (1)$$

if Σ is known. If Σ is unknown, then

$$P\left(n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu) \le \frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\right) = 1-\alpha \quad (2)$$

where S is the estimated covariance of p×n matrix X, $F_{p,n-p}(\alpha)$ is the upper (100α)th percentile of the $F_{p,n-p}$ distribution.

In the investigation of Boos and Oliver (2000) for univariate populations, it was found out that the skewness of an underlying distribution varies directly with sample size n for a fixed α so that,

$$P\left(\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \le \mu \le \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) = 1-\alpha . \quad (3)$$

This paper wishes to extend this finding to a bivariate setting. Specifically, it intends to answer the following problem: given a bivariate distribution, how large should n be so that

$$P\left(n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu) \le \frac{2(n-1)}{n-2}F_{n,n-2}(\alpha)\right) = 1-\alpha \quad (4)$$

for a fixed α. Toward this end, we first look into the notion of saddlepoint approximation of a distribution. We then provide a discussion of the results of a simulation experiment and show the minimum sample sizes needed for (4) to hold for certain simulated models. We end the paper with some possible directions for future research.

2. DENSITY APPROXIMATION

Let $K_X(\beta)$ be the cumulant generating function for the random variable X, β, an n×1 vector. To estimate the density $f_X(x)$ using saddlepoint approximation, $f_X(x)$ is embedded into an exponential family and a density in the exponential family is chosen to be approximated. An approximation of the chosen density results in an approximation of $f_X(x)$ since members of the exponential family differ only by a factor of $\exp(x\beta - K_X(\beta))$.

The approximation of $f_X(x)$ is finally accomplished upon expanding the chosen member using the Edgeworth series. The Edgeworth series approximation of a density of a multivariate random variable X is given as

$$f_X(x) = \phi(x,\Sigma)\sum_{j=0}^{\infty}\sum_{s\in S(j)}\frac{1}{j!}\mu^{*_{s_1\Lambda s_j}}(-1)^j(h_{s_1\Lambda s_j}(x,\Sigma)) \quad (5)$$

where $\phi(x,\Sigma)$ is the density of a multivariate normal distribution; s is a 1×p vector of integers; μ^{*s} are pseudo-moments, and; $h_s(x,\Sigma)$ are generalized Hermite polynomials.

Suppose $f_X(x)$ is the underlying density of a random sample $X_1, X_2, \ldots X_n$. Then embedding $f_X(x)$ into an exponential family, the following expression for $f_X(x)$ is obtained:

$$f_X(x) = f_X(x,\hat{\beta})\exp[K_X(\hat{\beta}) - \hat{\beta}^T x] \quad (6)$$

where $\hat{\beta}$ is the solution of $K'(\beta) = x$ and K_X is the cumulant generating function of $f_X(x,\hat{\beta})$, the chosen member from the exponential family with mean x.

Approximating $f_X(x,\hat{\beta})$ with a normal density with mean 0, we have the following:

$$f_X(x) = \exp\left(n[K_X(\hat{\beta}) - \hat{\beta}^T X]\right)\left(\frac{n}{2\pi}\right)^{p/2}\det[K_X''(\hat{\beta})]^{-1/2}\left[1 + \frac{b(\hat{\beta})}{2n} + O(n^{-2})\right] \quad (7)$$

where $b(\hat{\beta})$ is the tilt measure for $f_X(x,\hat{\beta})$ and n, the sample size. Then by Edgeworth expansion,

$$f_X(x) = \sum_{j=0}^{\infty}\sum_{s\in S(j)}\frac{1}{j!}\mu^{*s_1\cdots s_j}(-1)^j\frac{d^j}{dx^{s_1}\cdots dx^{s_j}}$$

$$= \sum_{j=0}^{\infty}\sum_{s\in S(j)}\frac{1}{j!}\mu^{*s_1\cdots s_j}(-1)^j[h_{s_1\cdots s_j}(0;K''(\hat{\beta}))(\phi(0,K''(\hat{\beta})))$$

$$= (\phi(0,K''(\hat{\beta}))(\sum_{j=0}^{\infty}\sum_{s\in S(j)}\frac{1}{j!}\mu^{*s_1\cdots s_j}(-1)^j[h_{s_1\cdots s_j}(0,K''(\hat{\beta}))]$$

$$= (\phi(0,K''(\hat{\beta}))\{1 + \sum_{j=1}^{\infty}\sum_{s\in S(j)}\frac{1}{j!}\mu^{*s_1\cdots s_j}(-1)^j[h_{s_1\cdots s_j}(0,K''(\hat{\beta}))]\}$$

Now, using cumulants instead of pseudo-moments, we get

$$f_X(x) = (\phi(0,K''(\hat{\beta}))(\exp(\sum_{j=3}^{\infty}\sum_{s\in S(j)}\frac{1}{j!}\kappa^{s1\cdots sj}(-1)^j[h_{s1\cdots sj}(0,K''(\hat{\beta}))]))$$

$$= (\phi(0,K''(\hat{\beta}))[1 + (\sum_{j=3}^{\infty}\sum_{s\in S(j)}\frac{1}{j!}\kappa^{s1\cdots sj}(-1)^j[h_{s1\cdots sj}(0,K''(\hat{\beta}))])$$

$$+\frac{1}{2}(\sum_{j=3}^{\infty}\sum_{s\in S(j)}\frac{1}{j!}\kappa^{s1\cdots sj}(-1)^j[h_{s1\cdots sj}(0,K''(\hat{\beta}))])^2$$

$$+\frac{1}{3}(\sum_{j=3}^{\infty}\sum_{s\in S(j)}\frac{1}{j!}\kappa^{*s1\cdots sj}(-1)^j[h_{s1\cdots sj}(0,K''(\hat{\beta}))])^3 + \cdots$$

$$= (\phi(0,K''(\hat{\beta}))[1 + (\frac{1}{3!}\hat{\kappa}^{ijk}(-1)h_{ijk}(0;K''(\hat{\beta})) + \frac{1}{4!}\hat{\kappa}^{ijkl}h_{ijkl}(0;K''(\hat{\beta})) + \cdots)$$

$$+\frac{1}{2}(\frac{1}{3!}\hat{\kappa}^{ijk}(-1)h_{ijk}(0;K''(\hat{\beta})) + \frac{1}{4!}\hat{\kappa}^{ijkl}h_{ijkl}(0;K''(\hat{\beta})) + \cdots)^2$$

$$+\cdots]$$

Since $h_{ijk}(0,K''(\hat{\beta})) = 0$, $\frac{1}{3!}\hat{\kappa}^{ijk}h_{ijk}(0;K''(\hat{\beta})) = 0$. The terms of order $O(n^{-1})$ are:

$$\frac{1}{4!}\hat{\kappa}^{ijkl}h_{ijkl}(0;K''(\hat{\beta})) + \cdots) + \frac{1}{2\times 3!3!}\hat{\kappa}^{ijk}\hat{\kappa}^{lmo}h_{ijklmo}(0;K''(\hat{\beta}))$$

Equating this with $2b(\hat{\beta})$, we get

$$2b(\hat{\beta}) = \frac{1}{4!}\hat{\kappa}^{ijkl}h_{ijkl}(0;K''(\hat{\beta})) + \cdots) + \frac{1}{2\times 3!3!}\hat{\kappa}^{ijk}\hat{\kappa}^{lmo}h_{ijklmo}(0;K''(\hat{\beta}))$$

Since

$$h_{ijkl}(0;\Sigma) = \kappa_{ij}\kappa_{kl}[3] = \kappa_{ij}\kappa_{kl} + \kappa_{ik}\kappa_{jl} + \kappa_{il}\kappa_{jk},$$

and we know (from McCullagh ,1987) that

$$h_{ijklmn}(0;\Sigma) = \kappa_{ij}\kappa_{kl}\kappa_{mn}[15],$$

we then have

$$2b(\hat{\beta}) = \frac{1}{4!}\hat{\kappa}^{ijkl}(\hat{\kappa}_{ij}\hat{\kappa}_{kl}[3]) - \frac{10}{6!}\kappa^{ijk}\hat{\kappa}^{lmo}(\hat{\kappa}_{ij}\hat{\kappa}_{kl}\hat{\kappa}_{mo}[15]).$$

Thus,

$$b(\hat{\beta}) = \frac{1}{4}\hat{\kappa}^{ijkl}(\hat{\kappa}_{ij}\hat{\kappa}_{kl}) - \frac{25}{12}\hat{\kappa}^{ijk}\hat{\kappa}^{lmo}(\hat{\kappa}_{ij}\hat{\kappa}_{kl}\hat{\kappa}_{mo})$$

Since

$$\hat{\kappa}^{ijk}\hat{\kappa}^{lmo}(\hat{\kappa}_{ij}\hat{\kappa}_{kl}\hat{\kappa}_{mo}) = b_{1,p}$$

and

$$\hat{\kappa}^{ijkl}(\hat{\kappa}_{ij}\hat{\kappa}_{kl}) = b_{2,p}$$

(see Mardia, 1970), we then have

$$b(\hat{\beta}) = \frac{1}{4}b_{2,p} - \frac{25}{12}b_{1,p} \tag{8}$$

Substituting this to (7), we get

$$f_X(x) = \exp\left(n[K_X(\hat{\beta}) - \hat{\beta}^T X]\right)\left(\frac{n}{2\pi}\right)^{p/2} \det[K_X''(\hat{\beta})]^{-1/2}\left[1 + \frac{\frac{1}{4}b_{2,p} - \frac{25}{12}b_{1,p}}{2n} + O(n^{-2})\right]$$

$$= \exp\left(n[K_X(\hat{\beta}) - \hat{\beta}^T X]\right)\left(\frac{n}{2\pi}\right)^{p/2} \det[K_X''(\hat{\beta})]^{-1/2}\left[1 + \frac{b_{2,p}}{8n} - \frac{25b_{1,p}}{24n} + O(n^{-2})\right]$$

Hence,

$$f_x(x) = MVN_p(\mu,\Sigma)\left[1 + \frac{b_{2,p}}{8n} - \frac{25b_{1,p}}{24n} + O(n^{-2})\right] \tag{9}$$

From (9) the following result is evident:

Theorem 1. If f is the underlying distribution of a random sample $X_1, X_2, \ldots, X_n$ with $K_X(\beta)$ as its cumulant generating function. then $f_X(x) \to MVN_p(\mu,\Sigma)$ as $n \to \infty$.

From (9), since $\frac{b_{2,p}}{8} = \frac{0.125}{n}b_{2,p}$, the contribution of $b_{2,p}$ when $n > 1$ is negligible. Thus, as $b_{1,p}$ approaches to 0, $f_x(x) \approx MVN_p(\mu,\Sigma)$. Consequently, we have the following result:

Theorem 2. Let f be the underlying distribution of a random sample $X_1, X_2, \ldots X_n$ and let $K_X(\beta)$ be a cumulant generating function of f. Then $f_X(x) \approx MVN_p(\mu,\Sigma)$ as $b_{1,p} \to 0$.

3. SIMULATION RESULTS

If $X_1, X_2, \ldots, X_n$ is a sample of size n from a multivariate normal population with mean μ and sample covariance S, then

$$P\left(n(\bar{x}-\mu)'S^{-1}(\bar{x}-\mu) \le \frac{p(n-1)}{n-p}F_{n,n-p}(\alpha)\right) = 1-\alpha \qquad (10)$$

for any n and a given α. By Theorem 2, it is expected that if the underlying distribution has some measure of skewness, then there exists an n_o such that for some $n > n_o$

$$P\left(n(\bar{x}-\mu)'S^{-1}(\bar{x}-\mu) \le \frac{p(n-1)}{n-p}F_{n,n-p}(\alpha)\right) = 1-\alpha \qquad (11)$$

for a given α.

A simulation experiment was conducted with the Hotelling's $T^2 = n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu)$ statistic computed for different sample sizes, viz., 10, 20, and 100 from 10 different skewed distributions. The skewness measures of the distributions were computed from Mardia's formula:

$$b_{1,p} = \frac{1}{n^2}\sum\sum g_{ij}^3 \qquad (12)$$

where

$$g_{ij} = (x_i - \overline{X})'\hat{\Sigma}^{-1}(x_j - \overline{X}), \quad \hat{\Sigma} = \sum_{i=1}^{n}(x_i - \overline{X})(x_i - \overline{X})'/n,$$

with n = 500 and μ was estimated using a sample size of 10,000 elements. Table 1 lists the bivariate distributions together with their corresponding skewness measures and estimated population means:

Table 1. Bivariate Distributions with Different Measures of Skewness.

Bivariate Density	*Skewness*	μ	*Label*
1. $f(x,y) = (50/\pi xy)e^{-50((\ln x)^2+(\ln y)^2)}$	0.8937	(1.3859,1.3904)	Lognorm
2. $f(x,y) = 1.35x^{-0.1}y^{0.5}e^{-(x^{0.9}+y^{1.5})}$	1.8196	(11.9398, 2.9580)	Weib-1
3. $f(x,y) = [(1+0.1x)(1+0.1y)-0.1]e^{-x-y-0.1xy}$	2.0928	(3.8084, 3.8265)	Expo-1
4. $f(x,y) = x^{-0.1}e^{-x^{0.9}-1.11y}$	2.4394	(10.4837,6.5920)	Weib-2
5. $f(x,y) = 0.000009736x^{-.999}y^{-.99}e^{-0.05(x+y)}$	2.6705	(97.4583, 6.4384)	Gamma
6. $f(x,y) = [(1+2.5x)(1+2.5y)-2.5]e^{-x-y-2.5xy}$	2.9400	(2.4860, 2.4350)	Expo-2
7. $f(x,y) = [(1+5x)(1+5y)-5]e^{-x-y-5xy}$	3.6278	(2.1653, 2.1033)	Expo-3
8. $f(x,y) = [(1+15x)(1+15y)-15]e^{-x-y-15xy}$	4.7480	(1.9965, 1.9503)	Expo-4
9. $f(x,y) = [(1+20x)(1+20y)-20]e^{-x-y-20xy}$	5.9813	(1.7397, 1.7054)	Expo-5
10. $f(x,y) = [(1+30x)(1+30y)-30]e^{-x-y-30xy}$	7.6703	(1.8586, 1.8589)	Expo-6

It is expected that when the significance criterion α is set the 0.05 level, only about 50 out of 1,000 will be greater than $\frac{2(n-1)F_{n,n-2}(\alpha)}{(n-2)}$. Table 2 shows the rate at which the Hotelling's

T^2 exceeds the tabular F for samples of sizes 10, 20, and 100. The error rate was calculated here as

$$ErrorRate = \frac{\#T^2 : [T^2 > \frac{2(n-1)F_{n,n-2}(\alpha)}{n-2}]}{no.\ of\ iterates}$$

Table 2. Error Rates of Sample Means on Samples from Skewed Distributions

Bivariate Density Function	*Skewness* b_{1p}	*Error Rate (α=0.05)*		
		n = 10	*n = 20*	*n = 100*
1. Lognorm	0.8937	0.071	0.053	0.046
2. Weib-1	1.8196	0.068	0.062	0.039
3. Expo-1	2.0928	0.064	0.057	0.038
4. Weib-2	2.4394	0.062	0.065	0.039
5. Gamma	2.6705	0.060	0.060	0.045
6. Expo-2	2.9400	0.127	0.074	0.038
7. Expo-3	3.6278	0.089	0.086	0.070
8. Expo-4	4.7480	0.175	0.100	0.062
9. Expo-5	5.9813	0.193	0.119	0.060
10. Expo-6	7.6703	0.211	0.119	0.068

Figure 1 illustrates the relationship between skewness with error rates for varying sample size.

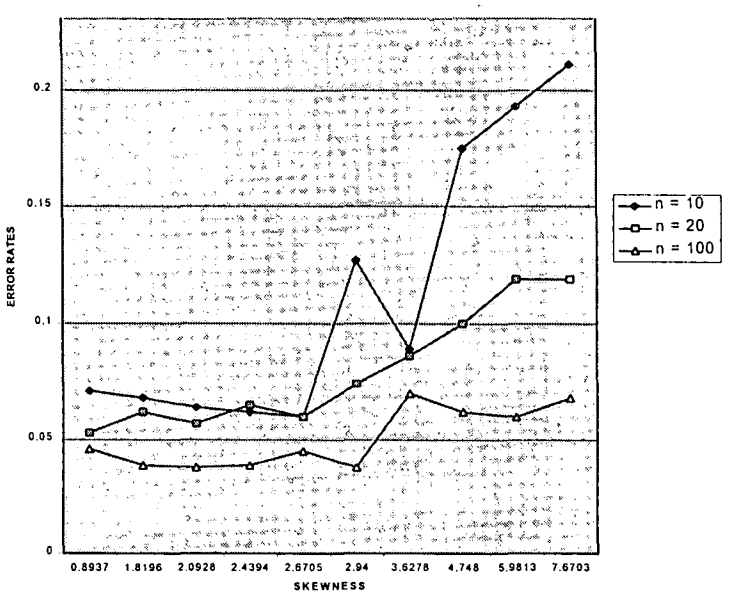

Figure 1. Relationship Between Error Rates and Skewness

Notice from both Table 1 and Figure 1 indicate that as the skewness coefficients increases, the error rate for a given sample size also increases. Furthermore, as the sample sizes are increased, the error rates decrease. We can therefore expect that for some sample size for a certain skewness coefficient, the 5% error rate can be reasonably attained.

Table 3 shows the minimum values of the sample size n needed in order for the following inequality to hold

$$P\left(n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu) \le \frac{2(n-1)F_{n,n-2}(\alpha)}{(n-2)} \right) \ge 1-\alpha$$

for α=0.05. Clearly, we see that when the skewness is rather small, the required sample size is nearly around the rule of thumb of 30, but with a rather large skewness coefficient, we would need much more than 30.

Table 3. Required Minimum Sample Size n from Skewed Distributions

Bivariate Density	*Skewness*	*Required n*
1. Lognorm	0.8937	27
2. Weib-1	1.8196	30
3. Expo-1	2.0928	34
4. Weib-2	2.4394	32
5. Gamma	2.6705	50
6. Expo-2	2.9400	70
7. Expo-3	3.6278	132
8. Expo-4	4.7480	150
9. Expo-5	5.9813	177
10. Expo-6	7.6703	200

4. FINAL REMARKS

In this study, we consider the effects of sample size and skewness of multivariate populations to confidence ellipsoids for population means from a simulation study and from the vantage point of saddlepoint approximation of the distribution. It was found out that error rates vary directly with skewness and inversely with sample size. One may want to extend the simulation runs to cases beyond the bivariate distributions examined here and moreover, investigate further how the sample size n is affected by both the skewness and kurtosis of a multivariate distribution.

ACKNOWLEDGMENTS

The author is grateful to Dr. Roberto N. Padua of Mindanao Polytechnic State College for helpful discussions on this topic, to the Statistical Research and Training Center for organizing the 2001 Student Paper Competition in Statistics. Thanks also to judges of the Student Paper Competition and to the anonymous referee for comments in improving this paper.

References

BOOS, D. D., and OLIVER, J. H. (2000), "How Large Does n Have to be for Z and t Intervals?" Journal of American Statistical Association, 54, 121-128.

BORENSTEIN, M., COHEN, J., and ROTHSTEIN, H. (1997), "Power Precision," Dataxiom, Inc., [On-line], http://www.dataxiom.com.

FELLER, W. (1971). An Introduction to Probability Theory and Its Application. 2nd Ed. John-Wiley & Sons.

HALL, P. (1992), "On the Removal of Skewness by Transformation," Journal Of Royal Statistical Society, Series B, 54, 221-228.

HOPKINS, W. G. (2000), "A New View of Statistics," Internet Society for Sport Science, http://www.sportsci.org/resource/stats/.

JOHNSON, N. J. (1978), "Modified t Tests and Confidence Intervals for Asymmetric Populations," Journal of American Association, 73, 536 – 547.

JOHNSON, R. A., and WICHERN, D. W. (1998). Applied Multivariate Statistical Analysis. Prentice-Hall International, Inc.

KOLASSA, J. E. (1997). Series Approximation Methods in Statistics. 2nd Ed. Springer-Verlag.

LOEVE, M. (1963). Probability Theory. D. Van Nostrand Company., Inc.

LUNFORD, B.R., and LUNFORD, T. R. (1998), "Research Forum – The Research Sample, Part II: Sample Size, "Journal of Prosthetics and Orthotics, 7, 137 – 141.

MARDIA, K. V. (1970), "Measures of Multivariate Skewness and Kurtosis with Applications," Biometrika, 57, 519 – 530.

MCCULLAGH, P. (1987). Tensor Methods in Statistics. Chapman and Hall.

MCNICKLE, D. C., PAWLIKOWSKIC, K., and EWING, G. (1996), "Experimental Evaluation of Confidence Interval Procedures in Sequential Steady-State Simulation," Proceedings of the 1996 Winter Simulation Conference, J. M. Charnes, D. J. Morrice, D.T. Brunner, and J. J. Swain, editors.

NEAL, D. K. (1993), "Determining Sample Sizes for Monte Carlo Integration," The College Mathematics Journal, 24, 254 – 259.

RENCHER, A. C. (1995). Methods of Multivariate Analysis. John-Wiley and Sons.

ZHOU, X. H., and GAO, S. (2000), "One-sided Confidence Intervals for Means and Positively Skewed Distributions," Journal of American Statistical Association, 54, 100 – 104.