ERRORS IN NORMAL APPROXIMATION TO CERTAIN TYPES OF DISTRIBUTION FUNCTIONS

by

J. T. Chu
University of North Carolina

Special report to the Office of Naval Research of work at Chapel Hill under Contract NR 042 031, Project N7-onr-284(02), for research in probability and statistics. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Institute of Statistics
Mimeograph Series No. 130
May 27, 1955

ERRORS IN NORMAL APPROXIMATIONS TO CERTAIN TYPES OF DISTRIBUTION FUNCTIONS

by
J. T. Chu
University of North Carolina

1. Summary. For certain types of sequences of distribution functions which are asymptotically normal, say with mean 0 and variance 1, a method is obtained for deriving upper and lower bounds in terms of the asymptotic distribution function. It is then shown that to many of these distribution functions, the errors are small in using the normal approximation.

2. Introduction. Let $F_n(x)$, $n = 1, 2, \ldots$, be a sequence of cdf's (cumulative distribution functions) such that for every fixed x, $F_n(x) \longrightarrow F(x)$ as $n \longrightarrow \infty$, where $F(x)$ is a cdf independent of n. From a practical point of view, it is desirable to know how large n has to be in order that $D_n(x) = |F_n(x) - F(x)|$ be small so that $F(x)$ may be used as an approximation to $F_n(x)$, although approximations are often used in practice without much knowledge about the magnitudes of the errors. The function $D_n(x)$ may vary, of course, considerably for different values of n and x. But the most interesting kind of $D_n(x)$'s is probably those which tend rapidly to 0, uniformly in x. In such cases, $F(x)$ provides for all n's greater than some minimum and for all x's a satisfactory approximation for $F_n(x)$. Generally, however, even though there is ample numerical evidence that as n increases $D_n(x)$ rapidly becomes uniformly small, it may not be easy to obtain a mathematical proof.

There are, on the other hand, types of sequences of cdf's for which we are able to confirm rigorously that they do tend rapidly to normality. If a cdf has

---

1. Work sponsored by the Office of Naval Research under Contract NR 042 031 at Chapel Hill.

one of the following forms:

$$F_n(x) = C_n \int_{-\infty}^{x} (1 \pm z^2/n)^{\mp m/2} \, dz \quad ,$$

where C_n and m depend only on n which takes integral values, and $\underset{n \longrightarrow \infty}{\text{Lim}}$ $m/n = 1$, then as $n \longrightarrow \infty$, $F_n(x) \longrightarrow \Phi(x)$, the normal cdf specified by (6), for every fixed x (Lemma 4). We find: by using simple transformations such as

$$u = \left[\pm n \log (1 \pm z^2/n) \right]^{\frac{1}{2}}$$

upper and lower bounds can be easily obtained for the integral in $F_n(x)$ in terms of $\Phi(x)$ (Lemma 3), and if C_n is not a very complicated function of n, then rather simple upper bounds can usually be derived for the error in using $\Phi(x)$ as an approximation to $F_n(x)$. If the bound is small, then so is the error. In § 4, applications are given to sequences of cdf's corresponding to the Student t-distribution, the τ-distribution of W. R. Thompson [4], and the distributions of the partial and total correlation coefficients when the variates involved are independently and normally distributed. For most of these cdf's, we are able to show that the errors are small in using the normal approximation, though the actual values of the errors seem even smaller.

Similar methods were used by the author [1] to derive upper and lower bounds for the cdf of the sample median $\tilde{x}$ in terms of its asymptotic distribution function (which is normal). There we also showed that if the parent distribution is normal, then even for samples of moderate sizes, the error is small in using the normal approximation to the cdf of $\tilde{x}$. The cdf of $\tilde{x}$ can be reduced to one of the forms given above by several transformations. But more re-

fined arguments are needed in order to get the bounds obtained in [1].

Some further applications and remarks are given at the end of §4 and §5.

3. Lemmas.

Lemma 1.

(1) $\quad 1 + x \leq e^x$, for all real x ,

(2) $\quad 1 + x \gtreqless e^{x - x^2/2}$, according as $x \gtreqless 0$.

If $x \geq 0$, then

(3) $\quad x / (1 - e^{-x^2})^{\frac{1}{2}} \geq 1$, and

(4) $\quad x\, e^{-x^2} / (1 - e^{-x^2})^{\frac{1}{2}} \leq 1$.

Proof. The function $e^x - x - 1$ has its minimum 0 at $x = 0$, hence we have (1).

(2) holds because $\log(1 + x) - x + x^2/2$ is monotonically increasing for all $x > -1$.

(3) is a consequence of (1). (4) follows from the facts that the LHS (left hand side) tends to 1 as $x \longrightarrow 0$ and is a monotonically decreasing function of x (Differentiate twice).

Lemma 2.

Let

$$b_n(c) = \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots (2n)} (n + c)^{\frac{1}{2}} , \quad n = 1, 2, \ldots,$$

where $c \geq -1$. Then

$$(5) \qquad \pi^{\frac{1}{2}} b_n(c) \leq 1 \quad , \text{ if } \quad c \leq \frac{1}{4} \quad ;$$

$$\geq 1 \quad , \text{ if } \quad c \geq \frac{2}{7} \quad .$$

Proof. $b_n(0)$ is known [5, p. 351] as the Wallis product and tends to $\pi^{-\frac{1}{2}}$ as $n \longrightarrow \infty$. Obviously $b_n(c)$ tends to the same limit for every fixed c. By examining the square of the ratio $b_{n+1}(c)/b_n(c)$, it can be shown that $b_n(c)$ is a monotonically increasing function of n if and only if $c \leq \frac{1}{4}$, and is a monotonically decreasing function of n if and only if $c \geq \frac{2}{7}$. Hence we have (5).

Lemma 3.

Let

$$(6) \qquad \Phi(x) = \int_{-\infty}^{x} (2\pi)^{-\frac{1}{2}} e^{-t^2/2} \, dt \quad ,$$

and $\Phi_0(x) = \Phi(x) - \frac{1}{2}$. Let (7) below be abbreviated as $A \leq B \leq C \leq D \leq E$, and let the inequalities $A \leq B$, $B \leq C$, $C \leq D$, and $D \leq E$ be respectively referred to as (7.1), (7.2), ..., and (7.4). Then (7.1) and (7.2) hold for all $m, n > 0$, and $0 \leq x \leq n^{1/2}$. (7.3) holds for all $m, n > 0$, and $0 \leq x < \infty$; and (7.4), all $m > 3$, $n > 0$, and $0 \leq x < \infty$.

$$(7) \qquad \sqrt{\frac{n}{m+2}} \, \Phi_0\left(x \sqrt{\frac{m+2}{n}}\right) \leq (2\pi)^{-\frac{1}{2}} \int_0^x (1 - z^2/n)^{m/2} \, dz$$

$$\leq \sqrt{\frac{n}{m}} \, \Phi_0\left(x \sqrt{\frac{m}{n}}\right) \leq (2\pi)^{-\frac{1}{2}} \int_0^x (1 + z^2/n)^{-m/2} \, dz$$

$$\leq \sqrt{\frac{n}{m-3}} \, \Phi_0\left(x \sqrt{\frac{m-3}{n}}\right) .$$

Proof. It is easy to see that (7.2) and (7.3) are immediate consequences of (1). Now use the transformation

$$v(z) = \left[n \log (1 + z^2/n) \right]^{\frac{1}{2}} ,$$

then

$$\int_0^x (1+z^2/n)^{-m/2} dz = \int_0^{v(x)} \exp \left[- \frac{(m-3)v^2}{2n} \right] h(v/n^{\frac{1}{2}}) dv ,$$

where $h(x)$ is the LHS of (4). By (2) and (4), $h(v/n^{\frac{1}{2}}) \leq 1$ and $v(x) \leq x$. Hence we have (7.4). Finally (7.1) can be obtained in a similar way by using (3) after applying to the integral the transformation

$$u(z) = \left[- n \log (1 - z^2/n) \right]^{\frac{1}{2}}$$

Lemma 4.

Suppose that for every integer $n \geq n_0 \geq 1$,

$$F_n(x) = C_n \int_{-\infty}^{x} (1 \pm z^2/n)^{\mp m/2} dz \tag{8}$$

is a cdf. where C_n and m depend only on n and $\lim_{n \to \infty} m/n = 1$ (If the integrand is $(1 - z^2/n)^{m/2}$, it should be replaced by 0 whenever $|z| \geq n^{\frac{1}{2}}$). Then, for every fixed x,

$$\lim_{n \to \infty} F_n(x) = \Phi(x) , \tag{9}$$

where $\Phi(x)$ is defined by (6).

Proof. By Lemma 3, we have $\lim_{n \to \infty} C_n = (2\pi)^{-\frac{1}{2}}$. Using the same lemma

once again, we obtain (9).

4. Normal approximations. In Lemma 4, we showed that if a cdf is of one of the types (8), then it tends to the Φ of (6) as $n \longrightarrow \infty$. In this section, we shall find upper and lower bounds for the C_n's corresponding to various well known cdfs of the types (8); then use Lemma 3 to obtain upper and lower bounds for these cdfs themselves, in terms of Φ ; and finally derive upper bounds for the proportional errors (18) in using Φ as an approximation to these cdfs.

A. t-distribution

The cdf of the t-distribution with n d.f. (degrees of freedom), $n = 1, 2, \ldots$, is given by

$$F_n(x) = \int_{-\infty}^{x} a_n(1 + z^2/n)^{-(n+1)/2}dz, \text{ where} \tag{10}$$

$$a_n = (n\pi)^{-1/2} \Gamma(\tfrac{n+1}{2}) / \Gamma(n/2) \quad . \tag{11}$$

It is well known that as $n \longrightarrow \infty$, $F_n(x) \longrightarrow \Phi(x)$ of (6) for every fixed x; and the "speed" of approaching the limit is rather fast. In fact, the normal approximation is often used in practice when $n \geq 30$. We shall derive for $F_n(x)$ upper and lower bounds in terms of $\Phi(x)$, then show that the proportional error in using $\Phi(x)$ as an approximation to $F_n(x)$ is less than $1/n$ for all $n \geq 8$.

Applying (7.4) to $F_n(y)-1/2$ and $1/2 - F_n(-x)$ and using the fact that $\Phi(-x) = 1 - \Phi(x)$, it can be shown easily that for arbitrary $x, y \geq 0$, and $n \geq 3$,

$$F_n(y) - F_n(-x) \leq (2\pi)^{\frac{1}{2}} a_n c_n^{-1} \left[\Phi(c_n y) - \Phi(-c_n x) \right] , \tag{12}$$

where $c_n = (1 - 2/n)^{\frac{1}{2}}$. Using (7.3), we obtain, in a similar way,

$$(13) \qquad F_n(y) - F_n(-x) \geq (2\pi)^{\frac{1}{2}} a_n d_n^{-1} \left[\Phi(d_n y) - \Phi(-d_n x) \right] ,$$

where $d_n = (1 + 1/n)^{1/2}$.

Using $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1/2) = \pi^{1/2}$, it can be seen that $a_{2m} = \left[m/(2m+2c) \right]^{1/2} b_m(c)$ and $a_{2m+1} = \left[(m+c)/(2m+1) \right]^{1/2} \pi^{-1} b_m^{-1}(c)$, where $m = 1, 2, \ldots$. Letting c be $1/4$ and $2/7$ in turn, we obtain, by (5), $(2\pi)^{1/2} a_n \leq \left[2n/(2n+1) \right]^{1/2}$ if $n = 2m$ and $\leq (1 - 3/(7n))^{1/2}$ if $n = 2m+1$, $m=1, 2, \ldots$. In general, for $n \geq 3$,

$$(14) \qquad (2\pi)^{1/2} a_n \leq \left(1 - 3/(7n) \right)^{1/2} .$$

Likewise, letting $c = 1/2$ and $1/4$ respectively, we obtain $(2\pi)^{1/2} a_n \geq \left[n/(n+1) \right]^{1/2}$ if $n = 2m$ and $\geq (1 - 1/2n)^{1/2}$ if $n = 2m + 1$, $m = 1, 2, \ldots$.

In general, for $n \geq 2$.

$$(15) \qquad (2\pi)^{1/2} a_n \geq \left[n/(n+1) \right]^{1/2} .$$

By (12) - (15), we have, for arbitrary $x, y \geq 0$, and $n \geq 3$.

$$(16) \qquad F_n(y) - F_n(-x) \leq \left[\frac{7n-3}{7n-14} \right]^{1/2} \left[\Phi(c_n y) - \Phi(-c_n x) \right] ,$$

$$(17) \qquad F_n(y) - F_n(-x) \geq d_n^{-2} \left[\Phi(d_n y) - \Phi(-d_n x) \right] .$$

The proportional error in using A as an approximation to B is defined to be

$$(18) \qquad E = \left| \frac{B}{A} - 1 \right| ,$$

where $|A|$ is the absolute value of A. Now omitting c_n and d_n in the

arguments in the Φ's of (16) and (17), we see that E is not more than the maximum of $[(7n-3)/(7n-14)]^{1/2} - 1$ and $1 - n/(n+1)$. For simplicity, we may state that $E < 1/n$ for all $n \geq 8$. It seems that the actual values of E are much smaller than 1/n. For example, if $n = 30$, and $y = x = 2.042$, then $F_n(y) - F_n(-x) = .95$ while $\Phi(y) - \Phi(-x) = .9588$ so $E \doteq .0092$. Nevertheless, the bound is of precise and general nature and small enough to justify the use of such approximations for large n.

B. Thompson's τ -distribution.

The cdf of the τ - distribution is given [2, p. 241] by

$$F_n(x) = \int_{-n^{1/2}}^{x} a_n' (1 - z^2/n)^{(n-3)/2} \, dz , \tag{19}$$

where $|x| \leq n^{1/2}$, $a_n' = (n\pi)^{-1/2} \Gamma(n/2)/\Gamma(\frac{n-1}{2})$, and $n = 2, 3, \ldots$. For applications of the τ - distribution the readers are referred to [2, p.390] and [4]. Obviously by (11), $a_n' = (1 - 1/n)^{1/2} a_{n-1}$. Using (7), (14), and (15), we obtain for $x, y \geq 0$, and $n \geq 4$.

$$F_n(y) - F_n(-x) \leq (\tfrac{7n-10}{7n-21})^{1/2} [\Phi(y\sqrt{\tfrac{n-3}{n}}) - \Phi(-x\sqrt{\tfrac{n-3}{n}})], \tag{20}$$

$$F_n(y) - F_n(-x) \geq (1 - 1/n)^{1/2} [\Phi(y\sqrt{\tfrac{n-1}{n}}) - \Phi(-x\sqrt{\tfrac{n-1}{n}})] \tag{21}$$

$$\geq (1 - 1/n)[\Phi(y) - \Phi(-x)] .$$

The second inequality of (21) is obtained by using the fact that $\Phi_0(ax) \geq a\,\Phi_0(x)$ if $0 \leq a \leq 1$. Thus in using $\Phi(y) - \Phi(-x)$ as an approximation to $F_n(y) - F_n(-x)$, the proportional error E, as defined by (18) is not more than the

maximum of $[(7n-10)/(7n-21)]^{1/2} - 1$ and $1 - (1 - 1/n)$. For $n \geq 13$, this maximum is $1/n$.

C. The correlation coefficients.

Let a sample of size $n + 1$ be drawn from each of k independently and normally distributed populations. The pdf of the partial correlation coefficients $r_{12\cdot34\ldots k}, (k \geq 2)$, corresponding to the k samples is then [2, p. 412]

$$n_k^{1/2}\, a'_{n_k} (1 - z^2)^{(n_k-3)/2} \quad , \quad |z| \leq 1 \quad , \tag{22}$$

where $n_k = n - k + 2$. If $k = 2$, then (22) reduces to the pdf of the total correlation coefficient r_{12}. The variance of $r_{12\cdot34\ldots k}$ is n_k^{-1}. Let $F_{n_k}(x)$ be the cdf of $n_k^{1/2}\, r_{12\cdot34\ldots k}$, then

$$F_{n_k}(x) = \int_{-n_k^{1/2}}^{x} a'_{n_k} (1 - z^2/n_k)^{(n_k-3)/2}\, dz \quad , \quad |x| \leq n_k^{1/2} \, . \tag{23}$$

This, obviously, is only a special case of (19). Therefore the proportional error in using $\Phi(y) - \Phi(-x)$ as an approximation to $F_{n_k}(y) - F_{n_k}(-x)$ is not more than $1/n_k$. Hotelling [3, p. 196] stated that (the normal approximation) is in ordinary cases the most convenient of all (methods for evaluating the integral (23)), but no suitable bound for the error has been available. The bound we obtain here seems acceptable, at least when n is large compared with k.

D. χ^2-distribution.

It is well known [2, p. 251] that if χ^2 has a χ^2-distribution with n d.f., then both $(\chi^2 - n)/(2n)^{1/2}$ and $(2\chi^2)^{1/2} - (2n)^{1/2}$ are asymptotically normally distributed with mean 0 and variance 1. According to Fisher, the distri-

bution of $(2\chi^2)^{1/2} - (2n-1)^{1/2}$ tends to normality even "faster". Unfortunately, we are not able to obtain upper and lower bounds for the cdfs of these distributions. We shall derive, as another application of (7), a lower bound, in terms of Φ_0 of (6), for $F_n(y) - F_n(0)$, where $y \geq 0$ and $F_n(y)$ is the cdf of $(2\chi^2)^{1/2} - (2n-2)^{1/2}$. If $y \geq 0$, then

$$(24) \qquad F_n(y) - F_n(0) = 2^{-n/2}\,\Gamma^{-1}(n/2)\int_m^{y_m} x^{m/2-1}\,e^{-x/2}\,dx$$

$$= \Gamma^{-1}(n/2)(m/2e)^{m/2}\int_0^{y}\left\{\left[1+z/(2m)^{1/2}\right]\exp\left[-z/(2m)^{1/2}-(1/2)z^2/2m\right]\right.$$

where $y_m = \left[(2m)^{1/2} + y\right]^{1/2}/2$, $m = n-1$, and $z = (2x)^{1/2} - (2m)^{1/2}$. For all practical purposes, we may use, for $n \geq 4$,

$$n! \sim (2\pi)^{1/2}\, n^{n+1/2} \exp\left[-n + 1/12n\right].$$

Using this approximation and (2), it can be shown that $\Gamma^{-1}(m/2)(m/2e)^{m/2} \geq (2\pi)^{-1/2}$ for $n \geq 10$. Now applying (2) to the second factor of the integrand in the second integral of (24), we see that the integrand is not less than $(1 - z^2/2m)^m$ (We assume that $y \leq (2m)^{1/2}$). Therefore a lower bound can be obtained for $F_n(y) - F_n(0)$ by using (7). However, a better lower bound can be obtained by applying (2) to the first factor of the integrand. We have, for all $y \geq 0$,

$$F_n(y) - F_n(0) \geq \Phi_0(y) \quad .$$

Similarly if $0 \leq x \leq (2m)^{1/2}$, then

$$F_n(0) - F_n(-x) \leq \Phi_0(x) \quad .$$

5. A remark. Sequences of distribution functions are often encountered which have asymptotic χ^2-distributions. For example, if x has a β-distribution with parameters m and n, then the c.d.f. of nx is given by [2. p. 243]

$$F_{m,n}(y) = \frac{\Gamma(\frac{m+n}{2})}{n^{m/2}\Gamma(m/2)\Gamma(n/2)} \int_0^y t^{m/2-1}(1 - z/n)^{n/2-1}\,dz,$$

where $0 \leq y \leq n$, and $m,n > 0$. As $n \longrightarrow \infty$, $F_{m,n}(y)$ tends to the c.d.f. $\chi^2_m(y)$ of the χ^2-distribution with m d.f. By methods similar to those used in deriving (7), we can obtain upper and lower bounds for $F_{m,n}(y)$ in terms of $\chi^2_m(y)$. Unfortunately, these bounds are not very close to each other unless m/n is very small; consequently they are of little practical interest. We omit the details

The author wishes to thank Professor Hotelling for his critical reading of the manuscript.

REFERENCES

[1] J. T. Chu, "On the distribution of the sample median," Annals of Mathematical Statistics, Vol. 26 (1955), pp. 112-116.

[2] H. Cramér, Mathematical Methods of Statistics, Princeton University Press 1946.

[3] Harold Hotelling, "New light on the correlation coefficient and its transforms," Journal of the Royal Statistical Society, Series B, Vol. 15 (1953), pp. 193-232.

[4] W. R. Thompson, "On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation," Annals of Mathematical Statistics, Vol. 6 (1935), pp. 214-219.

[5] J. V. Usponsky, Introduction to Mathematical Probability, McGraw-Hill, New York, 1937.