ON APPROXIMATIONS TO SAMPLING DISTRIBUTIONS OF THE MEAN FOR SAMPLES FROM NON-NORMAL POPULATIONS[1]

By A. Reitsma

University of the O.F.S., Bloemfontein, South Africa

1. Introduction and summary. In comparison with the vast number of distribution functions available for describing non-normal populations (see Haight [11]) a very few sampling distributions of the mean, when sampling from these, are known in exact form. The more important results have been derived by Baker [1], [2], Baten [3], [4], Bose [6], Church [7], Hall [12], Irwin [13], [14], Rao [16] and Shrivastava [18]. Except in the case of the well-known result which follows when sampling from a Pearson's Type III population, many of these are of exceptional form and hardly any of them have been tabulated.

Furthermore, since most of these results are only of practical use in the case of very small sample sizes, the need for approximations to the sampling distribution of the mean have been felt.

In the case of the normal approximation, which has been widely used in many cases in virtue of the Central Limit Theorem, Berry [5] has shown that it is not sufficient for moderate sample sizes. For the latter case, Esseen [9] has obtained approximations in terms of the normal distribution and its derivatives. These results have been extended by Gnedenko and Kolmogorov [10]. Other forms of approximations have been obtained by Daniels [8] and Welker [21].

Another approach to find approximations to fill the gap between the exact sampling distribution and its ultimate normal approximation is presented in this paper. A method developed by Steyn [19] in deriving a differential equation of the moment generating function of the sample mean and variance respectively for samples from a normal population, is used in deriving approximations to sampling distributions of the mean for samples from a number of Pearson's Type populations. Only first order approximations are considered and it will be shown that in the case of sampling from certain skew populations, the results lead to the Pearson's Type III-distribution before approaching normality, whereas, in the symmetrical cases the sampling distribution approaches normality directly.

2. The method of differential equations. Consider a random sample of size n from a population with a Pearson's Type I(Beta)-distribution. Let x_i, $i = 1, \cdots, n$, be n values of the stochastic variable X with distribution

$$[1/B(p, q)]x^{p-1}(1 - x)^{q-1}, \qquad p, q > 0; 0 \leqq x \leqq 1. \tag{2.1}$$

Received July 2, 1962; revised January 2, 1963.

[1] Research supported in part by a grant from the South African Council for Scientific and Industrial Research.

The m.g.f. of the mean $\bar{X} = (1/n)\sum_i X_i$, is given by

$$M(\alpha) = C \int_0^1 \cdots \int_0^1 \prod_i^n \{\phi(x_i)\}\, dx_1 \cdots dx_n, \tag{2.2}$$

where $\phi(x_i) = x_i^{p-1}(1 - x_i)^{q-1} \exp(\alpha x_i/n)$, $C = [1/B(p, q)]^n$ and α may be complex. Since this integral exists and is continuous in α and x ([22], p. 67), differentiation w.r.t. α gives

$$\frac{d}{d\alpha} M(\alpha) = \frac{C}{n} \int_0^1 \cdots \int_0^1 \prod_i^n \{\phi(x_i)\} \left(\sum_i x_i\right) dx_1 \cdots dx_n, \tag{2.3}$$

which may be written as

$$\frac{d}{d\alpha} M(\alpha) = \frac{C}{n} \sum_j \int_0^1 \cdots \int_0^1 \prod_{i \neq j}^n \{\phi(x_i)\} \cdot \left\{\int_0^1 \phi(x_j) x_j\, dx_j\right\} dx_1 \cdots (dx_j) \cdots dx_n, \tag{2.4}$$

where $dx_1 \cdots (dx_j) \cdots dx_n \equiv dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_n$.

Integration by parts of the simple integral

$$\int_0^1 \phi(x_j) x_j\, dx_j = \int_0^1 (1 - x_j)^{q-1} x_j^p e^{\alpha x_j/n}\, dx_j$$

gives

$$\int_0^1 \phi(x_j) x_j\, dx_j = (1/q) \int_0^1 \phi(x_j)(p - p x_j + \alpha x_j/n - \alpha x_j^2/n)\, dx_j. \tag{2.5}$$

By using (2.2) and (2.3) after substituting (2.5) in (2.4), the latter may be written as

$$\frac{d}{d\alpha} M(\alpha) = \frac{p}{q} M(\alpha) - \frac{p}{q} \frac{d}{d\alpha} M(\alpha) + \frac{1}{nq} \alpha \frac{d}{d\alpha} M(\alpha) - (C\alpha/n^2 q) \int_0^1 \cdots \int_0^1 \prod_i \{\phi(x_i)\} \left(\sum_j x_j^2\right) dx_1 \cdots dx_n. \tag{2.6}$$

Repeating the process (i.e., rewriting the latter integral in the form (2.4) and integrating the single integral, so formed, by parts) and retaining only terms of the order (n^{-1}), (2.6) becomes

$$\frac{d}{d\alpha} M(\alpha) = \frac{p}{q} M(\alpha) - \frac{p}{q} \frac{d}{d\alpha} M(\alpha) + \frac{1}{nq} \left\{1 - \frac{p+1}{q} + \left(\frac{p+1}{q}\right)^2 - \cdots\right\} \alpha \frac{d}{d\alpha} M(\alpha) + O\left(\alpha^2 n^{-2} \frac{d}{d\alpha} M(\alpha)\right), \tag{2.7}$$

where the series in brackets sums up to $1/(p+q+1)$, provided that $|(p+q)/q| < 1$, which leads to the conditions $p+1 < q$ and $p+q+1 > 0$, the latter of which is already contained in the conditions under which (2.1) is true.

The Equation (2.7) may, after a re-arrangement of terms, be written as

$$\frac{1}{(p+q+1)n}\alpha\frac{d}{d\alpha}M(\alpha) - \frac{p+q}{q}\frac{d}{d\alpha}M(\alpha) + \frac{p}{q}M(\alpha) + O\left(\alpha^2 n^{-2}\frac{d}{d\alpha}M(\alpha)\right) = 0, \tag{2.8}$$

a differential equation of the m.g.f. of the mean $\bar{X}$.

Now, consider the function $T = 2(\bar{X} - \mu_1')/(\mu_1'\gamma_1)$ of $\bar{X}$, in which γ_1 is chosen in a way such that $\gamma_1 = 2(\sigma/\mu_1' n^{\frac{1}{2}})$, where μ_1' and σ^2 are the mean $p/(p+q)$ and variance $pq(p+q)^{-2}(p+q+1)^{-1}$ of the parent population (2.1) respectively. Substituting for μ_1' and σ in the expression for γ_1, it may be written as $\gamma_1 = 2q^{\frac{1}{2}}/\{p(p+q+1)n\}^{\frac{1}{2}}$. The expression for T, after substituting for μ_1', may be written in the form

$$T = [2(p+q)/p\gamma_1]\bar{X} - 2/\gamma_1 = a\bar{X} + b, \quad \text{say},$$

where $a = 2(p+q)/p\gamma_1$ and $b = -2/\gamma_1$.

Denoting its m.g.f. by $M_T(\alpha)$ it follows by some well-known property that

$$M_T(\alpha) = e^{b\alpha}M(a\alpha), \tag{2.9}$$

where $M(a\alpha)$ denotes the m.g.f. of the variable $a\bar{X}$. Differentiation w.r.t. α gives

$$\frac{d}{d\alpha}M_T(\alpha) = bM_T(\alpha) + e^{b\alpha}\frac{d}{d\alpha}M(a\alpha). \tag{2.10}$$

Replacing α in (2.8) by $a\alpha = 2(p+q)\alpha/p\gamma_1$ and substituting for $M(a\alpha)$ and $(d/d\alpha)M(a\alpha)$ from (2.9) and (2.10) respectively into (2.8), the latter may, after a re-arrangement of terms be written as

$$\{\alpha/(p+q+1)n - p\gamma_1/2q\}(d/d\alpha)M_T(\alpha) + \{2\alpha/\gamma_1(p+q+1)n\}M_T(\alpha) + O(\gamma_1^{-1}n^{-2}) = 0.$$

Substituting for $1/(p+q+1)n$ from the expression for γ_1 and since γ_1 is of the order $(n^{-\frac{1}{2}})$, this equation reduces to the form

$$(1 - \gamma_1\alpha/2)(d/d\alpha)M_T(\alpha) - \alpha M_T(\alpha) - O(n^{-1}) = 0. \tag{2.11}$$

Hence, an approximation, including all terms of the order $(n^{-\frac{1}{2}})$ is given by (2.11), with the term $O(n^{-1})$ omitted. The latter equation has as solution

$$M_T(\alpha) = (1 - \gamma_1\alpha/2)^{-4/\gamma_1^2}\exp(-2\alpha/\gamma_1),$$

since $M_T(0) = 1$. This expression is the m.g.f. of the standardized Pearson's Type III-distribution with variable t,

$$f(t) = k(1 + \gamma_1 t/2)^{4/\gamma_1^2}\exp(-2t/\gamma_1), \quad -2/\gamma_1 \leqq t \leqq \infty, \tag{2.12}$$

with γ_1, as defined above, as its only parameter, and where

$$k = [(4/\gamma_1^2)^{4/\gamma_1^2-\frac{1}{2}} \exp(-4/\gamma_1^2)]/\Gamma(4/\gamma_1^2).$$

This expression as well as its c.d.f. has been tabulated by Salvosa [17] for $t = -4.99(0.01)9.99$ and $\gamma_1 = 0.0(0.1)1.1$. Furthermore, he has shown that the distribution (2.12) tends to normality when γ_1 tends to zero.

It may be noted that if in (2.7) only terms of $O(1)$ are retained, the differential equation becomes

$$(d/d\alpha)M(\alpha) = [p/(p+q)]M(\alpha) \quad \text{or} \quad M(\alpha) = \exp\{p\alpha/(p+q)\},$$

which is Khintchine's theorem for this distribution. Furthermore, the next "step" in limiting distributions is the Central Limit Theorem in virtue of which

TABLE 1

Parent Population		Parameter γ_1 of Approximation as represented by (2.12)	Conditions under which approx. is valid
Frequency function*	Mean		
Type I $\frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}$, $p, q > 0; 0 \leqq x \leqq 1$	$p/(p+q)$	$2q^{\frac{1}{2}}\{np(p+q+1)\}^{-\frac{1}{2}}$	$p, q > 0$; $q > p+1$
Type IV $k(1+x^2/a^2)^{-m} \exp(-\nu \tan^{-1} x/a)$, $m > 0; -\infty \leqq x \leqq \infty$	$-a\nu/2(m-1)$	$2\nu[\{4(m-1)^2+\nu^2\}(2m-3)n]^{-\frac{1}{2}}$	$2m > 3$†
Type V $\frac{\gamma^{p-1}}{\Gamma(p-1)} x^{-p} \exp(-\gamma/x)$, $0 \leqq x \leqq \infty$	$\gamma/(p-2)$	$2\{n(p-3)\}^{-\frac{1}{2}}$	$p > 3$
Type VI $\frac{1}{B(p,q)} x^{p-1}/(1+x)^{p+q}$, $p, q > 0; 0 \leqq x \leqq \infty$	$p/(q-1)$	$2\{(p+q-1)/np(q-2)\}^{\frac{1}{2}}$	$p > 0$; $q > 2$
Type IX $(m+1)(1+x/a)^m/a$, $-a \leqq x \leqq 0$, which transforms to $(m+1)(1-x/a)^m/a$, $0 \leqq x \leqq a$, if x is replaced by $-x$. This is then a special form of Type I.	$-a/(m+2)$ $a/(m+2)$	$2\{(m+1)/n(m+3)\}^{\frac{1}{2}}$	$m > 1$

* See Kendall ([15], Chapter 6).

† See closing sentences of Section 2.

the sampling distribution of the mean approaches normality. But, the distribution given by (2.12) also tends to normality when γ_1 tends to zero, i.e. when n tends to infinity, and at the same time is an approximation to the sampling distribution of the mean. Hence, this distribution may be considered as a transitional approximation between the exact sampling distribution as given by Irwin [14] and its normal approximation.

The use of (2.12) as an approximation to the sampling distribution of the mean when sampling from a Type I population has been investigated empirically and the results seem promising. Since this investigation is still underway, the results will be published in the near future.

It may be shown that in the case of sampling from Pearson's Type IV, V, VI and IX populations, the sampling distribution of the variable $T = 2(\bar{X} - \mu_1')/(\mu_1'\gamma_1)$, a function of the sample mean $\bar{X}$, is of the same form as (2.12). In all these cases, the latter distribution's only parameter γ_1, is given by $\gamma_1 = 2\sigma/(\mu_1' n^{\frac{1}{2}})$, where μ_1' and σ denote the mean and standard deviation of the sampled population respectively.

The results are summarised in Table 1.

In the case of sampling from a Type IV population, it may be noted that the approximation as given by (2.12) is only valid when $a\nu < 0$, a and ν being parameters of the parent distribution. Whenever $a\nu > 0$, the approximation is given by the expression (2.12), with t replaced by $-t$, defined over the interval $-\infty \leqq t \leqq 2/\gamma_1$.

3. Sampling from a symmetrical non-normal population. Consider a random sample of size n from a population with a Pearson's Type VII-distribution

$$[aB(1/2, m - 1/2)]^{-1}(1 + x^2/a^2)^{-m}, \qquad -\infty \leqq x \leqq \infty. \tag{3.1}$$

Proceeding on the same lines as in Section 2, it may be shown that, to terms of the order (n^{-1}), the differential equation of the m.g.f. of the mean $\bar{X}$ is given by

$$\frac{d}{d\alpha} M(\alpha) - \frac{a^2}{(2m - 3)n} \alpha M(\alpha) - O\left(\alpha^2 n^{-2} \frac{d}{d\alpha} M(\alpha)\right) = 0, \qquad m > \frac{3}{2}, \tag{3.2}$$

the latter being the condition under which the derivation of this equation is true.

The differential equation of the m.g.f. $M_T(\alpha)$ of the variable $T = \bar{X}n^{\frac{1}{2}}/\sigma$, where $\sigma = a/(2m - 3)^{\frac{1}{2}}$ is the standard deviation of the distribution (3.1), follows from (3.2) after the necessary substitutions as

$$(d/d\alpha)M_T(\alpha) - \alpha M_T(\alpha) - O(n^{-1}) = 0. \tag{3.3}$$

Taking the initial condition, $M_T(0) = 1$, in consideration, the differential Equation (3.3), when $n \to \infty$, has as solution

$$M_T(\alpha) = \exp(\alpha^2/2), \tag{3.4}$$

which is the m.g.f. of the normal distribution with zero mean and unit variance. Hence, when $2m > 3$, the normal distribution may be considered as a first order

approximation to the sampling distribution of the mean for samples of size n from a Pearson's Type VII population.

A similar result has been found for the approximation to the sampling distribution of the mean for samples from a Type II-population.

4. Higher order approximations. When, in the preceding sections, all terms of the order (n^{-2}) are retained in the evaluation of the expression for $(d/d\alpha)M(\alpha)$, a first order differential equation for the m.g.f. $M(\alpha)$, containing terms of the order (n^{-2}), is obtained after some tedious algebra.

In the case of sampling from a Pearson's Type VII population the differential equation of the m.g.f. of the mean $\bar{X}$ becomes

$$\left\{1 - \frac{a^2}{(2m-3)(2m-4)n^2}\alpha^2\right\}\frac{d}{d\alpha}M(\alpha) - \frac{a^2}{(2m-3)n}\alpha M(\alpha) = 0, \qquad m > 2, \tag{4.1}$$

which has as its solution

$$M(\alpha) = \left\{1 - \frac{a^2}{(2m-3)(2m-4)n^2}\alpha^2\right\}^{-n(m-2)}, \qquad m > 2, \tag{4.2}$$

since $M(0) = 1$.

From this expression the moments of the distribution for which it is the m.g.f., may be obtained in the usual way.

It is interesting to note that the coefficient of excess in this case is $\gamma_2 = 3/(m-2)n$. Since the same coefficient for (3.1) is given by $6/(2m-5)$, the coefficient of excess for the exact sampling distribution of the mean for samples of size n is given by $6/(2m-5)n$. The coefficient of excess of the normal approximation, as obtained in Section 3, is naturally zero. Hence, as far as peakedness is concerned, it is clear that the approximation of which (4.2) is the m.g.f., lies between the exact sampling distribution, as given by Irwin [14], and the normal approximation, since $6/(2m-5)n > 3/(m-2)n > 0$, for $2m > 5$.

Furthermore, by the use of a result published by Steyn [20] in 1960, the differential equation of the m.g.f., such as given by (4.1), may be transformed to a differential equation of the frequency function of the sample mean, from which the latter may be obtained as a solution. It is hoped to obtain higher order approximations on these lines.

5. Acknowledgment. The author wishes to thank Professor H. S. Steyn for his valuable discussions during the preparation of this paper and to Professor Douglas G. Chapman and the referee for their criticism and suggestions in improving the original version of the paper.

REFERENCES

[1] BAKER, G. A. (1930). Distribution of means of samples of n drawn at random from a population represented by a Gram-Charlier series. *Ann. Math. Statist.* **1** 199–204.

[2] BAKER, G. A. (1930). Random sampling from non-homogeneous populations. *Metron.* **8**, No 3 67–87.
[3] BATEN, W. D. (1933). Frequency laws for the sum of n variables which are subjected to given frequency laws. *Metron.* **10**, No 3 75–91.
[4] BATEN, W. D. (1934). The probability law for the sum of n independent variables, each subject to the law $(1/2h)$ sech $(\pi x/2h)$. *Bull. Amer. Math. Soc.* **40** 284–290.
[5] BERRY, A. C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. *Trans. Amer. Math. Soc.* **49** 122–136.
[6] BOSE, RAJ CHANDRA (1938). On the distribution of the means of samples drawn from a Bessel function population. *Sankhyā* **3** 262–264.
[7] CHURCH, A. E. R. (1926). On the means and squared standard-deviations of small samples from any population. *Biometrika* **18** 321–394 (correction, **24** 292).
[8] DANIELS, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25** 631–650.
[9] ESSEEN, C.-G. (1945). Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law. *Acta Math.* **77** 1–125.
[10] GNEDENKO, B. V. and KOLMOGOROV, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables.* Addison-Wesley, Cambridge.
[11] HAIGHT, FRANK A. (1961). Index to the distributions of mathematical statistics. *Nat. Bur. Standards Appl. Math. Ser.* **65**B 23–60.
[12] HALL, P. (1927). The distribution of means for samples of size N drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika* **19** 240–245.
[13] IRWIN, J. O. (1927). On the frequency-distribution of the means of samples from a population having any law of frequency with finite moments, with special reference to Pearson's Type II. *Biometrika* **19** 225–239, **21** 431–432.
[14] IRWIN, J. O. (1930). On the frequency-distribution of the means of samples from certain Pearson's types. *Metron.* **8**, No 4 51–105.
[15] KENDALL, M. G. (1945). *The Advanced Theory of Statistics.* Charles Griffin, London.
[16] RAO, C. RADHAKRISHNA (1942). On the sum of n observations from different gamma type populations. *Science and Culture* **7** 614–615.
[17] SALVOSA, L. R. (1930). Tables of Pearson's Type III function. *Ann. Math. Statist.* **1** 191–198, with the tables in an appendix, 1–187.
[18] SHRIVASTAVA, M. P. (1940). The distribution of the mean for certain Bessel function populations. *Science and Culture* **6** 244–245.
[19] STEYN, H. S. (1954). *'n Studie van Meerveranderlike Kansfunksies deur middel van Differensiaalvergelykinge vir die Momentefunksies.* (in Afrikaans). Unpublished doctoral thesis, Univ. of Pretoria, South Africa.
[20] STEYN, H. S. (1960). On regression properties of multivariate probability functions of Pearson's types. *Indag. Math.* **22** 302–311.
[21] WELKER, E. L. (1947). The distribution of the mean. *Ann. Math. Statist.* **18** 111–117.
[22] WHITTAKER, E. T. and WATSON, G. N. (1950). *A Course of Modern Analysis.* Cambridge Univ. Press.