COMPARISON OF PEARSONIAN APPROXIMATIONS WITH EXACT SAMPLING DISTRIBUTIONS OF MEANS AND VARIANCES IN SAMPLES FROM POPULATIONS COMPOSED OF THE SUMS OF NORMAL POPULATIONS

By G. A. Baker

1. Introduction. Biological and sociological data are often "non-homogeneous" and of such a nature as not to be easily separated into components. Non-homogeneous populations have been discussed by Karl Pearson, Charlier, and others. Non-normal material has been discussed by many writers. See for example, A. E. R. Church [1] and J. M. LeRoux [2] for a discussion of moments of the distributions of the means and variances for samples from non-normal material.

In a previous paper [3] the author has given the distributions of the means and standard deviations of samples from certain non-homogeneous populations. The purpose of the present paper is to extend the results given in [3] and to compare the moment approach of the Pearsonian school with the true distributions.

2. Moments of the distribution of means of samples of n from a non-homogeneous population. Consider a population with distribution

$$f(x) = \frac{1}{(1+k)\sqrt{2\pi}}\left[e^{-\frac{1}{2}x^2} + \frac{k}{\sigma}e^{-\frac{1}{2\sigma^2}(x-m)^2}\right]. \tag{2.1}$$

The first four moments of (2.1) about $x = 0$ are

$$\begin{aligned} v_1' &= \frac{km}{1+k} \\ v_2' &= \frac{1}{1+k}[1 + k(\sigma^2 + m^2)] \\ v_3' &= \frac{km}{1+k}[3\sigma^2 + m^2] \\ v_4' &= \frac{1}{1+k}[3 + k(3\sigma^4 + 6m^2\sigma^2 + m^4)]. \end{aligned} \tag{2.2}$$

The means of samples of n drawn at random from (2.1) are distributed according to

$$\frac{n}{\sqrt{2\pi}(1+k)^n}\left[\sum_{s=0}^{n}\binom{n}{s}\frac{k^s}{\sqrt{s\sigma^2 + n - s}}\exp\left\{-\frac{n^2\left(x - \frac{s}{n}m\right)^2}{s\sigma^2 + n - s}\right\}\right]. \tag{2.3}$$

Denote by m'_p the moments of (2.3) about $x = 0$ and by m_p the moments about the mean. Then in view of the relations

$$\frac{n!\, s^r}{(n-s)!\, s!} = \sum_{i=0}^{r-1} A_{ri} \frac{n!}{(n-s)!\,(s-r+i)!}$$

$$A_{r0} = 1, \quad A_{r(r-1)} = 1, \quad A_{31} = 3, \quad A_{41} = 6, \quad A_{42} = 7, \tag{2.4}$$
$$A_{51} = 10, \quad A_{52} = 25, \quad A_{53} = 15,$$

and similar relations, and reduction to moments about the mean we obtain

$$\begin{aligned}
m'_1 &= \frac{km}{1+k} = v'_1 \\
m_2 &= \frac{1}{n(1+k)}\left[1 + k\sigma^2 + \frac{km^2}{1+k}\right] \\
m_3 &= \frac{km}{n^2(1+k)^2}\left[3\sigma^2 - 3 + \frac{1-k}{1+k}m^2\right] \\
m_4 &= \frac{1}{n^3(1+k)^2}\Big[3k(nk+1)\sigma^4 + 3(n+k) + 6(n-1)k\sigma^2 \\
&\quad + \frac{6k}{1+k}\{k + (n-1)\}m^2 + \frac{6k}{1+k}\{(n-1)k+1\}m^2\sigma^2 \\
&\quad + \frac{k}{(1+k)^2}\{k^2 + (3n-4)\,k + 1\}m^4\Big] \qquad (2.5) \\
m_5 &= \frac{k}{n^4(1+k)^3}\Big[15\{(2n-1)k+1\}m\sigma^4 - 15\{k + (2n-1)\}m \\
&\quad + 30\,(n-1)(1-k)m\sigma^2 \\
&\quad + \frac{10}{1+k}\{-(n-1)k^2 + 4(n-1)k + 1\}m^3\sigma^2 \\
&\quad + \frac{10}{1+k}\{-k^2 - 4(n-1)k + (n-1)\}m^3 \\
&\quad + \frac{1}{(1+k)^2}\{-k^3 + (-10n+11)k^2 + (10n-11)k + 1\}m^5\Big].
\end{aligned}$$

The expressions for the first five moments agree with the results given by Church and Tchebycheff.

The betas of (2.4) are

$${}_1B_1 = \frac{k^2m^2}{n(1+k)} \frac{\left[3\sigma^2 - 3 + \frac{1-k}{1+k}m^2\right]^2}{\left[k\sigma^2 + 1 + \frac{k}{1+k}m^2\right]^3} \tag{2.6}$$

$$(2.7)\quad {}_1B_2 - 3 = \frac{k\left[3 + 3\sigma^4 - 6\sigma^2 - \frac{6}{1+k}m^2 + \frac{6}{1+k}m^2\sigma^2 + \frac{k^2 - 4k + 1}{(1+k)^2}m^4\right]}{n\left[k\sigma^2 + 1 + \frac{k}{1+k}m^2\right]^2}.$$

${}_1B_1$ vanishes if $k = 0$, $m = 0$, or $k = 1$ and $\sigma = 1$. If k and σ are constant and m approaches infinity ${}_1B_1$ approaches $(1 - k)^2/nk$. If k and m are constant and σ approaches infinity ${}_1B_1$ approaches zero. ${}_1B_2 - 3$ vanishes if $k = 0$, $k = \infty$, or if $m = 0$ and $\sigma = 1$. If k and σ are constant and m approaches

TABLE I

m_5 and ${}_pm_5$ compared for four sets of values of k, σ^2, and m

Sets of values k	σ^2	m	m_5	${}_pm_5$
1/2	1/4	1.1	$-\frac{4.599}{n^3} + \frac{1.228}{n^4}$	$\frac{1}{n^3}\frac{\left(-4.250 + \frac{.235}{n} - \frac{.115}{n^2}\right)}{\left(1 + \frac{.148}{n}\right)}$
1/3	1	3.2	$\frac{89.702}{n^3} - \frac{39.322}{n^4}$	$\frac{1}{n^3}\frac{\left(71.313 - \frac{26.709}{n} - \frac{4.886}{n^2}\right)}{\left(1 + \frac{.584}{n}\right)}$
−1/4	1/4	0.5	$\frac{4.640}{n^3} - \frac{1.744}{n^4}$	$\frac{1}{n^3}\frac{\left(4.746 - \frac{.095}{n} - \frac{.165}{n^2}\right)}{\left(1 + \frac{.199}{n}\right)}$
1	4	5.6	$\frac{1{,}302.840}{n^3} - \frac{646.060}{n^4}$	$\frac{1}{n^3}\frac{\left(762.035 - \frac{347.204}{n} - \frac{2.405}{n^2}\right)}{\left(1 + \frac{.277}{n}\right)}$

infinity then ${}_1B_2 - 3$ approaches $(k^2 - 4k + 1)/nk$. If k and m are constant and σ approaches infinity then ${}_1B_2 - 3$ approaches $3/nk$.

It is of interest to compare the higher moments of (2.3) with the higher moments calculated from the first four moments on the assumption of a Pearson curve in place of (2.3). On this assumption

$$(2.8)\quad {}_pm_5 = \frac{2m_3(m_4 + 7m_2^2m_4 - 3m_2m_3^2)}{9m_2^3 - m_2m_4 + 3m_3^2}.$$

It is seen that (2.8) bears little resemblance to m_5. If we consider the difference ${}_pm_5 - m_5$ we see that it is of the same order in $1/n$ as is m_5 and the

numerator is of the 16th degree in k, m, and σ; a very complicated locus. m_5 and ${}_pm_5$ are compared for certain values of the parameters of (2.1) in Table I.

Table I shows that the coefficients of $1/n^3$ in the expressions for m_5 and ${}_pm_5$ differ by from two to more than 40 per cent. The coefficients of $1/n^4$ differ even more. The assumption of Karl Pearson's curves to represent the distribution of means of samples of n from non-homogeneous populations seems to be adequate in some cases but inadequate in others even for moderate values of the parameters.

3. Moments of the distribution of variances. In [3] an estimate of n times the standard deviation squared is expressed as

$$\overline{W} = (n - s)\,\bar{\sigma}_1^2 + s\bar{\sigma}_2^2 + \frac{(n - s)s}{n}(\bar{m}_1 + \bar{m}_2)^2, \tag{3.1}$$

where a bar over a letter means an estimate of the corresponding population parameter and where $(n - s)$ denotes the number drawn from the first component of (2.1) and s denotes the number from the second component.

For the direct calculation of the moments of the distribution of variances it is easier not to use the distribution given in [3], but to proceed as follows. Put

$$(n - s)\,\bar{\sigma}_1^2 = y, \qquad s\bar{\sigma}_2^2 = x, \qquad \frac{(n - s)s}{n}(\bar{m}_1 + \bar{m}_2)^2 = z.$$

Of course, for population (2.1) $\sigma_1 = 1$, $\sigma_2 = \sigma$, $m_1 = 0$, $m_2 = m$. The variables, x, y, z are all independent in the probability sense and their probability distributions are well known. Hence the moments of

$$\frac{\overline{W}}{n} = \frac{x + y + z}{n} \tag{3.2}$$

can be directly calculated.

For instance, if $p = 1$ then

$$M_1' = \frac{n - 1}{n}\,\frac{1}{1 + k}\left[k\sigma^2 + 1 + \frac{k}{1 + k}\,m^2\right]. \tag{3.3}$$

In general, of course, the moments about the mean check with the values given by Church.

It is generally recommended to represent the distributions of variances of samples from non-normal parents by Pearson's curves. Let us examine the results of this procedure in a special case.

Suppose that the sampled population is

$$f(x) = \frac{1}{2\sqrt{2\pi}}\left[e^{-\frac{1}{2}x^2} + e^{-\frac{1}{2}(x-3.4)^2}\right]. \tag{3.4}$$

The first eight moments of (3.4) which are needed in the calculation of the first four moments of the variances are:

(3.5)
$$\begin{aligned} v_1' &= 1.7000 & v_5 &= 0 \\ v_2 &= 3.8900 & v_6 &= 294.47 \\ v_3 &= 0 & v_7 &= 0 \\ v_4 &= 28.692 & v_8 &= 3{,}818.4. \end{aligned}$$

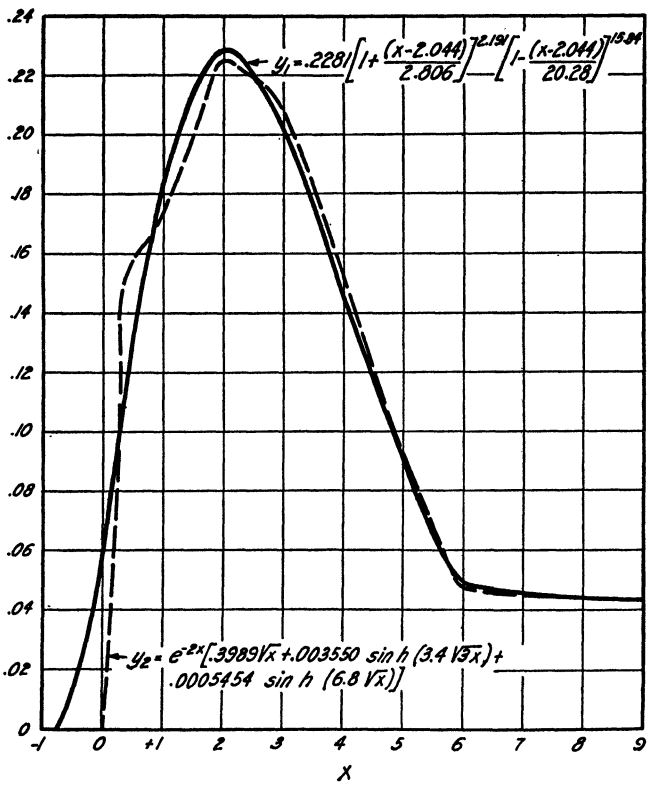

FIG. 1. COMPARISON OF THE TRUE DISTRIBUTION OF THE VARIANCES OF SAMPLES OF 4 DRAWN FROM THE NON-HOMOGENEOUS POPULATION (3.4) WITH THE CORRESPONDING EMPIRICAL PEARSON CURVE

The first four moments of the variances of samples of 4 from (3.4) are:

(3.6)
$$\begin{aligned} {}_2M_1' &= 2.918 & {}_2M_3 &= 4.745 \\ {}_2M_2 &= 3.396 & {}_2M_4 &= 41.52. \end{aligned}$$

Hence ${}_2B_1 = .60$ and ${}_2B_2 = 3.6$, $\kappa = -.87$ which calls for a type 1 curve. The equation of the curve is

(3.7)
$$y_1 = 0.2281\left(1+\frac{x}{2.806}\right)^{2.191}\left(1-\frac{x}{20.28}\right)^{15.84}$$

with its origin at its mode. The corresponding true distribution with the origin at the beginning of the range is

$$y_2 = e^{-2x}[.3989\sqrt{x} + .003550 \sinh (3.4\sqrt{3x}) + .0005454 \sinh (6.8\sqrt{x})]. \tag{3.8}$$

Distribution (3.8) differs slightly from the corresponding result given in [3] because of an error in that paper.

The two distributions are compared in Figure 1. It is seen that the two distributions are quite different. As the number of components of distributions similar to (3.8) increases, which is true as n increases, the distributions may be expected to become smoother and more closelv representable by a single smooth curve.

4. Summary. The moments of the distribution of the means of samples of n from a non-homogeneous population composed of two normal components are given up to and including the fifth. This fifth moment is compared with the fifth moment calculated on the assumption of Pearson's curves to represent the distribution of means. The B's of the distributions of the means are discussed in certain limiting cases. It appears that for small samples and extreme values of the parameters, and in some cases of moderate values of the parameters, the Pearsonian approximations give poor results.

Some identities involving the binomial coefficients are given which permit the reduction of the moments of the distribution of means calculated directly to forms given elsewhere [1]. A method is given for the direct calculation of the moments of the variances of samples from a non-homogeneous population composed of two normal components. An indication of the closeness with which a Pearson curve can be made to fit the distribution of variances in small samples from a non-homogeneous population is given in Figure 1.

REFERENCES

[1] A. E. R. CHURCH, "On the means and squared standard-deviations of small samples from any population," *Biometrika*, Vol. 18 (1926), pp. 321–394.

[2] J. M. LEROUX, "A study of the distribution of the variance in small samples," *Biometrika*, Vol. 23 (1931), pp. 134–190.

[3] G. A. BAKER, "Random Sampling from non-homogeneous populations," *Metron*, Vol. 8, no. 3, (1930).

UNIVERSITY OF CALIFORNIA,
DAVIS, CALIF.