APPROXIMATION OF THE DISTRIBUTION OF THE PRODUCT OF BETA VARIABLES BY A SINGLE BETA VARIABLE

By John W. Tukey and S. S. Wilks

Princeton University

1. Introduction. In an article published elsewhere in the present issue of the *Annals of Mathematical Statistics* [1] the g-th moments of two statistical test criteria L_{mvc} and L_{vc} were found to have the following expressions, respectively:

$$(1)\qquad (k-1)^{g(k-1)}\prod_{i=1}^{k-1}\left[\frac{\Gamma(\frac{1}{2}(n-1-i)+g)}{\Gamma(\frac{1}{2}(n-1-i))}\right]\cdot\frac{\Gamma(\frac{1}{2}n(k-1))}{\Gamma(\frac{1}{2}n(k-1)+g(k-1))}$$

and

$$(2)\qquad (k-1)^{g(k-1)}\prod_{i=1}^{k-1}\left[\frac{\Gamma(\frac{1}{2}(n-1-i)+g)}{\Gamma(\frac{1}{2}(n-1-i))}\right]\cdot\frac{\Gamma(\frac{1}{2}(n-1)(k-1))}{\Gamma(\frac{1}{2}(n-1)(k-1)+g(k-1))}\,.$$

If we denote by $(a)_g$ the expression $a(a+1)(a+2)\cdots(a+g-1)$ and make use of the fact that

$$(3)\qquad \Gamma(a+g) = \Gamma(a)\cdot(a)_g$$

and

$$(4)\qquad \Gamma(a+rg) = \Gamma(a)\cdot(a)_{rg} = \Gamma(a)\cdot r^{rg}\prod_{i=1}^{r}\left(\frac{a+i-1}{r}\right)_g$$

where r is a positive integer, the two moments (1) and (2) reduce to

$$(5)\qquad \frac{\prod_{i=1}^{k-1}\left(\frac{n}{2}+\frac{i-k-1}{2}\right)_g}{\prod_{i=1}^{k-1}\left(\frac{n}{2}+\frac{i-1}{k-1}\right)_g}\quad\text{and}\quad\frac{\prod_{i=1}^{k-1}\left(\frac{n-1}{2}+\frac{i-k}{2}\right)_g}{\prod_{i=1}^{k-1}\left(\frac{n-1}{2}+\frac{i-1}{k-1}\right)_g}$$

respectively.

For any given value of i ($i = 1, 2, \cdots, k-1$) the ratio

$$\frac{\left(\frac{n}{2}+\frac{i-k-1}{2}\right)_g}{\left(\frac{n}{2}+\frac{i-1}{k-1}\right)_g}\quad\text{or}\quad\frac{\left(\frac{n-1}{2}+\frac{i-k}{2}\right)_g}{\left(\frac{n-1}{2}+\frac{i-k}{k-1}\right)_g}$$

may be expressed in the form

$$\frac{\Gamma(p_i+g)}{\Gamma(p_i+q_i+g)}$$

which is the g-th moment of a beta variable u_i distributed according to

$$\frac{\Gamma(p_i+q_i)}{\Gamma(p_i)\Gamma(q_i)}\,u_i^{p_i-1}(1-u_i)^{q_i-1}\,du_i\,.$$

Each of the moments in (5) is therefore of the form

$$\prod_{i=1}^{k-1} \frac{\Gamma(p_i + g)}{\Gamma(p_i + q_i + g)}.$$

Thus, L_{mvc} and L_{vc} are each distributed like thę product of $k - 1$ independent beta variables.

Each of the moments in (5) can be expressed in the general form

$$M_g = \frac{\prod_{i=1}^{r'}\left(\frac{1}{x} - A_i + 1\right)_g}{\prod_{i=1}^{r'}\left(\frac{1}{x} - B_i + 1\right)_g} \tag{6}$$

where $x = \frac{2}{n}\left(\text{or } \frac{2}{n-1}\right)$, A_i and B_i are real numbers.

Other likelihood ratio statistical test criteria which have been discussed in the literature have moments which can be expressed in the general form (6). For example, the likelihood ratio criterion L_1 for testing the homogeneity of sample variances [2, Neyman and Pearson 1931] has moments of this type. The generalized L_1 criterion for samples from a normal multivariate population [3, Wilks 1933] has such moments. The criterion for testing sphericity [4, Mauchly 1940] of a normal multivariate distribution has moments of this kind. All test criteria having this type of moment lie on the interval (0, 1). The exact distribution functions of the criteria, except possibly for $r = 1$ or 2 in some cases, are very complicated.

The purpose of this note is to consider a method of finding a fractional power of the test criterion which is approximately distributed (in a sense to be described later) according to an incomplete beta (Pearson Type I) distribution function,

$$dF(u) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} u^{p-1}(1-u)^{q-1}du \tag{7}$$

and to find the appropriate values of p, q, and the exponent of the criterion.

2. Generalized hypergeometric series as moment generating functions. Suppose L is a statistical test criterion, or more generally a random variable having as its g-th moment the expression (6). The moment generating function $\varphi(t)$ of L can be expressed as

$$\varphi(t) = \sum_{g=0}^{\infty} M_g t^g = \frac{\prod_{i=1}^{r'}\left(\frac{1}{x} - A_i + 1\right)_g}{\prod_{i=1}^{r'}\left(\frac{1}{x} - B_i + 1\right)_g} \cdot t^g. \tag{8}$$

This can be written as

$$\varphi(t) = {}_{r'+1}F_{r'}\left[\begin{matrix} 1, \frac{1}{x} - A_1, \cdots, \frac{1}{x} - A_{r'}; t \\ \frac{1}{x} - B_1, \cdots, \frac{1}{x} - B_{r'} \end{matrix}\right] \tag{9}$$

where ${}_{r'+1}F_{r'}$ [ ] is a generalized hypergeometric series [5, Bailey 1935]. We shall not make explicit use of this fact; instead, we shall work with the coefficient of t^g in the series, i.e., M_g.

Let us consider

$$\ln M_g = \sum_{i=1}^{r'} \ln\left(\frac{1}{x} - A_i + 1\right)_g - \sum_{i=1}^{r'} \ln\left(\frac{1}{x} - B_i + 1\right)_g . \tag{10}$$

To expand this in a power series in x consider a single term

$$\begin{aligned} \ln\left(\frac{1}{x} - A + 1\right)_g &= \sum_{j=1}^{g} \ln\left(\frac{1}{x} - A + j\right) \\ &= -g \ln x + g \ln (1 - Ax) + \sum_{j=1}^{g} \ln\left(1 + \frac{jx}{1 - Ax}\right). \end{aligned} \tag{11}$$

Now

$$1 + \frac{jx}{1 - Ax} = 1 + jx + jAx^2 + jA^2x^3 + \cdots .$$

Writing

$$S_m(g) = \sum_{i=1}^{g} j^m ,$$

and using the usual expansion for $\ln (1 + x)$, we find

$$\begin{aligned} \ln\left(\frac{1}{x} - A + 1\right) = -g \ln x + [S_1(g) - Ag]x + [\tfrac{1}{2}A^2 + AS_1(g) - \tfrac{1}{2}S_2(g)]x^2 \\ + [-\tfrac{1}{3}A^3 + A^2S_1(g) - AS_2(g) + \tfrac{1}{3}S_3(g)]x^3 + \cdots . \end{aligned}$$

Applying this expansion to the separate terms in (10) and writing

$$C_m = \sum_{i=1}^{r'} A_i^m - \sum_{i=1}^{r'} B_i^m \tag{12}$$

the terms not involving A_i or B_i cancel out leaving

$$\begin{aligned} \ln M_g = (-C_1g)x + [\tfrac{1}{2}C_2 + C_1S_1(g)]x^2 \\ + [-\tfrac{1}{3}C_3 + C_2S_1(g) - C_1S_2(g)]x^3 + \cdots . \end{aligned} \tag{13}$$

We shall return to this expression later.

3. Powers of a beta variable. If u has (7) as its distribution function, then

$$E(u^h) = \frac{(p)_h}{(p + q)_h} . \tag{14}$$

If $v = u^r$, r integral, then its g-th moment is given by setting $h = rg$ in (14).

We have

$$E(v^g) = \frac{(p)_{rg}}{(p+q)_{rg}}.$$

But

$$(p)_{rg} = r^{rg}\left(\frac{p}{r}\right)_g \cdot \left(\frac{p+1}{r}\right)_g \cdots \left(\frac{p+r-1}{r}\right)_g$$

so that

$$E(v^g) = \frac{\prod_{i=1}^{r}\left(\frac{p+i-1}{r}\right)_g}{\prod_{i=1}^{r}\left(\frac{p+q+i-1}{r}\right)_g} \tag{15}$$

which is a special case of (6) when p is of order n.

Putting $\frac{1}{x} = \frac{p+q}{r}$, $A_i = 1 + (q - i + 1)/r$, and $B_i = 1 - (i - 1)/r$ we have

$$C_1 = q,$$

$$C_2 = \frac{q^2}{r} + q\left(1 + \frac{1}{r}\right).$$

For any given moment of the form (6), from which x, C_1, and C_2 can be computed, we determine p, q, and r so as to satisfy

$$\frac{p+q}{r} = \frac{1}{x} \tag{16}$$

$$q = C_1$$

and to satisfy, as nearly as possible, (with r integral)

$$\frac{q^2}{r} + q\left(1 + \frac{1}{r}\right) = C_2, \tag{17}$$

i.e.,

$$r = \frac{q(q+1)}{C_2 - q}.$$

The use of fractional r is obviously suggested, but its value and validity are not discussed here. Using the values of p, q and r thus obtained, the distribution of the criterion L (having moments (7)), is given approximately by

$$\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} (\sqrt[r]{L})^{p-1}(1 - \sqrt[r]{L})^{q-1}\, d(\sqrt[r]{L}) \tag{18}$$

where the approximation is such that all moments are correct through terms of order $\left(\frac{r}{p+q}\right)$ (when moments are expanded in series of $\frac{r}{p+q}$) and nearly (exactly if there is an integral value of r satisfying (17)) correct through terms of order $\left(\frac{r}{p+q}\right)^2$.

4. Examples. Returning to the g-th moment of L_{mvc} given by the first expression in (5) we have

$$x = \frac{2}{n}, \qquad r' = k - 1$$

$$A_i = \frac{k+3-i}{2}, \qquad B_i = \frac{k-i}{k-1},$$

$$C_1 = \sum_{i=1}^{k-1} A_i - \sum_{i=1}^{k-1} B_i = \tfrac{1}{4}(k^2 + 3k - 6)$$

$$C_2 = \sum_{i=1}^{k-1} A_i^2 - \sum_{i=1}^{k-1} B_i^2 = \frac{1}{24}[(k+2)(k+3)(2k+5) - 84] - \frac{k(2k-1)}{6(k-1)}.$$

To determine p, q and r for the fitted distribution of L_{mvc} we set

$$\frac{p+q}{r} = \frac{n}{2}$$

$$q = \tfrac{1}{4}(k^2 + 3k - 6)$$

$$r = \frac{q(q+1)}{C_2 - q}$$

and solve for p, q and r. We have the following table of values, p, q and r for various values of k (p being calculated by using the rounded values of r):

k	r	r (rounded)	p	q
3	2	2	$n - 3$	3
4	2.93	3	$1.5n - 5.5$	5.5
5	3.82	4	$2n - 8.5$	8.5
6	4.68	5	$2.5n - 12$	12
7	5.53	6	$3n - 16$	16
8	6.35	6	$3n - 20.5$	20.5
9	7.17	7	$3.5n - 25.5$	25.5
10	7.96	8	$4n - 31$	31
20	15.76	16	$8n - 113.5$	113.5

Thus, by rounding r off to the nearest integer and using this rounded value of r in determining p, we have values of p, q and r for each value of k, which, when substituted in (18) give us the desired fitted beta distribution for L_{mvc}. For $k = 3$, the fitted distribution is the exact distribution.

For the g-th moment of L_{vc} which is given by the second expression in (5), it is convenient to expand in powers of $\frac{2}{n-1}$. Hence we have

$$x = \frac{2}{n-1}, \qquad r' = k - 1$$

$$A_i = \frac{k+2-i}{2}, \qquad B_i = \frac{k-i}{k-1}$$

$$C_1 = \tfrac{1}{4}(k^2 + k - 4)$$

$$C_2 = \tfrac{1}{24}[(k+1)(k+2)(2k+3) - 30] - \frac{k(2k-1)}{6(k-1)}.$$

To determine p, q and r for fitting the distribution function of L_{vc} we put

$$\frac{p+q}{r} = \frac{n-1}{2}$$

$$q = \tfrac{1}{4}(k^2 + k - 4)$$

$$r = \frac{q(q+1)}{C_2 - q}.$$

We have the following table of values of p, q and r for several values of k:

k	r	r (rounded)	p	q
3	2	2	$n - 3$	2
4	2.88	3	$1.5n - 5.5$	4
5	3.71	4	$2n - 8.5$	6.5
6	4.52	5	$2.5n - 12$	9.5
7	5.32	5	$2.5n - 15.5$	13
8	6.14	6	$3n - 20$	17
9	6.88	7	$3.5n - 25$	21.5
10	7.82	8	$4n - 30.5$	26.5
20	15.26	15	$7.5n - 111.5$	104

By rounding r off to the nearest integer, and using the rounded value of r in determining p, we have values of p, q and r for each value of k which, when substituted in (18), give us the desired fitted beta distribution for L_{vc}. For $k = 3$, the fitted distribution is the exact distribution.

For a given value of k, approximate 5% and 1% points of $\sqrt[r]{L_{mvc}}$ and $\sqrt[r]{L_{vc}}$ can therefore be obtained from Thompson's [6] tables of the Incomplete Beta Function by entering the tables with $\nu_1 = 2q$, and $\nu_2 = 2p$. For example, for $k = 6$ the 5% and 1% points of $\sqrt[5]{L_{mvc}}$ are obtained by entering Thompson's tables with $\nu_1 = 24$, and $\nu_2 = 5n - 24$.

REFERENCES

[1] S. S. Wilks, "Sample criteria for testing equality of means, equality of variances and equality of covariances in a normal multivariate population," *Annals of Math. Stat.* Vol. 17 (1946) pp. 257–281.

[2] J. Neyman and E. S. Pearson, "On the problem of k samples," *Bulletin de l'Acadèmie Polonaise des Sciences et des Lettres, Sèrie A, Sciences Mathèmatiques, 1930 and 1931*, pp. 460–481.

[3] S. S. Wilks, "Certain generalizations in the analysis of variance," *Biometrika*, Vol. 24 (1932), pp. 471–494.

[4] John W. Mauchly, "Significance test for sphericity of a normal n-variate distribution," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 204–209.

[5] W. N. Bailey, *Generalized Hypergeometric Series*, Cambridge Tracts in Mathematics and Mathematical Physics, No. 32, Cambridge University Press, 1935.

[6] Catherine M. Thompson, "Tables of percentage points of the Incomplete Beta Function," *Biometrika*, Vol. 32, Part II (1941), pp. 187–191.