A k-SAMPLE EXTENSION OF THE ONE-SIDED TWO-SAMPLE SMIRNOV TEST STATISTIC[1]

By W. J. Conover

Kansas State University

1. Introduction. Let $F_{1,n}(x)$ and $F_{2,m}(x)$ be the empirical cdf's (cumulative distribution functions) of the random samples $(X_{1,1}, \cdots, X_{1,n})$ and $(X_{2,1}, \cdots, X_{2,m})$ drawn from populations with continuous cdf's $F_1(x)$ and $F_2(x)$, respectively. The one-sided, two sample Smirnov test statistic $D^+(n, m) = \sup_x (F_{1,n}(x) - F_{2,m}(x))$ provides a test of $H_0:F_1 = F_2$ that is consistent against all alternatives of the type $H_1:F_1(x) > F_2(x)$ for some x. A k-sample extension of $D^+(n, m)$ would be useful in situations where the experimenter has k populations, $k > 2$, and may legitimately assume, from biological or other nonmathematical considerations, that $F_1(x) \geqq F_2(x) \geqq \cdots \geqq F_k(x)$ for all x. Such k sample extensions have been obtained, in various forms, by Ozols (1956), Darling (1957), David (1958), Kiefer (1959), Dwass (1960), Birnbaum and Hall (1960), and Conover (1965), among others. However, each extension, except that of Conover (1965), fails to furnish the small sample distribution function of a test statistic that may be used for all $k \geqq 2$.

This paper furnishes the small sample, and asymptotic, distribution functions of a k-sample extension of $D^+(n, n)$, valid for all $k \geqq 2$, but restricted to equal sample sizes for all k samples. Such a restriction is not surprising since the distribution function in the two-sample case is still not known for all cases of $n \neq m$. Consistency is discussed in Section 5.

2. Principal results. While the interesting result is the corollary, the more general theorem is no more difficult to prove, although the notation may appear cumbersome. For $i = 1, 2, \cdots, k$, let $F_{i,n_i}(x)$ be the empirical cdf of a random sample $(X_{i,1}, \cdots, X_{i,n_i})$ of size n_i drawn from a population with the continuous cdf $F_i(x)$. Let $I_i(x) = n_i F_{i,n_i}(x)$ and let c_i' represent the smallest integer not less than c_i. Let

$$(2.1) \quad P_k^* = P(\sup_x (I_i(x) - I_{i+1}(x)) < c_i \,;\, i = 1, 2, \cdots, k - 1).$$

It is assumed hereafter that $H_0:F_1 \equiv F_2 \equiv \cdots \equiv F_k$ is true.

Theorem.

$$(2.2) \qquad P_k^* = |A^{k \times k}|$$

where $|A^{k\times k}|$ *is the determinant of the* $k \times k$ *matrix* A *whose elements are*

$$\begin{aligned} a_{ij} &= 0 \quad \textit{if} \quad n_j - \textstyle\sum_{\alpha=1}^{i-1} c_\alpha' + \sum_{\beta=1}^{j-1} c_\beta' < 0 \\ &= n_j!/(n_j - \textstyle\sum_{\alpha=1}^{i-1} c_\alpha' + \sum_{\beta=1}^{j-1} c_\beta')! \ \textit{otherwise}. \end{aligned}$$

Received 20 February 1967.

[1] Prepared with support from the Kansas Agricultural Experiment Station, Manhattan as Contribution No. 110, Department of Statistics and Statistical Laboratory.

COROLLARY. *Let* $F_{i,n}(x)$, $i = 1, 2, \cdots, k$, *be* k *empirical* cdf's *of* k *random samples of equal size* n *drawn from the same, or identical, populations. Then*

$$P(\sup_{x,i<k}(F_{i,n}(x) - F_{i+1,n}(x)) < c/n) = |B^{k\times k}| \tag{2.3}$$

where $|B^{k\times k}|$ *is the determinant of the* $k \times k$ *matrix* B *with elements*

$$\begin{aligned} b_{ij} &= 0 \quad \textit{if} \quad n - (i-j)c' < 0 \\ &= n!/(n - (i-j)c')! \ \textit{otherwise.} \end{aligned}$$

Note that for $k = 3$ the theorem gives, in matrix form, results identical with those derived by Ozols (1956), and for $k = 2$ the corollary gives the familiar distribution function of Smirnov's $D^+(n, n)$.

3. Preliminary lemma. The difficulty of proving the theorem is isolated in the following lemma.

LEMMA. *Let* n_j *and* c_j *be positive integers*, $j = 1, 2, \cdots, k$, *and satisfying* $n_{j+1} > n_j - c_j$, $1 \leqq j < k$, *and let*

$$M = \begin{pmatrix} A & R \\ S & T \end{pmatrix}$$

represent a partitioning of the square matrix M *of dimensions* $n_k - c_k + k + 1$, *where* A *is the* $k \times k$ *matrix described in Section* 2, R *is the* $k \times (n_k - c_k + 1)$ *matrix with elements*

$$r_{ij} = x_{k+1,j}^{c_i+\cdots+c_k+j-1}/(c_i + \cdots + c_k + j - 1)!;$$

S *is the* $(n_k - c_k + 1) \times k$ *matrix with elements*

$$\begin{aligned} s_{ij} &= 0 & &\textit{for } j \neq k \\ &= n_k!/(n_k - c_k - i + 1)! \ \textit{for } j = k; \end{aligned}$$

T *is the* $(n_k - c_k + 1) \times (n_k - c_k + 1)$ *matrix with elements*

$$\begin{aligned} t_{ij} &= 0 & &\textit{for } j < i \\ &= x_{k+1,j}^{j-i-1}/(j-1-i)! & &\textit{if } j \geqq i. \end{aligned}$$

Then

$$\prod_{j=1}^k n_j! \int_{x_{j+1,n_j-c_j+1}}^1 dx_{j,n_j} \int_{x_{j+1,n_j-c_j}}^{x_{j,n_j}} dx_{j,n_j-1} \cdots \int_{x_{j+1,2}}^{x_{j,c_j+2}} dx_{j,c_j+1} \int_{x_{j+1,1}}^{x_{j,c_j+1}} dx_{j,c_j} \int_0^{x_{j,c_j}} dx_{j,c_j-1} \cdots \int_0^{x_{j,2}} dx_{j,1} = |M|. \tag{3.1}$$

Note that the "product" notation in (3.1) is used to represent a multiple integral, where the variable of integration for $j = m + 1$ is the lower limit of the integral for $j = m$.

PROOF. Using induction, (3.1) will first be proved for $k = 1$. Integrating with respect to $x_{1,1}, x_{1,2}, \cdots, x_{1,c_1-1}$ gives $x_{1,c_1}^{c_1-1}/(c_1 - 1)!$ for an integrand. Integrating with respect to $x_{1,c_1}, x_{1,c_1+1}, \cdots, x_{1,n_1}$ gives integrands of

$$\begin{vmatrix} \frac{x_{1,c_1+1}^{c_1}}{c_1!} & \frac{x_{2,1}^{c_1}}{c_1!} \\ 1 & 1 \end{vmatrix}, \quad \begin{vmatrix} \frac{x_{1,c_1+2}^{c_1+1}}{(c_1+1)!} & \frac{x_{2,1}^{c_1}}{c_1!} & \frac{x_{2,2}^{c_1+1}}{(c_1+1)!} \\ \frac{x_{1,c_1+2}^{1}}{1!} & 1 & \frac{x_{2,2}}{1!} \\ 1 & 0 & 1 \end{vmatrix}, \cdots, |M|/n_1!.$$

This is easily seen if each integration is performed by expanding the integrand using the first column and the cofactors of the terms in the first column. The result of the integration is seen to be the next determinant in the above indicated sequence, expanded about its first and last columns. The last integral, with respect to x_{1,n_1}, when multiplied by $n_1!$, gives $|M|$ for $k = 1$.

If (3.1) is true for $k = m - 1$, then (3.1) is also true for $k = m$, as will be shown. Integration with respect to $x_{m,1}$ through x_{m,c_m-1} is simple if, each time, the integrand is expanded about its mth column, the integration is performed, and the result is reassembled into determinant form. Each time, the new mth column may be subtracted from the $(m + 1)$st column, leaving unity in row $m + 1$ of column $m + 1$, but zeros in the rest of column $m + 1$. Expanding about column $m + 1$ merely removes row $m + 1$ and column $m + 1$ from the determinant. Then the next variable of integration appears only in column m and the procedure is repeated. Note that if $c_m - 1 > n_{m-1} - c_{m-1}$, all but m columns have been removed on the $(n_{m-1} - c_{m-1})$th integration, so the aforementioned removal of row $m + 1$ is a vacuous operation for the last $c_m - 1 - n_{m-1} + c_{m-1}$ integrations.

If $c_m \leqq n_{m-1} - c_{m-1}$, the remaining integrations may be described by two procedures. In the first procedure, which applies to the integration of $x_{m,j}$ for $c_m \leqq j \leqq n_{m-1} - c_{m-1}$, the integration is performed by expanding the integrand along column m, integrating, and reassembling the integrand into determinant form, which resembles the former integrand except in the following ways: Column m, formerly the transpose of

$$(x_{m,j}^{c_1+\cdots+c_{m-1}+j-1}/(c_1 + \cdots + c_{m-1} + j - 1)!,$$
$$x_{m,j}^{c_2+\cdots+c_{m-1}+j-1}/(c_2 + \cdots + c_{m-1} + j - 1)!, \cdots,$$
$$x_{m,j}^{c_{m-1}+j-1}/(c_{m-1} + j - 1)!, x_{m,j}^{j-1}/(j - 1)!, 0, \cdots, 0)$$

now becomes the transpose of

$$(x_{m,j+1}^{c_1+\cdots+c_{m-1}+j}/(c_1 + \cdots + c_{m-1} + j)!, x_{m,j+1}^{c_2+\cdots+c_{m-1}+j}/(c_2 + \cdots + c_{m-1} + j)!,$$
$$\cdots, x_{m,j+1}^{c_{m-1}+j}/(c_{m-1} + j)!, x_{m,j+1}^{j}/(j)!, 0, \cdots, 0)$$

and a new column and new row are added to the right and bottom of the former integrand. The new column is the transpose of

$$(x_{m+1,j+1-c_m}^{c_1+\cdots+c_{m-1}+j}/(c_1 + \cdots + c_{m-1} + j)!, x_{m+1,j+1-c_m}^{c_2+\cdots+c_{m-1}+j}/(c_2 + \cdots + c_{m-1} + j)!,$$
$$\cdots, x_{m+1,j+1-c_m}^{c_{m-1}+j}/(c_{m-1} + j)!, x_{m+1,j+1-c_m}^{j}/(j)!, 0, 0, \cdots, 0, 1)$$

the last element being part of the new row, and the new row is all zeros except for unity in column m and in the new column, as just noted. This method of integration is merely a modification of the integration procedure described in the first part of this proof, for $k = 1$. The integration is followed by removing row and column $m + 1$ as previously described, that is, by subtracting column m from column $m + 1$ and expanding about column $m + 1$.

The second procedure describes the integration of $x_{m,j}$ for $n_{m-1} - c_{m-1} < j \leqq n_m$. It is identical with the first procedure, except that after the integration is performed, with the new row and new column added, the procedure is completed. That is, no row or column elimination is possible for the new integrand thus formed. After the final integration, with respect to x_{m,n_m}, multiplication by $n_m!$ gives $|M|$ for $k = m$.

If $c_m > n_{m-1} - c_{m-1}$, each successive integration is performed using the second procedure, as described above. Thus the lemma is proved.

4. Proof of the theorem. Let $Y_{j,1} < Y_{j,2} < \cdots < Y_{j,n_j}$ represent ordering of the random sample $X_{j,1}, X_{j,2}, \cdots, X_{j,n_j}$ for $j = 1, 2, \cdots, k$. Then

$$\begin{aligned} P_k^* &= P(Y_{j,c'_j} > Y_{j+1,1}, Y_{j,c'_j+1} > Y_{j+1,2}, \cdots, Y_{j,n_j} > Y_{j+1,n_j-c'_j+1} \\ &\qquad \text{for } j \leqq k-1) \quad \text{if} \quad n_{j+1} > n_j - c_j \quad \text{for all} \quad j \leqq k-1 \\ &= 0 \quad \text{if} \quad n_{j+1} \leqq n_j - c_j \quad \text{for some} \quad j \leqq k-1. \end{aligned} \tag{4.1}$$

P_k^* may be evaluated by integrating the joint density function of the Y_{ij}, $\prod_{j=1}^{k} n_j! \, dF(y_{j,1}) \, dF(y_{j,2}) \cdots dF(y_{j,n_j})$, over the proper limits. The proper limits are given by the left side of (3.1), after substituting $x_{i,j} = F(y_{i,j})$ for $i \leqq k$, and $x_{k+1,j} = 0$, for all j. Letting $x_{k+1,j} = 0$, for all j in the right side of (3.1) leaves only $|A|$, proving the theorem.

The corollary is proved by noting that $F_{i,n}(x) = I_i(x)/n$.

5. Asymptotic properties. As $n_j \to \infty$ and $c_j \to \infty$ for all j in such a way that $n_i/n_j \to r_{ij}$ and $c_j/n_j^{\frac{1}{2}} \to \lambda_j$, where r_{ij} and λ_j are constants, $1 \leqq (i, j) \leqq k$, then use of Stirling's formula and some algebra reveals the approximation

$$(5.1) \qquad a_{ij} \cong e^{-d_{ij}^2} n_j^{n_j^{\frac{1}{2}} d_{ij}} \qquad (5.1)$$

where $d_{ij} = \sum_{\alpha=1}^{i-1} \lambda_\alpha r_{\alpha,j} - \sum_{\beta=1}^{j-1} \lambda_\beta r_{\beta,j}$.

If $n_j = n$ for all j, the determinant of A may be simplified by multiplying the mth row by $n^{-\lambda_m n^{\frac{1}{2}}}$ and multiplying the mth column by $n^{\lambda_m n^{\frac{1}{2}}}$, for $m = 1, 2, \cdots, k$, which does not change the value of the determinant. Then $|A| \to |D|$, where D is the matrix whose (i, j)th element is $\exp\{-(\sum_{\alpha=1}^{i-1} \lambda_\alpha - \sum_{\beta=1}^{j-1} \lambda_\beta)^2\}$.

The asymptotic distribution function further simplifies if $c_j = c$ for all j. Then

$$\lim_{n\to\infty, c/n^{\frac{1}{2}} \to \lambda} P(\sup_{x, i<k} (F_{i,n}(x) - F_{i+1,n}(x) < \lambda/n^{\frac{1}{2}}) = |V^{k \times k}| \tag{5.2}$$

where the (i, j)th element of V is $\exp\{-(i - j)^2\lambda^2\}$.

From (5.2) it is seen that $\sup_{x,i<k} (F_{i,n}(x) - F_{i+1,n}(x))$ converges in prob-

ability to zero. However, if $F_i(x) - F_{i+1}(x) > 0$ for some x and some i, then there is no longer convergence in probability to zero. In fact, since $\sup_x |F_{i,n}(x) - F_i(x)|$ and $\sup_x |F_{i+1,n}(x) - F_{i+1}(x)|$ converge in probability to zero, use of the above statistic furnishes a test of H_0 that is consistent against all alternatives of the type just described. If an additional *a priori* assumption $F_1(x) \geqq F_2(x) \geqq \cdots \geqq F_k(x)$ is made, then the above statistic provides a test consistent against all alternatives of the type $F_i(x) \neq F_j(x)$ for some i, j, and x.

REFERENCES

Birnbaum, Z. W. and Hall, R. A. (1960). Small sample distributions for multisample statistics of the Smirnov type. *Ann. Math. Statist.* **31** 710–720.

Conover, W. J. (1965). Several k-sample Kolmogorov-Smirnov tests. *Ann. Math. Statist.* **36** 1019–1026.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramér-von Mises tests. *Ann. Math. Statist.* **28** 823–838.

David, H. T. (1958). A three-sample Kolmogorov-Smirnov test. *Ann. Math. Statist.* **29** 842–851.

Dwass, M. (1960). Some k-sample rank order tests. *Contributions to Probability and Statistics —Essays in Honor of Harold Hotelling.* Stanford Univ. Press.

Kiefer, J. (1959). k-sample analogues of the Kolmogorov-Smirnov and Cramér-v. Mises tests. *Ann. Math. Statist.* **30** 420–447.

Ozols, V. (1956). Generalization of the theorem of Gnedenko-Korolyuk to three samples in the case of two one-sided boundaries. *Latvijas PSR Zinātnu Akad. Vēstis,* No. 10 (**111**), 141–151.