A k-SAMPLE MODEL IN ORDER STATISTICS[1]

BY W. J. CONOVER

Kansas State University

1. Introduction and summary. Presented here is a new k-sample model in order statistics, in which k random samples of equal size are first ordered within themselves in the usual manner, and are then ordered among themselves by considering the size of the maximum value in each sample. The distribution functions and several probabilities relevant to the model are derived.

Consider k random samples of size n $(X_{11}, X_{21}, \cdots, X_{n1})$, $(X_{12}, X_{22}, \cdots, X_{n2})$, $\cdots$, $(X_{1k}, X_{2k}, \cdots, X_{nk})$, where the X_{ir} are independent and identically distributed according to the absolutely continuous distribution function $F(x)$. Arrange each sample in decreasing order, and let Z_{1r} be the greatest random variable in the sample $(X_{1r}, X_{2r}, \cdots, X_{nr})$, let Z_{2r} be the second greatest random variable in the same sample, and so on. The k samples, after ordering, are then $(Z_{11}, Z_{21}, \cdots, Z_{n1})$, $(Z_{12}, Z_{22}, \cdots, Z_{n2})$, $\cdots$, $(Z_{1k}, Z_{2k}, \cdots, Z_{nk})$. Form the set $S = \{Z_{11}, Z_{12}, \cdots, Z_{1k}\}$, so that the elements of S are the greatest random variables in each of the k samples. Order the random variables in S, and let Y_{11} denote the greatest, Y_{12} the second greatest, and so on to Y_{1k}. Then for each point in the sample space, Z_{1r} corresponds to some Y_{1j}. Define Y_{ij} as the ith ranked random variable from the same sample as the Z_{1r} mentioned above. In other words, for each point in the sample space where Z_{1r} corresponds to Y_{1j}, the sample $(Z_{1r}, Z_{2r}, \cdots, Z_{nr})$ will be denoted by $(Y_{1j}, Y_{2j}, \cdots, Y_{nj})$. Since $Z_{1r} > Z_{2r} > \cdots > Z_{nr}$, it follows that $Y_{1j} > Y_{2j} > \cdots > Y_{nj}$. Y_{ij} is called the ith order statistic in the jth sample. The number i is called the rank of Y_{ij} within the sample, and the number j is called the rank of the sample.

The above model is useful in flood frequency analysis, when it is desired to combine flood records at several independent stations in an attempt to study rare floods. Also the above model can be applied in the following situation. A large number (nk) of displays have been entered in a science fair, at which k prizes are to be awarded. The judge feels it is impractical to consider all nk displays at once, and so he randomly divides the displays into k groups of n elements in each group. He then judges each group separately, ordering the n displays in each group according to excellence. When each group has thus been ordered, he then considers the k best displays, one from each group, and awards the 1st prize, 2nd prize, $\cdots$, kth prize to these k displays according to their relative excellence. Y_{ij} represents the display that was ith best in its group where the best display in that group won jth prize.

Received 9 June 1964; revised 1 October 1964.

[1] This paper is part of a dissertation submitted to the Catholic University of America in partial fulfillment of the requirements for the Ph.D. degree. This research was performed with the support of the National Institutes of Health training grant NIH 5T1-GM-498, formerly 2G-498. Contribution No. 83, Department of Statistics, and Statistical Laboratory, Kansas Agricultural Experiment Station, Manhattan.

The distribution function of Y_{ij} is found in Section 2. A method of comparing Y_{ij} with X, an additional random variable with the same distribution function $F(x)$, is given by the equation for $P(X > Y_{ij})$, derived in Section 3. Section 4 contains $P(Y_{i_1 j_1} > Y_{i_2 j_2})$ where $Y_{i_1 j_1}$ and $Y_{i_2 j_2}$ are two ordered random variables from the same array of nk random variables. The results of Section 4 permit the comparison of any two of the ordered random variables. The comparison of $Y_{1,k}$ with $Y_{2,1}$ is related to the probability of $Y_{1,k}$ exceeding $\max_j Y_{2,j}$ which was considered by Cohn, Mosteller, Pratt, and Tatsuoka (1960).

In Section 5 the expected value of Y_{ij} is discussed. The fairness of the above method of awarding prizes is examined in Section 6.

2. The distribution function of Y_{ij}. The following results are well known in order statistics, and can be found in Fisz (1963), Sarhan and Greenberg (1962), and Wilks (1962). Let Z_i represent the ith order statistic in the single sample, and let $F_i(x) = P(Z_i < x)$. Assume $i_1 < i_2$. Then $Z_{i_1} > Z_{i_2}$ and, for absolutely continuous $F(x)$,

$$F_i(x) = \sum_{m=0}^{i-1} \binom{n}{m} [F(x)]^{n-m} [1 - F(x)]^m \tag{2.1}$$

$$F_1(x) = P(Z_1 < x) = F^n(x) \tag{2.2}$$

$$P(x < Z_i < x + dx) = i\binom{n}{i} [1 - F(x)]^{i-1} [F(x)]^{n-i}\, dF(x) \tag{2.3}$$

$$\begin{aligned} P(u < Z_{i_1} < u + du, v < Z_{i_2} < v + dv) &= [n!/(i_1 - 1)!(i_2 - i_1 - 1)!(n - i_2)!][1 - F(u)]^{i_1 - 1} \\ &\quad \cdot [F(u) - F(v)]^{i_2 - i_1 - 1} [F(v)]^{n - i_2}\, dF(u)\, dF(v) \quad \text{if} \quad v < u \\ &= 0 \quad \text{if} \quad v \geqq u \end{aligned} \tag{2.4}$$

$$\begin{aligned} P(u < Z_1 < u + du, v < Z_i < v + dv) &= [n!/(i - 2)!(n - i)!][F(u) - F(v)]^{i-2} [F(v)]^{n-i}\, dF(u)\, dF(v) \quad \text{if} \quad v < u \\ &= 0 \quad \text{if} \quad v \geqq u. \end{aligned} \tag{2.5}$$

It is easy to show that the joint distribution function of Z_1 and Z_i for $i > 1$ is given by

$$\begin{aligned} G_i(t, x) = P(Z_1 < t, Z_i < x) &= [F(t)]^n; \quad \text{if} \quad t \leqq x \\ &= \sum_{m=0}^{i-1} \binom{n}{m} [F(x)]^{n-m} [F(t) - F(x)]^m; \\ &\qquad \text{if} \quad x < t. \end{aligned} \tag{2.6}$$

LEMMA 1. *The distribution of Y_{1j} is given by*

$$F_{1j}(x) = P(Y_{1j} < x) = \sum_{\alpha=0}^{j-1} \binom{k}{\alpha} [F^n(x)]^{k-\alpha} [1 - F^n(x)]^\alpha \tag{2.7}$$

and the probability element is given by

$$P(x < Y_{1j} < x + dx) = j\binom{k}{j} [1 - F^n(x)]^{j-1} [F^n(x)]^{k-j} n [F(x)]^{n-1}\, dF(x). \tag{2.8}$$

PROOF. Consider again the set S consisting of $Z_{11}, Z_{12}, \cdots, Z_{1k}$. The set S

then can be regarded as a sample of size k, where the unranked element in S has the distribution function $F^n(x)$.

The set S is ordered the same way a sample is usually ordered and the random variable Y_{1j} corresponds to the random variable with rank j from a sample of size k. The distribution function of Y_{1j} and the probability element associated with Y_{1j} can then be obtained from (2.1) and (2.3) respectively. Instead of sample size n, we now use k, the rank is now j instead of i, and the population distribution function is now $F^n(x)$ instead of $F(x)$. Substitution of these values into (2.1) and (2.3) results directly in (2.7) and (2.8).

LEMMA 2. *Let j be a positive integer and k be a real number such that $k \geqq j$. Assume that $a \geqq 0$ and $n > 0$. Then*

$$(2.9)\quad \int_0^a j\binom{k}{j} n(1-w^n)^{j-1} w^{nk-nj+n-1}\,dw = \sum_{\alpha=0}^{j-1} [(k)_j(-1)^{j-1-\alpha}(a^n)^{k-\alpha}]/[\alpha!(j-1-\alpha)!(k-\alpha)] = \sum_{\alpha=0}^{j-1}\binom{k}{\alpha}(1-a^n)^{\alpha}(a^n)^{k-\alpha}$$

where $(k)_j$ represents the product $k(k-1)\cdots(k-j+1)$.

PROOF. To arrive at the first summation, the expansion

$$(1-w^n)^{j-1} = \sum_{\alpha=0}^{j-1}\binom{j-1}{\alpha}(w^n)^{j-1-\alpha}(-1)^{j-1-\alpha}, \qquad (j = 1, 2, 3, \cdots)$$

is substituted into the integrand of (2.9).

$$\begin{aligned}
&\int_0^a j\binom{k}{j} n(1-w^n)^{j-1} w^{nk-nj+n-1}\,dw \\
&\quad = \sum_{\alpha=0}^{j-1} j\binom{k}{j}\binom{j-1}{\alpha} n(-1)^{j-1-\alpha}\int_0^a w^{nk-n\alpha-1}\,dw \\
&\quad = \sum_{\alpha=0}^{j-1} [(k)_j(-1)^{j-1-\alpha} n/\alpha!(j-1-\alpha)!]\cdot[w^{nk-n\alpha}/(nk-n\alpha)]_0^a \\
&\quad = \sum_{\alpha=0}^{j-1} [(k)_j(-1)^{j-1-\alpha}(a^n)^{k-\alpha}/\alpha!(j-1-\alpha)!(k-\alpha)], \\
&\qquad\qquad (j = 1, 2, \cdots; n > 0)
\end{aligned}$$

This proves the first part of the lemma.

The second summation is obtained by using induction. For $j = 1$,

$$\int_0^a knw^{nk-1}\,dw = (a^n)^k, \qquad (n > 0, k \geqq 1)$$

is easily seen to be true. Assuming the relation to be true for j, and integrating by parts, we have for $(j+1)$,

$$\begin{aligned}
&\int_0^a (j+1)\binom{k}{j+1} n(1-w^n)^j w^{nk-n(j+1)+n-1}\,dw \\
&\quad = [(k)_{j+1}(1-w^n)^j nw^{nk-nj}/j!(nk-nj)]_0^a \\
&\quad + \int_0^a [(k)_{j+1} n^2 j(1-w^n)^{j-1} w^{nk-nj+n-1}/j!(nk-nj)]\,dw \\
&\quad = \binom{k}{j}(1-a^n)^j(a^n)^{k-j} + \int_0^a j\binom{k}{j} n(1-w^n)^{j-1} w^{nk-nj+n-1}\,dw \\
&\quad = \binom{k}{j}(1-a^n)^j(a^n)^{k-j} + \sum_{\alpha=0}^{j-1}\binom{k}{\alpha}(1-a^n)^{\alpha}(a^n)^{k-\alpha} \\
&\quad = \sum_{\alpha=0}^{j}\binom{k}{\alpha}(1-a^n)^{\alpha}(a^n)^{k-\alpha}, \qquad (n > 0, k \geqq j+1).
\end{aligned}$$

This is the desired relation.

THEOREM 1. *The distribution function of Y_{ij} is given by*

$$F_{ij}(x) = P(Y_{ij} < x) = \sum_{\alpha=0}^{j-1} \binom{k}{\alpha}[1 - F^n(x)]^\alpha [F^n(x)]^{k-\alpha}$$

$$(2.10) \qquad + \sum_{\alpha=0}^{j-1} \sum_{m=1}^{i-1} \sum_{\beta=0}^{m-1} j\binom{k}{j} m \binom{n}{m} \binom{j-1}{\alpha} \binom{m-1}{\beta} (-1)^{j+m-\beta-\alpha}$$

$$\cdot ([F(x)]^{n-1-\beta} - [F(x)]^{nk-n\alpha})/(nk - n\alpha + 1 - n + \beta)$$

where the triple summation in (2.10) is zero for $i = 1$.

PROOF. The case of $i = 1$ has already been proved in Lemma 1. For $i > 1$, consider again the k samples and also the ordered set S, consisting of $Y_{11}, Y_{12}, \cdots, Y_{1k}$. Let Y_{1r} be the random variable representing an unspecified element of S. Then Y_{ir} is the k sample order statistic with rank i, from the same unspecified sample as Y_{1r}.

The distribution function of Y_{1r}, for unspecified r, is the same as that for Z_1, given by (2.2), so that

$$(2.11) \qquad P(t < Y_{1r} < t + dt) = n[F(t)]^{n-1}\, dF(t).$$

Similarly, the joint distribution function of Y_{1r} and Y_{ir}, for unspecified r, and $i > 1$, is given by (2.6), from which it follows that

$$(2.12) \quad P(t < Y_{1r} < t + dt,\ Y_{ir} < x)$$

$$= G_i(t + dt, x) - G_i(t, x) = (\partial/\partial t)G_i(t, x)\, dt.$$

It is important to recall that the rank of the group containing Y_{ir} depends only on Y_{1r}, and is indpendent of whatever values the other random variables in the sample may assume. Mathematically stated,

$$(2.13) \quad P(r = j \mid t < Y_{1r} < t + dt,\ Y_{ir} < x) = P(r = j \mid t < Y_{1r} < t + dt).$$

Since $P(A \mid B) = P(AB)/P(B)$, each side of (2.13) can be rewritten to become

$$P(r = j,\ t < Y_{1r} < t + dt,\ Y_{ir} < x)/P(t < Y_{1r} < t + dt,\ Y_{ir} < x)$$

$$= P(r = j,\ t < Y_{1r} < t + dt)/P(t < Y_{1r} < t + dt)$$

which is equivalent to

$$(2.14) \quad P(t < Y_{1j} < t + dt,\ Y_{1j} < x)/P(t < Y_{1r} < t + dt,\ Y_{ir} < x)$$

$$= P(t < Y_{1j} < t + dt)/P(t < Y_{1r} < t + dt)$$

where j represents a specified value of the unspecified subscript r. Substituting (2.8), (2.11) and (2.12) into (2.14), and rearranging terms we obtain

$$(2.15) \quad P(t < Y_{1j} < t + dt,\ Y_{ij} < x)$$

$$= j\binom{k}{j}[1 - F^n(t)]^{j-1}[F^n(t)]^{k-j}(\partial/\partial t)G_i(t, x)\, dt.$$

To find the marginal distribution function of Y_{ij}, (2.15) is integrated over all

values of t. Then, using (2.6), we have

$$P(Y_{ij} < x) = \int_{-\infty}^{\infty} j\binom{k}{j}[1 - F^n(t)]^{j-1}[F^n(t)]^{k-j}(\partial/\partial t)G_i(t, x)\, dt, \qquad (i > 1)$$

$$= \int_{-\infty}^{x} j\binom{k}{j}[1 - F^n(t)]^{j-1}[F^n(t)]^{k-j}n[F(t)]^{n-1}\, dF(t) \tag{2.16}$$
$$+ \int_{x}^{\infty} j\binom{k}{j}[1 - F^n(t)]^{j-1}[F^n(t)]^{k-j}$$
$$\cdot \sum_{m=0}^{i-1} \binom{n}{m}[F(x)]^{n-m}m[F(t) - F(x)]^{m-1}\, dF(t).$$

Since the second integral in (2.16) is being multiplied by the factor m, the integral disappears when $m = 0$. Therefore one of the values of the index m can be eliminated. By changing the order of operations and rearranging terms, (2.16) can be written as

$$P(Y_{ij} < x) = \int_{-\infty}^{x} j\binom{k}{j}n[1 - F^n(t)]^{j-1}[F^n(t)]^{k-j}[F(t)]^{n-1}\, dF(t)$$
$$+ \sum_{m=1}^{i-1} \int_{x}^{\infty} j\binom{k}{j}m\binom{n}{m}[F(x)]^{n-m}[F^n(t)]^{k-j}[1 - F^n(t)]^{j-1} \tag{2.17}$$
$$\cdot [F(t) - F(x)]^{m-1}\, dF(t).$$

Since $F(-\infty) = 0$, Lemma 2 can be used to evaluate the first integral in (2.17) by letting $w = F(t)$ and $a = F(x)$. Then the first integral in (2.17) becomes

$$\sum_{\alpha=0}^{j-1} \binom{k}{\alpha}[1 - F^n(x)]^{\alpha}[F^n(x)]^{k-\alpha}. \tag{2.18}$$

Using the expansions

$$[1 - F^n(t)]^{j-1} = \sum_{\alpha=0}^{j-1} \binom{j-1}{\alpha}[F^n(t)]^{j-1-\alpha}(-1)^{j-1-\alpha}$$

and

$$[F(t) - F(x)]^{m-1} = \sum_{\beta=0}^{m-1} \binom{m-1}{\beta}[F(t)]^{\beta}[F(x)]^{m-1-\beta}(-1)^{m-1-\beta},$$

the second integral in (2.17) becomes

$$\sum_{m=1}^{i-1} \sum_{\alpha=0}^{j-1} \sum_{\beta=0}^{m-1} j\binom{k}{j}m\binom{n}{m}\binom{j-1}{\alpha}\binom{m-1}{\beta}(-1)^{j+m-\beta-\alpha-2}$$
$$\cdot [F(x)]^{n-1-\beta}\int_{x}^{\infty}[F(t)]^{nk-n-n\alpha+\beta}\, dF(t) \tag{2.19}$$
$$= \sum_{m=1}^{i-1} \sum_{\alpha=0}^{j-1} \sum_{\beta=0}^{m-1} j\binom{k}{j}m\binom{n}{m}\binom{j-1}{\alpha}\binom{m-1}{\beta}(-1)^{j+m-\beta-\alpha-2}$$
$$\cdot ([F(x)]^{n-1-\beta} - [F(x)]^{nk-n\alpha})/(nk - n - n\alpha + \beta + 1).$$

(2.18) and (2.19) combine to give (2.10), and the proof is complete.

3. $P(X > Y_{ij})$. In an effort to keep the proof to the following theorem as simple as possible, two lemmas will be introduced at this point.

LEMMA 3. *Let j be a positive integer and let k be a real number such that $k \geqq j$. Then*

$$j\binom{k}{j}\sum_{\alpha=0}^{j-1}[\binom{j-1}{\alpha}(-1)^{j-1-\alpha}/(k - \alpha)] = 1 \tag{3.1}$$

and

$$(3.2)\qquad \sum_{\alpha=0}^{j-1}[(-1)^{j-1-\alpha}/(\alpha!(j-1-\alpha)!(k-\alpha))] = 1/(k)_j.$$

Also, for $k + (1/n) \geqq j$, $n > 0$,

$$(3.3)\qquad \sum_{\alpha=0}^{j-1}[(-1)^{j-1-\alpha}/\alpha!(j-1-\alpha)!(nk-n\alpha+1)] = (1/n)[1/(k+(1/n))_j].$$

PROOF. Let $a = 1$ in Lemma 2. Then the first summation in (2.9) becomes the left side of (3.1), and the second summation in (2.9) becomes unity, completing (3.1). A rearrangement of terms in (3.1) leads to (3.2).

If n is factored from the denominator, and if the substitution

$$(3.4)\qquad k' = k + (1/n)$$

is used, the left side of (3.3) can be written as

$$(3.5)\qquad (1/n)\sum_{\alpha=0}^{j-1}[(-1)^{j-1-\alpha}/\alpha!(j-1-\alpha)!(k'-\alpha)].$$

The use of (3.2) and then (3.4) in (3.5) results in

$$(1/n)(1/(k')_j = (1/n)[1/(k+(1/n))_j],$$

which completes the proof of Lemma 3.

LEMMA 4. *For positive integer* j, *real number* $k \geqq j$, $n > 0$, *and* $a \geqq 0$,

$$(3.6)\qquad \int_0^1 \sum_{\alpha=0}^{j-1}\binom{k}{\alpha}(1-a^n)^\alpha(a^n)^{k-\alpha}\,da = 1 - [(k)_j/(k+(1/n))_j].$$

PROOF. Integrating both sums in (2.9) with respect to a gives

$$(3.7)\qquad \int_0^1 \sum_{\alpha=0}^{j-1}\binom{k}{\alpha}(1-a^n)^\alpha(a^n)^{k-\alpha}\,da = \int_0^1 \sum_{\alpha=0}^{j-1}[(k)_j(-1)^{j-1-\alpha}(a^n)^{k-\alpha}/\alpha!(j-1-\alpha)!(k-\alpha)]\,da.$$

The left side of (3.7) is seen to be the left side of (3.6). Term by term integration by parts applied to the right side of (3.7) results in

$$(3.8)\qquad (k)_j\sum_{\alpha=0}^{j-1}[(-1)^{j-1-\alpha}/\alpha!(j-1-\alpha)!(k-\alpha)] - n(k)_j\sum_{\alpha=0}^{j-1}[(-1)^{j-1-\alpha}/\alpha!(j-1-\alpha)!(nk-n\alpha+1)].$$

The first and second summations in (3.8) can be simplified by using (3.2) and (3.3), respectively, which gives the right side of (3.6) and completes the proof.

THEOREM 2. *The probability of an additional random variable* X, *with the distribution function* $F(x)$, *exceeding* Y_{ij} *is given by*

$$(3.9)\qquad P(X > Y_{ij}) = 1 - [1 - ((i-1)/n)][(k)_j/(k+(1/n))_j].$$

PROOF. The distribution function of X is $F(x)$, and the distribution function of Y_{ij} is given by (2.10). Then

$$P(X > Y_{ij}) = \int_{-\infty}^{\infty} F_{ij}(x)\,dF(x)$$

$$(3.10)\qquad = \int_{-\infty}^{\infty}\sum_{\alpha=0}^{j-1}\binom{k}{\alpha}[1-F^n(x)]^\alpha[F^n(x)]^{k-\alpha}\,dF(x) + \int_{-\infty}^{\infty}\sum_{\alpha=0}^{j-1}\sum_{m=1}^{i-1}\sum_{\beta=0}^{m-1} j\binom{k}{j}m\binom{n}{m}\binom{j-1}{\alpha}\binom{m-1}{\beta}(-1)^{j+m-\beta-\alpha} \cdot([F(x)]^{n-1-\beta} - [F(x)]^{nk-n\alpha})/(nk-n\alpha+1-n+\beta)\,dF(x).$$

The first integral in (3.10) is given by Lemma 4. Since the second integral disappears for $i = 1$, we have

$$P(X > Y_{1j}) = 1 - [(k)_j/(k + (1/n))_j]. \tag{3.11}$$

Integrating term by term and rearranging terms the second integral of (3.10) becomes

$$j\binom{k}{j} \sum_{\alpha=0}^{j-1} [\binom{j-1}{\alpha}(-1)^{j-1-\alpha}/(nk - n\alpha + 1)] \cdot \sum_{m=1}^{i-1} m\binom{n}{m} \sum_{\beta=0}^{m-1} [\binom{m-1}{\beta}(-1)^{m-1-\beta}/(n - \beta)]. \tag{3.12}$$

The use of (3.1) where $m = j$, $n = k$, and $\beta = \alpha$, reduces (3.12) to

$$(i - 1)j\binom{k}{j} \sum_{\alpha=0}^{j-1} \binom{j-1}{\alpha}(-1)^{j-1-\alpha}/(nk - n\alpha + 1)$$

which using (3.3) reduces to

$$(i - 1)(k)_j/n(k + (1/n))_j . \tag{3.13}$$

The substitution of (3.11) and (3.13) into (3.10) for the first and second integrals respectively, gives (3.9) and Theorem 2 is proved.

The following is an immediate result of Theorem 2.

COROLLARY.

$$P(Y_{i+1,j} < X < Y_{ij}) = (1/n)P(X < Y_{1j}), \qquad (i = 1, \cdots, n), \tag{3.14}$$

where $Y_{n+1,j}$ *is defined as* $-\infty$.

Therefore the probability of an additional random value falling between two adjacent members of the same sample is independent of the ranks of those members within the sample.

4. $P(Y_{i_1j_1} > Y_{i_2j_2})$. Since $Y_{i_1j_1}$ and $Y_{i_2j_2}$ are ordered random variables, they cannot be considered independent. For example, if $j_1 = j_2$, the two random variables come from the same sample, and we can say with probability 1 that $Y_{i_1j_1} < Y_{i_2j_1}$ if and only if $i_2 < i_1$. If $j_1 \neq j_2$, assume without loss of generality that $j_1 < j_2$. Then $Y_{i_1j_1}$ and $Y_{i_2j_2}$ have not been directly compared with each other, but have been compared with a third random variable namely Y_{1j_1}. It is then known that $Y_{i_1j_1}$ and $Y_{i_2j_2}$ are both less than Y_{1j_1}.

THEOREM 3. *When comparing sample values from the same collection of nk values, for* $j_1 < j_2$ *and* $i_1 \neq 1$,

$$P(Y_{i_1j_1} > Y_{i_2j_2}) = 1 - \binom{n-1}{i_2-1}\binom{n-2}{i_1-2}\binom{k-j_1}{j_2-j_1} \cdot \sum_{\alpha=0}^{i_1-2} [(n - 1)/(n - 1 - \alpha)]\binom{i_1-2}{\alpha}(-1)^{i_1-2-\alpha}/\binom{k-j_1+1-((\alpha+1)/n)}{j_2-j_1}\binom{2n-\alpha-2}{i_2-1}. \tag{4.1}$$

PROOF. Consider again the set S, consisting of $Y_{11}, Y_{12}, \cdots, Y_{1k}$. Let Y_{1r_1} and Y_{1r_2} be two unspecified elements of S. Then $Y_{i_1r_1}$ and $Y_{i_2r_2}$ are the elements of rank i_1 and i_2 from the same unspecified samples as Y_{1r_1} and Y_{1r_2}, respectively.

Since the samples are independent before they are ranked, Y_{1r_1} and Y_{1r_2} are mutually independent. With the aid of (2.11), this implies

$$\begin{aligned} P(w < Y_{1r_1} < w + dw,\ y < Y_{1r_2} < y + dy) &= P(w < Y_{1r_1} < w + dw)P(y < Y_{1r_2} < y + dy) \\ &= n^2[F(w)]^{n-1}[F(y)]^{n-1}\, dF(w)\, dF(y). \end{aligned} \tag{4.2}$$

For the same reason, we have in the case where $i_2 > 1$, using (2.5) and (2.12),

$$
\begin{aligned}
&P(w < Y_{1r_1} < w + dw,\ t < Y_{i_1r_1} < t + dt,\ y < Y_{1r_2} < y + dy,\ Y_{i_2r_2} < t) \\
&\quad = P(w < Y_{1r_1} < w + dw,\ t < Y_{i_1r_1} < t + dt) \\
&\quad\quad \cdot P(y - Y_{1r_2} < y + dy,\ Y_{i_2r_2} < t) \\
&\quad = (n!/(i_1 - 2)!(n - i_1)!)[F(w) - F(t)]^{i_1-2}[F(t)]^{n-i_1}\, dF(w)\, dF(t) \\
&\quad\quad \cdot (\partial/\partial y)G_{i_2}(y, t)\, dy; \quad \text{if} \quad t < w \\
&\quad = 0; \quad \text{if} \quad t \geqq w
\end{aligned} \tag{4.3}
$$

To find the joint probability element of Y_{1j_1} and Y_{1j_2}, for specified values j_1 and j_2, the same argument introduced in the proof of Lemma 1 will be used. The set S can be regarded as an ordered sample of size k, with elements Y_{11}, Y_{12}, $\cdots$, Y_{1k}, having ranks, $1, 2, \cdots, k$, respectively, and the distribution function of the unspecified element of S is $F^n(x)$. The joint probability element of Y_{1j_1} and Y_{1j_2} can be obtained from (2.4). Instead of sample size n we now use k. The ranks are now j_1 instead of i_1, and j_2 instead of i_2. The population distribution function is now $F^n(x)$ instead of $F(x)$. Substitution of these values into (2.4), and using variables w and y instead of u and v, gives the following equation.

$$
\begin{aligned}
&P(w < Y_{1j_1} < w + dw,\ y < Y_{1j_2} < y + dy) \\
&\quad = [k!/((j_1 - 1)!(j_2 - j_1 - 1)!(k - j_2)!)][1 - F^n(w)]^{j_1-1} \\
&\quad\quad \cdot [F^n(w) - F^n(y)]^{j_2-j_1-1}[F^n(y)]^{k-j_2}n^2[F(w)]^{n-1}[F(y)]^{n-1}\, dF(w)\, dF(y).
\end{aligned} \tag{4.4}
$$

At this point it is necessary to recall that the sample rank depends only on the greatest element of the sample, and is independent of whatever values the other elements of the sample may assume. Mathematically stated, this becomes

$$
\begin{aligned}
&P(r_1 = j_1, r_2 = j_2 \mid w < Y_{1r_1} < w + dw, \\
&\quad\quad t < Y_{i_1r_1} < t + dt,\ y < Y_{1r_2} < y + dy,\ Y_{i_2r_2} < t) \\
&\quad = P(r_1 = j_1, r_2 = j_2 \mid w < Y_{1r_1} < w + dw,\ y < Y_{1r_2} < y + dy).
\end{aligned} \tag{4.5}
$$

But since $P(A \mid B) = P(AB)/P(B)$ both sides of (4.5) can be rewritten as

$$
\begin{aligned}
&\frac{P(r_1 = j_1, r_2 = j_2, w < Y_{1r_1} < w + dw, t < Y_{i_1r_1} < t + dt,\ y < Y_{1r_2} < y + dy,\ Y_{i_2r_2} < t)}{P(w < Y_{1r_1} < w + dw, t < Y_{i_1r_1} < t + dt,\ y < Y_{1r_2} < y + dy,\ Y_{i_2r_2} < t)} \\
&\quad = \frac{P(r_1 = j_1, r_2 = j_2, w < Y_{1r_1} < w + dw,\ y < Y_{1r_2} < y + dy)}{P(w < Y_{1r_1} < w + dw,\ y < Y_{1r_2} < y + dy)}
\end{aligned} \tag{4.6}
$$

which is equivalent to

$$
\begin{aligned}
&\frac{P(w < Y_{1j_1} < w + dw, t < Y_{i_1j_1} < t + dt, y < Y_{1j_2} < y + dy,\ Y_{i_2j_2} < t)}{P(w < Y_{1r_1} < w + dw, t < Y_{i_1r_1} < t + dt, y < Y_{1r_2} < y + dy, Y_{i_2r_2} < t)} \\
&\quad = \frac{P(w < Y_{1j_1} < w + dw,\ y < Y_{1j_2} < y + dy)}{P(w < Y_{1r_1} < w + dw,\ y < Y_{1r_2} < y + dy)}.
\end{aligned} \tag{4.7}
$$

Substituting (4.2), (4.3), and (4.4) into (4.7), and rearranging terms, gives us

$$
\begin{aligned}
&P(w < Y_{1j_1} < w + dw,\ t < Y_{i_1 j_1} < t + dt,\ y < Y_{1j_2} < y + dy,\ Y_{i_2 j_2} < t) \\
&\quad = \{k!n![1 - F^n(w)]^{j_1 - 1}[F^n(w) - F^n(y)]^{j_2 - j_1 - 1}/(j_1 - 1)!(j_2 - j_1 - 1)! \\
&\quad\quad \cdot (k - j_2)!(i_1 - 2)!(n - i_1)!\}[F^n(y)]^{k - j_2}[F(w) - F(t)]^{i_1 - 2} \\
&\quad\quad \cdot [F(t)]^{n - i_1}(\partial/\partial y)G_{i_2}(y, t)\, dy\, dF(w)\, dF(t); \qquad \text{if } t < w \\
&\quad = 0; \qquad \text{if } t \geqq w
\end{aligned} \tag{4.8}
$$

where $G_{i_2}(y, t)$ is obtained from (2.6), for $i_2 > 1$, and

$$
\begin{aligned}
G_1(y, t) &= F^n(y); \qquad y \leqq t \\
&= F^n(t); \qquad t < y.
\end{aligned}
$$

The remainder of this proof consists of applying elementary principles of integral calculus to the expression given in (4.8) to arrive at the desired probability given in (4.1). First, the random variable Y_{1j_2} is removed by integrating (4.8) over all values of y from t to w. Then, the random variable Y_{1j_1} is removed by integrating over all values of w from t to ∞. Finally, the random variables $Y_{i_1 j_1}$ and $Y_{i_2 j_2}$ are removed by integrating over all values of t from $-\infty$ to $+\infty$. The results are simplified using the identities of the previous section, until (4.1) is obtained. The details are arithmetic in nature and are omitted.

COROLLARY.

$$
P(Y_{2j_1} > Y_{1j_2}) = 1 - [(k - j_1)_{j_2 - j_1}/(k - j_1 + 1 - (1/n))_{j_2 - j_1}]. \tag{4.9}
$$

PROOF. Substitution of $i_2 = 1$, $i_1 = 2$ into (4.1) gives

$$
P(Y_{2j_1} > Y_{1j_2}) = 1 - [\tbinom{k - j_1}{j_2 - j_1}/\tbinom{k - j_1 + 1 - (1/n)}{j_2 - j_1}]
$$

which is equivalent to (4.9).

Cohn, Mosteller, Pratt and Tatsuoka (1960) found the probability of Y_{1k} exceeding $\max_j Y_{2j}$ in the general case where the k samples were drawn from different populations, and in the special case where the k samples were drawn from the same population.

COROLLARY.

$$
P(Y_{ij_1} > Y_{1,j_1+1}) = (n - 1)_{i-1}/(nk - nj_1 + n - 1)_{i-1}\,. \tag{4.10}
$$

PROOF. Substitution of $i_1 = i$, $i_2 = 1, j_2 = j_1 + 1$ in (4.1) gives

$$
\begin{aligned}
P(Y_{ij_1} > Y_{1,j_1+1}) &= 1 - \tbinom{n}{i-2}(k - j_1) \textstyle\sum_{\alpha=0}^{i-2} [n(n - 1)\tbinom{i-2}{\alpha}(-1)^{i-2-\alpha}/ \\
&\qquad (n - 1 - \alpha)(nk - nj_1 + n - 1 - \alpha)] \\
&= 1 - \tbinom{n-2}{i-2}(n - 1)(i - 2)! \textstyle\sum_{\alpha=0}^{i-2} [(-1)^{i-2-\alpha}/\alpha!(i - 2 - \alpha)!] \\
&\qquad \cdot (1/(n - 1 - \alpha) - 1/(nk - nj_1 + n - \alpha - 1)).
\end{aligned}
$$

Using (3.2), the result in (4.10) is obtained.

As a special case of the above corollary, let $j_1 = 1$. Then

$$
P(Y_{i,1} > Y_{1,2}) = (n - 1)_{i-1}/(nk - 1)_{i-1} \tag{4.11}
$$

which was first obtained by Mosteller (1948) as the probability function for the test statistic i in a nonparametric test for a k-sample slippage problem.

5. $E(Y_{ij})$. The theorems of the two preceding sections can be used to compare one ordered random variable with another in two different ways, both distribution free. A third method of comparison, depending on the underlying distribution $F(x)$, is by comparing expected values of the two random variables. To obtain $dF_{ij}(x)$, Lemma 1 can be used in conjunction with (2.10) resulting in

$$dF_{ij}(x) = j\binom{k}{j}[1-F^n(x)]^{j-1}[F^n(x)]^{k-j}n[F(x)[^{n-1}dF(x) + \sum_{\alpha=0}^{j-1}\sum_{m=1}^{i-1}\sum_{\beta=0}^{m-1}$$

$$\frac{k!\,n!(-1)^{j+m-\beta-\alpha}[(n-1-\beta)[F(x)]^{n-2-\beta} - (nk-n\alpha)[F(x)]^{nk-n\alpha-1}]}{(k-j)!(n-m)!\alpha!(j-1-\alpha)!\beta!(m-1-\beta)!\cdot(nk-n\alpha+1-n+\beta)}\,dF(x). \tag{5.1}$$

If $F(x)$ is the exponential distribution function given by

$$F(x) = 1 - e^{-x} \quad \text{if} \quad x \geqq 0$$
$$= 0 \quad \text{if} \quad x < 0,$$

then Gumbel (1954), p. 82, gives

$$\int_{-\infty}^{\infty} x\,dF^r(x) = \sum_{\alpha=1}^{r} 1/\alpha. \tag{5.2}$$

The use of (5.2) and the expansion of $[1-F^n(x)]^{j-1}$ in (5.1) yield

$$E(Y_{ij}) = \sum_{\alpha=0}^{j-1}\sum_{\beta=1}^{nk-n\alpha}\frac{k!(-1)^{j-1-\alpha}}{(k-j)!\alpha!(j-1-\alpha)!(k-\alpha)\beta} + \sum_{\alpha=0}^{j-1}\sum_{m=1}^{i-1}\sum_{\beta=0}^{m-1}$$

$$\frac{k!\,n!(-1)^{j+m-\beta-\alpha}\left[\sum_{\delta=1}^{n-1-\beta}\frac{1}{\delta} - \sum_{\nu=1}^{nk-n\alpha}\frac{1}{\nu}\right]}{(k-j)!(n-m)!\alpha!(j-1-\alpha)!\beta!(m-1-\beta)!(nk-n\alpha+1-n+\beta)} \tag{5.3}$$

when $F(x)$ is the exponential distribution function.

In the simpler case where $F(x)$ is the uniform distribution function given by

$$F(x) = 1 \quad \text{if} \quad x \geqq 1$$
$$= x \quad \text{if} \quad 0 \leqq x < 1$$
$$= 0 \quad \text{if} \quad x < 0,$$

the use of (5.1) gives

$$E(Y_{ij}) = \sum_{\alpha=0}^{j-1}[k!n(-1)^{j-1-\alpha}/(k-j)!\alpha!(j-1-\alpha)!(nk-n\alpha+1)]$$
$$-j\binom{k}{j}\sum_{\alpha=0}^{j-1}\left[\binom{j-1}{\alpha}(-1)^{j-1-\alpha}/(nk-n\alpha+1)\right] \tag{5.4}$$
$$\cdot\sum_{m=1}^{i-1}m\binom{n}{m}\sum_{\beta=0}^{m-1}\left[\binom{m-1}{\beta}(-1)^{m-1-\beta}/(n-\beta)\right].$$

The first summation in (5.4) can be simplified by using (3.3). The remaining

term in (5.4) is reduced in the same way that (3.12) was reduced to (3.13). Thus a comparison with (3.9) shows

$$(5.5)\quad E(Y_{ij}) = (1 - [(i-1)/n])[(k)_j/(k + (1/n))_j] = P(X < Y_{ij})$$

when $F(x)$ is the uniform distribution function. In this special case a comparison of ordered random variables using expected values will give the same result as a comparison using (3.9).

6. Illustration. Suppose there are nk students at a science fair, each with a science display, competing for k prizes. The judge has divided the displays into k groups of n displays each. The displays within each group are ranked in the usual manner, and then the groups are ranked on the basis of the highest scoring display in each group. The random variable Y_{ij} then corresponds to the display of rank i in the group of rank j.

The k prizes are awarded to $Y_{1,1}$ through $Y_{1,k}$ in that order. However, if all nk displays had been considered at one time, it is possible that the prizes would have been awarded differently. It is possible that $Y_{2,1}$ is a better display than $Y_{1,k}$. If the ideal method of awarding prizes is by considering all nk displays at

TABLE 1

Ranks of nk random variables, based on $P(X > Y_{ij})$

n	i	$k=1$ j	$k=2$ j		$k=3$ j			$k=4$ j				$k=5$ j				
		1	1	2	1	2	3	1	2	3	4	1	2	3	4	5
1	1	1	1	2	1	2	3	1	2	3	4	1	2	3	4	5
2	1	1	1	2	1	2	3	1	2	3	5	1	2	3	4	7
	2	2	3	4	4	5	6	4	6	7	8	5	6	8	9	10
3	1	1	1	2	1	2	4	1	2	3	6	1	2	3	4	8
	2	2	3	4	3	5	6	4	5	7	8	5	6	7	9	10
	3	3	5	6	7	8	9	9	10	11	12	11	12	13	14	15
4	1	1	1	2	1	2	4	1	2	3	6	1	2	3	4	8
	2	2	3	4	3	5	6	4	5	7	9	5	6	7	9	12
	3	3	5	6	7	8	9	8	10	11	12	10	11	13	14	15
	4	4	7	8	10	11	12	13	14	15	16	16	17	18	19	20
5	1	1	1	2	1	2	4	1	2	3	6	1	2	3	4*	8
	2	2	3	4	3	5	6	4	5	7	9	5*	6	7	9	12
	3	3	5	6	7	8	9	8	10	11	12	10	11	13	14	15
	4	4	7	8	10	11	12	13	14	15	16	16	17	18	19	20
	5	5	9	10	13	14	15	17	18	19	20	21	22	23	24	25

* For the table of ranks based on $P(Y_{i_1j_1} > Y_{i_2j_2})$, interchange the two starred ranks and use the above table.

once, then perhaps several alternate methods should be considered and compared with the actual method of awarding the prizes, to see which method comes closest to the ideal.

The first alternate method is to rank the random variables Y_{ij} according to their ability to exceed the unranked random variable X drawn from the same original population. Then the random variable $Y_{i_1j_1}$ is considered better than the random variable $Y_{i_2j_2}$ only if $P(X > Y_{i_1j_1}) < P(X > Y_{i_2j_2})$, where the desired probabilities are computed using (3.9). The k prizes would then be awarded to the k best displays, represented by the k best random variables in the above ordering. For values of n and k from 1 to 5, this method results in the ordering shown by the table, the best displays having the lowest ranks.

The second alternate method is to declare $Y_{i_1j_1}$ better than $Y_{i_2j_2}$ only if the probability of $Y_{i_1j_1}$ exceeding $Y_{i_2j_2}$, as given by (4.1), is greater than $\frac{1}{2}$. Curiously, this method does not always yield the same results as the first alternate method. For n and k ranging from 1 to 5 the results using this method are the same as those given in the table, except for $n = 5$, $k = 5$, this second method would give $Y_{2,1}$ the rank 4 instead of 5 and $Y_{1,4}$ the rank 5 instead of 4.

The following unusual situation is now possible. In choosing the better display when $n = 5$ and $k = 5$, the judge might favor $Y_{2,1}$ over $Y_{1,4}$ since $P(Y_{2,1} > Y_{1,4}) = .53$. And yet, in choosing a display for touring the countryside and competing against all comers, the judge might favor $Y_{1,4}$ over $Y_{2,1}$ since $Y_{1,4}$ is more difficult to beat from the standpoint of the unranked display X, their respective probabilities of getting beat being .22 and .23.

As n and k increase, the differences between the two alternate methods become more apparent. Consider the case where $n = 50$ and $k = 20$. Under method one, the last nine of the twenty prizes would be awarded to $Y_{2,1}$, $Y_{2,2}$, $Y_{1,12}$, $Y_{2,3}$, $Y_{2,4}$, $Y_{1,13}$, $Y_{2,5}$, $Y_{2,6}$, and $Y_{1,14}$ in that order. Under the second method those same prizes would go to $Y_{1,12}$, $Y_{1,13}$, $Y_{2,1}$, $Y_{2,2}$, $Y_{1,14}$, $Y_{2,3}$, $Y_{2,4}$, $Y_{2,5}$, and $Y_{1,15}$ in that order.

An inconsistency sometimes arises when the second method is used. For example when $n = 40$ and $k = 12$, $P(Y_{4,1} > Y_{1,12}) = .505$ and $P(Y_{1,12} > Y_{2,11}) = .506$, but $P(Y_{2,11} > Y_{4,1}) = .808$ leading to an inconclusive ranking.

A third alternate method of assigning ranks is on the basis of the expected value of Y_{ij}. Knowledge of the underlying distribution function $F(x)$ is needed here. If $F(x)$ is the uniform distribution function, it was shown in the previous section that the Y_{ij} will be ranked exactly the same as by the first alternate method.

7. Acknowledgments. The author is grateful to M. A. Benson of the United States Geological Survey for his clear formulation of a problem in flood frequency analysis which led directly to the subject of this paper. Also the author is indebted to Professor E. Batschelet for his many helpful suggestions that led to some of the results of this paper and greatly simplified the proofs of other results, and for his careful supervision in directing the research involved. Credit for the reference to Cohn, Mosteller, Pratt and Tatsuoka (1960) and the idea behind Section 5 belong to a referee.

REFERENCES

COHN, R., MOSTELLER, F., PRATT, J. W. and TATSUOKA, M. (1960). Maximizing the probability that adjacent order statistics of samples from several populations form overlapping intervals. *Ann. Math. Statist.* **31** 1095–1104.

FISZ, M. (1963). *Probability Theory and Mathematical Statistics* (3rd ed.). Wiley, New York.

GUMBEL, E. J. (1954). The maxima of the mean largest value and of the range. *Ann. Math. Statist.* **25** 76–84.

MOSTELLER, F. (1948). A k-sample slippage test for an extreme population. *Ann. Math. Statist.* **19** 58–65.

SARHAN, A. E. and GREENBERG, B. G. (1962). *Contributions to Order Statistics*. Wiley, New York.

WILKS, S. S. (1962). *Mathematical Statistics*. Wiley, New York.