The Annals of Statistics
1976, Vol. 4, No. 6, 1190–1199

THE MULTIVARIATE INCLUSION-EXCLUSION FORMULA AND ORDER STATISTICS FROM DEPENDENT VARIATES

BY WILLI MAURER[1] AND BARRY H. MARGOLIN[2]

University of Florida and Yale University

A variant of the general multivariate inclusion-exclusion formula of Meyer (1969) is derived for the case where K classes of events are considered *and* specific subsets of the events, one from each class, are related to one another by set inclusion. This result, in turn, yields a formula for the cumulative distribution function of any subset of order statistics from dependent random variables in terms of cumulative distribution functions of subsets of the unordered variables. An important example of dependent random variables, where the variables are jointly distributed as a Dirichlet $D_n(1, 1, \cdots, 1)$, is discussed in detail; various authors' results for this distribution are extended, or rederived as special cases via the formulae presented.

1. Introduction. General results for order statistics from dependent random variables are rare when compared with those for the more restrictive independence case; this observation is corroborated by the paucity of references to papers dealing with dependence in the recent text on order statistics by David (1970). In formula 5.5.3 of his book, David presents one of these results, namely, an expression for the marginal cumulative distribution function of the rth order statistic from n exchangeable random variables in terms of the joint cumulative distribution functions of subsets of the unordered exchangeable variables. This expression is most easily deduced from the well-known inclusion-exclusion formula (e.g., page 109, Feller (1968)). The argument proceeds as follows.

The inclusion-exclusion formula states that if $A_1, \cdots, A_n$ are n events, and if the probability of realizing at least r of the events $\{A_i\}_{i=1}^n$ is denoted by $P(r; n)$ then

$$P(r; n) = \sum_{m=r}^{n} (-1)^{m-r}\binom{m-1}{r-1} S_m \,, \tag{1.1}$$

where

$$S_m = \sum_{i_1<i_2<\cdots<i_m} \Pr\{\textstyle\bigcap_{t=1}^{m} A_{i_t}\} \,.$$

If one now considers n random variables $Z_1, \cdots, Z_n$ that are exchangeable, i.e.,

Received August 1974; revised April 1976.

[1] Work was done while the author was a postdoctoral research fellow at the Department of Statistics, Yale University, and a visiting postdoctoral associate of the Statistical Laboratory, Iowa State University. Financial support was provided by the Swiss National Foundation.

[2] Prepared in connection with research supported by the Army, Navy, Air Force and NASA under a contract administered by the Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

AMS 1970 *subject classifications.* Primary 62G30; Secondary 60C05.

Key words and phrases. Multivariate inclusion-exclusion, dependent random variables, exchangeable random variables, order statistics, Dirichlet distribution.

the joint cumulative distribution function of $\{Z_i\}_{i=1}^n$ is symmetric in its arguments, and if one defines $A_i = \{Z_i \leqq y\}$, $i = 1, \cdots, n$, then it follows immediately from (1.1) that for the corresponding order statistics $Z_{(1)} \leqq Z_{(2)} \leqq \cdots \leqq Z_{(n)}$,

$$\text{(1.2)} \qquad \Pr\{Z_{(r)} \leqq y\} = \sum_{m=r}^{n} (-1)^{m-r}\binom{m-1}{r-1}\binom{n}{m} \Pr\{\bigcap_{i=1}^{m} (Z_i \leqq y)\}\,.$$

This argument has been repeated in its entirety because in Section 2 of this paper, (1.1) will be generalized to permit consideration of K classes of events for which specific subsets of the events, one from each class, are related to one another by set inclusion. This result will then lead to a generalization of (1.2) for joint behavior of subsets of the order statistics from dependent, not necessarily exchangeable random variables. The paper's final section will consider an example where the dependent variables are exchangeable and distributed jointly with a Dirichlet $D_n(1, 1, \cdots, 1)$ probability density function.

2. A special case of the multivariate inclusion-exclusion formula. The inclusion-exclusion formula in (1.1) has been extended by Meyer (1969) to the case of K (finite) classes. To extend (1.2) to the joint distribution function of subsets of order statistics from dependent random variables, one is led to consider an extension of (1.1) for events related by inclusion. The desired form does not follow trivially from Meyer's results; it is simpler to produce the desired generalization from first principles.

Consider collections of events $\mathscr{A}_i = \{A_{i,j}\}_{j=1}^n$, $i = 1, 2, \cdots, K$, such that for all j, $A_{1,j} \subseteqq A_{2,j} \subseteqq \cdots \subseteqq A_{K,j}$. Let $\mathbf{t} = (t_1, t_2, \cdots, t_K)$ denote a vector of K nonnegative integers, let $P[\mathbf{t}; n]$ denote the probability that *exactly* t_i events from $\mathscr{A}_i$ occur, and let $P(\mathbf{t}; n)$ denote the probability that *at least* t_i events from $\mathscr{A}_i$ occur, $i = 1, 2, \cdots, K$. For both probabilities, one's interest focuses on cases where $t_1 \leqq t_2 \leqq \cdots \leqq t_K$. The first probability is zero otherwise and the second is unchanged if t_i is replaced by $\max(t_1, t_2, \cdots, t_i)$, for all i. Moreover, it suffices to consider $t_i > 0$ for $P(\mathbf{t}; n)$; one can always discard $\mathscr{A}_i$ if $t_i = 0$.

The following notation will be employed. For each vector of length K, say $\mathbf{x} = (x_1, x_2, \cdots, x_K)$, define $x_0 = 0$, and $x_{K+1} = n$ unless specified otherwise. Then define $\mathbf{x}^* = (x_2, x_3, \cdots, x_{K+1})$; $\Delta x_i = x_i - x_{i-1}$, $i = 1, 2, \cdots, K+1$; $\Delta\mathbf{x} = (\Delta x_1, \Delta x_2, \cdots, \Delta x_K)$; $\Delta^*\mathbf{x} = (\Delta x_1, \Delta x_2, \cdots, \Delta x_{K+1})$; and $x_+ = \sum_{i=1}^K x_i$. A K-length vector whose elements are all ones will be denoted by $\mathbf{1}$. For two vectors of the same length, say $\mathbf{x}$ and $\mathbf{y}$, the interpretation of $\mathbf{x} \leqq \mathbf{y}$ will be elementwise weak inequality. Finally, $\sum_{\mathbf{x}}$ will denote summation over all vectors $\mathbf{x}$ *that are ordered* $0 \leqq x_1 \leqq x_2 \leqq \cdots \leqq x_K \leqq n$, and $\sum_{\mathbf{x}=\mathbf{s}}^{\mathbf{r}}$ will denote summation over all vectors $\mathbf{x}$ that are ordered as above but in addition satisfy $\mathbf{s} \leqq \mathbf{x} \leqq \mathbf{r}$ for $\mathbf{r}$ and $\mathbf{s}$ of length K.

The generalization of (1.1) follows:

THEOREM 2.1. *For a vector* $\mathbf{a} = (a_1, a_2, \cdots, a_K)$ *of integers such that* $0 = a_0 \leqq a_1 \leqq \cdots \leqq a_K \leqq a_{K+1} = n$ *define*

$$P_{\mathbf{a}} = \sum{}' \Pr\{\bigcap_{i=0}^{K-1} \bigcap_{m=a_i+1}^{a_{i+1}} A_{i+1,j_m}\}\,,$$

where $\sum'$ *denotes summation over all*

$$\binom{n}{\Delta^*\mathbf{a}}$$

sets of indices $(j_1, j_2, \cdots, j_{a_K})$ *such that* $j_r < j_{r+1}$ *for* $r \in (a_i + 1, a_i + 2, \cdots, a_{i+1} - 1)$, $i = 0, 1, \cdots, K-1$, *and* $j_r \neq j_s$ *when* $1 \leqq r < s \leqq a_K$.

Then

$$P[\mathbf{t}; n] = (-1)^{t_+} \sum_{\mathbf{a}=\mathbf{t}}^{\mathbf{t}^*} (-1)^{a_+} \prod_{i=1}^{K} \binom{\Delta a_i}{a_i - t_i} P_{\mathbf{a}} \tag{2.1}$$

and

$$P(\mathbf{t}; n) = (-1)^{t_+} \sum_{\mathbf{a}=\mathbf{t}}^{\mathbf{a}^*} (-1)^{a_+} \prod_{i=1}^{K} \binom{(\Delta a_i) - 1}{a_i - t_i} P_{\mathbf{a}} \,. \tag{2.2}$$

PROOF. For $\mathbf{w} = (w_1, \cdots, w_K)$, let $\varphi(\mathbf{w}) = \sum_{\mathbf{a}} P_{\mathbf{a}} \prod_{i=1}^{K} w_i^{\Delta a_i}$. It is easily verified that

$$\varphi(\mathbf{w}) = E[\prod_{j=1}^{n} \sum_{i=1}^{K} (I_{i,j} - I_{i-1,j})(1 + \sum_{r=i}^{K} w_r)] \tag{2.3}$$

where $I_{i,j}$ is the indicator function of set $A_{i,j}$. For this verification, note that if a point belongs to $A_{i,j}$ and not to $A_{i-1,j}$, then it belongs to none of the sets $A_{r,j}$ for $r < i$ and to all sets $A_{r,j}$ for $r \geqq i$ (j fixed).

Next for $\mathbf{z} = (z_1, \cdots, z_K)$ let

$$\psi[\mathbf{z}] = \sum_{\mathbf{t}} P[\mathbf{t}; n] \prod_{r=1}^{K} z_r^{t_r} \,. \tag{2.4}$$

It is easily verified that

$$\psi[\mathbf{z}] = E[\prod_{j=1}^{n} (\sum_{i=1}^{K} (I_{i,j} - I_{i-1,j})) \prod_{r=i}^{K} z_r] \,. \tag{2.5}$$

The right-hand sides of (2.3) and (2.5) are equal if, for all i,

$$\prod_{r=i}^{K} z_r = 1 + \sum_{r=i}^{K} w_r \,;$$

this implies that $w_i = (z_i - 1) \prod_{r=i+1}^{K} z_r$, $i < K$, and $w_K = z_K - 1$. Hence,

$$\begin{aligned} \psi[\mathbf{z}] &= \varphi((z_1 - 1) \textstyle\prod_{r=2}^{K} z_r, (z_2 - 1) \prod_{r=3}^{K} z_r, \cdots, z_K - 1) \\ &= \textstyle\sum_{\mathbf{a}} P_{\mathbf{a}} \prod_{r=1}^{K} (z_r - 1)^{\Delta a_r} z_r^{a_{r-1}} \,. \end{aligned} \tag{2.6}$$

Equating coefficients in (2.4) and (2.6) yields:

$$P[\mathbf{t}; n] = \sum_{\mathbf{a}=\mathbf{t}}^{\mathbf{t}^*} \prod_{i=1}^{K} \binom{\Delta a_i}{a_i - t_i} \cdot (-1)^{(a_+ - t_+)} P_{\mathbf{a}} \,,$$

which proves (2.1).

The proof of (2.2) is also based on (2.6). Let

$$\begin{aligned} \psi[\mathbf{z}] &= \textstyle\sum_{\mathbf{t}} P(\mathbf{t}; n) \prod_{i=1}^{K} z_i^{t_i} = \sum_{\mathbf{t}} \prod_{i=1}^{K} z_i^{t_i} \sum_{\mathbf{s}=\mathbf{t}} P[\mathbf{s}; n] \\ &= \textstyle\sum_{\mathbf{t}} P[\mathbf{t}; n] \prod_{i=1}^{K} (z_i^{t_i+1} - 1)(z_i - 1)^{-1} \,. \end{aligned}$$

The replacement of every polynomial $Q(z_i)$ in (2.4) and (2.6) by $(z_i - 1)(z_i Q(z_i) - Q(1))$ for the cases of interest with $\mathbf{a} \geqq \mathbf{1}$ yields

$$\psi[\mathbf{z}] = \sum_{\mathbf{a}} P_{\mathbf{a}} \prod_{i=1}^{K} (z_i - 1)^{(\Delta a_i) - 1} z_i^{a_{i-1}+1} \,. \tag{2.7}$$

Consequently $P(\mathbf{t}; n)$ is equal to the coefficient of $\prod_{i=1}^{K} z_i^{t_i}$ in (2.7). Using the extended definition of binomial coefficients, one finds

$$P(\mathbf{t}; n) = \sum_{\mathbf{a}} (-1)^{(a_+ - t_+)} P_{\mathbf{a}} \prod_{i=1}^{K} \binom{(\Delta a_i) - 1}{t_i - a_{i-1} - 1}$$
$$= (-1)^{t_+} \sum_{\mathbf{a}=\mathbf{t}}^{\mathbf{a}^*} (-1)^{a_+} \prod_{i=1}^{K} \binom{(\Delta a_i) - 1}{a_i - t_i} P_{\mathbf{a}} .$$

Note that the upper limit $\mathbf{a}^*$ in the sum cannot be replaced by $(\mathbf{t}^* - \mathbf{1})$ because for $a_i = a_{i-1}$,

$$\binom{(\Delta a_i) - 1}{a_i - t_i} = (-1)^{a_i - t_i} . \qquad \square$$

If, in addition to the assumptions of Theorem 2.1, the events are assumed *exchangeable* within classes, i.e.

$$\Pr\{\bigcap_{i=0}^{K-1} \bigcap_{m=a_i+1}^{a_{i+1}} A_{i+1,j_m}\} = \Pr\{\bigcap_{i=0}^{K-1} \bigcap_{m=a_i+1}^{a_{i+1}} A_{i+1,m}\} = p_{\mathbf{a}} \tag{2.8}$$

for all sets of indices considered in $\sum'$,

then

$$P_{\mathbf{a}} = \binom{n}{\Delta^* \mathbf{a}} p_{\mathbf{a}}$$

and this simplification can be introduced in (2.1) and (2.2).

For $K = 2$ and the exchangeability in (2.8), (2.1) and (2.2) reduce respectively to:

$$P[r, s; n] = (-1)^{r+s} \sum_{b=s}^{n} \sum_{a=r}^{s} (-1)^{a+b} \frac{n!}{(n-b)!\,(b-a)!\,a!} \times \binom{a}{a-r}\binom{b-a}{b-s} p_{a,b} \tag{2.9}$$

and

$$P(r, s; n) = (-1)^{r+s} \sum_{b=s}^{n} \sum_{a=r}^{b} (-1)^{a+b} \frac{n!}{(n-b)!\,(b-a)!\,a!} \times \binom{a-1}{a-r}\binom{b-a-1}{b-s} p_{a,b} . \tag{2.10}$$

The following recurrence relations then hold:

THEOREM 2.2. *For $K = 2$, and $1 \leqq r \leqq s \leqq n - 1$, if one assumes* (2.8) *and the assumptions of Theorem* 2.1, *then*

$$nP[r, s; n-1] = (r+1)P[r+1, s+1; n] + (s-r+1)P[r, s+1; n] + (n-s)P[r, s; n] \tag{2.11}$$

and

$$nP(r, s; n-1) = rP(r+1, s+1; n) + (s-r)P(r, s+1; n) + (n-s)P(r, s; n) . \tag{2.12}$$

PROOF. A "dropping" argument similar to one employed previously for order statistics (David (1970), page 83) will be constructed. Of the n paired events $(A_{1,j}, A_{2,j})$, $j = 1, \cdots, n$, consider dropping from consideration *one pair at random*. For the dropping to have resulted in the occurrence of exactly r A_1's and s A_2's from the remaining $n - 1$ pairs, one of three disjoint possibilities would have had to obtain:

(a) exactly $(r+1)$ A_1's and $(s+1)$ A_2's occurred initially, and then a pair $(A_{1,j}, A_{2,j})$, for which both events had occurred, was dropped;

(b) exactly r A_1's and $(s+1)$ A_2's occurred initially, and then a pair $(A_{1,j}, A_{2,j})$, for which $A_{2,j}$ had occurred but $A_{1,j}$ had not, was dropped;

or

(c) exactly r A_1's and s A_2's occurred initially, and then a pair $(A_{1,j}, A_{2,j})$, for which neither event had occurred, was dropped.

Since (a), (b) and (c) have respective probabilities $((r+1)/n)P[r+1, s+1; n]$, $((s-r+1)/n)P[r, s+1; n]$ and $((n-s)/n)P[r, s; n]$, the result for $P[r, s; n-1]$ follows.

Via straightforward algebraic reduction, the second portion of the theorem follows from the fact that

$$P(r, s; n-1) = \sum_{i=r}^{n-1} \sum_{j=s}^{n-1} P[i, j; n-1]\,. \qquad \square$$

Similar, albeit more complicated, recurrence relations can be derived for $K > 2$; for $K = 1$, it is easily verified that:

$$nP[r; n-1] = (r+1)P[r+1; n] + (n-r)P[r; n]\,, \tag{2.13}$$

$$nP(r; n-1) = rP(r+1; n) + (n-r)P(r; n)\,. \tag{2.14}$$

In the next section the connection between (2.12) and a recurrence relation known to exist for distribution functions of order statistics from exchangeable random variables will be discussed.

3. Order statistics from dependent random variables. In much the same way that (1.2) followed from (1.1), so one may deduce the following result for the joint distribution function of subsets of order statistics from dependent random variables.

THEOREM 3.1. *Let* $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_n)$ *be a vector of* n *dependent random variables and let* $\mathbf{c}$ *denote a vector of real numbers* $(c_1, c_2, \cdots, c_K)$ *with* $c_1 \leqq c_2 \leqq \cdots \leqq c_K$. *For a vector* $\mathbf{s} = (s_1, s_2, \cdots, s_K)$ *of integers such that* $1 \leqq s_1 \leqq s_2 \leqq \cdots \leqq s_K \leqq n$, *let* $\mathbf{Z}(\mathbf{s})$ *denote the vector* $(Z_{(s_1)}, Z_{(s_2)}, \cdots, Z_{(s_K)})$ *of* K *order statistics of* $\mathbf{Z}$. *Then*

$$\Pr\{\mathbf{Z}(\mathbf{s}) \geqq \mathbf{c}\} = (-1)^{t+} \sum_{\mathbf{a}=\mathbf{t}}^{\mathbf{a}^*} (-1)^{a+} \prod_{j=1}^{K} \binom{(\Delta a_j) - 1}{a_j - t_j} \times \sum{}' \Pr\{\textstyle\bigcap_{i=0}^{K-1} \bigcap_{m=a_i+1}^{a_{i+1}} (Z_{j_m} \geqq c_{K-i})\}\,, \tag{3.1}$$

where $\sum'$ denotes summation over the same set of indices as in Theorem 2.1 and where $t_j = n + 1 - s_{K+1-j}$, $j = 1, \cdots, K$. Equivalently,

$$\Pr\{\mathbf{Z}(\mathbf{s}) \leqq \mathbf{c}\} = (-1)^{s+} \sum_{\mathbf{a}=\mathbf{s}}^{\mathbf{a}^*} (-1)^{a+} \prod_{j=1}^{K} \binom{(\Delta a_j) - 1}{a_j - s_j} \times \sum{}' \Pr\{\textstyle\bigcap_{i=0}^{K-1} \bigcap_{m=a_i+1}^{a_{i+1}} (Z_{j_m} \leqq c_{i+1})\}\,. \tag{3.2}$$

PROOF. Both (3.1) and (3.2) follow immediately from (2.2). For (3.1) define the event $A_{i,j}$ by $A_{i,j} = \{Z_j \geqq c_{K+1-i}\}$, $j = 1, \cdots, n$, and let $\mathscr{A}_i = \{A_{i,j}, j = 1, \cdots, n\}$ for $i = 1, \cdots, K$. The events $\{A_{i,j}\}$ satisfy the assumptions in Theorem 2.1. Moreover $\Pr\{\mathbf{Z}(\mathbf{s}) \geqq \mathbf{c}\} = \Pr\{\bigcap_{i=1}^{K}$ (at least t_i events from $\mathscr{A}_i$ occur)$\}$, where $t_i = n + 1 - s_{K+1-i}$. From (2.2) and the fact that

$$P_{\mathbf{a}} = \sum{}' \Pr\{\textstyle\bigcap_{i=0}^{K-1} \bigcap_{m=a_i+1}^{a_{i+1}} (Z_{j_m} \geqq c_{K-i})\}\,,$$

one arrives at (3.1); (3.2) is proved in a similar fashion with $A_{i,j} = \{Z_j \leqq c_i\}$ and $\Pr\{\mathbf{Z}(\mathbf{s}) \leqq \mathbf{c}\} = \Pr\{\bigcap_{i=1}^{K}$ (at least s_i events from $\mathscr{A}_i$ occur)$\}$. □

Equation (2.1) may be used to derive a result similar to Theorem 3.1. The proof, which parallels the proof of Theorem 3.1, will not be given.

THEOREM 3.2 (without proof). *Let* $\mathbf{Z}$ *represent a vector of n dependent random variables. Consider a vector* $\mathbf{c}$ *of K constants* c_i, *such that* $c_1 \leqq c_2 \leqq \cdots \leqq c_K$, *and a vector* $\mathbf{s}$ *of K integers* s_i, $1 \leqq s_1 \leqq s_2 \leqq \cdots \leqq s_K \leqq n + 1$ *for* $1 \leqq K \leqq n$. $\mathbf{Z}(\mathbf{s})$ *will continue to be defined as in Theorem* 3.1; *in addition, define* $Z_{(0)} = -\infty$ *and* $Z_{(n+1)} = \infty$. *Then*

$$\Pr\{\mathbf{Z}(\mathbf{s}-1) < \mathbf{c} \leqq \mathbf{Z}(\mathbf{s})\} \\ = (-1)^{t_+} \sum_{\mathbf{a}=\mathbf{t}}^{\mathbf{t}^*} (-1)^{a_+} \prod_{j=1}^{K} \binom{\Delta a_j}{a_j - t_j} \\ \times \sum{}' \Pr\{\textstyle\bigcap_{i=0}^{K-1} \bigcap_{m=a_i+1}^{a_{i+1}} (Z_{j_m} \geqq c_{K-i})\} \tag{3.3}$$

where $t_j = n + 1 - s_{K+1-j}$, $j = 1, \cdots, K$.

If the random variables $Z_1, Z_2, \cdots, Z_n$ are *exchangeable*, then paralleling (2.8), one notes that

$$\sum{}' \Pr\{\textstyle\bigcap_{i=0}^{K-1} \bigcap_{m=a_i+1}^{a_{i+1}} (Z_{j_m} \geqq c_{K-i})\} \\ = \binom{n}{\Delta^* \mathbf{a}} \Pr\{\textstyle\bigcap_{i=0}^{K-1} \bigcap_{m=a_i+1}^{a_{i+1}} (Z_m \geqq c_{K-i})\}\,, \tag{3.4}$$

and this simplification can be introduced in (3.1) and (3.3). A similar simplification for exchangeability in (3.2) is clear.

This section concludes with comments on (3.1), (3.2) and (3.3) for the case $K = 2$.

(1) For $s_1 = r$, $s_2 = s$, (3.2) and (3.4) provide an expression for the bivariate cumulative distribution function of the rth and sth order statistics from a sample of *n exchangeable* random variables; this distribution function is usually denoted by $F_{r,s:n}(y, z)$.

The recurrence relation in (2.12) for $P(r, s; n-1)$ implies the following for $F_{r,s:n}(y, z)$:

$$nF_{r,s:n-1}(y, z) = rF_{r+1,s+1:n}(y, z) + (s - r)F_{r,s+1:n}(y, z) \\ + (n - s)F_{r,s:n}(y, z)\,. \tag{3.5}$$

The recurrence relation in (3.5) has been established previously (see David (1970), page 83); (2.12) is in one sense a generalization of equation (3.5) beyond order statistics.

(2) If one sets $s_1 = n - 1$, $s_2 = n$, and $\alpha \leqq \beta$, (3.2) yields:

$$\Pr\{Z_{(n-1)} \leqq \alpha, Z_{(n)} \leqq \beta\} = \sum_{j=1}^{n} \Pr\{\bigcap_{i\neq j} (Z_i \leqq \alpha) \cap (Z_j \leqq \beta)\} - (n-1)\Pr\{\bigcap_{i=1}^{n} (Z_i \leqq \alpha)\},$$

a result that is easily verified from first principles.

(3) If one sets $s_1 = 1$, $s_2 = n + 1$, and $\alpha \leqq \beta$, then (3.3) yields:

$$\begin{aligned} &\Pr\{\alpha \leqq Z_{(1)}, Z_{(n)} < \beta\} \\ &= \Pr\{\alpha \leqq Z_i < \beta, \text{ for all } i\} \\ &= \sum_{r=0}^{n} (-1)^r \sum{}' \Pr\{\bigcap_{i=1}^{r} (Z_{m_i} \geqq \beta) \cap \bigcap_{i=r+1}^{n} (Z_{m_i} \geqq \alpha)\}, \end{aligned} \tag{3.6}$$

where $\sum'$ denotes summation over all $\binom{n}{r}$ partitions of $(1, 2, \cdots, n)$ into $(m_1, m_2, \cdots, m_r)$ and $(m_{r+1}, m_{r+2}, \cdots, m_n)$.

4. Dirichlet $D_n(1, 1, \cdots, 1)$ order statistics and related problems. From Wilks (1962), page 179, one learns that if $X_1, \cdots, X_n$ are independent and identically distributed with an exponential probability density function

$$f(x; \theta) = \theta^{-1} e^{-x/\theta}, \qquad x > 0, \theta > 0,$$

then the transformation

$$Z_i = X_i / \sum_{j=1}^{n} X_j, \qquad i = 1, \cdots, n,$$

yields variables whose joint distribution is Dirichlet $D_n(1, 1, \cdots, 1)$. These random variables are *exchangeable* and singular with density

$$f(Z_1, \cdots, Z_n) = (n-1)!\, I[(\sum_{i=1}^{n} Z_i = 1) \cap \bigcap_{i=1}^{n} (Z_i > 0)], \tag{4.1}$$

where $I[S]$ denotes an indicator function for set S. The corresponding order statistics $\{Z_{(i)}\}_{i=1}^{n}$ then have a joint distribution

$$\begin{aligned} &f(Z_{(1)}, \cdots, Z_{(n)}) \\ &= n!\,(n-1)!\, I[(\sum_{i=1}^{n} Z_{(i)} = 1) \cap (0 < Z_{(1)} < \cdots < Z_{(n)})]. \end{aligned} \tag{4.2}$$

David and Johnson (1948) observed that the distributions in (4.1) and (4.2) are independent of the unknown θ.

The Dirichlet $D_n(1, 1, \cdots, 1)$ distribution and its corresponding order statistics arise in many contexts. David (1970), on pages 79–81, discusses the following related problems.

(a) Random divisions of a unit interval: Here $\{Z_i\}_{i=1}^{n}$ are distances between successive points that have been dropped at random in the interval [0, 1].

(b) Elementary coverages of the order statistics: Here, for $X_{(n)} = \infty$,

$$Z_i = F(X_{(i)}) - F(X_{(i-1)}), \quad i = 2, \cdots, n, \qquad Z_1 = F(X_{(1)}),$$

where $\{X_{(i)}\}_{i=1}^{n-1}$ are order statistics from a random sample of size $n - 1$ from a population with continuous distribution function $F(x)$.

(c) Harmonic analysis: Here Z_i is the percentage of the total variation accounted for by the ith harmonic, $i = 1, \cdots, n$.

(d) Coverage of a circular perimeter: Here $\{Z_i\}_{i=1}^{n}$ are distances between the midpoints of successive arcs of equal length that have been randomly placed on the perimeter of a circle of unit circumference.

In all five problems, interest frequently centers on the ordered values $\{Z_{(i)}\}_{i=1}^{n}$. For example, using the easily verified fact that for $d_i > 0$, $i = 1, \cdots, n$

$$\Pr\{\textstyle\bigcap_{i=1}^{n}(Z_i > d_i)\} = \{\max(0, 1 - \sum_{i=1}^{n} d_i)\}^{n-1} = (1 - \sum_{i=1}^{n} d_i)^{n-1} I[\sum d_i \leqq 1], \tag{4.3}$$

Stevens (1939) proved via (1.2) that

$$\Pr\{Z_{(r)} > y\} = \textstyle\sum_{m=n-r+1}^{n} (-1)^{m-n+r-1}\binom{m-1}{n-r}\binom{n}{m}\{(1 - my)^{n-1} I[y < m^{-1}]\} \tag{4.4}$$

and applied the result to problems in (d).

One can now apply the result in (3.1), in conjunction with (4.3), to obtain joint behavior of subsets of K of the order statistics from a set of random variables whose joint distribution is $D_n(1, 1, \cdots, 1)$. For $1 \leqq s_1 < s_2 < \cdots < s_K \leqq n$; $c_1 \leqq c_2 \leqq \cdots \leqq c_K$; and $t_i = n + 1 - s_{K+1-i}$, $i = 1, \cdots, K$, one obtains the result

$$\Pr\{\mathbf{Z}(\mathbf{s}) > \mathbf{c}\} = (-1)^{t_+} \textstyle\sum_{\mathbf{a}=\mathbf{t}}^{\mathbf{t}^*} (-1)^{a_+} \binom{n}{\Delta^*\mathbf{a}} \prod_{j=1}^{K} \binom{(\Delta a_j) - 1}{a_j - t_j} \times (1 - \sum_{i=1}^{K} c_{K+1-i}\Delta a_i)^{n-1} I[\sum_{i=1}^{K} c_{K+1-i}\Delta a_i \leqq 1]. \tag{4.5}$$

For $K = 2$, $s > r$ and $\beta > \alpha$, (4.5) reduces to

$$\begin{aligned} &\Pr\{Z_{(s)} > \beta, Z_{(r)} > \alpha\} \\ &\quad = (-1)^{r+s} \textstyle\sum_{b=n-r+1}^{n} \sum_{a=n-s+1}^{b} (-1)^{a+b} \dfrac{n!}{(n-b)!\,(b-a)!\,a!} \\ &\qquad \times \binom{a-1}{a-n+s-1}\binom{b-a-1}{b-n+r-1}(1 - a\beta - (b-a)\alpha)^{n-1} \\ &\qquad \times I[a\beta + (b-a)\alpha \leqq 1]. \end{aligned} \tag{4.6}$$

Via (4.6) one can tie results in this paper to special cases that have appeared in the literature.

(1) If, in (4.6), one sets $r = 1$, $s = 2$, then for $\alpha \leqq \beta$ one obtains,

$$\Pr\{Z_{(1)} > \alpha, Z_{(2)} > \beta\} = n(1 - (n-1)\beta - \alpha)^{n-1} I[(n-1)\beta + \alpha \leqq 1] - (n-1)(1 - n\beta)^{n-1} I[n\beta \leqq 1]. \tag{4.7}$$

This last expression corrects an error in exercise 5.4.3(b) in David (1970), and extends the domain of (α, β) to include $\{(\alpha, \beta): (\alpha \leqq \beta) \cap (n\beta > 1) \cap ((n-1)\beta + \alpha \leqq 1)\}$. In this case, the second term in (4.7) is zero, but the first is not.

(2) If, in (4.6), one assumes $\alpha \leqq \beta$, and sets $r = 1$, $s = n$, one obtains

$$\Pr\{Z_{(1)} > \alpha, Z_{(n)} > \beta\} = \textstyle\sum_{j=1}^{n} (-1)^{j+1}\binom{n}{j}(1 - j\beta - (n-j)\alpha)^{n-1} I[j\beta + (n-j)\alpha \leqq 1].$$

It then follows that

$$\Pr\{Z_{(1)} > \alpha, Z_{(n)} < \beta\} = \textstyle\sum_{j=0}^{n} (-1)^{j}\binom{n}{j}(1 - j\beta - (n-j)\alpha)^{n-1} I[j\beta + (n-j)\alpha \leqq 1]. \tag{4.8}$$

The result in (4.8) appears as Theorem 8.4 in a paper on random divisions of an interval by Darling (1953).

(3) If one differentiates the results in (4.5) or (4.6) to obtain density functions, one recovers the results reported by Barton and David (1956); the reverse integration operation is not easily performed.

The general result in (4.5) is used extensively by Margolin and Maurer (1975) to obtain exact expressions and bounds for cumulative distribution functions of Kolmogorov–Smirnov type one-sample statistics used to test goodness of fit of the exponential distribution with unknown scale. Margolin and Maurer also indicate that result (3.6) of Durbin (1975) for the same goodness of fit problem is equivalent to (4.5) when $K = n$.

Consider, finally, an earlier application of the $\{Z_{(i)}\}_{i=1}^{n}$. In problem (d), i.e., coverage of a circular perimeter by randomly placed arcs of equal length, if the circle has unit perimeter and α is the common length of the arcs, then Stevens (1939) showed that

$$\Pr\{\text{at most } r \text{ breaks in the coverage}\} = \Pr\{Z_{(n-r)} \leqq \alpha\}.$$

Given (4.6), one can now evaluate:

$$\begin{aligned}\Pr\{&\text{at least } r \text{ breaks in the coverage and the largest break is} \\ &\text{larger than } \delta - \alpha, \text{ for } \delta > \alpha\} \\ &= \Pr\{(Z_{(n-r+1)} > \alpha) \cap (Z_{(n)} - \alpha \geqq \delta - \alpha)\} \\ &= \Pr\{(Z_{(n-r+1)} > \alpha, Z_{(n)} > \delta\}.\end{aligned}$$

From this expression one can compute, with the aid of (4.4), answers to conditional questions such as

$$\begin{aligned}\Pr\{&\text{largest break is smaller than } \delta - \alpha \\ &\text{given that there are at most } r \text{ breaks}\}.\end{aligned}$$

Acknowledgment. The authors wish to thank Professor S. J. Press for providing an important reference. A referee contributed substantially to the improvement of an earlier version of this paper.

REFERENCES

[1] Barton, D. E. and David, F. N. (1956). Some notes on ordered random intervals. *J. Roy. Statist. Soc. Ser. B* **18** 79–94.

[2] Darling, D. A. (1953). On a class of problems related to the random division of an interval. *Ann. Math. Statist.* **24** 239–253.

[3] David, F. N. and Johnson, N. L. (1948). The probability integral transformation when parameters are estimated from the sample. *Biometrika* **35** 182–190.

[4] David, H. A. (1970). *Order Statistics.* Wiley, New York.

[5] Durbin, J. (1975). Kolmogorov–Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings. *Biometrika* **62** 5–22.

[6] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* **1** (3rd ed.). Wiley, New York.

[7] Margolin, B. H. and Maurer, W. (1975). Tests of the Kolmogorov–Smirnov type for exponential data with unknown scale, and related problems. *Biometrika* **63** 149–160.

[8] MEYER, R. M. (1969). Note on a "multivariate" form of Bonferroni's inequalities. *Ann. Math. Statist.* **40** 692–693.

[9] STEVENS, W. L. (1939). Solution to a geometrical problem in probability. *Ann. Eugenics* **9** 315–320.

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

FORSCHUNGSINSTITUT FÜR MATHEMATIK
ETHZ
RÄMISTR. 101
8006 ZÜRICH, SWITZERLAND

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520