The Annals of Statistics
1981, Vol. 9, No. 4, 822–833

ON DISTRIBUTIONS DETERMINED BY RANDOM VARIABLES DISTRIBUTED OVER THE n-CUBE

By Iain D. Currie

Heriot-Watt University

The distribution function of a random variable of the form $\sum_{i=1}^{n} a_i Y_1 Y_2 \cdots Y_i$ where $a_i > 0$ and $0 \le Y_i \le 1$ is considered. A geometric argument is used to obtain the distribution function as a repeated integral. The result is used first to obtain the distribution function of a linear combination of variables defined over the simplex $X_i \ge 0$, $\sum_{i=1}^{n} X_i \le 1$. As a second application the distribution of certain quadratic forms over the simplex is obtained. This result yields as a special case the distribution of the internally studentized extreme deviate; the cases of normal and exponential samples are considered in detail and the required distributions obtained.

1. Introduction. Let $a_i > 0$, $i = 1, 2, \cdots, n$ be real numbers. We consider the distribution of the random variable

$$g(\mathbf{Y}) = \sum_{i=1}^{n} a_i Y_1 Y_2 \cdots Y_i \tag{1.1}$$

where $Y_1, Y_2, \cdots, Y_n$ are distributed over the unit n-cube, $0 \le Y_i \le 1$, $i = 1, 2, \cdots, n$. It is not assumed that $Y_1, Y_2, \cdots, Y_n$ are independent or identically distributed. We will obtain the distribution function of the random variable $g(\mathbf{Y})$ as a repeated integral; a statement of this theorem is given in Section 2 although we delay a proof until Section 4. Most of Section 2 is given over to an informal discussion of the geometry of random variables of the form (1.1) and an interpretation of the theorem and an outline of its proof are given in geometric terms.

The motivation for studying this random variable is as follows: we show in Section 3 that the solution to the distributional problem posed by (1.1) is equivalent to finding the distribution of a linear combination of variables defined over the simplex $X_i \ge 0$, $\sum_{i=1}^{n} X_i \le 1$. This problem has as a special case the widely studied problem of obtaining the distribution of a linear combination of the order statistics of a sample from the uniform distribution, a problem approached from a geometric point of view by Hall (1927) and later by Dempster and Kleyle (1968) and others. The Laplace transform provides a particularly effective way of tackling the problem of finding the distribution of a linear combination of order statistics from the uniform distribution, and recently Margolin (1977) used this method to obtain the distribution of a linear combination of Dirichlet variables. Closely related to this result is one first obtained by Fisher (1929), generalized first by Cochran (1941) and more recently by Lewis and Fieller (1979). Here we are concerned with the distribution of $X_{(n)}/\sum_{i=1}^{n} X_i$ based on gamma samples. As well as linear forms, we may use the result on the distribution of a random variable of the form (1.1) to obtain the distribution of certain quadratic forms over the simplex. As a special case of this we consider the distribution of the extreme studentized deviate, a statistic commonly used in simple data monitoring; with an underlying normal population we have the problem studied by Pearson and Chandra Sekar (1936), Grubbs (1950) and Borenius (1959), while with an underlying exponential distribution we require the distribution of the statistic proposed by Shapiro and Wilk (1972) as a test of exponentiality.

Received August, 1979; revised August, 1980.

AMS 1970 *subject classifications.* Primary 60D05; secondary 62E15, 62G30.

Key words and phrases. Geometric probability, order statistics, linear combinations, internally studentized deviate.

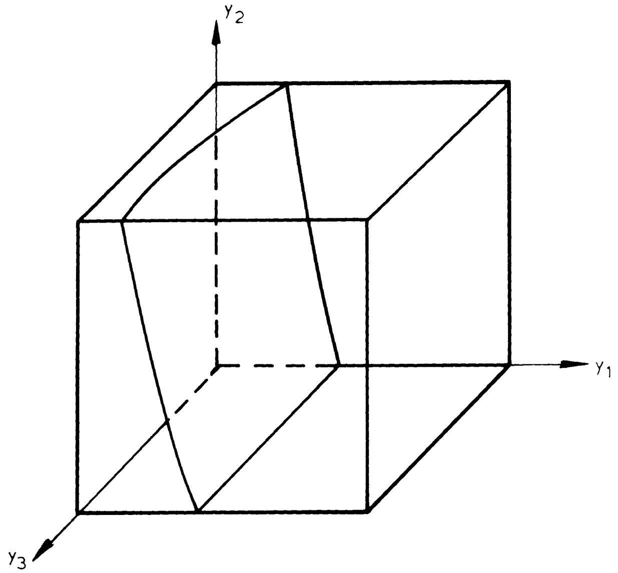

FIG. 1. *The surface* $g(\mathbf{y}) = c, 0 < c < 1$.

The paper concludes with a proof of the theorem and a few remarks on some of the geometric aspects of the results obtained in the main part of the paper.

2. The main theorem.

2.1 *Geometrical Considerations.* Random variables of the form (1.1) are in fact particularly convenient to deal with when the variables $Y_1, Y_2, \cdots, Y_n$ are defined over the n-cube. To illustrate the simplicity of the geometry involved consider an example with $n = 3$. Let $g(\mathbf{Y}) = Y_1 + Y_1Y_2 + Y_1Y_2Y_3$, $0 \le Y_i \le 1$, $i = 1, 2, 3$ and suppose that Y_1, Y_2, Y_3 have a continuous joint probability density function $f(\mathbf{y})$. In order to compute $\Pr(g(\mathbf{Y}) \le c)$ we must consider the surface $g(y_1, y_2, y_3) = c$. If $0 < c < 1$ this surface is shown in Figure 1 and the diagram tells us that

$$\Pr(g(Y_1, Y_2, Y_3) > c) = \int_0^1 \int_0^1 \int_{\alpha_1(y_2,y_3)}^1 f(y_1, y_2, y_3)\, dy_1\, dy_2\, dy_3 \tag{2.1}$$

where

$$\alpha_1(y_2, y_3) = c/(1 + y_2 + y_2y_3). \tag{2.2}$$

The first critical value of c is $c = 1$. Notice that $g(\mathbf{y}) = 1$ along the edge $y_1 = 1, y_2 = 0$ and so the next case to consider is $1 < c < 2$ when the surface is shown in Figure 2.

This time we find

$$\Pr(g(Y_1, Y_2, Y_3) > c) = \int_0^1 \int_{\alpha_2(y_3)}^1 \int_{\alpha_1(y_2,y_3)}^1 f(y_1, y_2, y_3)\, dy_1\, dy_2\, dy_3 \tag{2.3}$$

where

$$\alpha_2(y_3) = (c - 1)/(1 + y_3). \tag{2.4}$$

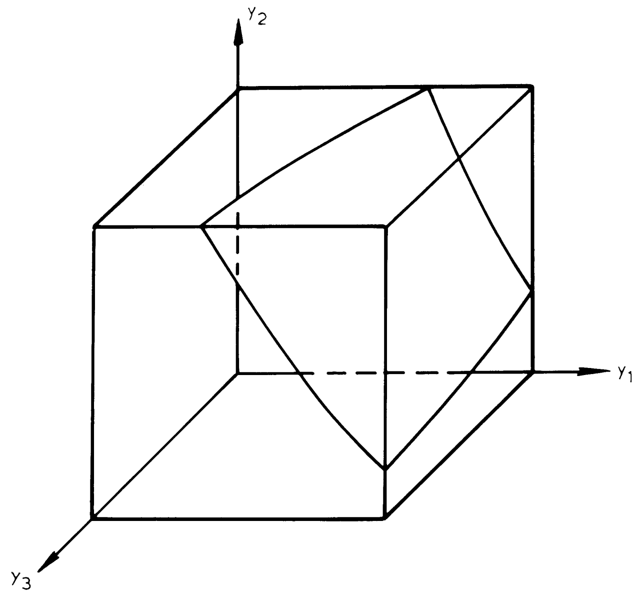

FIG. 2. *The surface* $g(\mathbf{y}) = c$, $1 < c < 2$.

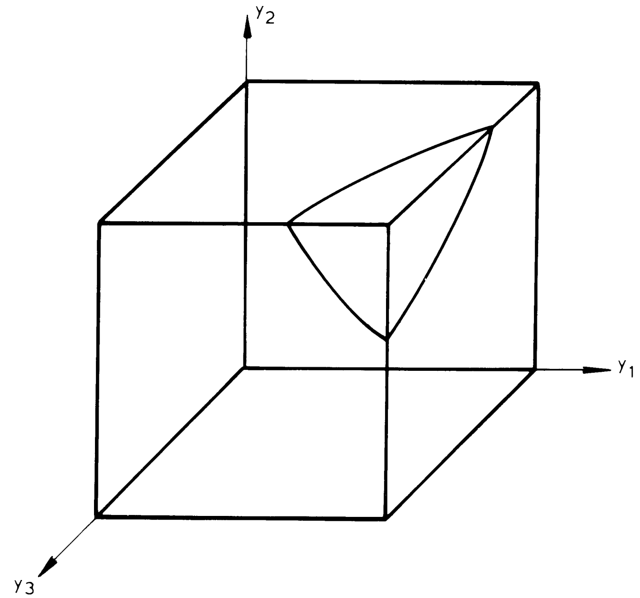

FIG. 3. *The surface* $g(\mathbf{y}) = c$, $2 < c < 3$.

The next critical value of c is $c = 2$. The form of the function g ensures that as c increases through the value 2 the surface $g(\mathbf{y}) = c$ changes form from that in Figure 2 straight to that in Figure 3.

The final form for the distribution function of $g(\mathbf{Y})$ is given for $2 < c < 3$ by

$$\Pr(g(Y_1, Y_2, Y_3) > c) = \int_{\alpha_3}^{1} \int_{\alpha_2(y_3)}^{1} \int_{\alpha_1(y_2, y_3)}^{1} f(y_1, y_2, y_3)\, dy_1\, dy_2\, dy_3 \tag{2.5}$$

where

$$\alpha_3 = c - 2. \tag{2.6}$$

This example illustrates a common feature of the distribution function of a random variable of the form $g(\mathbf{Y})$, namely, that it has a different functional form in each of n intervals. Distributions of this type are not uncommon in statistics but often they are awkward to deal with. There are two features of the variable $g(\mathbf{Y})$ that make it easy to study. The functional form of the distribution function is determined by the surface $g(\mathbf{y}) = c$; as c increases this form changes whenever the surface passes through a vertex of the cube. What is not so obvious is that the functional form should change only n times; this is a consequence of the 'nested' form of $g(\mathbf{y})$ which ensures that the surface $g(\mathbf{y}) = c$ coincides with whole faces of the n-cube for certain values of c. This dramatically reduces the number of cases to be considered. The other feature of the geometry of this problem worth noting is that the condition $a_i > 0$ ensures that the surface $g(\mathbf{y}) = c$ consists of a single piece inside the cube; this simplifies the regions of integration involved and means that the upper limits of integration are always unity.

2.2 *Statement of the Main Theorem.* We will obtain the distribution function of the random variable $g(\mathbf{Y})$ defined in (1.1) in terms of functions $\alpha_i = \alpha_i(n; x; y_{i+1}, \cdots, y_n)$; these functions are natural generalizations of those defined in (2.2), (2.4) and (2.6) and enable the distribution function of $g(\mathbf{Y})$ to be obtained as integrals of the form (2.1), (2.3) and (2.5).

Let $a_0 = 0$ and define

$$A(i) = a_i + \textstyle\sum_{j=i+1}^{n} a_j y_{i+1} y_{i+2} \cdots y_j, \qquad i = 0, 1, \cdots, n \tag{2.7}$$

where we have adopted the convention that $\sum_{j=i}^{n}$ with $i > n$ is identically zero. Clearly we have $A(0) = g(\mathbf{y})$ from (1.1), $A(n) = a_n$ and

$$A(i) = a_i + y_{i+1} A(i + 1), \qquad i = 0, 1, \cdots, n - 1. \tag{2.8}$$

Thus the functions $A(i)$ enable us to deal conveniently with the nested form of $g(\mathbf{y})$. Further, define

$$b_i = \textstyle\sum_{j=0}^{i} a_j, \qquad i = 0, 1, 2, \cdots, n. \tag{2.9}$$

Then, if $x \geq b_{i-1}$, define

$$\alpha_i = (x - b_{i-1})/A(i), \qquad i = 1, 2, \cdots, n. \tag{2.10}$$

Notice that the functional form of α_i factorises into a function of x and a function of $y_{i+1}, \cdots, y_n$; this has important consequences when the evaluation of the integrals is performed in particular applications. Observe also that we have simplified notation by dropping the variables on which $A(i)$ and α_i depend. Finally, let

$$I_n(i) = [b_{i-1}, b_i], \qquad i = 1, 2, \cdots, n. \tag{2.11}$$

The set of intervals $\{I_n(i)\}_{1 \leq i \leq n}$ covers the range of $g(\mathbf{Y})$ and these are nonoverlapping (apart from the end points). We can now state our main result.

THEOREM. *If $Y_1, Y_2, \cdots, Y_n$ have a continuous joint probability density function $f(\mathbf{y}) = f(y_1, y_2, \cdots, y_n)$ and if $x \in I_n(k)$ then*

$$\Pr(g(\mathbf{Y}) \geq x) = \int_0^1 \int_0^1 \cdots \int_0^1 \int_{\alpha_k}^1 \int_{\alpha_{k-1}}^1 \cdots \int_{\alpha_1}^1 f(\mathbf{y})\, d\mathbf{y} \tag{2.12}$$

The proof of this theorem is in two parts: first, we show how to define the set required in the integral (2.12); secondly, we show that this set coincides with the required set $\{\mathbf{y}; g(\mathbf{y}) \geq x\}$. The proof will be given in Section 4.

We now give a number of simple generalizations of the main theorem.

Remark 1. The condition that $Y_1, Y_2, \cdots, Y_n$ have a density function is not necessary. The result holds for an arbitrary distribution for $Y_1, Y_2, \cdots, Y_n$. We have that $P(g(\mathbf{Y}) \geq x)$ is given simply by adding up all the probability inside and on the boundary of the region of integration in the theorem. However, the applications given in this paper all have $\mathbf{Y}$ with a continuous probability density function.

Remark 2. The condition $a_i > 0$, $i = 1, \cdots, n$ may be relaxed to $a_i \geq 0$, $i = 1, \cdots, n$ with no difficulty. If only r of the n coefficients a_i are nonzero then the problem is easily expressed in the standard form for r variables with all $a_i > 0$.

Remark 3. Although we have stated the main result for a random variable of the form (1.1) where $0 \leq Y_i \leq 1$, $i = 1, 2, \cdots, n$ the essential assumption is that $Y_i \geq 0$. For example, if we have $0 \leq Y_i \leq \delta_i$ then the substitution $Y_i^* = Y_i/\delta_i$ recovers the problem in the standard form. If however we have only that $Y_i \geq 0$ then the geometry simplifies and it is easily seen that $P\{g(\mathbf{Y}) \geq x\}$ is given by the form

$$\int_0^\infty \cdots \int_0^\infty \int_{\alpha_1}^\infty f(\mathbf{y})\, d\mathbf{y}.$$

whatever the value of x.

3. Applications.

Example 1. We consider the distribution of $X = \sum_{i=1}^n c_i X_i$ where $\mathbf{X}' = (X_1, X_2, \cdots, X_n)$ is distributed over the simplex $X_i \geq 0$, $\sum_{i=1}^n X_i \leq 1$. We assume without loss of generality that $c_1 > c_2 > \cdots > c_n > 0$ and transform $\mathbf{X} \to \mathbf{Y}$ as follows:

$$(3.1)\qquad \begin{aligned} X_1 &= 1 - Y_1 \\ X_i &= (1 - Y_i) Y_1 Y_2 \cdots Y_{i-1}, \qquad i = 2, \cdots, n. \end{aligned}$$

The Jacobian of this transformation is $y_1^{n-1} y_2^{n-2} \cdots y_{n-1}$ and the joint probability density function of $Y_1, Y_2, \cdots, Y_n$ is easily found. The region $X_i \geq 0$, $\sum_{i=1}^n X_i \leq 1$ transforms into $0 \leq Y_i \leq 1$, $i = 1, 2, \cdots, n$ and the random variable X transforms

$$(3.2)\qquad X \to d_0 - \textstyle\sum_{i=1}^n d_i Y_1 Y_2 \cdots Y_i$$

where

$$(3.3)\qquad d_0 = c_1; \qquad d_i = c_i - c_{i+1}, \quad i = 1, 2, \cdots, n-1; \qquad d_n = c_n.$$

Thus

$$(3.4)\qquad P(X \leq x) = P(\textstyle\sum_{i=1}^n d_i Y_1 Y_2 \cdots Y_i \geq d_0 - x)$$

and since $d_i > 0$ (2.12) may be applied immediately. In this case we have $\alpha_i = (c_i - x)/A(i)$ where

$$A(i) = d_i + \textstyle\sum_{j=i+1}^n d_j Y_{i+1} Y_{i+2} \cdots Y_j.$$

In any particular application we still have the problem of evaluating the n-dimensional repeated integral in (2.12). This involves firstly the computation of the joint probability density function of $Y_1, Y_2, \cdots, Y_n$ and secondly the evaluation of the resulting integral. This evaluation may well have to be attempted numerically but there do exist a number of problems of interest where these integrals may be obtained explicitly. We illustrate the kind of calculation involved by considering the distribution of the sum in a sample from the uniform distribution on [0, 1]. This is a well known problem whose solution goes back to Lagrange. To apply the present method we note that if $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ are the order statistics from a sample of size n from the uniform distribution on [0, 1] with $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ and if we let $U_1 = X_{(1)}$, $U_i = X_{(i)} - X_{(i-1)}$, $i = 2, 3, \cdots, n$ then we have $f(u_1, u_2, \cdots, u_n) = n!$ where $u_1, u_2, \cdots, u_n$ are distributed over the simplex $u_i \ge 0$, $\sum_{i=1}^n u_i \le 1$ as required. In terms of $u_1, u_2, \cdots, u_n$ we have $\sum_{i=1}^n X_i = \sum_{i=1}^n X_{(i)} = \sum_{i=1}^n (n - i + 1) U_i$ and we may apply the preceding general result with $c_i = n - i + 1$. We find from (3.2) and (3.3) that

$$\textstyle\sum_{i=1}^n X_i \to n - Y_1 - Y_1 Y_2 - \cdots - Y_1 Y_2 \cdots Y_n$$

where $f(y_1, y_2, \cdots, y_n) = n!\, y_1^{n-1} y_2^{n-2} \cdots y_{n-1}$ over $0 \le y_i \le 1$, $i = 1, \cdots, n$. Applying the main result (2.12) as in (3.4) we find when $x \in [n - i, n - i + 1]$, $i = 1, 2, \cdots, n$ that $P(\sum_{i=1}^n X_i \le x)$ is given by

$$\int_0^1 \int_0^1 \cdots \int_0^1 \int_{\alpha_i}^1 \int_{\alpha_{i-1}}^1 \cdots \int_{\alpha_1}^1 n!\, y_1^{n-1} y_2^{n-2} \cdots y_{n-1}\, dy_1\, dy_2 \cdots dy_n, \tag{3.5}$$

where $\alpha_k = (n - k + 1 - x)/(1 + \sum_{j=k+1}^n y_{k+1} y_{k+2} \cdots y_j)$, $j = 1, \cdots, i$. We omit the evaluation of this integral and remark only that the required result may be obtained by using a technique similar to that used in Currie (1978, page 40 ff). We also remark that the more general results of Dempster and Kleyle (1968), Weisberg (1971), and Margolin's (1977) result for Dirichlet variables, may all be obtained in a similar fashion, though the calculations involved are rather formidable.

Closely related to the distribution of the sum in uniform samples is the distribution of $T_n = X_{(n)}/(X_{(1)} + \cdots + X_{(n)})$. The distribution of T_n for an exponential population was obtained first by Fisher (1929) using an ingenious geometrical argument. Cochran (1941) extended this result to a gamma population. Recently, Lewis and Fieller (1979) gave a recursive algorithm that obtains these results in a very neat way. The link between the distribution of T_n and the distribution of the sum in uniform samples was given by Darling (1952). Using the Laplace transform, he found that the distribution of T_n^{-1}, based on a sample from the uniform distribution on $[0, a]$, was the same as $1 + S_{n-1}$ where S_{n-1} has the distribution of the sum of the observations in a sample of size $n - 1$ from the uniform distribution on [0, 1]. This curious result has an important consequence in the present context: since we know that S_{n-1} has a distribution of the form (1.1) we can conclude that T_n^{-1} also has a distribution of this form. The result (2.12) can now be applied to give the distribution of T_n^{-1} as a repeated integral. The integrand depends on the sampled population. If we supply the integrands appropriate for uniform, exponential or gamma sampling we obtain the results of Darling, Fisher and Cochran respectively. Again it is only fair to remark that the integrations required are rather daunting.

EXAMPLE 2. In the previous example we considered linear functions of variables distributed over the simplex; we turn now to the distribution of certain quadratic forms in such variables. To be precise, we consider the distribution of the random variable X where

$$X = a_1 + \textstyle\sum_{i=1}^n a_i X_i^2 - 2 \sum_{i=1}^n a_i X_i + 2 \sum_{i=1}^n \sum_{j>i}^n a_j X_i X_j \tag{3.6}$$

where $a_1 > a_2 > \cdots > a_n > 0$ and $X_i \ge 0$, $\sum_{i=1}^n X_i \le 1$. This quadratic form has particular statistical interest since it turns out that the studentized extreme deviate, $(\bar{X} - X_{(1)})/s$ can be written in this way. We first show that, by using suitable transformations the quadratic form can be put into the form (1.1) and so the main result can be applied. Secondly, we show that $(\bar{X} - X_{(1)})/s$ does indeed have a distribution of the tvpe (1.1) by transforming it into an example of the quadratic form being discussed here. Lastly, we specialize to samples from (a) normal and (b) exponential populations and give explicit results for the distribution of $(\bar{X} - X_{(1)})/s$ in both cases.

We have the following identity for X as given by (3.6):

$$X = \sum_{i=1}^{n} b_i(1 - X_1 - X_2 - \cdots - X_i)^2,$$

where $a_i = \sum_{j=i}^{n} b_j$. We now apply the transformation (3.1) and find

$$X = \sum_{i=1}^{n} b_i Y_1^2 Y_2^2 \cdots Y_i^2 \tag{3.7}$$

which has the required form (1.1) and the distribution function is obtained.

As an illustration of this result we consider the distribution of

$$W = \frac{n(\bar{X} - X_{(1)})^2}{(n-1)S^2}$$

where W is based on a random sample $X_1, X_2, \ldots, X_n$ with $\bar{X} = (\sum X_i)/n$, $S^2 = \sum (X_i - \bar{X})^2$ and $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$. This statistic, in the equivalent form $R = (\bar{X} - X_{(1)})/S$, is widely used in detecting various forms of abnormality in the sample: for example, it can be used in the detection of outliers; in this case the sampled population is assumed to be normal. This distribution was studied first by Pearson and Chandra Sekar (1936) who obtained upper tail probabilities. Grubbs (1950) obtained an integral form for the distribution function of the statistic θ_n where $\sin^2\theta_n = 1 - W$; Grubbs evaluated his integrals numerically. Borenius (1959) effectively obtained the density function of W. When the underlying distribution is assumed to be exponential, we have the distributional problem for W, the test statistic proposed by Shapiro and Wilk (1972) as a test for the exponential distribution.

To obtain W in the required form (3.6) or (3.7) we proceed as follows. If $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are the order statistics let $G_1 = X_{(1)}$; $G_i = X_{(i)} - X_{(i-1)}$, $i = 2, \ldots, n$. Then W depends only on $G_2, G_3, \ldots, G_n$. In fact $W^{-1} = (n-1)\, Y/Z^2$ where

$$Z = n(\bar{X} - X_{(1)}) = \sum_{i=2}^{n} q_i G_i$$

and

$$Y = nS^2 = \sum_{i=2}^{n} p_i q_i G_i^2 + 2 \sum_{i=2}^{n} \sum_{j>i}^{n} p_i q_j G_i G_j$$

where $p_i = i - 1$ and $q_i = n - i + 1$. Now transform as follows:

$$U_1 = \sum_{i=2}^{n} q_i G_i; \; U_1 U_i = q_i G_i, \qquad i = 2, \ldots, n-1.$$

Then W depends only on $U_2, \ldots, U_{n-1}$ and we find

$$W^{-1} = (n-1)(n - 1 - 2n \sum_{i=2}^{N} a_i U_i + n \sum_{i=2}^{N} a_i U_i^2 + 2n \sum_{i=2}^{N} \sum_{j>i}^{N} a_j U_i U_j) \tag{3.8}$$

where $a_i = (n-i)/(n-i+1)$, $i = 2, \ldots, n-1$ and $N = n - 1$. It is a routine calculation to show that $U_2, \ldots, U_{n-1}$ are distributed over the simplex $U_i \geq 0$, $\sum_{i=2}^{n-1} U_i \leq 1$ and so the problem is essentially of the form (3.6). It is worth remarking that the transformation $\mathbf{X} \to \mathbf{G}$ which gives W in terms of $G_2, \ldots, G_n$ uses the location invariance of W; the transformation $\mathbf{G} \to \mathbf{U}$ which gives W in terms of $U_2, \ldots, U_{n-1}$ depends on the scale invariance of W.

We now have the following integral form for the distribution function of W. For $k = 1, 2, \ldots, n-2$ let

$$I_n(k) = \left[\frac{n-k-1}{(k+1)(n-1)}, \frac{n-k}{k(n-1)}\right]. \tag{3.9}$$

Then for $w \in I_n(k)$ we have

$$P(W \leq w) = \int_0^1 \cdots \int_0^1 \int_{\alpha_k}^1 \int_{\alpha_{k-1}}^1 \cdots \int_{\alpha_1}^1 f(y_2, \cdots, y_{n-1})\, dy_2\, dy_3 \cdots dy_{n-1} \tag{3.10}$$

where $\alpha_k = \psi_k(w)/A(k)$ with

$$\psi_1(w)^2 = \{(n-2)/n\}(w^{-1}-1);$$

$$\psi_k(w)^2 = \psi_1(w)^2 - (n-2)(k-1)/(n-k), \qquad k = 2, \dots, n-1$$

and

$$A(k)^2 = \phi(k)^2 + \textstyle\sum_{i=k+2}^{N} \phi(i-1)^2 y_{k+2}^2 y_{k+3}^2 \cdots y_i^2$$

with

$$\phi(k)^2 = (n-2)(n-1)/\{(n-k-1)(n-k)\}, \qquad k = 1, \dots, n-2.$$

We remark that (3.10) gives the distribution function of W whatever the underlying sampled population is. We now specialise to the exponential and normal cases. If $X_1, \dots, X_n$ is a random sample from the exponential distribution, then we may show that the joint distribution of $Y_2, \dots, Y_{n-1}$ is

$$f(y_2, \dots, y_{n-1}) = (n-2)!\, y_2^{n-3} y_3^{n-4} \cdots y_{n-2}; \qquad 0 \le y_i \le 1. \tag{3.11}$$

In this case the variables $Y_2, \dots, Y_{n-1}$ are independently distributed over the $n-2$ cube. The integral (3.10) may now be evaluated in the same way as the integral (3.5). Details are contained in Currie (1978, page 40 ff). We report only the final result. In (3.10) we first replace $\int_{\alpha_i}^1$ by $\int_0^1 - \int_0^{\alpha_i}$ and hence obtain (3.10) in terms of 2^k integrals of the form

$$\int_0^1 \cdots \int_0^1 \int_0^{\alpha_{\kappa(l)}} \int_0^1 \cdots \int_0^1 \int_0^{\alpha_{\kappa(l-1)}} \cdots \int_0^{\alpha_{\kappa(2)}} \int_0^1 \cdots \int_0^1 \int_0^{\alpha_{k(1)}} \int_0^1 \cdots \int_0^1 f(\mathbf{y})\, d\mathbf{y} \tag{3.12}$$

$$\leftarrow k_{l+1} \rightarrow \leftarrow \quad k_l \quad \rightarrow \qquad \leftarrow \quad k_2 \quad \rightarrow \leftarrow \quad k_1 \quad \rightarrow$$

where $\kappa(i) = \sum_{j=1}^{i} k_j$ and $k_1, k_2, \dots, k_l$ are any positive integers such that $\sum_{j=1}^{l} k_j \le k$ and $k_{l+1} = n - 2 - \kappa(l)$. We denote the integral (3.12) by

$$J_n(k_1, k_2, \dots, k_l) = J_n(\mathbf{k}_l). \tag{3.13}$$

Now define

$$T_k(l) = \{\mathbf{k}_l : \mathbf{k}_l' = (k_1, k_2, \dots, k_l);\ k_i \ge 1, \textstyle\sum_{j=1}^{l} k_j \le k\}.$$

Then the integral (3.10) is given by

$$1 + \textstyle\sum_{l=1}^{k} (-1)^l \sum_{k_l \in T_k(l)} J_n(\mathbf{k}_l) \tag{3.14}$$

with the density function f given by (3.11). The integral (3.12) can be evaluated as a product of l functions and it is worth noting that this factorisation depends critically on the factorisation of α_i remarked on earlier.

For $n \ge 4$, $k = 1, 2, \dots, n-3$ and $x \ge 0$ we define

$$\begin{aligned} G_{n,0}(x) &= 1 \\ G_{n,k}(x) &= \int_0^{\tan^{-1} x} G_{n,k-1}\left(\frac{n-k+1}{n-k-1} \sec\theta\right) \sin^{n-k-3}\theta\, d\theta. \end{aligned} \tag{3.15}$$

It is interesting to note that the functions $G_{n,k}(x)$ are very similar to the functions $\beta_p(r)$ used by Borenius (1959) in deriving the density function of W in the normal case. We note that $G_{n,k}(x)$ is essentially a k dimensional integral. For any $\mathbf{k}_l' = (k_1, \dots, k_l)$ let $\kappa = \sum_{j=1}^{l} k_j$ and

$$\gamma_n(\mathbf{k}_l) = \frac{2(n-k_1)^{1/2}}{\Gamma((n-\kappa-1)/2)} \cdot \frac{(n-k_1-1)!\,(n-k_1-2)!}{(n-\kappa)!} \cdot \frac{\pi^{(n-\kappa-1)/2}}{\{(n-1)(n-2)\}^{(n-k_1-1)/2}},$$

then

$$J_n(\mathbf{k}_l) = \gamma_n(\mathbf{k}_l)\psi_{k_1}(w)^{n-k_1-1} \prod_{i=1}^{l-1} G_{n-\kappa(i)+1,k_{i+1}}\left(\frac{\psi_{\kappa(i+1)}(w)}{\phi\{\kappa(i+1)-1\}}\right). \tag{3.16}$$

Note that we have adopted the convention that a product Π_k^l with $k > l$ is identically one.

We recall that $G_{n,k}(x)$ is effectively a k-dimensional integral. This limits the practical usefulness of the above result but exact percentage points for W have been found for small values of n and for part of the range of W for larger values of n. These results are reported in Table 1. We remark that the original values of Shapiro and Wilk (1972) obtained by simulation agree with these exact results within the limits of their experimental error.

We now specialise in (3.10) to the case when the underlying sampled population is normal. We may then show that the joint distribution of $Y_2, Y_3, \ldots, Y_{n-1}$ is

$$f(y_2, \ldots, y_{n-1}) = c_n \frac{y_2^{n-3}y_3^{n-4} \cdots y_{n-2}}{\{1 + n(n-1)\sum_{i=2}^{N} b_i y_2^2 y_3^2 \cdots y_i^2\}^{N/2}} \tag{3.17}$$

with $b_i = \{(n-i)(n-i+1)\}^{-1}$, $N = n - 1$ and

$$c_n = \tfrac{1}{2}\, n^{n/2}(N/\pi)^{N/2}\Gamma(N/2)$$

over the region $0 \leq y_i \leq 1$, $i = 2, \ldots, n-1$.

In this integral form, the result is equivalent to that obtained by Grubbs (1950). We can however evaluate the integral in terms of the function $G_{n,k}(x)$ defined in (3.15); for details

TABLE 1
The Percentage Points of W-exponential

n	0.5	1.0	2.5	5.0	10.0	50.0	90.0	95.0	97.5	99.0	99.5
3	.2519	.2538	.2596	.2697	.2915	.5714	.9709	.9926	.9981	.9997	.99993
4	.1242	.1302	.1433	.1604	.1891	.3768	.7514	.8581	.9236	.9680	.9837
5	.0831	.0899	.1036	.1198	.1449	.2868	.5564	.6657	.7597	.8534	.9024
6	.0638	.0704	.0831	.0973	.1176	.2284	.4341	.5204	.6054	.7081	.7743
7	.0528	.0588	.0702	.0823	.0992	.1896	.3541	.4211	.4897	.5806	.6463
8	.0455	.0510	.0611	.0716	.0860	.1617	.2960	.3509	.4062	.4815	.5392
9			.0544	.0636	.0762	.1408	.2528	.2983	.3442	.4065	.4550
10				.0574	.0685	.1245	.2197	.2579	.2965	.3488	.3897
11					.0624	.1115	.1935	.2261	.2590	.3035	.3382
12						.1010	.1725	.2007	.2289	.2671	.2969
13						.0922	.1553	.1799	.2044	.2376	.2634
14						.0848	.1409	.1626	.1842	.2132	.2357
15						.0784	.1288	.1481	.1672	.1928	.2127
16							.1185	.1358	.1528	.1756	.1932
17							.1096	.1252	.1405	.1609	.1766
18							.1018	.1160	.1299	.1483	.1624
19							.0950	.1079	.1205	.1372	.1500
20							.0890	.1008	.1124	.1276	.1392
21							.0837	.0946	.1052	.1191	.1297
22								.0890	.0987	.1115	.1212
23									.0930	.1048	.1137
24										.0987	.1070
25										.0933	.1010
26										.0884	.0955
27											.0906
28											.0860
29											.0819
30											.0781

of this evaluation see Currie (1978; page 64 ff). We denote the integral (3.12) by $J_n^*(\mathbf{k}_l)$ when the density function (3.17) is used. Let

$$\delta_n(\mathbf{k}_l) = \frac{n!}{(n-\kappa)!} \cdot \frac{\Gamma((n-1)/2)}{\Gamma((n-\kappa-1)/2)} \cdot \pi^{-\kappa/2}$$

Then

$$J^*(\mathbf{k}_l) = \delta_n(\mathbf{k}_l) \textstyle\prod_{i=0}^{l-1} G_{n-\kappa(i)+1,k_{i+1}} \left(\dfrac{\psi_{\kappa(i+1)}(w)}{\phi(\kappa(i+1)-1)} \right) \tag{3.18}$$

where $\kappa(0) = 0$.

The distribution function of W in the normal case is now given by (3.14) with $J_n(\mathbf{k}_l)$ replaced by $J_n^*(\mathbf{k}_l)$. It is interesting that the formulae (3.16) and (3.18) for $J_n(\mathbf{k}_l)$ and $J_n^*(\mathbf{k}_l)$ each consist of a product of l functions of w, $l-1$ of which are identical. This result for the distribution function of W is equivalent to the result of Borenius (1959) for the density function of W.

4. Proof of the theorem. First, we define

$$S(x) = \{\mathbf{y}' = (y_1, y_2, \ldots, y_n) : g(\mathbf{y}) \geq x, 0 \leq y_i \leq 1, 1 \leq i \leq n\} \tag{4.1}$$

and observe that $S(x)$ is the set over which we require to integrate the density $f(\mathbf{y})$ in order to evaluate the required probability. The set $S(x)$ is given implicitly by the condition $g(\mathbf{y}) \geq x$; we show first how to construct a set $S_k(x)$ for $x \in I_n(k)$ that will turn out to be $S(x)$. The construction of $S_k(x)$ is aimed at producing the set of points in the region of integration in the integral (2.12). We choose the coordinates of the points $\mathbf{y}$ to be included in $S_k(x)$ in the order y_n, then $y_{n-1}, \ldots$, and finally y_1. We need to check that if $y_n, y_{n-1}, \ldots, y_j$ have been chosen using the construction where $j \leq k$, then $0 \leq \alpha_{j-1} \leq 1$.

Suppose that $x \in I_n(k)$ for some $k = 1, 2, \ldots, n-1$ where $n \geq 2$. Let $y_i \in [0, 1]$, $i = k+1, \ldots, n$. Let $\alpha_k = \alpha_k(n; x; y_{k+1}, \ldots, y_n)$ be defined as in (2.10). Then

$$\alpha_k = \frac{x - b_{k-1}}{A(k)} \leq \frac{b_k - b_{k-1}}{A(k)} = \frac{a_k}{a_k + y_{k+1}A(k+1)} \leq 1.$$

Clearly, $\alpha_k \geq 0$ and so $0 \leq \alpha_k \leq 1$ as required. Similarly, if $x \in I_n(n)$ then $0 \leq \alpha_n \leq 1$.

Assume now that $y_n, y_{n-1}, \ldots, y_{n-l}$ have been chosen and are such that $\alpha_{n-l-1} = \alpha_{n-l-1}(n; x; y_{n-l}, \ldots, y_n)$ is defined and $0 \leq \alpha_{n-l-1} \leq 1$ for some l where $0 \leq l \leq n-3$, $n \geq 3$. Let $y_{n-l-1} \in [\alpha_{n-l-1}, 1]$. Then we have

$$\begin{aligned} \alpha_{n-l-1} \geq 0 &\Rightarrow x - b_{n-l-2} \geq 0 \quad \text{by (2.10)} \\ &\Rightarrow x - b_{n-l-3} \geq a_{n-l-2} > 0 \quad \text{by (2.9)} \\ &\Rightarrow \alpha_{n-l-2} > 0 \quad \text{by (2.10).} \end{aligned}$$

We also have, using (2.10), (2.9) and (2.8) that

$$\begin{aligned} \alpha_{n-l-2} &= \frac{x - b_{n-l-2} + a_{n-l-2}}{a_{n-l-2} + y_{n-l-1}A(n-l-1)} \\ &= \frac{a + \alpha_{n-l-1}}{a + y_{n-l-1}} \quad \text{where} \quad a = \frac{a_{n-l-2}}{A(n-l-1)} > 0 \\ &\leq 1 \qquad \text{since} \quad y_{n-l-1} \geq \alpha_{n-l-1} \geq 0. \end{aligned}$$

And so $0 < \alpha_{n-l-2} \leq 1$ as required. In a similar way if α_n is defined and $0 \leq \alpha_n \leq 1$ where $n \geq 2$, and if $y_n \in [\alpha_n, 1]$ then $0 < \alpha_{n-1} \leq 1$.

These results enable us to define the sets $S_k(x)$ that will turn out to give the required set $S(x)$.

DEFINITION. (a) Let $x \in I_n(k)$, for some $k = 1, 2, \ldots, n - 1$. The set $S_k(x)$ consists of all the points $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$ constructed as follows: for $i = k + 1, \ldots, n$ choose any $y_i \in [0, 1]$. Now form α_k using (2.10). We have just proved that $[\alpha_k, 1] \subseteq [0, 1]$. Choose any $y_k \in [\alpha_k, 1]$ and form α_{k-1}. Again we know $[\alpha_{k-1}, 1] \subseteq [0, 1]$. We may continue in this fashion choosing $y_j \in [\alpha_j, 1]$ and forming α_{j-1} in the order $j = k - 1, k - 2, \ldots, 2$. At each stage we are ensured that $[\alpha_{j-1}, 1] \subseteq [0, 1]$ and the choice $y_{j-1} \in [\alpha_{j-1}, 1]$ can be made. Lastly, choose any $y_1 \in [\alpha_1, 1]$.

(b) If $x \in I_n(n)$ then $S_n(x)$ consists of all the points $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$ constructed as follows: we know that $0 \leq \alpha_n \leq 1$ and so we may choose any $y_n \in [\alpha_n, 1]$ and form α_{n-1}. Then $[\alpha_{n-1}, 1] \subseteq [0, 1]$ and the construction of $S_n(x)$ is completed as in (a).

We can now prove that the set of interest $S(x)$ is given by $S_k(x)$ whenever $x \in I_n(k)$. We begin by proving the following: if $x \geq b_{k-1}$ then $S(x) \subseteq S_k(x)$. Suppose $\mathbf{y} \in S(x)$. We must show that $y_i \in [\alpha_i, 1]$ for $i = 1, 2, \ldots, k$. We prove this by induction on k.

Case 1. $k = 1$. If $\mathbf{y} \in S(x)$ then by definition $x \leq g(\mathbf{y}) = y_1 A(1)$ and so $y_1 \geq x/A(1) = \alpha_1$. Since we also have $y_1 \leq 1$ we have $y_1 \in [\alpha_1, 1]$ as required.

Case 2. We assume the result is true for $k = m < n$. We further assume that $x \geq b_m$; hence $x - b_{m-1} \geq a_m > 0$ and so by induction $y_i \in [\alpha_i, 1]$, $i = 1, \ldots, m$. Further

$$\alpha_m \leq 1$$

yields

$$\frac{x - b_{m-1}}{a_m + y_{m+1}A(m + 1)} \leq 1$$
$$\Rightarrow x - b_{m-1} - a_m \leq y_{m+1}A(m + 1)$$
$$\Rightarrow y_{m+1} \geq \frac{x - b_m}{A(m + 1)} = \alpha_{m+1}$$

since we have assumed $x - b_m \geq 0$. Since $y_{m+1} \leq 1$ we have $y_{m+1} \in [\alpha_{m+1}, 1]$ as required.

Now suppose that $x \in I_n(k)$. Then $x \geq b_{k-1}$ and so we have $S(x) \subseteq S_k(x)$. Conversely, suppose that $\mathbf{y} \in S_k(x)$. Then by the construction of $S_k(x)$ we have $y_1 \in [\alpha_1, 1]$. Thus by (2.10) $y_1 \geq (x - b_0)/A(1) \Rightarrow y_1 A(1) \geq x$ since $b_0 = 0$; and so by (1.1) and (2.7) $g(\mathbf{y}) \geq x$ and this is the condition (4.1) that $\mathbf{y} \in S(x)$. Thus $S_k(x) \subseteq S(x)$ and so $x \in I_n(k)$ implies that $S(x) = S_k(x)$.

We can now prove the main result. We assume $x \in I_n(k)$ and so $P(g(\mathbf{Y}) \geq x)$ is given by

$$\int_{S(x)} \cdots \int f(\mathbf{y})\, d\mathbf{y} = \int_{S_k(x)} \cdots \int f(\mathbf{y})\, d\mathbf{y}$$

By the construction of $S_k(x)$ we may replace the multiple integral over $S_k(x)$ with the repeated integral in (2.12) and this proves the result.

5. Discussion. In Section 2.1 we noted the connection between the changing form of the density function of $g(\mathbf{Y})$ and the geometry of a surface passing through a cube. We now give a statistical interpretation of this for one of our examples. Pearson and Chandra Sekar (1936) obtained the upper tail of the probability density function of the extreme studentized deviate from a normal population as follows: let $x_1, x_2, \ldots, x_n$ be a series of observed values of a variable x. Let $\bar{x} = \sum x_i/n$, $s^2 = \sum (x_i - \bar{x})^2/n$ and $\tau_i = (x_i - \bar{x})/s$. Let $\tau_{(1)} \leq \tau_{(2)} \leq \cdots \leq \tau_{(n)}$ be the ordered values of $\tau_1, \tau_2, \ldots, \tau_n$. Pearson and Chandra Sekar show that max $\tau_{(n-1)} <$ max $\tau_{(n)}$ and they deduce from the general relation $nf(\tau) = \sum f_i(\tau)$ that, for $\tau >$ max $\tau_{(n-1)}$, $f_n(\tau) = nf(\tau)$ where $f_j(\tau)$ represents the (unknown) probability density function of $\tau_{(j)}$ and $f(\tau)$ the (known) probability density function of the unordered τ_i. It is simple to verify that the value of max $\tau_{(n-1)}$ coincides with the left end of the interval $I_n(1)$ defined in (3.9). This correspondence extends as follows: we may show as in Currie (1978, page 77 ff) that max $\tau_{(i)} = \{(i-1)/(n-i+1)\}^{1/2}$; it is now a simple matter to verify that this value coincides with the left hand end of the interval $I_n(n-i)$. We do not have to look far for an explanation of this apparent set of coincidences. From the relation $nf(\tau) = \sum f_i(\tau)$ we have $nf(\tau) = f_n(\tau)$ for $\tau \geq$ max $\tau_{(n-1)}$ and $nf(\tau) = f_n(\tau) + f_{n-1}(\tau)$ for max $\tau_{(n-1)} \geq \tau \geq$ max $\tau_{(n-2)}$. Now $f(\tau)$ has a single functional form over its entire range, at least in sampling from a normal population. Hence, for $f(\tau)$ to maintain its single form, $f_n(\tau)$ must change form at exactly the point that $f_{n-1}(\tau)$ becomes nonzero, and an obvious extension of the argument provides the general explanation.

REFERENCES

Borenius, G. (1959). On the distribution of the extreme values in a sample from a normal distribution. *Skand. Actuarietidskr.* **42** 131-166.

Cochran, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Ann. Eugenics* **11** 47-52.

Currie, I. D. (1978). The distribution of extreme studentized deviates in normal and exponential samples. Ph.D. thesis. Heriot-Watt University, Edinburgh.

Darling, D. A. (1952). On a test for homogeneity and extreme values. *Ann. Math. Statist.* **23** 450-456.

Dempster, A. P. and Kleyle, R. M. (1968). Distributions determined by cutting a simplex with hyperplanes. *Ann. Math. Statist.* **39** 1473-1478.

Fisher, R. A. (1929). Tests of significance in harmonic analysis. *Proc. Roy. Soc. A* **125** 54-59.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Ann. Math. Statist.* **21** 27-58.

Hall, P. (1927). The distribution of means for samples of size N drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika* **19** 240-244.

Lewis, T. and Fieller, N. R. J. (1979). A recursive algorithm for null distributions for outliers: 1. gamma samples. *Technometrics* **21** 371-376.

Margolin, B. H. (1977). The distribution of internally studentized statistics via Laplace transform inversion. *Biometrika* **64** 573-582.

Pearson, E. S. and Chandra Sekar, C. (1936). The efficiency of statistical tools and a criteria for the rejection of outlying observations. *Biometrika* **28** 308-320.

Shapiro, S. S. and Wilk, M. B. (1972). An analysis of variance test for the exponential distribution (complete samples). *Technometrics* **14** 355-370.

Weisberg, H. (1971). The distribution of linear combinations of order statistics from the uniform distribution. *Ann. Math. Statist.* **42** 704-709.

Department of Actuarial Mathematics and Statistics
Heriot-Watt University
Edinburgh EH14 4AS, Scotland