MULTIVARIATE STATISTICAL OUTLIERS*

By S. S. WILKS

Princeton University

SUMMARY. This paper deals with the problem of identifying and testing a candidate set of a small number t of extreme sample elements as significant outliers in a sample of size n from a k-dimensional normal distribution with unknown parameters. The problem is considered in detail for $t=1, 2, 3, 4$, that is, for sets of 1, 2, 3, and 4 outliers. The criterion for indentifying and testing a single observation as a significant outlier is r_1 as defined in Section 3(b) and that for a pair of outliers is r_2 as defined in Section 4, small values of r_1 or r_2 being critical values. In the absence of exact values for the extremely complicated probabilities $P(r_1<r)$ and $P(r_2<r)$ upper bounds for these probabilities are given by (2.15) and (3.2) respectively. These upper bounds are suggested for *a fortiori* significance testing of observed values of r_1 and r_2. Some evidence of the closeness of these upper bounds obtained for the probabilities $P(r_1<r)$ and $P(r_2<r)$ is given in Table 1 for $k=1$, that is, for a sample from a one-dimensional normal distribution. In this case exact values of r_α for which $P(r_1<r_\alpha)=\alpha$ are available from Grubbs' (1950) tables for certain values of α. These are compared with the upper bounds of $P(r_1<r_\alpha)$ for several values of n in Table 1.

Values of r_α for which the upper bound of $P(r_1<r_\alpha)$ has the value α are given in Table 2 for $\alpha = 0.010, 0.025, 0.050, 0.100$; $k=1, 2, 3, 4, 5$; and $n=5(1)30(5)100(100)500$. Table 3 gives values of $\sqrt{r_\alpha}$ for which the upper bound of $P(r_2<r_\alpha)$ has the value α for the same values of α, k and n.

Extension of r_1 and r_2 to the case of t outliers is r_t as defined in Section 5. Expressions are given for the cases $t=3$ and 4 from which values of r_α can be determined so that the upper bound of $P(r_t<r_\alpha)$ is α. No tabulations have been made, however, for the cases of three and four outliers.

In the more general problem of t outliers a procedure is outlined as to how one could obtain the value of r_α for which the upper bound of $P(r_t<r_\alpha)$ has the value α.

1. Introduction

Studies of criteria for the rejection of extreme observations as significant outliers in a single sample from a one-dimensional normal distribution with unknown parameters have been made by various authors during the last thirty years.

If $(x_1, \ldots, x_n)$ is a sample from such a distribution and if $\bar{x}$ and s^2 are the sample mean and sample variance, Thompson (1935) has determined the distribution of $(x_\xi-\bar{x})/s$ for an arbitrary ξ. He has proposed that for a given α, values of x_ξ for which $|x_\xi-\bar{x}|/s > \tau_\alpha$ be rejected as significant outliers in a sample from a normal distribution where τ_α is chosen so that for any ξ, $P(|x_\xi-\bar{x}|/s>\tau_\alpha)=\alpha$. He determined τ_α for $\alpha=\frac{0.05}{n}, \frac{0.10}{n}, \frac{0.20}{n}$ and for $n=3(1)22, 32, 42, 102, 202, 1002$. Thus, for instance, if $\alpha=\frac{0.10}{n}$ the expected number of observations which would be falsely rejected as outliers, (that is, would be rejected if all elements of the sample were actually from the same normal distribution) would be 1 per 10 samples of size n.

* Research partially supported by the Office of Naval Research while the author was a Fellow of the Center for Advanced Study in the Behavioral Sciences in the Fall of 1961. Presented at the International Congress of Mathematicians, Stockholm, August 20, 1962.

Pearson and Chandra Sekar (1936) considered $(x_{(n)}-\bar{x})/s$ and $(\bar{x}-x_{(1)})/s$ as criteria for rejecting individual observations as significantly high and low outliers respectively, where $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ are the order statistics of the sample. In particular, they showed that the upper tail of the distribution of $(x_{(n)}-\bar{x})/s$ (or of $(\bar{x}-x_{(1)})/s$) has a density function $xf_n(\tau)$ on the interval $(\sqrt{(n-2)/2}, \sqrt{n-1})$, where $f_n(\tau)$ is the probability density function of $(x_\xi-\bar{x})/s = \tau$, say. From this fact they found the upper 1%, 2.5% and 10% points of the distribution of $(x_{(n)}-\bar{x})/s$ (or of $(\bar{x}-x_{(1)})/s$) for values of n ranging from 11 to 19, that is, for all values of n such that the specified upper percentage point falls in the interval $(\sqrt{(n-2)/2}, \sqrt{n-1})$.

Grubbs (1950) extended the work of Pearson and Chandra Sekar (1936) for individual outliers by actually determining the distribution of $(x_{(n)}-\bar{x})/s$ (or of $(\bar{x}-x_{(1)}/s)$ in a sample from a normal distribution with unknown parameters. He tabulated the upper 1%, 2.5%, 5% and 10% points of the distribution of $(x_{(n)}-\bar{x})/s$ (or of $(\bar{x}-x_{(1)})/s$) for all $n \leqslant 25$. He also tabulated the lower 1 %, 2.5%, 5 % and 10% points of the distribution of $\sum_{\xi=1}^{n-1}(x_{(\xi)}-\bar{x}_n)^2/[(n-1)s^2]$ (or of $\sum_{\xi=2}^{n}(x_{(\xi)}-\bar{x}_1)^2/[(n-1)s^2]$) for all $n \leqslant 25$ where $\bar{x}_n$ is the mean of $x_{(1)}, \ldots, x_{(n-1)}$ and $\bar{x}_1$ is the mean of $x_{(2)}, \ldots, x_{(n)}$. Grubbs also considered the case of two high (or two low) outliers, using as the criterion of rejection $\sum_{\xi=1}^{n-2}(x_{(\xi)}-\bar{x}_{n,\,n-1})/[(n-2)s^2]$(or $\sum_{\xi=3}^{n}(x_{(\xi)}-\bar{x}_{1,2})/[(n-2)s^2]$) where $\bar{x}_{n,\,n-1}$, is the mean of $x_{(1)}, \ldots, x_{(n-2)}$ and $\bar{x}_{1,2}$ is the mean of $x_{(\xi)}, \ldots, x_{(n)}$. He tabulated the lower 1%, 2.5%, 5% and 10% points of the distribution of these quantities for all $n \leqslant 20$. He mentioned, but did not go into the details of $\sum_{\xi=2}^{n-1}(x_{(\xi)}-\bar{x}_{1,n})^2/[(n-2)s^2]$ as a two-outlier test, where $\bar{x}_{1,n}$ is the mean of $x_{(2)}, \ldots, x_{(n-1)}$.

Dixon (1951) has considered ratios of form $(x_{(n)}-x_{(n-j)})/(x_{(n)}-x_{(i)})$ [or $(x_{(j+1)}-x_{(1)})/(x_{(n-i+1)}-x_{(1)})$], $i = 1, 2, 3$; $j = 1, 2$, as criteria for testing extreme observations as outliers and he has tabulated the 0.5%, 1%, 2%, 5%, 10(10)90%, 95% points of the distributions of these quantities. Dixon (1950) has also studied the power functions of all of the criteria mentioned above against alternatives in which it is assumed that the outliers are from normal distributions of form $N(\mu+\lambda\sigma, \sigma^2)$ or $N(\mu, \lambda^2\sigma^2)$ for various values of λ and for unknown μ and σ^2.

All of the studies mentioned above deal with the problem of testing one or two extreme observations as significant outliers in a sample from a one-dimensional normal distribution with unknown parameters.

Problems of outliers in samples from normal distributions for which one or both of the parameters are known or are estimated from independent samples have been considered by various authors, including Irwin (1925), McKay (1935), Newman (1940), Pearson and Hartley (1942), Nair (1948, 1952), David (1956), Pillai and Tienzo (1959) and Pillai (1959). Rider (1932) has given a survey of the literature on outliers prior to 1932.

The purpose of the present paper is to discuss in detail and present tables for the problem of selecting and testing one or two extreme observations as significant outliers in a sample from a multivariate normal distribution, with unknown parameters. The mathematical theory of selecting and testing three or more extreme observations as significant outliers is discussed, but no tables are given.

No attempt has been made to study the power of the outlier tests discussed in this paper under various possible alternatives to the null hypothesis that all of the elements of the sample are independently drawn from a common k-dimensional normal distribution with unknown parameters. This would be a much more extensive investigations than the study of the tests presented in this paper under the null hypothesis. Such a study remains to be done. Some of the power properties of a test equivalent to r_1 the test for the problem of one-outlier, have been investigated by Karlin and Truax (1960), and by Ferguson (1961).

2. The case of a single outlier

(a) *The one-outlier scatter ratios of a sample.* Let $(x_{1\xi}, \ldots, x_{k\xi}\ ;\ \xi = 1, \ldots, n)$ be a sample of size n from a k-dimensional normal distribution $N(\{\mu_i\}, \|\sigma_{ij}\|)$ where $\{\mu_i\}$ is the vector of means $(\mu_1, \ldots, \mu_k)$ and $\|\sigma_{ij}\|$ is the covariance matrix of the distribution. It is assumed that the vector of means and covariance matrix of the distribution are unknown. Let $(\bar{x}_1, \ldots, \bar{x}_k)$ be the vector of sample means, where $n\bar{x}_i = \sum\limits_{\xi=1}^{n} x_{i\xi}$ and let

$$a_{ij} = \sum_{\xi=1}^{n} (x_{i\xi} - \bar{x}_i)(x_{j\xi} - \bar{x}_j), \qquad i, j = 1, \ldots, k. \qquad \ldots \quad (2.1)$$

The sample can be represented as a cluster of n points in a k-dimensional euclidean space R_k. Any k of these n points together with the sample center of gravity point $(\bar{x}_1, \ldots, \bar{x}_k)$ forms a simplex. If the volume of this simplex is squared and if the sum of squares is taken of the volumes of all possible simplexes which can be formed in this manner, it can be shown (see Wilks, 1962, for instance) that this sum of squared volumes is

$$(k!)^{-2}|a_{ij}|, \qquad \ldots \quad (2.2)$$

where $|a_{ij}|$ is the determinant of the matrix $\| a_{ij} \|$. It is convenient to call $|a_{ij}|$ the *internal scatter* of the sample $(x_{1\xi}, \ldots, x_{k\xi}\ ;\ \xi = 1, \ldots, n)$; if $n > k$, $|a_{ij}| > 0$ with probability 1.

If we delete the ξ-th element of the sample we obtain a cluster of $n-1$ points in R_k. Let the internal scatter of these $n-1$ points be $|a_{ij\xi}|$ which will be > 0 with probability 1 if $n > k+1$.

Let

$$R_\xi = \frac{|a_{ij\xi}|}{|a_{ij}|}, \quad \xi = 1, \ldots, n. \qquad \ldots \quad (2.3)$$

The quantities $R_1, \ldots, R_n$ will be called *one-outlier scatter ratios* of the sample $(x_{1\xi}, \ldots, x_{k\xi}\ ;\ \xi = 1, \ldots, n)$.

It can be verified that

$$a_{ij\xi} = a_{ij} - b_{i\xi} b_{j\xi} \qquad \dots \quad (2.4)$$

where

$$b_{i\xi} = \sqrt{\frac{n}{n-1}}\,(x_{i\xi} - \bar{x}_i).$$

Thus we have

$$|a_{ij\xi}| = |a_{ij} - b_{i\xi}\, b_{\xi}| = |a_{ij}|\,.\,[1 - \sum_{i,j=1}^{k} a^{ij} b_{i\xi} b_{j\xi}], \qquad \dots \quad (2.5)$$

where

$$\| a^{ij} \| = \| a_{ij} \|^{-1}.$$

Hence

$$R_{\xi} = 1 - \sum_{i,j=1}^{k} a^{ij} b_{i\xi} b_{j\xi}$$

and since

$$\sum_{\xi=1}^{n} a^{ij} b_{i\xi} b_{j\xi} = \frac{nk}{(n-1)},$$

we have

$$\sum_{\xi=1}^{n} R_{\xi} = n\left(1 - \frac{k}{n-1}\right). \qquad \dots \quad (2.6)$$

Now, it is known in multivariate statistical analysis (see Wilks (1962), for example) that for any ξ the ratio R_{ξ} has the beta distribution $B_e\left(\frac{n-k-1}{2}, \frac{k}{2}\right)$, where a random variable z is said to have the beta distribution $B_e(\nu_1, \nu_2)$ if the probability density function of z is

$$f(z) = \frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)}\, z^{\nu_1 - 1}\ (1-z)^{\nu_2 - 1} \qquad \dots \quad (2.7)$$

on the interval (0, 1) and $f(z) = 0$ outside the interval.

Under the null hypothesis (that is, assuming that all elements of the sample are independently drawn from a common k-dimensional normal distribution), the one-outlier scatter ratios $R_1, \dots, R_n$ are random variables having a distribution which is symmetric over the n-dimensional space of $R_1, \dots, R_n$ for which

$$R_1 + \dots + R_n = n\left(1 - \frac{k}{n-1}\right) \qquad \dots \quad (2.8)$$

$$0 \leqslant R_{\xi} \leqslant 1, \qquad \xi = 1, \dots, n,$$

where the (marginal) distribution of each R_{ξ} is identical with the distribution of a random variable u having the beta distribution $B_e\left(\frac{n-k-1}{2}, \frac{k}{2}\right)$.

(b) *The ordered values of one-outlier scatter ratios.* Let $R_{(1)} < \dots < R_{(n)}$ be the ordered values of $R_1, \dots, R_n$. The criterion we propose for selecting and testing a single extreme observation as a significant outlier is $R_{(1)}$ which we shall denote by r_1. In other words the strongest candidate for being a significant outlier is identified as the sample element whose deletion gives the scatter ratio $r_1 = \min_{\xi} \{R_{\xi}\}$. It is the one to be tested as a significant outlier, with r_1 being the test criterion. It is evident that the critical values of r_1 are those in the left tail of its distribution.

The joint distribution of $R_{(1)}, \ldots, R_{(n)}$, or even of $R_1, \ldots, R_n$ for that matter, is very complicated. However, one can readily obtain moments of any one of the random variables $R_1, \ldots, R_n$ and also certain low joint moments of two or more of these random variables. For instance,

$$\mathcal{E}(R_\xi) = \frac{n-k-1}{n-1}, \ \operatorname{var}(R_\xi) = \frac{2k(n-k+1)}{(n-1)^2(n+1)}$$

$$\operatorname{cov}(R_\xi, R_\eta) = -\frac{2k(n-k+1)}{(n-1)^3(n+1)}. \qquad \ldots \quad (2.9)$$

Even though it does not appear feasible to determine exact percentage points in the lower tail of the distribution of r_1, except for $k = 1$ and then only for small values of n as we shall see later, we can determine upper bounds for the amount of probability in the lower tail of the distribution of r_1 which should be useful, at least for small values of k and small percentage points, for *a fortiori* significance testing of r_1.

First let us examine the lower limits of the ranges of $R_{(1)}, \ldots, R_{(n)}$. If we consider the space of $(R_1, \ldots, R_n)$ remembering that $R_1, \ldots, R_n$ must each lie on the interval (0, 1) it will be seen that not more than $n-k-1$ of the R's, in the set $\{R_1, \ldots, R_n\}$ can be 1 simultaneously. For if this were possible the average of the remaining R's in this set would be negative. This means that $R_{(1)}$ would be negative contrary to the fact that each R in the set $\{R_1, \ldots, R_n\}$ must lie on the interval (0, 1). Thus if $n-k-1$ R's in the set $\{R_1, \ldots, R_n\}$ are simultaneously equal to 1, we would have $R_{(1)}+\cdots+R_{(k+1)} = k+1-\frac{nk}{n-1}$ and hence the average of $R_{(1)}, \ldots, R_{(k+1)}$ would be $1-\frac{nk}{(k+1)(n-1)}$ which implies that $R_{(k+1)} \geqslant 1-\frac{nk}{(k+1)(n-1)}$. Similarly, if we put $n-k-2$ of the R's in the set $\{R_1, \ldots, R_n\}$ equal to 1, it will be seen that $R_{(k+2)} \geqslant 1-\frac{nk}{(k+2)(n-1)}$. Continuing this process, if we put only one R in the set $\{R_1, \ldots, R_n\}$ equal to 1, we obtain

$$R_{(n)} \geqslant 1-\frac{k}{n-1}.$$

Note that it is possible for $R_{(1)}, \ldots, R_{(k)}$ to be 0 simultaneously, in which case $R_{(k+1)} = 1-\frac{nk}{(k+1)(n-1)}$. Therefore for left-hand end points of the distribution of $R_{(1)}, \ldots, R_{(n)}$ we have :

$$\begin{aligned} R_{(1)} &\geqslant 0, \ldots, R_{(k)} \geqslant 0 \\ R_{(k+1)} &\geqslant 1-\frac{nk}{(k+1)(n-1)} \\ R_{(k+2)} &\geqslant 1-\frac{nk}{(k+2)(n-1)} \\ &\vdots \\ R_{(n)} &\geqslant 1-\frac{k}{n-1}. \end{aligned} \qquad \ldots \quad (2.10)$$

(c) *Upper bound for* $P(r_1 < r)$. For a fixed number r let us consider the problem of finding an upper bound for $P(r_1 < r)$. Let $E_1, \ldots, E_n$ denote the events for which $R_1 < r, \ldots, R_n < r$ respectively. Then

$$P(r_1 < r) = P(E_1 \, U \cdots \, UE_n). \qquad \ldots \quad (2.11)$$

But
$$P(E_1 \, U \cdots \, UE_n) \leqslant P(E_1)+\cdots+P(E_n), \qquad \ldots \quad (2.12)$$

and
$$P(E_1) = \cdots = P(E_n) = P(u < r), \qquad \ldots \quad (2.13)$$

where
$$P(u < r) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-k-1}{2}\right)\Gamma\left(\frac{k}{2}\right)} \int_0^r u^{\frac{n-k-1}{2}-1}(1-u)^{\frac{k}{2}-1} du, \qquad \ldots \quad (2.14)$$

since, as stated in Section 2(a), u is a random variable having the beta distribution $B_e\left(\frac{n-k-1}{2}, \frac{k}{2}\right)$.

Therefore
$$P(r_1 < r) \leqslant nP(u < r), \qquad \ldots \quad (2.15)$$

that is, $nP(u < r)$ *is an upper bound for* $P(r_1 < r)$. In particular, if we choose $r = r_\alpha$ so that $nP(u < r_\alpha) = \alpha$ we obtain

$$P(r_1 < r_\alpha) \leqslant \alpha. \qquad \ldots \quad (2.16)$$

(d) *The upper bound* $nP(u < r)$ *as the expected number of scatter ratios with values* $< r$. The quantity $nP(u < r)$ has another useful interpretation. Suppose δ_ξ is a random variable which has the value 1 if $R_\xi < r$ and 0 otherwise, $\xi = 1, \ldots, n$. Let

$$N(r) = \sum_{\xi=1}^{r} \delta_\xi, \qquad \ldots \quad (2.17)$$

that is, $N(r)$ is the number of the one-outlier scatter ratios which have values less than r. We have

$$\mathcal{E}(N(r)) = \mathcal{E}(\delta_1)+\ldots+\mathcal{E}(\delta_n) = nP(u < r) \qquad \ldots \quad (2.18)$$

since $\mathcal{E}(\delta_\xi) = P(u < r)$, $\xi = 1, \ldots, n$. Thus, *the expected number of the one-outlier scatter ratios* $R_1, \ldots, R_n$ *having values less than r is equal to the upper bound* $nP(u < r)$ *of* $P(r_1 < r)$.

In particular, we have

$$\mathcal{E}(N(r_\alpha)) = P(r_1 < r_\alpha) = \alpha. \qquad \ldots \quad (2.19)$$

(e) *Comparison of values of upper bound of* $P(r_1 < r)$ *with Grubbs' exact values of* $P(r_1 < r)$ *for a sample from a one-dimensional normal distribution.* For the case $k = 1$ it will be seen from (3.10) that $R_{(1)} \geqslant 0$ (i.e. $r_1 \geqslant 0$), and $R_{(2)} \geqslant 1-\frac{n}{2(n-1)}$. Hence for any value of r on the interval $\left(0, 1-\frac{n}{2(n-1)}\right)$ the expression (2.15) is an equality. For a value of r which exceeds $1-\frac{n}{2(n-1)}$ expression (2.15) is a strict inequality. In this case r_1 is the smaller of the two quantities $\sum_{\xi=1}^{n-1}(x_{(\xi)}-\bar{x}_n)^2/[(n-1)s^2]$ and $\sum_{\xi=2}^{n}(x_{(\xi)}-\bar{x}_1)^2/[(n-1)s^2]$ which were considered by Grubbs (1950) as criteria for upper and lower outliers in a sample from a one-dimensional distribution. Thus, if r_α is the lower $100\alpha\%$ point of r_1 and lies on the interval $\left(0, 1-\frac{n}{2(n-1)}\right)$ it is the lower $100\frac{\alpha}{2}\%$ of each of the two criteria considered by Grubbs. For the case $k = 1$ Table 1 gives a comparison between the probability $P(r_1 < r_\alpha)$ and its upper bound $nP(u < r_\alpha)$ for $\alpha = 0.02$, 0.05, 0.10 and 0.20 and for certain values of n from Grubbs' tables for which inequality (2.15) is a strict inequality.

TABLE 1. COMPARISON OF $P(r_1 < r_\alpha)$ WITH ITS UPPER BOUND $nP(u < r_\alpha)$

n	α	r_α	exact probability (by Grubbs) $P(r_1 < r_\alpha)$	upper bound $nP(u < r_\alpha)$ of $P(r_1 < r_\alpha)$ [or equivalently, $E(N(r_\alpha))$]
20	.02	.5393	.020	.020
25		.6071	.020	.021
15		.5030	.050	.050
20	.05	.5937	.050	.050
25		.6544	.050	.052
15		.5558	.100	.100
20	.10	.6379	.100	.100
25		.6922	.100	.103
10		.4881	.200	.200
15	.20	.6134	.200	.200
20		.6848	.200	.206
25		.7319	.200	.210

(f) *Tables of values of r_α for which upper bound $nP(u < r_\alpha) = \alpha$.* For the case $k \geqslant 2$ it will be seen from (3.10) that the left hand endpoints of the distributions of $R_{(1)}, \ldots, R_{(k)}$ are all 0. Therefore for $k \geqslant 2$ expression (2.15) is a strict inequality; and there exists no value of r for which $nP(u < r)$ provides an exact value of $P(r_1 < r)$. The problem of determining exact values of $P(r_1 < r)$ for $k \geqslant 2$ does not seem feasible at present because of the complexity of the distribution of r_1. We therefore resort to the use of the upper bound $nP(u < r)$.

Table 2 gives values of r_α for which the upper bound $nP(u < r_\alpha)$ of $P(r_1 < r_\alpha)$ has the value α[or equivalently, values of r_α for which $\mathscr{E}(N(r_\alpha)) = \alpha$] for $\alpha = 0.010$, 0.025, 0.050, 0.100; $k = 1, 2, 3, 4, 5$; and $n = 5(1)30(5)100(100)500$.

3. THE CASE OF TWO OUTLIERS

Suppose we delete two elements, say $(x_{1\xi}, \ldots, x_{k\xi})$ and $(x_{1\eta}, \ldots, x_{k\eta})$ from the sample defined in Section 2 and denote the internal scatter of the resulting cluster of $n-2$ points by $|a_{ij\xi\eta}|$ which is positive with probability 1 if $n > k+2$. Let

$$R_{\xi\eta} = \frac{|a_{ij\xi\eta}|}{|a_{ij}|}, \; \eta > \xi = 1, \ldots, n. \qquad \ldots \quad (3.1)$$

The quantities $\{R_{\xi\eta}\}$ will be called *two-outlier scatter ratios* of the sample $(x_{1\xi}, \ldots x_{k\xi}$; $\xi = 1, \ldots, n)$. The conditions satisfied by the $\{R_{\xi\eta}\}$ except that each must lie on (0, 1) appear rather complicated and no attempt will be made here to state them.

It can be shown (see Wilks (1962), for instance) that for $n > k+2$ each of the $\binom{n}{2}$ scatter ratios in the set $\{R_{\xi\eta}\}$ has the property that its distribution is identical with that of a random variable u^2 where u has the beta distribution $B_e(n-k-2, k)$.

Let $r_2 = \min\limits_{\eta > \xi} \{R_{\xi\eta}\}$. The criterion proposed here for selecting and testing the strongest candidate pair of sample elements as significant outliers is r_2, that is, the candidate pair whose deletion in computing two-outlier scatter ratios produces the smallest scatter ratio.

No attempt is made here to give inequalities for these ordered scatter ratios analogous to those for the $\{R_{(1)}, \ldots, R_{(n)}\}$ as given in (2.10).

Under the null hypothesis, (that is, assuming that all elements in the sample are independently drawn from a common k-dimensional normal distribution) the joint distribution of $\{R_{\xi\eta}, \eta > \xi = 1, \ldots, n\}$ is symmetric in the $R_{\xi\eta}$, although apparently very complicated. However, an upper bound for the probability $P(r_2 < r)$ can be found by a procedure similar to that by which (2.15) was established, namely

$$P(r_2 < r) \leqslant \binom{n}{2} P(u^2 < r) \qquad \ldots \quad (3.2)$$

where
$$P(u^2 < r) = \frac{\Gamma(n-2)}{\Gamma(n-k-2)\,\Gamma(k)} \int_0^r (\sqrt{u})^{n-k-3}(1-\sqrt{u})^{k-1} d\sqrt{u}, \qquad \ldots \quad (3.3)$$

remembering that each $R_{\xi\eta}$ is a random variable having a distribution identical to that of a random variable u^2 where u has the beta distribution $B_e\,(n-k-2,\ k)$.

In particular if we choose r_α such that

$$\binom{n}{2} P(u^2 < r_\alpha) = \alpha, \qquad \ldots \quad (3.4)$$

we have
$$P(r_2 < r_\alpha) \leqslant \alpha. \qquad \ldots \quad (3.5)$$

As in the one-outlier problem, if we let $N(r)$ be the number of the $\binom{n}{2}$ two-outlier scatter ratios $\{R_{\xi\eta}\}$ which have values less than r, then

$$\mathcal{E}(N(r)) = \binom{n}{2} P(u^2 < r). \qquad \ldots \quad (3.6)$$

In particular, we have

$$\mathcal{E}(N(r_\alpha)) = \binom{n}{2} P(u^2 < r_\alpha) = \alpha. \qquad \ldots \quad (3.7)$$

Values of $\sqrt{r_\alpha}$ for which the upper bound $\binom{n}{2} P(u^2 < r_\alpha)$ of $P(r_2 < r_\alpha)$ has the value α [or equivalently, values of $\sqrt{r_\alpha}$ for which $\mathcal{E}[N(r_\alpha)) = \alpha$] are given in Table 3 for $\alpha = 0.010,\ 0.025,\ 0.050,\ 0.100$; $k = 1, 2, 3, 4, 5$; and $n = 5(1)\,30\,(5)\,100(100)500$.

4. The case of three or more outliers

The scatter ratio criteria for selecting and testing outliers can be extended to the case of three or more outliers in a fairly straightforward way.

For t outliers, we define the $\binom{n}{t}$ t-outlier scatter ratios as

$$R_{\xi_1 \dots \xi_t} = \frac{|a_{ij\xi_1 \cdots \xi_t}|}{|a_{ij}|}, \qquad \dots \quad (4.1)$$

$\xi_t < \dots < \xi_1 = 1, \dots, n$ where $|a_{ij\xi_1 \cdots \xi_n}|$ is the internal scatter of the $n-t$ points remaining in the sample after deletion of $(x_{1\xi_1} \dots, x_{k\xi_1}), \dots, (x_{1\xi_t}; \dots, x_{\mu k\xi_t})$. The scatter ratio $R_{\xi_1 \dots \xi_t}$ is positive with probability 1 if $n > k+t$. The smallest of these scatter ratios, which we denote by r_t is the proposed criterion for selecting the t most extreme observations in the sample and for testing this set of t observations as a set of significant outliers.

Under the assumption that the n elements in the sample are independently drawn from a common k-dimensional normal distribution any one of the scatter ratios, say $R_{\xi_1 \dots \xi_t}$ is a random variable whose k-th moment is given by (see Wilks (1962))

$$\mathscr{E}(R^h_{\xi_1 \cdots \xi_t}) = \prod_{i=1}^{t} \frac{\Gamma\left(\frac{n-i}{2}\right)\Gamma\left(\frac{n-k-i}{2}+h\right)}{\Gamma\left(\frac{n-i}{2}+h\right)\Gamma\left(\frac{n-k-i}{2}\right)}, \qquad \dots \quad (4.2)$$

$$h = 0, 1, 2, \dots$$

Note that the h-th moment of $R_{\xi_1 \cdots \xi_t}$ is identical with the h-th moment of the product $z_1 \dots z_t$ where $z_1, \dots, z_t$ are independent random variables having beta distributions $B_e\left(\frac{n-k-1}{2}, \frac{k}{2}\right), \dots, B_e\left(\frac{n-k-t}{2}, \frac{k}{2}\right)$, respectively. The distribution of $R_{\xi_1 \dots \xi_t}$, is uniquely determined by its moments (see Cramér (1943)). Hence the distribution of $R_{\xi_1 \dots \xi_t}$ is identical with the distribution of the product $z_1 \dots z_t$ and hence

$$P(R_{\xi_1 \cdots \xi_t} < r) = P(z_1 \dots z_t < r). \qquad \dots \quad (4.3)$$

As in the one- and two-outlier problems we find that

$$P(r_t < r) \leqslant \binom{n}{t} P(z_1 \dots z_t < r), \qquad \dots \quad (4.4)$$

the probability $P(z_1 \dots z_t < r)$ to be determined from the joint distribution of $z_1, \dots, z_t$ as described above.

If $N(r)$ is the number of the $\binom{n}{t}$ scatter ratios in $\{R_{\xi_1} \cdots {\xi_t}\}$ which are less than r we have, as in the one- and two-outlier cases,

$$\mathscr{E}(N(r)) = \binom{n}{t} P(z_1 \dots z_t < r). \qquad \dots \quad (4.5)$$

If r_α is chosen so that $\binom{n}{t} P(z_1 \ldots z_t < r_\alpha) = \alpha$ then

$$P(r_t < r_\alpha) \leqslant \binom{n}{t} P(z_1 \ldots z_t < r_\alpha) = \mathcal{E}(N(r_\alpha)) = \alpha. \qquad \ldots \quad (4.6)$$

As a matter of fact, the probability $P(z_1 \ldots z_t < r)$ can be reduced to a probability involving fewer than t independent beta variables if $t > 1$. More precisely, if t is even $P(z_1 \ldots z_t < r)$ reduces to an expression involving $\frac{1}{2}t$ independent beta variables, and if t is odd it reduces to one involving $\frac{1}{2}t+\frac{1}{2}$ independent beta variables. In the case of two outliers $P(z_1 z_2 < r)$ reduces to $P(u^2 < r)$ as given by (3.3). We shall now consider the cases of three and four outliers.

In the three-outlier problem, by making use of the relation

$$\sqrt{\pi}\,\Gamma(2m) = 2^{2m-1}\Gamma(m)\,\Gamma(m+\tfrac{1}{2}) \qquad \ldots \quad (4.7)$$

in (4.2) for $t = 3$, we find

$$\mathcal{E}(R^h_{\xi_1\xi_2\xi_3}) = \frac{\Gamma(n-2)\Gamma\left(\frac{n-3}{2}\right)\Gamma(n-k-2+2h)\,\Gamma\left(\frac{n-k-3}{2}+h\right)}{\Gamma(n-2+2h)\Gamma\left(\frac{n-3}{2}+h\right)\Gamma(n-k-2)\Gamma\left(\frac{n-k-3}{2}\right)}, \qquad \ldots \quad (4.8)$$

from which it is seen that the distribution of $R_{\xi_1\xi_2\xi_3}$ is identical with that of the product u^2v where u and v are independent random variables having beta distributions $B_e(n-k-2, k)$ and $B_e\left(\frac{n-k-3}{2}, \frac{k}{2}\right)$, respectively. Therefore,

$$P(R_{\xi_1\xi_2\xi_3} < r) = P(u^2v < r), \qquad \ldots \quad (4.9)$$

and denoting $\min\limits_{\xi_3>\xi_2>\xi_1} \{R_{\xi_1\xi_2\xi_3}\}$ by r_3

we have
$$P(r_3 < r) \leqslant \binom{n}{3} P(u^2v < r) \qquad \ldots \quad (4.10)$$

where, omitting details, we find

$$P(u^2v < r) = \frac{\Gamma(n-2)\,\Gamma\left(\frac{n-3}{2}\right)}{\Gamma(n-k-2)\Gamma\left(\frac{n-k-3}{2}\right)\Gamma(k)\,\Gamma\left(\frac{k}{2}\right)} \int_0^r s^{\frac{n-k-5}{2}} \int_{\sqrt{s}}^1 (1-u)^{k-1}\left(1-\frac{s}{u^2}\right)^{\frac{k}{2}-1} du\, ds$$

$$\ldots \quad (4.11)$$

For $k = 1$ this expression reduces to

$$P(u^2v < r) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-4}{2}\right)} \int_0^r s^{\frac{n-5}{2}} \int_s^1 v^{-\frac{3}{2}}(1-v)^{-\frac{1}{2}} dv\, ds \qquad \ldots \quad (4.12)$$

and for $k = 2$ it reduces to

$$P(u^2v < r) = \frac{(n-3)(n-4)(n-5)}{2}(\sqrt{r})^{n-5}\left[\frac{1}{n-5} - \frac{2\sqrt{r}}{n-4} + \frac{r}{n-3}\right]. \qquad \ldots \quad (4.13)$$

If we choose r_α so that
$$\binom{n}{3} P(u^2v < r_\alpha) = \alpha. \qquad \dots \quad (4.14)$$

where $P(u^2v < r)$ is given by (4.10), we have
$$P(r_3 < r_\alpha) \leqslant \alpha.$$

If $N(r)$ is the number of the scatter ratios in the set $\{R_{\xi_1\xi_2\xi_3}\}$ having values $< r$ we note that $\mathcal{E}(N(r_\alpha)) = \alpha$.

In the four-outlier case by making use of (4.7) in (4.2) for $t = 4$ we find

$$\mathcal{E}(R^h_{\xi_1\xi_2\xi_3\xi_4}) = \frac{\Gamma(n-2)\Gamma(n-4)\Gamma(n-k-2+2h)\Gamma(n-k-4+2h)}{\Gamma(n-k-2)\Gamma(n-k-4)\Gamma(n-2+2h)\Gamma(n-4+2h)} \qquad \dots \quad (4.15)$$

from which we note that the distribution of $R_{\xi_1\xi_2\xi_3\xi_4}$ is identical with that of the product u^2w^2 where u and w are independent random variables having the beta distributions $B_e(n-k-2, k)$ and $B_e\ (n-k-4, k)$, respectively. Therefore

$$P(R_{\xi_1\xi_2\xi_3\xi_4} < r) = P(u^2w^2 < r) \qquad \dots \quad (4.16)$$

and denoting $\min\limits_{\xi_4>\xi_3>\xi_2>\xi_1} \{R_{\xi_1\xi_2\xi_3\xi_4}\}$ by r_4

we have
$$P(r_4 < r) \leqslant \binom{n}{4} P(u^2w^2 < r) \qquad \dots \quad (4.17)$$

where, omitting details, we find that

$$P(u^2w^2 < r) = \frac{\Gamma(n-2)\Gamma(n-4)}{2\Gamma(n-k-2))\Gamma(n-k-4)\Gamma^2(k)} \int_0^r s^{\frac{n-k-6}{2}} \int_{\sqrt{s}}^1 u\left(1+\sqrt{s}-u-\frac{\sqrt{s}}{u}\right)^{k-1} du\, ds$$
$$\dots \quad (4.18)$$

For $k = 1$ (4.18) reduces to

$$P(u^2w^2 < r) = \tfrac{1}{2}(\sqrt{r})^{n-5}[(n-3)-(n-5)r], \qquad \dots \quad (4.19)$$

and for $k = 2$ we find

$$P(u^2w^2 < r) = \frac{(n-3)!}{6(n-7)!}(\sqrt{r})^{n-6}\left[\frac{1}{n-6}-\frac{3\sqrt{r}}{n-5}+\frac{3r}{n-4}-\frac{\sqrt{r^3}}{n-3}\right]. \qquad \dots \quad (4.20)$$

Again note that if we choose k_α so that

$$\binom{n}{4} P(u^2w^2 < r_\alpha) = \alpha, \qquad \dots \quad (4.21)$$

where $P(u^2w^2 < r)$ is given by (4.18) we obtain
$$P(r_4 < r_\alpha) \leqslant \alpha. \qquad \dots \quad (4.22)$$

as in the case for $k = 3$ if $N(r)$ is the number of scatter ratios in the set $\{R_{\xi_1\xi_2\xi_3\xi_4}\}$ which have values less than r we have $\mathcal{E}(N(r_\alpha)) = \alpha$.

TABLE 2. VALUES OF r_α FOR WHICH THE UPPER BOUND $nP(u<r_\alpha)$ OF $P(r_1<r_\alpha)$ HAS THE VALUE α [OR EQUIVALENTLY, VALUES OF r_α FOR WHICH $E(N(r_\alpha))=\alpha$] FOR THE CASE OF ONE OUTLIER

$\alpha = 0.010$

sample size n	number of dimensions k: 1	2	3	4	5
5	0.02795	0.00200	0.00000		
6	.06592	.01406	.00111	0.00000	
7	.11026	.03780	.00893	.00071	0.00000
8	.15547	.06898	.02593	.00632	.00050
9	.19888	.10358	.04987	.01937	.00476
10	.23942	.13895	.07781	.03866	.01523
11	.27678	.17364	.10755	.06200	.03129
12	.31103	.20689	.13765	.08757	.05126
13	.34238	.23835	.16726	.11407	.07362
14	.37107	.26790	.19590	.14065	.09723
15	.39738	.29556	.22330	.16678	.12128
16	.42156	.32141	.24936	.19215	.14525
17	.44383	.34555	.37404	.21657	.16878
18	.46440	.36810	.29737	.23996	.19167
19	.48344	.38919	.31940	.26228	.21378
20	.50112	.40893	.34019	.28354	.23506
21	.51757	.42743	.35982	.30376	.25547
22	.53292	.44480	.37835	.32298	.27501
23	.54727	.46113	.39588	.34125	.29370
24	.56071	.47651	.41246	.35861	.31155
25	.57334	.49102	.42815	.37513	.32861
26	.58521	.50471	.44304	.39084	.34491
27	.59641	.51767	.45716	.40580	.36048
28	.60698	.52994	.47057	.42006	.37536
29	.61697	.54158	.48333	.43365	.38959
30	.62644	.55263	.49547	.44663	.40320
35	.66716	.60048	.54835	.50344	.46318
40	.69944	.63870	.59091	.54949	.51217
45	.72567	.66994	.62588	.58754	.55284
50	.74745	.69598	.65514	.61947	.58711
55	.76583	.71803	.67997	.64666	.61636
60	.78157	.73694	.70133	.67009	.64162
65	.79521	.75336	.71990	.69050	.66366
70	.80715	.76775	.73620	.70843	.68306
75	.81769	.78048	.75062	.72432	.70026
80	.82708	.79181	.76348	.73851	.71563
85	.83549	.80197	.77503	.75124	.72944
90	.84308	.81115	.78545	.76274	.74192
95	.84995	.81946	.79490	.77319	.75326
100	.85622	.82704	.80352	.78271	.76361
200	.92016	.90435	.89155	.88018	.86361
300	.94392	.93293	.92411	.91625	.90899
400	.95652	.94801	.94125	.93525	.92969
500	.96439	.95739	.95190	.94704	.94254

TABLE 2. VALUES OF r_α FOR WHICH THE UPPER BOUND $nP(u<r_\alpha)$ OF $P(r_1<r_\alpha)$ HAS THE VALUE α [OR EQUIVALENTLY, VALUES OF r_α FOR WHICH $E(N(r_\alpha))=\alpha$] FOR THE CASE OF ONE OUTLIER—(*Continued*)

$\alpha = 0.025$

sample size n	number of dimensions k: 1	2	3	4	5
5	0.05124	0.00500	0.00002		
6	.10353	.02589	.00278	0.00001	
7	.15787	.05976	.01647	.00179	0.00000
8	.20934	.09953	.04111	.01166	.00125
9	.25636	.14057	.07219	.03075	.00879
10	.29873	.18053	.10601	.05606	.02420
11	.33677	.21834	.14030	.08466	.04541
12	.37094	.25361	.17380	.11452	.07008
13	.40170	.28629	.20589	.14441	.09644
14	.42950	.31647	.23627	.17360	.12331
15	.45471	.34433	.26485	.20171	.14998
16	.47768	.37007	.29165	.22854	.17601
17	.49867	.39387	.31674	.25400	.20114
18	.51794	.41593	.34022	.27811	.22525
19	.53569	.43642	.36221	.30089	.24827
20	.55208	.45547	.38281	.32239	.27020
21	.56727	.47324	.40213	.34269	.29106
22	.58139	.48984	.42028	.36187	.31088
23	.59455	.50538	.43735	.37999	.32971
24	.60685	.51996	.45343	.39713	.34760
25	.61836	.53367	.46860	.41336	.36460
26	.62917	.54657	.48292	.42873	.38076
27	.63934	.55874	.49647	.44332	.39614
28	.64891	.57025	.50930	.45717	.41079
29	.65796	.58113	.52147	.47034	.42475
30	.66651	.59144	.53303	.48287	.43806
35	.70317	.63587	.58306	.53737	.49626
40	.73208	.67113	.62301	.58117	.54333
45	.75551	.69982	.65567	.61711	.58213
50	.77492	.72365	.68286	.64715	.61466
55	.79128	.74378	.70589	.67264	.64232
60	.80527	.76102	.72564	.69454	.66613
65	.81738	.77596	.74278	.71357	.68685
70	.82798	.78904	.75780	.73027	.70504
75	.83733	.80060	.77108	.74503	.72116
80	.84566	.81088	.78291	.75820	.73553
85	.85312	.82010	.79352	.77001	.74844
90	.85984	.82841	.80308	.78067	.76009
95	.86594	.83595	.81176	.79034	.77066
100	.87150	.84281	.81967	.79916	.78030
200	.92829	.91280	.90028	.88914	.87885
300	.94949	.93871	.93008	.92240	.91530
400	.96077	.95240	.94579	.93993	.93450
500	.96783	.96093	.95555	.95082	.94642

TABLE 2. VALUES OF r_α FOR WHICH THE UPPER BOUND $nP(u<r_\alpha)$ OF $P(r_1<r_\alpha)$ HAS THE VALUE α [OR EQUIVALENTLY, VALUES OF r_α FOR WHICH $E(N(r_\alpha)) = \alpha$] FOR THE CASE OF ONE OUTLIER—(*Continued*)

$\alpha = 0.050$					
sample size n	number of dimensions k				
	1	2	3	4	5
5	0.08083	0.01000			
6	.14529	.04110	0.00556		
7	.20661	.08452	.02620	0.00358	
8	.26161	.13133	.05831	.01856	0.00251
9	.31006	.17711	.09559	.04367	.01400
10	.35261	.22007	.13408	.07438	.03440
11	.39008	.25965	.17171	.10731	.06033
12	.42325	.29584	.20751	.14050	.08896
13	.45277	.32886	.24112	.17285	.11850
14	.47921	.35897	.27245	.20383	.14785
15	.50302	.38650	.30154	.23319	.17642
16	.52457	.41171	.32855	.26086	.20386
17	.54417	.43487	.35361	.28686	.23002
18	.56208	.45620	.37690	.31124	.25486
19	.57852	.47591	.39857	.33412	.27837
20	.59365	.49417	.41876	.35558	.30060
21	.60764	.51113	.43761	.37573	.32160
22	.62061	.52692	.45525	.39467	.34145
23	.63267	.54166	.47178	.41249	.36021
24	.64391	.55545	.48729	.42929	.37796
25	.65443	.56838	.50188	.44513	.39477
26	.66429	.58053	.51563	.46010	.41069
27	.67355	.59197	.52860	.47426	.42580
28	.68226	.60276	.54086	.48767	.44014
29	.69048	.61296	.55247	.50040	.45377
30	.69825	.62260	.56347	.51248	.46674
35	.73146	.66402	.61090	.56478	.52314
40	.75758	.69675	.64857	.60654	.56843
45	.77872	.72330	.67924	.64067	.60557
50	.79621	.74532	.70472	.66909	.63659
55	.81094	.76388	.72624	.69314	.66289
60	.82354	.77975	.74467	.71377	.68548
65	.83444	.79351	.76065	.73167	.70511
70	.84398	.80554	.77464	.74735	.72232
75	.85240	.81616	.78700	.76122	.73754
80	.85989	.82561	.79800	.77356	.75111
85	.86661	.83408	.80786	.78463	.76328
90	.87267	.84172	.81674	.79462	.77426
95	.87816	.84864	.82480	.80367	.78421
100	.88317	.85494	.83214	.81192	.79329
200	.93447	.91924	.90696	.89602	.88591
300	.95372	.94310	.93463	.92711	.92013
400	.96399	.95573	.94924	.94351	.93817
500	.97043	.96361	.95833	.95370	.94938

TABLE 2. VALUES OF r_α FOR WHICH THE UPPER BOUND $nP(u<_\alpha)$ OF $P(r_1<r_\alpha)$ HAS THE VALUE α [OR EQUIVALENTLY, VALUES OF r_α FOR WHICH $E(N(r_\alpha)) = \alpha$] FOR THE CASE OF ONE OUTLIER—(*Continued*)

$\alpha = 0.100$

sample size n	number of dimensions k: 1	2	3	4	5
5	0.10000	0.02000	0.00025		
6	.20000	.06525	.01114	0.00012	
7	.26960	.11952	.04172	.00717	0.00007
8	.32610	.17328	.08282	.02959	.00502
9	.37418	.22314	.12675	.06216	.02234
10	.41540	.26827	.16978	.09888	.04901
11	.45106	.30878	.21038	.13629	.08032
12	.48221	.34511	.24801	.17267	.11319
13	.50966	.37776	.28264	.20723	.14593
14	.53405	.40719	.31442	.23967	.17764
15	.55586	.43383	.34358	.26995	.20789
16	.57550	.45804	.37037	.29813	.23651
17	.59328	.48014	.39502	.32433	.26346
18	.60948	.50038	.41777	.34870	.28878
19	.62428	.51899	.43881	.37139	.31254
20	.63789	.53615	.45832	.39255	.33484
21	.65043	.55205	.47645	.41231	.35578
22	.66205	.56680	.49334	.43079	.37545
23	.67282	.58053	.50912	.44812	.39396
24	.68286	.59335	.52389	.46438	.41140
25	.69223	.60535	.53774	.47967	.42784
26	.70101	.61660	.55075	.49408	.44337
27	.70925	.62717	.56301	.50767	.45806
28	.71699	.63713	.57457	.52052	.47197
29	.72429	.64653	.58549	.53268	.48516
30	.73119	.65540	.59583	.54420	.49768
35	.76063	.69342	.64023	.59385	.55182
40	.78375	.72335	.67531	.63326	.59499
45	.80245	.74758	.70379	.66532	.63023
50	.81792	.76763	.72738	.69195	.65955
55	.83094	.78451	.74728	.71443	.68435
60	.84208	.79895	.76430	.73369	.70561
65	.85173	.81145	.77904	.75037	.72404
70	.86017	.82238	.79193	.76497	.74019
75	.86763	.83203	.80331	.77787	.75446
80	.87427	.84061	.81344	.78935	.76717
85	.88023	.84830	.82251	.79963	.77856
90	.88560	.85524	.83069	.80891	.78883
95	.89048	.86152	.83811	.81731	.79814
100	.89492	.86725	.84486	.82497	.80663
200	.94067	.92574	.91371	.90300	.89308
300	.95796	.94751	.93923	.93186	.92503
400	.96722	.95908	.95272	.94711	.94189
500	.97304	.96631	.96113	.95661	.95238

TABLE 3. VALUES OF $\sqrt{r_\alpha}$ FOR WHICH THE UPPER BOUND $\binom{n}{2}$ $P(u^2<r_\alpha)$ OF $P\,(r_2<r_\alpha)$ HAS THE VALUE α [OR EQUIVALENTLY, VALUES OF $\sqrt{r_\alpha}$ FOR WHICH $E(N(r_\alpha))=\alpha$] FOR THE CASE OF TWO OUTLIERS

$\alpha = 0.010$

sample size n	number of dimensions k				
	1	2	3	4	5
5	0.03162	0.00050			
6	.08736	.01498	0.00022		
7	.14772	.04982	.00896	0.00012	
8	.20444	.09374	.03349	.00601	0.00007
9	.25544	.13926	.06744	.02449	.00433
10	.30069	.18308	.10490	.05181	.01887
11	.34076	.22397	.14265	.08337	.04152
12	.37636	.26160	.17912	.11626	.06861
13	.40812	.29606	.21363	.14889	.09761
14	.43660	.32758	.24595	.18044	.12699
15	.46228	.35641	.27605	.21050	.15589
16	.48553	.38284	.30403	.23893	.18382
17	.50670	.40713	.33002	.26568	.21057
18	.52604	.42952	.35419	.29082	.23601
19	.54380	.45019	.37667	.31441	.26013
20	.56016	.46935	.39764	.33655	.28296
21	.57528	.48714	.41721	.35735	.30454
22	.58930	.50371	.43551	.37689	.32494
23	.60234	.51918	.45266	.39528	.34423
24	.61451	.53365	.46876	.41261	.36248
25	.62588	.54722	.48390	.42895	.37975
26	.63655	.55996	.49816	.44439	.39611
27	.64657	.57196	.51162	.45899	.41164
28	.65599	.58328	.52434	.47282	.42637
29	.66489	.59397	.53637	.48594	.44037
30	.67329	.60409	.54778	.49839	.45370
35	.70925	.64753	.59694	.55228	.51159
40	.73753	.68184	.63596	.59527	.55804
45	.76043	.70968	.66772	.63038	.59611
50	.77937	.73276	.69411	.65962	.62789
55	.79532	.75222	.71638	.68436	.65482
60	.80897	.76887	.73547	.70556	.67796
65	.82077	.78328	.75201	.72396	.69804
70	.83111	.79590	.76649	.74009	.71566
75	.84023	.80704	.77928	.75434	.73124
80	.84835	.81695	.79066	.76703	.74512
85	.85563	.82583	.80086	.77840	.75757
90	.86219	.83384	.81007	.78866	.76880
95	.86814	.84110	.81841	.79797	.77899
100	.87356	.84771	.82601	.80645	.78828
200	.92902	.91518	.90349	.89291	.88304
300	94974	.94023	.93218	.92489	.91808
400	.96076	.95349	.94734	.94176	.93655
500	.96766	.96180	.95679	.95226	.94804

TABLE 3. VALUES OF $\sqrt{r_\alpha}$ FOR WHICH THE UPPER BOUND $\binom{n}{2}$ $P(u^2<r_\alpha)$ OF $P(r_2<r_\alpha)$ HAS THE VALUE α [OR EQUIVALENTLY, VALUES OF $\sqrt{r_\alpha}$ FOR WHICH $E(N(r_\alpha)) = \alpha$] FOR THE CASE OF TWO OUTLIERS—(*Continued*)

	$\alpha = 0.025$				
	number of aimensions k				
sample size n	1	2	3	4	5
5	0.05000	0.00125			
6	.11856	.02376	0.00056		
7	.18575	.06794	.01422	0.00030	
8	.24556	.11852	.04575	.00954	0.00018
9	.29758	.16820	.08546	.03346	.00687
10	.34274	.21444	.12703	.06572	.02579
11	.38212	.25661	.16755	.10110	.05269
12	.41670	.29479	.20581	.13678	.08327
13	.44728	.32932	.24141	.17138	.11495
14	.47453	.36058	.27432	.20427	.14634
15	.49896	.38897	.30468	.23522	.17670
16	.52099	.41483	.33266	.26419	.20568
17	.54097	.43848	.35849	.29124	.23314
18	.55918	.46017	.38237	.31647	.25905
19	.57585	.48014	.40450	.34003	.28346
20	.59118	.49859	.42504	.36203	.30642
21	.60532	.51567	.44415	.38260	.32802
22	.61842	.53155	.46197	.40187	.34835
23	.63058	.54633	.47863	.41995	.36751
24	.64191	.56014	.49423	.43694	.38557
25	.65250	.57307	.50887	.45292	.40262
26	.66242	.58520	.52263	.46799	.41874
27	.67173	.59660	.53560	.48221	.43399
28	.68049	.60734	.54784	.49566	.44844
29	.68874	.61748	.55941	.50839	.46215
30	.69653	.62707	.57036	.52046	.47517
35	.72985	.66814	.61742	.57252	.53152
40	.75603	.70050	.65465	.61389	.57651
45	.77720	.72671	.68487	.64757	.61326
50	.79471	.74841	.70994	.67555	.64386
55	.80946	.76669	.73108	.69919	.66975
60	.82207	.78233	.74917	.71944	.69195
65	.83300	.79586	.76484	.73699	.71121
70	.84255	.80770	.77855	.75236	.72808
75	.85099	.81815	.79066	.76593	.74299
80	.85851	.82746	.80144	.77801	.75628
85	.86524	.83579	.81109	.78884	.76818
90	.87132	.84330	.81980	.79861	.77892
95	.87683	.85012	.82769	.80746	.78866
100	.88185	.85632	.83487	.81552	.79753
200	.93335	.91971	.90818	.89774	.88801
300	.95267	.94330	.93537	.92818	.92146
400	.96298	.95582	.94976	.94426	.93912
500	.96944	.96350	.95873	.95427	.95010

TABLE 3. VALUES OF $\sqrt{r_\alpha}$ FOR WHICH THE UPPER BOUND $\binom{n}{2}$ $P(u^2<r_\alpha)$ OF $P(r_2<r_\alpha)$ HAS THE VALUE α [OR EQUIVALENTLY, VALUES OF $\sqrt{r_\alpha}$ FOR WHICH $E(N(r_\alpha))=\alpha$] FOR THE CASE OF TWO OUTLIERS—(*Continued*)

$\alpha = 0.050$.

sample size n	number of dimensions k: 1	2	3	4	5
5	0.07071	0.00250	0.00000		
6	.14938	.03372	.00111		
7	.22090	.08601	.02019	0.00060	
8	.28207	.14167	.05800	.01355	0.00036
9	.33403	.19419	.10237	.04245	.00975
10	.37841	.24188	.14702	.07881	.03273
11	.41670	.28462	.18945	.11716	.06322
12	.45006	.32286	.22884	.15489	.09657
13	.47939	.35712	.26504	.19086	.13029
14	.50539	.38792	.29819	.22462	.16313
15	.52863	.41573	.32853	.25609	.19451
16	.54952	.44096	.35634	.28532	.22417
17	.56842	.46394	.38188	.31244	.25207
18	.58562	.48496	.40540	.33763	.27823
19	.60134	.50426	.42712	.36103	.30274
20	.61578	.52205	.44722	.38282	.32571
21	.62908	.53849	.46588	.40313	.34723
22	.64139	.55374	.48324	.42210	.36743
23	.65282	.56793	.49944	.43986	.38641
24	.66346	.58117	.51459	.45651	.40426
25	.67339	.59354	.52878	.47215	.42108
26	.68269	.60515	.54211	.48687	.43695
27	.69141	.61604	.55465	.50076	.45194
28	.69962	.62630	.56648	.51386	.46612
29	.70735	.63598	.57765	.52625	.47955
30	.71465	.64513	.58821	.53799	.49230
35	.74583	.68424	.63351	.58850	.54730
40	.77032	.71501	.66926	.62850	.59106
45	.79013	.73992	.69824	.66101	.62671
50	.80652	.76052	.72224	.68798	.65635
55	.82032	.77787	.74247	.71073	.68138
60	.83213	.79271	.75978	.73021	.70284
65	.84236	.80555	.77476	.74708	.72143
70	.85132	.81678	.78787	.76185	.73772
75	.85922	.82670	.79944	.77489	.75210
80	.86627	.83553	.80974	.78650	.76491
85	.87259	.84343	.81896	.79689	.77639
90	.87829	.85056	.82728	.80627	.78674
95	.88346	.85703	.83482	.81477	.79612
100	.88817	.86292	.84169	.82251	.80467
200	.93664	.92316	.91177	.90145	.89181
300	.95490	.94564	.93780	.93069	.92405
400	.96466	.95759	.95160	.94616	.94107
500	.97080	.96500	.96021	.95580	.95167

TABLE 3. VALUES OF $\sqrt{r_\alpha}$ FOR WHICH THE UPPER BOUND $\binom{n}{2}$ $P(u^2<r_\alpha)$ OF $P(r_2<r_\alpha)$ HAS THE VALUE α [OR EQUIVALENTLY, VALUES OF $\sqrt{r_\alpha}$ FOR WHICH $E(N(r_\alpha))=\alpha$] FOR THE CASE OF TWO OUTLIERS—(*Continued*)

$\alpha = 0.100$

sample size n	number of dimensions k: 1	2	3	4	5
5	0.10000	0.00501			
6	.18821	.04791	0.00223		
7	.26269	.10904	.02872	0.00119	
8	.32402	.16955	.07368	.01927	0.00072
9	.37493	.22444	.12283	.05397	.01386
10	.41780	.27305	.17039	.09468	.04161
11	.45442	.31592	.21449	.13599	.07599
12	.48609	.35381	.25473	.17565	.11219
13	.51380	.38747	.29126	.21282	.14791
14	.53826	.41753	.32440	.24728	.18212
15	.56006	.44453	.35452	.27909	.21439
16	.57962	.46892	.38196	.30842	.24461
17	.59728	.49106	.40705	.33548	.27282
18	.61332	.51125	.43006	.36047	.29911
19	.62797	.52974	.45124	.38360	.32362
20	.64140	.54676	.47079	.40506	.34649
21	.65378	.56246	.48889	.42501	.36785
22	.66522	.57700	.50571	.44360	.38783
23	.67584	.59051	.52137	.46095	.40655
24	.68572	.60310	.53598	.47720	.42412
25	.69494	.61487	.54966	.49243	.44064
26	.70357	.62589	.56250	.50675	.45619
27	.71167	.63623	.57456	.52023	.47086
28	.71928	.64596	.58592	.53293	.48472
29	.72646	.65513	.59664	.54494	.49784
30	.73323	.66380	.60677	.55631	.51026
35	.76216	.70082	.65015	.60508	.56375
40	.78489	.72990	.68431	.64361	.60614
45	.80328	.75343	.71196	.67486	.64061
50	.81850	.77288	.73485	.70074	.66921
55	.83133	.78926	.75413	.72257	.69335
60	.84231	.80327	.77061	.74124	.71401
65	.85183	.81539	.78488	.75740	.73191
70	.86017	.82600	.79736	.77155	.74758
75	.86754	.83537	.80837	.78403	.76141
80	.87410	.84370	.81817	.79514	.77372
85	.87999	.85118	.82696	.80509	.78475
90	.88531	.85791	.83488	.81407	.79470
95	.89014	.86402	.84205	.82220	.80372
100	.89454	.86960	.84859	.82961	.81193
200	.93994	.92664	.91538	.90518	.89565
300	.95713	.94799	.94025	.93322	.92666
400	.96635	.95936	.95345	.94807	.94305
500	.97215	.96650	.96169	.95733	.95326

5. Acknowledgement

The author is grateful to Mr. Paul Raynault for programming and carrying out the computations involved in Tables 2 and 3 on the IBM 7090 electronic computer. The author is also glad to acknowledge several interesting and useful discussions about multidimensional outliers with Professor Henry F. Kaiser of the University of Illinois who was also a Fellow of the Center for Advanced Study in the Behavioral Sciences in the fall of 1961.

References

David, H. A. (1956): Revised upper percentage points of the extreme studentized deviate from the sample mean. *Biometrika*, **43**, 449–452.

Dixon, W. J. (1950): Analysis of extreme values. *Ann. Math. Stat.*, **21**, 488–506.

——— (1951): Ratios involving extreme values. *Ann. Math. Stat.*, **22**, 68–78.

Ferguson, T. S. (1961): On the rejection of outliers. *Proceedings Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, University of California Press.

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

Irwin, J. O. (1925): On a criterion for the rejection of outlying observations. *Biometrika*, **17**, 238–250.

Karlin, S. and Truax, D. (1960): Slippage problems. *Ann. Math. Stat.*, **31**, 296–324.

McKay, A. T. (1935): The distribution of the difference between the extreme observation and the sample mean in samples of n from a normal universe. *Biometrika*, **27**, 466–471.

Nair, K. R. (1952): Tables of percentage points of studentized extreme deviate from the sample mean. *Biometrika*, **39**, 189–193.

——— (1948): The distribution of the extreme deviate from the sample mean and its studentized form. *Biometrika*, **35**, 118–144.

Newman, D. (1940): The distribution of ranges in samples from a normal population, expressed in terms of an independent estimate of the standard deviation. *Biometrika*, **31**, 20–30.

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

Pearson, E. S. and Hartley, H. O. (1942): Tables of probability integral of studentized ranges. *Biometrika*, **33**, 89–99.

——— (1942): The probability integral of the range in samples of n observations from a normal population. *Biometrika*, **32**, 301–310.

Pillai, K. C. S. and Tienzo, B. P. (1959): On the distribution of the extreme studentized deviate from the sample mean. *Biometrika*, **46**, 467–472.

Pillai, K. C. S. (1959): Upper percentage points of the extreme studentized deviate from the sample mean. *Biometrika*, **46**, 473–474.

Rider, P. R. (1932): Criteria for rejection of observations. *Washington University Studies*, No. 8.

Thompson, W. R. (1935): On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviates, *Ann. Math. Stat.*, **6**, 214–219.

Wilks, S. S. (1962): *Mathematical Statistics*, John Wiley and Sons, Inc., New York.

Paper received : February, 1963.