TECHNOMETRICS ©, VOL. 25, NO. 2, MAY 1983

Percentage Points for a Generalized ESD Many-Outlier Procedure

Bernard Rosner

Channing Laboratory
Harvard Medical School
Boston, MA 02115

A generalized (extreme Studentized deviate) ESD many-outlier procedure is given for detecting from 1 to k outliers in a data set. This procedure has an advantage over the original ESD many-outlier procedure (Rosner 1975) in that it controls the type I error both under the hypothesis of no outliers and under the alternative hypotheses of 1, 2, ..., k-1 outliers. A method is given for approximating percentiles for this procedure based on the t distribution. This method is shown to be adequately accurate using Monte Carlo simulation, for detecting up to 10 outliers in samples as small as 25. Tables are given for implementing this method for $n = 25(1)50(10)100(50)500$; $k = 10$, $\alpha = .05, .01, .005$.

KEY WORDS: Outliers; Multiple outliers; t distribution; Detection of outliers; Extreme Studentized deviate.

1. INTRODUCTION

Detecting outliers can be important either because the outlying observations themselves are of interest (for example, they might represent significant mineral deposits), or because one wants to prevent outlier contamination of subsequent estimates. Here the problem addressed is detection, not estimation, under the assumption that the main body of data comes from a normal distribution.

For a prespecified upper limit, k, on the number of outliers, one approach to this problem is the *extreme Studentized deviate* (ESD) "*many outlier*" *procedure* (Rosner 1975; Prescott 1979). This procedure is based on the statistics $R_1, \ldots, R_k$, which are the extreme Studentized deviates computed from successively reduced samples of size n, $n-1, \ldots, n-k+1$ respectively. Specifically, for the complete sample, $R_1 = (\max |x_i - \bar{x}|)/s$, where $\bar{x} = (\sum x_i)/n$ and $s^2 = (\sum (x_i - \bar{x})^2)/(n-1)$. R_2 is then computed in an analogous way from the reduced sample of size $n-1$ obtained from deleting the observation corresponding to $\max |x_i - \bar{x}|$ from the full sample and similarly for $R_3, \ldots, R_k$. Critical values of the test are determined by specifying α and then finding $\beta, \lambda(\beta)$ such that

$$\Pr [R_i > \lambda_i(\beta) | H_0] = \beta, \qquad i = 1, \ldots, k$$

and (1.1)

$$\Pr \left\{ \bigcup_{i=1}^{k} [R_i > \lambda_i(\beta) | H_0] \right\} = \alpha.$$

The *ESD many outlier procedure* then has the following form:

> If all of the R_i are $\leq \lambda_i(\beta)$, then declare that no outliers are present.
>
> If some of the R_i are $> \lambda_i(\beta)$, then define $l = \max \{i: R_i > \lambda_i(\beta)\}$ and declare $x^{(0)}, x^{(1)}, \ldots, x^{(l-1)}$ as outliers where $x^{(0)}, x^{(1)}, \ldots, x^{(l-1)}$ correspond to the most extreme observations (i.e. the observations corresponding to $\max |x_i - \bar{x}|$ in the successively reduced samples. (1.2)

Although this procedure has good power against a variety of outlier alternatives, it has a number of shortcomings.

First, the procedure requires a table of estimated percentiles for $R_1, \ldots, R_k$ for each combination of sample size (n) and maximum number of outliers (k). These percentiles must be estimated from Monte Carlo simulations for each (n, k) (Rosner 1975; Prescott 1979). The unavailability of these percentiles for a wide variety of (n, k) combinations has hindered the usefulness of the procedure, particularly for large n.

Second, in the initial development of the procedure, the percentiles of $R_1, \ldots, R_k$ were all fixed at the same level (β) for convenience. The problem with this formulation (see Hawkins 1978) is that although the procedure has appropriate type I error when no outliers are present, it may not have appropriate type I error when there are some outliers present. For example, if there are l outliers in the data and they have

been detected and removed from the sample, where $1 \le l < k$, then it would be desirable to fix the probability of declaring more than l outliers at α. We denote these alternative hypotheses by H_l, $l = 1, \ldots, k$ and require that $\Pr\{\bigcup_{i=l+1}^{k} [R_i > \lambda_i \mid H_l]\} = \alpha$, $l = 0, 1, \ldots, k-1$. This property is not necessarily achieved by the procedure in (1.2). Indeed, if we use the ESD many-outlier procedure with a nominal type I error of α and there are l outliers present in the data ($1 \le l < k$), then the probability of declaring more than l outliers is greater than α. Thus, if there are some outliers present in the data, this procedure will tend to detect more than the appropriate number of outliers.

We propose a *generalized ESD many-outlier procedure* to deal specifically with the latter problem by specifying critical values as follows:
Find $\lambda_i, i = 1, \ldots, k$ such that

$$\Pr\left\{\bigcup_{i=l+1}^{k} (R_i > \lambda_i \mid H_l)\right\} = \alpha \quad \text{for} \quad l = 0, 1, \ldots, k-1. \tag{1.3}$$

A procedure similar to that given in (1.3) has been proposed in Hawkins (1980) and critical values are presented there for $n = 10, 15, 20, 30, 50$ using simulation procedures. An alternative to this method is to approximate critical values based on the t distribution.

In Section 2 we investigate the type I error for this procedure and provide a more extensive table of critical values. In Section 3 a power study is presented comparing the original and generalized ESD procedures. A table of estimated percentiles for the generalized ESD procedure utilizing approximate critical values based on the t distribution is given in Section 4. Finally, an example is given in Section 5.

2. ESTIMATION OF THE PERCENTILES OF THE GENERALIZED ESD MANY-OUTLIER PROCEDURE

From (1.3) we require λ_i such that $\Pr\{\bigcap_{i=l+1}^{k} [(R_i \le \lambda_i) \mid H_l]\} = 1 - \alpha$. We shall conjecture that this probability is essentially determined by R_{l+1}; that is,

$$\Pr\left\{\bigcap_{i=l+1}^{k} [(R_i \le \lambda_i) \mid H_l]\right\} \approx \Pr[(R_{l+1} \le \lambda_{l+1}) \mid H_l], \quad l = 0, 1, \ldots, k-1, \tag{2.1}$$

and will test the validity of this conjecture subsequently. If we denote the original sample by I_0, the successively reduced samples obtained by deleting the most extreme points by $I_1, \ldots, I_k$, and the mean and standard deviation from I_l by $\bar{x}^{(l)}$, $s^{(l)}$, $l = 0, 1, \ldots, k$, then we have that under H_l, $R_{l+1} = \max\{|y_i| : i \in I_l\}$ where $y_i = [x_i - \bar{x}^{(l)}]/s^{(l)}$ for all $i \in I_l$. The exact distribution of $\max[|x_i - \bar{x}^{(l)}|/s^{(l)}]$ was studied by Grubbs (1950) using numerical integration techniques. However, for simplicity we will use an approximate method due to Quesenberry and David (1961) whereby the distribution is approximated by a transformation of the t distribution. First, the distribution of y_i is given by Thompson (1935) as follows:

$$y_i \sim \frac{t_{n-l-2}(n-l-1)}{\{[n-l-2+t^2_{n-l-2}](n-l)\}^{1/2}}, \qquad i \in I_l \tag{2.2}$$

Second, using Bonferroni inequality techniques as given, for example, in Equation 4.1 of Quesenberry and David (1961) we will approximate $\Pr(\max_{i \in I_l} y_i > \lambda_{l+1})$ by $(n-l)\Pr(y_i > \lambda_{l+1})$ or equivalently λ_{l+1} is approximated from the equation:

$$1 - [\alpha/(n-l)] = \Pr(y_i \le \lambda_{l+1}). \tag{2.3}$$

Thus, from (2.2) and (2.3), λ_{l+1} is given for the one-sided outlier problem by

$$\lambda_{l+1} = \frac{t_{n-l-2,\,p}(n-l-1)}{\{[n-l-2+t^2_{n-l-2,\,p}](n-l)\}^{1/2}}, \quad l = 0, 1, \ldots, k-1, \tag{2.4}$$

where $p = 1 - [\alpha/(n-l)]$ and $t_{d,\,p}$ represents the pth percentile of a t distribution with d degrees of freedom. For the two-sided outlier problem, which is the problem actually addressed in (1.3), we substitute $\alpha/2$ for α in (2.4) and obtain

$$\lambda_{l+1} = \frac{t_{n-l-2,\,p}(n-l-1)}{\{[n-l-2+t^2_{n-l-2,\,p}](n-l)\}^{1/2}}, \quad l = 0, 1, \ldots, k-1, \tag{2.5}$$

where $p = 1 - [(\alpha/2)/(n-l)]$.

To evaluate the accuracy of the approximation in (2.1) and the resulting approximation to the percentiles of the R_i given in (2.5), a Monte Carlo study was conducted. The cases simulated were $n = 10(5)30, 50, 100$, and $k = 1, 2, \ldots, k^*$, where $k^* = \min([n/2], 10)$. The limit, k^*, was chosen since it was felt that this would be the maximum number of outliers one would want to detect in practice. Two thousand iterations were performed for each n. For each iteration,

1. a random sample of n, $N(0, 1)$ random deviates was generated using the IMSL subroutine GGNML (International Mathematical and Statistical Libraries, 1979);
2. the $R_i, i = 1, \ldots, k^*$ were computed as in (1.1);
3. for each $k = 1, \ldots, k^*$, it was noted whether $\{\bigcup_{i=1}^{k} (R_i > \lambda_i)\}$ was true or false using the approximate percentiles in (2.5) with a nominal α level of .05.

Table 1. *Estimation of the True* α *Levels* ($\hat{\alpha}(n, k)$) *Resulting From Using the Generalized ESD Procedure in* (1.3) *With the Approximate Percentiles in* (2.5), $n = 10(5)30,50,100$, $k = 1, \ldots,$ min ($[n/2]$, 10), $\alpha = .05$, 2000 *Iterations*

	k									
n	1	2	3	4	5	6	7	8	9	10
10	.06	.07	.09	.11	.13					
15	.05	.06	.07	.08	.08	.09	.10			
20	.06	.07	.07	.07	.07	.07	.07	.07	.08	.09
25	.05	.06	.06	.06	.06	.06	.06	.06	.06	.06
30	.06	.06	.06	.06	.06	.06	.06	.06	.06	.06
50	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05
100	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05

The actual α level (denoted by $\hat{\alpha}(n, k)$) was estimated by the proportion of samples for which $\{\bigcup_{i=1}^{k} (R_i > \lambda_i)\}$ was true. The results of this study, given in Table 1, show that the approximation clearly does not work well for very small samples; it yields true α levels of about 2 to 2.5 times the nominal α level of .05 for $n \leq 15$. However, for $n \geq 25$ the estimated α levels are quite close to .05 and the approximated percentiles can be confidently used in this case. Similar results were obtained for $\alpha = .01$.

3. POWER STUDIES

In this section we compare the power of the standard (ESD) and generalized (GEN) ESD many-outlier procedures given in (1.2) and (1.3) for the case $n = 25$, $k = 2$ using Monte Carlo simulation methods. In particular, 2,000 iterations were performed of the following algorithm to address this question:

1. a random sample of 25 standard normal random deviates $x_1, \ldots, x_{25}$ was generated using the IMSL subroutine GGNML (International Mathematical and Statistical Libraries, 1979);
2. x_{24} and x_{25} were then perturbed by adding the constants γ_1 and γ_2 respectively where $(\gamma_1, \gamma_2) = (0, 0)$, $(0, 2)$, $(0, 4)$, $(0, 6)$, $(2, 2)$, $(2, 4)$, $(2, 6)$, $(4, 4)$, $(4, 6)$, $(6, 6)$, $(-2, 2)$, $(-2, 4)$, $(-2, 6)$, $(-4, 4)$, $(-4, 6)$, $(-6, 6)$; thus $x_{24} \sim N(\gamma_1, 1)$, $x_{25} \sim N(\gamma_2, 1)$ (note that the same random sample in (1) was used on a given iteration for each (γ_1, γ_2) to provide consistency of results);
3. R_1 and R_2 were then computed and the procedures in (1.2) and (1.3) were implemented using $\alpha = .05$ to identify both the number of outliers in a data set and which specific points were outliers for both the standard and generalized ESD many-outlier procedures. The results are presented in Table 2.

Table 2. *Power Comparisons of the Standard* (ESD) *and Generalized* (GEN) ESD *Many-Outlier Procedures Given in* (1.2) *and* (1.3), $\alpha = .05$, 2000 *Iterations*

number of true outliers	alternative (γ_1, γ_2)	method	number of outliers detected 0	1	2	number of true outliers	alternative (γ_1, γ_2)	method	number of outliers detected 0	1	2
0	(0, 0)	ESD	.95	.02	.03	1	(0, 2)	ESD	.90	.06	.04
		GEN	.95	.05	.01*			GEN	.88	.11	.01
1	(0, 4)	ESD	.42	.48	.10	1	(0, 6)	ESD	.03	.85	.12
		GEN	.33	.63	.04			GEN	.02	.93	.05
2	(2, 2)	ESD	.87	.06	.06	2	(-2, 2)	ESD	.86	.07	.07
		GEN	.86	.11	.03			GEN	.83	.14	.03
2	(2, 4)	ESD	.49	.34	.18	2	(-2, 4)	ESD	.44	.38	.19
		GEN	.42	.48	.10			GEN	.37	.54	.09
2	(2, 6)	ESD	.05	.73	.22	2	(-2, 6)	ESD	.04	.75	.22
		GEN	.03	.84	.13			GEN	.02	.86	.12
2	(4, 4)	ESD	.26	.11	.64	2	(-4, 4)	ESD	.24	.14	.62
		GEN	.30	.22	.48			GEN	.23	.30	.47
2	(4, 6)	ESD	.04	.18	.79	2	(-4, 6)	ESD	.03	.21	.76
		GEN	.04	.29	.67			GEN	.03	.33	.65
2	(6, 6)	ESD	.01	.01	.99	2	(-6, 6)	ESD	.00	.01	.98
		GEN	.01	.02	.97			GEN	.00	.03	.97

* The probabilities do not always add to 1.00 due to round-off error.

We see that for the one-outlier alternatives $(\gamma_1, \gamma_2) = (0, 2), (0, 4), (0, 6)$ the ESD procedure detects two outliers with probability .04, .10, and .12, respectively, which is at least twice as often as the GEN procedure, which detects two outliers with probability .01, .04, and .05, respectively. However, for the two-outlier alternatives we see that the GEN procedure has less power than the ESD procedure in detecting two outliers. This is to be expected since the ESD procedure overestimates the number of outliers if there are some outliers in the data.

4. TABLES

Although the approximate percentiles in (2.5) can be generated from the percentiles of a t distribution, we felt that many researchers may not have direct access to all but the most common percentage points. Thus, we used a double precision adaptation of the IMSL subroutine MDSTI to compute the required percentage points of a t distribution (International Mathematical and Statistical Libraries 1979), then calculated the ESD percentiles using (2.5) for $n = 25(1)50(10)100(50)500$; $k = 10$, $l + 1 = 1(1)5$, 10, $\alpha = .05$, .01, .005. These percentiles are presented in Table 3. The entries for $l + 1 = 6(1)9$ have not been provided since they can be adequately approximated by linear interpolation. Also, for the larger values of n the percentiles depend weakly on $l + 1$, so only entries for $l + 1 = 1$, 5, and 10 are given. Linear interpolation

Table 3. *Approximate Percentage Points for the Generalized* ESD *Many-Outlier Procedure, Two-Tailed Test*, $n = 25(1)50(10)100(50)500$, $k = 10$, $\alpha = .05, .01, .005$, $\ell + 1 = 1(1)5, 10$

		α					α		
n	ℓ + 1	0.05	0.01	0.005	n	ℓ + 1	0.05	0.01	0.005
25	1	2.82	3.14	3.25	31	1	2.92	3.25	3.38
	2	2.80	3.11	3.23		2	2.91	3.24	3.36
	3	2.78	3.09	3.20		3	2.89	3.22	3.34
	4	2.76	3.06	3.17		4	2.88	3.20	3.32
	5	2.73	3.03	3.14		5	2.86	3.18	3.30
	10	2.59	2.85	2.95		10	2.76	3.06	3.17
26	1	2.84	3.16	3.28	32	1	2.94	3.27	3.40
	2	2.82	3.14	3.25		2	2.92	3.25	3.38
	3	2.80	3.11	3.23		3	2.91	3.24	3.36
	4	2.78	3.09	3.20		4	2.89	3.22	3.34
	5	2.76	3.06	3.17		5	2.88	3.20	3.32
	10	2.62	2.89	2.99		10	2.78	3.09	3.20
27	1	2.86	3.18	3.30	33	1	2.95	3.29	3.41
	2	2.84	3.16	3.28		2	2.94	3.27	3.40
	3	2.82	3.14	3.25		3	2.92	3.25	3.38
	4	2.80	3.11	3.23		4	2.91	3.24	3.36
	5	2.78	3.09	3.20		5	2.89	3.22	3.34
	10	2.65	2.93	3.03		10	2.80	3.11	3.23
28	1	2.88	3.20	3.32	34	1	2.97	3.30	3.43
	2	2.86	3.18	3.30		2	2.95	3.29	3.41
	3	2.84	3.16	3.28		3	2.94	3.27	3.40
	4	2.82	3.14	3.25		4	2.92	3.25	3.38
	5	2.80	3.11	3.23		5	2.91	3.24	3.36
	10	2.68	2.97	3.07		10	2.82	3.14	3.25
29	1	2.89	3.22	3.34	35	1	2.98	3.32	3.44
	2	2.88	3.20	3.32		2	2.97	3.30	3.43
	3	2.86	3.18	3.30		3	2.95	3.29	3.41
	4	2.84	3.16	3.28		4	2.94	3.27	3.40
	5	2.82	3.14	3.25		5	2.92	3.25	3.38
	10	2.71	3.00	3.11		10	2.84	3.16	3.28
30	1	2.91	3.24	3.36	36	1	2.99	3.33	3.46
	2	2.89	3.22	3.34		2	2.98	3.32	3.44
	3	2.88	3.20	3.32		3	2.97	3.30	3.43
	4	2.86	3.18	3.30		4	2.95	3.29	3.41
	5	2.84	3.16	3.28		5	2.94	3.27	3.40
	10	2.73	3.03	3.14		10	2.86	3.18	3.30

Table 3. *Continued*

		α					α		
n	ℓ + 1	0.05	0.01	0.005	n	ℓ + 1	0.05	0.01	0.005
37	1	3.00	3.34	3.47	44	1	3.08	3.43	3.56
	2	2.99	3.33	3.46		2	3.07	3.41	3.55
	3	2.98	3.32	3.44		3	3.06	3.40	3.54
	4	2.97	3.30	3.43		4	3.05	3.39	3.52
	5	2.95	3.29	3.41		5	3.04	3.38	3.51
	10	2.88	3.20	3.32		10	2.98	3.32	3.44
38	1	3.01	3.36	3.49	45	1	3.09	3.44	3.57
	2	3.00	3.34	3.47		2	3.08	3.43	3.56
	3	2.99	3.33	3.46		3	3.07	3.41	3.55
	4	2.98	3.32	3.44		4	3.06	3.40	3.54
	5	2.97	3.30	3.43		5	3.05	3.39	3.52
	10	2.89	3.22	3.34		10	2.99	3.33	3.46
39	1	3.03	3.37	3.50	46	1	3.09	3.45	3.58
	2	3.01	3.36	3.49		2	3.09	3.44	3.57
	3	3.00	3.34	3.47		3	3.08	3.43	3.56
	4	2.99	3.33	3.46		4	3.07	3.41	3.55
	5	2.98	3.32	3.44		5	3.06	3.40	3.54
	10	2.91	3.24	3.36		10	3.00	3.34	3.47
40	1	3.04	3.38	3.51	47	1	3.10	3.46	3.59
	2	3.03	3.37	3.50		2	3.09	3.45	3.58
	3	3.01	3.36	3.49		3	3.09	3.44	3.57
	4	3.00	3.34	3.47		4	3.08	3.43	3.56
	5	2.99	3.33	3.46		5	3.07	3.41	3.55
	10	2.92	3.25	3.38		10	3.01	3.36	3.49
41	1	3.05	3.39	3.52	48	1	3.11	3.46	3.60
	2	3.04	3.38	3.51		2	3.10	3.46	3.59
	3	3.03	3.37	3.50		3	3.09	3.45	3.58
	4	3.01	3.36	3.49		4	3.09	3.44	3.57
	5	3.00	3.34	3.47		5	3.08	3.43	3.56
	10	2.94	3.27	3.40		10	3.03	3.37	3.50
42	1	3.06	3.40	3.54	49	1	3.12	3.47	3.61
	2	3.05	3.39	3.52		2	3.11	3.46	3.60
	3	3.04	3.38	3.51		3	3.10	3.46	3.59
	4	3.03	3.37	3.50		4	3.09	3.45	3.58
	5	3.01	3.36	3.49		5	3.09	3.44	3.57
	10	2.95	3.29	3.41		10	3.04	3.38	3.51
43	1	3.07	3.41	3.55	50	1	3.13	3.48	3.62
	2	3.06	3.40	3.54		2	3.12	3.47	3.61
	3	3.05	3.39	3.52		3	3.11	3.46	3.60
	4	3.04	3.38	3.51		4	3.10	3.46	3.59
	5	3.03	3.37	3.50		5	3.09	3.45	3.58
	10	2.97	3.30	3.43		10	3.05	3.39	3.52

Table 3. *Continued*

n	ℓ + 1	α = 0.05	α = 0.01	α = 0.005
60	1	3.20	3.56	3.70
	2	3.19	3.55	3.69
	3	3.19	3.55	3.69
	4	3.18	3.54	3.68
	5	3.17	3.53	3.67
	10	3.14	3.49	3.63
70	1	3.26	3.62	3.76
	2	3.25	3.62	3.76
	3	3.25	3.61	3.75
	4	3.24	3.60	3.75
	5	3.24	3.60	3.74
	10	3.21	3.57	3.71
80	1	3.31	3.67	3.82
	2	3.30	3.67	3.81
	3	3.30	3.66	3.81
	4	3.29	3.66	3.80
	5	3.29	3.65	3.80
	10	3.26	3.63	3.77
90	1	3.35	3.72	3.86
	2	3.34	3.71	3.86
	3	3.34	3.71	3.85
	4	3.34	3.70	3.85
	5	3.33	3.70	3.84
	10	3.31	3.68	3.82
100	1	3.38	3.75	3.90
	2	3.38	3.75	3.90
	3	3.38	3.75	3.89
	4	3.37	3.74	3.89
	5	3.37	3.74	3.89
	10	3.35	3.72	3.87
150	1	3.52	3.89	4.04
	2	3.51	3.89	4.04
	3	3.51	3.89	4.03
	4	3.51	3.88	4.03
	5	3.51	3.88	4.03
	10	3.50	3.87	4.02
200	1	3.61	3.98	4.13
	2	3.60	3.98	4.13
	3	3.60	3.97	4.12
	4	3.60	3.97	4.12
	5	3.60	3.97	4.12
	10	3.59	3.96	4.11
250	1	3.67	4.04	4.19
	5	3.67	4.04	4.19
	10	3.66	4.03	4.18
300	1	3.72	4.09	4.24
	5	3.72	4.09	4.24
	10	3.71	4.09	4.23
350	1	3.77	4.14	4.28
	5	3.76	4.13	4.28
	10	3.76	4.13	4.28
400	1	3.80	4.17	4.32
	5	3.80	4.17	4.32
	10	3.80	4.16	4.31
450	1	3.84	4.20	4.35
	5	3.83	4.20	4.35
	10	3.83	4.20	4.34
500	1	3.86	4.23	4.38
	5	3.86	4.23	4.37
	10	3.86	4.22	4.37

can also be used to estimate the appropriate percentiles for those large values of n ($n > 50$) not presented in the table. For values of n greater than 500, the normal distribution, for which tables giving extreme percentiles are generally available, can be used in place of the t to obtain approximate percentiles. Furthermore, although the table enables one to detect up to 10 outliers in a sample after suitable interpolation, one can use these tables to detect up to a maximum of k outliers for any k, $1 \leq k \leq 10$, by stopping at the appropriate column in the table. Thus for $k = 2$ one would only use the first two columns of the table. Finally, if one wishes to use (2.5) with a critical value for α other than those given in Table 3, then one can refer to the charts given by Moses (1978) to obtain the appropriate upper percentage points for the t distribution.

To facilitate using (1.3), a FORTRAN IV subroutine is provided in the Appendix, which computes $R_1, \ldots, R_k$ for an arbitrary unordered sample $x_1, \ldots, x_n$.

5. EXAMPLE

In 1980 a study was performed (Sacks, Ornish, Rosner, McLanahan, and Kass 1982) in a group of lactovegetarians (persons eating a primarily vegetarian diet in which the only consumption of animal foods permitted is dairy products other than eggs). Blood pressures were measured and a dietary assessment was obtained from a group of 54 persons using a food-frequency questionnaire consisting of more than 100 individual food items. Standard food composition tables were then used to compute total nutrient scores including total protein, fat consumption, and vitamin intake. The ordered distribution of daily dietary vitamin E intake is given in Table 4. The log transformation was used to better normalize the underlying distribution.

We will assume a maximum of 10 outliers and will compute R_i, $i = 1, \ldots, 10$ as given in the GEN procedure in (1.3). The percentiles, λ_i, were estimated by using linear interpolation between the entries in Table 3 for $n = 50, 60$, respectively. The computations, given in Table 5, show that R_3 is statistically significant at the 5 percent level while $R_4, \ldots, R_{10}$ are not. Thus, we declare the values 6.01, 5.42, and 5.34 as outliers. It happens that these data points correspond to persons taking megadoses of vitamin E (at least 200 IU daily) in the form of vitamin E capsules. By way of contrast, if we had used a one-outlier procedure, then we would not have detected *any* outliers ($R_1 = 3.119 < \lambda_1 = 3.157$) due to the masking effect of multiple outliers.

Table 4. *Ordered Distributions of* $\log_e$ *(daily dietary vitamin E intake) in* 54 *Lactovegetarians*

i	$x_{[i]}$	i	$x_{[i]}$	i	$x_{[i]}$	i	$x_{[i]}$
1	-0.25	15	1.58	29	2.14	43	2.92
2	0.68	16	1.65	30	2.15	44	2.93
3	0.94	17	1.69	31	2.23	45	3.21
4	1.15	18	1.70	32	2.24	46	3.26
5	1.20	19	1.76	33	2.26	47	3.30
6	1.26	20	1.77	34	2.35	48	3.59
7	1.26	21	1.81	35	2.37	49	3.68
8	1.34	22	1.91	36	2.40	50	4.30
9	1.38	23	1.94	37	2.47	51	4.64
10	1.43	24	1.96	38	2.54	52	5.34
11	1.49	25	1.99	39	2.62	53	5.42
12	1.49	26	2.06	40	2.64	54	6.01
13	1.55	27	2.09	41	2.90		
14	1.56	28	2.10	42	2.92		

Table 5. *Computations for Vitamin E Example*

i	n-i	$\bar{x}^{(i)*}$	$s^{(i)*}$	$x^{(i)+}$	$R_{i+1} = \lvert x^{(i)} - \bar{x}^{(i)} \rvert / s^{(i)}$	λ_{i+1} ($\alpha = .05$)
0	54	2.321	1.183	6.01	3.119	3.157
1	53	2.251	1.077	5.42	2.943	3.149
2	52	2.190	.991	5.34	3.179	3.142
3	51	2.128	.894	4.64	2.810	3.134
4	50	2.078	.827	−.25	2.816	3.126
5	49	2.126	.763	4.30	2.848	3.117
6	48	2.080	.702	3.68	2.279	3.108
7	47	2.046	.668	3.59	2.310	3.100
8	46	2.013	.634	0.68	2.102	3.091
9	45	2.042	.608	3.30	2.067	3.083

* $\bar{x}^{(i)}$, $s^{(i)}$ = mean, standard deviation from successively reduced samples after deleting the most outlying observations using the procedure in Section 1, $i = 0, \ldots, 9$.

\+ $x^{(i)}$ = most outlying observation from successively reduced samples = observation in I_i corresponding to $\max \lvert x - \bar{x}^{(i)} \rvert$, $i = 0, \ldots, 9$.

6. SUMMARY

The generalized procedure for detecting from one to k outliers in a data set, given in (1.3), has the advantage over the original procedure in (1.2) of controlling the type I error both under the null hypothesis of no outliers and under each of the alternative hypotheses of 1, 2, ..., $k - 1$ outliers respectively. Thus the two methods will both falsely detect at least one outlier with equal probability. However, when any outliers are detected, the generalized procedure will tend to detect a smaller and more accurate number of outliers than will the original procedure.

The generalized procedure also has the advantage of yielding approximate percentiles based on the t distribution. Unfortunately, this approximation does not work well for small samples ($n < 25$). True significance levels can be as much as two times as large as nominal significance levels for $n \leq 15$. Further work needs to be done in this area either to modify this approximate method so that it will work well for small samples or to conduct Monte Carlo studies to estimate the exact percentiles in this case. Additional work is needed to compare the estimation properties resulting from using the many-outlier procedure in (1.3) with those obtained from using robust estimators such as the median, trimmed mean, and so on. Finally, further work is also needed to study the robustness of the many-outlier procedures in this article in the presence of nonnormal distributions.

APPENDIX: FORTRAN IV PROGRAM TO COMPUTE $R_1, \ldots, R_{NOUT}$ FOR AN ARBITRARY UNORDERED SAMPLE $x_1, \ldots, x_{NSAM}$

```
      SUBROUTINE WT(X,N,NSAM,
     NOUT,WTVEC,N1,Q)
C COMPUTE ESD OUTLIER STATISTICS
      DIMENSION X(N),WTVEC(N1),Q(N)
      II=1
      SUM=0.0
      SUMSQ=0.0
      FN=0.0
      DO 2 I=1,NSAM
      Q(I) = 0.0
      SUM=SUM+X(I)
      SUMSQ=SUMSQ+X(I)**2
      FN=FN+1.0
    2 CONTINUE
    1 SS=SUMSQ-(SUM**2)/FN
      S=(SS/(FN-1.0))**.5
      XBAR=SUM/FN
      BIG=0.0
      IBIG=0
      DO 3 I=1,NSAM
      IF(Q(I).EQ.1.0) GO TO 3
      A=ABS(X(I)-XBAR)
      IF(A.LE.BIG) GO TO 3
      BIG=A
      IBIG=I
    3 CONTINUE
      WTVEC(II)=BIG/S
      Q(IBIG)=1.0
      II=II+1
      IF (II.GT.NOUT) GO TO 999
      SUM=SUM-X(IBIG)
      SUMSQ=SUMSQ-X(IBIG)**2
      FN=FN-1.0
      GO TO 1
  999 RETURN
      END
```

where

NSAM = sample size of data set

NOUT = maximum number of outliers to be detected

N = maximum sample size for all data sets used with this subroutine on a given computer run

N1 = maximum number of outliers to be detected for all data sets used with this subroutine on a given computer run

X = single precision input vector of dimension N, whose first NSAM elements represents the unordered data set to which this subroutine is to be applied

WTVEC = single precision output vector of dimension N1 whose first NOUT elements are $R_1, \ldots, R_{NOUT}$ in (1.1)

Q = single precision input vector of dimension N used internally in the subroutine.

[*Received November 1980. Revised September 1982.*]

REFERENCES

GRUBBS, F. E. (1950), "Sample Criteria for Testing Outlying Observations," *Annals of Mathematical Statistics*, 21, 27–58.

HAWKINS, D. (1978), Letter to the Editor, *Technometrics*, 20, 218.

——— (1980), *Identification of Outliers*, London: Methuen, p. 71.

INTERNATIONAL MATHEMATICAL AND STATISTICAL LIBRARIES (1979), *IMSL Library Reference Manual* (Volume 2, 7th ed.), Houston: IMSL.

MOSES, L. (1978), "Charts for Finding Upper Percentage Points of Student's t in the Range .01 to .00001," *Communications in Statistics—Simulation and Computation*, B7, 479–490.

PRESCOTT, P. (1979), "Critical Values for a Sequential Test for Many Outliers," *Applied Statistics*, 28, 36–39.

QUESENBERRY, C. P., and DAVID, H. A. (1961), "Some Tests for Outliers," *Biometrika*, 48, 379–390.

ROSNER, B. (1975), "On the Detection of Many Outliers," *Technometrics*, 17, 221–227.

SACKS, F., ORNISH, D., ROSNER, B., MCLANAHAN, S., and KASS, E. H. (1982), "Blood Pressure and Plasma Lipoproteins in Strict Vegetarians, Lactovegetarians and Nonvegetarians," unpublished manuscript.

THOMPSON, W. R. (1935), "On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation," *Annals of Mathematical Statistics*, 6, 214–219.