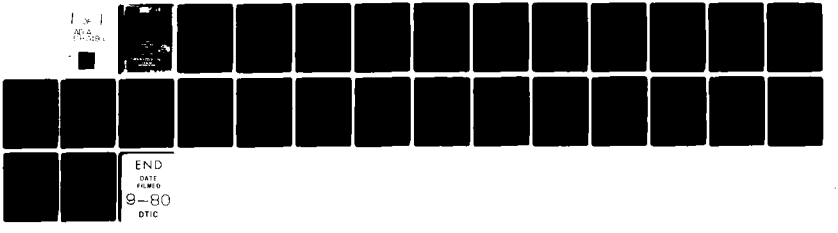
AD-A087 491 DESMATICS INC STATE COLLEGE PA F/G 12/1
COMPARISON OF SEVERAL ESTIMATORS FOR THE VARIANCE OF A NORMAL D--ETC(U)
JUL 80 D E SMITH, D D SCHMOYER N00014-75-C-1054
UNCLASSIFIED TR-106-10 NL



END
DATE FILMED
9-80
DTIC

DESMATICS, INC.

P. O. Box 618
State College, Pa. 16801
Phone: (814) 238-9621

Applied Research in Statistics - Mathematics - Operations Research

(6) COMPARISON OF SEVERAL ESTIMATORS FOR THE VARIANCE OF A NORMAL DISTRIBUTION WHEN OUTLIERS MAY BE PRESENT

by

(10) Dennis E. Smith
Denise D. Schmoyer

(14) TR-106-10

(9) TECHNICAL REPORT NO. 106-10

(11) July 1980 (12) 28

(15) This study was supported by the Office of Naval Research under Contract No. N00014-75-C-1054, Task No. NR 042-334

3 11 15

ABSTRACT

Although much research has been directed at dealing with outliers, particularly for a $N(\mu, \sigma^2)$ distribution, little work has been devoted to estimating σ^2 when outliers may be present. This report describes a comparison of some proposed estimators of σ^2 for the case of data which, except for spurious observations, results from a $N(\mu, \sigma^2)$ distribution. All the estimators considered are nonadaptive and can accommodate up to n/2 outliers from a sample of size n. Monte Carlo simulation was used for the comparison of MSE, which was selected as a measure of performance. Interval estimates based on several of the estimators are also considered.

Key Words

Variance Estimation

Outliers

Nonadaptive Estimators

TABLE OF CONTENTS

I. INTRODUCTION

Much research has been directed at dealing with outliers, particularly in the case of data which, except for spurious observations, results from a $N(\mu, \sigma^2)$ distribution. Major emphasis has been placed on techniques for use when μ is to be estimated while little research has been devoted to estimating σ^2. Primary research for the estimation of σ^2 when outliers may be present has been done by Guttman and Smith (1971) and Johnson, McGuire and Milliken (1978). This technical report describes the results of a small scale comparison of some of their proposed variance estimators and one additional estimator. Both point estimates and interval estimates are considered.

II. DEFINITION OF ESTIMATORS

The variance estimators proposed by Guttman and Smith (GS) were developed for samples of observations hopefully all from a $N(\mu, \sigma^2)$ distribution, but with at most one observation from a $N(\mu + a\sigma, \sigma^2)$ or $N(\mu, (1 + b)\sigma^2)$, $b > 0$. These estimators are adaptive in that the value of a sample statistic determines the form of the estimator. Three different adaptive estimators or rules were developed: A-Rule, W-Rule, and S-Rule. Each rule essentially uses s^2, the usual variance estimator, with a modified data set or the original data set depending on whether or not there is sufficient evidence to conclude that the suspect observation is an outlier. When a modified data set is used, the A-Rule (Anscombe's method) eliminates the suspected outlier, the W-Rule (Winsorization) replaces the suspected outlier with the nearest retained observation and the S-Rule (Semi-Winsorization) replaces the suspected outlier with the critical value for the rule.

The estimators proposed by Johnson, McGuire and Milliken (JMM) are also for samples from a $N(\mu, \sigma^2)$ distribution, but with as many as 50% of the observations arising from a $N(\mu + a\sigma, \sigma^2)$. Although JMM discussed several forms for a variance estimator, only their estimator V_k will be considered here. This estimator, like those of GS, was developed as a modification of s^2. Consider a sample $x_1, x_2 \ldots x_n$, let

$$u_{ij} = |x_i - x_j| \text{ for } i < j = 2, 3 \ldots n$$

and let $u_{(1)} \geq u_{(2)} \geq \ldots \geq u_{(n(n-1)/2)}$

be the ordered u_{ij}'s. Then s^2 can be written as

$$s^2 = \sum_{1}^{n(n-1)/2} u^2_{(i)}/n(n-1).$$

V_k, which assumes k outliers in a sample of size n, modifies s^2 by eliminating the $k(n-k)$ largest u_{ij}'s.

Thus,

$$V_k = \sum_{k(n-k)+1}^{n(n-1)/2} u^2_{(i)}/[k(k-1) + (n-k)(n-k-1)].$$

V_k is not adaptive but rather requires the experimenter to specify the suspected number of outliers (k).

Assuming the same framework as JMM (nonadaptive estimators for use with samples where the outliers result from a $N(\mu + a\sigma, \sigma^2)$ distribution), the principles of the GS rules can be extended to accommodate samples with up to 50% outliers. Here, as with the V_k estimator, the experimenter would specify the number of suspected outliers.

Under the assumption that the outliers are from a normal distribution with shifted mean, the suspected outliers must be either the largest or smallest k observations. In practice the experimenter may know whether the largest or smallest observations are suspect. However, it will be assumed that in general this must be determined. Thus, one of two possible groupings must be selected: the smallest k as outliers and the remaining $n - k$ as $N(\mu, \sigma^2)$ or the largest k as outliers and the remaining $n - k$ as $N(\mu, \sigma^2)$. The ratio of the between-sum-of-squares to the within-sum-of-squares (B/W), for identifying clusters from normal distributions with the same variance but different means, can be used for this purpose. [See

Engelman and Hartigan (1969).] This ratio is determined for each of the two partitions, {k, n - k} and {n - k, k}. The grouping with the maximum B/W represents the most likely clustering of samples from two normal distributions.

The case with an even sample size where n/2 outliers are assumed requires special consideration. Here there is only one possible grouping {n/2, n/2} and thus the B/W approach for identifying the set of outliers can not be applied. If there actually were n/2 outliers, then under the assumption of equal variance for the two distributions it would not matter which half of the sample was designated to be the outliers as either set could be used to obtain an estimate of σ^2. However, if the assumption of n/2 outliers is incorrect, then the "mixed" half, that containing some observations from a $N(\mu, \sigma^2)$ and some outliers, would have a larger sample variance and would lead to an overestimation of σ^2. Therefore, for this special case the sample variances are calculated for each half of the data set and that half with the larger sample variance is labeled as the set of outliers.

Having determined which group of k observations to consider as the outliers, the extensions for two of the GS rules are straightforward. A_k corresponding to the A-Rule eliminates the k suspected outliers and calculates s^2 based on the remaining n - k observations. W_k corresponding to the W-Rule replaces the k suspected outliers with the closest retained observation and calculates s^2 using the modified sample of n observations. Because a critical value has not been specified, an extension to the S-Rule is not so easily defined. Due to the preliminary nature of this study, no attempt was made to define S_k.

An additional estimator for σ^2 considered in this study is the pooled sample variance for the group of k assumed outliers and the group of the remaining n - k observations. This estimator will be denoted by P_k.

III. MONTE CARLO COMPARISONS

A Monte Carlo simulation was conducted as a means for comparing the different nonadaptive variance estimators, which require that the number of outliers must be specified. In addition to evaluating the estimators using the correct number of outliers, the simulation, which was set up to parallel that done by JMM, looks at the performance of the estimators in several cases where the number of outliers is specified incorrectly.

The factors considered in the simulation were:

n - the sample size

k_1 - the number of actual outliers

a - the bias in the mean (in units of σ)

k - the assumed number of outliers

and N - the number of random samples simulated.

Given a sample size (n), an actual number of outliers (k_1), and a bias (a), N random samples were generated with $n - k_1$ observations from the N(0, 1) and k_1 observations from the N(a, 1). For each sample the values of A_k, W_k, P_k and V_k were determined for all k included in the study. Based on these simulated values an estimate of the expected value and MSE for each estimator was obtained. The results of this simulation can be easily transformed to the appropriate values for sampling from a $N(\mu, \sigma^2)$ with outliers from a $N(\mu + a\sigma, \sigma^2)$.

Due to the exploratory nature of this study not all the cases included in the JMM paper were considered. Only sample sizes of ten and twenty-five

were used, and for each of these only a subset of the JMM cases were completed. The cases included in the study are outlined in Figure 1. The number of random samples generated for each case was the same as that used by JMM (10,000/n). This number of samples provides reasonable confidence limits for the estimates in an acceptable amount of computer time.

Samples With n = 10 Observations:

k_1 = 1, 2, 3, 5

a = 3.0, 6.0, 9.0

k = 1, 2, 3, 5

Samples with n = 25 Observations:

k_1 = 0, 1, 2, 3, 5, 7, 10, 12

a = 3.0, 9.0

k = 12

---

k_1 is the number of actual outliers,

a is the bias in the mean (in units of σ),

and k is the assumed number of outliers.

Figure 1: Simulation Cases Completed

IV. RESULTS OF THE SIMULATION

The expected values and mean square errors obtained from the simulation are presented in Figures 2 to 5. The estimated maximum standard error (in percentage) observed for each estimate is included to provide an indication of the accuracy of the simulation. As a check on the simulation procedure the values for V_k and s^2 were calculated and are presented in the tables. These values are in accordance with those presented by JMM.

For some cases it is also possible to check the expected values of the A, W and P estimators obtained from the simulation. When the bias is large (e.g., 9.0), it is reasonable to assume that the simulated outliers are larger than the N(0, 1) observations. Thus, the ordered observations can be split into two groups, the largest k_1 making up a random sample from N(9, 1) and the smallest $n - k_1$ making up a random sample from N(0, 1). The tabulated values of the moments of the order statistics for samples from the normal distribution [Sarhan and Greenberg (1962)] can then be used to obtain the expected values of the A_k, W_k and P_k estimators. Unfortunately, the moments needed to calculate the MSE for these estimators have not been tabulated. The results of the analytical check for samples of ten observations containing one outlier with bias of 9.0 are presented in Figure 6.

The MSE was used to compare the performance of the estimators. From Figures 2 to 4 it can be seen that V_k is the best estimator when the number of outliers is underestimated. When the number of outliers is overestimated

		V_k		A_k		W_k		P_k		s^2	
k	k_1	E	MSE	E	MSE	E	MSE	E	MSE	E	MSE
1	1	.791	.174	.953	.231	1.083	.322	.953	.231	1.872	1.395
	2	1.214	.277	1.581	.803	1.952	1.726	1.581	.803	2.578	3.448
	3	1.635	.745	2.216	2.188	2.601	3.560	2.216	2.188	3.148	5.745
	5	1.937	1.369	2.683	3.847	2.968	5.089	2.683	3.847	3.489	7.543
2	1	.460	.343	.704	.226	.794	.227	.828	.181		
	2	.661	.197	.923	.231	1.084	.339	.914	.197		
	3	.948	.126	1.505	.691	1.933	1.707	1.388	.482		
	5	1.187	.220	2.106	1.978	2.447	3.112	1.910	1.401		
3	1	.303	.511	.569	.311	.599	.288	.797	.209		
	2	.427	.368	.688	.261	.739	.261	.906	.170		
	3	.600	.220	.938	.260	1.073	.378	.924	.180		
	5	.770	.127	1.585	.854	1.905	1.637	1.344	.407		
5	1	.207	.641	.368	.477	.284	.556	.913	.257		
	2	.291	.523	.433	.423	.338	.498	1.188	.372		
	3	.404	.388	.510	.365	.413	.427	1.217	.322		
	5	.505	.283	.550	.328	.477	.375	.893	.188		
Maximum Estimated Standard Error		1.7%	5.8%	2.4%	5.8%	2.3%	7.5%	1.7%	7.0%		

Figure 2: Estimated Expected Values and Mean Square Errors
For Sample Size n = 10 and Bias a = 3.0

		V_k		A_k		W_k		P_k		s^2	
k	k_1	E	MSE	E	MSE	E	MSE	E	MSE	E	MSE
1	1	.989	.217	1.010	.252	1.147	.360	1.010	.252	4.617	14.871
	2	3.301	6.003	4.235	11.841	6.163	29.873	4.235	11.841	7.377	43.736
	3	5.343	20.374	7.015	38.936	8.540	60.874	7.015	38.936	9.534	76.967
	5	7.050	39.012	9.677	79.689	10.253	90.427	9.677	79.689	11.094	106.753
2	1	.522	.298	.757	.242	.838	.243	2.017	1.516		
	2	.954	.190	.985	.244	1.171	.410	.983	.220		
	3	2.635	3.058	4.310	12.408	7.079	41.187	3.856	9.240		
	5	4.303	11.876	8.356	57.883	9.399	75.194	7.378	43.550		
3	1	.335	.473	.628	.331	.641	.295	2.505	3.037		
	2	.548	.279	.715	.253	.762	.243	2.467	2.702		
	3	.990	.213	1.032	.352	1.187	.528	1.041	.260		
	5	2.358	2.090	6.616	34.365	8.169	55.759	5.117	18.514		
5	1	.226	.615	.438	.449	.333	.515	3.163	5.879		
	2	.357	.449	.473	.420	.367	.477	4.423	13.325		
	3	.591	.267	.579	.373	.464	.417	4.373	12.740		
	5	.948	.167	.636	.305	.570	.347	1.009	.245		
Maximum Estimated Standard Error		1.8%	5.2%	2.5%	6.9%	2.5%	7.2%	1.6%	6.2%		

Figure 3: Estimated Expected Values and Mean Square Errors
For Sample Size n = 10 and Bias a = 6.0

		V_k		A_k		W_k		P_k		s^2	
k	k_1	E	MSE	E	MSE	E	MSE	E	MSE	E	MSE
1	1	.999	.241	1.000	.241	1.130	.336	1.000	.241	9.078	68.908
	2	6.982	37.451	8.756	63.186	13.381	160.628	8.756	63.186	15.241	209.555
	3	12.058	126.052	15.327	211.353	18.720	323.012	15.327	211.353	20.028	370.576
	5	16.017	230.683	21.497	428.937	22.364	465.740	21.497	428.937	23.499	515.588
2	1	.518	.305	.754	.236	.839	.237	4.272	11.867		
	2	1.011	.255	1.011	.273	1.190	.455	1.014	.245		
	3	5.727	23.264	9.328	72.378	16.479	249.713	8.236	54.620		
	5	10.040	84.061	19.251	341.098	21.083	412.836	16.906	259.087		
3	1	.331	.479	.613	.319	.630	.297	5.627	23.338		
	2	.570	.271	.747	.254	.798	.251	5.473	21.318		
	3	.998	.263	1.000	.324	1.163	.517	.998	.251		
	5	5.173	18.015	15.843	226.652	19.324	345.414	12.038	125.332		
5	1	.224	.618	.432	.447	.334	.514	7.144	40.680		
	2	.370	.436	.488	.426	.384	.474	10.200	88.537		
	3	.575	.277	.545	.393	.437	.426	10.176	87.635		
	5	.998	.247	.636	.302	.565	.348	1.000	.257		
Maximum Estimated Standard Error		1.8%	6.1%	2.7%	6.8%	2.6%	7.5%	1.6%	6.6%		

Figure 4: Estimated Expected Values and Mean Square Errors
For Sample Size n = 10 and Bias a = 9.0

a	k_1	V_{12} E	V_{12} MSE	A_{12} E	A_{12} MSE	W_{12} E	W_{12} MSE	P_{12} E	P_{12} MSE	s^2 E	s^2 MSE
3.0	0	.144	.735	.309	.498	.265	.552	.368	.414	1.017	.080
	1	.167	.698	.424	.396	.344	.457	.599	.210	1.380	.295
	2	.195	.652	.445	.372	.376	.419	.768	.129	1.717	.728
	3	.220	.612	.477	.359	.406	.392	.890	.093	1.988	1.215
	5	.295	.504	.504	.323	.453	.340	1.033	.086	2.507	2.587
	7	.371	.406	.541	.269	.507	.285	1.047	.080	2.883	3.980
	10	.470	.293	.668	.196	.660	.183	.908	.061	3.237	5.470
	12	.499	.263	.791	.128	.800	.103	.842	.071	3.312	5.833
9.0	0	.137	.747	.291	.518	.247	.575	.352	.434	.961	.079
	1	.170	.692	.447	.645	.374	.499	3.108	4.864	4.301	11.432
	2	.208	.631	.420	.375	.382	.409	5.178	18.243	7.192	39.426
	3	.245	.576	.430	.370	.392	.401	6.804	34.509	9.859	79.830
	5	.356	.427	.479	.320	.444	.346	8.574	58.465	14.437	182.944
	7	.513	.261	.545	.264	.521	.273	8.402	55.746	17.958	290.672
	10	.793	.096	.686	.194	.716	.180	4.891	15.486	21.286	414.770
	12	1.017	.080	1.010	.168	1.326	.503	1.020	.082	22.056	447.288
Maximum Estimated Standard Error		1.9%	7.0%	6.2%	17.6%	4.2%	11.1%	1.9%	7.4%		

Figure 5: Estimated Expected Values and Mean Errors
For Sample Size n = 25 and k = 12

k	A_k		W_k		P_k	
	Analytic	Simulation	Analytic	Simulation	Analytic	Simulation
1	1.000	1.000 ± .016*	1.134	1.130 ± .018	1.000	1.000 ± .016
2	.749	.754 ± .013	.829	.839 ± .015	4.270	4.272 ± .034
3	.610	.613 ± .013	.626	.630 ± .013	5.652	5.627 ± .044
5	.436	.432 ± .011	.333	.334 ± .008	7.171	7.144 ± .054

*The simulation entries indicate estimates of the expected value ± one standard deviation.

Figure 6: Analytical Check on Expected Values Obtained From the Simulation For Sample Size n = 10, k_1 = 1 and a = 9.0

and the bias is small (a = 3.0), P_k has the smallest MSE. However, for larger bias (a = 6.0 or a = 9.0), P_k breaks down. As can be seen from Figures 3, 4 and 5 no single estimator is unconditionally superior when the bias is large and the number of outliers is overestimated.

Despite the breakdown of P_k with large biases, its exceptional performance with a small bias suggests that a modification of the estimator should be considered. A simplistic form of an adaptive rule to replace P_k was defined and briefly evaluated. This modified pooled estimator is

$$MP_k = \begin{cases} P_k \text{ if } s_1^2/s_2^2 < F_{1-\alpha}(\nu_1, \nu_2), \\ A_k \text{ otherwise} \end{cases}$$

where s_1^2 is the sample variance of the k suspected outliers and s_2^2 is the sample variance of the n - k remaining observations. The MSE of this modified estimator is bounded by that of P_k and A_k.

This estimator was evaluated for sample sizes of 25 assuming twelve outliers and $\alpha = .05$. The performance of P_{12} and MP_{12} is presented in Figure 7. It can be seen that the MP_k rule does not do as well as P_k in the case of a small bias; however, it does much better than P_k with large biases. Thus, even a very simple adaptive rule can greatly improve an estimate. This suggests that adaptive rules should be further investigated, particularly rules which would not require specification of the number of outliers.

When assuming twelve outliers with large bias in a sample of twenty-five observations, the MSE values (Figure 5) display some peculiar results which warrant explanation. Specifically, there is an unexpected jump in the MSE

a	k_1	P_{12} E	P_{12} MSE	MP_{12} E	MP_{12} MSE
3.0	0	.368	.414	.333	.462
	1	.599	.210	.429	.371
	2	.768	.129	.470	.341
	3	.890	.093	.500	.327
	5	1.033	.086	.559	.294
	7	1.047	.080	.629	.245
	10	.908	.061	.745	.156
	12	.842	.071	.806	.095
9.0	0	.352	.434	.316	.484
	1	3.108	4.864	.411	.467
	2	5.178	18.243	.420	.375
	3	6.804	34.509	.430	.370
	5	8.574	58.465	.479	.320
	7	8.402	55.746	.545	.264
	10	4.891	15.486	.686	.194
	12	1.020	.082	.984	.097
Maximum Estimated Standard Error		1.9%	7.4%	4.2%	6.7%

Figure 7: Estimated Expected Values and Mean Square Errors For Sample Size n = 12 and k = 12

of the A estimator for the case of one actual outlier and an unexpected jump in the MSE of the W estimator for the case of twelve actual outliers. The problem with the A estimator is a result of the method used to identify the set of outliers. When there is actually only one outlier but twelve outliers are assumed the B/W ratio for the two groupings {12, 13} and {13, 12} are very close in value; thus the wrong set of twelve observations is sometimes designated as the outliers. When the bias in the outliers is large and the wrong set of observations (i.e., the set containing the outlier) is used, the A estimator overestimates σ^2 leading to a large MSE. Once again this points out the need for adaptive rules which do not require the specification of the number of outliers.

The large MSE for the W_{12} estimator when twelve outliers with a large bias are present is not a result of incorrectly identifying the outliers but rather it is an inherent property of the estimator. In this case the B/W ratio method correctly selects the set of twelve outliers and the W estimator is applied to the sample of thirteen observations from a N(0, 1). Depending on whether the bias is positive or negative, the largest or smallest ordered observation ($x_{(13)}$ or $x_{(1)}$) is weighted heavily by the W estimator. The heavy weighting of this extreme observation can lead to overestimation of σ^2 and consequently a large MSE. Based on these results the W estimator appears to be inappropriate when outliers make up a large proportion of the sample.

V. USING THE NONADAPTIVE ESTIMATORS

Due to the bias of the estimator, V_k, and the difficulty of specifying the number of outliers, JMM suggested that V_k should be used in conjunction with an estimator V_k^*. V_k^* is simply V_k modified to be unbiased in the case of no outliers. That is, $V_k^* = V_k/v_k$ where $v_k = E(V_k/\sigma^2)$ given no outliers.

When the number of outliers is overestimated, V_k underestimates σ^2 whereas V_k^* overestimates σ^2. JMM suggested that when outliers are suspected, (V_L, V_L^*) should be used as an interval estimate for σ^2, where L is an upper bound for the number of outliers in the sample. If the experimenter has no estimate for L, JMM propose using n/2.

It can be seen from Figures 2 through 5 that when the number of outliers is overestimated, A_k and W_k also underestimate σ^2. Thus, the intervals (A_L, A_L^*) and (W_L, W_L^*) could also serve as interval estimates for σ^2. These interval estimates were evaluated for a sample size of ten with five assumed outliers and for a sample size of twenty-five with twelve assumed outliers. The expected values of V_k, A_k and W_k in the null case, which are needed to calculate V_5^*, A_5^*, W_5^*, V_{12}^*, A_{12}^* and W_{12}^*, were obtained by Monte Carlo simulation. These values are 0.141, 0.224, 0.176, 0.141, 0.300 and 0.256, respectively. The interval estimates are presented in Figures 8 and 9. A comparison of interval lengths reveals that for two or more actual outliers, the A intervals put the tightest bounds on σ^2, followed by the W and then the V intervals.

k_1	Bias	(V_5, V_5^*)	Length	(A_5, A_5^*)	Length	(W_5, W_5^*)	Length
1	3.0	(.207, 1.468)	1.261	(.368, 1.643)	1.275	(.284, 1.614)	1.330
	6.0	(.226, 1.603)	1.377	(.438, 1.955)	1.517	(.333, 1.892)	1.559
	9.0	(.224, 1.589)	1.365	(.432, 1.929)	1.497	(.334, 1.898)	1.564
2	3.0	(.291, 2.064)	1.773	(.433, 1.933)	1.500	(.338, 1.920)	1.582
	6.0	(.357, 2.532)	2.175	(.473, 2.112)	1.639	(.367, 2.085)	1.718
	9.0	(.370, 2.624)	2.254	(.488, 2.179)	1.691	(.384, 2.182)	1.798
3	3.0	(.404, 2.865)	2.461	(.510, 2.277)	1.767	(.413, 2.347)	1.934
	6.0	(.591, 4.191)	3.600	(.579, 2.585)	2.006	(.464, 2.636)	2.172
	9.0	(.575, 4.078)	3.503	(.545, 2.433)	1.888	(.437, 2.483)	2.046
5	3.0	(.505, 3.582)	3.077	(.550, 2.455)	1.905	(.477, 2.710)	2.233
	6.0	(.948, 6.723)	5.775	(.636, 2.839)	2.203	(.570, 3.239)	2.669
	9.0	(.998, 7.078)	6.080	(.636, 2.839)	2.203	(.565, 3.210)	2.645

Figure 8: Interval Estimates For σ^2 With Sample Size n = 10

k_1	Bias	(V_{12}, V_{12}^*)	Length	(A_{12}, A_{12}^*)	Length	(W_{12}, W_{12}^*)	Length
1	3.0	(.167, 1.184)	1.017	(.424, 1.413)	.989	(.344, 1.344)	1.000
	9.0	(.170, 1.206)	1.036	(.447, 1.490)	1.043	(.374, 1.461)	1.087
2	3.0	(.195, 1.383)	1.188	(.445, 1.483)	1.038	(.376, 1.469)	1.093
	9.0	(.208, 1.475)	1.267	(.420, 1.400)	.980	(.382, 1.492)	1.110
3	3.0	(.220, 1.560)	1.340	(.477, 1.590)	1.113	(.406, 1.586)	1.180
	9.0	(.245, 1.738)	1.493	(.430, 1.433)	1.003	(.392, 1.531)	1.139
5	3.0	(.295, 2.092)	1.797	(.504, 1.680)	1.176	(.453, 1.770)	1.317
	9.0	(.356, 2.525)	2.169	(.479, 1.597)	1.118	(.444, 1.734)	1.290
7	3.0	(.371, 2.631)	2.260	(.541, 1.803)	1.262	(.507, 1.980)	1.473
	9.0	(.513, 3.638)	3.125	(.545, 1.817)	1.272	(.521, 2.035)	1.514
10	3.0	(.470, 3.333)	2.863	(.668, 2.227)	1.559	(.660, 2.578)	1.918
	9.0	(.793, 5.624)	4.831	(.686, 2.287)	1.601	(.716, 2.797)	2.081
12	3.0	(.499, 3.539)	3.040	(.791, 2.637)	1.846	(.800, 3.125)	2.325
	9.0	(1.017, 7.213)	6.196	(1.010, 3.367)	2.357	(1.326, 5.180)	3.854

Figure 9: Interval Estimates For σ^2 With Sample Size n = 25

VI. CONCLUSIONS

This Monte Carlo investigation has shown that the performance of the nonadaptive estimators greatly depends on the assumed number of outliers in relation to the true number. This points out the need for adaptive procedures which could be applied without requiring the experimenter to estimate the number of outliers.

When nonadaptive estimators are to be used, an interval estimation procedure may prove useful. This Monte Carlo study has indicated that interval estimates based on A_k or W_k tend to be superior (i.e., produce tighter bounds) to the V_k interval estimate suggested by JMM.

VII. REFERENCES

Engelman, L. and Hartigan, J. A. (1969). "Percentage Points of a Test For Clusters," Journal of the American Statistical Association, Vol. 64, pp. 1647-1648.

Guttman, I. and Smith, D. E. (1971). "Investigation of Rules For Dealing With Outliers in Small Samples From the Normal Distribution II: Estimation of the Variance," Technometrics, Vol. 13, pp. 101-111.

Johnson, D. E., McGuire, S. A. and Milliken, G. A. (1978). "Estimating σ^2 in the Presence of Outliers," Technometrics, Vol. 20, pp. 441-456.

Sarhan, A. E. and Greenberg, B. G. (Eds.). (1962). Contributions to Order Statistics, Wiley, New York.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER: 106-10 ✓	2. GOVT ACCESSION NO.: AD-A087491
3 RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): COMPARISON OF SEVERAL ESTIMATORS FOR THE VARIANCE OF A NORMAL DISTRIBUTION WHEN OUTLIERS MAY BE PRESENT	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
	6. PERFORMING ORG. REPORT NUMBER:
7 AUTHOR(s): Dennis E. Smith and Denise D. Schmoyer	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-1054 ✓
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Desmatics, Inc ✓ P. O. Box 618 State College, PA 16801	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 042-334
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research Arlington, VA 22217	12. REPORT DATE: July 1980
	13. NUMBER OF PAGES: 22
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this report is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Variance Estimation
Outliers
Nonadaptive Estimators

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

Although much research has been directed at dealing with outliers, particularly for a $N(\mu, \sigma^2)$ distribution, little work has been devoted to estimating σ^2 when outliers may be present. This report describes a comparison of some proposed estimators of σ^2 for the case of data which, except for spurious observations, results from a $N(\mu, \sigma^2)$ distribution. All the estimators considered are nonadaptive and can accommodate up to n/2 outliers from a sample of size n. Monte Carlo simulation was used for the

(over)

DD FORM 1 JAN 7 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20.
comparison of MSE, which was selected as a measure of performance. Interval estimates based on several of the estimators are also considered.