The Annals of Mathematical Statistics
1968, Vol. 39, No. 2, 567–582

JACKKNIFING VARIANCES[1]

BY RUPERT G. MILLER, JR.

Stanford University

1. Introduction. The Tukey jackknife [18], [19], [11], which is an extension of an idea of Quenouille [12], is a rough-and-ready statistical tool which (a) reduces bias and (b) produces approximate confidence intervals. It exactly eliminates a $1/n$ bias term. Its approximate confidence intervals are a godsend in problems where messy distribution theory prohibits the formation of exact confidence limits.

It has been demonstrated that the jackknife can be beneficially applied in ratio problems (Quenouille [13], Durbin [6], Rao [14], Rao and Webster [15], Deming [5]), in maximum likelihood estimation (Brillinger [3]), and in transformations of statistics (Miller [9]). A recent proposed application is the construction of confidence limits for estimates of parameters in a functional relationship (Brillinger [4]).

Indiscriminate universal application of the jackknife can be hazardous. This is illustrated in the case of interval estimation for a truncation point (Miller [9]), although under restrictions on the probability density the jackknife still performs satisfactorily (Robson and Whitlock [16]). Lincoln Moses in unpublished work has shown that the jackknife runs into trouble for interval estimation on the median.

The purpose of this paper is to examine how the jackknife performs in testing hypotheses on variances. It is well known for this problem that it is disastrous to base a test on the χ^2 or F distribution because of extreme sensitivity of the distribution to nonnormality. A variety of alternatives to the classical techniques have been proposed. Some of these involve arbitrarily dividing the data into groups. As soon as the idea of division into groups creeps forth, the jackknife cries out to be tested.

Two objectives are accomplished in this paper: (1) Another technique is added to the short list of tests which are robust and reasonably powerful for testing variances. (2) Another problem is recorded in which the jackknife performs admirably so statisticians should be imbued with courage to try the jackknife elsewhere.

Section 2 contains a description of the jackknife technique. The asymptotic distribution theory is worked out in Section 3. In Section 4 the jackknife is compared with other techniques for testing variances, both for large samples and for small samples. Section 5 closes the paper with a discussion of the performance of the jackknife with regard to testing variances and in general. The Appendix contains a proof that Levene's z test is not asymptotically distribution-free.

Received 9 June 1967.

[1] Research supported in part by Public Health Service Grant USPHS-5T1 GM 25–09.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

2. The jackknife. For the reader unfamiliar with the jackknife a description is included: Let θ be an unknown parameter, and let $(X_1, \cdots, X_N)$ be a sample of N independent, identically distributed observations with cdf F_θ, which depends on θ. Suppose a method (biased or unbiased) is available for estimating θ. Further suppose that the data is divided into n groups of size $k(N = nk)$; i.e., $(X_1, \cdots, X_k), (X_{k+1}, \cdots, X_{2k}), \cdots, (X_{(n-1)k+1}, \cdots, X_{nk})$. This division may be determined by the structure of the experiment or arbitrarily imposed by the statistician. Let $\hat{\theta}$ denote the estimate of θ based on all $N = nk$ observations, and let $\hat{\theta}_{-i}, i = 1, \cdots, n$, denote the estimate of θ obtained by deleting the ith group and estimating θ from the remaining $(n - 1)k$ observations. Form the new estimates (called "pseudo-values" by Tukey)

$$(1) \qquad \tilde{\theta}_i = n\hat{\theta} - (n - 1)\hat{\theta}_{-i}, \qquad i = 1, \cdots, n.$$

The jackknife estimate of θ is the average of the $\tilde{\theta}_i$; i.e.,

$$(2) \qquad \tilde{\theta}. = n^{-1} \textstyle\sum_{i=1}^{n} \tilde{\theta}_i = n\hat{\theta} - (n - 1)\hat{\theta}_{-}.,$$

where $\hat{\theta}_{-}. = (\sum_{i=1}^{n} \hat{\theta}_{-i})/n$.

The jackknife exactly eliminates a n^{-1} bias term. Namely, if

$$(3) \qquad E(\hat{\theta}) = \theta + cN^{-1} + O(N^{-2}),$$

then

$$(4) \qquad E(\tilde{\theta}.) = \theta + O(n^{-2}).$$

Quenouille conceived the jackknife to achieve this reduction in bias. Tukey went one step further and proposed that in many instances the $\tilde{\theta}_i$ are approximately *independently*, identically distributed. If this proposal is correct, then

$$(5) \qquad (n(n - 1))^{-1} \textstyle\sum_{i=1}^{n} (\tilde{\theta}_i - \tilde{\theta}.)^2$$

should be an (approximate) estimate of Var $(\tilde{\theta}.)$, and

$$(6) \qquad (\tilde{\theta}. - \theta)[(n(n - 1))^{-1} \textstyle\sum_{i=1}^{n} (\tilde{\theta}_i - \tilde{\theta}.)^2]^{-\frac{1}{2}}$$

should be approximately distributed as a t variate with $n - 1$ d.f. The ratio (6) could therefore be used to construct an approximate confidence interval for θ or perform an approximate significance test on θ.

For the problem considered specifically in this paper let $X_1, \cdots, X_N$ be independently, identically distributed according to $F((x - \mu_x)\sigma_x^{-1})$, and $Y_1, \cdots, Y_M$ be independently, identically distributed according to $F((y - \mu_y)\sigma_y^{-1})$. The X_i and Y_j are governed by the same cdf except for possibly different location and scale parameters. The location and scale parameters ($\mu_x, \mu_y; \sigma_x, \sigma_y$, respectively) are unknown, and the cdf F is unknown with the exception that it will be assumed to have finite fourth moments.

The one sample problem for variances is to construct a confidence interval for $\sigma_x^2 = E(X - E(X))^2$, or to test the hypothesis $H_0 : \sigma_x^2 = \sigma_0^2$, where σ_0^2 is a specified constant. The two sample problem is to test the hypothesis $H_0 : \sigma_y^2 =$

σ_x^2, or to construct a confidence interval on the ratio or difference of the variances.

If the variance σ_x^2 is an unknown parameter (as well as $\mu_x = E(X)$), then a classically "good" estimator of σ_x^2 is $s_x^2 = \sum_{i=1}^N (X_i - \bar{X})^2/(N - 1)$. If σ_x^2 is identified with θ and s_x^2 with $\hat{\theta}$, then the jackknife for group size $k = 1$ $(n = N)$ gives

$$\begin{aligned} \tilde{\theta}_i &= s_x^2 + n(n-2)^{-1}[(X_i - \bar{X})^2 - n^{-1} \\ &\quad \cdot \textstyle\sum_{j=1}^n (X_j - \bar{X})^2], \\ \tilde{\theta}. &= s_x^2, \\ \textstyle\sum_{i=1}^n (\tilde{\theta}_i - \tilde{\theta}.)^2 &= n^2(n-2)^{-2} \textstyle\sum_{i=1}^n ((X_i - \bar{X})^2 - n^{-1} \\ &\quad \cdot \textstyle\sum_{j=1}^n (X_j - \bar{X})^2)^2. \end{aligned} \tag{7}$$

Construction of a test or confidence interval from the quantities in (7) is almost equivalent to using Levene's [8] s procedure which treats $(X_i - \bar{X})^2$, $i = 1, \cdots, N$, as independently identically distributed observations. The equivalence is not exact because the constants involving n appearing in the t ratios are not exactly the same.

For group size $k > 1$ the jackknife does not give simple expressions as appear in (7). In particular,

$$\begin{aligned} \tilde{\theta}_i &= s_x^2 + (n-1)((n-1)k - 1)^{-1} \\ &\quad \cdot [\textstyle\sum_{j \epsilon I_i} (X_j - \bar{X}_i)^2 + nk(n-1)^{-1}(\bar{X}_i - \bar{X})^2 - k(nk-1)^{-1} \\ &\quad \cdot \textstyle\sum_j (X_j - \bar{X})^2], \\ \tilde{\theta}. &= s_x^2 + ((n-1)/n)((n-1)k - 1)^{-1} \\ &\quad \cdot [k \textstyle\sum_{i=1}^n (\bar{X}_i - \bar{X})^2(n-1)^{-1} - \textstyle\sum_{j=1}^{nk} (X_j - \bar{X})^2(nk-1)^{-1}], \end{aligned} \tag{8}$$

where I_i denotes the integers of the ith group (i.e., $I_i = \{(i-1)k + 1, (i-1)k + 2, \cdots, ik\}$) and $\bar{X}_i$ denotes the mean of the observations in the ith group (i.e., $\bar{X}_i = \sum_{j \epsilon I_i} X_j/k$). The sum of squares $\sum_{i=1}^n (\tilde{\theta}_i - \tilde{\theta}.)^2$ is whatever it is—a mess.

Statistical lore tells us that applying the log transformation to s^2 very often produces beneficial results. It stabilizes the variance, creates a more normal-looking distribution, etc. A reasonable procedure might then be to jackknife log s^2 instead of just s^2. This is the procedure that is proposed and analyzed in this paper. Specifically,

$$\begin{aligned} \theta &= \log \sigma^2, \qquad \hat{\theta} = \log s^2, \\ \tilde{\theta}_i &= n \log s^2 - (n-1) \log s_{-i}^2, \\ \tilde{\theta}. &= n \log s^2 - (n-1)n^{-1} \textstyle\sum_{i=1}^n \log s_{-i}^2, \\ (n-1)^{-1} \textstyle\sum_{i=1}^n (\tilde{\theta}_i - \tilde{\theta}.)^2 &= (n-1) \textstyle\sum_{i=1}^n (\log s_{-i}^2 - n^{-1} \textstyle\sum_{j=1}^n \log s_{-j}^2)^2, \end{aligned} \tag{9}$$

where s_{-i}^2 is the sample variance for the $(n - 1)k$ observations obtained by deleting the ith group of k observations.

For the one sample problem the statistician would use (6) with $\tilde{\theta}_i$ and $\tilde{\theta}.$ given by (9) to test or put a confidence interval on $\log \sigma_x^2$. For the two sample problem the statistic

$$[{}_x\tilde{\theta}. - \theta_x) - ({}_y\tilde{\theta}. - \theta_y)][(n(n - 1))^{-1} \textstyle\sum_{i=1}^n ({}_x\tilde{\theta}_i - {}_x\tilde{\theta}.)^2 \tag{10}$$
$$+ (m(m - 1))^{-1} \textstyle\sum_{i=1}^m ({}_y\tilde{\theta}_i - {}_y\tilde{\theta}.)^2]^{-\frac{1}{2}}$$

is used for the confidence interval or test where $\theta_x = \log \sigma_x^2$, $\theta_y = \log \sigma_y^2$, and ${}_x\tilde{\theta}_i$, ${}_y\tilde{\theta}_i$ are based on $\hat{\theta}_x = \log s_x^2$, $\hat{\theta}_y = \log s_y^2$, respectively. For equal sample sizes $N = M$ ($n = m$ and k are the same in each sample) the appropriate degrees of freedom for (10) would be $n + m - 2$. The best choice for k would be 1, but for large samples economy of computing time may dictate choosing a larger k. For unequal sample sizes $N \neq M$ the appropriate degrees of freedom is somewhat ambiguous just as for the two sample t statistic of mean differences with unequal variances.

3. Asymptotic distribution theory. It will be demonstrated in this section that with the identifications (9), (6) is asymptotically normally distributed (mean 0, variance 1) as $n \to +\infty$. The method of proof is to show that $\tilde{\theta}.$ is asymptotically normally distributed and the denominator in (6) converges in probability to the correct standard deviation. The extension to the asymptotic normality of (10) is immediate.

The proof will be given for the case $k = 1$ $(n = N)$. The proof for $k > 1$ is completely analogous, but the algebra is considerably messier. For simplicity the subscript x will be dropped from s^2, etc., because only a single sample is involved.

A power series expansion is used to prove the asymptotic normality of $\tilde{\theta}.$. The technique is identical to that employed in [9]. For some ζ_i between s^2 and s_{-i}^2

$$\log s_{-i}^2 = \log s^2 + s^{-2}(s_{-i}^2 - s^2) + \tfrac{1}{2}(s_{-i}^2 - s^2)(-\zeta_i^{-2}). \tag{11}$$

Thus,

$$\tilde{\theta}_i = \log s^2 - (n - 1)s^{-2}(s_{-i}^2 - s^2) + \tfrac{1}{2}(n - 1)((s_{-i}^2 - s^2)\zeta_i^{-1})^2,$$
$$\tilde{\theta}. = \log s^2 - s^{-2}(n - 1)n^{-1} \textstyle\sum_{i=1}^n (s_{-i}^2 - s^2) + (n - 1)(2n)^{-1} \tag{12}$$
$$\cdot \textstyle\sum_{i=1}^n ((s^2_{-i} - s^2)\zeta_i^{-1})^2.$$

The reader can verify with a little bit of algebra that

$$s_{-i}^2 - s^2 = - n[(n - 1)(n - 2)]^{-1}[(X_i - \bar{X})^2 - n^{-1}\textstyle\sum_{j=1}^n(X_j - \bar{X})^2]. \tag{13}$$

Consequently, $\sum_{i=1}^n(s^2_{-i} - s^2) = 0$, and $\tilde{\theta}.$ in (12) reduces to

$$\tilde{\theta}. = \log s^2 + (n - 1)(2n)^{-1}\textstyle\sum_{i=1}^n((s_{-i}^2 - s^2)\zeta_i^{-1})^2. \tag{14}$$

Since s^2 is asymptotically normally distributed with mean σ^2 and variance $(\mu_4 - \sigma^4)/n$ where $\mu_4 = E(X - E(X))^4$, the variable $\log s^2$ is asymptotically normal with mean $\log \sigma^2$ and variance $((\mu_4/\sigma^4) - 1)/n$. If the remainder term in (14) multiplied by $n^{\frac{1}{2}}$ converges in probability to zero, then $n^{\frac{1}{2}}(\tilde{\theta}. - \theta)$ will be asymptotically normal with mean 0, variance $(\mu_4/\sigma^4) - 1$ by Slutsky's theorem.

From (13) it is easy to show (cf., [9], p. 1599) that the s^2_{-i} are uniformly close to s^2 with probability tending to 1 as $n \to \infty$; i.e., for $\epsilon > 0$, $P\{|s^2_{-i} - s^2| \leqq \epsilon, i = 1, \dots, n\} \to 1$ as $n \to +\infty$. Since $s^2 \to_p \sigma^2$,

$$(15) \quad n^{\frac{1}{2}}(n-1)(2n)^{-1} \sum_{i=1}^n ((s^2_{-i} - s^2)\zeta_j^{-1})^2 \leqq (n-1)(\tfrac{1}{2}n^{-\frac{1}{2}})K \sum_{i=1}^n (s^2_{-i} - s^2)^2,$$

with probability tending to 1. K is a constant chosen arbitrarily subject to $1/\sigma^4 < K < +\infty$. Again, a little algebra verifies

$$(16) \quad (n-1) \sum_{i=1}^n (s^2_{-i} - s^2)^2 = n^2(n-2)^{-2}(n-1)^{-1}[\sum_{i=1}^n U_i^2 - n\bar{U}^2],$$

where $U_i = (X_i - \bar{X})^2$. As $n \to +\infty$, (16) converges in probability to $\mu_4 - \sigma^4$. Consequently, the right hand side of (15) converges in probability to zero, and the remainder term in $n^{-\frac{1}{2}}(\tilde{\theta}. - \theta)$ is forced to zero.

To complete the proof that (6) has a limiting unit normal distribution, it is necessary for the convergence

$$(17) \qquad (n-1)^{-1} \sum_{i=1}^n (\tilde{\theta}_i - \tilde{\theta}.)^2 \to_p \mu_4\sigma^{-4} - 1$$

to hold. From the power series expansions (12)

$$(18) \qquad \tilde{\theta}_i - \tilde{\theta}. = -(n-1)s^{-2}(s^2_{-i} - s^2) + r_i - \bar{r}.,$$

where

$$(19) \qquad r_i = \tfrac{1}{2}((n-1)(\zeta_i^{-2}))(s^2_{-i} - s^2)^2.$$

From (18) the sample variance of the pseudo-values $\tilde{\theta}_i$ equals

$$(20) \quad (n-1)^{-1} \sum_{i=1}^n (\tilde{\theta}_i - \tilde{\theta}.)^2 = (n-1)s^{-4} \sum_{i=1}^n (s^2_{-i} - s^2)^2 - 2s^{-2} \sum_{i=1}^n (s^2_{-i} - s^2)(r_i - \bar{r}.) + (n-1)^{-1} \sum_{i=1}^n (r_i - \bar{r}.)^2.$$

The first term on the right hand side of (20) converges to $(\mu_4 - \sigma^4)/\sigma^4$ by virtue of (16). If $\sum_{i=1}^n (r_i - \bar{r}.)^2/(n-1) \to_p 0$, then the last term in (20) vanishes, and the cross-product term will as well by the Cauchy-Schwarz inequality.

With probability converging to 1 as $n \to +\infty$, the following inequality holds:

$$(21) \quad (n-1)^{-1} \sum_{i=1}^n r_i^2 = \tfrac{1}{4}(n-1) \sum_{i=1}^n (s^2_{-i} - s^2)^4 \zeta_i^{-4} \leqq \tfrac{1}{4}K^2 \max_{1 \leqq i \leqq n} \{(s^2_{-i} - s^2)^2\}(n-1) \sum_{i=1}^n (s^2_{-i} - s^2)^2.$$

The term $\max_{1 \leqq i \leqq n}\{(s^2_{-i} - s^2)^2\}$ converges to zero in probability, and $(n-1) \sum_{i=1}^n (s^2_{-i} - s^2)^2 \to_p \mu_4 - \sigma^4$. Thus, both sides of (21) converge to zero. Since

$$(22) \qquad 0 \leqq \bar{r}.^2 \leqq n^{-1} \sum_{i=1}^n r_i^2,$$

the convergence $\sum_{i=1}^{n} (r_i - \bar{r}.)^2/(n - 1) \rightarrow_p 0$ is proved, and the proof is complete.

The arguments just presented validate the approximate distribution claims for (6) and (10) when n and m are large. These claims for small n and m are supported by Monte Carlo results which appear in the next section.

4. Comparison with alternative techniques. The classical tests and confidence intervals on variances in the one and two sample problems are based on the χ^2 and F distributions. However, it has been known for some time that these tests and intervals are extremely sensitive to non-normality. A good summary of the work in this direction appears in Box [1].

To remedy this situation in the two sample problem an array of non-parametric tests based on ranks has accumulated in the literature. The list includes tests proposed by (a) Lehmann, (b) Sukhatme, (c) Mood, (d) Barton–David–Ansari–Freund–Siegel–Tukey, (e) Klotz, and (f) Capon. The Sukhatme test (b) requires that the medians of both samples be known and equal to zero. Tests (c), (d), (e), and (f) allow the medians to be unknown but require that they be equal. For a description and discussion of these tests the reader is referred to Klotz [7].

These tests will not be considered competitors to the jackknife because they have difficulties inherent in their universal application. The Lehman test (a) is not distribution-free. Tests (b) through (f) are distribution-free, but the assumption of known, or at least equal, medians is unrealistic in most applications. These tests can be modified by centering each sample at its sample median or mean. However, the tests are then not distribution-free; they are not even distribution-free asymptotically. Specifically, it is known that tests (b), (c), and (e) are asymptotically distribution-free if the density function is symmetric, but are not asymptotically distribution-free for general densities. No results have been published on the asymptotically distribution-free character, or lack thereof, for tests (d) and (f) when they are centered at sample medians or means.

Moses [10] has also voiced additional objections to the use of these rank tests.

A small collection of techniques that can be applied safely and universally to variance problems is available. Attention will be focused on these as competitors to the jackknife. Each technique will be described and discussed briefly. With one exception they, like the jackknife, are based on approximate significance levels. However, they are asymptotically distribution-free, and small sample Monte Carlo experiments indicate that their small sample size significance levels are not very sensitive to the form of the underlying distribution.

The earliest technique in this group is the *Box test* [1]. For the two sample problem divide each sample into subsamples of size k $(k > 1)$. Compute $\log s^2$ for each subsample. There will be n such numbers for the X sample, m for the Y sample. Compare these two sets of values by a two sample t test for location. The t statistic will not have exactly a t distribution since $\log s^2$ is not exactly normally distributed, but the significance or confidence level should be closely approximate because of the robustness of the t statistic. The choice of k rests on the shoulders of the statistician. The main disadvantage of the Box test is the loss of information in subdividing the samples. A confidence interval for σ_y^2/σ_x^2 is easily constructed, and application of this technique in the one sample problem is straightforward.

The *Moses test* in the two sample problem is the Wilcoxon two sample rank test applied to the values log s^2 (or s^2 since only the ranking is important) obtained from the subsamples as in the Box test. This was proposed by Moses [10] and studied in detail by Shorack [17]. These authors also consider applying rank tests to other measures of dispersion (e.g., the range). This procedure yields an exact significance level, but it still suffers from loss of information in the sample subdivision. A confidence interval can be constructed graphically (see Shorack [17]), and an analogous signed rank procedure would yield tests and intervals in the one sample problem.

The *Box-Andersen test* [2] adjusts the degrees of freedom for the classical F or beta test so that the first two moments of the beta distribution agree with the first two moments of the beta statistic under the permutation distribution. You obtain the same adjustment in degrees of freedom if you equate the asymptotic variance of the F statistic under normal theory to the asymptotic variance of F under sampling from a distribution with general kurtosis. Specifically, the test or confidence interval is obtained by comparing the F ratio s_y^2/s_x^2 with the upper and lower critical points from an F distribution with $d(M-1)$, $d(N-1)$ degrees of freedom. The adjustment d is

$$d = [1 + \tfrac{1}{2}(b_2 - 3)]^{-1} = [1 + \tfrac{1}{2}c_2]^{-1}, \tag{23}$$

where

$$\begin{aligned} b_2 &= \hat{\mu}_4\hat{\sigma}^{-4} = (N+M)^{-1}[\textstyle\sum_{i=1}^{N}(X_i - \bar{X})^4 + \sum_{i=1}^{M}(Y_i - \bar{Y})^4] \\ &\quad \cdot ((N+M)^{-1}[\textstyle\sum_{i=1}^{N}(X_i - \bar{X})^2 + \sum_{i=1}^{M}(Y_i - \bar{Y})^2])^{-2}, \\ c_2 &= b_2 - 3. \end{aligned} \tag{24}$$

The quantities b_2 and c_2 are the sample estimates of the two commonest measures of population kurtosis, $\gamma_2 = \mu_4/\sigma^4$ and $\beta_2 = (\mu_4/\sigma^4) - 3$, respectively, which involve the fourth central moment $\mu_4 = E(X - E(X))^4$. The significance level will not be exact because of the approximations involved, but it should be reasonably accurate. Adaptation of the Box-Andersen technique to confidence interval construction is virtually impossible because of the amount of tedious calculation involved. Application of this principle to the χ^2 distribution in the one sample problem is immediate.

Levene [8] proposed four variations of the same technique. The *Levene s test* compares the variances from two populations by treating $U_i = (X_i - \bar{X})^2$, $i = 1, \cdots, N$, as N *independently*, identically distributed observations, and $V_i = (Y_i - \bar{Y})^2$, $i = 1, \cdots, M$, as M *independently*, identically distributed observations in a two sample t test for location. The $U_i(V_i)$ are not normally or independently distributed. On the other hand, the robustness of the t statistic protects us from nonnormality, and the correlation between U_i and $U_{i'}$, $i \neq i'$, is sufficiently small, $O(N^{-2})$, as to not likely cause trouble. The significance level would, of course, be approximate. Construction of a confidence interval from an s test would be tedious. The form of Levene's procedure in the one sample problem is self-evident.

Levene also proposes applying the two sample t test to the variables $|X_i - \bar{X}|$, $i = 1, \cdots, N$, and $|Y_i - \bar{Y}|$, $i = 1, \cdots, M$. This is referred to as Levene's z test. In the Monte Carlo studies run by Levene the z test was found to be preferable (in terms of power) to the s test. However, a proof that the z test is not asymptotically distribution-free is attached to this paper in the Appendix so the z test will not be considered a competitor.

Other variations of Levene are to consider $\log (X_i - \bar{X})^2$ and $|X_i - \bar{X}|^{\frac{1}{2}}$. However, these transformations do not seem particularly suitable for normal or near-normal distributions, and Levene's Monte Carlo sampling substantiates this. These tests are also not asymptotically distribution-free.

In addition to the jackknife the preceding list contains four adversaries: (1) Box-Andersen, (2) Levene s, (3) Box, (4) Moses. How can they be compared? Asymptotically, the Pitman relative efficiency is a sensible and convenient comparative index. For small samples the exact distributions are untractable (except possibly in isolated cases), but the tests can be compared through Monte Carlo sampling.

Consider the asymptotic comparisons first. The Pitman efficacies of the five tests in the two sample problem are readily calculated. The efficacies for the Box-Andersen, Levene s, and jackknife (any k) tests are identical and equal to

$$(25) \qquad 4MN/(M + N)(\mu_4\sigma^{-4} - 1).$$

All three procedures are asymptotically equivalent; each in its own way studentizes the F test by the appropriate estimate of its asymptotic variance. This would still hold true for the jackknife if it were applied directly to s^2 instead of $\log s^2$.

The efficacy for the Box test with $M = mk$, $N = nk$ is

$$(26) \qquad 4MN/(M + N)k \operatorname{Var} (\log s_k^2)$$

where s_k^2 is a sample variance based on k observations. The efficacy of the Moses test is the same as the Wilcoxon efficacy for location applied to $\log s_k^2$; namely,

$$(27) \quad 12MN((M + N)k)^{-1}\left[\int_0^{+\infty} p^2_{\log s_k^2}(u)\, du\right]^2 = 12MN((M + N)k)^{-1}\left[\int_0^{+\infty} v p^2_{s_k^2}(v)\, dv\right]^2.$$

The comparison of Moses (27) versus Box (26) is the same as comparing Wilcoxon and t for shift in location when the variables are $\log s_k^2$. Numerical values are presented for the asymptotic relative efficiency (ratio of the efficacies) of Moses to Box in Shorack [17]. The ARE's under the assumption of X, Y normally distributed are close to 1 for various values of k with some above and some below 1.

Shorack also gives numerical values for the ARE of Moses versus Box-Andersen (Levene s and jackknife) when X, Y are (i) normally distributed, (ii) uniformly distributed, and (iii) double exponentially distributed. For all three distributions and all selected values of k the Box-Andersen (Levene s and jackknife) procedure is more efficient. However, Shorack conjectures that for heavier tailed distributions such as a contaminated normal the reverse should be true.

The variance of s_k^2 is $\sigma^4(2(k-1)^{-1} + \gamma_2 k^{-1})$. If k is sufficiently large for the approximation Var $(\log s_k^2) \cong (2(k-1)^{-1} + \gamma_2 k^{-1})$ to be accurate, then the efficacy of the Box test is approximately

$$4MN/(M+N)(2k(k-1)^{-1} + \gamma_2). \tag{28}$$

Expression (28) is always smaller than (25) because the factor $k/(k-1)$ is greater than 1, which seems to indicate that Box-Andersen (Levene s and jackknife) are always more asymptotically efficient than Box. However, $k/(k-1)$ for $k \geqq 5$ is not appreciably larger than 1. It is conceivable that the inaccuracy of the approximation could mask some cases in which (26) is larger than (25) although this goes against one's intuition.

What about small samples? Two Monte Carlo studies were run on the Burroughs B5500 at Stanford University. The first study compared samples of size 25 ($M = 25$, $N = 25$), and the results are exhibited in Table I. One thousand pairs of samples of 25 pseudo-random numbers were generated in the computer. The pseudo-random numbers represent samples from a uniform distribution. In addition, the numbers were transformed to obtain samples from a normal distribution, a double exponential distribution, a skew double exponential distribution with density

$$\begin{aligned} p_X(x) &= \tfrac{2}{3}e^{x}, \qquad x < 0, \\ &= \tfrac{1}{6}e^{-x/2}, \qquad x > 0, \end{aligned} \tag{29}$$

and a distribution with heavy tails which vanished like a sixth power (just sufficient to have a fourth moment). The density of this sixth power distribution is

$$p_X(x) = \tfrac{5}{2}(1 + |x|)^{-6}. \tag{30}$$

After transformation the Y sample was scaled by the factor Δ so that the ratio of the Y to X variances was Δ^2 for each distribution. Different values of Δ^2 were selected and applied to the same samples.

Seven tests were applied to each of the 1000 pairs of samples. The standard normal theory F test was included as one of the seven to illustrate how terribly non-robust it is, and consequentially to drive one more nail into its coffin. The Box-Andersen and Levene s tests were two of the seven. The jackknife test was applied with subsample size $k = 1$ ($n = m = 25$) and with subsample size $k = 5$ ($n = m = 5$). Box and Moses were both used with subsample size $k = 5$ ($n = m = 5$).

The entries in Table I are the proportions of samples in 1000 trials that the

TABLE I

Monte Carlo Power Functions for Tests on Variances

$M = N = 25$

	$\alpha = .05$					$\alpha = .01$				
$\Delta^2 = \sigma_y^2/\sigma_x^2$	1	2	4	6	10	1	2	4	6	10
Uniform Distribution										
F	.007	.501	.997	1.00	1.00	.001	.127	.926	1.00	1.00
Box-Andersen	.053	.794	.999	1.00	1.00	.012	.516	.982	.999	1.00
Jackknife $k = 1$	.029	.786	1.00	1.00	1.00	.005	.498	.992	1.00	1.00
Jackknife $k = 5$	.036	.710	.997	1.00	1.00	.009	.367	.935	.993	1.00
Levene s	.041	.781	.999	1.00	1.00	.010	.485	.973	.998	1.00
Box $k = 5$	.054	.498	.910	.978	.995	.013	.231	.681	.865	.969
Moses $k = 5$	.040	.402	.801	.920	.970	.004	.137	.514	.707	.874
Normal Distribution										
F	.060	.496	.937	.994	1.00	.017	.252	.818	.966	.999
Box-Andersen	.047	.477	.920	.982	1.00	.016	.195	.687	.895	.970
Jackknife $k = 1$	.050	.465	.907	.963	.989	.014	.217	.767	.929	.968
Jackknife $k = 5$	.050	.417	.868	.950	.986	.015	.178	.586	.823	.935
Levene s	.041	.453	.901	.974	.995	.011	.149	.639	.812	.916
Box $k = 5$	.049	.369	.798	.935	.985	.009	.135	.479	.706	.891
Moses $k = 5$	.035	.264	.659	.845	.946	.007	.078	.319	.503	.719
Double Exponential Distribution										
F	.127	.494	.864	.954	.995	.063	.337	.743	.899	.981
Box-Andersen	.053	.315	.715	.878	.971	.012	.090	.372	.596	.800
Jackknife $k = 1$	.069	.313	.690	.847	.950	.019	.133	.470	.656	.849
Jackknife $k = 5$	.058	.290	.654	.803	.918	.016	.103	.373	.537	.739
Levene s	.043	.282	.677	.824	.932	.008	.055	.261	.406	.565
Box $k = 5$	.051	.266	.616	.801	.942	.010	.080	.305	.496	.712
Moses $k = 5$	.031	.180	.474	.657	.854	.007	.040	.186	.299	.491
Skew Double Exponential Distribution										
F	.177	.499	.809	.929	.983	.098	.346	.682	.854	.964
Box-Andersen	.065	.249	.579	.733	.899	.013	.083	.271	.441	.626
Jackknife $k = 1$	.068	.242	.579	.740	.873	.024	.116	.330	.520	.705
Jackknife $k = 5$	.072	.242	.517	.691	.841	.020	.098	.268	.405	.580
Levene s	.054	.214	.506	.654	.805	.008	.046	.158	.275	.388
Box $k = 5$	.059	.210	.498	.661	.845	.011	.072	.211	.351	.549
Moses $k = 5$	.030	.135	.368	.540	.720	.004	.032	.123	.209	.332
Sixth Power Distribution										
F	.218	.489	.776	.881	.952	.126	.380	.685	.813	.924
Box-Andersen	.054	.223	.531	.683	.846	.010	.054	.222	.374	.553
Jackknife $k = 1$	.082	.269	.545	.673	.800	.023	.116	.327	.486	.637
Jackknife $k = 5$	.076	.243	.506	.649	.770	.018	.088	.252	.383	.536
Levene s	.038	.185	.436	.576	.708	.005	.024	.104	.195	.299
Box $k = 5$	.048	.198	.497	.665	.845	.011	.058	.217	.352	.545
Moses $k = 5$	.030	.131	.369	.521	.719	.006	.028	.118	.213	.332

tests rejected the null hypothesis $\sigma_y^2 = \sigma_x^2$ for the various distributions and combinations of Δ^2 and α (the significance level of the test). For the $\Delta^2 = 1$ columns (i.e., when the null hypothesis is true) the proportions should be close to $\alpha = .05$ or $\alpha = .01$. For $\Delta^2 > 1$ the proportions are Monte Carlo estimates of the power of the tests at the particular alternatives Δ^2 for the various distributions.

Inspection of the table leads one to the following conclusions:

(i) The F test is extremely non-robust. Its actual significance level under the null hypothesis is much smaller than indicated for short-tailed distributions (uniform), and it gives far too many significant results for long-tailed distributions (double exponential, sixth power).

(ii) The Box-Andersen test and the jackknife test with $k = 1$ have about the same power, and in general they are the most powerful tests in the group. Box-Andersen has slightly better power for $\alpha = .05$ level tests while the jackknife is slightly better for $\alpha = .01$ level tests. The true significance levels of these two tests when $\Delta^2 = 1$ are slightly more sensitive to the form of the distribution than the Levene s, Box, and Moses tests.

(iii) The jackknife with $k = 5$ is just not as powerful as the jackknife with $k = 1$, but its actual significance level tends to be close to the nominally indicated level.

(iv) The Levene s test is robust, but it is not as powerful as the Box-Andersen, jackknife, and Box tests for long-tailed distributions, particularly at significance level $\alpha = .01$.

(v) The Box test ($k = 5$) is robust, but its power is not as good as Box-Andersen and the jackknife with $k = 1$. In fact, its performance is very similar to the jackknife with $k = 5$. (If the choice of tests was between Box with $k = 5$ and the jackknife with $k = 5$, one would certainly choose Box because it is easier to compute.)

(vi) The Moses test ($k = 5$) lags behind the Box test ($k = 5$) in power. It seems to be the least powerful of all the tests.

A curiosity of the Monte Carlo sampling experiment is the observed significance levels for the Moses test. Of all the tests the Moses test is the only one which should have true significance levels exactly equal to $\alpha = .05$ or $\alpha = .01$ for all distributions. Yet the observed levels range from .03 to .04 for $\alpha = .05$ and .004 to .007 for $\alpha = .01$. The discrepancies are not quite as large as indicated because the true α's corresponding to the critical points used in the test were .0476 and .0079 due to the discreteness of the distribution. The binomial sampling standard errors at these points are $[(.0476)(.9524)/1000]^{\frac{1}{2}} = .0067$ and $[(.0079)(.9921)/1000]^{\frac{1}{2}} = .0028$. In view of the size of these standard errors one would have expected the observed α's to be somewhat closer to their true values.

The F test should also have had exact levels under the null hypothesis for the normal distribution. In this case the levels are high, .06 and .017. One can only conclude either that these discrepancies are purely an accidental phenomenon of the Monte Carlo sampling or that the pseudo-random numbers are not as random as they are purported to be.

A second Monte Carlo study identical to the first except that the sample sizes were reduced to $M = N = 10$ was also run on the computer. One thousand pairs of samples based on pseudo-random numbers were again selected. The sequence of pseudo-random numbers did not overlap with the sequence used in the first study. The same distributions were sampled, and the same values for α and Δ^2 were selected.

The Box and Moses tests were not applied to the samples because of the small sample size. The jackknife with subsample size $k > 1$ was not tried for the same

TABLE II

Monte Carlo Power Functions for Tests on Variances

$M = N = 10$

	$\alpha = .05$					$\alpha = .01$				
$\Delta^2 = \sigma_y^2/\sigma_x^2$	1	2	4	6	10	1	2	4	6	10
	Uniform Distribution									
F	.017	.181	.686	.912	.986	.002	.036	.268	.607	.907
Box-Andersen	.077	.364	.784	.904	.973	.016	.135	.431	.626	.819
Jackknife $k = 1$	.036	.281	.785	.914	.983	.008	.106	.458	.725	.903
Levene s	.052	.361	.765	.886	.952	.011	.118	.395	.535	.661
	Normal Distribution									
F	.063	.259	.633	.827	.947	.018	.092	.347	.562	.818
Box-Andersen	.075	.256	.586	.771	.906	.013	.077	.253	.383	.589
Jackknife $k = 1$	.062	.219	.550	.745	.885	.018	.090	.274	.462	.690
Levene s	.050	.232	.519	.693	.828	.008	.048	.157	.242	.336
	Double Exponential Distribution									
F	.125	.324	.598	.746	.891	.053	.173	.397	.549	.742
Box-Andersen	.077	.199	.431	.595	.778	.018	.056	.154	.249	.391
Jackknife $k = 1$	.074	.205	.415	.562	.716	.026	.084	.205	.302	.470
Levene s	.045	.168	.335	.448	.580	.005	.030	.068	.113	.169
	Skew Double Exponential Distribution									
F	.173	.343	.595	.727	.850	.094	.211	.409	.551	.721
Box-Andersen	.093	.203	.370	.503	.655	.021	.062	.144	.211	.331
Jackknife $k = 1$	.085	.194	.353	.472	.610	.036	.082	.176	.256	.376
Levene s	.067	.144	.268	.350	.465	.007	.025	.043	.062	.091
	Sixth Power Distribution									
F	.195	.373	.580	.696	.825	.104	.238	.421	.536	.696
Box-Andersen	.076	.191	.368	.480	.640	.019	.045	.122	.186	.295
Jackknife $k = 1$	.099	.202	.365	.478	.614	.029	.082	.168	.249	.368
Levene s	.044	.135	.254	.339	.432	.003	.015	.048	.069	.109

reason. The tests that were applied are the F test, Box-Andersen, the jackknife with $k = 1$, and the Levene s test.

The results of the second study are exhibited in Table II. The conclusions one can draw from the table are very similar to those reached in the first study.

(i) The F test is extremely non-robust.

(ii) Box-Andersen and the jackknife ($k = 1$) are about equally powerful with Box-Andersen slightly better at $\alpha = .05$ and the jackknife slightly better at $\alpha = .01$.

(iii) The observed significance levels under the null hypothesis for the jackknife and Box-Andersen are more sensitive to the form of the distribution than in the case of the larger sample sizes.

(iv)' The Levene s test is quite robust, but it lags far behind the jackknife and Box-Andersen in power.

5. Discussion. The problem of testing equality of variances has now been added to the list of problems in which the jackknife proves to be a robust and powerful statistical tool. Previously recorded instances of other such problems were cited in the introduction.

In general, the jackknife seems to work well if it works at all. The basic ingredient which appears necessary for the jackknife's success is for the unmodified estimator $\hat{\theta}(X_1, \ldots, X_N)$ to be asymptotically locally linear in each observation or some convenient function of each observation. By "asymptotically locally linear" it is meant that $\hat{\theta}$ can be expanded in a power series for each observation X_i where (a) the second and higher order terms are negligible and (b) the first order term is linear in the observation X_i or some nice, simple function of X_i (viz., X_i^2 in this paper). Tukey more or less conjectured this at the outset.

When $\hat{\theta}$ has this property, the jackknife estimator $\tilde{\theta}.$ can be conveniently expanded in a power series in order to apply large sample theory in establishing asymptotic normality with the correct mean and variance. The linear quality of $\hat{\theta}$ will impart asymptotic normality to it, and this normality is preserved under jackknifing. As yet no example has been given in which $\hat{\theta}$ is not asymptotically normally distributed but $\tilde{\theta}.$ is. It would seem that normality can only be preserved, and not created, by jackknifing.

As a technique for testing variances, the jackknife fares well in comparison with the other available procedures. For small samples (N, $M \leqq 15$, say) the jackknife ($k = 1$) and Box-Andersen are the best choices, and they are about equally powerful. Box and Moses do not seem feasible in this sample range because either the size of the subsample or the number of subsamples would have to be very small. Levene s is available and is easy to use. However, it lacks the power of Box-Andersen and the jackknife.

Even for larger samples there does not seem to be any technique more powerful than Box-Andersen or the jackknife ($k = 1$) although Levene s and the jackknife ($k > 1$) are asymptotically equivalent ~~to~~ them. However, as the sample size increases, the amount of computation involved in Box-Andersen or the jackknife ($k = 1$) becomes enormous. If an electronic digital computer is available, there is no problem, but it would be extremely time-consuming to use them with hand computation. Box and Moses require far less computation and are not agging far behind in power. When using these tests there is no escape from the worry that the inference may depend upon the particular grouping into subsamples, but for large samples (N, $M \geqq 25$) this should be a rather small worry. The jackknife with $k > 1$ could be employed, but Box or Moses should be almost as powerful, easier to use, and more easily explained to the client. Levene s is also available, and may be the most convenient to use.

The jackknife, Box, and Moses easily provide confidence intervals whereas Box-Andersen and Levene s do not.

6. Appendix: Levene z test. In the one sample problem the Levene z test treats the variables $Z_i = |X_i - \bar{X}|$, $i = 1, \cdots, N$, as *independently*, identically distributed random variables. A proof will be given that the variance of $\bar{Z} = \sum_{i=1}^{N} Z_i/N$ is not estimated correctly by $\sum_{i=1}^{N} (Z_i - \bar{Z})^2/N(N-1)$. The argument extends immediately to the two sample problem.

Assume the observations come from a continuous distribution with $E(X) = 0$.

The proof is based on the following representation of $\sum_{i=1}^{N} Z_i$ which the reader can verify with a little thought. For $\bar{X} > 0$

$$(31)\quad \sum_{i=1}^{N} |X_i - \bar{X}| = \sum_{0<x_i} X_i - \sum_{x_i<0} X_i - 2\sum_{0<x_i<\bar{x}} X_i + \bar{X}(N_{x_i<\bar{x}} - N_{x_i>\bar{x}}),$$

where $N_{x_i<\bar{x}}(N_{x_i>\bar{x}})$ is the number of X_i less (greater) than $\bar{X}$. The analogous representation for $\bar{X} < 0$ is

$$(32)\quad \sum_{i=1}^{N} |X_i - \bar{X}| = \sum_{0<x_i} X_i - \sum_{x_i<0} X_i + 2\sum_{\bar{x}<x_i<0} X_i + \bar{X}(N_{x_i<\bar{x}} - N_{x_i>\bar{x}}).$$

These two representations can be combined into one expression for $\bar{Z}$:

$$(33)\quad \bar{Z} = N^{-1}\sum_{i=1}^{N} |X_i| \pm 2N^{-1}\sum_{\bar{x}<x_i<0,\ 0<x_i<\bar{x}} X_i + \bar{X}(N_{x_i<\bar{x}}/N - N_{x_i>\bar{x}}/N).$$

The pair $(N^{-1}\sum_{i=1}^{N} |X_i|, N^{-1}\sum_{i=1}^{N} X_i)$ has a limiting bivariate normal distribution with means $E|X|$ and 0, respectively, and with variances-covariances given by

$$(34)\quad \begin{aligned} \operatorname{Var}(N^{-1}\textstyle\sum_{i=1}^{N} |X_i|) &= N^{-1}[E(X^2) - (E|X|)^2], \\ \operatorname{Var}(N^{-1}\textstyle\sum_{i=1}^{N} X_i) &= N^{-1}E(X^2), \\ \operatorname{Cov}(N^{-1}\textstyle\sum_{i=1}^{N} |X_i|, N^{-1}\sum_{i=1}^{N} X_i) &= N^{-1}E(X^2 \operatorname{sgn} X), \end{aligned}$$

where sgn X denotes the sign of X. Since $\bar{X} \to_p 0$, the normalized middle term

$$(35)\quad \pm 2N^{-\frac{1}{2}}\sum_{\bar{x}<x_i<0,\ 0<x_i<\bar{x}} X_i \to_p 0.$$

The fractions $N_{x_i<\bar{x}}/N$ and $N_{x_i>\bar{x}}/N$ converge in probability to $P\{X < 0\}$ and $P\{X > 0\}$, respectively. Combining these convergences via Slutsky's theorem proves that $\bar{Z}$ has a limiting normal distribution with mean $E|X|$ and variance

$$(36)\quad N^{-1}[E(X^2) - (E|X|)^2 + E(X^2)(P\{X < 0\} - P\{X > 0\})^2 \\ + 2E(X^2 \operatorname{sgn} X)(P\{X < 0\} - P\{X > 0\})].$$

If $\sum_{i=1}^{N} (Z_i - \bar{Z})^2/(N - 1)$ is written as

$$(37)\quad (N-1)^{-1} \sum_{i=1}^{N} (X_i - \bar{X})^2 - N(N-1)^{-1}(N^{-1} \sum_{i=1}^{N} |X_i - \bar{X}|)^2,$$

then it is easily seen that this converges in probability (and a.s.) to

$$(38)\quad E(X^2) - (E|X|)^2.$$

The convergence of the first factor is immediate, and the second follows from the representation (33).

Comparison of expression (38) with expression (36) without the N^{-1} factor reveals that the two will not be the same in general. Thus, the z test is not asymptotically distribution-free.

The z test will be asymptotically distribution-free if the distribution is symmetric or at least has its median equal to its mean. The factor $P\{X < 0\} - P\{X > 0\}$ will be zero in this case, and (36) and (38) will agree.

If the reader were to carry through an argument essentially identical to the above, he would be able to verify that the z test is asymptotically distribution-free if the sample is centered at its median instead of its mean. That is, the random variables $Z_i = |X_i - m|$, $i = 1, \cdots, N$, can be treated as independently, identically distributed from an asymptotic point of view. The variable m is the sample median.

The difference between this and the original z test is that the term $N_{x_i<\bar{x}} - N_{x_i>\bar{x}}$, which causes all the trouble, is replaced by $N_{x_i<m} - N_{x_i>m}$, which is zero.

This modified z test was not included in the paper partly because this property of the z test was not discovered till all of the other work in the paper had been completed, partly because dispersion about a median is not a variance, and partly because means are easier to work with on computers.

If the observations come from a non-continuous distribution, then expressions (31), (32) have to be modified to account for the possibility $X_i = \bar{X}$, and the asymptotic distribution theory becomes more complicated.

7. Acknowledgment. The author would like to thank Judith Grindle for programming the Monte Carlo study in this paper. The author would also like to thank Lincoln Moses for his valuable discussions on problems with variances.

REFERENCES

[1] Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40** 318–335.

[2] Box, G. E. P. and Andersen, S. L. (1955). Permutation theory in the derivation of

robust criteria and the study of departures from assumption. *J. Roy. Statist. Soc. Ser. B.* **17** 1-26.

[3] Brillinger, D. R. (1964). The asymptotic behaviour of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates. *Rev. Inst. Internat. Statist.* **32** 202-206.

[4] Brillinger, D. R. (1966). Discussion on Mr. Sprent's paper. *J. Roy. Statist. Soc. Ser. B* **28** 294.

[4a] Brillinger, D. R. (1966). The application of the jackknife to the analysis of sample surveys. *Commentary* **8** 74-80.

[5] Deming, W. E. (1963). On the correction of mathematical bias by use of replicated designs. *Metrika* **6** 37-42.

[6] Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika* **46** 477-480.

[7] Klotz, J. (1962). Nonparametric tests for scale. *Ann. Math. Statist.* **33** 498-512.

[8] Levene, H. (1960). Robust tests for equality of variances. *Contributions to Probability and Statistics* (I. Olkin, et al, editors) 278-292.

[9] Miller, R. G., Jr. (1964). A trustworthy jackknife. *Ann. Math. Statist.* **35** 1594-1605.

[10] Moses, L. E. (1963). Rank tests of dispersion. *Ann. Math. Statist.* **34** 973-983.

[11] Mosteller, F., and Tukey, J. W. (1965). Data analysis, including statistics. An article in preparation for the revised *Handbook of Social Psychology.*

[12] Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *J. Roy. Statist. Soc. Ser. B* **11** 68-84.

[13] Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika* **43** 353-360.

[14] Rao, J. N. K. (1965). A note on estimation of ratios by Quenouille's method. *Biometrika* **52** 647-649.

[14a] Rao, J. N. K. (1967). The precision of Mickey's unbiased ratio estimator. *Biometrika* **54** 321-324.

[15] Rao, J. N. K., and Webster, J. T. (1966). On two methods of bias reduction in the estimation of ratios. *Biometrika* **53** 571-577.

[16] Robson, D. S., and Whitlock, J. H. (1964). Estimation of a truncation point. *Biometrika* **51** 33-39.

[17] Shorack, G. R. (1965). Nonparametric tests and estimation of scale in the two sample problem. Technical Report No. 10 (USPHS-5T1GM 25-07), Stanford University.

[18] Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Ann. Math. Statist.* **29** 614.

[19] Tukey, J. W. (1962). Data analysis and behavioral science. Unpublished manuscript.