Approximate Tests and Confidence Intervals Using the Jackknife

by

James N. Arvesen*

and

David S. Salsburg**

Department of Statistics

Division of Mathematical Sciences

Mimeograph Series # 267

October, 1971

* Department of Statistics, Purdue University. Research supported by the Purdue Research Foundation.

** Pfizer Pharmaceuticals, Groton, Conn.

To appear in Perspectives in Biometrics, edited by R. Elashoff, published by Academic Press, 1972.

What follows is a discussion of relevant papers which have been significant in bringing the jackknife technique to its present status. It enjoys the position of being an almost universally applicable tool in aiding a researcher to obtain approximate tests or confidence intervals for parameters of interest. However to understand its potential pitfalls, it is important to understand the development of some of the theory behind it. This development is far from complete as of this date.

1. Introduction. Let $X_1,\ldots,X_N$ be independent identically distributed (iid) random variables (vectors), and let the real-valued parameter θ be associated with their distribution. Group the N observations into n groups of k observations each, $N = nk$. Let $\hat{\theta}_n^0$ be some estimate of θ based on all n groups of observations, and let $\hat{\theta}_{n-1}^i$ be the estimate of θ based on $\hat{\theta}_n^0$ after deletion of the ith group of observations (delete $X_{(i-1)k+1},\ldots,X_{ik}$). Then let

$$\hat{\theta}_i = n\,\hat{\theta}_n^0 - (n-1)\,\hat{\theta}_{n-1}^i, \qquad i = 1,\ldots,n,$$

$$\hat{\theta} = n^{-1}\sum_{i=1}^{n}\hat{\theta}_i, \tag{1.1}$$

$$s^2_{\hat{\theta}} = (n-1)^{-1}\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta})^2.$$

The jackknife estimate of θ based on $\hat{\theta}_n^0$ is $\hat{\theta}$, and was first introduced by Quenouille [1949] in the special case $n = 2$, and subsequently discussed in Quenouille [1956]. The jackknife estimate of θ possesses the interesting property that if $\hat{\theta}_n^0$ is biased of order $1/N$, then $\hat{\theta}$ reduces the bias to order $1/N^2$. That is, if

$$E\,(\hat{\theta}_n^{O}) = \theta + a/kn + b/(kn)^2 + \ldots\,,$$

(1.2)
$$E\,(\hat{\theta}) = n(\theta + a/kn + b/(kn)^2 + \ldots)$$
$$-(n-1)\,(\theta + a/k(n-1) + b/(k(n-1))^2 + \ldots)$$
$$= \theta - b/k^2\, n(n-1) + \ldots\,.$$

It is possible to employ the jackknife to reduce bias of order $1/N^2$, $1/N^3$, etc. A set of excellent references for the jackknife to reduce bias is Mantel [1967], Schucany [1971], Schucany, Gray, and Owen [1971], Adams, Gray, and Watkins [1971], and the monograph by Gray and Schucany [1971].

It is also tempting to consider $\hat{\theta}_1,\ldots,\hat{\theta}_n$ as approximately <u>independent</u> identically distributed random variables (they are exchangeable random variables for each n, however the marginals typically depend on n). Then one might consider

(1.3)
$$t = \sqrt{n}\,(\hat{\theta}-\theta)/s_{\hat{\theta}}$$

to be approximately t distributed with $n-1$ degrees of freedom. This observation is due to Tukey [1958]. It is this last suggestion that we propose to analyze in detail.

2. Approximate tests and confidence intervals. Almost simultaneously, two rigorous justifications of Tukey's conjecture appeared in the literature. These were the situation where the original estimate $\hat{\theta}_n^{O}$ is a maximum likelihood

estimate (MLE) (Brillinger [1964]), and $\hat{\theta}_n^0$ is a transformation of a sample mean (Miller [1964]). Brillinger showed that if $X_1,\ldots,X_N$ are iid from a distribution for which the MLE satisfies the standard regularity conditions for asymptotic normality, then (1.3) has asymptotically (as $N \to \infty$) a t distribution with n-1 degrees of freedom. What this probably means in practice is that unless N is large, only a small number of degrees of freedom can be obtained for the asymptotic distribution of (1.3).

Fryer [1970] claims to have overcome the difficulty that n remain finite as $N \to \infty$, and also claims to have extended Brillinger's results to the multiparameter case. Unfortunately, these results have not yet appeared in published form. These results (as well as those to be subsequently discussed) open up a question regarding the relative sizes of n and k if N is large. At present, the only solution seems to be that n should be large enough to yield adequate degrees of freedom for the t-statistic, but not too large to cause excessive computing time.

Brillinger's result can be extended to the case of non-identically distributed random variables. As long as the MLE based on $X_1,\ldots,X_N$ has an asymptotic normal distribution (when normalized), the jackknife will produce the desired result in (1.3). For $N = nk$, $k \to \infty$ and n finite, Brillinger's proof is adequate. It is unclear whether Fryer's claims would cover this case.

As an interesting example, consider the bio-assay problem considered in Berkson [1955]. That is the model is that one observes $Y_1,\ldots,Y_N$ independent, Y_i having a Bernoulli distribution with parameter depending on a continuous predictor variate X_i (usually a dose level),

$$P(X_i) = \left(1 + e^{-(\alpha + \beta X_i)}\right)^{-1}, \quad i = 1,\ldots,N .$$

Let us assume one is interested in a confidence interval for β based on its MLE. Then if $\hat{\alpha},\hat{\beta}$ denotes the MLE's of α,β, they may obtained by maximizing

$$L(X_1, Y_1; X_2, Y_2;\ldots; X_N, Y_N)$$

$$= \Pi_{i=1}^{N} \left(\frac{1}{1 + e^{-(\alpha + \beta X_i)}} \right)^{Y_i} \left(\frac{e^{-(\alpha + \beta X_i)}}{1 + e^{-(\alpha + \beta X_i)}} \right)^{1-Y_i} .$$

This is equivalent to finding α,β to minimize

$$L^*(X_1, Y_1; X_2, Y_2;\ldots; X_N, Y_N)$$

$$= \sum_{i=1}^{N} \ell n\{(1 + e^{-(\alpha + \beta X_i)})\} + (1 - Y_i)(\alpha + \beta X_i) .$$

This is a relatively easy task given any good non-linear minimization computer program. Most computers seem to have these available as standard packages.

Thus, let $N = nk$, $\hat{\theta}_n^0 = \hat{\beta}$, $\hat{\theta}_{n-1}^i$ denote the MLE of β after deleting $X_{(i-1)k+1},\ldots,X_{ik}$, $i = 1,\ldots,n$, and repeat (1.1). Then as long as $\hat{\beta}$ is asymptotically normal, and n remains finite as $k \to \infty$, one obtains (1.3). To see this in a special case, a Monte Carlo simulation was performed on the CDC 6500 computer at Purdue University.

The model considered was the following:

$$P(X_{ij}) = \left(1 + e^{-(\alpha+\beta X_{ij})}\right)^{-1}, \quad i = 1,\ldots,5, \quad j = 1,\ldots,20 ,$$

and one observes Y_{ij}, independent Bernoulli random variables given by this

model. Five levels of X were taken, namely, $X_{ij} = -4$, $X_{2j} = -2$, $X_{3j} = 0$, $X_{4j} = 2$, $X_{5j} = 4$, $j = 1,\ldots,20$. Finally, due to time constraints, only the parameter values $(\alpha = 0, \beta = 0)$, $(\alpha = 0, \beta = .10)$ and $(\alpha = 0, \beta = .25)$ were simulated. Each of these three sets was simulated 1000 times. Thus, for each of the three sets, 1000 x 100 = 100,000 Bernoulli trials were simulated. Finally, if the data are presented in the tableau,

$$\begin{matrix} Y_{11} & Y_{12} & \cdots & Y_{1,20} \\ Y_{21} & Y_{22} & & Y_{2,20} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ Y_{51} & Y_{52} & & Y_{5,20} \end{matrix} ,$$

a 5 x 20 matrix results. The data were then grouped into

$$Z_1 \quad Z_2 \quad Z_3 \quad Z_4 \quad Z_5 ,$$

where each Z_i is a 5 x 4 matrix. Thus in (1.1), $n = 5$ $(N = 100, k = 20)$, and first Z_1 is deleted, then $Z_2,\ldots,$ and finally Z_5.

Thus to test

H_0: $\beta = 0$

H_A: $\beta \neq 0$

using (1.1) and (1.3), one rejects H_0 at the $\alpha = .10$ level if $| \sqrt{5}\ \hat{\theta}/s_{\hat{\theta}} | > 2.132$. Note that α was also considered unknown, and no transformation of $\hat{\beta}$ was used. The results were (recall $\alpha = 0$ for all three runs)

$\beta =$	0	.10	.25
Empirical Power Function	.095	.286	.823

Thus, of the 1000 simulations when $\beta = 0$, the actual number of rejections was 95, and similarly for the other two. This appears to be quite good.

In addition, for the three values of β, 90% confidence intervals were obtained. What follows is the empirical percent of intervals that covered the true β.

$\beta =$	0	.10	.25
Empirical Probability of Coverage	.905	.907	.915

The total amount of computer time used was just under 20 minutes, or 1200 seconds. Thus to obtain the 3000 tests (and confidence intervals), an average of .4 seconds was used on each one. To obtain only one test and confidence interval took just under 2 seconds. This would of course be insignificant.

Miller considered the case where $\hat{\theta}_n^0$ was a "nice" function of the sample mean $\bar{X}$(nice here means one with a bounded second derivative in a neighborhood of θ). Such problems might arise in situations such as the aresine transformation of a binomial proportion, or the square root of a Poisson mean. Miller also gives examples of the result in (1.3) not holding. In fact, $\hat{\theta}_n^0, \hat{\theta}_1, \ldots, \hat{\theta}_n$ can have a joint normal distribution, and the asymptotic t distribution for (1.3) may not necessarily hold.

In a subsequent paper, Miller [1968] extended his successful result to the potentially more useful case in which $\hat{\theta}_n^0$ is the sample variance, or the log of the sample variance. The motivation is to obtain a good competitor to the usual χ^2 test (or confidence interval) if the $X_1,\ldots,X_N$ are normal, and to obtain a robust procedure if the data are not normal. One only needs the mathematical nicety of finite fourth moments for the data, and the result holds.

The idea is as follows, let $\mu = E(X_i)$, $\sigma^2 = \text{Var}\,(X_i)$, $i = 1,\ldots,N$,

$$\bar{X} = N^{-1} \sum_{j=1}^{N} X_i, \quad s^2 = (N-1)^{-1} \sum_{j=1}^{N} (X_j - \bar{X})^2 ,$$

(2.1)

$$\hat{\theta}_n^0 = \log s^2 .$$

Then follow the steps outlined in (1.1). Presumably these calculations will be done on a computer, especially if n is moderately large. Here (and elsewhere to follow) this is a relatively easy programming task for anyone with a minimal capability of writing a computer program. Now a two-sided $(1-\alpha) \times 100\%$ asymptotic confidence interval for $\theta = \log \sigma^2$ is obtained from

$$(2.2) \qquad L = \hat{\theta} - t_{\alpha/2;n-1}\, s_{\hat{\theta}}/\sqrt{n} \leq \theta \leq \hat{\theta} + t_{\alpha/2;n-1}\, s_{\hat{\theta}}/\sqrt{n} = U ,$$

where $t_{\alpha/2;n-1}$ is the upper $\alpha/2$ point of a t distribution with $n-1$ degrees of freedom. Hence a $(1-\alpha) \times 100\%$ level asymptotic confidence interval for σ^2 is

$$(2.3) \qquad e^L \leq \sigma^2 \leq e^U .$$

If one is interested in testing

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_A: \sigma^2 \neq \sigma_0^2 ,$$

H_0 is rejected if the interval in (2.3) does not contain σ_0^2. An analogous procedure holds for one-sided confidence intervals or tests.

The jackknife technique was also applied by Miller to obtain tests and confidence intervals in the two sample case. Miller felt that equal sample sizes were necessary, but this restriction was later shown to be unnecessary (see Layard [1971] and Hall [1971] mentioned below). For moderate sample sizes, Monte Carlo results for several distributions are given which indicate that the jackknife technique is a valid competitor to the F-test if the data are normal. Moreover, unlike the F-test, the jackknife gives robust significance levels if the data are not normal (as they undoubtedly will be in practice).

In Miller [1964], one of the situations in which (1.3) did not have an asymptotic t distribution was the following. Let $X_1,\ldots,X_N$ be uniform on $[0,\theta]$, and $\hat{\theta}_n^0 = \max(X_1,\ldots,X_N)$. A slightly different version of the jackknife was shown by Robson and Whitlock [1964] to behave satisfactorily.

An ingenious extension of Robson and Whitlock's use of the jackknife was done by Schucany [1971], and Schucany, Gray, and Owen [1971]. They considered two estimators of the parameter θ, each of which is biased such that

$$E[t_K(X_1,\ldots,X_n)] - \theta = b_K(n,\theta) \neq 0,\ k = 1,2 .$$

Letting $R = \frac{b_1(n,\theta)}{b_2(n,\theta)}$, they define an unbiased estimate of θ by $\hat{\theta} = \frac{t_1 - Rt_2}{1-R}$. If as in (1.1),

$$t_1 = \hat{\theta}^0_n,\ t_2 = n^{-1} \sum_{i=1}^{n} \hat{\theta}^i_{n-1},\ R = (n-1)/n\ ,$$

then $\hat{\theta}$ as defined by Schucany is the usual jackknife.

However, his form of the estimator is much more general. For example, if $X_1,\ldots,X_n$ are uniform $(0,\theta)$, and if $\hat{\theta}^0_n = \max(X_1,\ldots,X_n)$ and t_1,t_2 as above, with $R = n/(n+1)$, then Schucany's estimate (which agrees with Robson and Whitlock) is $\hat{\theta} = 2X_{(n)} - X_{(n-1)}$ where $X_{(i)}$ is the ith order statistic. $\hat{\theta}$ is of course unbiased.

Schucany, Gray, and Owen also comment that the asymptotic t distribution results of Miller and Arvesen hold for their modified form of the jackknife estimator.

Dempster [1966] proposes another modification of the jackknife for problems dealing with canonical correlation coefficients. Instead of eliminating observations, Dempster advocates elimination of single degrees of freedom of the sample covariance matrix. Asymptotically, these two methods should be equivalent, although this has not been shown.

Brillinger [1966] proposes a general class of estimates to which the jackknife may be applied to obtain an estimate of the standard error of an estimate. Unfortunately n must remain finite as $N \to \infty$, and use of the t distribution is not obtained.

Mosteller and Tukey [1968] give several examples where they would propose use of the jackknife method to obtain tests or confidence intervals. An interesting use of the method is proposed in discriminant analysis, discriminating

between Hamilton and Madison as authors of the Federalist papers.

In Arvesen [1969] is found a class of statistics to which the jackknife may be profitably applied to obtain asymptotic tests or confidence intervals. If θ is estimated by $\hat{\theta}_n^0$, and $\hat{\theta}_n^0$ is based on a U-statistic, or a function of several U-statistics, then the asymptotic t-distribution of (1.3) is valid. For an earlier paper exploiting another relation between the jackknife and U-statistics, see Mantel [1967]. The best reference to U-statistics remains Hoeffding's original paper (Hoeffding [1948]).

One of the most striking discoveries that an applied statistician soon makes is the nice statistical behavior of the mean. One reason why the central limit conjecture was believed and used as a theorem long before the Lindeberg-Lévy proofs is that means from reasonably homogeneous data do behave like normal variates: they tend to have symmetric distributions, and confidence bounds computed from t and normal tables are both tight and believable.

There are many problems, however, that involve parameters other than the mean. For instance, (1) in drug screening it is often desirable to derive conjectures about the tails of a distribution from relatively small samples, (2) attempts to fit data to entire distributions usually involve knowledge of the first few central moments, and (3) regression problems rapidly get into higher mixed moments and the shape of conditional distributions whenever one attempts to fit real life data. It would be nice to have the central limit theorem and resultant robust estimates of the uncertainty working for the statistician in these more complicated problems. With certain caveats about error in very small samples, the jackknife offers just such a tool.

Hoeffding [1940] extended the central limit theory to a large class of statistics, but his U-statistics remained primarily a theoretical tool, because it is quite difficult to derive usable estimates of the variance of a U-statistic

from the sample. In fact, it is usually possible to calculate the theoretical variance of a U-statistic only when one knows at least the term of the underlying distribution. This difficulty was first overcome by Sen [1960].

Arvesen, by extending Miller's work on functions of the mean to functions of U-statistics, has simply produced a kind of "studentization" of Hoeffding's extension of the central limit theorems. It now becomes possible to use the jackknife to estimate the variance of a U-statistic or function of U-statistics from the data, and the t tables now become a useful set of probability bounds for constructing believable confidence intervals.

Arvesen was primarily interested in applying these results to the problem

$$Y_{ij} = \mu + a_i + e_{ij},\ i = 1,\ldots,n,\ j = 1,\ldots,k\ , \tag{2.4}$$

the $\{a_i\}$ independent $\eta(0, \sigma_A^2)$, the $\{e_{ij}\}$ independent $\eta(0, \sigma_e^2)$, test

$$H_0:\ \theta = \sigma_A^2/\sigma_e^2 \leq \theta_0 \text{ vs.} \tag{2.5}$$

$$H_A:\ \theta > \theta_0\ .$$

It is well-known that the standard F-test for (2.5) is non-robust against non-normality, especially of the random effect terms. Even the significance levels are incorrect unless $\theta_0 = 0$.

Letting

$$\text{MSA} = \textstyle\sum_{i=1}^{n} k(Y_{i.}-Y_{..})^2/(n-1)\ ,$$

$$\text{MSE} = \textstyle\sum_{i=1}^{n} \sum_{j=1}^{k} (Y_{ij}-Y_{i.})^2/n(k-1))$$

denote the usual mean squares, $\hat{\theta}_n^0 = \text{MSA/MSE}$, one can follow (1.1). Each interaction, the k observations associated with a main random effect are eliminated.

Also, note that

$$MSA = \binom{n}{2}^{-1} \sum_{\alpha_1<\alpha_2} k\,(Y_{\alpha_1 .} - Y_{\alpha_2 .})^2/2 ,$$

$$MSE = \binom{n}{1}^{-1} \sum_{\alpha_1} \sum_{j=1}^{k} (Y_{\alpha_{1j}} - Y_{\alpha_1 .})^2/(k-1) ,$$

and hence they are U-statistics.

In Arvesen and Schmitz [1970], it was shown by Monte Carlo techniques that one would prefer to jackknife $\hat{\theta}_n^0 = \log(MSA/MSE)$, especially with moderate sized samples. Again this variance stabilizing transformation (in the case of normality), has the advantage of symmetrizing the distribution. Since it is possible to deal with functions of U-statistics, the intraclass correlation coefficient, $\rho = \sigma_A^2/(\sigma_A^2 + \sigma_e^2)$ is also treated.

Finally, in Arvesen and Schmitz, the following problem is treated. Consider $S_1 \sim \mathcal{W}(k \Sigma_1 + \Sigma_2, n-1)$, $S_2 \sim \mathcal{W}(\Sigma_2, n(k-1))$ independently where $\mathcal{W}(\Sigma, a)$ denotes the Wishart distribution with expectation $a.\Sigma$ and a degrees of freedom. Then if Σ_1, Σ_2 are bivariate positive semidefinite and positive definite covariance matrices respectively, one may be interested in the between group correlation coefficient

$$\rho = \frac{(\Sigma_1)_{12}}{((\Sigma_1)_{11}(\Sigma_1)_{22})^{1/2}} \quad , \text{ where}$$

$(\Sigma_1)_{ij}$ denotes the element in the ith row and jth column of Σ_1. Let

$r = \dfrac{(S_1-S_2)_{12}}{[(S_1-S_2)_{11}\ (S_1-S_2)_{22}]^{1/2}}$. This is

useful to geneticists as a genetic correlation coefficient between traits. Results of Brown [1969] illustrate that the standard $\tanh^{-1}$ transformation of Fisher is not always satisfactory with this estimate (assuming normality). Unfortunately, the appropriate transformation has not been found.

Results given by Arvesen and Schmitz show that the jackknife confidence intervals for ρ as defined above (when using Fisher's $\tanh^{-1}$ transformation) may be drastically different than those given by the standard procedure of quoting the estimate and its estimated standard error (where the estimated standard error assumes normality).

Another possible application of the results of Arvesen [1969] is to obtain a confidence interval interval for the correlation coefficient of a bivariate sample (without the assumption of normality). The standard confidence interval for ρ is based on work of F. N. David [1938] (also reported in Anderson [1958], Ch. 42). It is not robust against nonnormality. Let $\binom{X_i}{Y_i}$, $i = 1,\ldots,N$ have mean $\binom{\mu_X}{\mu_Y}$, covariance matrix

$$\begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

(one again needs the mathematical nicety of finite fourth moments).

Let

(2.7) $$\hat{\theta}_n^0 = \tanh^{-1} r ,$$

where r is the sample correlation coefficient. Then if as in (2.2), (2.3) $L = \hat{\theta} - (t_{\alpha/2;n-1} \; s_{\hat{\theta}}/\sqrt{n}$, $U = \hat{\theta} + t_{\alpha/2;n-1} \; s_{\hat{\theta}}/\sqrt{n}$, where $N = nk$, a $(1-\alpha) \times 100\%$ confidence interval for ρ is given by

(2.8) $$\tanh L \leq \rho \leq \tanh U .$$

Layard [1971] and Hall [1971], in separate papers have shown how the jackknife may be profitably used in testing homogeneity of variances of p univariate populations. The standard Bartlett test is shown to behave quite poorly when applied to nonnormal data. Let $X_{i1},\ldots,X_{in_i}$ be independent with cumulative distribution function $F(\frac{x_i-\mu_i}{\sigma_i})$, finite fourth moments, F, μ_i, σ_i unknown, $i = 1,\ldots,p$. If one wants to test

(2.9) $$H_0: \; \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_p^2 ,$$

a test based on the jackknife is proposed. Let

$$s_i^2 = (n_i-1)^{-1} \sum_{m=1}^{n_i} (X_{im}-\bar{X}_{i.})^2 ,$$

$$\bar{X}_{i.} = n_i^{-1} \sum_{m=1}^{n_i} X_{im} ,$$

$$s_{i(j)}^2 = (n_i-2)^{-1} \sum_{\ell \neq j} (X_{i\ell} - \bar{X}_{i(j)})^2 ,$$

$$\bar{X}_{i(j)} = (n_i-1)^{-1} \sum_{\ell \neq j} X_{i\ell} ,$$

$$U_{ij} = n_i \log s_i^2 - (n_i-1) \log s_{i(j)}^2 .$$

Then as $\min_{1\le i\le p} (n_i) \to \infty$, if H_0 is true,

$$(2.10) \qquad \frac{\sum_{i=1}^{p} n_i (\bar{U}_{i.} - \bar{U}_{..})^2/(p-1)}{\sum_{i=1}^{p} \sum_{j=1}^{n_i} (U_{ij} - \bar{U}_{i.})^2/(N^*-p)} \xrightarrow{\mathcal{L}} \chi^2_{p-1}/(p-1) \quad ,$$

where $\bar{U}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} U_{ij}$,

$\bar{U}_{..} = (N^*)^{-1} \sum_{i=1}^{p} \sum_{j=1}^{n_i} U_{ij}$,

$N^* = \sum_{i=1}^{p} n_i$.

Hence, one rejects H_0 in (2.9) if the statistic in (2.10) is larger than the appropriate critical point of a χ^2 distribution with p-1 degrees of freedom (normalizing by the p-1 factor). Extensive Monte Carlo results indicate favorable results with moderate sample sizes.

One frequently encounters the situation where counts are taken of a relatively rare (or highly probable) event in the presence of differing levels of some suspected causative agent. For instance, a major element of controversy in public policy involves the use of suspected or proven carcinogens or mutagens in the environment. One view holds that there is a dose-response relationship involved and that one can calculate minimum acceptable levels. Another view is that there exists a quantal relationship and one molecule of the material is as dangerous as a kilogram. On a less cataclysmal level, table (1) shows the results of complete "cures" using a drug against three organisms. The agent was known to kill organism A and was suspected of being effective against organism B. It was not thought to be useful against organism C. Three doses were available to the investigators, equally spaced on a logarithmic scale. The dose

chosen for a given patient was based upon the investigator's clinical evaluation of the seriousness of the disease and the capacity of the patient to handle treatment. Ex post facto, the question arose whether there might not be a dose-response curve involved and that more heroic doses could be expected to affect organism C. The jackknife was used on the numerator of the least squares slope estimator of the original 0,1 variable against dose. That is, we jackknifed the estimator

$$(2.11) \qquad \hat{\beta} = \sum (Y_i - \bar{Y})(d_i - \bar{d}), \qquad Y_i = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} \; d_i = \log \text{dose}.$$

This analysis was used to conclude that there might be a dose response relationship against organism A, but the cure rate using the three doses now available was so high that it did not warrant the potential problems involved in trying higher levels against this organism. It was also decided that the evidence was sufficiently strong against there being a dose response relationship in the case of organism C, and further investigation of this drug in such cases would not be pursued. There was a clear dose response with respect to organism B. However, there already exist treatments superior to this one for organism B, and this line of investigation was dropped.

Since there were only three doses involved and the response variables were dichotomous, there are only six possible values that the pseudo-values can take on. These are displayed in table (1) along with the results of the jackknife.

The behavior of this particular use of the jackknife was investigated in Salsburg [1971]. Under the null hypothesis of no difference in response, exact probability levels were computed for the "t statistic" that results from the jackknife for three doses, equally separated, equal numbers of observations at

Table (1) Dose-Responses Using the Jackknife

	Organism A			Organism B			Organism C		
Log Dose	1	2	3	1	2	3	1	2	3
Exposed per dose	62	38	27	25	5	42	62	38	49
Successes per dose	55	38	26	7	5	39	0	0	1
Pseudo-value for Y=1	-5.799	2.201	10.201	-25.960	-4.960	16.036	-135.086	12.913	160.913
Pseudo-value for Y=0	86.205	-32.795	151.795	63.040	12.040	-38.964	0.914	-0.087	-1.087
Mean $\hat{\beta}$ (jackknife est.)		3.791			19.960			1.087	
Standard Deviation		24.876			29.619			13.244	
t Statistic		1.718			5.717			1.005	
Total Observations		127			72			149	

each dose, with the number of observations per dose running from 5 to 20. Computations were made for underlying probabilities of response of 0.90, 0.95, 0.99. The results indicate that the one-sided t test based on the jackknife is a conservative test of alpha level. That is, if the table of t statistic probabilities are used to compute rejection regions, the true probability of rejection will be less than or equal to the nominal alpha. However, the test approximates its true alpha level at 20 observations per dose and at an underlying probability of 0.90. Monte Carlo studies on several different alternate hypotheses indicate that the test has low power (as might be expected from its conservative size).

Except for the earlier discussion of the bio-assay problem in conjunction with the extension of Brillinger's result for MLE's, we have restricted attention to the identically distributed case. Again working with U-statistics (or functions of U-statistics), Arvesen and Layard [1971] were able to obtain procedures based on the jackknife when the observations are not necessarily identically distributed.

The procedure was applied to the test (2.5) in the model of (2.4) when the groups were unbalanced, that is for $i = 1,\ldots,n$, $j = 1,\ldots,k_i$, $n_i \geq 2$ for all i. Spjøtvoll [1967] has shown that in this case (with the assumption of normality), that an optimal test of (2.5) is given by the following.

Let

$$T = \sum_{i=1}^{n} k_i(\theta_0 k_i+1)^{-1} (\bar{Y}_{i.}-\bar{Y}_{..})^2 \left(\sum_{i=1}^{n} \sum_{j=1}^{k_i} (Y_{ij}-\bar{Y}_{i.})^2\right)^{-1}$$

where $\bar{Y}_{i.} = k_i^{-1} \sum_{j=1}^{k_i} Y_{ij}$, and

$$\bar{Y}_{..} = \left(\sum_{i=1}^{n} k_i(\theta_0 k_i+1)^{-1}\right)^{-1} \sum_{i=1}^{n} k_i(\theta_0 k_i+1)^{-1} \bar{Y}_{i.} ,$$

$$K = \sum_{i=1}^{n} k_i ,$$

and then one rejects H_0 at the α-level if

$$(K-n)\ (n-1)^{-1}\ T > F_{\alpha;\ K-n,\ n-1} , \text{ where}$$

$F_{\alpha;\ \nu_1,\nu_2}$ denotes the upper α point of an F distribution with ν_1, ν_2 degrees of freedom. Spjøtvoll explains that this test is optimal for "distant" alternatives. For "close" alternatives, a more complicated test statistic arises. Without the normality assumption, Spjøtvoll's test is not robust.

The use of the jackknife is possible if one defines

$$\hat{\theta}_n^0 = \log\ ((K-n)(n-1)^{-1}\ T), \text{ and the}$$

jackknife procedure as in (1.1). Monte Carlo results are given which indicate that this jackknife procedure is a good competitor to Spjøtvoll's procedure even when the data are moderate and normality is present.

Two recent papers open up entire new areas to which the jackknife may be applied with success. They deal with multivariate problems (Layard [1972]), and stochastic processes (Gray, Watkins, and Adams [1972]). We treat them in

that order. Consider $X_1,\ldots,X_{N_1} \sim \eta_p(\mu_1, \Sigma_1)$, $Y_1,\ldots,Y_{N_2} \sim \eta_p(\mu_2, \Sigma_2)$, and test

(2.13) $$H_0: \Sigma_1 = \Sigma_2 .$$

Layard shows that standard tests based on normal theory are not robust as to significance level when the $\{X_i\}$ and $\{Y_j\}$ are nonnormal (as they must be in the real world).

To illustrate the use of the jackknife to test H_0 in (2.13), let $N_1 = N_2 = N$, and in the grouping $N = nk$, $k = 1$. Let

(2.14)
$$T = \sum_{i=1}^{N} (X_i-\bar{X})\ (X_i-\bar{X})' ,$$

$$U = \sum_{i=1}^{N} (Y_i-\bar{Y})\ (Y_i-\bar{Y})' ,$$

and $T_{\ell m}$, $U_{\ell m}$ denote the entries in the ℓth row and mth columns, $\ell,m = 1,2$. Let

$$g(T) = (\log T_{11},\ \log T_{22},\ \tanh^{-1}\ (\frac{T_{12}}{\sqrt{T_{11}\ T_{22}}}))' ,$$

with $g(U)$ defined similarly. Then if $\hat{\theta}_n^0(T) = g(T)$, $\hat{\theta}_i(T)$, $\theta(T)$ as in (1.1) (except here we are dealing with vectors), and

$$S^2_{\hat{\theta}(T)} = (N-1)^{-1} \sum_{i=1}^{N} (\hat{\theta}_i(T) - \hat{\theta}(T))\ (\hat{\theta}_i(T) - \hat{\theta}(T))' ,$$

with similar expressions based on $\hat{\theta}_n^0(U)$. Then Layard shows that if H_0 in (2.13) is true,

$$(2.15)\qquad N(\hat{\theta}(T) - \hat{\theta}(U))' \left(S^2_{\hat{\theta}(T)} + S^2_{\hat{\theta}(U)}\right)^{-1} (\hat{\theta}(T) - \hat{\theta}(U)) \xrightarrow{\mathcal{L}} \chi^2_3 .$$

Hence one rejects H_0 if this test statistic is larger than the appropriate critical value of a χ^2_3 distribution (in general the appropriate statistic is a $\chi^2_{p(p+1)/2}$). If $N_1 \neq N_2$, the procedure is readily extended.

Another promising extension of the jackknife is found in Gray, Watkins, and Adams [1972]. This study appears to have been motivated by earlier work of Gaver and Hoel [1970]. In this latter paper, the authors were concerned with estimation of the reliability function associated with a Poisson process. That is, if $N(t)$ is a Poisson process with parameter λ on $[0,T]$, estimate $f(\lambda) = e^{-\lambda\chi}$, $\chi > 0$. A standard estimate (the MLE) is to estimate $f(\lambda)$ by $e^{-\hat{\lambda}\chi}$, $\hat{\lambda} = \frac{N(T)}{T}$. Let $[0,T]$ be partitioned into n intervals of equal length, $0 = t_n < t_1 < t_2 < \ldots < t_n = T$.

Let $\hat{\theta} = \hat{\theta}(0,T) = \frac{N(T)}{T} = \hat{\lambda}$.

$$\hat{\theta}_i = \hat{\theta}(t_{i-1}, t_i) = \frac{N(t_i) - N(t_i - 1)}{t_i - t_{i-1}} = \hat{\lambda}_i$$

$$\hat{\theta}^i_n = \frac{n}{n-1}\hat{\theta} - \frac{1}{n-1}\hat{\theta}_i , \qquad i = 1,\ldots,n .$$

Then a jackknife estimate of θ, based on $\hat{\theta}$, is suggested and is defined to be

$$J_n(e^{-\hat{\lambda}\chi}) = ne^{-\hat{\theta}\chi} - \left(\frac{n-1}{n}\right) \sum_{i=1}^{n} e^{-\hat{\theta}^i_n \chi} .$$

As the partition becomes finer, or more precisely, as $n \to \infty$,

$$(2.16) \qquad J_n(e^{-\hat{\lambda}\chi}) \Rightarrow e^{-\hat{\lambda}\chi} \{ 1-N(T)[e^{\chi/T} - 1 - \chi/T] \} .$$

This estimator was then studied with respect to its robustness in case of departures from the Poisson assumption.

Gray, Watkins, and Adams extend to above results to arbitrary piecewise continuous stochastic processes. In addition, the original estimate must be of the form, for $t_1, t_2 \in [0,T]$,

$$\hat{\theta}\ (t_1,t_2) = \frac{I_G\ (t_2) - I_G\ (t_1)}{t_2 - t_1} \quad , \text{ where}$$

if $\{G(t) \mid t \in [0,T]\}$ is the original stochastic process, $\{I_G(t) \mid t \in [0,T]\}$ is a stochastic process determined by the original stochastic process, and almost every realization is piecewise continuous. Unfortunately, an important example, to be discussed in section 3, does not satisfy these restrictions.

3. Some possible extensions. Basically, we have seen two types of situations to which the jackknife seems to be a useful tool in obtaining approximate tests and confidence intervals. One of these situations was where the distribution theory associated with typical estimates is difficult or perhaps impossible to obtain. This situation was exemplified by the MLE of the slope coefficient in the logistic bio-assay model, and in the situation proposed by Salsburg [1971]. The other situation was one in which the usual techniques (which assume normality) behave in a non-robust fashion when normality is not present. This situation was illustrated by examples dealing with variances, variance components, and covariance matrices. At present, there seem to be far too few examples where the jackknife procedure was successfully applied in a situation where distribution

theory is difficult or impossible - that is in the former case. Some examples of these will now briefly be discussed.

The theory of classifying a multivariate observation into one of two multivariate normal populations is well worked out when the population means are known and the population covariance matrices are known and equal (see Anderson [1958], Ch. 6). In fact, it is possible to obtain error rates for misclassifying observations. Unfortunately the population parameters are rarely known, and must be estimated from some preliminary sample. If

$$X_{1i} \sim \mathcal{N}_p(\mu_1, \Sigma) ,$$

$$X_{2i} \sim \mathcal{N}_p(\mu_2, \Sigma) , \quad i = 1,\ldots,N \text{ all independent}$$

one quantity of interest is the population Mahalanobis distance $\alpha = (\mu_1-\mu_2)' \Sigma^{-1}(\mu_1-\mu_2)$. The MLE of α is $(\hat{\mu}_1-\hat{\mu}_2)' \hat{\Sigma}^{-1} (\hat{\mu}_1 \hat{\mu}_2)$,

where $\hat{\mu}_1 = N^{-1} \sum_{i=1}^{N} X_{1i}$,

$$\hat{\mu}_2 = N^{-1} \sum_{i=1}^{N} X_{2i} ,$$

$$\hat{\Sigma} = \sum_{i=1}^{N} \frac{[(X_{1i} - \hat{\mu}_1)(X_{1i}-\hat{\mu}_1)' + (X_{2i} - \hat{\mu}_2)(X_{2i} - \hat{\mu}_2)']}{2N}$$

The distribution theory for $\hat{\alpha}$ is very complicated. A related problem is found in Sitgreaves [1952].

Now, following Brillinger (or Fryer's extension to a function of several MLE's), it is possible to follow (1.1) and obtain a confidence interval for α.

Actually, in practice one might prefer to take $\hat{\theta}_n^0 = \log(\hat{\alpha})$. Of course, $\hat{\alpha}$ is also a function based on U-statistics. The lower bound on the confidence interval would be of interest in obtaining probabilities of misclassification. Actually, this idea goes back go Mosteller and Tukey [1968], although not quite in this form.

Mosteller and Tukey describe the jackknife as having as general use as the Boy Scout's trusty tool after which it was named. This generality makes the jackknife particularly handy in the early creative part of data analysis or when conjecturing hypotheses from a preliminary set of information. In such situations, one usually starts with a large class of models which might, a priori, apply to the problem at hand. If there is some parameter associated with this class for which a consistent estimator exists which is a sufficiently smooth function of U-statistics or for which a computable maximum likelihood estimator exists, then the jackknife can be used to produce confidence bounds on this parameter or a rejection region that will enable the investigator to restrict his attention to a sub-class of the models.

For instance, in the development of drugs, a new compound might be created which appears, in animal trials, to produce its toxic effects by some unanticipated mode of action. Dose-response curves are straight lines developed by regressing some function of the response variable (such as the arc-sine of a proportion) on some other function of the dose (such as the logarithm.) But, the choice of transformations can be made arbitrarily from a wide range of such functions, and a test which enables the investigator to reject linearity in favor of convexity (or concavity), such as described below, is a useful tool for reducing the class of transformations under consideration.

The jackknife has been applied to the early phases of data analysis in several problems arising in the drug industry. Some examples and empirical results are described below.

Testing for Symmetry

One is frequently tempted to apply a paired t test to paired before-after observations, justifying the use of the t distribution with the vague statement that the difference between two identically distributed random variables is symmetric and that Efron [1969] indicates that the t is robust under symmetry. Under a null hypothesis of no difference between observations, the paired differences are, in fact, differences of pairs of identically distributed random variables. But, what if the null hypothesis is wrong? How good are the resulting confidence bounds? One way of checking on the assumption of symmetry is to examine the sample third central moment - a notoriously variable variate. Table (2) shows some results of applying the jackknife to the third central moment of before-after data taken from clinical trials of new drugs. As a comparison for the jackknifed t test, the sample third central moment has been divided by a sample estimate of its standard error (based upon the assumption of normality) which uses the first six sample central moments. Although all three of these measures have been shown not to have normal distributions, the normalized third central moment frequently agrees with the jackknifed t (especially when one cannot reject the hypothesis of symmetry). This suggests a similar degree of robustness for both methods of estimation when the data is symmetric and divergence when it is not. Theoretical considerations also suggest that the jackknifed t may be the more robust for non-symmetric distributions.

Table (2)

Small Sample Performance of Jackknifed Third Central Moment on Before Treatment-After Treatment Differences

	N		$\sum \frac{(X_i-\bar{X})^3}{N}$	Jackknifed Estimates	Normalized 3rd Central Moment	Jackknifed "t" Statistic
	13		15.10	18.76	0.60	0.89
Psychoneurotic Anxiety	9		-2.11	-2.86	-0.44	-0.93
Scores	13		23.17	28.78	0.92	1.48
	15		3.53	4.26	0.067	0.44
	13		-28.55	-35.46	-1.00	-1.80
	13		-8.05	-10.00	-1.12	-1.35
	9		-4.85	-2.86	-0.44	-0.93
	9		0.021	0.014	0.65	0.14
Fasting Blood Sugar	9		0.016	0.0006	1.03	0.169
	9		0.0001	0.0002	0.403	0.644
	9		-0.238	-0.033	-1.07	-1.28
	9		-0.0001	-0.0002	-0.23	-0.44
	9		0.00	0.0001	0.46	0.90
Systolic Blood Pressures	24		-107.0	-139.0	-0.0001	-0.0002
	24		-9×10^{6}	-10^{8}	-6.41	-9.88

Testing Dose Response - Linear versus Concave (Convex)

Thornby and Rao [1969] have proposed a class of non-parametric estimators and tests that involve the convexity of a regression. Heuristically, the procedure divides the independent variable's range into several regions. Thornby has formalized a specific test that uses three regions, and so we will illustrate with three regions in figure (1).

For a typical point in region I (X_i in the figure), we compute the slope of the line to a typical point in region II (X_j in the figure) -- call it B_{12} -- and the slope of the line to a typical point in region III (X_k in the figure) -- call it B_{13}. Then, define

$$h(X_i, X_j, X_k) = \begin{cases} -1 & \text{if } B_{12} < B_{13} \\ +1 & \text{if } B_{12} > B_{13} \\ 0 & \text{otherwise.} \end{cases}$$

With a little clevel juggling of indices to identify the kernel, it can be shown that the average of the $h(.,.,.)$ is a linear function of a U-statistic, and that its expectation is 0 if the regression is linear.

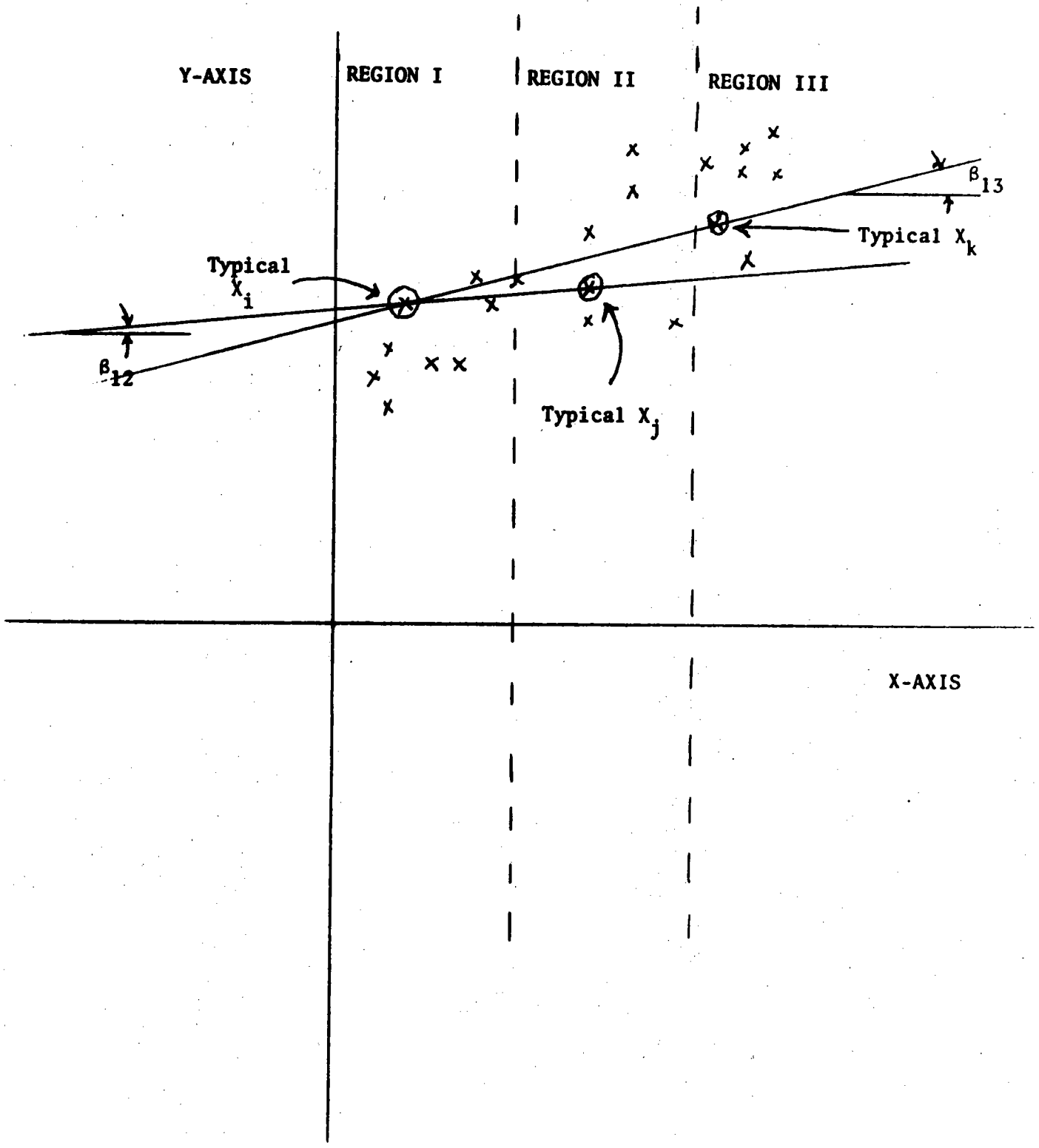
Monte Carlo studies have been run by one of the authors in an attempt to understand the effects of jackknifing this estimator. Two kinds of errors were imposed on a linear regression of the form

$$y = x \ .$$

The first error was normal with mean 0, variance 1; the second error, a mixture of normals, N (0, 1) with $p = 0.8$ and N (0, 10) with $p = 0.2$.

FIGURE (1) - THORNBY - RAO TEST

OF LINEARITY



Equal groups of observations were generated at $x = 5, 10, 15$. The concatenation of a Monte Carlo generation and the jackknife used excessive amounts of computer time (at 10 observations per group, a Monte Carlo study of 200 sample runs took 44 minutes on a PDP-10), and so the investigation was limited to 200 runs per trial.

Initial results on 10 observations per group indicated that the distribution of the resulting 't' statistics were essentially the same for both types of random error. (A Smirinov test comparing the two distributions produced an alpha level of 0.75.) For this reason, the bulk of the investigation was done for the first type of error [N (0,1)] only. Tables 3 and 4 display the results for 2 observations per group, 3 per group, and 5 and 10 per group. At 2 and 3 observations per group, the resulting 't' statistics had discrete distributions with a small number of possible values. The resulting frequency counts for each of these possible values are shown in table 3. Theoretically, the distribution should have been symmetric and the results appeared to corroborate this, so the observed frequency counts were 'folded' to produce more stable point probability estimates, which are also displayed. Note that at $n_i = 3$ there is a positive probability of infinite value. This occurs when all pseudo variates are equal and the denominator of the 't' statistic is zero. Note also that the tail probabilities are close to what one might expect from the t distribution. However, 'improbable' tail values are unusually large.

Table 4 shows the results for $n_i = 5$ and $n_i = 10$. Here, the number of possible values (for a 't' statistic which is still discrete) was much too large to display, and so counts are shown of percentile cells taken from the tabled values of the appropriate t statistics. Chi squared goodness of fit tests were then run to compare the results to the predicted counts for the 12 cells displayed, and these are also shown. At $n_i = 10$, the fit is quite good,

Table (3)

Thornby-Rao Test Jackknifed

Estimated Discrete Distribution, Small Sample Sizes

$n_i = 2, \sum n_i = 6$			$n_i = 3, \sum n_i = 9$		
Discrete "t" values	Counts N=200	folded prob. estimate	Discrete "t" values	Counts N=200	folded prob. estimate
			$-\infty$	2	.0150
-1.7×10^5	4	0.0325	-2.143	8	.0450
-1.643	6	0.0250	-1.250	8	.0275
-0.885	3	0.0225	-0.968	16	.0775
-0.775	22	0.1075	-0.581	14	.0875
-0.701	12	0.0525	-0.530	4	.0150
-0.245	26	0.1175	-0.433	16	.0675
-0.149	13	0.0800	-0.194	4	.0225
0.000	25	0.1250	-0.177	13	.0475
0.149	19	0.0800	-0.153	17	.0950
0.245	21	0.1175	0.153	21	.0950
0.701	9	0.0525	0.177	6	.0475
0.775	21	0.1075	0.194	5	.0225
0.885	6	0.0225	0.433	11	.0675
1.643	4	0.0250	0.530	2	.0150
1.7×10^5	9	0.0325	0.581	21	.0875
			0.968	15	.0775
			1.250	3	.0275
			2.143	10	.0450
			∞	4	.0150

Table (4)

Thornby-Rao Test Jackknifed Medium Sized Sample Properties

Goodness of Fit to Tabled t Distribution (200 Trials per Model)

t Statistic Tabled Percentiles	Frequency Counts	
	$n_i = 5$, $\sum n_i = 15$	$n_i = 10$, $\sum n_i = 30$
[0, .05]	13	9
(.05, .10]	9	11
(.10, .20]	24	21
(.20 .30]	26	27
(.30 .40]	16	20
(.40 .50]	25	20
(.50 .60]	20	23
(.60 .70]	22	24
(.70 .80]	15	17
(.80 .90]	14	16
(.90 .95]	4	4
(.95 1.0]	12	8
$\chi^2_{(12)}$	12.700	9.200

suggesting that for samples of this size or larger one can rely on the asymptotic properties of the jackknife. The 'fat' tails that were so obvious for smaller counts still influence the distribution at $n_i = 5$. In fact, there were four runs which produced 't' statistics greater in absolute value than 7.0, and one of these was was infinite (zero estimated variance).

Thus, it would appear that one is reasonably safe to jackknife this estimator with moderately sized samples provided he has a fixed alpha level determined before the investigation. Attempts to attribute something remarkable to extremely large absolute values can lead to error.

Trends in Variance During the Course of a Clinical Trial

In clinical trials of psychotherapeutic drugs, an important measure of effect is the psychiatric traing scale. The clinician records his impression of a large number of specific patterns of patient behavior on a three to seven point scale. That is, he may record that the patient fidgits to a moderate degree, suffers from early insomnia to a severe degree, complains of cardiovascular symptoms to a mild degree, etc. By the use of principal component analysis, factors have been derived from these scales that appear to represent stable and independent aspects of patient response. There is a considerable placebo response in this kind of trial, and over a course of time many of the items of a rating scale tend to be reduced to the lowest category (usually coded as a zero). Thus, it is only natural to expect a decrease in both mean and variance.

However, general "eyeballing" of data may suggest that factors associated with drug activity tended to show a more precipitous drop in variance than did factors not associated with a drug's activity or factors derived from patients on placebo. Here is a typical problem that can arise early in the analysis of

data. Without the underpinning of any mathematical model, a consistent pattern begins to emerge in the numbers. Should this be pursued? Is the eye being fooled by purely random phenomena, or can this be used as a hint out of which one might generate a mathematical model and gain further insight into differences between treatments?

In this case, there do not exist any statistical tests ready to examine the randomness of such an observation. The variances being observed are derived from highly correlated time series, the underlying distributions are clearly skewed, and there is an essential infimum which is clearly having an effect on both mean and variance. On the other hand, the jackknife calls for only an estimator that can be written as a sufficiently smooth function of U-statistics. Tables (5) and (6), display some typical results of the investigation. Least squares linear regressions were computed on the model

$$\log(\text{variance}) = A + B(\text{time})$$

and the slope estimator, B, was jackknifed.

In both examples displayed in the tables, patient improvement (as measured by baseline-final mean differences) was highly significant. Two sample tests comparing that improvement between groups failed to show any significance (most likely due to the small sample size). The jackknife suggests a difference in response not only because of the resulting t statistics - one is significant at 5% and the other is not - but also because 11 out of 16 pseudo variates in one group are positive and only 6 out of 23 are positive in the other group. Well over half of the patients who contributed positive pseudo variates in both groups showed some deterioration in condition towards the end of the study. All of those contributing negative pseudo variates showed continual improvement throughout the study. It

Table (5) - JACKKNIFE APPLIED TO SLOPE

ESTIMATE - I

MEAN	NO. OF PAT.	VARIANCE	STD ERROR
1.747	17.	0.756	0.211
0.675	16.	0.486	0.174
0.506	16.	0.429	0.164
0.660	15.	0.401	0.164
0.270	10.	0.189	0.137

INITIAL ESTIMATES

SLOPE-LOG-VAR	VAR(X(1))	VAR(X(2))	VAR(X(3))
-0.1603E 00	-0.5266E 00	-0.7215E 00	-0.8472E 00

INDIVIDUAL PSUEDO VARIATES

-0.637E 00
0.228E 01
0.163E 00
0.163E 00
-0.385E 01
0.322E-01
-0.507E 00
-0.709E 00
0.322E-01
0.322E-01
0.163E 00
0.163E 00
-0.114E 01
0.163E 00
0.163E 00
0.322E-01

SUMMARY STATISTICS

MEANS	VAR	T TEST
-0.2038E 00	0.1372E 01	-0.717

Table (6) - JACKKNIFE APPLIED TO SLOPE

ESTIMATE - II

MEAN	NO. OF PAT.	VARIANCE	STD ERROR
2.466	23.	1.418	0.248
0.913	23.	0.956	0.204
0.425	23.	0.661	0.170
0.103	23.	0.201	0.094
0.030	21.	0.167	0.089

INITIAL ESTIMATES

SLOPE-LOG-VAR	VAR(X(1))	VAR(X(2))	VAR(X(3))
-0.3816E 00	0.3490E 00	-0.4512E-01	-0.4142E 00

INDIVIDUAL PSUEDO VARIATES

-0.982E 00
0.787E 00
0.504E 00
-0.580E 00
-0.999E 00
-0.249E 01
-0.180E 01
-0.162E 00
-0.461E 00
-0.646E 00
0.262E 00
-0.510E 00
0.172E 01
-0.388E 00
0.346E 00
0.334E 00
-0.245E 00
-0.299E 00
-0.444E 00
-0.859E 00
-0.610E 00
-0.620E 00
-0.605E 00

SUMMARY STATISTICS

MEANS	VAR
-0.3803E 00	0.7178E 00

T TESTS -2.153

would appear that the jackknife can be used in situations like this to detect subtle perturbations in the results of continuing treatment. Attempts are now being made to link this empirical observation to a reasonable mathematical model.

Now follows an example with no solution. This is the first example in which the observations were not independent. Consider normal observations, $X_1,\ldots,X_n$ means 0, variances σ^2, and Corr $(X_i, X_{i+j}) = \rho^j$. Think of $X_1,\ldots,X_n$ being responses from some phenomena placed on a line, each response has the same marginal distribution, but neighbors are correlated with a correlation that decreases geometrically with an exponent proportional to the distance of separation. There is essentially no problem to obtain MLE's of ρ or σ^2, at least iteratively on a computer. However, the method of Brillinger to obtain confidence intervals fails due to the non-independence of the observations. It remains to be seen, but perhaps the stochastic process approach of Gray et al may prove successful when modified to treat this type of case. This approach may also be helpful in making adjustments for the fact that in general, $\hat{\theta}_1,\ldots,\hat{\theta}_n$ of (1.1) are correlated.

One final example in which a solution may be more readily available is to obtain confidence intervals for variance components in the complicated mixed ANOVA models considered for example by Hartley and Rao [1967]. Using a formulation in terms of linear models, they obtain MLE's of variance components in what could be complicated models involving several classifications, possibly unbalanced, and possibly both fixed and random effects. The technique of Brillinger could probably be extended to obtain confidence intervals for the variance components in these complicated models. However, the question arises as to whether one wants to obtain intervals for all the parameters simultaneously. In much simpler situations, Miller [1966] shows that this is not an easy problem to resolve. Hence

the use of the jackknife in such a problem of simultaneous inference may prove very challenging as well as worthwhile. Of course, the numerical analysis problem involved in using iterative techniques (essentially non-linear programming) is a problem that might prove equally difficult in actually obtaining Hartley and Rao's MLE's.

Finally, we note that there are two problems that are very important from a practical viewpoint, and have received virtually no attention. These are the problem of deciding if one's data is adequate for the asymptotic result of (1.3) to hold, and what (if any) transformation should be used in conjunction with the jackknife. Only limited Monte Carlo results give an indication of an answer to the former question. Until a better answer comes along, the answer to the second question will probably remain that the so-called variance stabilizing transformation is the appropriate one. Actually however, the method proposed by Box and Cox [1964], namely to let the data give you the transformation, may be preferable.

REFERENCES

[1] Adams, J., Gray, H., and Watkins, T. [1971]. An asymptotic characterization of bias reduction by jackknifing. To appear in Ann. Math. Statist.

[2] Anderson, T. W. [1958]. An Introduction to Multivariate Statistical Analysis. Wiley, New York.

[3] Arvesen, J. [1969]. Jackknifing U-statistics. Ann. Math. Statist. 40, 2076-2100.

[4] Arvesen, J., and Layard, M. [1971]. Asymptotically robust tests in unbalanced variance component models. Mimeograph series #263, Purdue University Statistics Department.

[5] Arvesen, J., and Schmitz [1970]. Robust procedures for variance component problems using the jackknife. Biometrics 26, 677-686.

[6] Berkson, J. [1955]. Maximum likelihood and minimum χ^2 estimates of the logistic function. Jour. Amer. Statist. Assoc. 50, 130-162.

[7] Box, G. E. P., and Cox, D. R. [1964]. An analysis of transformations. Jour. Roy. Statist. Soc. Ser. B 26, 211-252.

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

[9] Brillinger, D. [1966]. The application of the jackknife to the analysis of consumer surveys. Commentary. The Journal of the Market Res. Soc. 8, 74-80.

[10] Brown, G. H. [1969]. An empirical study of the distribution of the sample genetic correlation coefficient. Biometrics 25, 63-72.

[11] David, F. N. [1938]. Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples, Biometrika. London.

[12] Dempster, A. [1966]. Estimation in multivariate analysis. Multivariate Analysis, V. I. Academic Press, New York.

[13] Efron, B. [1969]. Student's t-test under symmetry conditions. Jour. Amer. Statist. Assoc. 64, 1278-1302.

[14] Fryer, J. [1970]. Jackknifing maximum likelihood estimates in the multi-parameter case (preliminary report) (abstract) Ann. Math. Statist. 41, 1392.

[15] Gaver, D. P. and Hoel, D. G. [1970]. Comparison of certain small-sample Poisson probability estimates. Technometrics 12, 835-850.

[16] Gray, H. L., and Schucany, W. R. [1971]. The Generalized Jackknife Statistic. Marcel Dekker, Inc., New York.

[17] Gray, H. L., Watkins, T. A., and Adams, J. E. [1972]. A generical development of the generalized jackknife. To appear in Ann. Math. Statist., Feb. issue.

[18] Hall, I. J. [1971]. Some comparisons of tests for equality of variances. To appear in Jour. of Comp. and Sim.

[19] Hartley, H. O., and Rao, J. N. K. [1967]. Maximum-likelihood estimation for the mixed analysis of variance model. Biometrika 54 93-108.

[20] Hoeffding, W. [1948]. A class of statistics with asymptotically normal distribution. Ann. Math. Statist. 19, 293-326.

[21] Layard, M. [1971]. Robust large-sample tests for homogeneity of variances. University of California, Davis. Technical report.

[22] Layard, M. [1972]. Asymptotically robust tests, about covariance matrices. To appear in Ann. Math. Statist., Feb. issue.

[23] Mantel, N. [1967]. Assumption-free estimators using U-statistics and a relationship to the jackknife method. Biometrics 23, 567-571.

[24] Miller, R. G., Jr. [1964]. A trustworthy jackknife. Ann. Math. Statist. 35, 1594-1605.

[25] Miller, R. G., Jr. [1966]. Simultaneous Statistical Inference. Mc-Graw-Hill, New York.

[26] Miller, R. G., Jr. [1968]. Jackknifing variances. Ann. Math. Statist. 39, 567-582.

[27] Mosteller, F., and Tukey, J. W. [1968]. Data analysis, including statistics. Handbook of Social Psychology. G. Lindzey and E. Aronson, Eds., Addison-Wesley, Reading, Mass.

[28] Quenouille, M. [1949]. Approximate tests of correlation in time-series. Jour. Roy. Statist. Soc., Ser. B 11, 68-84.

[29] Quenouille, M. [1956]. Notes on bias in estimation. Biometrika 43, 353-360.

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

[31] Salsburg, D. [1971]. Testing dose responses on proportions near zero or one with the jackknife. To appear in Biometrics.

[32] Sen, P. K. [1960]. On some convergence properties of U-statistics. Calcutta Statist. Assoc. Bull. 10, 1-18.

[33] Schucany, W. R. [1971]. The reduction of bias in parametric estimation. SMU Statistics Department Technical Report No. 93.

[34] Schucany, W. R., Gray, H. L., and Owen, D. B. [1971]. Bias reduction in estimation. Jour. Amer. Statist. Assoc. 66, 524-533.

[35] Sitgreaves, R. [1952]. On the distribution of two random matrices used in classification procedures. Ann. Math. Statist. 23, 263-270.

[36] Spjøtvoll, E. [1967]. Optimal invariant tests in unbalanced variance component models. Ann. Math. Statist. 38, 422-429.

[37] Thornby, J. I., and Rao, P. V. [1969]. A robust point estimator in generalized regression model. Ann. Math. Statist. 40, 1784-1790.

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