A TRUSTWORTHY JACKKNIFE[1]

By Rupert G. Miller, Jr.[2]

The Johns Hopkins University

1. Introduction. A method for the reduction of bias in parametric estimation was introduced by Quenouille [6], and its properties were studied in some specific situations by Quenouille [7] and Durbin [4]. Tukey [9] proposed the general use of this technique in order to (a) reduce the bias and (b) obtain approximate confidence intervals in problems where standard statistical procedures may not exist or are difficult to apply. Tukey adopted the name of "jackknife" for this procedure, since a boy scout's jackknife is symbolic of a rough-and-ready instrument capable of being utilized in all contingencies and emergencies.

The jackknife procedure depends upon judiciously dividing the data into groups, obtaining estimates from combinations of these groups, and then averaging the estimates. Let θ be the unknown parameter, and let $(X_1, \cdots, X_N)$ be a sample of N independent, identically distributed observations with cdf F_θ, which depends upon θ. Suppose a reasonably good (but biased) method of estimating θ is available. Further suppose the data can be divided into n groups of size k $(N = nk)$, i.e., $(X_1, \cdots, X_k ; X_{k+1}, \cdots, X_{2k} ; \cdots ; X_{(n-1)k+1}, \cdots, X_{nk})$. Denote by $\hat{\theta}^i_{n-1}$, $i = 1, \cdots, n$, the estimate of θ obtained by deleting the ith group and estimating θ from the other $(n-1)k$ observations, e.g., $\hat{\theta}^1_{n-1} = \hat{\theta}(X_{k+1}, \cdots, X_{kn})$. Let $\hat{\theta}^0_n$ be the estimate of θ based on all nk observations. Form the new estimates (called "pseudo-values" by Tukey)

$$\hat{\theta}_i = n\hat{\theta}^0_n - (n-1)\hat{\theta}^i_{n-1}, \qquad i = 1, \cdots, n. \tag{1}$$

The jackknife estimate of θ is the average of the $\hat{\theta}_i$, $i = 1, \cdots, n$, i.e.,

$$\hat{\theta} = n^{-1}\sum_1^n \hat{\theta}_i = n\hat{\theta}^0_n - (n-1)\hat{\theta}^{\cdot}_{n-1}, \tag{2}$$

where $\hat{\theta}^{\cdot}_{n-1} = (\sum_1^n \hat{\theta}^i_{n-1})/n$.

The jackknife $\hat{\theta}$ exactly eliminates the $1/n$ term from any bias. For if

$$E(\hat{\theta}^0_n) = \theta + a/kn + b/(kn)^2 + \cdots, \tag{3}$$

for all n and k, then

$$\begin{aligned} E(\hat{\theta}) &= n(\theta + a/kn + b/(kn)^2 + \cdots) - (n-1)(\theta + a/k(n-1) \\ &\quad + b/(k(n-1))^2 + \cdots) = \theta - b/k^2 n(n-1) + \cdots. \end{aligned} \tag{4}$$

Quenouille [7] and Durbin [4] have shown that in certain ratio problems this

Received April 10, 1964; revised May 25, 1964.

[1] This research was supported in part by NSF Grant No. G-25218 awarded to The Johns Hopkins University.

[2] Permanent address: Stanford University.

reduction in bias is not accompanied by an appreciable increase in variance, and, in fact, in some instances the variance is reduced as well.

Tukey has gone one step farther and suggested that the jackknife will reduce most forms of bias. This is true for a $1/n^{\frac{3}{2}}$ term, for instance, but not a $1/n^2$ term. He has also proposed that in most instances $\hat{\theta}_1, \cdots, \hat{\theta}_n$ can be treated as n approximately *independent* (identically distributed) observations from which an approximate t-statistic confidence interval or test for θ can be constructed. The purpose of this paper is to see how true this latter proposition is. In Section 2 some simple examples are given in which this proposal is false—mildly so and then wildly so. Sections 3 and 4 are devoted to rigorous justification of this conjecture in two general situations, thereby giving to the jackknife a degree of "trustworthiness", the first scout law.

There are no mathematical guidelines yet on the relative choice of k and n, but there may be external design reasons for their choice. Throughout this paper it will be assumed that $k = 1$ for simplicity. The arguments can be readily duplicated for $k > 1$.

Quenouille [7] has pointed out that to compute $\hat{\theta}^i_{n-1}$ for all possible combinations of $(n - 1)k$ observations and then average the $\binom{N}{k}$ estimates would also work, and may be advantageous in reducing the variance. For a large scale experiment the amount of computation involved in this could become prohibitive and/or detract from the quick, rough-and-ready quality of the technique. Quenouille [7] has also extended his technique to eliminate terms of the order $1/n^2$. These modifications will not be considered in this paper.

2. Three counterexamples. The motivation of this section is to produce a simple estimation problem in which the jackknife technique of constructing confidence intervals or tests goes awry.

Let $X_1, \cdots, X_n$ be independently, identically distributed according to F_θ which has a positive density p_θ on the interval $[0, \theta]$, $\theta > 0$, and no probability outside the interval. A common example would be the uniform, i.e., $p_\theta(x) = 1/\theta$, $0 \leqq x \leqq \theta$; $= 0$, otherwise. In general, the parameter θ might be allowed to influence the shape as well as the spectrum of F_θ.

Suppose because of optimality considerations (e.g., the uniform case) or because of the intractability of any other method, it is decided to estimate θ by $\max\{X_1, \cdots, X_n\}$, a not irrational thing to do. Let $X_{(1)} \leqq X_{(2)} \leqq \cdots \leqq X_{(n)}$ be the order statistics corresponding to $(X_1, \cdots, X_n)$. Then, the jackknife gives

$$\begin{aligned} \hat{\theta}^i_{n-1} &= X_{(n)} \qquad \text{if } i \neq (n), \text{ i.e., } n - 1 \text{ times}, \\ &= X_{(n-1)} \qquad \text{if } i = (n), \text{ i.e., once}, \end{aligned} \tag{5}$$

and

$$\begin{aligned} \hat{\theta}_i &= X_{(n)} \qquad\qquad\qquad\qquad\qquad\quad \text{if } i \neq (n), \\ &= nX_{(n)} - (n - 1)X_{(n-1)} \qquad \text{if } i = (n), \end{aligned} \tag{6}$$

so

$$\hat{\theta} = X_{(n)} + [(n-1)/n](X_{(n)} - X_{(n-1)}). \tag{7}$$

The estimated variance is

$$s^2 = (n-1)^{-1}\sum_1^n (\hat{\theta}_i - \hat{\theta})^2 = [(n-1)^2/n](X_{(n)} - X_{(n-1)})^2, \tag{8}$$

so the supposed approximate t-statistic is

$$\begin{aligned} T_n &= \frac{n^{1/2}(\hat{\theta} - \theta)}{s} = \left(\frac{n}{n-1}\right)\left(\frac{X_{(n)} + ((n-1)/n)(X_{(n)} - X_{(n-1)}) - \theta}{X_{(n)} - X_{(n-1)}}\right) \\ &\sim 1 - \frac{\theta - X_{(n)}}{X_{(n)} - X_{(n-1)}}. \end{aligned} \tag{9}$$

Define $R_n = (\theta - X_{(n)})/(X_{(n)} - X_{(n-1)})$. Even without additional evaluation it is obvious the upper tail of the distribution of T_n is suffering a departure from the conjecture since $R_n \geqq 0$.

The asymptotic distribution ($n \to +\infty$) of R_n, and hence T_n, will be considered in the following three cases, each giving a quite different result. The parameter θ is now assumed to be fixed and positive.

Case A. p_θ is *asymptotically uniform* in the sense that $p_\theta(\theta - x)/p_\theta(\theta - cx) \to 1$ as $x \to 0$ for all $c > 1$. An example is the uniform distribution. In this case R_n has a limiting distribution:

$$\lim_{n\to\infty} P\{R_n \leqq r\} = r/(1+r), \qquad r \geqq 0. \tag{10}$$

Case B. p_θ is *asymptotically diminishing* in the sense that $p_\theta(\theta - x)/p_\theta(\theta - cx) \to 0$ as $x \to 0$ for all $c > 1$. An example is $p_\theta(\theta - x) = e^{-1/x}$ for x near 0. In this case R_n drifts off to $+\infty$ in probability, i.e., $\lim_{n\to\infty} P\{R_n \leqq r\} \to 0$ for all $r > 0$.

Case C. p_θ is *asymptotically exploding* in the sense that $p_\theta(\theta - x)/p_\theta(\theta - cx) \to c$ as $x \to 0$ for all $c > 1$. An example is $p_\theta(\theta - x) = 1/x(\log x)^2$ for x near 0. In this case R_n degenerates to zero in probability, i.e., $\lim_{n\to\infty} P\{R_n \leqq r\} \to 1$ for all $r > 0$.

For simplicity of notation in the succeeding derivations θ will be taken equal to 1 and dropped as a subscript on the density function.

The joint density function of $(X_{(n-1)}, X_{(n)})$ is

$$p_{X_{(n-1)},X_{(n)}}(x, y) = n(n-1)p(x)p(y)(F(x))^{n-2}, \tag{11}$$

for $0 \leqq x \leqq y \leqq 1$, zero otherwise, and the joint density of $U = 1 - X_{(n)}$ and $V = X_{(n)} - X_{(n-1)}$ is

$$p_{U,V}(u, v) = n(n-1)p(1-u-v)p(1-u)(F(1-u-v))^{n-2}, \tag{12}$$

for $0 \leqq v \leqq 1 - u$, $0 \leqq u \leqq 1$, zero otherwise. The probability $P\{U/V \geqq r\}$

for $r > 0$ is obtained by integrating (12) over the triangular intersection of the regions $0 \leqq v \leqq 1 - u$, $0 \leqq u \leqq 1$, and $u \geqq rv$.

$$P\{U/V \geqq r\} = \left(\int_0^{r/(1+r)} \int_0^{u/r} + \int_{r/(1+r)}^1 \int_0^{1-u}\right) n(n-1) \cdot p(1-u-v)\, p(1-u)\, (F(1-u-v))^{n-2}\, dv\, du. \tag{13}$$

As $n \to +\infty$ the asymptotic behavior of (13) is determined solely by the limiting behavior of the integral over a fixed, but arbitrarily small neighborhood $(0 \leqq u \leqq \epsilon,\ 0 \leqq v \leqq u/r)$ of the origin.

$$\lim_{n\to\infty} P\{U/V \geqq r\} = \lim_{n\to\infty} \int_0^{\epsilon} \cdot \int_0^{u/r} n(n-1)p(1-u-v)p(1-u)\,(F(1-u-v))^{n-2}\, dv\, du \tag{14}$$

(assuming limits exist); the remainder of the integral converges to zero (Lebesque dominated convergence theorem). Integration over v in (14) gives

$$\lim_{n\to\infty} P\{U/V \geqq r\} = 1 - \lim_{n\to\infty} \int_0^{\epsilon} \left(\frac{p(1-u)}{p(1-u(1+1/r))}\right) np(1-u(1+1/r)) \cdot (F(1-u(1+1/r)))^{n-1}\, du. \tag{15}$$

Define $I_n(\epsilon)$ to be the integral in (15). The constant c in the definition of Cases A, B, and C is now identified with $(1 + 1/r)$. Since ϵ can be fixed arbitrarily small, the following limits pertain.

Case A.

$$\lim_{n\to\infty} I_n(\epsilon) = \lim_{n\to\infty} \int_0^{\epsilon} np(1-u(1+1/r))(F(1-u(1+1/r)))^{n-1}\, du = \lim_{n\to\infty} \left(\frac{-1}{1+1/r}\,(F(1-u(1+1/r)))^n \Big|_0^{\epsilon}\right) = \frac{r}{1+r}.$$

Case B. For $\delta > 0$,

$$\lim_{n\to\infty} I_n(\epsilon) \leqq \delta \lim_{n\to\infty} \int_0^{\epsilon} np(1-u(1+1/r))(F(1-u(1+1/r)))^{n-1}\, du \to 0 \text{ as } \delta \to 0.$$

Case C.

$$\lim_{n\to\infty} I_n(\epsilon) = (1+1/r) \lim_{n\to\infty} \int_0^{\epsilon} np(1-u(1+1/r)) \cdot (F(1-u(1+1/r)))^{n-1}\, du = (1+1/r) \cdot r/(1+r) = 1.$$

Thus Case A produces a limiting distribution for T_n, which is not normal and has all its mass below $+1$, and Cases B and C have extreme degenerate behavior, i.e., degeneracy at $-\infty$ and $+1$, respectively. The t-confidence intervals could not in any sense be regarded as approximate (asymptotically) in Cases B and C, and would require a rather loose interpretation in Case A.

For the uniform distribution $n(\theta - X_{(n)})$ has a limiting negative-exponential distribution, and for Cases B and C no sequence $\{a_n\}$ of normalizing constants exists such that $a_n(\theta - X_{(n)})$ has a limiting non-degenerate distribution. The critical reader might complain that it is therefore senseless to investigate asymptotic normality. But this is just the point—to see whether the jackknife can create asymptotic normality when none exists prior to it.

It is interesting to note that for the uniform distribution the jackknife corrects the expectation to

$$E(\hat{\theta}) = \frac{n\theta}{n+1} + \left(\frac{n-1}{n}\right)\frac{\theta}{n+1} = \theta\left(1 - \frac{1}{n(n+1)}\right), \tag{16}$$

whereas the constant multiplier $(n+1)/n$ corrects the unjackknifed $X_{(n)}$ exactly. To get an idea of the performance of the jackknife estimator in the uniform case, it is informative to compare it risk-wise with the standard optimal estimators. The estimator $\hat{\theta}_u = (n+1)X_{(n)}/n$, mentioned above, is the UMV unbiased estimator of θ, and, for mean squared error, $\hat{\theta}_a = (n+2)X_{(n)}/(n+1)$ is the unique admissible estimator of the form $cX_{(n)}$, c constant (see Karlin [5]). It is clear $\hat{\theta}$ is in no sense optimal for squared error loss since it does not depend solely on the sufficient statistic $X_{(n)}$. Due to the strict convexity of the loss function, the Rao-Blackwellized estimator

$$\hat{\theta}_c = E\{\hat{\theta} \mid X_{(n)}\} = (n^2 + n - 1)X_{(n)}/n^2$$

constitutes a strict improvement. The respective mean squared errors for these various estimators are:

$$\begin{aligned}
E(\hat{\theta}_n^0 - \theta)^2 &= 2\theta^2/[(n+1)(n+2)], \\
E(\hat{\theta} - \theta)^2 &= [2\theta^2(n^2 - n + 1)]/[n^2(n+1)(n+2)], \qquad n \geqq 2, \\
E(\hat{\theta}_c - \theta)^2 &= [\theta^2(n^3 + n^2 - n + 1)]/[n^3(n+1)(n+2)], \quad n \geqq 2, \\
E(\hat{\theta}_u - \theta)^2 &= \theta^2/[n(n+2)], \\
E(\hat{\theta}_a - \theta)^2 &= \theta^2/(n+1)^2.
\end{aligned} \tag{17}$$

For $n \geqq 3$, MSE $(\hat{\theta}_n^0)$ > MSE $(\hat{\theta})$ > MSE $(\hat{\theta}_u)$ > MSE $(\hat{\theta}_c)$ > MSE $(\hat{\theta}_a)$, and for $n = 2$, MSE $(\hat{\theta}_n^0)$ > MSE $(\hat{\theta})$ = MSE $(\hat{\theta}_u)$ > MSE $(\hat{\theta}_c)$ > MSE $(\hat{\theta}_a)$. For n large the mean squared errors of $\hat{\theta}_n^0$ and $\hat{\theta}$ are roughly twice as large as those for $\hat{\theta}_c$, $\hat{\theta}_u$, and $\hat{\theta}_a$. Thus, although the jackknife estimator improves on the unadjusted maximum estimator, it does not do as well as the standard optimal estimators or its conditional expectation.

3. Transformation of means. Transformations of statistics are frequently used to stabilize variances and/or produce linearity or additivity. Reduction in non-normality might also give just cause for their use. These transformations are frequently of the form $f(\bar{x})$ where $\bar{x}$ is a sample mean. Witness, in evidence, arc sin $\hat{p}^{\frac{1}{2}}$ where $\hat{p}$ is a binomial estimate, $\bar{x}^{\frac{1}{2}}$ where $\bar{x}$ is Poisson, and log s^2 where s^2 is a normal sample variance. The family of powers, $(\bar{x} + c)^p$, $e^{c\bar{x}}$, and log $(c + \bar{x})$, often constitues an array from which a suitable choice can be made.

Although $\bar{x}$ is an unbiased estimate of μ, $f(\bar{x})$ will usually be a biased estimate of $f(\mu)$ because of the non-linearity of the transformation. For example,

$$E(\text{arc sin } \hat{p}^{\frac{1}{2}}) \cong \text{arc sin } p^{\frac{1}{2}} + (4n)^{-1}\{(p - \tfrac{1}{2})/[p(1 - p)]^{1/2}\}. \tag{18}$$

The jackknife procedure could be applied to reduce this bias.

$$\text{Let } \bar{X} = n^{-1}\sum_1^n X_j, \qquad \bar{X}^i = (n-1)^{-1}\sum_{j\neq i} X_j = (n\bar{X} - X_i)/(n-1).$$

The jackknife estimator of $\theta = f(\mu)$ is

$$\hat{\theta} = nf(\bar{X}) - (n-1)n^{-1}\sum_1^n f(\bar{X}^i). \tag{19}$$

The theorem below gives simple conditions under which $\hat{\theta}$ is asymptotically normally distributed.

THEOREM 1. *Let $\{X_i\}$ be a sequence of independent, identically distributed random variables with mean $\mu = 0$ and variance $0 < \sigma^2 < +\infty$. Let f be a function defined on the real line which, in a neighborhood of the origin, has a bounded second derivative. Then, as $n \to \infty$, $n^{\frac{1}{2}}(\hat{\theta} - \theta)$ is asymptotically normally distributed with mean zero and variance $\sigma^2(f'(0))^2$.*

PROOF. Let $I = (-3\Delta, +3\Delta)$, $\Delta > 0$, be any neighborhood of zero in which f'' is bounded. As $n \to +\infty$, $\bar{X} \to_p 0$ so $P\{\bar{X} \varepsilon (-\Delta, +\Delta)\} \to 1$. Also,

$$P\{\max\{|X_1|/n, \cdots, |X_n|/n\} > \Delta\} = 1 - (F_{|X|}(n\Delta))^n \to 0, \tag{20}$$

as $n \to +\infty$, because, as $x \to +\infty$,

$$\begin{aligned} &x \log (1 - (1 - F_{|X|}(x))) \\ &\qquad = -x(1 - F_{|X|}(x)) + xO((1 - F_{|X|}(x))^2) \to 0, \end{aligned} \tag{21}$$

since $E|X| < +\infty$. From $\bar{X}^i = (n/(n-1))(\bar{X} - (X_i/n))$ it therefore follows that, as $n \to +\infty$,

$$P\{\bar{X}, \bar{X}^1, \cdots, \bar{X}^n \varepsilon I \text{ simultaneously}\} \to 1. \tag{22}$$

For a double sequence of events $\{A_n\}$ and $\{E_n\}$ in which $P\{E_n\} \to 1$ it follows that

$$\lim P\{A_n\} = \lim [P\{A_nE_n\} + P\{A_nE_n^c\}] = \lim P\{A_nE_n\}, \tag{23}$$

so that the imposition or removal of the condition E_n has no effect on the limit-

ing probabilities. In the probability argument to follow it will be convenient to assume $\bar{X}, \bar{X}^1, \cdots, \bar{X}^n \varepsilon I$ in some statements and to ignore it in others and by virtue of (22) and (23) the limit probabilities will be unaffected by this.

For $\bar{X}, \bar{X}^1, \cdots, \bar{X}^n \varepsilon I$,

$$f(\bar{X}^i) = f(\bar{X}) + (\bar{X}^i - \bar{X})f'(\bar{X}) + [(\bar{X}^i - \bar{X})^2/2]f''(\zeta_i), \tag{24}$$

where ζ_i lies between $\bar{X}^i$ and $\bar{X}$. From (19) and (24)

$$\hat{\theta} = f(\bar{X}) - [(n-1)/n]f'(\bar{X})\sum_1^n (\bar{X}^i - \bar{X}) - [(n-1)/n]\tfrac{1}{2}\sum_1^n (\bar{X}^i - \bar{X})^2 f''(\zeta_i), \tag{25}$$

where the middle term is actually zero since $\bar{X}^i - \bar{X} = -(X_i - \bar{X})/(n-1)$. Expression (25) can thus be rewritten as

$$n^{\frac{1}{2}}(\hat{\theta} - \theta) = n^{\frac{1}{2}}(f(\bar{X}) - f(0)) - 1/[2(n^{\frac{1}{2}})(n-1)]\sum_1^n (X_i - \bar{X})^2 f''(\zeta_i). \tag{26}$$

For $\bar{X} \varepsilon I$, $f(\bar{X}) = f(0) + \bar{X}f'(\xi_{\bar{x}})$, where $|\xi_{\bar{x}}| < |\bar{X}|$. Asymptotically $n^{\frac{1}{2}}\bar{X}$ has the distribution $N(0, \sigma^2)$, and $f'(\xi_{\bar{x}}) \to_p f'(0)$ (with $\xi_{\bar{x}}$ defined arbitrarily when $\bar{X} \varepsilon I$). Hence, by Slutsky's theorem the first term on the right in (26) is asymptotically $N(0, \sigma^2(f'(0))^2)$ with or without the condition $\bar{X}, \bar{X}^1, \cdots, \bar{X}^n \varepsilon I$.

It remains to be shown that the second term $\to_p 0$. For $\bar{X}, \bar{X}^1, \cdots, \bar{X}^n \varepsilon I$, $|f''(\zeta_i)| < M$, $i = 1, \cdots, n$, for some $0 < M < +\infty$. Since $\sum_1^n (X_i - \bar{X})^2/(n-1) \to_p \sigma^2$, the extra $n^{\frac{1}{2}}$ in the denominator of the second term makes it $\to_p 0$ with or without the condition $\bar{X}, \bar{X}^1, \cdots, \bar{X}^n \varepsilon I$ (where ζ_i is chosen arbitrarily if $\bar{X}, \bar{X}^i \varepsilon I$). ||

For the unjackknifed estimate $f(\bar{x})$ all that is required for asymptotic normality is a continuous first derivative near 0, which is weaker than what is required here, namely, a bounded second derivative. The world abounds with functions that have continuous first derivatives at zero but non-existent or unbounded second derivatives so presumably, unless the theorem can be strengthened, there exists an example in which $f(\bar{x})$ is asymptotically normal but the jackknifed $\hat{\theta}$ is not. Unfortunately, this author has been unsuccessful in his attempts to find a function for which it could be proved the convergence breaks down.

From a practical point of view the restriction to f with bounded second derivative near the X mean is a mere bagatelle. It is essentially always satisfied. In the three specific examples listed the condition is fulfilled provided the degenerate end-points are omitted, i.e., $p = 0, 1$ in the binomial, $\lambda = 0$ in the Poisson, and $\sigma^2 = 0$ in the χ^2.

The asymptotic variance of the statistic $n^{\frac{1}{2}}(\hat{\theta} - \theta)$ is $\sigma^2(f'(\mu))^2$. In some cases, such as when stabilizing the variance, this is a known constant. In others where it is not known, it can be consistently estimated by $\hat{\sigma}^2(f'(\bar{X}))^2$ where $\hat{\sigma}^2$ is a consistent estimate of σ^2 based directly on $(X_1, \cdots, X_n)$ such as s^2, $\hat{p}(1 - \hat{p})$

in the binomial, or $\bar{X}$ in the Poisson. It can also be consistently estimated from

$$(27)\quad s_f^2 = (n-1)^{-1}\sum_1^n (\hat{\theta}_i - \hat{\theta})^2 = (n-1)\sum_1^n (f(\bar{X}^i) - n^{-1}\sum_1^n f(\bar{X}^j))^2,$$

as suggested by the Tukey conjecture. The only condition required on f is the weaker one of a continuous first derivative at the mean.

THEOREM 2. *Let $\{X_i\}$ be a sequence of independent, identically distributed random variables with $\mu = 0$, $0 < \sigma^2 < +\infty$. Let f be a function with a continuous first derivative near 0. Then, as $n \to +\infty$, $s_f^2 \to_p \sigma^2(f'(0))^2$.*

PROOF. Let $I = (-\Delta, +\Delta)$ be any neighborhood of zero in which f' exists and is continuous. By the argument in Theorem 1 $P\{\bar{X}, \bar{X}^1, \cdots, \bar{X}^n \varepsilon I\} \to 1$ as $n \to +\infty$.

For $\bar{X}, \bar{X}^1, \cdots, \bar{X}^n \varepsilon I$, $f(\bar{X}^i) = f(\bar{X}) + (\bar{X}^i - \bar{X})f'(\zeta_i)$, with $\bar{X} \leqq \zeta_i \leqq \bar{X}^i$ or $\bar{X}^i \leqq \zeta_i \leqq \bar{X}$ and $(\bar{X}^i - \bar{X}) = -(X_i - \bar{X})/(n-1)$, so

$$\begin{aligned} s_f^2 &= (n-1)^{-1}\sum_1^n ((X_i - \bar{X})f'(\zeta_i) - n^{-1}\sum_1^n (X_j - \bar{X})f'(\zeta_j))^2 \\ &= (n-1)^{-1}\sum_1^n ((X_i - \bar{X})f'(0) + (X_i - \bar{X})(f'(\zeta_i) - f'(0)) \\ &\quad - n^{-1}\sum_1^n (X_j - \bar{X})(f'(\zeta_j) - f'(0)))^2 \\ &= (f'(0))^2(n-1)^{-1}\sum_1^n (X_i - \bar{X})^2 \\ &\quad + (n-1)^{-1}\sum_1^n ((X_i - \bar{X})(f'(\zeta_i) - f'(0)) \\ &\quad - n^{-1}\sum_1^n (X_j - \bar{X})(f'(\zeta_j) - f'(0)))^2 + X\text{-product term.} \end{aligned} \tag{28}$$

The quantity $(f'(0))^2\sum_1^n (X_i - \bar{X})^2/(n-1) \to_p \sigma^2(f'(0))^2$ with or without the condition $\bar{X}, \bar{X}^1, \cdots, \bar{X}^n \varepsilon I$. Let $g(x) = f'(x) - f'(0)$. If it can be shown that

$$(29)\qquad (n-1)^{-1}\sum_1^n (X_i - \bar{X})^2(g'(\zeta_i))^2 \to_p 0,$$

then the second term in the final expression in (28) $\to_p 0$, and the X-product term does as well by the Cauchy-Schwarz inequality. But for $\epsilon > 0$ there exists a $\Delta_\epsilon > 0$ such that $|g'(x)| < \epsilon$ for $x \varepsilon I_\epsilon = (-\Delta_\epsilon, +\Delta_\epsilon)$. For $\bar{X}, \bar{X}^1, \cdots, \bar{X}^n \varepsilon I_\epsilon$,

$$(30)\quad (n-1)^{-1}\sum_1^n (X_i - \bar{X})^2(g'(\zeta_i))^2 \leqq \epsilon^2(n-1)^{-1}\sum_1^n (X_i - \bar{X})^2 \to_p \epsilon^2\sigma^2,$$

so, since ϵ is arbitrary, the left hand side must $\to_p 0$. ||

No attempt has been made to investigate whether it would be better to estimate $\sigma^2(f'(\mu))^2$ by s_f^2 or by $\hat{\sigma}^2(f'(\bar{X}))^2$ since this paper is concerned with asymptotics. For small sample problems it would be worth investigating.

The classical method of eliminating the $1/n$ bias term from transformations is to use the modified estimator

$$\hat{\theta}_b = f(\bar{X}) - [\hat{\sigma}^2/(2n)]f''(\bar{X}), \tag{31}$$

and would constitute an alternative to the jackknife procedure. This author has not obtained any general results on which of the two estimators, $\hat{\theta}$ and $\hat{\theta}_b$, is better in some sense. For each of the commoner transformations a study into which estimator had a better mean squared error or better distributional properties would be worthwhile.

4. Preservation of normality. Suppose that normality is already present in the sample estimates; i.e., suppose $(\hat{\theta}^1_{n-1}, \cdots, \hat{\theta}^n_{n-1}, \hat{\theta}^0_n)$ has a multivariate normal distribution. Is it automatically true that the studentized jackknife statistic has a t-distribution and a limiting normal distribution? The answer is no in general. The statistic is a t-statistic except for a multiplicative constant, but this constant can tend to 0 or $+\infty$ or anything in between.

THEOREM 3. *Let* $(\hat{\theta}^1_{n-1}, \cdots, \hat{\theta}^n_{n-1}, \hat{\theta}^0_n)$ *have a multivariate normal distribution with means* θ, *variances* $\sigma^2_{n-1}, \cdots, \sigma^2_{n-1}, \sigma^2_n$, *respectively, and correlation matrix*

$$\begin{pmatrix} 1 & \rho_{n-1} & \cdot & \cdot & \rho_{n-1} & \tau_n \\ \rho_{n-1} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \rho_{n-1} & \cdot \\ \rho_{n-1} & \cdot & \cdot & \rho_{n-1} & 1 & \tau_n \\ \tau_n & \cdot & \cdot & \cdot & \tau_n & 1 \end{pmatrix}. \tag{32}$$

Then

$$\frac{n^{\frac{1}{2}}(\hat{\theta} - \theta)}{\left[\sum_1^n (\hat{\theta}_i - \hat{\theta})^2/(n-1)\right]^{\frac{1}{2}}} \sim t_{n-1}\left[\frac{nV_n}{(n-1)^2(1-\rho_{n-1})\,\sigma^2_{n-1}}\right]^{\frac{1}{2}}, \tag{33}$$

where $\sim$ *denotes "is distributed as",* t_{n-1} *is a generic t-variable with* $n - 1$ d.f., *and*

$$\begin{aligned} V_n = n^2\sigma_n^2 - 2n(n-1)\tau_n\sigma_n\sigma_{n-1} + (n-1)^2\rho_{n-1}\sigma^2_{n-1} \\ + (n-1)^2\sigma^2_{n-1}(1-\rho_{n-1})/n. \end{aligned} \tag{34}$$

PROOF. From its definition and straightforward covariance computations $\hat{\theta} \sim N(\theta, V_n)$.

Under the transformation $W_i = \hat{\theta}^i_{n-1} - c\hat{\theta}_{n-1}$, $i = 1, \cdots, n$, where c is chosen so that Cov $(W_i, W_j) = 0$, $W_i \sim N(\theta(1-c), \sigma^2_{n-1}(1-\rho_{n-1}))$ and $W_1, \cdots, W_n$ are independent, so

$$\begin{aligned} (n-1)^{-1}\sum_1^n (\hat{\theta}_i - \hat{\theta})^2 &= (n-1)\sum_1^n (\hat{\theta}^i_{n-1} - \hat{\theta}_{n-1})^2 \\ &= (n-1)\sum_1^n (W_i - \bar{W})^2 \\ &\sim (n-1)\sigma^2_{n-1}(1-\rho_{n-1})\chi^2_{n-1}, \end{aligned} \tag{35}$$

where χ^2_{n-1} is a generic χ^2-variable with $n-1$ d.f. Furthermore, $\hat{\theta}$ and $\sum_1^n (\hat{\theta}_i - \hat{\theta})^2$ are independently distributed since Cov $(n\hat{\theta}^0_n - (n-1)\hat{\theta}^{\cdot}_{n-1}, \hat{\theta}^{i}_{n-1} - \hat{\theta}^{\cdot}_{n-1}) = 0$. Hence,

$$\frac{n^{\frac{1}{2}}(\hat{\theta} - \theta)}{\left[\sum_1^n (\hat{\theta}_i - \hat{\theta})^2/(n-1)\right]^{\frac{1}{2}}} = \frac{(\hat{\theta} - \theta)}{(V_n)^{\frac{1}{2}}}\left(\frac{(n-1)}{\chi^2_{n-1}}\right)^{\frac{1}{2}}\left(\frac{nV_n}{(n-1)^2\,\sigma^2_{n-1}(1-\rho_{n-1})}\right)^{\frac{1}{2}} \tag{36}$$

$$= t_{n-1}\left(\frac{nV_n}{(n-1)^2\,\sigma^2_{n-1}(1-\rho_{n-1})}\right)^{\frac{1}{2}}. \|$$

For confidence intervals and tests of significance based on $n^{\frac{1}{2}}(\hat{\theta} - \theta)/(\sum (\hat{\theta}_i - \hat{\theta})^2/(n-1))^{\frac{1}{2}} \sim t_{n-1}$ to be approximately correct, the constant $nV_n/(n-1)^2\sigma^2_{n-1}(1-\rho_{n-1})$ must be approximately one. This may be the case, but it is not necessarily so.

In view of the practical context from which the estimators $(\hat{\theta}^1_{n-1}, \cdots, \hat{\theta}^n_{n-1}, \hat{\theta}^0_n)$ were supposedly generated, certain natural restrictions can be put on the var-covariance parameters. These are:

$$\sigma_n \downarrow 0, \qquad \rho_n \uparrow 1, \qquad \tau_n \uparrow 1, \quad \text{and} \quad \rho_{n-1} < \tau_n, \tag{37}$$

where the monotonicity is strict. Still not every matrix of the form (32) with restrictions (37) is a correlation matrix as it must also be positive semi-definite. It can be verified that a necessary and sufficient condition for positive semi-definiteness is

$$\tau_n^2 \leqq \rho_{n-1} + n^{-1}(1 - \rho_{n-1}), \tag{38}$$

so this condition will also be imposed on the parameters ρ_n and τ_n.

The multiplicative constant $nV_n/(n-1)^2\sigma^2_{n-1}(1-\rho_{n-1})$ can achieve any value in the interval $[0, +\infty)$ by varying the parameters under the restrictions (37) and (38). By choosing the parameters so that

$$\tau_n^2 = \rho_{n-1} + n^{-1}(1 - \rho_{n-1}) = \left(\frac{n\sigma_n}{(n-1)\,\sigma_{n-1}}\right)^2, \tag{39}$$

the variance V_n has the value zero. Also, rearrangement and collection of terms gives

$$\frac{nV_n}{(n-1)^2\,\sigma^2_{n-1}(1-\rho_{n-1})}$$

$$= 1 + n\left\{\frac{1}{1-\rho_{n-1}}\left(\frac{n\sigma_n}{(n-1)\sigma_{n-1}} - 1\right)\left(\frac{n\sigma_n}{(n-1)\sigma_{n-1}} - \rho_{n-1}\right)\right. \tag{40}$$

$$\left. + \left(\frac{n\sigma_n}{(n-1)\sigma_{n-1}}\right)\left[2\left(\frac{1-\tau_n}{1-\rho_{n-1}}\right) - 1\right]\right\},$$

so it is clear that for $1 - \rho_{n-1}$ very small and the other parameters chosen suitably the ratio can be made arbitrarily large.

For the ratio in (40) to be asymptotically equal to 1 the quantity in braces

must be $o(1/n)$. This cannot happen if, for instance, $\sigma_n/\sigma_{n-1} \to \alpha$, $0 \leqq \alpha < 1$, or the first term is $o(1/n)$ but $(1 - \tau_n)/(1 - \rho_{n-1}) \to \beta$, $\frac{1}{2} < \beta \leqq 1$. In these two instances the ratio would diverge to $+\infty$.

Loosely speaking, one should have $\sigma_n/\sigma_{n-1} \cong 1$ and $(1 - \tau_n)/(1 - \rho_{n-1}) \cong \frac{1}{2}$ for the multiplicative constant to be $\cong 1$. This is the behavior exhibited by the linear case where $\hat{\theta}_n^0 = (\sum_1^n X_i)/n$:

$$(41) \quad \sigma_n^2 = \sigma^2/n, \quad \tau_n = (1 - n^{-1})^{\frac{1}{2}} \sim 1 - (2n)^{-1}, \quad \text{and} \quad \rho_{n-1} = 1 - (n - 1)^{-1}.$$

Trivially $\hat{\theta}$ has a t-distribution for the linear case since $\hat{\theta}_i = X_i$, or it can be checked by the theorem.

It is interesting to note in closing that depending on the relative sizes of the parameters, the variance of $\hat{\theta}$ may exceed, equal, or be smaller than the variance of $\hat{\theta}_n^0$. Thus, jackknifing could reduce or increase the variance depending on the parameters.

5. Discussion. As Tukey has pointed out, the studentized jackknife has approximately a t-distribution (exactly under normal theory) in the linear case $\hat{\theta}_n^0 = (\sum_1^n X_i)/n$ since $\hat{\theta}_i = X_i$. Intuitively, this should extend to estimators which are locally linear in the observations so that in a power series expansion the linear term, which behaves nicely under jackknifing, would play the dominant role.

To get counterexamples the author chose one of the simplest non-linear estimators he could think of—the maximum of $X_1, \cdots, X_n$. There would be many other possible choices, but this was sufficient to produce the three different types of asymptotic behavior one might expect—a non-normal distribution, degeneracy at a point, and drift to infinity.

Trustworthiness for the jackknife was established in two situations where the estimators had a linear quality to them. The first was where the estimator was a twice-differentiable function of the sample mean. A power series argument, as suggested above, established the limiting normality of the jackknife statistic. In the second case the estimators had a linear quality through the behavior of their variances and correlation coefficients. The theorem was stated for normal variables, but presumably if this were approximate, the result would also be approximate.

Both of the above situations were ones in which the unjackknifed estimator had a proper finite or limiting distribution under weaker conditions than required for the jackknife. The hope would be the reverse, namely, that the jackknife would create an approximate t or normal distribution where none existed for the unjackknifed statistic. The lack of theorems in this area is directly related to a lack of suitable theorems on exchangeable or interchangeable random variables. The random variables $(\hat{\theta}_1, \cdots, \hat{\theta}_n)$ are interchangeable, but they do not sum to a constant [3] nor are they a portion of an infinite sequence of interchangeable random variables [1] so the asymptotic normality in these contexts is of no help.

The reader's attention is also directed to two forthcoming papers on the jackknife by Brillinger [2] and Robson and Whitlock [8].

6. Acknowledgment. The author would like to thank Professor G. S. Watson for some useful discussions on the jackknife.

REFERENCES

[1] BLUM, J. R., CHERNOFF, H., ROSENBLATT, M., and TEICHER, H. (1958). Central limit theorems for interchangeable processes. *Canad. J. Math.* **10** 222–229.

[2] BRILLINGER, D. R. (1964). *Rev. Inst. Internat. Statist.*

[3] CHERNOFF, H. and TEICHER, H. (1958). A central limit theorem for sums of interchangeable random variables. *Ann. Math. Statist.* **29** 118–130.

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

[5] KARLIN, S. (1958). Admissibility for estimation with quadratic loss. *Ann. Math. Statist.* **29** 406–436.

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

[7] QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360.

[8] ROBSON, D. S., and WHITLOCK, J. H. (1964). Estimation of a truncation point. *Biometrika.* **51** 33–39.

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