The Annals of Statistics
1990, Vol. 18, No. 3, 1438–1452

ON THE ASYMPTOTIC PROPERTIES OF THE JACKKNIFE HISTOGRAM[1]

BY C. F. J. WU

University of Waterloo

We study the asymptotic normality of the jackknife histogram. For one sample mean, it holds if and only if r, the number of observations retained, and d $(= n - r)$, the number of observations deleted, both diverge to infinity. The best convergence rate $n^{-1/2}$ is obtained when $r = O(n)$ and $d = O(n)$. For U statistics of degree 2 and nonlinear statistics admitting the expansion (3.1), similar results are obtained under conditions on r and d. A second order approximation based on the Edgeworth expansion is discussed briefly.

1. Introduction. The Quenouille–Tukey jackknife based on deleting one observation each time is known to be effective for bias reduction and variance estimation in many situations. However, it does not provide enough information for investigating the distribution of the estimator $\hat{\theta}$. Such an investigation is necessary if interval estimation or inference on some functional of the distribution of $\hat{\theta}$ is desired. It also fails to give a consistent variance estimator for nonsmooth $\hat{\theta}$ such as the sample quantiles. Wu (1986) suggested that both problems could be resolved by relaxing the delete-one recipe to allow a larger number of observations to be deleted each time. For the second problem Shao and Wu (1989) showed that the consistency of jackknife variance estimation is restored by using a delete-d jackknife with d depending on a smoothness measure of $\hat{\theta}$. As $\hat{\theta}$ becomes less smooth (such as the sample median), d increases. In this paper we address the first problem of the jackknife, i.e., we show that the normalized histogram of the delete-d jackknife with properly chosen d converges to the distribution of $\hat{\theta}$. Results along this line are given for three classes of statistics. See Theorems 1 to 3 and the summary at the end of the section.

Another resampling method, the bootstrap (Efron, 1979), does not have the same problems as the delete-one jackknife. Several methods based on the bootstrap histogram are known to possess some desirable asymptotic properties for estimating the distribution of $\hat{\theta}$ for a variety of statistics [Bickel and Freedman (1981); Singh (1981); Beran (1987); Efron (1987); Hall (1986)]. For the sample quantiles, the bootstrap variance estimator is consistent if the underlying distribution has a finite αth moment, $\alpha > 0$ [Ghosh, Parr, Singh and Babu (1984)]. Because of these two problems, the jackknife is perceived to be less versatile than the bootstrap. Given the long history of the jackknife and

Received September 1987; revised June 1989.

[1]Research supported by NSF-AFORS grants DMS-85-02303 and ISSSA-860068 and the Natural Sciences and Engineering Research Council of Canada.

AMS 1980 *subject classification*. 62G05.

Key words and phrases. Jackknife histogram, bootstrap, asymptotic normality, Edgeworth expansion, simple random sampling without replacement.

its frequent use in practice, attempts such as ours to improve its utility should be worthwhile.

Since the jackknife employs simple random sampling without replacement for drawing resamples (see Section 2), the jackknife histogram does not in general approximate the distribution of $\hat{\theta}$ to the second order term for independent and identically distributed samples. This will be studied in Section 4. In spite of this, there are situations in which the jackknife is preferred. This and the potential values of the present asymptotic study will be discussed in Section 5.

Several definitions are required for the general jackknife. Let $\mathbf{x} = (x_1, \dots, x_n)$ be independent and identically distributed (i.i.d.) with finite variance and $\hat{\theta} = \hat{\theta}(\mathbf{x})$ be an estimator of an unknown quantity θ. We assume that $\hat{\theta}$ is asymptotically normal with limiting variance σ^2, i.e.,

$$\sqrt{n}\,(\hat{\theta} - \theta) \to_d N(0, \sigma^2). \tag{1.1}$$

The jackknife method resamples from $\mathbf{x}$ by taking each subset $\mathbf{x}_s$ (of size r) of $\mathbf{x}$ with equal probability $\binom{n}{r}^{-1}$. Taking a subset of size r is equivalent to deleting its complement of size $d = n - r$. Denote this jackknife sampling by $*$. Notations such as E_*, v_* refer to probability calculations under $*$. For each selected subset $\mathbf{x}_s$, we calculate $\hat{\theta}_s = \hat{\theta}(\mathbf{x}_s)$ and define

$$\tilde{\theta}_s = \hat{\theta} + \sqrt{\frac{r}{d}}\,(\hat{\theta}_s - \hat{\theta}). \tag{1.2}$$

For variance estimation,

$$E_*(\tilde{\theta}_s - \hat{\theta})^2 = \frac{r}{d} E_*(\hat{\theta}_s - \hat{\theta})^2 \tag{1.3}$$

is called a delete-d jackknife variance estimator [Shao and Wu (1989)].

The definition (1.2) is motivated by moment-matching for the one-sample mean $\hat{\theta} = \bar{x} = n^{-1}\Sigma_1^n x_i$. It is easy to see that, for $\hat{\theta} = \bar{x}$,

$$E_*\tilde{\theta}_s = \hat{\theta} = \bar{x},$$

$$E_*(\tilde{\theta}_s - \hat{\theta})^2 = \frac{1}{n(n-1)} \sum_1^n (x_i - \bar{x})^2.$$

Note that for the delete-one jackknife ($r = n - 1$), Tukey's (1958) pseudovalues

$$\tilde{\theta}_i = \hat{\theta} - (n-1)(\hat{\theta}_s - \hat{\theta}), \qquad s = \{1, \dots, n\}/\{i\},$$

are different from (1.2).

We can construct a *jackknife histogram* from the normalized $\tilde{\theta}_s$ values, that is, its cumulative distribution is

$$J(t) = P_*\left\{\frac{\tilde{\theta}_s - \hat{\theta}}{\hat{v}^{1/2}} \le t\right\} = P_*\left\{\left(\frac{nr}{d}\right)^{1/2} \frac{\hat{\theta}_s - \hat{\theta}}{\hat{\sigma}} \le t\right\}, \tag{1.4}$$

where $\hat{v} = n^{-1}\hat{\sigma}^2$ and $\hat{\sigma}^2$ is a *consistent* estimate of the limiting variance σ^2 in (1.1). The estimate $\hat{v}$ can be obtained by several methods, e.g., linearization, jackknife and bootstrap. Consistency of the jackknife and the bootstrap variance estimators are studied, respectively, in Shao and Wu (1989) and Shao (1987). A major purpose of resampling is to use the observed data to construct a distribution that mimics the unknown distribution of $\hat{\theta}$. Since the limiting distribution of $\hat{\theta}$ is normal, a central question is whether $J(t)$, (1.4), will converge to $N(0,1)$. For the one-sample mean, $r \to \infty$ and $d \to \infty$ is shown to be necessary and sufficient for the asymptotic normality of $J(t)$ [Theorem 1(i) and (ii)]. For U statistics of degree 2, it is sufficient for the asymptotic normality of $J(t)$ [Theorem 3(i)]. Convergence rate of $J(t)$ to normality is studied in Theorems 1(iii) and 3(ii). For the one-sample problem, the best convergence rate $n^{-1/2}$ is obtained when $r = O(n)$ and $d = O(n)$, that is to delete a fraction of n observations. Results on asymptotic normality are obtained in Theorem 2 for nonlinear statistics admitting the expansion (3.1).

2. Asymptotic normality of the jackknife histogram: Linear statistics. We first study the problem for linear statistics. Let $x_1, \dots, x_n$ be i.i.d. with mean μ, finite variance σ^2 and distribution $F((x-\mu)/\sigma)$. For $\hat{\theta} = \bar{x}$, $J(t)$, (1.4), becomes

$$P_*\left\{\left(\frac{nr}{d}\right)^{1/2}\frac{\bar{x}_r - \bar{x}}{\hat{\sigma}} \le t\right\}, \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2, \tag{2.1}$$

where $\bar{x}_r$ is the mean of the x_i's in the subset s. We note that the jackknife sampling $*$ is the same as simple random sampling (srs) without replacement from the population $\{x_1, \dots, x_n\}$. Therefore, in this context, $\bar{x}$ is the population mean and $\bar{x}_r$, is the mean of an srs sample without replacement. We can rewrite (2.1) as

$$P_*\left\{\frac{\sqrt{r}\,(\bar{x}_r - \bar{x})}{[(1-f)\hat{\sigma}^2]^{1/2}} \le t\right\}, \tag{2.2}$$

where $f = r/n$ is the sampling fraction and $(1-f)\hat{\sigma}^2/r$ is the variance of $\bar{x}_r$ under srs without replacement [Cochran (1977)]. This connection with finite population sampling enables us to obtain simpler proofs.

From the central limit theorem,

$$P_F\left\{\frac{\sqrt{n}\,(\bar{x} - \mu)}{\sigma} \le t\right\} \to \Phi(t) \quad \text{for each } t, \tag{2.3}$$

where $\Phi(t)$ is the standard normal distribution. Therefore a requirement of $J(t)$, (2.1) or (2.2), is its convergence to the same limit. In Theorem 1(ii) and (iii), a strong version will be proved, that is, $J(t)$ as a function of $\mathbf{x}$ converges to $\Phi(t)$ a.s. in $\mathbf{x}$.

From Theorem 1(iii) the best convergence rate $n^{-1/2}$ is obtained when $r = O(n)$ and $d = O(n)$.

THEOREM 1. (i) *If either d or r is bounded, $J(t) \to \Phi(t)$ for each t a.s. does not hold except for normal F.*

(ii) *Under $0 < E_F(X-\mu)^2 < \infty$, $r \to \infty$ and $d \to \infty$ imply*

$$\sup_t |J(t) - \Phi(t)| \to 0 \quad a.s. \tag{2.4}$$

(iii) *Under $E_F|X-\mu|^3 < \infty$ and $\sigma > 0$, the upper bound in (2.5) converges to zero a.s. at the rate $\max(r^{-1/2}, d^{-1/2})$ as $r \to \infty$ and $d \to \infty$,*

$$\sup_t |J(t) - \Phi(t)| \le \frac{C}{\min(\sqrt{r}, \sqrt{d})} \frac{n^{-1}\sum_1^n |x_i - \bar{x}|^3}{\hat{\sigma}^3}, \tag{2.5}$$

where C is a constant independent of r and n.

PROOF. (i) Let $\psi_n(\xi)$ be the characteristic function of

$$\left(\frac{nr}{d}\right)^{1/2} \frac{(\bar{x}_r - \bar{x})}{\hat{\sigma}}$$

under $*$, that is,

$$\begin{aligned} \psi_n(\xi) &= \frac{1}{\binom{n}{r}} \sum_r \exp\left\{ i\xi\sqrt{\frac{nr}{d}} \frac{(\bar{x}_r - \bar{x})}{\hat{\sigma}} \right\} \\ &= \frac{1}{\binom{n}{r}} \sum_r \prod_{j=1}^r \exp\left\{ i\xi\sqrt{\frac{n}{rd}} \frac{1}{\hat{\sigma}}(x_{k_j} - \bar{x}) \right\}, \end{aligned} \tag{2.6}$$

where $\sum_r$ denotes summation over all the subsets $\{k_1, \dots, k_r\}$ of size r. Using $r(\bar{x}_r - \bar{x}) = -d(\bar{x}_d - \bar{x})$, where $\bar{x}_d$ is the mean of the x_i's in the complement $\bar{s}$ of s, $\psi_n(\xi)$ can be written as

$$\frac{1}{\binom{n}{d}} \sum_d \prod_{j=1}^d \exp\left\{ -i\xi\left(\frac{n}{rd}\right)^{1/2} \frac{1}{\hat{\sigma}}(x_{k_j} - \bar{x}) \right\}. \tag{2.7}$$

Consider the case of bounded d. We have $n/r \to 1$, $\bar{x} \to \mu$, $\hat{\sigma} \to \sigma$ and (2.7) is asymptotically equivalent to

$$\frac{1}{\binom{n}{d}} \sum_d \prod_{j=1}^d \exp\left\{ -i\xi \frac{1}{d^{1/2}} \frac{1}{\sigma}(x_{k_j} - \mu) \right\}, \tag{2.8}$$

which is a U statistic of degree d with kernel

$$h(x_1, \dots, x_d) = \prod_{j=1}^d \exp\left\{ -i\xi \frac{1}{d^{1/2}} \frac{1}{\sigma}(x_j - \mu) \right\}.$$

Since $E|h| < \infty$, from the strong law of large numbers of U statistics [Serfling (1980)], (2.8) converges a.s. to

$$Eh = \left\{\phi\left(-\frac{\xi}{\sqrt{d}}\right)\right\}^d, \tag{2.9}$$

where $\phi(n)$ is the characteristic function of $(X - \mu)/\sigma$ under F. Except when F is normal, (2.9) cannot be the characteristic function of a normal distribution. Therefore, convergence to a normal limit holds only for normal F. Similarly for bounded r, it can be shown that (2.6) converges a.s. to $\{\phi(\xi/\sqrt{r})\}^r$ by following the same proof.

(ii) The expression (2.2) for $J(t)$ is for the mean of an srs sample (without replacement) from the population $\{x_1, \ldots, x_n\}$. Hájek's (1960) result on the necessary and sufficient condition for the asymptotic normality of $\bar{x}_r$ (in conjunction with Polya's theorem) establishes (2.4) if $\{x_1, \ldots, x_n\}$ satisfy the condition

$$\lim_{n\to\infty} \frac{1}{(n-1)\hat{\sigma}^2} \sum_1^n (x_i - \bar{x})^2 I_{(|x_i - \bar{x}| \geq \tau(rd/n)^{1/2}\hat{\sigma})} = 0 \quad \text{a.s.} \tag{2.10}$$

for any $\tau > 0$. Since $E_F|X - \mu|^2 < \infty$, $\bar{x} \to \mu$ and $\hat{\sigma} \to \sigma$ a.s., (2.10) is equivalent to

$$\lim_{n\to\infty} \frac{1}{n} \sum_1^n (x_i - \mu)^2 I_{(|x_i - \mu| \geq \tau(rd/n)^{1/2})} = 0 \quad \text{a.s. for any } \tau > 0. \tag{2.11}$$

From the inequality

$$\min(r, d) \geq \frac{rd}{n} \geq \frac{1}{2}\min(r, d), \tag{2.12}$$

$\min(r, d) \to \infty$ iff $rd/n \to \infty$ as $n \to \infty$. For any positive constant k, choose a large m such that $rd/n > k^2$ for $n \geq m$. Then the left-hand side of (2.11) is bounded above by

$$\lim_{n\to\infty} \frac{1}{n} \sum_1^n (x_i - \mu)^2 I_{(|x_i - \mu| \geq \tau k)} = E_F(X - \mu)^2 I_{(|X - \mu| \geq \tau k)},$$

which is arbitrarily small since k can be arbitrarily large and $E_F(X - \mu)^2 < \infty$.

(iii) Using the following Berry–Esseen bound for srs without replacement [Höglund (1978)],

$$\sup_t |J(t) - \Phi(t)| \leq \frac{C}{(rd/n)^{1/2}} \frac{n^{-1}\sum_1^n |x_i - \bar{x}|^3}{\hat{\sigma}^3}, \tag{2.13}$$

where C is a constant independent of $\{x_1, \ldots, x_n\}$, r and n. Since $E_F|X - \mu|^3 < \infty$, $n^{-1}\sum_1^n |x_i - \bar{x}|^3 \to E_F|X - \mu|^3$ and $\hat{\sigma} \to \sigma$. From (2.12) and (2.13), (2.5) and the conclusion in (iii) follow.

3. Asymptotic normality of the jackknife histogram: Nonlinear statistics. In this section we extend Theorem 1 to nonlinear statistics. First we consider the class of estimators $\hat{\theta}(\mathbf{x})$ which admit the expansion

$$\hat{\theta} = \theta + \frac{1}{n}\sum_{1}^{n} \phi_F(x_i) + R_n, \tag{3.1}$$

where ϕ_F is measurable in x with $E_F(\phi_F(X)) = 0$, $0 < E_F(\phi_F^2(X)) = \sigma^2 < \infty$ and the remainder term R_n satisfies

$$\sqrt{n}\, R_n \to 0 \quad \text{in probability}. \tag{3.2}$$

From (3.2), $\sigma^2 = E_F(\phi_F^2(X))$ is the limiting variance of $\sqrt{n}\,(\hat{\theta} - \theta)$. Any consistent estimator $\hat{\sigma}$ in (1.4) can be used for estimating this σ.

Conditions (3.1) and (3.2) give a broad class of estimators $\hat{\theta}$. Noting that the linear term $n^{-1}\Sigma\, \phi_F(x_i)$ in (3.1) is of the order $O_p(n^{-1/2})$, (3.1) and (3.2) imply that the estimator $\hat{\theta}$ can be approximated by a linear statistic with the approximation error being of lower order. This very general property covers many smooth and not-so-smooth estimators, including the sample quantiles, L statistics, M estimators, smooth functions of the sample mean and U statistics [detail in Serfling (1980) and Shao and Wu (1989)]. If $\hat{\theta} = T(F_n)$ is a functional of the empirical distribution function F_n of $(x_1, x_2, \dots, x_n)$, a sufficient condition for (3.1) and (3.2) is that T is quasi-Fréchet differentiable with respect to a norm $\|\cdot\|$ for which $\|F_n - F\| = O_p(n^{-1/2})$ [Serfling (1980), page 221].

Similarly for any subset of size r, $\hat{\theta}_s$ has the expansion

$$\hat{\theta}_s = \theta + \frac{1}{r}\sum_{i \in s} \phi_F(x_i) + R_{n,s},$$

where the remainder $R_{n,s} = o_p(r^{-1/2})$. To study the limiting behavior of $J(t)$, (1.4), we decompose $\hat{\theta}_s - \hat{\theta}$ into the linear part and the remainder part, i.e.,

$$\hat{\theta}_s - \hat{\theta} = \left[\frac{1}{r}\sum_{i \in s} \phi_F(x_i) - \frac{1}{n}\sum_{1}^{n} \phi_F(x_i)\right] + [R_{n,s} - R_n].$$

The limiting behavior of $J(t)$ for the linear part was considered in Theorem 1. In Lemma 1 we give conditions under which the contribution of the remainder part of $J(t)$ is asymptotically negligible.

LEMMA 1. (i) *Under* (3.2), $d/n \geq \lambda$ *for some* $\lambda > 0$ *and* $r \to \infty$ *ensure that*

$$Q(\mathbf{x}) = P_*\left\{\left(\frac{nr}{d}\right)^{1/2} |R_{n,s} - R_n| \geq \varepsilon\right\} \to 0 \quad \textit{in probability}. \tag{3.3}$$

(ii) *Under*

$$ER_n^2 = \frac{a}{n^2} + o(n^{-2}), \qquad E(R_r R_n) = \frac{a}{rn} + o(r^{-2}), \tag{3.4}$$

where a is independent of n, $d/n \to 0$ *and* $r \to \infty$ *ensure that* (3.3) *holds.*

PROOF. (i) Since $Q(\mathbf{x})$ is nonnegative, (3.3) follows from

$$E_F\{Q(\mathbf{x})\} = P_F\left\{\left(\frac{nr}{d}\right)^{1/2} |R_r - R_n| \geq \varepsilon\right\} \to 0,$$

which follows from

$$\left(\frac{nr}{d}\right)^{1/2} |R_r - R_n| \leq \lambda^{-1/2}\sqrt{r}|R_r - R_n| \to 0 \quad \text{in probability}$$

under (3.2).

(ii) Under (3.4) and $r \to \infty$, Lemma 2 of Shao and Wu (1989) shows that

$$\frac{nr}{d} E_F(R_r - R_n)^2 \to 0,$$

which implies

$$\left(\frac{nr}{d}\right)^{1/2} |R_r - R_n| \to 0 \quad \text{in probability.}$$

The rest of the proof is the same as in (i). □

By combining Theorem 1 and Lemma 1, we have the following result for the nonlinear estimators (3.1) and (3.2). Unlike Theorem 1, it is a weak result because the $Q(\mathbf{x})$ in Lemma 1 converges to zero weakly.

THEOREM 2. *For $\hat{\theta}$ satisfying* (3.1) *and* (3.2), *the following results hold for* $J(t)$, (1.4).

(i) *If d is bounded and* (3.4) *holds, $J(t) \to \Phi(t)$ for each t in probability does not hold except when $\phi_F(X)$ is normally distributed.*

(ii) *If $d \to \infty$, $d/n \to 0$ and* (3.4) *holds,*

$$\sup_t |J(t) - \Phi(t)| \to 0 \quad \textit{in probability.}$$

(iii) *If $d/n \geq \lambda$ for some $\lambda > 0$ and $r \to \infty$,*

$$\sup_t |J(t) - \Phi(t)| \to 0 \quad \textit{in probability.}$$

PROOF. Since $\hat{\sigma}$ in (1.4) converges to σ in probability, we will consider the version of $J(t)$ with $\hat{\sigma}$ replaced by σ. Note that Theorem 1 applies to $\hat{\theta} = n^{-1}\sum_1^n \phi_F(x_i)$, since $0 < E_F(\phi_F^2(X)) < \infty$. By writing

$$\left(\frac{nr}{d}\right)^{1/2} \frac{\hat{\theta}_s - \hat{\theta}}{\sigma}$$

as

$$\left(\frac{nr}{d}\right)^{1/2} \frac{1}{\sigma}\left[\frac{1}{r}\sum_{i \in s} \phi_F(x_i) - \frac{1}{n}\sum_1^n \phi_F(x_i)\right] + \left(\frac{nr}{d}\right)^{1/2} \frac{1}{\sigma}[R_{n,s} - R_n], \tag{3.5}$$

we can apply Theorem 1 and Lemma 1 to the first and second terms of (3.5), respectively. Part (i) follows from Theorem 1(i) and Lemma 1(ii). Part (ii) follows from Theorem 1(ii) and Lemma 1(ii). Part (iii) follows from Theorem 1(ii) and Lemma 1(i). □

Theorem 2(iii) gives the weak consistency of the jackknife histogram for any $\hat{\theta}$ satisfying (3.1) and (3.2). This is a significant result since, as discussed after (3.2), it covers a very broad class of estimators including many smooth and not-so-smooth estimators. It is worth noting that, under the conditions of Theorem 2(iii), the delete-d jackknife variance estimator (1.3) is weakly consistent and asymptotically unbiased [Shao and Wu (1989), Corollary 1].

While $d = O(n)$ in Theorem 2(iii), the complementary case of $d = o(n)$ and $d \to \infty$ is addressed in Theorem 2(ii). An additional condition (3.4) on $\hat{\theta}$ is required. Roughly speaking, (3.4) imposes a more severe requirement on the smoothness of $\hat{\theta}$. Shao and Wu [(1989), Theorem 5 and Example 5] gave two classes of $\hat{\theta}$ satisfying (3.4): (i) $\hat{\theta} = T(F_n)$, where T is second order Fréchet differentiable with respect to the supremum norm and $\mathrm{Var}(\hat{\theta}) = \sigma^2/n + o(n^{-1})$, σ^2 given in (3.1); (ii) U statistics of any degree [see (3.6) and Remark 2 after Theorem 3] satisfying a condition analogous to (3.7) for the case of degree 2. It is not surprising that a more severe condition on the smoothness of $\hat{\theta}$ is imposed when $d \to \infty$ at a slower rate than n, since the theory of Shao and Wu (1989) for jackknife variance estimation points to an inverse relationship between d and the smoothness of $\hat{\theta}$.

The weak result in Theorem 2(ii) and (iii) can be improved to obtain a strong version if a finite population central limit theorem for $\hat{\theta}$ is available. Similarly, by using a finite population Berry–Esseen bound for $\hat{\theta}$, a convergence rate result analogous to Theorem 1(iii) can be obtained. One such example is the U statistics. For the simplicity of presentation, we will study in detail the special case of degree 2 with the symmetric kernel ϕ,

$$\hat{\theta}(\mathbf{x}) = \binom{n}{2}^{-1} \sum_{1 \le j < k \le n} \phi(x_j, x_k). \tag{3.6}$$

For the extension to any general degree, see Remark 2 after Theorem 3. Under $E_F|\phi(X_1, X_2)| < \infty$, where X_1 and X_2 are independent with distribution F, $\hat{\theta}$ is a strongly consistent estimator for the parameter $\theta = E_F\phi(X_1, X_2)$. Define $G(X_1) = E_F(\phi(X_1, X_2)|X_1)$. We assume

$$0 < \xi_1^2 = \mathrm{Var}_F(G(X_1)) \quad \text{and} \quad E_F\phi^2(X_1, X_2) < \infty, \tag{3.7}$$

so that $\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, 4\xi_1^2)$. Let $(Y_1, \dots, Y_r)$ denote the random vector of r distinct elements selected randomly from the population $\{x_1, \dots, x_n\}$, $d = n - r$. Define

$$\begin{aligned} g(Y_1) &= E_*\big(\phi(Y_1, Y_2)|Y_1\big), \\ g_j(\mathbf{x}) &= E_*\big(\phi(Y_1, Y_2)|Y_1 = x_j\big) = \frac{1}{n-1}\sum_{k \ne j}^{n} \phi(x_j, x_k), \\ \hat{\sigma}_g^2 &= \mathrm{Var}_*(g(Y_1)) = \frac{1}{n}\sum_{1}^{n}\big(g_j(\mathbf{x}) - \hat{\theta}\big)^2, \end{aligned} \tag{3.8}$$

where $E_* g(Y_1) = \hat{\theta}$.

Under (3.7), we will show that

$$\hat{\sigma}_g^2 \to \xi_1^2 \quad \text{a.s. as } n \to \infty. \tag{3.9}$$

From (3.8),

$$\begin{aligned}
\hat{\sigma}_g^2 &= \frac{1}{n-1}\binom{n}{2}^{-1} \sum_{1 \le j < k \le n} \phi^2(x_j, x_k) \\
&\quad + \frac{n-2}{n-1}\binom{n}{3}^{-1} \sum_{1 \le j < k < l \le n} \phi(x_j, x_k)\phi(x_j, \chi_l) - \hat{\theta}^2 \\
&= \frac{1}{n-1}\binom{n}{2}^{-1} \sum_{1 \le j < k \le n} \phi^2(x_j, x_k) \\
&\quad + \frac{n-2}{n-1}\binom{n}{3}^{-1} \sum_{1 \le j < k < l \le n} \Psi(x_j, x_k, x_l) - \hat{\theta}^2,
\end{aligned} \tag{3.10}$$

where

$$\begin{aligned}
\Psi(x_1, x_2, x_3) = \tfrac{1}{3}\{&\phi(x_1, x_2)\phi(x_1, x_3) + \phi(x_1, x_2)\phi(x_2, x_3) \\
&+ \phi(x_1, x_3)\phi(x_2, x_3)\}
\end{aligned}$$

is a symmetric function in x_1, x_2 and x_3. From the strong law of large numbers for U statistics, the first, second and third terms of (3.10) converge, respectively, to

$$\lim_{n \to \infty} \frac{1}{n-1} E_F\big(\phi^2(X_1, X_2)\big) = 0,$$

$$E_F \Psi(X_1, X_2, X_3) = E_F\big(G^2(X_1)\big)$$

and θ^2, which establishes $\hat{\sigma}_g^2 \to \xi_1^2 = \mathrm{Var}_F(G(X_1))$ a.s.

Using this $\hat{\sigma}_g$ as an estimate of ξ_1, the jackknife histogram $J(t)$, (1.4), takes the form

$$J(t) = P_*\left(\left(\frac{nr}{d}\right)^{1/2} \frac{\hat{\theta}_s - \hat{\theta}}{2\hat{\sigma}_g} \le t\right), \tag{3.11}$$

where

$$\hat{\theta}_s = \binom{r}{2}^{-1} \sum_{1 \le i < j \le r} \phi(x_{k_i}, x_{k_j}), \qquad s = (k_1, \dots, k_r).$$

For $\hat{\theta}$ and $J(t)$ given in (3.6) and (3.11), we have the following result.

THEOREM 3. (i) *Under* (3.7), $r \to \infty$ *and* $d \to \infty$ *imply*

$$\sup_t |J(t) - \Phi(t)| \to 0 \quad a.s. \tag{3.12}$$

(ii) *Under* $\xi_1^2 > 0$ *and* $E_F|\phi(X_1, X_2)|^3 < \infty$, $r \to \infty$ *and* $r/n \le \lambda < 1$ *for some constant* λ *imply*

$$\sup_t |J(t) - \Phi(t)| \le \frac{C}{\sqrt{r}} A(\mathbf{x}), \tag{3.13}$$

where C *is a constant solely dependent on* λ *and* $A(\mathbf{x})$ *converges a.s. to a finite constant.*

PROOF. (i) From Theorem 2.1 of Zhao and Chen (1990), (3.12) holds under the following conditions:

(a) For any $\tau > 0$,

$$\lim_{n\to\infty} \frac{1}{n\hat{\sigma}_g^2} \sum_{j=1}^{n} \left(g_j(\mathbf{x}) - \hat{\theta}\right)^2 I_{\{|g_j(\mathbf{x}) - \hat{\theta}| \ge \tau\hat{\sigma}_g (rd/n)^{1/2}\}} = 0 \quad \text{a.s.}$$

(b) $\lim_{n\to\infty} (d/rn)\hat{\sigma}_g^{-2} E_*[\phi(Y_1, Y_2) - \hat{\theta}]^2 = 0$ a.s.

The expression in (b) is bounded above by

$$\lim_{n\to\infty} \frac{d}{rn} \frac{1}{\hat{\sigma}_g^2} \binom{n}{2}^{-1} \sum_{1\le j<k\le n} \phi^2(x_j, x_k),$$

which converges to zero a.s. because of $r \to \infty$, $\hat{\sigma}_g^2 \to \xi^2$ a.s. and $E_F\phi^2(X_1, X_2) < \infty$.

To prove (a), we note that $g_j(\mathbf{x})$ are not independent since each $g_j(\mathbf{x})$ depends on all the x_i except x_j. From $\hat{\theta} \to \theta$ and $\hat{\sigma}_g^2 \to \xi_1^2 > 0$, (a) follows from

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \left(g_j(\mathbf{x}) - \theta\right)^2 I_{\{|g_j(\mathbf{x}) - \theta| \ge \tau (rd/n)^{1/2}\}} = 0 \quad \text{for any } \tau > 0 \text{ a.s.} \tag{3.14}$$

Since $\min(r, d) \to \infty$ implies $rd/n \to \infty$ [see (2.12)], (3.14) is bounded above by

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \left(g_j(\mathbf{x}) - \theta\right)^2 I_{\{|g_j(\mathbf{x}) - \theta| \ge 2A\}} \quad \text{for any constant } A > 0. \tag{3.15}$$

Define the three functions

$$h_1(x) = \begin{cases} 0, & \text{if } |x| < 2A, \\ x^2, & \text{if } |x| \ge 2A, \end{cases}$$

$$h_2(x) = \begin{cases} 0, & \text{if } |x| \le A, \\ \text{interpolating linear functions}, & \text{if } A < |x| < 2A, \\ x^2, & \text{if } |x| \ge 2A, \end{cases}$$

$$h_3(x) = \begin{cases} 0, & \text{if } |x| < A, \\ x^2, & \text{if } |x| \ge A. \end{cases}$$

It is clear that $h_1 \le h_2 \le h_3$ and h_2 is convex. From (3.8), the expression

(3.15) equals

$$\limsup_{n\to\infty} E_* h_1(g(Y_1) - \theta) \le \limsup_{n\to\infty} E_* h_2(g(Y_1) - \theta). \tag{3.16}$$

From the Jensen inequality,

$$\begin{aligned} h_2(g(Y_1) - \theta) &= h_2\big(E_*\{\phi(Y_1, Y_2) - \theta | Y_1\}\big) \\ &\le E_*\{h_2(\phi(Y_1, Y_2) - \theta) | Y_1\}. \end{aligned}$$

Therefore, the right-hand expression of (3.16) is bounded above by

$$\begin{aligned} &\limsup_{n\to\infty} E_* h_2(\phi(Y_1, Y_2) - \theta) \\ &\quad \le \limsup_{n\to\infty} E_* h_3(\phi(Y_1, Y_2) - \theta) \\ &\quad = \limsup_{n\to\infty} \binom{n}{2}^{-1} \sum_{1\le j<k\le n} \big(\phi(x_j, x_k) - \theta\big)^2 I_{(|\phi(x_j, x_k) - \theta| \ge A)} \\ &\quad = E_F\big(\phi(X_1, X_2) - \theta\big)^2 I_{(|\phi(X_1, X_2) - \theta \ge A)} \quad \text{a.s.} \end{aligned}$$

The last expression can be made arbitrarily small by choosing a large A since $E_F\phi^2(X_1, X_2) < \infty$. This proves (a).

(ii) Using Theorem 3.1 of Zhao and Chen (1990),

$$\sup_t |J(t) - \Phi(t)| \le \frac{C}{\sqrt{r}} \left\{ \frac{E_*|g(Y_1) - \hat{\theta}|^3}{\hat{\sigma}_g^3} + \frac{E_*\big(\phi(Y_1, Y_2) - \hat{\theta}\big)^2}{\hat{\sigma}_g^2} \right\}, \tag{3.17}$$

where C is a constant solely dependent on λ. Since $\hat{\theta} \to \theta$, $\hat{\sigma}_g^2 \to \xi_1^2$ and $E_*\phi^2(Y_1, Y_2) \to E_F(\phi^2(X_1, X_2)) < \infty$, what remains is the term

$$\begin{aligned} E_*|g(Y_1) - \theta|^3 &= E_*\big|E_*(\phi(Y_1, Y_2) - \theta | Y_1)\big|^3 \\ &\le E_*|\phi(Y_1, Y_2) - \theta|^3 = \binom{n}{2}^{-1} \sum_{1\le j<k\le n} |\phi(x_j, x_k) - \theta|^3, \end{aligned}$$

which converges a.s. to $E_F|\phi(X_1, X_2) - \theta|^3 < \infty$ from the strong law of large numbers for U statistics. This proves (3.13). □

REMARK 1. Extension to a general variance estimate $\tilde{\sigma}_g^2$. As long as $\tilde{\sigma}_g^2 \to \xi_1^2 > 0$ a.s., Theorem 3(i) holds if $\hat{\sigma}_g$ in $J(t)$ is replaced by such $\tilde{\sigma}_g$. This follows easily from Corollary 2.2 of Zhao and Chen (1990). For Theorem 3(ii) to hold, an additional condition on $\tilde{\sigma}_g$ is required. From Corollary 3.2 of Zhao and Chen (1990), a sufficient condition on $\tilde{\sigma}_g$ is

$$\begin{aligned}\frac{\tilde{\sigma}_g^2}{\hat{\sigma}_g^2} - 1 &= O\left\{r^{-1/2}\left(\frac{E_*|g(Y_1) - \hat{\theta}|^3}{\hat{\sigma}_g^3} + \frac{E_*(\phi(Y_1, Y_2) - \hat{\theta})^2}{\hat{\sigma}_g^2}\right)\right\} \\ &= O\left\{r^{-1/2}\binom{n}{2}^{-1} \sum_{1 \le j < k \le n} \frac{|\phi(Y_1, Y_2)|^3}{\hat{\sigma}_g^3}\right\} \\ &= O(r^{-1/2}).\end{aligned}$$

REMARK 2. Extension to U statistics of any fixed degree. This extension is straightforward but tedious. Note that the proof of Theorem 3 is based on Theorems 2.1 and 3.1 of Zhao and Chen (1990), which hold for U statistics of any degree. Verification of the corresponding conditions is more cumbersome.

4. A higher order consideration. In the previous sections, no or weak condition on the growth rate of r is required for the first order asymptotics. A natural question is how to choose r to ensure a better approximation of $J(t)$ to the distribution of $\hat{\theta}$? First we recall the Edgeworth expansion for the distribution of $\bar{x}$:

$$\begin{aligned}H(t) &= P_F\left\{\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \le t\right\} \\ &= \Phi(t) + (1 - t^2)\phi(t)\frac{E_F(X - \mu)^3}{6\sqrt{n}\,\sigma^3} + o\left(\frac{1}{\sqrt{n}}\right).\end{aligned} \tag{4.1}$$

Better approximation of $J(t)$ to $H(t)$ obtains if the $n^{-1/2}$ term of $J(t)$ converges to the $n^{-1/2}$ term of $H(t)$ as $n \to \infty$. Using the Edgeworth expansion for srs without replacement [rigorous justification in Babu and Singh (1985); Bickel and van Zwet (1978); Robinson (1978)], $J(t)$ in the form (2.2) can be expanded as

$$J(t) = \Phi(t) + (1 - t^2)\phi(t)\frac{1 - 2f}{6\sqrt{f(1 - f)}}\frac{n^{-1}\Sigma_1^n(x_i - \bar{x})^3}{\sqrt{n}\,\hat{\sigma}^3} + o\left(\frac{1}{\sqrt{r}}\right), \tag{4.2}$$

where $f = r/n$. Since $n^{-1}\Sigma_1^n(x_i - \bar{x})^3 \to E_F(X - \mu)^3$ and $\hat{\sigma} \to \sigma$, by comparing (4.1) and (4.2),

$$|J(t) - H(t)| = o\left(\frac{1}{\sqrt{n}}\right) \tag{4.3}$$

iff

$$\frac{1 - 2f}{\sqrt{f(1 - f)}} \to 1$$

iff

$$f = \frac{r}{n} \to \frac{5 - \sqrt{5}}{10} = 0.2764.$$

That is, in order to achieve a higher order approximation of the jackknife histogram $J(t)$ to the true distribution $H(t)$, one should retain about 27.64% of the data in computing $\hat{\theta}_s = \bar{x}_r$.

In general, choosing the r value alone will not achieve the second order approximation (4.3). Take, for example, the t statistic $(\bar{x} - \mu)/\hat{\sigma}$, $\hat{\sigma}$ given in (2.1). Its Edgeworth expansion is

$$P_F\left\{\frac{\sqrt{n}(\bar{x} - \mu)}{\hat{\sigma}} \le t\right\} = \Phi(t) + (2t^2 + 1)\phi(t)\frac{E_F(X - \mu)^3}{6\sqrt{n}\,\sigma^3} + o\left(\frac{1}{\sqrt{n}}\right), \tag{4.4}$$

while $J(t)$ admits the expansion [Babu and Singh (1985)]

$$\begin{aligned} J(t) &= P_*\left\{\frac{\sqrt{r}(\bar{x}_r - \bar{x})}{\left[(1 - f)\hat{\sigma}_r^2\right]^{1/2}} \le t\right\} \\ &= \Phi(t) + \left(3t^2 - \frac{1 - 2f}{1 - f}(t^2 - 1)\right)(1 - f)^{1/2}\phi(t) \\ &\quad \times \frac{n^{-1}\sum_1^n (x_i - \bar{x})^3}{6\sqrt{r}\,\hat{\sigma}^3} + o\left(\frac{1}{\sqrt{r}}\right), \end{aligned} \tag{4.5}$$

where $\hat{\sigma}_r^2 = (r - 1)^{-1}\sum_{i \in s}(x_i - \bar{x}_r)^2$. It is obvious that, no matter how f is chosen in (4.5), the difference between (4.5) and (4.4) cannot be of the order $o(n^{-1/2})$. This should not be too surprising since (4.4) is for sampling with replacement and (4.5) is for sampling without replacement. Because of the difference in the probability mechanisms, the polynomials in the second order terms of the Edgeworth expansions are in general different.

If the original data come from i.i.d. sampling, jackknife resampling, which does not mimic i.i.d. sampling, will not in general give a histogram that approximates the original distribution to the second order. On the other hand, if the data are obtained from simple random sampling without replacement from a finite population, the bootstrap histogram does not approximate the original distribution to the second order since bootstrap resampling does not mimic without-replacement sampling. A method that accomplishes the desired second order approximation first replicates the observed sample to the size of the population and then resamples from this enlarged sample using without-replacement sampling [Gross (1980); Babu and Singh (1985)]. So the relative advantage of the two methods in terms of second order approximation depends on the nature of the original sampling plan.

5. Concluding remarks. As discussed in Section 1, it is of considerable theoretical importance to study the asymptotic properties of the jackknife histogram. Our results show that the jackknife histogram has desirable first order asymptotic properties for a variety of statistics. In particular, Theorem 2(iii) ensures its consistency for a very general class of $\hat{\theta}$ satisfying (3.1) and (3.2) if d, the number of deleted observations, is of the order $O(n)$ and $r = n - d \to \infty$. This resampling can be easily implemented by deleting a fixed fraction of the sample. On the other hand, there is no result available on the bootstrap histogram that covers such a broad class of statistics. Our study opens up the possibility of using the jackknife histogram $J(t)$ as an optional method for resampling inference. Although for i.i.d. samples the jackknife does not do as well as the bootstrap in terms of second order approximation, we do not think a method can be solely judged by this criterion.

Regarding the choice of d, our asymptotic normality results suggest that d be chosen to be λn, $0 < \lambda < 1$, but cannot specify the value of λ. As shown in Section 4, use of a second order approximation succeeds in finding such a λ value only in a special case. Because of the success in using the balanced half-samples method for variance estimation in complex surveys [Kish and Frankel (1974)], we think $\lambda = \frac{1}{2}$ deserves special attention. In general one can consider choosing λ between $\frac{1}{4}$ and $\frac{3}{4}$, but the best choice of λ seems to depend on the particular problem. In this regard, the jackknife is not alone. For example, for the bootstrap method in complex surveys, the best resample size depends on the sampling design [Rao and Wu (1988)].

There is no question on the versatility of the bootstrap for statistical estimation and inference. Its theoretical properties have been studied by many investigators. Due to the problems mentioned before, the jackknife has not been received with the same degree of interests. We think the significance of the jackknife as a resampling method may have been overlooked. In some situations such as the following, the jackknife may even be a preferred method.

1. Unlike the bootstrap, the jackknife either retains or omits an observation in its resamples. This take-or-not-take feature of the jackknife makes it more robust than the bootstrap against heteroscedasticity [Wu (1986); Shao and Wu (1987)].
2. As the counterexample of Ghosh, Parr, Singh and Babu (1984) demonstrated, the bootstrap variance estimator for the sample quantile is inconsistent if the underlying distribution F has very heavy tails. Recently Shi (1988) proved the consistency of the delete-d jackknife variance estimator for the sample quantile without any conditions on the tails or moments of F if d/n is bounded away from 0 and 1. This can be explained by the fact that the jackknife point estimate $\hat{\theta}_s$ does not take extreme order statistics while the bootstrap does so, albeit with small probability.
3. The jackknife employs a more systematic method of sampling than the bootstrap. Intuitively, it may be a more efficient method for Monte Carlo approximation, but no theory is available yet.

Acknowledgments. I would like to thank L. C. Zhao for his generous help in the proof of Theorem 3 and Youyi Chen, Xiquan Shi and the referee for helpful comments.

REFERENCES

BABU, G. J. and SINGH, K. (1985). Edgeworth expansions for sampling without replacement for finite populations. *J. Multivariate Anal.* **17** 261–278.

BERAN, R. J. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74** 457–468.

BICKEL, P. J. and FREEDMAN, D. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BICKEL, P. J. and VAN ZWET, W. (1978). Asymptotic expansion for the power of distribution free tests in the two-sample problem. *Ann. Statist.* **6** 937–1004.

COCHRAN, W. G. (1979). *Sampling Techniques*, 2nd ed. Wiley, New York.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

EFRON, B. (1987). Better bootstrap confidence intervals (with discussion). *J. Amer. Statist. Assoc.* **82** 171–185.

GHOSH, M., PARR, W. C., SINGH, K. and BABU, G. J. (1984). A note on bootstrapping the sample median. *Ann. Statist.* **12** 1130–1135.

GROSS, S. (1980). Median estimation in sample surveys. *Proc. Sect. Survey Res. Methods* 181–184. American Statistical Association, Washington, D.C.

HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hungarian Acad. Sci.* **5** 361–374.

HALL, P. (1986). On the bootstrap and confidence intervals. *Ann. Statist.* **14** 1431–1452.

HÖGLUND, T. (1978). Sampling from a finite population: A reminder estimate. *Scand. J. Statist.* **5** 69–71.

KISH, L. and FRANKEL, M. R. (1974). Inference from complex samples (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 1–37.

RAO, J. N. K. and WU, C. F. J. (1988). Resampling inference with complex survey data. *J. Amer. Statist. Assoc.* **83** 231–241.

ROBINSON, J. (1978). An asymptotic expansion for samples from a finite population. *Ann. Statist.* **5** 1005–1011.

SERFLING, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SHAO, J. (1987). Bootstrap variance estimations. Ph.D. thesis, Chap. 4, Univ. Wisconsin.

SHAO, J. and WU, C. F. J. (1987). Heteroscedasticity–robustness of jackknife variance estimators in linear models. *Ann. Statist.* **15** 1563–1579.

SHAO, J. and WU, C. F. J. (1989). A general theory for jackknife variance estimation. *Ann. Statist.* **17** 1176–1197.

SHI, X. Q. (1988). Some asymptotic results for jackknifing the sample quantiles. Unpublished.

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195.

TUKEY, J. (1958). Bias and confidence in not quite large samples (abstract). *Ann. Math. Statist.* **29** 614.

WU, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Ann. Statist.* **14** 1261–1350.

ZHAO, L. C. and CHEN, X. R. (1990). Normal approximation for finite-population U statistics. *Acta Math. Appl. Sinica*. To appear.

FACULTY OF MATHEMATICS
DEPARTMENT OF STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF WATERLOO
WATERLOO, ONTARIO
CANADA N2L 3G1