JACKKNIFE VARIANCE ESTIMATOR FOR TWO SAMPLE LINEAR RANK STATISTICS[1]

by

Jun Shao

Purdue University

Technical Report #88-61

PURDUE UNIVERSITY

DEPARTMENT OF STATISTICS

(I)

JACKKNIFE VARIANCE ESTIMATOR FOR TWO SAMPLE LINEAR RANK STATISTICS[1]

by

Jun Shao

Purdue University
Technical Report #88-61

Department of Statistics
Purdue University

November 1988

[1] The research of this author was partially supported by the Office of Naval Research Contract N00014-88-K-0170 and NSF Grants DMS-8606964, DMS-8702620 at Purdue University.

JACKKNIFE VARIANCE ESTIMATORS FOR TWO SAMPLE LINEAR RANK STATISTICS

Jun Shao*

Purdue University

ABSTRACT

The jackknife estimator of the asymptotic variance of a two sample linear rank statistic is shown to be strongly consistent. Statistical applications of the result are discussed. The technique used in proving the consistency of the jackknife variance estimator can be applied to general situations.

Keywords: strong consistency; linear rank test; influence function.

* The research of this author was partially supported by the Office of Naval Research Contract N00014-88-K-0170 and NSF Grant DMS-8606964, DMS-8702620 at Purdue University.

1. Introduction and the main result

Consider the following test problem concerning two (not necessarily continuous) population distributions F and G:

$$\mathrm{H}_0\colon F=G \quad \text{vs.} \quad \mathrm{H}_1\colon F\neq G. \tag{1.1}$$

Let $\{X_1,...,X_n\}$ and $\{Y_1,...,Y_m\}$ be independent samples from F and G, respectively. For simplicity, we assume that $m=n$. The results obtained in the following can be extended to the case $n/m\to\lambda$, $0<\lambda<1$, with some modifications. The statistic for the test problem (1.1) is the following two-sample simple linear rank statistic (see, e.g., Hájek and Sidák, 1967; Huber, 1981):

$$S(F_n,G_n) = \int J[\tfrac{1}{2}F_n(x)+\tfrac{1}{2}G_n(x)]\mathrm{d}F_n(x), \tag{1.2}$$

where F_n and G_n are empirical distribution functions corresponding to the samples $\{X_1,...,X_n\}$ and $\{Y_1,...,Y_n\}$, respectively, and J is a score function satisfying $J(1-t) = -J(t)$, $t\in[0,1]$. Let $H = \tfrac{1}{2}F+\tfrac{1}{2}G$ and $H_n = \tfrac{1}{2}F_n+\tfrac{1}{2}G_n$. $S(F_n,G_n)$ can be used as a point estimator of the quantity

$$S(F,G) = \int J[H(x)]\mathrm{d}F(x).$$

We assume that $S(F,G) = 0$ under the null hypothesis H_0 (which is satisfied if F is symmetric or F is continuous). Thus, we reject H_0 if $|S(F_n,G_n)|$ is large.

An asymptotic analysis of the sampling distribution of $S(F_n,G_n)$ is needed for obtaining the critical value for the test problem (1.1) and for calculating the power of the test procedure. Chernoff and Savage (1958) showed that under certain conditions (see also Hájek and Sidák, 1967, pp.233-237), $(2n)^{1/2}[S(F_n,G_n)-S(F,G)]$ converges in distribution to $N(0, \sigma^2)$ with

$$\sigma^2 = Var_F\,\phi(X_1) + Var_G\,\psi(Y_1), \tag{1.3}$$

where

$$\begin{aligned}\phi(x) &= \tfrac{1}{2}\int J'[H(y)][I_{(x\le y)}-F(y)]\mathrm{d}F(y) + J[H(x)] - \int J[H(y)]\mathrm{d}F(y)\\ \psi(x) &= \tfrac{1}{2}\int J'[H(y)][I_{(x\le y)}-G(y)]\mathrm{d}F(y),\end{aligned} \tag{1.4}$$

I_A is the indicator function of the set A and J' is the derivative of J. Note that $\phi(x)$ and $\psi(x)$ in (1.4) are influence functions of $S(F,G)$ by using a statistical functional approach (see

Hampel, 1974; Huber, 1981).

Suppose that we have a consistent estimator s^2 of σ^2 given in (1.3), i.e., $s^2 \to \sigma^2$ *a.s.* Then

$$(2n)^{1/2}[S(F_n,G_n) - S(F,G)]/s \to N(0,1) \quad \text{in distribution.}$$

Hence a test procedure with approximate level α ($0<\alpha<1/2$) concludes H_1 if

$$(2n)^{1/2}|S(F_n,G_n)|/s > \Phi^{-1}(1-\alpha/2), \tag{1.5}$$

where Φ is the distribution function of $N(0,1)$. (1.5) gives the critical region of the test for (1.1).

In this note we prove that an estimator of σ^2 obtained by using the jackknife method (Tukey, 1958) is strongly consistent and therefore can be used in the above test procedure.

For $i=1,...,n$, let F_{ni} and G_{ni} be the empirical distribution functions corresponding to the samples $\{X_1,...,X_{i-1},X_{i+1},...,X_n\}$ and $\{Y_1,...,Y_{i-1},Y_{i+1},...,Y_n\}$, respectively, and $H_{ni} = 1/2 F_{ni} + 1/2 G_{ni}$. Let $S(F_{ni},G_{ni})$ be defined as in (1.2) with F_n and G_n replaced by F_{ni} and G_{ni}. The jackknife estimator of σ^2 is defined to be

$$s_J^2 = (n-1)\sum_{i=1}^n [S(F_{ni},G_{ni}) - S(F_n,G_n)]^2.$$

We shall assume the following condition.

Condition A. J' is continuous on $[0,1]$ and $\|J'\|_V$ is finite, where $\| \; \|_V$ is the total variation norm (see Natanson, 1961).

Note that J' satisfies $J'(1-t) = J'(t)$ for $t \in [0,1]$. Hence the condition $\|J'\|_V < \infty$ is satisfied if J' is monotone on $[0,1/2]$. If J'' exists, then $\|J'\|_V = \int_0^1 |J''(t)|dt$ and therefore $\|J'\|_V < \infty$ if J'' is integrable. An example of J satisfying condition A is $J(t) = t - 1/2$ (corresponding to Wilcoxon statistic).

The following is our main result.

Theorem. Assume condition A. Then the jackknife estimator is strongly consistent, i.e.,

$$s_J^2 \to \sigma^2 \quad a.s.$$

2. Proof of the theorem

Let $\phi_n(x)$ and $\psi_n(x)$ be defined as in (1.4) with F, G and H replaced by F_n, G_n and H_n, respectively. We prove the following result first.

Lemma. Assume condition A. Then

$$\|\phi_n - \phi\|_\infty \to 0 \ \ a.s. \quad \text{and} \quad \|\psi_n - \psi\|_\infty \to 0 \ \ a.s., \tag{3.1}$$

where $\| \ \|_\infty$ is the sup norm.

Proof. Under condition A, $\|J'\|_\infty < \infty$. From $\|F_n - F\|_\infty \to 0$ and $\|G_n - G\|_\infty \to 0 \ a.s.$,

$$|J[H_n(x)] - J[H(x)]| \le \|J'\|_\infty \ \|H_n - H\|_\infty \to 0 \ \ a.s.$$

and

$$\int |J[H_n(x)] - J[H(x)]| \mathrm{d}F_n(x) \le \|J'\|_\infty \ \|H_n - H\|_\infty \to 0 \ \ a.s.$$

From the strong law of large numbers (SLLN),

$$\int J[H(x)] \mathrm{d}[F_n(x) - F(x)] \to 0 \ \ a.s.$$

Hence

$$\|J[H_n] - \int J[H_n(x)] \mathrm{d}F_n(x) - J[H] - \int J[H(x)] \mathrm{d}F(x)\|_\infty \to 0 \ \ a.s.$$

For the first assertion in (3.1), it remains to show that

$$\sup_x \left| \int J'[H_n(y)][I_{(x \le y)} - F_n(y)] \mathrm{d}F_n(y) - \int J'[H(y)][I_{(x \le y)} - F(y)] \mathrm{d}F(y) \right| \to 0 \ \ a.s. \tag{3.2}$$

The quantity in (3.2) is bounded by

$$\begin{aligned} &\left| \int J'[H_n(y)][F(y) - F_n(y)] \mathrm{d}F_n(y) \right| + \sup_x \left| \int J'[H(y)][I_{(x \le y)} - F(y)] \mathrm{d}[F_n(y) - F(y)] \right| \\ &+ \sup_x \left| \int \{J'[H_n(y)] - J'[H(y)]\}[I_{(x \le y)} - F(y)] \mathrm{d}F_n(y) \right|. \end{aligned} \tag{3.3}$$

The first term in (3.3) is bounded by $\|J'\|_\infty \|F_n - F\|_\infty \to 0 \ a.s.$ The third term in (3.3) is bounded by $\|J'[H_n] - J'[H]\|_\infty$, which $\to 0 \ a.s.$ since J' is continuous on [0,1]. From the SLLN, $\int J'[H(y)]F(y) \mathrm{d}[F_n(y) - F(y)] \to 0 \ a.s.$ Hence (3.2) follows from

$$\sup_x \left| \int J'[H(y)] I_{(x \le y)} \mathrm{d}[F_n(y) - F(y)] \right| \to 0 \ \ a.s. \tag{3.4}$$

Let $I_x(y) = I_{(x \le y)}$ and $g_x(y) = J'[H(y)]I_x(y)$. From Natanson (1961, p.232),

$$\left| \int J'[H(y)] I_{(x \le y)} \mathrm{d}[F_n(y) - F(y)] \right| \le \|g_x\|_V \|F_n - F\|_\infty.$$

Note that $\| g_x \|_V \le \| J' \|_V \| I_x \|_\infty + \| J' \|_\infty \| I_x \|_V \le \| J' \|_V + \| J' \|_\infty$. Hence (3.4) holds and the first assertion follows. The proof of the second assertion is similar. □

Proof of Theorem. Let

$$V_{ni} = \int \phi(x) d[F_{ni}(x) - F_n(x)] + \int \psi(x) d[G_{ni}(x) - G_n(x)],$$

$$U_{ni} = \int [\phi_n(x) - \phi(x)] d[F_{ni}(x) - F_n(x)] + \int [\psi_n(x) - \psi(x)] d[G_{ni}(x) - G_n(x)],$$

$$W_{ni} = \int \{J[H_{ni}(x)] - J[H_n(x)]\} dF_{ni}(x) - \int J'[H_n(x)][H_{ni}(x) - H_n(x)] dF_n(x)$$

and $R_{ni} = U_{ni} + W_{ni}$. Then

$$S(F_{ni}, G_{ni}) - S(F_n, G_n) = V_{ni} + R_{ni}$$

and therefore

$$s_J^2 = (n-1)\sum_{i=1}^n (V_{ni}^2 + R_{ni}^2 + 2V_{ni}R_{ni}).$$

Let $\xi_i = \phi(X_i)$, $\zeta_i = \psi(Y_i)$, $\bar{\xi} = n^{-1}\sum_{i=1}^n \xi_i$ and $\bar{\zeta} = n^{-1}\sum_{i=1}^n \zeta_i$. Then

$$\begin{aligned}(n-1)\sum_{i=1}^n V_{ni}^2 &= (n-1)\{\sum_{i=1}^n [(n-1)^{-1}\sum_{j\ne i}\xi_j - \bar{\xi}]^2 + \sum_{i=1}^n [(n-1)^{-1}\sum_{j\ne i}\zeta_j - \bar{\zeta}]^2 \\ &\quad + 2\sum_{i=1}^n [(n-1)^{-1}\sum_{j\ne i}\xi_j - \bar{\xi}][(n-1)^{-1}\sum_{j\ne i}\zeta_j - \bar{\zeta}]\} \\ &= (n-1)^{-1}\sum_{i=1}^n (\xi_i - \bar{\xi})^2 + (n-1)^{-1}\sum_{i=1}^n (\zeta_i - \bar{\zeta})^2 + 2(n-1)^{-1}\sum_{i=1}^n (\xi_i - \bar{\xi})(\zeta_i - \bar{\zeta}),\end{aligned}$$

which converges *a.s.* to σ^2 according to the SLLN. From Cauchy-Schwarz inequality, it remains to show that

$$(n-1)\sum_{i=1}^n R_{ni}^2 \to 0 \ \ a.s.,$$

which is implied by

$$\max_{i\le n} |U_{ni}| = o(n^{-1}) \ \ a.s. \tag{3.5}$$

and

$$\max_{i\le n} |W_{ni}| = o(n^{-1}) \ \ a.s. \tag{3.6}$$

Since

$$\begin{aligned}\left|\int [\phi_n(x) - \phi(x)] d[F_{ni}(x) - F_n(x)]\right| &= (n-1)^{-1} |\phi(X_i) - \phi_n(X_i) - n^{-1}\sum_{i=1}^n \phi(X_i)| \\ &\le (n-1)^{-1} \|\phi_n - \phi\|_\infty + [n(n-1)]^{-1} |\sum_{i=1}^n \xi_i|,\end{aligned}$$

$\max_{i\le n} \left|\int [\phi_n(x) - \phi(x)] d[F_{ni}(x) - F_n(x)]\right| = o(n^{-1})$ *a.s.* follows from $\|\phi_n - \phi\|_\infty \to 0$ *a.s.* (Lemma) and $n^{-1}\sum_{i=1}^n \xi_i \to 0$ *a.s.* (SLLN). Similarly, we can prove that

$$\max_{i\le n} \left|\int[\psi_n(x)-\psi(x)]d[G_{ni}(x)-G_n(x)]\right| = o(n^{-1}) \quad a.s.$$

Hence (3.5) holds. From the continuity of J' and $\|H_{ni} - H_n\|_\infty \le n^{-1}$,

$$\max_{i\le n} \left|\int\{J[H_{ni}(x)]-J[H_n(x)] - J'[H_n(x)][H_{ni}(x)-H_n(x)]\}dF_n(x)\right| = o(n^{-1}) \quad a.s.$$

Then (3.6) follows from

$$\max_{i\le n} \left|\int\{J[H_{ni}(x)]-J[H_n(x)]\}d[F_{ni}(x)-F_n(x)]\right| = o(n^{-1}) \quad a.s. \tag{3.7}$$

Again from the continuity of J', (3.7) follows from

$$\max_{i\le n} \left|\int J'[H_n(x)][H_{ni}(x)-H_n(x)]d[F_{ni}(x)-F_n(x)]\right| = o(n^{-1}) \quad a.s. \tag{3.8}$$

Note that

$$\left|\int J'[H_n(x)][H_{ni}(x)-H_n(x)]d[F_{ni}(x)-F_n(x)]\right| \le \|F_{ni} - F_n\|_\infty \|J'[H_n][H_{ni}-H_n]\|_V$$
$$\le n^{-1}\|J'[H_n][H_{ni}-H_n]\|_V \le n^{-1}(\|J'\|_V \|H_{ni}-H_n\|_\infty + \|J'\|_\infty \|H_{ni}-H_n\|_V).$$

Since $F_{ni}(y)-F_n(y) = (n-1)^{-1}[F_n(y)-I_{X_i}(y)]$, where $I_{X_i}(y) = I_{(X_i\le y)}$,

$$\|F_{ni} - F_n\|_V = (n-1)^{-1}\|F_n - I_{X_i}\|_V \le (n-1)^{-1}(\|F_n\|_V + \|I_{X_i}\|_V) = 2(n-1)^{-1}.$$

Similarly, $\|G_{ni} - G_n\|_V \le 2(n-1)^{-1}$ and therefore $\|H_{ni} - H_n\|_V \le 2(n-1)^{-1}$. Hence (3.8) holds. This proves the theorem. □

3. Comments

In some situations (e.g., F is continuous), the variance of $S(F_n,G_n)$ under the null hypothesis H_0 can be calculated using theory of rank statistics (see Hájek and Sidák, 1967). Hence the critical region of the test procedure for (1.1) can be constructed by using $s^2 = (2n)Var[S(F_n,G_n)|H_0]$, provided $(2n)Var[S(F_n,G_n)|H_0] \to \sigma^2$. In these situations, the jackknife just gives an alternative method. Computing the jackknife estimator is routine and simple and does not require a theoretical derivation of $Var[S(F_n,G_n)|H_0]$. Furthermore, the consistency of the jackknife estimator holds under both null and alternative hypotheses and therefore the jackknife may provide other statistical analysis procedures in some situations. For example, suppose that under the alternative H_1, $G(x) = qF + pF^2$ (see Serfling, 1980, p.293), where p may or may not be known, $0<p\le 1$ and $q=1-p$. Suppose also that F is continuous. Then

$$S(F,G) = \int_0^1 J[t-p(t-t^2)]\mathrm{d}t.$$

Denote this quantity by $g(p)$. Then the power of the test at $p = p_1$ is approximately

$$1 - \Phi[\Phi^{-1}(1-\alpha/2)-n^{1/2}g(p_1)/s_J] + \Phi[\Phi^{-1}(\alpha/2)-n^{1/2}g(p_1)/s_J].$$

Assume that J is strictly increasing. Then $g(p)$ is strictly decreasing in p. If p is unknown, an approximate 100(1−α)% confidence interval for p has limits

$$g^{-1}[S(F_n,G_n) \pm \Phi^{-1}(1-\alpha/2)n^{-1/2}s_J].$$

Finally, the technique used in the proof of the consistency of jackknife estimator can be applied to general situations where $S(F,G)$ is a functional with inference functions satisfying (3.1).

References

Chernoff, H. and Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Statist.* **29**, 972-994.

Hájek, J. and Sidák, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69**, 383-397.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Natanson, I. P. (1961). *Theory of Functions of a Real Variable*. Vol. 1, rev. ed., Ungar, New York.

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

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

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release, distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Report #88-61	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Purdue University / 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Department of Statistics West Lafayette, IN 47907	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Office of Naval Research / 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-88-K-0170 and NSF Grants DMS-8606964, DMS-8702620
8c. ADDRESS (City, State, and ZIP Code) Arlington, VA 22217-5000	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

JACKKNIFE VARIANCE ESTIMATOR FOR TWO SAMPLE LINEAR RANK STATISTICS (unclassified)

12. PERSONAL AUTHOR(S)

Jun Shao

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM ______ TO ______	November 1988	6

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

Strong consistency; linear rank test; influence function.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The jackknife estimator of the asymptotic variance of a two sample linear rank statistic is shown to be strongly consistent. Statistical applications of the result are discussed. The technique used in proving the consistency of the jackknife variance estimator can be applied to general situations.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Jun Shao	22b. TELEPHONE (Include Area Code) 317-494-6039 / 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED