LIETUVOS MATEMATIKOS RINKINYS
Proc. of the Lithuanian Mathematical Society, Ser. A
Vol. **53**, 2012, 19–23

ISSN 0132-2818
www.mii.lt/LMR/

The consistency of bootstrap and jackknife variance estimators for finite population L-statistics

Andrius Čiginas

Institute of Mathematics and Informatics, Vilnius University
Akademijos 4, LT-08663, Vilnius
E-mail: andrius.ciginas@mif.vu.lt

Abstract. For the linear combinations of order statistics (L-statistics), we present conditions sufficient for the consistency of their finite-population bootstrap variance estimator and the classical jackknife variance estimator.

Keywords: finite population, sampling without replacement, L-statistic, Hoeffding decomposition, bootstrap, jackknife, consistency.

1 Results

Let $\mathcal{X} = \{x_1, \dots, x_N\}$ denote measurements of the study variable x of the population $\{1, \dots, N\}$. Let $\mathbb{X} = \{X_1, \dots, X_n\}$ denote measurements of units of the simple random sample of size $n < N$ drawn without replacement from the population. Let $X_{1:n} \leqslant \dots \leqslant X_{n:n}$ denote the order statistics of $\mathbb{X}$. Define the L-statistic

$$L_n = L_n(\mathbb{X}) = \frac{1}{n} \sum_{j=1}^{n} c_j X_{j:n},$$

and define its normalized version

$$S_n = S_n(\mathbb{X}) = n^{1/2}(L_n - \mathbf{E}\, L_n). \tag{1}$$

Here $c_1, \dots, c_n$ is a given sequence of real numbers called weights. It is convenient (and always possible) to determine these weights by a weight function $J\colon (0,1) \to \mathbb{R}$ as follows:

$$c_j = J\left(\frac{j}{n+1}\right), \quad 1 \leqslant j \leqslant n.$$

Denote $\tilde{\sigma}_n^2 = \mathbf{Var}\, S_n$.

Note that for correct formulation of any statement on the consistency of finite population statistics, we need to consider a sequence of populations $\mathcal{X}_r = \{x_{r,1}, \dots, x_{r,N_r}\}$, with $N_r \to \infty$ as $r \to \infty$, and a sequence of statistics $L_{n_r}(\mathbb{X}_r)$, based on simple random samples $\mathbb{X}_r = \{X_{r,1}, \dots, X_{r,n_r}\}$ drawn without replacement from $\mathcal{X}_r$. In order to keep the notation simple, we shall skip the subscript r in what follows. Denote $n_* = \min\{n, N-n\}$. Then the population size N and the sample size n tend to infinity as $n_* \to \infty$.

The bootstrap estimator of variance. We consider here the finite population bootstrap of [5]. It is important to mention that there are more adaptations of Efron's bootstrap to the case of finite populations, see, e.g., [1, 6, 12, 14]. Write $N = mn + t$, where $0 \leqslant t < n$. The empirical population $\mathcal{X}^*$ is defined by taking m copies $\mathcal{X}_j = \{X_{j1}, \dots, X_{jn}\}$, $1 \leqslant j \leqslant m$ of $\mathbb{X}$ and, if $t > 0$, drawing the simple random sample $\mathcal{Y} = \{Y_1, \dots, Y_t\}$ of size t without replacement from $\mathbb{X}$. If $t = 0$, then put $\mathcal{Y} = \emptyset$. Then

$$\mathcal{X}^* = \left(\bigcup_{j=1}^{m} \mathcal{X}_j \right) \cup \mathcal{Y}. \tag{2}$$

For the population parameter of interest $\tilde{\sigma}_n^2 = \tilde{\sigma}_n^2(\mathcal{X})$, the bootstrap estimator is then defined as the conditional expectation

$$\tilde{\sigma}_{nB}^2 = \mathbf{E}\left(\tilde{\sigma}_n^2(\mathcal{X}^*) \mid \mathbb{X} \right), \tag{3}$$

i.e., the expectation is taken over all $\binom{n}{t}$ empirical populations conditional on $\mathbb{X}$.

Theorem 1. *Assume that $n_* \to \infty$ and $\tilde{\sigma}_n \geqslant c_1 > 0$ for all n_*. Suppose that $\mathbf{E}\, X_1^2 \leqslant C_1 < \infty$ for all n_* and that $J(\cdot)$ is bounded and satisfies the Hölder condition of order $\delta > 1/2$ on $(0,1)$. Then*

$$\tilde{\sigma}_{nB}^2 \xrightarrow{a.s.} \tilde{\sigma}_n^2 \quad \text{as } n_* \to \infty.$$

Let us mention that asymptotic properties of the bootstrap variance estimator for in a sense similar statistics (U-statistics) were studied in [3]. In the case of L-statistics, an exact approximation to $\tilde{\sigma}_{nB}^2$ is proposed in [8], i.e., the error is eliminated, which typically appears in resampling approximations of (3).

In the case of the m out of n bootstrap (independent and identically distributed (i.i.d.) observations), a similar result obtained in [10].

The jackknife estimator of variance. We define the jackknife variance estimator of (1) in the same way as it is done in [4] for symmetric finite population statistics: consider the extended sample $\mathbb{X}_1 = \{X_1, \dots, X_{n+1}\}$ drawn without replacement from the population; then

$$\tilde{\sigma}_{nJ}^2 = \left(1 - \frac{n}{N}\right) \sum_{k=1}^{n+1} (S_{(k)} - \overline{S})^2, \quad \overline{S} = \frac{1}{n+1} \sum_{k=1}^{n+1} S_{(k)}.$$

Here $S_{(k)} = S_n(\mathbb{X}_1 \backslash \{X_k\})$, $1 \leqslant k \leqslant n+1$. In comparison to the classical Quenouille–Tukey estimator in the case of independent observations, $\tilde{\sigma}_{nJ}^2$ additionally includes the finite population correction factor.

Theorem 2. *Assume that $n_* \to \infty$ and $\tilde{\sigma}_n \geqslant c_1 > 0$ for all n_*. Suppose that, for some $\theta > 0$, $\mathbf{E}\,|X_1|^{2+\theta} \leqslant C_1 < \infty$ for all n_* and that $J(\cdot)$ is bounded and satisfies the Hölder condition of order $\delta > 1/2$ on $(0,1)$. Then*

$$\tilde{\sigma}_{nJ}^2 \xrightarrow{P} \tilde{\sigma}_n^2 \quad \text{as } n_* \to \infty.$$

In the case of finite population symmetric statistics, properties of $\tilde{\sigma}_{nJ}^2$ were studied in [2] (see also [4]). In the proof of Theorem 2, we apply some of these general results.

For a comparison with the i.i.d. case, see [11]. See also [13].

2 Proofs

Proof of Theorem 1. Since L-statistic is a symmetric statistic (symmetric function of observations), results on the Hoeffding decomposition from [4] are applicable, i.e., we write $S_n = U_1 + R_1$, where $U_1 = \sum_{i=1}^{n} g_1(X_i)$ is a linear statistic and R_1 is a remainder term. More specifically, by [4], the components U_1 and R_1 are centered and uncorrelated, and, see [9], for $1 \leqslant k \leqslant N$,

$$g_1(x_k) = -n^{-1/2} \sum_{i=1}^{N-1} \left(\mathbb{I}\{i \geqslant k\} - \frac{i}{N} \right) a_i \, \triangle_i, \tag{4}$$

with

$$a_i = a_{N,n,i} = \sum_{j=1}^{n} c_j \binom{i-1}{j-1} \binom{N-i-1}{n-j} \binom{N-2}{n-1}^{-1},$$

where it is assumed that, without loss of generality, $x_1 \leqslant \cdots \leqslant x_N$. Here we denote $\triangle_i = x_{i+1} - x_i$ and $\mathbb{I}\{\cdot\}$ is the indicator function. Since $J(\cdot)$ satisfies the Hölder condition of order $\delta > 1/2$ on $(0,1)$, we have $|c_j - c_{j-1}| \leqslant B(n+1)^{-\delta}$, for some finite constant $B > 0$. By Theorem 1 of [4] we have $\mathbf{E}\, R_1^2 \leqslant \delta_2(S_n)$. Here $\delta_2(S_n) = \mathbf{E}(n_* \mathbb{D}_2 S_n)^2$, where

$$\begin{aligned} \mathbb{D}_2 S_n = {} & S_n\bigl(\mathbb{X}_2 \backslash \{X_{n+1}, X_{n+2}\}\bigr) - S_n\bigl(\mathbb{X}_2 \backslash \{X_1, X_{n+2}\}\bigr) \\ & - S_n\bigl(\mathbb{X}_2 \backslash \{X_2, X_{n+1}\}\bigr) + S_n\bigl(\mathbb{X}_2 \backslash \{X_1, X_2\}\bigr) \end{aligned}$$

with the extended sample $\mathbb{X}_2 = \{X_1, \dots, X_{n+2}\}$, see [4]. Since it is proved in [7, p. 42] that

$$\delta_2(S_n) \leqslant 24B^2 \frac{n_*^2 n^{-1}}{(n+1)^{2\delta}} \, \mathbf{Var}\, X_1, \tag{5}$$

we get

$$\mathbf{E}\, R_1^2 \leqslant 24B^2 n^{1-2\delta} \, \mathbf{Var}\, X_1. \tag{6}$$

Next, we conclude from $\tilde{\sigma}_n^2 = \mathbf{Var}\, U_1 + \mathbf{Var}\, R_1$ and the conditions of the theorem that

$$\tilde{\sigma}_n^2 \longrightarrow \mathbf{Var}\, U_1 \quad \text{as } n_* \to \infty. \tag{7}$$

Consider bootstrap population (2) and draw a simple random sample without replacement $\mathbb{X}^* = \{X_1^*, \dots, X_n^*\}$ from this population. Then the bootstrap estimator of statistic (1) is $S_n^* = S_n(\mathbb{X}^*)$. We analogously decompose $S_n^* = U_1^* + R_1^*$, with $U_1^* = \sum_{i=1}^{n} g_1(X_i^*)$, where the possible realizations $g_1(x_k^*)$, $1 \leqslant k \leqslant N$ of $g_1(X_1^*)$ are based on the bootstrap population $\mathcal{X}^* = \{x_1^*, \dots, x_N^*\}$. Then, by (6), we get $\mathbf{E}(R_1^{*2} \mid \mathbb{X}, \mathcal{Y}) \leqslant 24B^2 n^{1-2\delta}\, \mathbf{Var}(X_1^* \mid \mathbb{X}, \mathcal{Y})$, and then, taking the conditional expectation given $\mathbb{X}$, we obtain

$$\mathbf{E}\left(R_1^{*2} \mid \mathbb{X}\right) \leqslant 24B^2 n^{1-2\delta} h(\mathbb{X}), \tag{8}$$

where we denote $h(\mathbb{X}) = \mathbf{E}[\mathbf{Var}(X_1^* \mid \mathbb{X}, \mathcal{Y}) \mid \mathbb{X}]$. Let us establish an asymptotic behaviour of $h(\mathbb{X})$ as $n_* \to \infty$. We have

$$\mathbf{Var}\left(X_1^* \mid \mathbb{X}, \mathcal{Y}\right) = \frac{1}{N^2} \sum_{1 \leqslant k < l \leqslant N} \left(x_l^* - x_k^*\right)^2$$

and find that, for $1 \leqslant i \neq j \leqslant n$,

$$
\begin{aligned}
p_{klij} &= \mathbf{P}\bigl\{x_k^* = X_i, x_l^* = X_j \bigm| \mathbb{X}\bigr\} = \frac{m^2}{N(N-1)}\mathbf{P}\{X_i, X_j \notin \mathcal{Y} \mid \mathbb{X}\} \\
&\quad + 2\frac{m(m+1)}{N(N-1)}\mathbf{P}\{X_i \in \mathcal{Y}, X_j \notin \mathcal{Y} \mid \mathbb{X}\} + \frac{(m+1)^2}{N(N-1)}\mathbf{P}\{X_i, X_j \in \mathcal{Y} \mid \mathbb{X}\} \\
&= \frac{m^2(n-t)(n-t-1) + 2m(m+1)t(n-t) + (m+1)^2 t(t-1)}{N(N-1)n(n-1)} \\
&= \frac{N(N-m) - (m+1)t}{N(N-1)n(n-1)}.
\end{aligned}
$$

Therefore, we obtain

$$
\begin{aligned}
h(\mathbb{X}) &= \frac{1}{N^2}\sum_{1 \leqslant k < l \leqslant N}\Biggl[\sum_{1 \leqslant i \neq j \leqslant n}(X_j - X_i)^2 p_{klij}\Biggr] \\
&= \frac{N(N-m) - (m+1)t}{N^2}\,\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,
\end{aligned}
$$

where $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$. It follows from here and by the law of large numbers that

$$
h(\mathbb{X}) \xrightarrow{a.s.} \mathbf{Var}\, X_1 \quad \text{as } n_* \to \infty. \tag{9}
$$

Thus, by (8) and from $\tilde{\sigma}_{nB}^2 = \mathbf{Var}(U_1^* \mid \mathbb{X}) + \mathbf{E}(R_1^{*2} \mid \mathbb{X})$, we get

$$
\tilde{\sigma}_{nB}^2 \xrightarrow{a.s.} \mathbf{Var}\left(U_1^* \mid \mathbb{X}\right) \quad \text{as } n_* \to \infty. \tag{10}
$$

Since, by (2.6) in [4], $\mathbf{Var}\, U_1 = n(N-n)\sigma_1^2/(N-1)$, where $\sigma_1^2 = \mathbf{E}\, g_1^2(X_1)$, we find, using (4),

$$
\sigma_1^2 = \frac{1}{n}\Biggl[\sum_{i=1}^{N-1}\frac{i}{N}\biggl(1 - \frac{i}{N}\biggr)a_i^2 \,\triangle_i^2 + 2\sum_{1 \leqslant i < j \leqslant N-1}\frac{i}{N}\biggl(1 - \frac{j}{N}\biggr)a_i a_j \,\triangle_i \triangle_j\Biggr].
$$

Observe that $\sigma_1^2 = n^{-1}\,\mathbf{Var}\, Z_1$, where Z_1 is drawn from the new population with values: $z_{i+1} = z_i + a_i \triangle_i$, $i = 1, \ldots, N-1$, and $z_1 := 0$. Since $J(\cdot)$ is bounded, there exists an absolute constant $a > 0$ that

$$
\max_{1 \leqslant j \leqslant n} |c_j| \leqslant a \tag{11}
$$

for all n. Therefore $\mathbf{Var}\, Z_1 \leqslant a^2\, \mathbf{Var}\, X_1 < \infty$. Then the fact

$$
\mathbf{Var}\left(U_1^* \bigm| \mathbb{X}\right) \xrightarrow{a.s.} \mathbf{Var}\, U_1 \quad \text{as } n_* \to \infty \tag{12}
$$

follows from the same arguments as that of the proof of (9).

Finally, the theorem follows from (7), (10) and (12).

Proof of Theorem 2. Since $J(\cdot)$ is bounded, condition (11) is satisfied for some $a > 0$. Then the condition $\tilde{\sigma}_n^2 \leqslant c_2$, for some $c_2 > 0$, of Proposition 2 in [4] follows from the bound $\tilde{\sigma}_n^2 \leqslant 2a^2(1 - n/N)\,\mathbf{Var}\, X_1$, see [7, p. 41].

Bound (5) implies that the condition $\delta_2(S_n) = o(1)$ of Proposition 2 in [4] is satisfied.

Next, as it is pointed in [4, p. 904], condition (3.3) (ibidem) can be replaced by the condition $\limsup_n \mathbf{E}(n_* g_1^2(X_1))^{1+\theta} < \infty$, for some $\theta > 0$. Using (2.31) from [7], we get the bound $\mathbf{E}(n_* g_1^2(X_1))^{1+\theta} \leqslant (2a)^{2(1+\theta)}\,\mathbf{E}\,|X_1|^{2(1+\theta)}$. The theorem is proven.

References

[1] P.J. Bickel and D.A. Freedman. Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.*, **12**:470–482, 1984.

[2] M. Bloznelis. A note on the bias and consistency of the jackknife variance estimator in stratified samples. *Statistics*, **37**:489–504, 2003.

[3] M. Bloznelis. Bootstrap approximation to distributions of finite population U-statistics. *Acta Appl. Math.*, **96**:71–86, 2007.

[4] M. Bloznelis and F. Götze. Orthogonal decomposition of finite population statistics and its applications to distributional asymptotics. *Ann. Statist.*, **29**:899–917, 2001.

[5] J.G. Booth, R.W. Butler and P. Hall. Bootstrap methods for finite populations. *J. Amer. Statist. Assoc.*, **89**:1282–1289, 1994.

[6] M.T. Chao and S.H. Lo. A bootstrap method for finite population. *Sankhyā, Ser. A*, **47**:399–405, 1985.

[7] A. Čiginas. *Approximations to Distributions of Linear Combinations of Order Statistics in Finite Populations.* PhD thesis, Vilnius University, 2011. Available from Internet: http://www.mii.lt/files/a_ciginas_phd.pdf.

[8] A. Čiginas. An exact bootstrap for variance of finite-population L-statistic. *Lith. Math. J.*, **51**:322–329, 2011.

[9] A. Čiginas. An Edgeworth expansion for finite-population L-statistics. *Lith. Math. J.*, **52**:40–52, 2012. See also arXiv:1103.4220v2.x

[10] N.V. Gribkova. Bootstrap approximation of distributions of the L-statistics. *J. Math. Sci.*, **109**:2088–2102, 2002.

[11] W.C. Parr and W.R. Schucany. Jackknifing L-statistics with smooth weight functions. *J. Amer. Statist. Assoc.*, **77**:629–638, 1982.

[12] J.N.K. Rao and C.F.J. Wu. Resampling inference with complex survey data. *J. Amer. Statist. Assoc.*, **83**:231–241, 1988.

[13] J. Shao and C.F.J. Wu. A general theory for jackknife variance estimation. *Ann. Statist.*, **17**:1176–1197, 1989.

[14] R.R. Sitter. A resampling procedure for complex survey data. *J. Amer. Statist. Assoc.*, **87**:755–765, 1992.

REZIUMĖ

Baigtinių populiacijų L-statistikų savirankos ir visrakčio dispersijos įvertinių pagrįstumas

A. Čiginas

Pozicinių statistikų tiesinėms kombinacijoms (L-statistikoms) pateikiame pakankamas sąlygas, kurioms esant statistikų baigtinės populiacijos savirankos dispersijos įvertinys ir klasikinis visrakčio dispersijos įvertinys yra pagrįstieji.

Raktiniai žodžiai: baigtinė populiacija, ėmimas be grąžinimo, L-statistika, Hoeffding'o skleidinys, saviranka, visraktis, pagrįstumas.