Lietuvos matematikos rinkinys. LMD darbai
51 tomas, 2010, 391–396

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

Bootstrap, jackknife and Edgeworth approximations for finite population L-statistics

Andrius Čiginas

Vilnius University, Faculty of Mathematics and Informatics
Naugarduko 24, LT-03225 Vilnius
E-mail: andrius.ciginas@mif.vu.lt

Abstract. In this paper we give exact bootstrap estimators for the parameters defining one-term Edgeworth expansion of distribution function of finite population L-statistic and compare these estimators with corresponding jackknife estimators. We also compare 'true' distribution of L-statistic with its normal approximation, Edgeworth expansion, empirical Edgeworth expansion and bootstrap approximation.

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

1 Introduction

Consider a population $\mathcal{X} = \{x_1, \dots, x_N\}$ of size N. We assume without loss of generality that $x_1 \leqslant \cdots \leqslant x_N$. Let $\mathbb{X} = \{X_1, \dots, X_n\}$ be the simple random sample of size $n < N$ *drawn without replacement* from $\mathcal{X}$ and let $X_{1:n} \leqslant \cdots \leqslant X_{n:n}$ denote the order statistics of $\mathbb{X}$. Consider a linear combination $L_n = L_n(\mathbb{X}) = \frac{1}{n}\sum_{j=1}^{n} c_j X_{j:n}$ of the order statistics with coefficients determined by a weight function $J\colon (0,1) \to \mathbb{R}$ as follows $c_j = J(j/(n+1))$, $j = 1, \dots, n$.

In the case where random variables $X_1, \dots, X_n$ are independent and identically distributed (i.i.d.) asymptotic properties of the distribution function $F_n(x) = \mathbf{P}\{S_n \leqslant x\}$ of $S_n = (L_n - \mu)/\sigma$ were widely studied. Here $\mu = \mathbf{E}L_n$ and $\sigma^2 = \mathbf{Var}L_n$. For results on Berry–Esseen bounds we refer to [9, 13], for questions about Edgeworth expansion and empirical Edgeworth expansions we refer to [10, 11, 2, 12].

For samples drawn without replacement from finite population the most general results on one-term Edgeworth expansion and empirical Edgeworth expansions are obtained in [5, 3]. Similarly as in the i.i.d. case (see [2, 12]), these works are devoted to general symmetric asymptotically linear statistics. The analysis in these papers is based on Hoeffding's decomposition of symmetric statistics (see, e.g. [5]) $L_n = \mathbf{E}L_n + U_1 + U_2 + R_n$, where $U_1 = \sum_{1\leqslant i\leqslant n} g_1(X_i)$ and $U_2 = \sum_{1\leqslant i<j\leqslant n} g_2(X_i, X_j)$ are the linear and quadratic parts of the decomposition and the remainder term R_n is negligible under appropriate smoothness conditions. It is shown in [5] that

$$G_n(x) = \Phi(x) - \frac{(q-p)\alpha + 3\kappa}{6\tau}\Phi^{(3)}(x) \tag{1}$$

provides the one term Edgeworth expansion to the distribution function $F_n(x)$. Here $\Phi^{(3)}(x)$ denotes the third derivative of the standard normal distribution function $\Phi(x)$,

$\tau^2 = Npq$, $p = n/N$, $q = 1 - p$ and

$$\alpha = \sigma_1^{-3}\frac{1}{N}\sum_{k=1}^{N} g_1^3(x_k), \qquad \kappa = \sigma_1^{-3}\tau^2\frac{1}{\binom{N}{2}}\sum_{1\leqslant k<l\leqslant N} g_2(x_k,x_l)g_1(x_k)g_1(x_l) \tag{2}$$

and $\sigma_1^2 = \frac{1}{N}\sum_{k=1}^{N} g_1^2(x_k)$. Note that the parameters α, κ and σ_1 defining G_n depend on the kernels $g_1(\cdot)$ and $g_2(\cdot\,,\cdot)$ only. Usualy α, κ and σ_1 are unknown population characteristics and one can not apply (1) directly. One way to overcome this problem is to replace these parameters by their estimators thus obtaining empirical Edgeworth expansion. In [3] it was done by using jackknife estimators. In the case of L-statistics explicit expressions of the kernels $g_1(\cdot)$ and $g_2(\cdot\,,\cdot)$ are available, see [7]. Using these expressions one can also construct bootstrap estimators.

Here we study the bootstrap estimators of α, κ and σ_1. We consider the finite population bootstrap of [6]. In particular, we compare them (their efficiency) with the jackknife estimators. Also, we aim to compare 'true' distribution function F_n (obtained by Monte Carlo (M–C) method) with its normal approximation, Edgeworth expansion, two empirical Edgeworth expansions (with parameters estimated in two ways mentioned) and bootstrap approximation of F_n. We note that the accuracy and features of the latter approximation is not completely understood for L-statistics (on bootstrap for U-statistics see [4]).

2 Estimators for parameters α, κ and σ_1

2.1 Bootstrap estimators

In order to estimate parameters defining Edgeworth expansion (1) we shall consider bootstrap method proposed in [6]. Generally, let $\theta = \theta(\mathcal{X})$ be a characteristic of the population $\mathcal{X}$. Assume that $N = mn + t$, where $0 \leqslant t < n$. Given the sample $\mathbb{X}$ we construct an empirical population $\mathcal{X}^*$ by combining m copies of $\mathbb{X}$ and a simple random sample without replacement $\mathbb{Y} = \{Y_1, \dots, Y_t\}$ of size t from $\mathbb{X}$. Then bootstrap estimator of θ is conditional expectation

$$\hat{\theta} = \mathbf{E}\big(\theta(\mathcal{X}^*) \mid \mathbb{X}\big), \tag{3}$$

i.e. expectation over all empirical populations conditional on $\mathbb{X}$. Practically one can obtain bootstrap estimates of parameters of interest by using M-C method (see, e.g. [6]), but we shall give exact expressions of estimators $\hat{\alpha}_B$, $\hat{\kappa}_B$ and $\hat{\sigma}_{1B}$. Our approach is the following. First we give bootstrap estimator (3) for any of population $\mathcal{X}$ characteristics $\theta_k = g_1(x_k)$, $1 \leqslant k \leqslant N$ and $\theta_{kl} = g_2(x_k, x_l)$, $1 \leqslant k < l \leqslant N$ (see Theorem 1 of [7]). Denote $\mathbf{\Delta}_{j:n} = X_{j+1:n} - X_{j:n}$, $1 \leqslant j \leqslant n-1$ and write $\triangle_i = x_{i+1} - x_i$, $1 \leqslant i \leqslant N-1$. Denote $\mathcal{H}_{N,n,i}(r) = \binom{i}{r}\binom{N-i}{n-r}/\binom{N}{n}$ the probability that a hypergeometric random variable with parameters N, n and i is equal to r. Denote $u_i(k) = -n^{-1}\varphi_k(i)\sum_{p=1}^{n} c_p\mathcal{H}_{N-2,n-1,i-1}(p-1)$ and $v_i(k,l) = -n^{-1}\phi_{k,l}(i)\sum_{p=2}^{n}(c_p - c_{p-1})\mathcal{H}_{N-4,n-2,i-2}(p-2)$, where $\varphi_k(i) = \mathbb{I}\{i \geqslant k\} - i/N$ (here $\mathbb{I}\{\cdot\}$ is the indicator set function) and

$$\phi_{k,l}(i) = \begin{cases} i(i-1)/[(N-1)(N-2)] & \text{if } 1 \leqslant i < k, \\ -(i-1)(N-i-1)/[(N-1)(N-2)] & \text{if } k \leqslant i < l, \\ (N-i-1)(N-i)/[(N-1)(N-2)] & \text{if } l \leqslant i < N. \end{cases}$$

Then we write $\theta_k = \sum_{i=1}^{N-1} u_i(k)\ \triangle_i$ and $\theta_{kl} = \sum_{i=1}^{N-1} v_i(k,l)\ \triangle_i$.

Proposition 1. *We have*

$$\hat{\theta}_k = \sum_{j=1}^{n-1}\sum_{s=0}^{t} u_{mj+s}(k)\mathcal{H}_{n,t,j}(s)\boldsymbol{\Delta}_{j:n}, \quad 1 \leqslant k \leqslant N, \tag{4}$$

$$\hat{\theta}_{kl} = \sum_{j=1}^{n-1}\sum_{s=0}^{t} v_{mj+s}(k,l)\mathcal{H}_{n,t,j}(s)\boldsymbol{\Delta}_{j:n}, \quad 1 \leqslant k < l \leqslant N. \tag{5}$$

Proof. We prove formula (4). Consider empirical population $\mathcal{X}^* = \{x_1^*, \ldots, x_N^*\}$, where $x_1^* \leqslant \cdots \leqslant x_N^*$. Denote $\triangle_i^* = x_{i+1}^* - x_i^*$, $i = 1, \ldots, N-1$. Introduce random variables i_j, $j = 1, \ldots, n-1$, where i_j is a number from the set $\{1, \ldots, N-1\}$ such that $\triangle_{i_j}^* = \boldsymbol{\Delta}_{j:n}$. Clearly, drawing $\mathbb{Y}$ without replacement from $\mathbb{X}$ we have $\mathbf{P}\{i_j = mj+s\} = \mathcal{H}_{n,t,j}(s)$, $s = 0, \ldots, t$ for each $j = 1, \ldots, n-1$. Observe that $g_1(x_k^*) = \sum_{i=1}^{N-1} u_i(k)\ \triangle_i^* = \sum_{j=1}^{n-1} u_{i_j}(k)\ \triangle_{i_j}^*$. Therefore (4) follows from $\mathbf{E}(u_{i_j}(k)\ \triangle_{i_j}^* |\ \mathbb{X}) = \boldsymbol{\Delta}_{j:n}\sum_{s=0}^{t} u_{mj+s}(k)\mathcal{H}_{n,t,j}(s)$, for $j = 1, \ldots, n-1$. The proof of (5) is the same. □

Now we obtain bootstrap estimators $\hat{\alpha}_B$, $\hat{\kappa}_B$ and $\hat{\sigma}_{1B}$ of α, κ and σ_1 by substitution of (4) and (5) into (2).

2.2 Jackknife estimators

We define jackknife estimators for the parameters (2) as follows (cf. [12, 3]). For $1 \leqslant i \leqslant n$ and $1 \leqslant k \neq l \leqslant n$ denote $V_i = \bar{L} - L_{(i)}$ and $W_{kl} = \tilde{L} - \bar{L}_{(k)} - \bar{L}_{(l)} + L_{(k,l)}$, where

$$\bar{L} = \frac{1}{n}\sum_{j=1}^{n} L_{(j)}, \qquad \bar{L}_{(i)} = \frac{1}{n-1}\sum_{1\leqslant j\leqslant n, j\neq i} L_{(i,j)}, \qquad \tilde{L} = \frac{1}{\binom{n}{2}}\sum_{1\leqslant i<j\leqslant n} L_{(i,j)},$$

and where $L_{(i)} = L_{n-1}(\mathbb{X}\backslash\{X_i\}) = \frac{1}{n-1}\sum_{j=1}^{n-1} J(j/n)X_{j:n-1}^{(i)}$ with order statistics $X_{1:n-1}^{(i)} \leqslant \cdots \leqslant X_{n-1:n-1}^{(i)}$, which correspond to the set $\mathbb{X}\backslash\{X_i\}$, and $L_{(k,l)} = L_{n-2}(\mathbb{X}\backslash\{X_k, X_l\}) = \frac{1}{n-2}\sum_{j=1}^{n-2} J(j/(n-1))X_{j:n-2}^{(k,l)}$ with order statistics $X_{1:n-2}^{(k,l)} \leqslant \cdots \leqslant X_{n-2:n-2}^{(k,l)}$, which correspond to the set $\mathbb{X}\backslash\{X_k, X_l\}$. Then jackknife estimators of interest are

$$\hat{\sigma}_{1J}^2 = \frac{1}{n}\sum_{i=1}^{n} V_i^2, \qquad \hat{\alpha}_J = \hat{\sigma}_{1J}^{-3}\frac{1}{n}\sum_{i=1}^{n} V_i^3, \qquad \hat{\kappa}_J = \hat{\sigma}_{1J}^{-3}\tau^2\frac{1}{\binom{n}{2}}\sum_{1\leqslant k<l\leqslant n} W_{kl}V_kV_l.$$

3 Approximations to distribution function F_n

Replacing the moments (2) in (1) by their bootstrap and jackknife estimators given in Section 2 we obtain two empirical Edgeworth expansions. Denote them by $\hat{G}_{nB}(x)$ and $\hat{G}_{nJ}(x)$ respectively.

We will now consider the bootstrap approximation to distribution function. Let $\mathcal{X}^*$ be the empirical population defined in Section 2.1. We draw simple random sample

without replacement $\mathbb{X}^*$ (it is called resample) from $\mathcal{X}^*$. Define $S_n^* = (L_n(\mathbb{X}^*) - \mu(\mathcal{X}^*))/\sigma(\mathcal{X}^*)$ and consider the bootstrap estimator $\hat{\theta} = \mathbf{P}(S_n^* \leqslant x \mid \mathbb{X})$ of $\theta = F_n(x)$. In general the bootstrap distribution is difficult, if not impossible, to calculate, therefore we need to approximate it in order to apply it in our simulations. In that purpose we employ the same M-C method as in [6] as follows. Given the sample $\mathbb{X}$ we construct independently C empirical populations $\mathcal{X}^*_{(1)}, \dots, \mathcal{X}^*_{(C)}$. For each $c = 1, \dots, C$ we draw independently R resamples $\mathbb{X}^*_{(c,1)}, \dots, \mathbb{X}^*_{(c,R)}$ from $\mathcal{X}^*_{(c)}$. Then M-C approximation to $\hat{\theta}$ is $\tilde{F}_{nB}(x) = (CR)^{-1} \sum_{c=1}^{C} \sum_{r=1}^{R} \mathbb{I}\{S^*_{n;(c-1)R+r} \leqslant x\}$, where $S^*_{n;(c-1)R+r}$ is the value of S_n^* for $\mathcal{X}^*_{(c)}$ and $\mathbb{X}^*_{(c,r)}$.

We note that parameters $\mu = \mu(\mathcal{X})$ and $\sigma^2 = \sigma^2(\mathcal{X})$ can be expressed as follows. Simple combinatorial calculations give $\mu = n^{-1} \sum_{p=1}^{n} c_p \mathbf{E} X_{p:n}$,

$$\sigma^2 = n^{-2} \Bigg[\sum_{p=1}^{n} c_p^2 \mathbf{Var} X_{p:n} + 2 \sum_{1 \leqslant p < r \leqslant n} c_p c_r \mathbf{Cov}(X_{p:n}, X_{r:n}) \Bigg],$$

where

$$\mathbf{Var} X_{p:n} = \binom{N}{n}^{-1} \sum_{i=1}^{N} \binom{i-1}{p-1} \binom{N-i}{n-p} x_i^2 - (\mathbf{E} X_{p:n})^2, \quad 1 \leqslant p \leqslant n$$

and

$$\mathbf{Cov}(X_{p:n}, X_{r:n}) = \binom{N}{n}^{-1} \sum_{1 \leqslant i < j \leqslant N} \binom{i-1}{p-1} \binom{j-i-1}{r-p-1} \binom{N-j}{n-r} x_i x_j - \mathbf{E} X_{p:n} \mathbf{E} X_{r:n}, \quad 1 \leqslant p < r \leqslant n,$$

with $\mathbf{E} X_{p:n} = \binom{N}{n}^{-1} \sum_{i=1}^{N} \binom{i-1}{p-1} \binom{N-i}{n-p} x_i$, $1 \leqslant p \leqslant n$.

4 Simulation study and conclusions

L-statistics are applied for estimation of location and scale parameters of distribution of X_1, see, e.g. [8]. For more examples of L-statistics see, e.g. [1]. For our simulation study we choose well known trimmed mean, for $0 \leqslant p_1 < p_2 \leqslant 1$ defined by the weight function $J(u) = (p_2 - p_1)^{-1} \mathbb{I}\{p_1 < u < p_2\}$. Here we take $p_1 = 0.25$ and $p_2 = 0.75$.

We consider two different populations of size $N = 60$. The populations $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ were simulated from exponential $\mathcal{E}(0.5)$ and Cauchy $\mathcal{C}(2, 1)$ distributions respectively. In both cases the sample size is $n = 24$. Table 1 presents simulation results for population $\mathcal{X}^{(1)}$ and Table 2 – for $\mathcal{X}^{(2)}$.

In each table we give $q = 0.05, 0.1, 0.9, 0.95$ quantiles of distribution functions F_n, Φ, G_n and empirical distribution functions $\hat{G}_{nB}$, $\hat{G}_{nJ}$, $\tilde{F}_{nB}$. Instead of single estimate of quantile for the last three functions we give two characteristics: estimated values of the expectation and standard error of empirical quantile, which were calculated from 30 estimates of quantile (we draw independently 30 samples from $\mathcal{X}^{(i)}$, $i = 1, 2$). We note that 'true' distribution F_n by drawing independently 10^6 samples from $\mathcal{X}^{(i)}$, $i = 1, 2$ was obtained. Also we note that in order to obtain $\tilde{F}_{nB}$ by M-C simulations (discussed in Section 3) we choose $C = 10$ and $R = 10^5$.

Table 1. The case of population $\mathcal{X}^{(1)}$.

$q=$	0.05		0.10		0.90		0.95		α	0.24
$F_n^{-1}(q)$	-1.52		-1.23		1.32		1.75		κ	0.49
$\Phi^{-1}(q)$	-1.64		-1.28		1.28		1.64			MSE
$G_n^{-1}(q)$	-1.55		-1.24		1.33		1.77		$\hat{\alpha}_B$	0.02
$\hat{G}_{nB}^{-1}(q)$	-1.56	0.04	-1.25	0.01	1.33	0.02	1.76	0.06	$\hat{\alpha}_J$	0.04
$\hat{G}_{nJ}^{-1}(q)$	-1.58	0.07	-1.26	0.03	1.32	0.04	1.74	0.10	$\hat{\kappa}_B$	0.05
$\tilde{F}_{nB}^{-1}(q)$	-1.53	0.06	-1.23	0.03	1.32	0.02	1.74	0.05	$\hat{\kappa}_J$	0.16

Table 2. The case of population $\mathcal{X}^{(2)}$.

$q=$	0.05		0.10		0.90		0.95		α	0.15
$F_n^{-1}(q)$	-1.57		-1.22		1.28		1.71		κ	0.31
$\Phi^{-1}(q)$	-1.64		-1.28		1.28		1.64			MSE
$G_n^{-1}(q)$	-1.58		-1.26		1.31		1.72		$\hat{\alpha}_B$	0.09
$\hat{G}_{nB}^{-1}(q)$	-1.62	0.12	-1.27	0.04	1.31	0.09	1.71	0.17	$\hat{\alpha}_J$	0.07
$\hat{G}_{nJ}^{-1}(q)$	-1.63	0.20	-1.29	0.10	1.35	0.19	1.73	0.26	$\hat{\kappa}_B$	0.42
$\tilde{F}_{nB}^{-1}(q)$	-1.58	0.17	-1.22	0.09	1.26	0.06	1.67	0.15	$\hat{\kappa}_J$	1.46

In both tables we also give the values of the parameters α and κ, and estimated values of the mean square errors (MSEs) of their estimators $\hat{\alpha}_B$, $\hat{\alpha}_J$ and $\hat{\kappa}_B$, $\hat{\kappa}_J$.

Table 1 shows that approximations G_n, $\hat{G}_{nB}$, $\hat{G}_{nJ}$, $\tilde{F}_{nB}$ outperform Φ. Also, $\hat{G}_{nB}$ is more stable compared to $\hat{G}_{nJ}$. There is no leading approximation of F_n in Table 2, but $\hat{G}_{nB}$ also improves upon $\hat{G}_{nJ}$.

Acknowledgement. I am grateful to M. Bloznelis for useful comments.

References

[1] B.C. Arnold, N. Balakrishnan and H.N. Nagaraja. *A First Course in Order Statistics*. Wiley, New York, 1992.

[2] V. Bentkus, F. Götze and W.R. van Zwet. An Edgeworth expansion for symmetric statistics. *Ann. Statist.*, **25**(2):851–896, 1997.

[3] M. Bloznelis. Empirical Edgeworth expansion for finite population statistics I, II. *Lith. Math. J.*, **41**(2,3):120–134, 207–218, 2001.

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

[5] M. Bloznelis and F. Götze. Orthogonal decomposition of finite population statistics and its applications to distributional asymptotics. *Ann. Statist.*, **29**(3):899–917, 2001.

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

[7] A. Čiginas. Orthogonal decomposition of finite population L-statistics. *Liet. Mat. Rink.*, **50**:287–292, 2009.

[8] H. Chernoff, J.L. Gastwirth and M.V. Johns Jr. Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *Ann. Math. Statist.*, **38**(1):52–72, 1967.

[9] R. Helmers. The order of the normal approximation for linear combinations of order statistics with smooth weight functions. *Ann. Probab.*, **5**(6):940–953, 1977.

[10] R. Helmers. Edgeworth expansions for linear combinations of order statistics with smooth weight functions. *Ann. Statist.*, **8**(6):1361–1374, 1980.

[11] H. Putter. *Consistency of Resampling Methods*. PhD thesis, Leiden University, 1994.

[12] H. Putter and W.R. van Zwet. Empirical Edgeworth expansions for symmetric statistics. *Ann. Statist.*, **26**(4):1540–1569, 1998.

[13] W.R. van Zwet. A Berry–Esseen bound for symmetric statistics. *Z. Wahrsch. Verw. Gebiete*, **66**(3):425–440, 1984.

REZIUMĖ

Baigtinių populiacijų *L*-statistikų savirankos, visrakčio ir Edžvorto aproksimacijos

A. Čiginas

Darbe tiriami baigtinių populiacijų *L*-statistikos Edžvorto skleidinio parametrų įvertiniai. Pateikiami tikslūs šių parametrų savirankos įvertiniai, kurie palyginami su atitinkamais visrakčio įvertiniais. Be to, „tikroji“ *L*-statistikos pasiskirstymo funkcija palyginama su jos normaliąja, Edžvorto, empirine Edžvorto ir savirankos aproksimacijomis.

Raktiniai žodžiai: baigtinė populiacija, ėmimas be grąžinimo, *L*-statistika, Hoeffding'o skleidinys, savirankos metodas, visrakčio metodas, Edžvorto skleidinys.