The Exact Bootstrap Mean and Variance of an L-Estimator

Author(s): Alan D. Hutson and Michael D. Ernst
Source: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 62, No. 1 (2000), pp. 89-94
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2680679
Accessed: 12-07-2019 11:35 UTC

J. R. Statist. Soc. B (2000)
62, *Part* 1, *pp.* 89–94

The exact bootstrap mean and variance of an *L*-estimator

Alan D. Hutson and Michael D. Ernst

University of Florida, Gainesville, USA

[Received October 1998. Revised March 1999]

Summary. Exact analytic expressions for the bootstrap mean and variance of any *L*-estimator are obtained, thus eliminating the error due to bootstrap resampling. The expressions follow from the direct calculation of the bootstrap mean vector and covariance matrix of the whole set of order statistics. By using these expressions, recommendations can be made about the appropriateness of bootstrap estimation under given conditions.

Keywords: Kernel quantile estimators; *L*-statistic; Median; Order statistics; Quantile function; Quick estimators; Trimmed mean

1. Introduction

Let $X_{1:n} \leqslant X_{2:n} \leqslant \ldots \leqslant X_{n:n}$ denote the order statistics from an independent and identically distributed sample of size n from a continuous distribution F with support over the entire real line. An L-estimator (or L-statistic) is defined as

$$T_n = \sum_{i=1}^{n} c_i X_{i:n}, \tag{1.1}$$

where the choice of constants c_i determines the properties and functionality of T_n. Special cases of note include the mean, trimmed mean, median, quick estimators of location and scale, and the upper and lower quartiles. The large sample properties of L-estimators have been examined extensively and many important theoretical results exist; see for example Reiss (1989) and Serfling (1980). However, owing to the difficulty in obtaining variance estimators for many of the common L-estimators, such as the trimmed mean, their use has been limited in practice. Parr and Schucany (1982) used the jackknife to obtain variance estimates of L-estimators. Another method which has aided the practitioner in the difficult theoretical problems of L-estimation is the bootstrap, introduced by Efron (1979). More recently Shao and Tu (1995) provided an up-to-date detailed theoretical investigation of the bootstrap; also see Lepage and Billard (1992). Efron and Tibshirani (1993) and Davison and Hinkley (1997) provide a practical approach to the bootstrap.

The bootstrap commonly employs resampling methods to estimate some exact bootstrap solution. The only exact solutions that are available for the bootstrap variance of an L-estimator are for the specific cases of the sample mean and the sample median (only for odd sample sizes). As an alternative, Huang (1991) has provided an algorithm for directly

Address for correspondence: Alan D. Hutson, Division of Biostatistics, Department of Statistics, Health Science Center, University of Florida, Box 100212, Gainesville, FL 32610-0212, USA.
E-mail: AHUTSON@STAT.UFL.EDU

1369–7412/00/62089

calculating the bootstrap mean and variance for any T_n. Although his combinatoric approach is theoretically interesting and practical for very small samples, it is computer intensive and not well suited for general use.

Our approach, which is based on making direct use of the bootstrap distribution of a single order statistic in conjunction with the joint distribution of two order statistics, directly yields the exact bootstrap mean and bootstrap variance for any L-estimator. The key is in obtaining the bootstrap mean vector and covariance matrix of the whole set of order statistics and then obtaining the specific linear combination corresponding to the L-estimator of interest. The method is easily programmed and takes advantage of commonly built-in functions that are available in most statistical software packages, such as S-PLUS and SAS. These exact expressions eliminate the computer-intensive methods of the resampling approximation to the bootstrap and its associated error. This opens the door to

(a) better estimation of the bootstrap standard error and better naïve bootstrap confidence intervals,
(b) improvements in adaptive bootstrap estimation,
(c) expanded use of robust estimators in other applications such as analysis of variance and
(d) a new theoretical examination of the bootstrap estimators.

The basic approach follows from the well-known expressions for the moments of order statistics; for example see David (1981). Specifically, we derive the *exact* nonparametric bootstrap estimators for

$$\mu_{r:n} = E(X_{r:n}),$$
$$\sigma^2_{r:n} = \mathrm{var}(X_{r:n}),$$
$$\sigma_{rs:n} = \mathrm{cov}(X_{r:n}, X_{s:n})$$

and denote them by $\hat{\mu}_{r:n}$, $\hat{\sigma}^2_{r:n}$ and $\hat{\sigma}_{rs:n}$ respectively, where $1 \leqslant r < s \leqslant n$. The bootstrap mean and variance of any T_n follow directly from these estimates.

Let $\mathbf{c} = (c_1, c_2, \ldots, c_n)'$ denote the $n \times 1$ vector of constants corresponding to a specific L-estimator (1.1) and let

$$\hat{\boldsymbol{\mu}} = (\hat{\mu}_{1:n}, \hat{\mu}_{2:n}, \ldots, \hat{\mu}_{n:n})'$$

denote the bootstrap mean vector of the order statistics. Furthermore, denote the $n \times n$ bootstrap covariance matrix of the order statistics as

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\sigma}^2_{1:n} & \hat{\sigma}_{12:n} & \cdots & \hat{\sigma}_{1n:n} \\ \hat{\sigma}_{21:n} & \hat{\sigma}^2_{2:n} & \cdots & \hat{\sigma}_{2n:n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{n1:n} & \hat{\sigma}_{n2:n} & \cdots & \hat{\sigma}^2_{n:n} \end{pmatrix}. \tag{1.2}$$

It follows that the exact bootstrap mean and variance of T_n are given by

$$\hat{\mu}_{T_n} = \mathbf{c}'\hat{\boldsymbol{\mu}} = \sum_{i=1}^{n} c_i \hat{\mu}_{i:n} \tag{1.3}$$

and

$$\hat{\sigma}^2_{T_n} = \mathbf{c}'\hat{\boldsymbol{\Sigma}}\mathbf{c} = \sum_{i=1}^{n} c_i^2 \hat{\sigma}^2_{i:n} + 2\sum_{i<j}\sum c_i c_j \hat{\sigma}_{ij:n} \tag{1.4}$$

respectively.

The details on how to calculate the elements of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ directly without resampling are provided in Section 2 and Section 3 respectively. In Section 4 we offer some concluding remarks and discuss some possible generalizations and applications of the results given here.

2. Bootstrap mean

Given that the sample order statistic $\mathbf{X} = (X_{1:n}, X_{2:n}, \ldots, X_{n:n})$ is held fixed, then a non-parametric bootstrap replication is generated by taking a sample of size n with replacement from $\mathbf{X}$. This is equivalent to generating a random sample of size n from a uniform(0, 1) distribution, with corresponding order statistic $\mathbf{U} = (U_{1:n}, U_{2:n}, \ldots, U_{n:n})$, and applying the sample quantile function

$$\hat{Q}(u) = \hat{F}^{-1}(u) = X_{[nu]+1:n} \tag{2.1}$$

to each element of $\mathbf{U}$, where $0 < u < 1$ and $[\cdot]$ denotes the floor function, i.e. $\hat{Q}(u) = X_{i:n}$ in the region given by $(i-1)/n \leqslant u < i/n$, $i = 1, 2, \ldots, n$. It follows that the individual elements of $\hat{\boldsymbol{\mu}}$ are given in the following theorem.

Theorem 1. The exact bootstrap estimate of $\mu_{r:n}$, $1 \leqslant r \leqslant n$, is

$$\hat{\mu}_{r:n} = E_{\hat{Q}}(X_{r:n}) = \sum_{j=1}^{n} w_{j(r)} X_{j:n}, \tag{2.2}$$

where

$$w_{j(r)} = r\binom{n}{r}\left\{B\left(\frac{j}{n}; r, n-r+1\right) - B\left(\frac{j-1}{n}; r, n-r+1\right)\right\}, \tag{2.3}$$

and

$$B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$$

is the incomplete beta function.

Proof. The deriviation of equation (2.2) follows by noting that

$$\begin{aligned}
\mu_{r:n} = E(X_{r:n}) &= \int_{-\infty}^{\infty} x_r f_r(x_r)\,\mathrm{d}x_r \\
&= \int_0^1 r\binom{n}{r} Q(u_r)u_r^{r-1}(1-u_r)^{n-r}\,\mathrm{d}u_r \\
&= r\binom{n}{r}\sum_{j=1}^{n}\int_{(j-1)/n}^{j/n} Q(u_r)u_r^{r-1}(1-u_r)^{n-r}\,\mathrm{d}u_r.
\end{aligned}$$

Substituting the sample quantile function $\hat{Q}(u_r) = X_{[nu_r]+1:n}$ for $Q(u_r)$ and conditioning on the data yields equation (2.2). □

$\hat{\mu}_{r:n}$ is the Harrell–Davis kernel quantile function estimator evaluated at $r/(n+1)$; see Sheather and Marron (1990).

It follows from equation (1.3) and theorem 1 that the exact bootstrap mean of T_n is given by

$$\hat{\mu}_{T_n} = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i w_{j(i)} X_{j:n}. \tag{2.4}$$

Remark 1. It is interesting and important theoretically to note that $\hat{\mu}_{T_n}$ is also an L-estimator of the form $\Sigma_{j=1}^{n} k_j X_{j:n}$, where $k_j = \Sigma_{i=1}^{n} c_i w_{j(i)}$. If $c_i \geqslant 0$, $i = 1, 2, \ldots, n$ (with at least one inequality being strict), as is the case with many location estimators, then $k_j > 0$, $j = 1, 2, \ldots, n$, i.e. $\hat{\mu}_{T_n}$ is a weighted sum of all n order statistics. Therefore, the existence of the rth moment of T_n does not necessarily imply the existence of the rth moment of $\hat{\mu}_{T_n}$.

3. Bootstrap variances and covariances

The elements of the bootstrap covariance matrix $\hat{\Sigma}$ at expression (1.2) are given by the following two theorems.

Theorem 2. The exact bootstrap estimate of $\sigma_{r:n}^2$ is given by

$$\hat{\sigma}_{r:n}^2 = \mathrm{var}_{\hat{Q}}(X_{r:n}) = \sum_{j=1}^{n} w_{j(r)} (X_{j:n} - \hat{\mu}_{r:n})^2, \tag{3.1}$$

where $w_{j(r)}$ is defined at expression (2.3).

Proof. The proof is similar to that for theorem 1. □

Maritz and Jarrett (1978) and Efron (1979) independently proposed the specific case of equation (3.1) for the median from an odd-sized sample.

Theorem 3. The exact bootstrap estimate of $\sigma_{rs:n}$ for $r < s$ is given by

$$\hat{\sigma}_{rs:n} = \mathrm{cov}_{\hat{Q}}(X_{r:n}, X_{s:n}) = \sum_{j=2}^{n} \sum_{i=1}^{j-1} w_{ij(rs)} (X_{i:n} - \hat{\mu}_{r:n})(X_{j:n} - \hat{\mu}_{s:n})$$
$$+ \sum_{j=1}^{n} v_{j(rs)} (X_{j:n} - \hat{\mu}_{r:n})(X_{j:n} - \hat{\mu}_{s:n}), \tag{3.2}$$

where the weights are given by

$$w_{ij(rs)} = \int_{(j-1)/n}^{j/n} \int_{(i-1)/n}^{i/n} f_{rs}(u_r, u_s) \, \mathrm{d}u_r \, \mathrm{d}u_s, \tag{3.3}$$

$$v_{j(rs)} = \int_{(j-1)/n}^{j/n} \int_{(j-1)/n}^{u_s} f_{rs}(u_r, u_s) \, \mathrm{d}u_r \, \mathrm{d}u_s, \tag{3.4}$$

and

$$f_{rs}(u_r, u_s) = {}_nC_{rs} u_r^{r-1} (u_s - u_r)^{s-r-1} (1 - u_s)^{n-s} \tag{3.5}$$

is the joint distribution of two uniform order statistics $U_{r:n}$ and $U_{s:n}$ with ${}_nC_{rs} = n!/(r-1)!\,(s-r-1)!\,(n-s)!$.

Proof. The expression for $\hat{\sigma}_{rs:n}$ follows directly from observing that

$$\begin{aligned}\sigma_{rs:n} &= E\{(X_{r:n}-\mu_{r:n})(X_{s:n}-\mu_{s:n})\} \\ &= \int_0^1 \int_0^{u_s} \{Q(u_r)-\mu_{r:n}\}\{Q(u_s)-\mu_{s:n}\} f_{rs}(u_r, u_s)\,\mathrm{d}u_r\,\mathrm{d}u_s \\ &= \sum_{j=2}^{n}\sum_{i=1}^{j-1} \int_{(j-1)/n}^{j/n} \int_{(i-1)/n}^{i/n} \{Q(u_r)-\mu_{r:n}\}\{Q(u_s)-\mu_{s:n}\} f_{rs}(u_r, u_s)\,\mathrm{d}u_r\,\mathrm{d}u_s \\ &\quad + \sum_{j=1}^{n} \int_{(j-1)/n}^{j/n} \int_{(j-1)/n}^{u_s} \{Q(u_r)-\mu_{r:n}\}\{Q(u_s)-\mu_{s:n}\} f_{rs}(u_r, u_s)\,\mathrm{d}u_r\,\mathrm{d}u_s. \qquad (3.6)\end{aligned}$$

Substituting $\hat{Q}$, $\hat{\mu}_{r:n}$ and $\hat{\mu}_{s:n}$ for Q, $\mu_{r:n}$ and $\mu_{s:n}$ respectively and noting that $\hat{Q}(u) = X_{[nu]+1:n}$ is a constant over each region of integration results in equation (3.2). □

The weights $w_{ij(rs)}$ and $v_{j(rs)}$ in equations (3.3) and (3.4) can be easily calculated by evaluating the integrals and writing them in closed form. The key is to note that the binomial series expansion of $(u_s-u_r)^{s-r-1}$ in equation (3.5) is

$$(u_s-u_r)^{s-r-1} = \sum_{k=0}^{s-r-1} \binom{s-r-1}{k} (-1)^{s-r-1-k} u_s^k u_r^{s-r-1-k}.$$

Then, $f_{rs}(u_r, u_s)$ can be written as

$$f_{rs}(u_r, u_s) = {}_nC_{rs} \sum_{k=0}^{s-r-1} \binom{s-r-1}{k} (-1)^{s-r-1-k} u_r^{s-k-2} u_s^k (1-u_s)^{n-s}$$

which is easily integrated. Using this expression in equation (3.3) results in

$$\begin{aligned} w_{ij(rs)} &= {}_nC_{rs} \sum_{k=0}^{s-r-1} \binom{s-r-1}{k} (-1)^{s-r-1-k} \int_{(i-1)/n}^{i/n} u_r^{s-k-2}\,\mathrm{d}u_r \int_{(j-1)/n}^{j/n} u_s^k (1-u_s)^{n-s}\,\mathrm{d}u_s \\ &= {}_nC_{rs} \sum_{k=0}^{s-r-1} \binom{s-r-1}{k} \frac{(-1)^{s-r-1-k}}{s-k-1} \left\{ \left(\frac{i}{n}\right)^{s-k-1} - \left(\frac{i-1}{n}\right)^{s-k-1} \right\} \\ &\quad \times \left\{ B\left(\frac{j}{n}; k+1, n-s+1\right) - B\left(\frac{j-1}{n}; k+1, n-s+1\right) \right\}, \end{aligned}$$

which can be readily calculated by using the incomplete beta function

$$B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t.$$

Applying the same technique to equation (3.4) leads to

$$\begin{aligned} v_{j(rs)} = {}_nC_{rs} \sum_{k=0}^{s-r-1} \binom{s-r-1}{k} \frac{(-1)^{s-r-1-k}}{s-k-1} \Bigg[& B\left(\frac{j}{n}; s, n-s+1\right) - B\left(\frac{j-1}{n}; s, n-s+1\right) \\ & - \left(\frac{j-1}{n}\right)^{s-k-1} \left\{ B\left(\frac{j}{n}; k+1, n-s+1\right) - B\left(\frac{j-1}{n}; k+1, n-s+1\right) \right\} \Bigg]. \end{aligned}$$

Therefore, the exact bootstrap variance of any L-estimator, given by equation (1.4), can be calculated directly from the variances and covariances given in equations (3.1) and (3.2) respectively.

4. Concluding remarks

Some researchers suggest the use of the *symmetric bootstrap* if the assumption that the data come from a symmetric population seems reasonable. The exact computations discussed here can be carried out under the symmetric bootstrap by replacing $\hat{Q}(u)$ in equation (2.1) with

$$\hat{Q}_{\text{sym}}(u) = T_n + (X_{[nu]+1:n} - X_{[n(1-u)]+1:n})/2.$$

Whereas we have provided a method for calculating the exact bootstrap mean and variance of an L-estimator through the elimination of the bootstrap resampling error, we have not addressed the statistical error of the estimators. As an alternative to the approach of Silverman and Young (1987) to bootstrapping through the use of smoothed empirical estimators of the cumulative distribution function F, we conjecture that by utilizing more efficient and smooth estimators of the quantile function Q in the variance expressions in Section 3 it will be possible similarly to derive new exact bootstrap variance estimators for T_n, which will benefit from improved efficiency over equation (1.4).

Acknowledgements

Alan Hutson's work was supported in part by National Institutes of Health General Clinical Research Center grant RR0082. The authors gratefully acknowledge the helpful comments and suggestions of the Joint Editor and two referees.

References

David, H. A. (1981) *Order Statistics*, 2nd edn. New York: Wiley.
Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1–26.
Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.
Huang, J. S. (1991) Efficient computation of the performance of bootstrap and jackknife estimators of the variance of L-statistics. *J. Statist. Computn Simuln*, **38**, 45–56.
Lepage, R. and Billard, L. (eds) (1992) *Exploring the Limits of Bootstrap*. New York: Wiley.
Maritz, J. S. and Jarrett, R. G. (1978) A note on estimating the variance of the sample median. *J. Am. Statist. Ass.*, **73**, 194–196.
Parr, W. C. and Schucany, W. R. (1982) Jackknifing L-statistics with smooth weight functions. *J. Am. Statist. Ass.*, **77**, 629–638.
Reiss, R.-D. (1989) *Approximate Distributions of Order Statistics: with Applications to Nonparametric Statistics*. New York: Springer.
Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. New York: Wiley.
Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. New York: Springer.
Sheather, S. J. and Marron, J. S. (1990) Kernel quantile estimators. *J. Am. Statist. Ass.*, **85**, 410–416.
Silverman, B. W. and Young, G. A. (1987) The bootstrap: to smooth or not to smooth? *Biometrika*, **74**, 469–479.