

Applications of the Exact Bootstrap

Michael D. Ernst

Alan D. Hutson*

Indiana University–Purdue University Indianapolis

University of Florida

Abstract

For the vast majority of statistics, the only recourse for calculating nonparametric bootstrap percentiles is through the use of resampling methods, leading to the common misconception that the bootstrap is strictly a simulation technique. A well-known exception to this is that bootstrap percentiles of any order statistic may be obtained analytically. In this note we show that the nonparametric bootstrap percentiles and moment estimates of any function of the order statistics from an i.i.d. sample may be obtained analytically through the reformulation of the resampling problem as a minimization problem, thereby eliminating the simulation error.

KEY WORDS: L -estimator; L -statistic; median; quantile function; quick estimator; trimmed mean.

1 Introduction

Let $\mathbf{X} = (X_{1:n}, X_{2:n}, \dots, X_{n:n})$ denote the sample order statistic of an i.i.d. sample of size n from a continuous distribution F . In this note we provide a minimization algorithm

*Michael D. Ernst is Assistant Professor, Department of Mathematical Sciences, Indiana University–Purdue University Indianapolis, 402 North Blackford Street, LD 270, Indianapolis, IN 46202-3216. Alan D. Hutson is Research Assistant Professor, University of Florida, Division of Biostatistics, Department of Statistics, Health Science Center, P.O. Box 100212, Gainesville, FL 32610-0212. Alan Hutson’s work was supported in part by NIH General Clinical Research Center Grant RR0082.

for obtaining the *exact* nonparametric bootstrap percentiles for any statistic

$$T_n = T_n(\mathbf{X}) = T(X_{i_1:n}, X_{i_2:n}, \dots, X_{i_k:n}), \quad (1.1)$$

that is a function of a subset of order statistics from $(X_{1:n}, X_{2:n}, \dots, X_{n:n})$, where $1 \leq i_1 < i_2 < \dots < i_k \leq n$ and $1 \leq k \leq n$. Specific examples of T_n include L -estimators of the form $T_n = \sum_{i=1}^n c_i X_{i:n}$, products of order statistics such the geometric mean written as $T_n = (\prod_{i=1}^n X_{i:n})^{1/n}$, functions of the sample spacings of the form $T_n = \sum_{i=1}^{n-m} \Phi(X_{i+m:n} - X_{i:n})$ (Van Es 1992), and generalized L -estimators (Serfling 1984). Expressions for the bootstrap mean and variance are readily available in closed form for the specific T_n such as the mean, the median (odd sample size case) and quartiles (when the quantile function is defined by a single order statistic), e.g. see Shao and Tu (1995). Huang (1991) and Hutson and Ernst (2000) define two alternative methods for calculating the exact bootstrap mean and variance of any L -estimator. Historically, resampling methods have been used in lieu of an analytical solution for estimating bootstrap percentiles. To the best of our knowledge the only case where the exact percentiles have been demonstrated to have analytical solutions is the the case where $T_n = X_{i:n}$ is exactly equal to a specific order statistic, e.g. see David (1981). Recently, Hutson (1999) has generalized this result for any nonextreme sample quantile estimated using the linear interpolation of adjacent order statistics.

The method we will present for calculating the exact bootstrap percentiles for any statistic T_n defined by (1.1) follows from a version of the definition of the α -quantile for the random variable X given by the value of θ satisfying

$$\inf_{\theta \in \mathcal{R}} E \left\{ \frac{|X - \theta| + (2\alpha - 1)(X - \theta)}{2} - \frac{|X| + (2\alpha - 1)X}{2} \right\}. \quad (1.2)$$

The properties of (1.2) are outlined in Abdous and Theodorescu (1992) who generalize the definition of the α -quantile to \mathcal{R}^k space, $k \geq 1$. Most notably, the assumption that $E(X) < \infty$, $\alpha \in (0, 1)$, in (1.2) is not a necessary condition for the existence of the α -quantile.

The 100α bootstrap percentile of T_n follows straightforward from (1.2) and is given by the value of θ satisfying the bootstrap estimate of

$$\inf_{\theta \in \mathcal{R}} E \left\{ \frac{|T_n - \theta| + (2\alpha - 1)(T_n - \theta)}{2} - \frac{|T_n| + (2\alpha - 1)T_n}{2} \right\}. \quad (1.3)$$

Analytic expressions for the bootstrap estimators of the k th moment, $E(T_n^k)$, may be obtained using the same technique. In Section 2, we derive the exact bootstrap percentiles of T_n using expression (1.3) in conjunction with a specific quantile function estimator, $\hat{Q}(u) = \hat{F}^{-1}(u)$, which is defined to correspond to sampling with replacement from the order statistic, i.e. the standard bootstrap resampling scheme. In Section 3, we illustrate the exact percentiles using the sample median, Tukey's trimean, and the interquartile range and compare this to the traditional resampling approach.

2 Bootstrap Percentiles

In this section we detail the steps for calculating the nonparametric bootstrap percentiles of T_n . The 100α bootstrap percentile of T_n will be denoted by the quantile function estimator $\hat{Q}_{T_n}(\alpha)$, $\alpha \in (0, 1)$, and is obtained through the bootstrap estimate of the linear component of (1.3) given by

$$E \{ |T_n - \theta| + (2\alpha - 1)(T_n - \theta) \}, \quad (2.1)$$

and then minimizing the estimate of (2.1) with respect to θ . The empirical bootstrap estimates of the other linear components of (1.3) do not factor into the minimization/estimation procedure and can be ignored.

Traditionally, the nonparametric estimate $\hat{Q}_{T_n}(\alpha)$ is estimated by ordering the values of the B bootstrap replicates, T_n^* , obtained by sampling with replacement from the sample order statistic. Then, $\hat{Q}_{T_n}(\alpha)$ is estimated as the $([B\alpha] + 1)$ th ordered value of T_n^* , where $[\cdot]$ denotes the floor function, e.g. see Efron and Tibshirani (1993). The limiting case ($B \rightarrow \infty$) is identical to substituting $\hat{Q}(u)$ for $Q(u)$ in (2.1) and minimizing with respect

to θ , where

$$\hat{Q}(u) = \hat{F}^{-1}(u) = X_{[nu]+1:n}. \quad (2.2)$$

This method is often times colorfully referred to as the “plug-in” method when $\hat{F}(x)$ is substituted for $F(x)$.

The bootstrap percentile estimation procedure corresponding to expression (2.1) consists of two components; the bootstrap estimation of (2.1) followed by its minimization with respect to θ . First we have,

Theorem 2.1. The Estimation. Let

$$g(u_1, u_2, \dots, u_k) = n! \prod_{j=0}^k \frac{(u_{j+1} - u_j)^{i_{j+1} - i_j - 1}}{(i_{j+1} - i_j - 1)!} \quad (2.3)$$

be the joint density of $U_{i_1:n}, U_{i_2:n}, \dots, U_{i_k:n}$; a subset of k order statistics from a sample of size n from a uniform(0, 1) distribution, where $u_{k+1} = 1$, $u_0 = 0$, $i_{k+1} = n + 1$ and $i_0 = 0$.

The exact bootstrap estimate of (2.1) is given by

$$\sum_{i_k=1}^n \sum_{i_{k-1}=1}^{i_k} \dots \sum_{i_1=1}^{i_2} \{|T_n(\mathbf{X}) - \theta| + (2\alpha - 1)[T_n(\mathbf{X}) - \theta]\} C(i_1, \dots, i_{k-1}, i_k), \quad (2.4)$$

where

$$C(i_1, \dots, i_{k-1}, i_k) = \int_{\frac{i_{k-1}}{n}}^{\frac{i_k}{n}} \int_{\frac{i_{k-2}}{n}}^{I_{i_{k-1}}} \dots \int_{\frac{i_1-1}{n}}^{I_{i_1}} g(u_1, \dots, u_{k-1}, u_k) du_1 \dots du_{k-1} du_k, \quad (2.5)$$

and

$$I_{i_j} = \begin{cases} u_{j+1} & \text{if } i_j = i_{j+1} \\ i_j/n & \text{otherwise.} \end{cases} \quad (2.6)$$

Proof. The expression (2.4) follows from the re-expression of (2.1) using the quantile function $Q(u)$. Specifically, if h is the joint p.d.f. of $X_{i_1:n}, X_{i_2:n}, \dots, X_{i_k:n}$, then

$$\begin{aligned} & E\{|T_n - \theta| + (2\alpha - 1)(T_n - \theta)\} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{x_k} \dots \int_{-\infty}^{x_2} |T_n - \theta| + (2\alpha - 1)[T_n - \theta] h(x_1, \dots, x_k) dx_1 \dots dx_{k-1} dx_k \end{aligned}$$

$$\begin{aligned}
&= \int_0^1 \int_0^{u_k} \cdots \int_0^{u_2} \left\{ |T(Q(u_{i_1:n}), \dots, Q(u_{i_{k-1}:n}), Q(u_{i_k:n})) - \theta| \right. \\
&\quad \left. + (2\alpha - 1)[T(Q(u_{i_1:n}), \dots, Q(u_{i_{k-1}:n}), Q(u_{i_k:n})) - \theta] \right\} \\
&\quad \times g(u_1, u_2, \dots, u_k) du_1 \cdots du_{k-1} du_k.
\end{aligned} \tag{2.7}$$

The theorem follows by substituting $\hat{Q}(u)$ given by (2.2) into (2.7) and noting that $\hat{Q}(u)$ is constant in the region given by $(i-1)/n \leq u < i/n$, $i = 1, 2, \dots, n$. \square

For the specific case of L -estimators, $T_n = \sum_{i=1}^n c_i X_{i:n}$, the exact bootstrap estimate of the second component of (2.7), namely $E[(2\alpha - 1)(T_n - \theta)]$, reduces to

$$(2\alpha - 1)(\hat{\mu}_{T_n} - \theta), \tag{2.8}$$

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

$$w_{j(i)} = i \binom{n}{i} \left[B\left(\frac{j}{n}; i, n - i + 1\right) - B\left(\frac{j-1}{n}; i, n - i + 1\right) \right], \tag{2.9}$$

and $B(x; a, b) = \int_0^x t^{a-1} (1-t)^{b-1} dt$ is the incomplete beta function. See Hutson and Ernst (2000) for a complete description.

Definition 2.2. The Minimization. The *exact* 100α bootstrap percentile of T_n , $\hat{Q}_{T_n}(\alpha)$, is the value of θ that minimizes (2.4). Note that the component $(|T_n| + (2\alpha - 1)T_n)/2$ in (1.3) is a constant in the bootstrap estimation/minimization procedure and can be ignored when solving for θ . It is straightforward yet cumbersome to carry out the above calculations using standard software packages such as Mathematica (1996). We are currently working toward an efficient routine that will reduce the computing time needed to carry out the calculations.

Remark 2.3. Moment Estimators. The exact bootstrap estimates of $E(T_n^k)$ are given directly by

$$\sum_{i_k=1}^n \sum_{i_{k-1}=1}^{i_k} \cdots \sum_{i_1=1}^{i_2} T_n^k(\mathbf{X}) C(i_1, \dots, i_{k-1}, i_k), \tag{2.10}$$

where $C(i_1, \dots, i_{k-1}, i_k)$ is defined by (2.5).

Three Special Cases. We give the expressions needed to obtain the exact 100α bootstrap percentiles, $\hat{Q}_{T_n}(\alpha)$, for the mean, median, and trimmed mean below. It is

obvious the calculations become infeasible very quickly for the mean and certain trimmed means as a function of n and therefore Monte Carlo methods are recommended. The utility of the procedure is for obtaining exact percentile intervals for statistics involving a few order statistics such as the median.

1. *Mean.* The exact 100α bootstrap percentile of the mean is given by the value of θ satisfying

$$\inf_{\theta \in \mathcal{R}} \left\{ \sum_{i_n=1}^n \sum_{i_{n-1}=1}^{i_n} \cdots \sum_{i_1=1}^{i_2} \left| \frac{1}{n} \sum_{j=1}^n x_{i_j:n} - \theta \right| \right. \\ \left. \times \int_{\frac{i_n-1}{n}}^{\frac{i_n}{n}} \int_{\frac{i_{n-1}-1}{n}}^{I_{i_{n-1}}} \cdots \int_{\frac{i_1-1}{n}}^{I_{i_1}} n! du_1 \cdots du_{n-1} du_n + (2\alpha - 1)(\bar{x} - \theta) \right\}.$$

2. *Median ($n = 2r - 1$ odd).* The exact 100α bootstrap percentile of the median is given by the value of θ satisfying

$$\inf_{\theta \in \mathcal{R}} \left\{ \sum_{j=1}^n |x_{j:n} - \theta| w_{j(r)} + (2\alpha - 1)(\hat{\mu}_{T_n} - \theta) \right\},$$

where $\hat{\mu}_{T_n}$ and $w_{j(r)}$ are given by (2.8) and (2.9), respectively. More specifically, $\hat{Q}_{T_n}(\alpha) = \hat{Q}(P(\alpha))$, where $P(\cdot)$ denotes the quantile function of a beta random variable with parameters r and $n - r + 1$, and $\hat{Q}(\cdot)$ is given by (2.2). See Hutson (1999) for alternative methods for constructing confidence intervals for quantiles.

3. *Trimmed Mean.* Let β be the trimming proportion and $m = [\beta n]$. The exact 100α bootstrap percentile of the β -trimmed mean is given by the value of θ satisfying

$$\inf_{\theta \in \mathcal{R}} \left\{ \sum_{i_{n-2m}=1}^n \sum_{i_{n-2m-1}=1}^{i_{n-2m}} \cdots \sum_{i_1=1}^{i_2} \left| \frac{1}{n-2m} \sum_{j=1}^{n-2m} x_{i_j:n} - \theta \right| \right. \\ \left. \times \int_{\frac{i_{n-2m}-1}{n}}^{\frac{i_{n-2m}}{n}} \int_{\frac{i_{n-2m-1}-1}{n}}^{I_{i_{n-2m-1}}} \cdots \int_{\frac{i_1-1}{n}}^{I_{i_1}} u_1^{m-1} (1 - u_{n-2m})^{n-2m} du_1 \cdots du_{n-2m-1} du_{n-2m} \right. \\ \left. + (2\alpha - 1)(\hat{\mu}_{T_n} - \theta) \right\},$$

where $\hat{\mu}_{T_n}$ is given by (2.8).

3 Example

In this section we use the data in Table 1 from Caudill *et al.* (1998) to illustrate the behavior of the bootstrap percentiles based on resampling, compared to the exact bootstrap percentiles given by Definition 2.2. These data are measurements of the urinary folate catabolites, acetamidobenzolyglutamate (apABG), used to assess folate requirements in both pregnant and nonpregnant women. Table 2 shows the exact 2.5% and 97.5% bootstrap percentiles for the median, trimean, and interquartile range (IQR) for the data in Table 1. The resample approximations of these percentiles are also given based on $B = 100, 500,$ and 1000 resamples. We can see from Table 2 that the lower and upper terminals of the simulated percentile intervals vary quite a bit over the replications sizes and that it is not quite clear when convergence is obtained.

Table 1: Patient’s baseline urinary apABG(nmol/d) ($n = 24$).

67.9	7.1	14.0	10.9	3.1	8.5	646.3	0.5	6.2	9.4	10.3	4.9
136.0	138.5	297.7	184.3	10.6	433.5	275.7	3.3	230.8	12.0	7.8	21.4

Table 2: Exact and approximate 95% bootstrap percentile intervals.

Statistic	T_n	Exact	$B = 100$	$B = 500$	$B = 1000$
Median	11.45	(8.50, 136.00)	(9.40, 136.00)	(8.15, 136.00)	(8.50, 136.00)
Trimean*	54.03	(10.60, 144.38)	(9.75, 139.58)	(11.33, 140.75)	(10.43, 144.20)
IQR**	176.50	(9.10, 289.91)	(6.20, 289.90)	(9.10, 289.90)	(7.80, 290.60)

*Tukey’s Trimean defined as $\hat{Q}(1/4)/4 + \hat{Q}(1/2)/2 + \hat{Q}(3/4)/4$.

**Interquartile Range defined as $\hat{Q}(3/4) - \hat{Q}(1/4)$.

4 Summary

This note demonstrates that nonparametric bootstrap moment and percentile estimates can be obtained for a large class of statistics without resorting to resampling. The method has utility for statistics involving a function of a few order statistics such as the median and quartiles, the interquartile range and quick estimators of location. The advantages of the minimization procedure are as follows:

1. Reduction in computational errors(simulation versus minimization).
2. Reproducibility.
3. Straightforward extensions to two-sample problems.

References

- Abdous, B. and Theodorescu, R. (1992). “Note on the spatial quantile of a random vector,” *Statistics & Probability Letters*, **13**, 333–336.
- Caudill, M. A., Gregory, J. F., Hutson, A. D., and Bailey, L. B. (1998). “Folate Catabolism in Pregnant and Nonpregnant Women Consuming Controlled Folate Intakes,” *Journal of Nutrition*, **128**, 204–208.
- David, H. A. (1981). *Order Statistics*, 2nd Ed., New York: John Wiley.
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, New York: Chapman & Hall.
- Huang, J. S. (1991). “Efficient computation of the performance of bootstrap and jack-knife estimators of the variance of L -statistics,” *Journal of Statistical Computation and Simulation*, **38**, 45–56.
- Hutson, A. D. (1999). “Calculating Nonparametric Confidence Intervals for Quantiles Using Imaginary Order Statistics,” *Journal of Applied Statistics*, **26**, 339-349.

Hutson, A. D. and Ernst, M. D. (2000). “The Exact Bootstrap Mean and Variance of an L -estimator,” *Journal of the Royal Statistical Society–Series B*, **62**, 89–94.

Mathematica 3.0 (1996). Wolfram Research, Inc., Champaign, IL.

Serfling, R. J. (1984). “Generalized L-, M-, and R-Statistics,” *The Annals of Statistics*, **12**, 76–86.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, New York: Springer-Verlag.

Van Es, B. (1992). “Estimating Functionals Related to a Density by a Class of Statistics Based on Spacings,” *Scandinavian Journal of Statistics*, **19**, 61–72.