NONPARAMETRIC BOOTSTRAP CONFIDENCE INTERVAL FOR MEDIAN

Bambang Suprihatin, Suryo Guritno, and Sri Haryatmi

Mathematics Department, University of Sriwijaya, Palembang, Indonesia
E-mail: bambangs@unsri.ac.id

Abstract. Given sample $X=(X_1,X_2,\ldots,X_n)$ of size *n* from an unknown distribution *F*. If all elements of *X* are distinct, then the number of different possible resamples X^* with replacement equals n^n. In general, this number obvious very large in amount. For *n* = 10, think of the number 10^{10}, which is an enormous number. Let $\hat{\theta}^*=t(X_1^*,X_2^*,\ldots,X_n^*)$ be the estimate value of statistic computed from X^*, where *t* is functional. In most cases of practical interest, each distinct X^* (without regard for order), gives rise to a distinct $\hat{\theta}^*$. Accordingly, we concern only on so-called *atoms* of nonparametric bootstrap. The number of atoms is far less than n^n. Based on these atoms, the nonparametric bootstrap used to estimate a statistic computed from *X*. This paper presents how to find the number of atoms. The implementation of the uses of atoms is applied in bootstrapping bias estimate of sample median. Bootstrap version of standar error as a measure of accuracy of estimator is considered, as well. The main purpose of this paper is to construct a confidence interval for median. Results from Monte Carlo simulation for these cases are also presented.

Keywords and phrases: Nonparametric bootstrap, atoms, bias, sample median, confidence interval, Monte Carlo simulation.

1. INTRODUCTION

A typical problem in applied statistics involves the estimation of the unknown parameter θ. The two main questions asked are: (1) what estimator $\hat{\theta}$ should be used or choosen? (2) Having choosen to use particular $\hat{\theta}$, how accurate is it as an estimator of θ? The bootstrap is a general methodology for answering the second question. We use the standard error and confidence interval as measures of statistical accuracy. This paper deal with confidence interval for the population median based on atoms of nonparametric bootstrap. We will investigate how to find the number of atoms and discuss on finding the bootstrap estimate for standard error of the sample median.

2010 Mathematics Subject Classification: 62F40

The sample median and its estimate of standard error play important role in constructing a confidence interval for population median. Main purpose of this paper is to construct a confidence interval for population median based on atoms of nonparametric bootstrap. For small sample, Maritz and Jarrett [13] gave a good approximation for variance of the sample median. However, for larger sample, we handle it by using Monte Carlo simulation for producing good approximation of boootstrap standard error. The bootstrap is then extended to other measures of statistical accuracy such as estimates of bias and confidence interval. Suprihatin *et.al* (2012a, 2012b) showed that the bootstrap work well for estimating the statistics mean and parameter of autoregressive model respectively.

We describe algoritms for constructing a confidence interval for population median. Section 2 reviews how to find the the number of atoms of nonparametric bootstrap, and discuss the implementation of the uses of atoms is applied in bootstrapping bias estimate of sample median. Section 3 describes the bootstrap estimate for standard error of sample median. Section 4 deal with confidence intervals and explores the results of Monte Carlo simulation for bootstrap estimates of standard error and confidence interval for median. Section 5, is the last section, briefly describes summary of this paper as concluding remarks.

2. THE ATOMS OF NONPARAMETRIC BOOTSTRAP

Let X^* denotes a same-size resample drawn, with replacement, from a given sample $X=(X_1,X_2,..,X_n)$, and let $\hat{\theta}^*=t(X_1^*,X_2^*,\ldots,X_n^*)$ be the estimate value of statistic computed from X^*, where t is functional. In most cases of practical interest, each distinct X^* (without regard for order) gives rise to a distinct $\hat{\theta}^*$, as Hall [11] has elaborated it.

If the sample X is of size n, and if all elements of X are distinct, the number of different possible resamples X^* equals n^n. But, if without regard for order, then the number of different possible resamples X^* equals the number $N(n)$, of distinct ways of placing n indistinguishable objects into n numbered boxes, the boxes being allowed to contain any number of objects. Accordingly, we concern only on so-called *atoms* of nonparametric bootstrap. The number of atoms is far less than n^n.

To derive the number $N(n)$, let m_i denote the number of times X_i is repeated in X^*. The number $N(n)$ equals the number of different ways of choosing the ordered *n*-vector $(m_1, m_2, \ldots, m_n)$ such that each $m_i \geq 0$ and $m_1 + m_2 + \ldots + m_n = n$. We imagine that m_i as the number of objects in box *i*.

Calculation of $N(n)$ is a problem in combinatorial. For this purpose, we start with sample size *n* = 2. Meantime, for *n* = 1 is trivial. Let $X = (X_1, X_2)$, then the atoms are: $(X_1, X_1), (X_1, X_2)$ and $(X_2, X_2) = X$, and yields $N(2) = 3$. For sample size *n* = 3, let $X = (X_1, X_2, X_3)$, then the atoms are: (X_1, X_1, X_1), (X_1, X_1, X_2), (X_1, X_1, X_3), (X_1, X_2, X_2), (X_1, X_3, X_3), $(X_2, X_2, X_2), (X_2, X_2, X_3)$, $(X_2, X_3, X_3), (X_3, X_3, X_3)$ and (X_1, X_2, X_3) = *X*, yields $N(3) = 10$. Finding the numbers $N(2)$ and $N(3)$ can be described as follows. We can check that $N(2) = 1 \cdot 2 + 1 \cdot 1 = 3$ and $N(3) = 1 \cdot 3 + 2 \cdot 3 + 1 \cdot 1 = 10$. Analogy to these calculations, for *n* = 4, we obtain $N(4) = 1 \cdot 4 + 3 \cdot 6 + 3 \cdot 4 + 1 \cdot 1 = 35$. For general *n*, by inductively, we conclude that

$$N(n) = \binom{n-1}{0}\binom{n}{1} + \binom{n-1}{1}\binom{n}{2} + \cdots + \binom{n-1}{n-1}\binom{n}{n}$$

$$= \sum_{i=1}^{n} \binom{n-1}{i-1}\binom{n}{i} \qquad (1)$$

Fisher and Hall [9] showed that the number of atoms $N(n)$ equals $\binom{2n-1}{n}$. This formula looks simpler, but difficult in proving. Thus, in order to show that (1) is actually the number of atoms $N(n)$, it suffices to show, by using mathematical induction principle, that

$$\sum_{i=1}^{n} \binom{n-1}{i-1}\binom{n}{i} = \binom{2n-1}{n}.$$

Not all of the atoms of the bootstrap distribution $\hat{\theta}^*$ have equal probability mass. To compute probabilities, let $X^*(m_1, m_2, \ldots, m_n)$ denotes the resample drawn from X in which X_i out of *X* on any (with replacement) result in $X^*(m_1, m_2, \ldots, m_n)$ equals the multinomial probability

$$\binom{n}{m_1, m_2, \ldots, m_n}\left(n^{-1}\right)^{m_1}\left(n^{-1}\right)^{m_2}\cdots\left(n^{-1}\right)^{m_n} = \frac{n!}{n^n m_1! m_2! \cdots m_n!}.$$

Thus, if $\hat{\theta}^*(m_1, m_2, \ldots, m_n)$ denotes the value of the statistic $\hat{\theta}^*$ when the resample is $X^*(m_1, m_2, \ldots, m_n)$ then

$$P\left(\hat{\theta}^* = \hat{\theta}^*(m_1, m_2, \ldots, m_n) \middle| X\right) = \frac{n!}{n^n m_1! m_2! \cdots m_n!}. \tag{2}$$

Here is an example of the uses of the atoms of nonparametric bootstrap, which can also be found in Lehmann [12]. Let θ be the median of distribution *F*, the bias of the sample median is

$$bias = E(\hat{\theta}) - \theta. \tag{3}$$

The counterpart of bias (3) is the bootstrap estimator for bias

$$bias^* = E(\hat{\theta}^*) - \hat{\theta}, \tag{4}$$

where (3) and (4) follow the bootstrap terminology that says the population is to the sample as the sample is to the bootstrap samples. Let us consider the case $n = 3$, and $X_{(1)} \le X_{(2)} \le X_{(3)}$ denote the order statistics. To obtain the probabilities for the corresponding ordered triples $\left(X^*_{(1)} \le X^*_{(2)} \le X^*_{(3)}\right)$, we must count the number of cases with these values. For instance, $P\left(X^*_{(1)} = X_1 \le X^*_{(2)} = X_1 \le X^*_{(3)} = X_2\right) = \frac{3}{27}$, since this probability is the sum of probabilities of the triples $(X_1, X_1, X_2), (X_1, X_2, X_1)$, (X_2, X_1, X_1) for $\left(X^*_{(1)}, X^*_{(2)}, X^*_{(3)}\right)$. By using (2), we obtain the distribution for atoms $\left(X^*_{(1)}, X^*_{(2)}, X^*_{(3)}\right)$ is

$X_1X_1X_1$	$X_1X_1X_2$	$X_1X_1X_3$	$X_1X_2X_2$	$X_1X_2X_3$
1/27	3/27	3/27	3/27	6/27
$X_1X_3X_3$	$X_2X_2X_2$	$X_2X_2X_3$	$X_2X_3X_3$	$X_3X_3X_3$
3/27	1/27	3/29	3/27	1/27

The median $X^*_{(2)}$ of $\left(X^*_{(1)}, X^*_{(2)}, X^*_{(3)}\right)$ is X_1 for the triples $X_1X_1X_1$, $X_1X_1X_2$, and $X_1X_1X_3$, and so on. Hence, the distribution of $X^*_{(2)}$ is

$$P\left(X^*_{(2)} = X_1\right) = \frac{7}{27}, P\left(X^*_{(2)} = X_2\right) = \frac{13}{27}, P\left(X^*_{(2)} = X_3\right) = \frac{7}{27}.$$

Therefore, the bootstrap estimator for the bias of $\hat{\theta}^* = X^*_{(2)}$ is by (4)

$$bias^* = E\left(X^*_{(2)}\right) - X_2$$

$$= \frac{7}{27}\cdot X_{(1)} + \frac{13}{27}\cdot X_{(2)} + \frac{7}{27}\cdot X_{(3)} - X_{(2)}$$

$$= \frac{14}{27}\left(\frac{X_{(1)} + X_{(3)}}{2} - X_{(2)}\right).$$

3. THE BOOTSTRAP ESTIMATE FOR STANDARD ERROR

Let we have a random sample $X = (X_1, X_2, \ldots, X_n)$ from distribution function F having a positive derivative f continous in a neighborhood of its median $\theta = \inf\left\{t \,|\, F(t) \geq \frac{1}{2}\right\}$. Let $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t)$ for all real t be the empirical distribution. Define the sample median as $\hat{\theta} = \inf\left\{t \,|\, F_n(t) \geq \frac{1}{2}\right\}$. Recall that $\hat{\theta}$ is the $(m+1)$th order statistic $X_{(m+1)}$ for n is odd, $n = 2m+1$ for positive integers m, and $\hat{\theta} = \frac{1}{2}\left(X_{(m)} + X_{(m+1)}\right)$ for n is even.

How accurate is $\hat{\theta}$ as an estimator for the actual θ? To answer this question, we use two measures of statistical accuracy, i.e. standard error and confidence interval. For small sample, Maritz and Jarrett [13] suggested a good approximation for the standard error of sample median. Consider the case $n = 2m + 1$. If $f_{(r)}$ denotes the pdf of $X_{(r)}$, from David and Nagaraja [2] we have

$$f_{(r)} = \frac{1}{B(r, n-r+1)} F^{r-1}(1 - F(x))^{n-r} f(x). \tag{5}$$

For $r = m + 1$, $B(r, n-r+1) = \frac{(m!)^2}{(2m+1)!}$ and denote $\hat{\theta} = \hat{X}_n$, by (5) we have

$$E(\hat{\theta}^r) = \frac{(2m+1)!}{(m!)^2}\int_{-\infty}^{\infty} x^r \left(F(x)(1 - F(x))\right)^m f(x)dx. \tag{6}$$

If we let $y = F(x)$, $x = F^{-1}(y) = \psi(y)$, then (6) becomes

$$E(\hat{\theta}^r) = \frac{(2m+1)!}{(m!)^2}\int_0^1 \psi(y)^r \left(y(1-y)\right)^m dy.$$

Maritz and Jarrett [13] estimated $E(\hat{\theta}^r)$ by $A_{rn} = \sum_{j=1}^{n} X_{(j)}{}^r W_j$, where

$$W_j = \frac{(2m+1)!}{(m!)^2} \int_{(j-1)/n}^{j/n} y^m (1-y)^m \, dy .$$

Then the value of $Var(\hat{\theta})$ is estimated by $V_n = A_{2n} - A_{1n}{}^2$. Hence, an estimate for standard error of $\hat{\theta}$ is

$$\widehat{se}(\hat{\theta}) = \sqrt{V_n / n} . \tag{7}$$

Meantime, the bootstrap is a tool for answering the accuracy of estimator, which is based on computer-intensive, even for handling larger sample. To proceed how the bootstrap works, consider n = 3. By using the distribution of $(X^*_{(1)}, X^*_{(2)}, X^*_{(3)})$ as we have discussed in Section 2, we have bootstrap estimator for variance of $X^*_{(2)}$ is

$$Var(X^*_{(2)}) = E(X^{*2}_{(2)}) - (X^*_{(2)})^2$$

$$= \frac{7}{27} \cdot X^2_{(1)} + \frac{13}{27} \cdot X^2_{(2)} + \frac{7}{27} \cdot X^2_{(3)} - \left(\frac{7}{27} \cdot X_{(1)} + \frac{13}{27} \cdot X_{(2)} + \frac{7}{27} \cdot X_{(3)} \right)^2 .$$

As customary, the bootstrap estimate for standard error of $X^*_{(2)}$ is square root of $Var(X^*_{(2)})$. But, this calculation is not sufficient for larger n. However, we still can do this calculation even for larger n using Monte Carlo simulation. For simulation theory, Efron and Tibshirani [8] is a good reference.

4. CONFIDENCE INTERVALS AND MONTE CARLO SIMULATION

So far, we have discussed the computation of bootstrap standard errors which are often used to construct approximate confidence intervals for a statistic of interest θ. Given an estimate $\hat{\theta}$ and an estimated standard error $\widehat{se}$, the standard $(1-2\alpha)\cdot 100\%$ confidence interval for θ is

$$(\hat{\theta} - z_\alpha \cdot \hat{se}, \hat{\theta} + z_{1-\alpha} \cdot \hat{se}), \tag{8}$$

where z_α is the $100 \cdot \alpha$th percentile point of standard normal distribution. For case θ is a population median, we use $\widehat{se}$ is Maritz and Jarrett's approximation of standard error for sample median. Consider the sample is 21.2, 22.5, 20.3, 21.4, 22.8, 21.6, 20.5, 21.3, 21.6. It's obvious that $\hat{\theta}$ = 21.40. Since n = 9, Maritz and Marrett [13] gave the weights

W_1 = 0.00145, W_2 = 0.02892, W_3 = 0.11447, W_4 = 0.22066, W_5 = 0.26899, and the rests can be found using the fact that $W_1 = W_n, W_2 = W_{n-2}, \ldots, W_m = W_{m-n+1}$. Then by (7) we obtain an estimate of standard error $\widehat{se}\left(\hat{\theta}\right)$ = 0.2903. Hence, using α = 5% we obtain the standard 90% confidence interval for θ is $(20.92, 21.88)$. DiCiccio and Tibshirani [5], and Davison and Hinkley [3] also reported a good bootstrap confidence interval.

Hall [10], DiCiccio and Efron [4], and Efron and Tibshirani [7] suggested to use the number of B larger than 1000 to construct an confidence interval. Using bootstrap sample size of B = 2000, Monte Carlo simulation gives mean of $\hat{\theta}^*$ = 21.44 compared with $\hat{\theta}$ = 21.40 and $\widehat{se}\left(\hat{\theta}^*\right)$ = 0.2925, which is closed to $\widehat{se}\left(\hat{\theta}\right)$. Moreover, resulting 90% bootstrap percentile confidence interval for θ is $(20.27, 21.60)$. Again, this result also closed to the 90% standard interval. This results agree to results of Efron [6], and Brown and Hall [1].

5. CONCLUDING REMARKS

A number of points arise from the consideration of Section 2, 3, and 4, amongst which we note as follows.

1. For small sample, it is often feasible to calculate a bootstrap estimate exactly, by computing all the atoms of the bootstrap samples. Unfortunately, this calculation is not sufficient for larger samples. However, we remain handle it by using Monte Carlo simulation.
2. Only using atoms, we obtain a good approximation of standard error and confidence interval for statistics of interest.
3. The largest probability of (2) is reached when each $m_i = 1$, in which case X^* is identical to X. Hence, the most likely resample to be drawn is the original sample, with probability $n!/n^n$. This probability is very small, being only $9.4x10^{-4}$ when n = 9 and decreases exponentially quickly. Thus, we infer that the probability that one or more repeats occur in the B values of $\hat{\theta}^*$ converges to zero as $n \to \infty$.

REFERENCES

[1] BROWN, B. M., HALL, P., AND YOUNG, G. A., The smoothed median and the bootstrap, *Biometrika*, **88**, 519-534, 2001.

[2] DAVID, H. A. AND NAGARAJA, H. N., *Order Statistics*, Wiley, New Jersey, 2003.

[3] DAVISON, A. C. AND HINKLEY, D. V., *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge, 2006.

[4] DICICCIO, T. J. AND EFRON, B., Bootstrap confidence intervals, *Statistical Science*, **11**, 198-228, 1996

[5] DICICCIO, T. J. AND TIBSHIRANI, R., Bootstrap confidence intervals and bootstrap approximations, *J. Amer. Statist. Ass.*, **82**, 163-170, 1987.

[6] EFRON, B., Better bootstrap confidence intervals, *J. Amer. Statist. Ass.*, **82**, 171-185, 1987.

[7] EFRON, B. AND TIBSHIRANI, R., Bootstrap methods for standard errors, confidence intervals, and others measures of statistical accuracy, *Statistical Science*, **1**, 54-77, 1986.

[8] EFRON, B. AND TIBSHIRANI, R., *An Introduction to the Bootstrap*, Chapman & Hall, New York, 1993.

[9] FISHER, N. I. AND HALL, P., Bootstrap algorithms for small sample, *J. Stat. Planning and Inf.*, **27**,157-169, 1991.

[10] HALL, P., On the number of bootstrap simulations required to construct a confidence interval, *Ann. Statist.*, **14**, 1453-1462, 1986.

[11] HALL, P., *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York, 1992.

[12] LEHMANN, E. L., *Element of Large-Sample Theory*, Springer-Verlag, New York, 1999.

[13] MARITZ, J. S. AND JARRETT, R. G., A note on estimating the variance of the sample median, *J. Amer. Statist. Ass.*, **73**, 194-196, 1998.

[14] SUPRIHATIN, B., GURITNO, S. AND HARYATMI, S., Consistency and accuracy of the bootstrap estimator for mean under Kolmogorov metric. *The 6th SEAMS-GMU 2011 International Conference on Mathematics and Its Applications*, 679-688, 2012.

[15] SUPRIHATIN, B., GURITNO, S. AND HARYATMI, S., Delta method for deriving the consistency of bootstrap estimator for parameter of autoregressive model. *Proceeding of the 8th World Congress on Probability and Statistics*, *Istanbul*, 2012