GLOBAL JOURNAL OF PURE AND APPLIED SCIENCES VOL. 14, NO. 2, 2008: 217 - 220
 ISSN 1118-0579

COMPARISON OF BOOTSTRAP AND JACKKNIFE METHODS OF RE-SAMPLING IN ESTIMATING POPULATION PARAMETERS

G. M. OYEYEMI

(Received 2 December 2006; Revision Accepted 26 July 2007)

ABSTRACT

Re-sampling methods have been found to be useful for several purposes such as model selection, linear regression, and estimation of sampling variances or standard errors and confidence intervals. In estimating the population coefficient of variation and its standard error, two methods of re-sampling, Bootstrap and Jackknife, are compared in this paper. The Jackknife method is found to require relatively small sample size to attain consistency in its estimate while Bootstrap requires large sample size. Bootstrap is also found to always underestimate the standard error of its estimates.

KEY WORDS: Bootstrap, Coefficient of variation, Jackknife, Sample size, Parametric.

1.0 INTRODUCTION

Many conventional statistical methods of analysis such as; correlation, regression, t-tests, and analysis of variance make some assumptions about normality. When these assumptions are violated, such methods of analysis may fail. Re-sampling method (Bootstrap or Jackknife) is steadily becoming more popular as a statistical methodology to overcome this problem (Efron, 1979). It is intended to simplify the calculation of statistical estimates, sometimes in situations where analytical answer cannot be obtained. With computer processors becoming faster and more powerful, the time and effort needed for re-sampling (Bootstrap or Jackknife) decreases to levels where it becomes a viable alternative to standard parametric statistical techniques (Barker, 2005)

In statistics, there are lots of methods that are practically guaranteed to work well if the distribution of the data is known especially, if it is normally distributed and all we are interested in are linear combinations of these normally distributed variables. If the sample size is large enough, central limit theorem can be used, in that we expect means to converge to normality, and hence do not need to do re-sampling from a normal distribution as N increases.

Suppose we want to make inference about the data when one of the following is true;

-Small sample size where assumption of normality does not hold

-A non-linear combination of variables (eg. Ratio)

-A location statistic other than the mean (Correlation or coefficient of variation)

Data-based simulation (Bootstrap or Jackknife) method for assigning measures of accuracy to statistical estimates can be used to produce inferences such as confidence intervals without knowing the type of distribution from which a sample has been taken. The use of statistical research is becoming more and more sophisticated. It is increasingly common for many proposed methodology to go beyond standard parametric analysis and in addition, costly or expensive data collection with its extremely non-normal nature are regularly encountered. Re-sampling method has become a recognized technique for dealing with these problems.

Bootstrap and Jackknife procedures as discussed by Efron (1982) and Miller (1974) are non-parametric statistical techniques which can be used to reduce the bias of point estimates and construct approximate confidence intervals for the population parameters and other summary statistics. These procedures require no assumption regarding the statistical distribution (e.g. Normal, Lognormal, Gamma etc.) for the underlying population, and can be applied to a variety of situations no matter how complicated.

Though it should be pointed out that, use of parametric statistical method (depending upon its distributional assumptions), when appropriate, is more efficient than its non-parametric counterpart, in practice, parametric assumptions are often difficult to justify or meet, hence, non-parametric methods are valuable tools for obtaining reliable estimates of parameters of interest (Singh et al, 1997).

This paper will examine and compare the two methods of re-sampling, Bootstrap and Jackknife, in estimating population parameters or their derivatives and the standard error of estimate. The standard error is used to construct the confidence intervals and statistical test of hypothesis. Also the minimum Bootstrap sampling required to obtain a stable and reliable estimate will be determined.

2. BOOTSTRAP METHOD

This method was first introduced by Efron (1979) to derive the estimate of standard error of an arbitrary estimator. Finding standard error of an estimator is an important activity for every statistician as it is rarely enough to find a point estimate. Statisticians will always want to know how reasonable an estimator is, by establishing or finding its variability. The Bootstrap is a type of Monte Carlo method based on observed data (Efron and Tibshirani, 1993).

In the Bootstrap procedure, repeated sample of size n are drawn with replacement from the given set of observations of size n. The process is repeated a large number of times, and each time an estimate of θ is computed. The estimates obtained are used to compute the estimate, $\hat{\theta}$, of the population parameter and its standard error. Bootstrap is not a way of reducing the errors, but it only tries to estimate the error

G. M. Oyeyemi, Department of Statistics, University of Ilorin, Ilorin, Kwara State, Nigeria

though it is often found to under estimate the errors. Bootstrap method of re-sampling is found to be less efficient when the data set is small, the observations are dependent and when the data contains outlier(s).

A general description of Bootstrap is as follow;

Step 1: Let $x_{i1}, x_{i2}, x_{i3}, \ , \ , \ x_{in}$ represent the i^{th} sample of size n with replacement from the original data set $x_1, x_2, x_3, \ , \ , \ x_n$. Then, compute the estimate of the population parameter and denote it as $\hat{\theta}_i$.

Step 2: Perform step 1 independently N times (500, 1000 or more), each time calculate a new estimate $\hat{\theta}$. Denote these estimates by $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ , \ , \ \hat{\theta}_N$.

The Bootstrap estimate of the population parameter is the arithmetic mean

$$\hat{\theta}_B = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i$$

And the Bootstrap estimate of the standard error is given by;

$$\hat{\sigma}_B = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\theta}_i - \bar{\theta}_B\right)^2}$$

The confidence interval is obtained by using the Bootstrap estimate and its standard error assuming normality as;

$$C.I = (\hat{\theta}_B + z_{(1-\alpha/2)}\sigma_B(\hat{\theta}),\ \hat{\theta}_B - z_{(1-\alpha/2)}\sigma_B(\hat{\theta}))$$

Where z_α denotes the $\alpha - th$ quartile of the standard normal distribution.

3.0 JACKKNIFE METHOD

Jackknife is one of the most commonly used methods for constructing simple and efficient variance estimators. For small sample size, Jackknife is simple and quicker to use. Moreover, Jackknife produces nonrandom estimates and it is easy to work with its closed-form solution. For small sample, such as n = 3 or 4, there is high probability that the Bootstrap samples contain only one distinct unit, which may lead to some estimates or summary statistics undefined (variance or its derivatives) (Variyath et al, 2005).

This method also involves two steps;

Step 1: Given a random sample $x_1, x_2, x_3, \ , \ , \ x_n$, we generate Jackknife samples which has value x_i removed from the data set, i = 1, 2, 3, , , n.

Step 2: For each Jackknife sample, the i^{th} partial estimate of the population parameter using the sample generated is obtained;

$$\hat{\theta}_i = (x_1, x_2, x_3, \ , \ , \ x_{i-1}, x_{i+1}, \ , \ , \ , x_n)$$

The n estimates of the population parameter, θ, are computed by deleting one observation at a time. If we denote the arithmetic mean of the n estimates by;

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$$

A quantity which shall be called the i^{th} "pseudo-value" is defined by;

$$J_i = n\theta - (n-1)\hat{\theta}_i$$

Where θ is the estimate from the original observations. The Jackknife estimate of θ is then given by;

$$J(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}J_i$$

If the original estimate of θ is biased, then, under certain conditions, part of the bias is removed by the Jackknife procedure, and an estimate of the standard error of the Jackknife estimate, $J(\hat{\theta})$, is given by;

$$\hat{\sigma}_{J(\hat{\theta})} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(J_i - J(\hat{\theta}))^2}$$

Another application or usefulness of pseudo-values, suggested by J. Turkey (see Miller, 1974), is using it to obtain the confidence intervals for the parameter θ based on the following pivotal quantity;

$$t = \frac{J(\hat{\theta}) - \theta}{\hat{\sigma}_{J(\hat{\theta})}}$$

The statistic, t, in the above equation has an approximate student's t distribution with n-1 degrees of freedom (Standard Normal for large N) which can be used to derive the $(1-\alpha)$ confidence interval for θ;

$$[J(\hat{\theta}) + t_{1-\alpha/2,n-1}\hat{\sigma}_{J(\hat{\theta})}\ ,\ J(\hat{\theta}) - t_{1-\alpha/2,n-1}\hat{\sigma}_{J(\hat{\theta})}]$$

4.0 SIMULATION

A random sample of size n (n = 5, 10, 20, 30, 50, 80, 100, 120, and 150) from normal variable with mean $\mu = 10$ and variance, $\sigma^2 = 3$ are simulated using R codes. The estimate of interest is the coefficient of variation, which is a function of mean and variance of the data. For simplicity, a function is defined in R such that if it is called, it will compute the coefficient of variation for a given data set. This function is used for both Bootstrap and Jackknife methods. Assume that x is the sample of size n generated from N(10, 3)

The codes for Bootstrap method

```
- cv = function(x) { sqrt(var(x))/mean(x)      # function to compute the cv
- boots = <- numeric(1000)                     # number of bootstrap sampling
- for (i in 1:1000) {
- boots[i] = cv(sample(x, replace=TRUE))}      # obtain the estimate of each sample
- mean(boots)                                  # compute the mean of the estimates
- var(boots)                                   # compute the variance of the estimates
```

The codes for Jackknife method

```
- jack = numeric(length(x)-1)
- pseudo = numeric(length(x))
- for (i in 1:length(x))
- { for (j in 1:length(x))
- { if (j < i) jack[j] = x[j] else if (j >i) jack[j-1] = x[j] }
- pseudo[i] = length(x)*cv(x)-(length(x)-1)*cv(jack) }
- mean(pseudo)
- var(pseudo)
```

Table 1: Summary of simulated results for Bootstrap and Jackknife.

	Bootstrap Method		Jackknife Method	
N	CV	Std Error	CV	Std Error
5	0.145	0.047	0.191	0.066
10	0.162	0.033	0.177	0.039
20	0.168	0.020	0.175	0.020
30	0.170	0.017	0.174	0.018
50	0.171	0.014	0.174	0.012
80	0.172	0.012	0.173	0.012
100	0.172	0.010	0.173	0.011
120	0.172	0.009	0.173	0.010
150	0.173	0.009	0.173	0.009

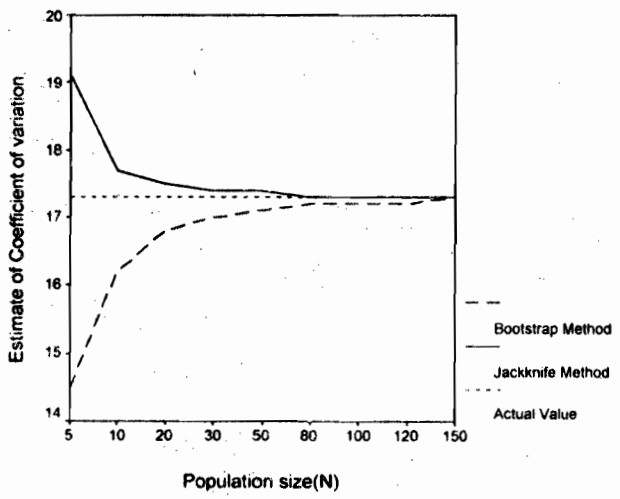

Fig. 1a. Estimate of Coefficient of Variation (CV)

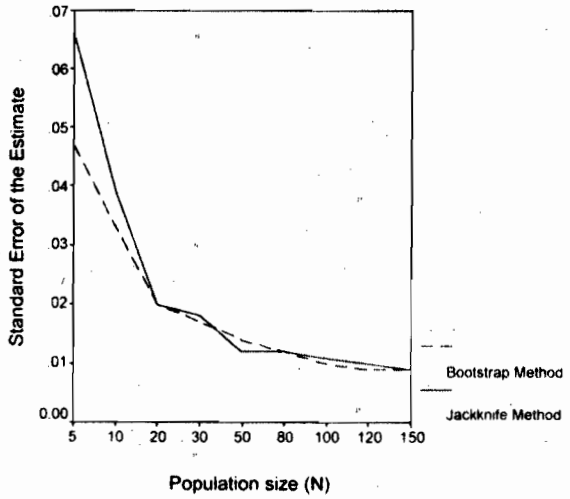

Fig. 1b. Standard Error of the estimates

The actual value of the population coefficient of variation (CV = 0.173) is attained, to two places of decimal, when the sample size is 30 for both methods (Table 1). Since sample size of 50 gave a reasonable estimate of the population parameter by the two methods, this sample size is used to determine the appropriate number of bootstrap samples required to attain a reasonable estimate of population parameter when Bootstrap method of re-sampling is used. The summary of the results is shown in Table 2.

Table 2: Summary of simulated results for Bootstrap method

Bootstrap samples	CV	Std Error
200	0.170	0.012
500	0.171	0.012
1000	0.171	0.012
2000	0.171	0.012
4000	0.171	0.012
6000	0.171	0.012
8000	0.171	0.012
10000	0.171	0.012
20000	0.171	0.012
30000	0.171	0.012

5.0 DISCUSSION AND CONCLUSION

It has been demonstrated that both methods of re-sampling technique are very efficient in estimating the population parameters and their standard errors, especially when population distribution is not specified or when estimation of summary statistics and the standard errors of such estimates are needed for inferential purposes. With sample size of 30 both methods gave an unbiased estimate of CV to 2 places of decimal. While a sample of size 80 is required by the Jackknife method to give the precise estimate of CV to 3 decimal places, Bootstrap method required at least a sample size of 150 to attain the same precision (Table 1 and Figure 1a). Generally, Bootstrap method is found to underestimate the standard error of its estimate when compared with Jackknife method, though, they both converge with large sample sizes (Table 1 and Figure 1b).

Bootstrap method seems to give consistent estimates of the population parameter and its standard error when the sample number of the observations it is sampling from is large. Therefore, when the number of observations is large enough, small Bootstrapping will give consistent estimates. With sample size of 50, Bootstrap sampling of 500 gave a consistent estimate of population parameter and its standard error (Table 2). It follows that when the sample size of observed data is small (say less than 30) large Bootstrap sampling (at least 1000) will be required to obtain consistent estimates, but large sample of observed data does not require such a large Bootstrap sampling.

REFERENCES

Barker, N., 2005. A Practical Introduction to the Bootstrap Using SAS System. www.ops-web.com

Efron, B., 1979. Bootstrap Methods: Another look at the Jackknife, Annals of Stat, 7: 1 - 26.

Efron, B., 1982. The Jackknife, the Bootstrap, and other Re-sampling Plans, Philadelphia, SIAM.

Efron, B. and Tibshirani, R. J., 1993. An Introduction to the Bootstrap. Chapman and Hall. New-York

Miller, R., 1974. The Jackknife – A Review. Biometrica, 61, 1-15.

Singh, A. K, Singh, A. and Engelhardt, M., 1997. The Lognormal Distribution in Environmental Applications. EPA/600/R-97/006. www.epa.gov.

Variyath, A. M, Abraham, B. and Chen, J., 2005. Analysis of Performance Measures in Experimental Designs Using Jackknife. Journal of Quality Technology, 37: 91-100.