Journal of the Nigerian Association of Mathematical Physics
***Volume 29, (March, 2015*), pp 245 – 250**
© J. of NAMP

K-Repeated Jackknife Algorithm for Small-size Population Inference

Adewara Johnson Ademola and Mbata Ugochukwu Ahamefuna

[1]Distance Learning Institute, University of Lagos, Akoka-Lagos, Nigeria.
[2]Department of Mathematics, University of Lagos, Akoka-Lagos, Nigeria.

Abstract

This paper proposed a modified Jackknife re-sampling technique called k-repeated Jackknife method for small-size population inference. The proposed method is aimed at improving the efficiency of the sample estimator and reducing bias estimate of the population parameter and inference. The method was compared with the usual Jackknife method to find the best minimum variance unbiased estimator. The results showed a high error reduction in the application of K-repeated jackknife method than the usual Jackknife method.

Keywords: Jackknife, K-Repeated Jackknife, Mean Square Error (MSE), Bias Reduction, Confidence Interval (C.I) and Coefficient of Variation (C.V)

1.0 Introduction

In statistical applications involving the point estimation of an unknown parameter θ, there is need to estimate the accuracy of $\ddot{\theta}$ as an estimator of θ by taking cognizance of error reduction. The variance or the mean square error and standard error are commonly used to describe the efficiency of an estimate. However, many statistical tests are based on the assumption that the data follow a particular distribution, either normal distribution, exponential distribution, binomial distribution or other type of distribution. The stability and efficiency of an estimator of parameter distribution has been a major course of study in both descriptive and inferential statistics. Many scholarly works have been carried out to bring into view a bias reduction estimation technique. More recently, the act of re-sampling which is based on repeated sampling within the same sample: either computed with or without replacement) became an important procedure for parameter estimation. Some common re-sampling techniques in use include randomization exact test, cross-validation, bootstrapping, jackknifing, training, importance sampling, Monte Carlo, Markov Chain Monte Carlo, and Gibbs sampling. The objective of this paper is to use a modified Jackknife bias reduction technique for parameter estimation and compared the result to the usual jackknife re-sampling procedures. The use of variance, standard error of estimates and coefficient of variation will be employed to detect a highly bias reduction technique for sample inference. The remainder of the paper is organized as follows: Sections 2.0 discuss *the Methodology under Jackknife and the proposed K-repeated Jackknife,* Section 3.0 discuss the Methods and data, Section 4.0 deals with *Data analysis,* Section 5.0 *Discussion of results* while section 6.0 presents the *conclusion***.**

2.0 Jackknife Method

The jackknife takes the entire sample except for 1 value, and then calculates the test statistic of interest. It repeats the process, each time leaving out a different value, and each time recalculating the test statistic. The jackknife assumes that the values came from a sample that was collected randomly and that the observations in the sample are independent.

3.0 Estimation under Jackknife Method

Given a set of observed random samples, let $\ddot{\theta}$ be an estimator of the parameter θ based on the complete sample of size n with g subgroups. Computed as

$$\ddot{\theta} = \frac{1}{n}\sum_{i=1}^{n} t_i \qquad (1)$$

Let $\ddot{\theta}_{-i}$ be the corresponding estimator based on the sample at the i-th deletion. Defined as

$$\bar{\theta}_i = g\ddot{\theta} - (g-1)\ddot{\theta}_{-i}(i = 1, 2, \ldots, g) \qquad (2)$$

Corresponding author: Adewara Johnson Ademola, E-mail: , Tel.: +2348023875722.

The i-th deletion of the total could be one individual observation or several observations [1-5]. The latter case is called group- or block-based jackknife if one replication or one block observations are deleted [1,4, 6]. In equation (2) estimation $\bar{\theta}_i$ is called the i-th pseudo value and the estimator in equation (3) is the jackknife estimator for the parameter θ, where θ can be a variance component, covariance component, correlation coefficient, or any other parameter of interest.

$$\bar{\theta} = \frac{1}{g}\sum_{j-1}^{g}\bar{\theta}_i = g\ddot{\theta} - (g-1)\frac{1}{g}\sum_{j-1}^{g}\ddot{\theta}_{-i} \qquad (3)$$

We call $\bar{\theta}$ in equation (3) a pseudo jackknife estimate. Then a t-test can be used to test significant deviation from a given parameter value, $\theta_0$with the degrees of freedom $g-1$[4]. The equation (2) can be rewritten as follows,

$$\bar{\theta}_i = g\ddot{\theta} - (g-1)\ddot{\theta}_{-i} = \ddot{\theta} + (g-1)\left(\ddot{\theta} - \ddot{\theta}_{-i}\right)(i = 1, 2, \dots, g) \qquad (4)$$

Thus, it is obvious that pseudo value $\bar{\theta}_i$ in equation (4) is related to choices for g. When $g$large, slight is is difference between $\ddot{\theta}$ and $\ddot{\theta}_{-i}$will cause unfavorable values. More importantly, it will potentially cause a large standard error for an estimate and thus decrease the power for the parameter being tested. If we assume that the estimate $\ddot{\theta}_{-i}$ in equation (2) for the i-th deletion is unbiased, then it is easy to prove that $\bar{\ddot{\theta}}$ in equation (5) is unbiased too. It is often true if $\ddot{\theta}$ is an unbiased estimate of θ, then $\ddot{\theta}_{-i}$ will be unbiased after a few individuals in the original data are deleted.

$$\bar{\ddot{\theta}}_{j} = \frac{\sum_{i=1}^{g}\ddot{\theta}_{-i}}{g} \qquad (5)$$

We call $\bar{\ddot{\theta}}$ in equation (5) a non-pseudo jackknife estimate of the parameter θ. For each non-normally distributed variable, based on the Central Limit Theorem, $\bar{\ddot{\theta}}$ is approximately normally distributed when g is large [8]. Thus, an approximate z-test can be used when g is large or t-test can be used to test significant deviation from a given parameter value,θ_0,with the degrees of freedom $g-1$ [7].

Estimate of the variance of the jackknife estimate, $\bar{\ddot{\theta}}_{j\ k}$ is given by

$$V\left(\bar{\ddot{\theta}}_{j}\right) = \frac{1}{g(g-1)}\sum_{i-1}^{g}\left(\ddot{\theta}_{-i} - \bar{\ddot{\theta}}\right)^2 \qquad (6)$$

Estimate of the standard error of the jackknife estimate, $\bar{\ddot{\theta}}_{j}$ is given by

$$S^{-}\left(\bar{\ddot{\theta}}_{j}\right) = \sqrt{\frac{1}{g(g-1)}\sum_{i=1}^{g}\left(\ddot{\theta}_{-i} - \bar{\ddot{\theta}}\right)^2} \qquad (7)$$

100(1 -α) % confidence interval for θ is given by

$$\bar{\ddot{\theta}}_{j} \pm t_{\alpha/2,\ g-1}S^{-}\left(\bar{\ddot{\theta}}_{j}\right) \qquad (8)$$

The coefficient of variation is obtained as

$$C = \frac{S^{-}\left(\bar{\bar{\theta}}_{j}\right)}{\bar{\bar{\theta}}_{j}} * 100\% \qquad (9)$$

4.0 K – Repeated Jackknife Method

K – Repeated jackknife procedure is a re-sampling iterative scheme for mean square error (MSE) reduction. This involves jackknifing the observed data k-time, where k equals the sample size of the observed data. The procedure is conveniently applied when the sample size is small. The stopping rule for the repeated jackknife replications depends on the sample size of the original data. The procedure converges before or at k-th time, where the estimate from the jackknife replications is the same as estimator of the parameter θ based on the complete sample of size n. At the K-th time, the k-th – repeated jackknife estimate of bias is highly negligible.

5.0 Estimation under K – Repeated Jackknife Method

The method involves the following steps from the usual jackknife procedure.

Step 1: Observe a random sample T = (t_1, t_2, . . . ,t_n)

Step 2: Compute $\ddot{\theta}(t)$ a function of the data which estimates the parameter θ of the model.

$$\ddot{\theta} = \frac{1}{n}\sum_{i-1}^{n} t_i \qquad i = 1, 2, \dots, n \qquad (10)$$

Step 3: For i up to n

- generate a jackknife sample $T_{-i} = (t_1, t_{i-1}, t_{i+1}, \dots, t_n)$ by leaving out the i-th observation
- calculate $\ddot{\theta}_{-i}$ from each of the Jackknife sample T_{-i} by

$$\ddot{\theta}_{-i} = \frac{1}{n-1}\sum_{i-1}^{n-1} T_{-i} \qquad (11)$$

Step 4: Repeat step 3 using the estimates from $\ddot{\theta}_{-i}$ to form pseudo samples. The new pseudo samples are used to generate another set of jackknife estimates; this is continued until the k-th time. This implies that the process is repeated k times, and at any given stage the preceding jackknife estimates are used as new samples in the next stage until the k-th time.

Step 5: At the k-th time the K-repeated Jackknife mean estimate is calculated as

$$\bar{\theta}^{K} = \frac{1}{k}\sum_{i-1}^{n}\tilde{\theta}_{i-1}^{K} \tag{12}$$

The variance is obtained as

$$V\quad(\bar{\theta}^{K}) = \frac{1}{k(k-1)}\sum_{i-1}^{k}\left(\tilde{\theta}_{i-1}^{K} - \bar{\theta}^{K}\right)^{2} \tag{13}$$

The standard error of the mean estimate is given by

$$S^{-}\left(\bar{\theta}^{K}\right) = \sqrt{\frac{1}{n\,(n-1)}\sum_{i-1}^{n}\left(\tilde{\theta}_{i-1}^{K} - \bar{\theta}^{K}\right)^{2}} \tag{14}$$

An approximate (1–α) % confidence interval for θ is given by

$$\bar{\theta}^{K} \pm t_{\alpha/2,\ K-1}S^{-}\left(\bar{\theta}^{K}\right) \tag{15}$$

The coefficient of variation is obtained as

$$C\ = \frac{S^{-}\left(\bar{\theta}^{K}\right)}{\bar{\theta}^{K}} * 100\% \tag{16}$$

The general iterative scheme is as follows: from a random sample $T = (t_1, t_2, \ldots, t_n)$

1. $\tilde{\theta}_{(-1)}^{1} = \frac{1}{n-1}\sum_{i-1}^{n-1} T_{-i}$
2. $\tilde{\theta}_{(-1)}^{2} = \frac{1}{n-1}\sum_{i-1}^{n-1} \tilde{\theta}_{-1}^{1}$
3. $\tilde{\theta}_{(-1)}^{3} = \frac{1}{n-1}\sum_{i-1}^{n-1} \tilde{\theta}_{-1}^{2}$

. . . .

. . . .

. . . .

K. $\tilde{\theta}_{(-1)}^{K} = \frac{1}{n-1}\sum_{i-1}^{n-1} \tilde{\theta}_{-1}^{K-1}$

Thus, $\bar{\theta}^{K} = \frac{1}{n}\sum_{i-1}^{n}\tilde{\theta}_{i-1}^{K}$

Where K = n (sample size) indicates the stopping rule. Other estimators such as variance, standard error and confidence interval can be estimated as in equation (13), (14), (15) and (16).

6.0 Data

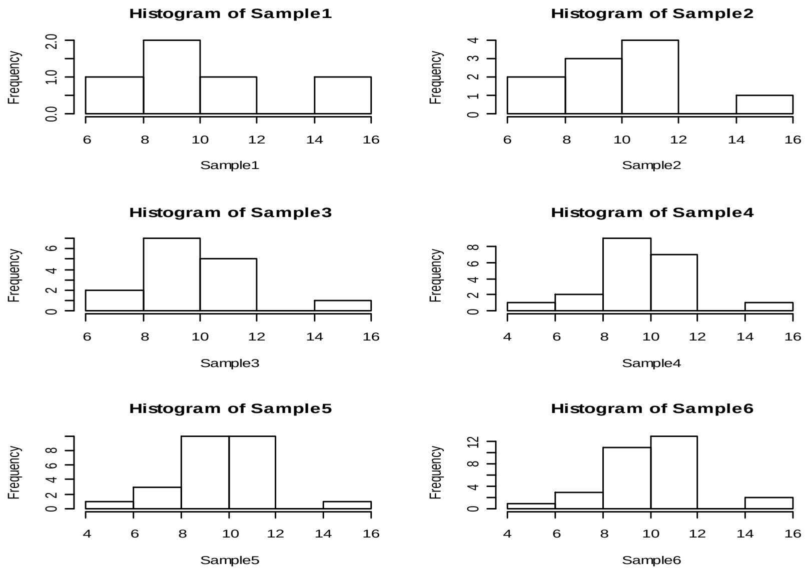
The data used for this work is simulated from a normal distribution, that is arbitrarily $t_i \sim N(10, 2.5)$, of sample sizes n = 5, 10, 15, 20, 25 and 30, using R statistical software. The data analysis was done by carrying out first an exploratory data analysis of the simulated data to validate the statistical assumptions of normality and constant variance. These have to be satisfied for the corresponding confidence interval to cover the true mean with the prescribed probability.

7.0 Results

It can be observed from the descriptive statistics in Table 1 that the means are close to ten and there is approximately constant variance as the sample size increases. These somewhat agree with the normal distribution assumptions. Figure 1 showed approximately normal distribution, which supports the random draw from the normal distribution. Hence, the exploratory data analysis has apparently revealed that the randomly generated samples approximately exhibited characteristics of the normal model N (10, 2.5).

Table 1: Descriptive Statistics of Randomly Generated Samples from Normal (10, 2.5)

Sample (i)	N	Mean	Variance	Std. Dev.	Lower limit	Upper limit
Sample 1	5	10.297	18.641	4.317537	4.936263	15.658125
Sample 2	10	10.021	10.251	3.201782	7.730870	12.311704
Sample 3	15	9.379	8.333	2.886634	7.780560	10.977691
Sample 4	20	9.785	6.801	2.607817	8.564622	11.005614
Sample 5	25	9.674	6.413	2.532476	8.628795	10.719505
Sample 6	30	9.680	7.807	2.794116	8.637094	10.723774



Figures 1: Histogram for each of the Randomly Generated Sample

8.0 Results of Parameter Estimations

The results of estimation of the parameters are displayed in Table 2, based on the randomly generated samples drawn from a normal distribution. The results of the parameter estimation from Jackknife method and K-repeated Jackknife method include point and interval estimations.

Table 2: Parameter Estimation Using the Two Methods

	JACKKNIFE METHOD				
Sample Size	$\bar{\bar{\theta}}_J$	Variance	SE	CI	CV
5	10.297	0.2330	0.4827	(9.205**,** 11.389)	0.0469
10	10.021	0.0127	0.1125	(9.766, 10.275)	0.0112
15	9.379	0.0028	0.0532	(9.267, 9.490)	0.0057
20	9.785	0.0009	0.0307	(9.720, 9.849)	0.0031
25	9.674	0.0004	0.0211	(9.630, 9.717)	0.0022
30	9.680	0.0003	0.0176	(9.644, 9.716)	0.0018

	K-REPEATED JACKKNIFE METHOD				
Sample Size	$\bar{\bar{\theta}}^K$	Variance	SE	CI	CV
5	10.297	3.56E-06	1.89E-03	(10.292, 10.301)	1.83E-04
10	10.021	8.43E-20	2.90E-10	(10.021, 10.021)	2.90E-11
15	9.379	4.51E-32	2.12E-16	(9.379, 9.379)	2.26E-17
20	9.785	1.08E-34	1.04E-17	(9.785, 9.785)	1.06E-18
25	9.674	2.63E-35	5.13E-18	(9.674, 9.674)	5.30E-19
30	9.68	8.70E-36	2.95E-18	(9.680, 9.680)	3.05E-19

SE: standard error. CI: 95% confidence interval. CV: coefficient of variation

9.0 Discussion of Results

Figure 1 showed approximately normal distribution, which supports the random draw from the normal distribution. Hence, the exploratory data analysis has apparently revealed that the randomly generated samples approximately exhibited characteristics of the normal model N (10, 2.5). The mean of the jackknife and the k repeated are equal but when comparing the variances, the result shows that variances of k repeated jackknife method is better than the jackknife.

Also, the standard errors obtained of the k repeated jackknife method performed better that the jackknife approach. The confidence interval for the k repeated jackknife was better that the jackknife method.

The two methods showed unbiased estimation of the population parameter with minimum variance. The estimated coefficient of variation is more homogeneous in the k-repeated Jackknife method than the usual Jackknife method in parameter estimation. Also, K-Repeated Jackknife method has a more minimum variance unbiased estimator (MVUE) than the usual Jackknife method; irrespective of the sample size, and it converges faster as the sample size increases. This is evident in the standard error and coefficient of variation. This result is an indication of strong influence of k-repeated Jackknife procedure on error reduction in estimating population parameter over the usual Jackknife method. This is well depicted in Figures 2 and 3.

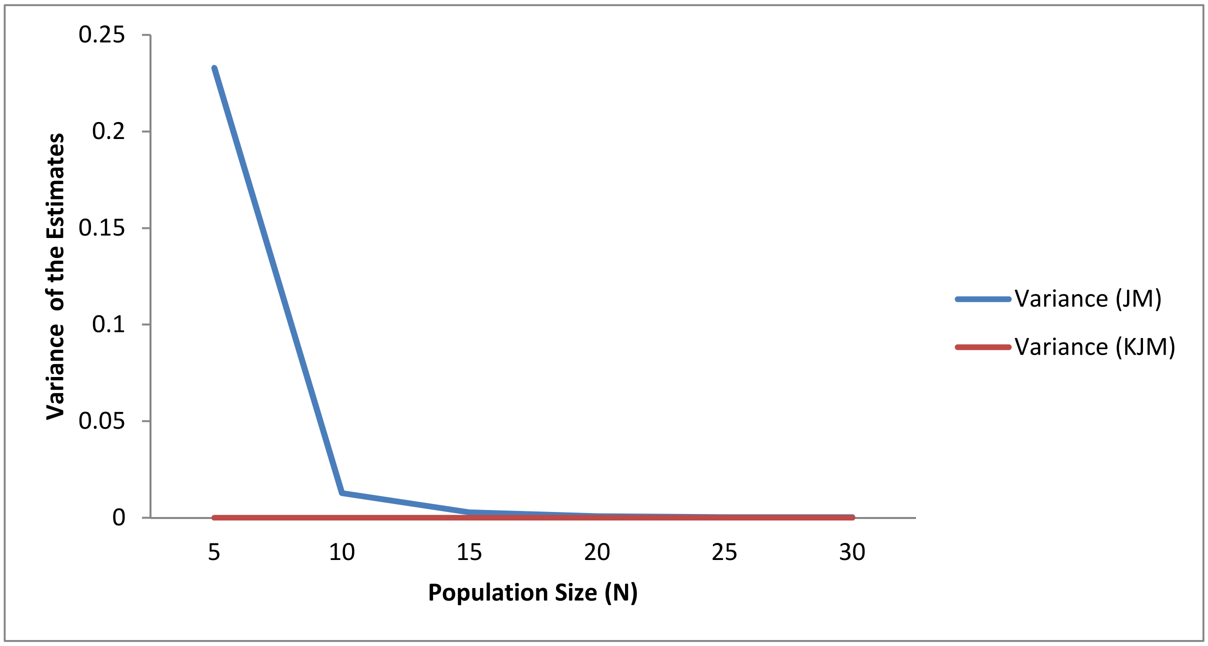


Figure 2: Variance of the Estimates

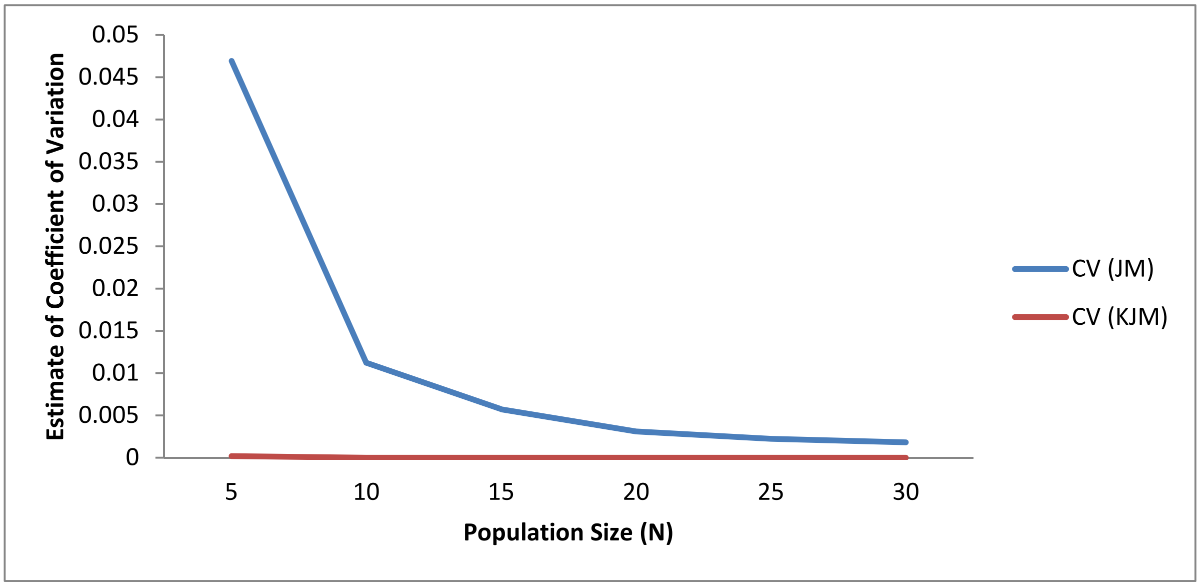


Figure 3: Estimates of Coefficient of Variation (CV)

Finally, the coefficients of variation of the two approaches were considered and the result showed that k repeated jackknife performed better than the jackknife approach

10.0 Conclusion

It has been demonstrated that both methods of re-sampling technique are very efficient in estimating the population parameters and their standard error especially when the sample size is small. The two techniques share the same methodological framework, technical features, limitations and inherent assumptions. Also, they are computationally intensive; but the jackknife iteration scheme is more in K-repeated method than in usual jackknife method. This is because in K-repeated method, the jackknife process is repeated K-times (where K is the number of observation in the original data set). However, these techniques represent an important step in refining the process of data analysis more especially the k-repeated procedure in bias reduction. Hence, K-repeated Jackknife is a method for further reduction of error accompanying a point estimator, and evaluation of variance of an estimator. This is clearly shown in the variance, standard error, coefficient of variation and confidence interval results in Table 2. The estimator error is highly reduced in the application of the K–repeated jackknife method compared to the usual Jackknife method.

11.0 References

[1] Quenouille, M. H.(1949): Approximate tests of correlation in time-series. Journal of the Royal Statistical Society, Series B, 11: 68-84.

[2] Quenouille, M. H. (1956): Notes on bias in estimation. Biometrika, 43:353-360.

[3] Miller, R. G.(1974b): The jackknife-a review. Biometrika, 61:1-15.

[4] Miller, R. G.(1974a): An unbalanced jackknife. Annals of Statistics, 2: 880-891.

[5] Wu, J., Jenkins, J. N., McCarty, J. C. and Wu, D.(2006): Variance component estimation using the additive, dominance, and additive × additive model when genotypes vary across environments. Crop Sci. 46: 174-179.

[6] Wu, C. J. (1990): On the asymptotic properties of the jackknife histogram. Annals of Statistics, 18: 1438 – 1452.

[7] W. W. Daniel. (1995): Biostatistics: A foundation for analysis in the health sciences. New York: John Wiley & Sons.

[8] Wackerly, D. D., Mendenhall, W. S. and Schaeffer, R. L. (1996): Mathematical statistics with applications. Belmont, California: Wadsworth Publishing Company