International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:6, No:2, 2012

A Bootstrap's Reliability Measure on Tests of Hypotheses

Al Jefferson J. Pabelic, Dennis A. Tarepe

***Abstract*—**Bootstrapping has gained popularity in different tests of hypotheses as an alternative in using asymptotic distribution if one is not sure of the distribution of the test statistic under a null hypothesis. This method, in general, has two variants – the parametric and the nonparametric approaches. However, issues on reliability of this method always arise in many applications. This paper addresses the issue on reliability by establishing a reliability measure in terms of quantiles with respect to asymptotic distribution, when this is approximately correct. The test of hypotheses used is F-test. The simulated results show that using nonparametric bootstrapping in F-test gives better reliability than parametric bootstrapping with relatively higher degrees of freedom.

***Keywords*—**F-Test, Nonparametric Bootstrapping, Parametric Bootstrapping, Reliability Measure, Tests of Hypotheses

I. Introduction

THE fundamental notion of every hypothesis test is to assess the position of the computed value of a test statistic, say t, using the distribution of T under the assumed null hypothesis and perform an inference whether there is sufficient evidence not to reject the null relative to certain margins [5]. Whilst many distributions of test statistics have proved to be reliable when the underlying distribution under the null hypothesis is known and can perform exact tests, there are still cases in which the characteristics of the latter distribution are quite unclear or, in extreme case, unknown. In consequence, one will have to assess t in a distribution that is just approximately correct which may lead to unreliable results

Consider the observed values of a sample data $\boldsymbol{y} = (y_1, \dots, y_k)$ which is an outcome of independent and identically distributed random variable $\boldsymbol{Y} = (Y_1, \dots, Y_k)$. Assume that T is an estimator, satisfying some established properties, of a parameter θ. If $\boldsymbol{Y}$ is distributed to some known probability distribution, then one is confident to use in inference a valid asymptotic distribution that T follows under a null hypothesis. Indeed, the suitable assumption on $\boldsymbol{Y}$ is the ideal thing to do to get reliable results. However, it is evident that in many applications, one can be fairly, even not, confident in a particular known distribution $\boldsymbol{Y}$ has. Hence, in this case, using standard asymptotic distribution in inference would give undesirable results [6].

Simulation based testing is a straightforward approach to address these limitations which makes advantage of the very fast development of computing.

A. J. Pabelic is with the Northern Consortium UK, Shenyang City, PRC (phone: 13897976427; e-mail: alpabelic@yahoo.com).

D. A. Tarepe is with the Mathematical Sciences Department, Mindanao University of Science and Technology, Cagayan de Oro City, Philippines 9000 (phone: 09228025324; e-mail: da_tarepe@must.edu.ph).

This approach is to generate a large number of simulated values of the test statistic and compare t with the generated distribution. A particular method of this testing is bootstrapping. Politis [8] has exemplified that "the availability of valid nonparametric inference procedures based on re-sampling and/or sub-sampling has freed practitioners from the necessity of resorting to simplifying assumptions such as normality or linearity that may be misleading." Apparently, this method gains attractiveness on its applications. However it is, in general, neither as easy nor as reliable as users often seem to believe [5].

This paper tries to deal with the issue on reliability of bootstrapping, both parametric and nonparametric, on tests of hypotheses relative to the established distributions from asymptotic theory. The test to be considered is the F-test which is widely used in many tests of hypotheses. In particular, the objectives of this paper are as follows: (a) determine the bias and variance of F^*, where F^* is the set of f bootstrap replicates, from the two bootstrapping approaches; (b) estimate the quantiles of F^* on different number of iterations; (c) assess the reliability of F^* relative to asymptotic distribution in terms of quantiles and (d) make some inferences based on the reliability percentage results.

This paper is organized as follows. Section 2 describes the tests of hypotheses and how these tests are performed. The principles behind bootstrapping are presented in section 3. It also illustrates parametric and nonparametric methods. In section 4, simulation methodology is given. Section 5 provides results and discussions and the summary, conclusion and recommendation are given in Section 6.

II. Testing A Hypothesis

A statistical hypothesis test is a method of making statistical decisions using sample data. This will be done by computing a statistic and examine its position to the theoretical distribution it will follow if the null assumption is true. There are certain levels of significance a test can be evaluated by which decisions can then be drawn whether or not to reject a null hypothesis. Hypothesis testing defined in this general procedure follows a "frequentist" statistical inference framework.

Common tests are one-sample & two-sample z-tests, one-sample & paired t-tests, pooled t-test, one proportion z-test, pooled z-test, two-sample F-test and many more.

The model of the F-test which will be used throughout this paper is in a form

$$F = \frac{Q_1/n}{Q_2/m} \qquad (1)$$

where $Q_1 = \sum_{i=1}^{n} X_i^2$, $Q_2 = \sum_{i=1}^{m} Y_i^2$, $X \sim iidN(0,1)$ and $Y \sim iidN(0,1)$. The variables n and m are the degrees of freedom for X and Y respectively. Also, $Q_1 \sim \chi^2(n)$ and $Q_2 \sim \chi^2(m)$.

III. Bootstrapping

A. Basic Concepts

Bootstrapping is a direct approach in generating a probability distribution for a statistic that can be used for statistical inference [3]. Given a sample data $y = (y_1, \dots, y_n)$, a statistic t can be computed. Bootstrap samples denoted as $y^* = (y_1^*, \dots, y_n^*)$ are generated from y in performing bootstrapping procedure. These bootstrap samples are then mapped to a functional form of t to produce the bootstrap replicates denoted as t^*. The number of replicates depends on how many times the iteration is being performed in the process. In building a distribution of bootstrap replications, the nonparametric and parametric bootstrapping are the two general approaches to use.

B. Nonparametric Bootstrapping

Suppose $y = (y_1, \dots, y_n)$ is an outcome of independent and identically distributed random variable $Y = Y_1, \dots, Y_n$. If the distribution function of Y, say G, is unknown, then a sensible estimate of G is the empirical distribution function (EDF) $\hat{G}$ [7]. The role of the EDF is the foundation of nonparametric bootstrapping. This is defined as $\hat{G}(u) = n^{-1} \sum_{i=1}^{n} h(y_i \leq u)$, where $h(\cdot)$ is an indicator function. Since $\hat{G}$ places equal probabilities on the original sample y, then each element in y^* is independently sampled at random from these data values. Therefore the simulated sample $Y_1^*, \dots, Y_n^*$ is a random sample taken with replacement from the data. This simplicity is special to the case of a homogenous sample but many extensions are straightforward [3].

C. Parametric Bootstrapping

Moreover, if y assumes a particular parametric model there exists an estimate $\hat{\psi}$ of the parameter ψ of G. This estimate serves as a substitute parametrically in the fitted distribution $\hat{G}_{par}$. Thus, $\hat{G}_{par}$ will be used in generating bootstrap samples y^*. For instance, consider y as an outcome from a normal distribution. $\hat{G}_{normal}$ with parameters $\bar{x}$ and s^2 generates $y^* = (y_1^*, \dots, y_n^*)$ which in turn used to compute the replicates. This approach is parametric bootstrapping.

D. Empirical Mean, Bias and Variance of the Replications

Getting an unbiased and consistent estimator is one of the main goals in statistical estimation. Consider again the random sample $y = (y_1, \dots, y_n)$. Estimating the parameter θ can be done by calculating a statistic T from the random sample. The value of statistic from the random sample is denoted as t. For every bootstrap sample, the same statistic can be calculated to obtain the bootstrap replications of T as $\widehat{\theta_i^*} = T_i^* = t(X_{i1}^*, \dots, X_{in}^*)$, $i = 1, \dots, B$. Hence, a straightforward computation of the empirical mean of the replications is by using the formula

$$\overline{\widehat{\theta^*}} = \frac{1}{B} \sum_{i=1}^{B} \widehat{\theta_i^*}, i = 1, \dots, B \tag{2}$$

Furthermore, this empirical mean can be used to compute the empirical bias which is defined as

$$\widehat{bias}_B = \overline{\widehat{\theta^*}} - \hat{\theta}, i = 1, \dots, B \tag{3}$$

Finally, the empirical variance is denoted as the plug-in formula

$$\hat{\sigma}^2_{\widehat{\theta^*}} = \frac{1}{B-1} \sum_{i=1}^{B} \left(\widehat{\theta_i^*} - \overline{\widehat{\theta^*}} \right)^2, i = 1, \dots, B \tag{4}$$

E. A Reliability Measure on Bootstrapping

There is a variety of labels in literature designated to problems on reliability. Such labels are reliability, availability, interval availability, efficiency, effectiveness, etc. Unfortunately, the definitions given in the literatures are sometimes unclear and vary among different writers.

Barlow and Proschan have defined mathematically, a single generalized quantity which, when appropriately specialized, will yield most of the fundamental quantities of reliability theory [2]. Their definition is

Assume a system whose state at time t is described by $X(t) = (X_1(t), \dots, X_n(t))$, a vector-valued random variable. $X(t)$, being a random variable, will be governed by a distribution function, $F = (x_1, \dots, x_n : t)$; explicitly, $F = (x_1, \dots, x_n : t)$ equals the probability that $X_1(t) \leq x_1, \dots, X_n(t) \leq x_n$. Now, corresponding to any state $x = (x_1, \dots, x_n)$, there is a gain, or payoff, $g(x)$. The expected gain $G(t)$ at time t will be the quantity of interest; it may be calculated from $G(t) = E_g(X(t)) = \int \dots \int g(x_1, \dots, x_n) dF = (x_1, \dots, x_n : t)$

For the purpose of this paper, the bootstrap reliability measure relative to F asymptotic distribution $R(b)$ maybe thought of as a state at bootstrap iteration b, where $R(b) = \left(1 - \left| f_{(asy)} - f^*_{(par/non)b} \right| / f_{(asy)}\right) \times 100$. The notation $f_{(asy)}$ is denoted as the critical value at $\alpha = 1 - p$ of F distribution. While $f^*_{(par/non)b}$ is the $(B+1)p^{th}$ ordered value of f^* from bootstrap empirical distribution, where $p = j/(B+1)$ and $f^*_{(j)}$ denotes the j^{th} ordered value. It is also of interest to calculate the $E(R(b))$, but this is not included in the scope of this paper.

IV. Algorithms for Simulation

The following are the algorithms established to guide the simulation process.

A. Algorithm in Constructing Baseline Quantiles based from Asymptotic Distribution

1. Select desired degrees of freedom.
2. Compute the critical values of f in the 90th, 95th and 99th quantiles. These computed values will serve as the baseline values in computing for bootstrap's reliability.

B. Algorithm in Constructing Quantiles from Parametric Bootstrapping

1. Draw random samples $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ from $N(0,1)$. These sets are assumed to be the observed samples.
2. Calculate $\bar{X}, s_x^2, \bar{Y}, s_y^2$ and $\hat{f}$. These values will serve as observed statistics.
3. Sample $x^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$ and $y^{*b} = (y_1^{*b}, \ldots, y_n^{*b})$ from $N(\bar{X}, s_x^2)$ and $N(\bar{Y}, s_y^2)$ respectively. These sets are the bootstrap samples.
4. Compute $\hat{f}^{*b}$. This value is a bootstrap replicate.
5. Repeat steps 3 through 4, B times.
6. Create $\hat{F}_B(\hat{f}^*)$, using the replicates, which is the empirical distribution function of $\hat{f}^*$.
7. Find $\hat{f}^*_{((B+1)p)}$ from $\hat{F}_B^{-1}(p)$ where B is chosen so that $(B+1)p$ is an integer. This value is the estimated j^{th} $(B+1)$ - quantiles.

C. Algorithm in Constructing Quantiles from Nonparametric Bootstrapping

1. Use the samples obtained from B.1.
2. Sample the sets in step 1 with replacement to get $x^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$ and $y^{*b} = (y_1^{*b}, \ldots, y_n^{*b})$
3. Compute $\hat{f}^{*b}$.
4. Repeat steps 2 through 3, B times.
5. Create $\hat{F}_B(\hat{f}^*)$, using the replicates, which is the empirical distribution function of $\hat{f}^*$.
6. Find $\hat{f}^*_{((B+1)p)}$ from $\hat{F}_B^{-1}(p)$ where B is chosen so that $(B+1)p$ is an integer. This value is the estimated j^{th} $(B+1)$ - quantiles.

D. Algorithm in Computing the Empirical Biases and Variances of F*

1. Use $\hat{f}$ and $\hat{f}^{*b}$ to compute the empirical biases of F^* given in the formula $\widehat{bias} = \frac{\sum_{i=1}^{b} \hat{f}^{*i}}{b} - \hat{f}$.
2. Use $\hat{f}^{*b}$ to compute the empirical variances of F^* given in the formula $\widehat{\sigma^2} = \frac{1}{b-1}\sum_{i=1}^{b}\left(\hat{f}^{*i} - \frac{\sum_{i=1}^{b}\hat{f}^{*i}}{b}\right)^2$.

E. Algorithm in Computing the Reliability of Bootstrapping Approaches

1. Denote $f_{(asy)}$ as the critical value at $\alpha = 1 - p$ of F distribution and $f^*_{(par/non)b}$ as the $(B+1)p^{th}$ ordered value of f^* from bootstrap empirical distribution.
2. Compute $R(b)$ for parametric and nonparametric bootstrapping.
$R(b) = \left(1 - \left|f_{(asy)} - f^*_{(par/non)b}\right| / f_{(asy)}\right) \times 100$ is the bootstrap reliability measure at b in percent with respect to F asymptotic distribution.

V. Results and Discussions

*A. Empirical Bias of F**

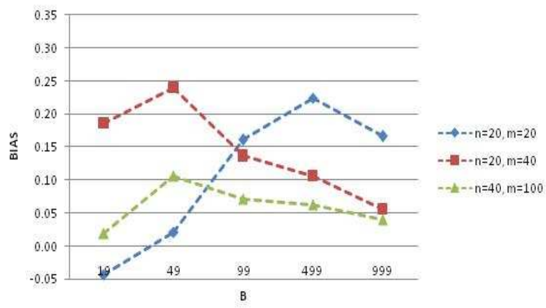

Fig. 1 Empirical biases of F^* for p=0.90 under parametric bootstrapping

Fig. 1 shows the behavior of the biases under parametric bootstrapping. The bias ranges from -0.05 to 0.35 and the bootstrap replications are 19, 49, 99, 499 and 999. The bias on (n=20, m=20) departs from zero as b increases. Contrary to this, bias on (n=20, m=40) approaches to zero as b increases. Among the three, bias on (n=40, m=100) converges to zero the fastest as b increases. This implies that the higher the degrees of freedom, parametric bootstrapping produces unbiased replicates as b increases.

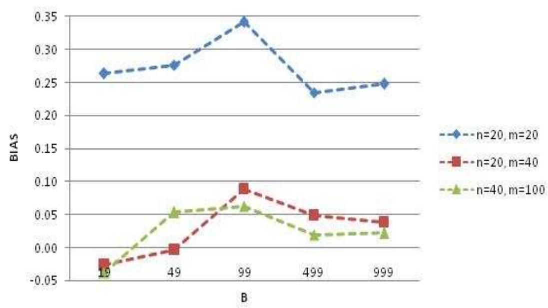

Fig. 2 Empirical biases of F^* for p=0.90 under nonparametric bootstrapping

On the other hand, fig. 2 shows the empirical biases of F^* for p=0.90 under nonparametric bootstrapping. The behavior of the biases of nonparametric bootstrapping is much likely different from that of parametric bootstrapping. The distance from the bias on (n=20, m=20) to the bias of the two groups is very wide. This implies that relatively lower degrees of freedom produces biased replicates at all values of b. Evidently, still the fastest rate of convergence to zero is the bias from (n=40, m=100). It gives biases almost zero on all values of b.

B. Empirical Variance of F^*

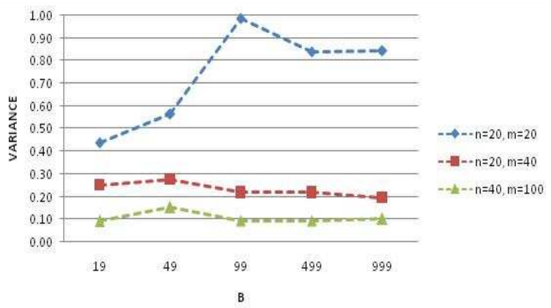

Fig. 3 Empirical variances of F^* for p=0.9 under parametric bootstrapping

Fig. 3 displays the empirical variances of F^* for p=0.90 under parametric bootstrapping. The distances of empirical variances among the three dimensions of degrees of freedom are clearly wide. Each dimension reveals characteristics of consistency of the bootstrapping processes. The attribute of consistency implies that if the variance approaches to zero then the estimator is consistent. The empirical variances on (n=40, m=100) display consistency as the data points are near to zero at all b.

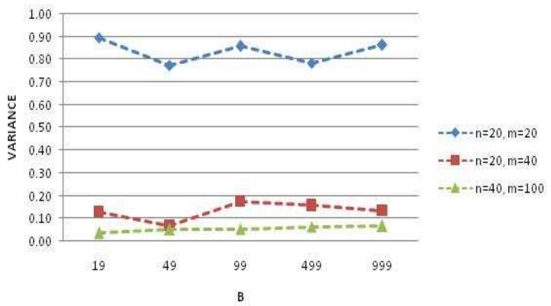

Fig. 4 Empirical variances of F^* for p=0.90 under nonparametric bootstrapping

Fig. 4 displays the empirical variances of F^* for p=0.90 under nonparametric bootstrapping. Apparently, the distances of the empirical variances from (n=20, m=20) to (n=20, m=40) and (n=40, m=100) are wider compare to the parametric bootstrapping counterpart. When b=49, empirical variances of (n=20, m=40) and (n=40, m=100) are close while the other values of b maintain a distance from each and the other points. Nevertheless, empirical variances of (n=40, m=100) are nearer to zero at all b than its counterpart. Hence, nonparametric bootstrapping gives favorable results on consistency.

C. Empirical Quantile of F^*

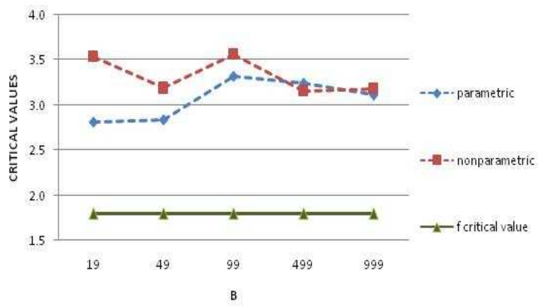

Fig. 5 Empirical quantiles of F^* versus f quantile from asymptotic distribution for p=0.90 at n=20 and m=20

The graphs of the quantiles of F^* seem to have an equal level with respect to the critical value *f* based from the *F* distribution as shown in fig. 5. Clearly, the empirical quantiles from both bootstrapping approaches do not converge to the *f* critical value at all b. This means that the bootstrapping processes do not give good estimators of the exact *f* quantile. The *f* critical value is referred to as the solid line in figures 5, 6 & 7.

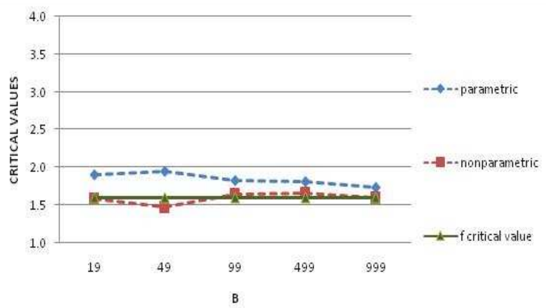

Fig. 6 Empirical quantiles of F^* versus f quantile from asymptotic distribution for p=0.90 at n=20 and m=40

Contrary to fig. 5, fig. 6 gives a different view on the empirical quantiles approaching to the baseline *f*. Nonparametric bootstrapping outperforms parametric bootstrapping in terms of giving the right quantiles relative to baseline *f*. Although there is a "burst" at b=49, nonparametric bootstrapping can give quantiles close to baseline *f*, even at b=19.

International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:6, No:2, 2012

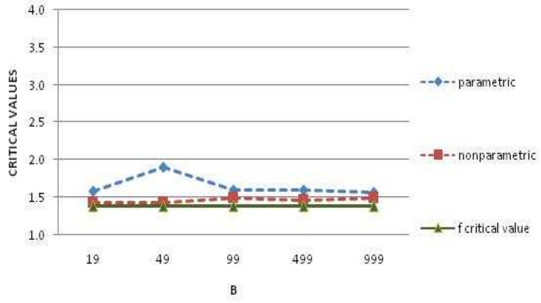

Fig. 7 Empirical quantiles of F^* versus f quantile from asymptotic distribution for p=0.90 at n=40 and m=100

Fig. 7 displays a surprising result. Parametric bootstrapping is very slow in converging to *f* critical value even the dimension of degrees of freedom is relatively high. Nonparametric bootstrapping maintains its close distance, for all b, from the *f* critical value.

D.Reliability Measure

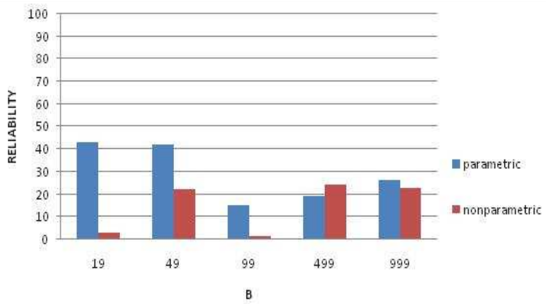

Fig. 8 Reliability measures for p=0.90 at n=20 and m=20

Fig. 8 shows the reliability measures with range from 0 to 100 in percent. Should there be a negative reliability measure, it simply implies that there is a wide difference between the *f* quantile from asymptotic distribution and f^* quantile from bootstrapping. Hence, the reliability in this case is not good. All of the measures on both approaches are less than fifty percent. Bootstrapping at relatively lower dimension of degrees of freedom does not give a satisfactory result. Parametric bootstrapping performs better than nonparametric bootstrapping in this case.

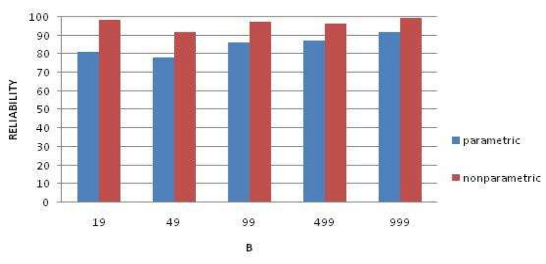

Fig. 9 Reliability measures for p=0.90 at n=20 and m=40

Moreover, fig. 9 shows the reliability measures for p=0.90 at n=20 and m=40. All of the reliability measures are approaching to 100. Congruent to their variances' consistency, these reliability measures coincide with their results at all b. Nonparametric bootstrapping gives higher reliability measures compared to parametric bootstrapping where almost all of the measures are above 90.

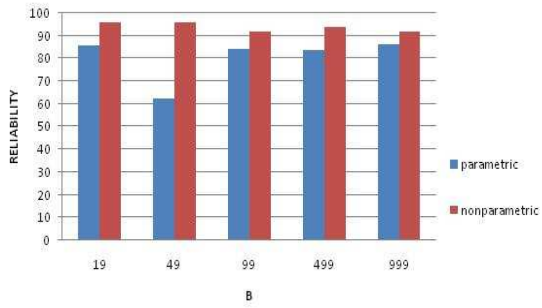

Fig. 10 Reliability measures for p=0.90 at n=40 and m=100

Lastly, in fig. 10, there is a slight change of the levels of the measures under both approaches. For instance, the reliability measure under nonparametric bootstrapping at n=20 and m=40 with b=999 is 99.48 while at n=40 and m=100 lowers to 91.65. Further, the reliability measure under parametric bootstrapping at n=20 and m=40 with b=999 is 91.68 while at n=40 and m=100 lowers to 86.23. In general, nonparametric bootstrapping performs well in this case at all b.

VI. SUMMARY, CONCLUSION AND RECOMMENDATION

Bootstrapping has gained popularity in different tests of hypotheses as an alternative in using asymptotic distribution if one is not sure of the test statistic's distribution under a null hypothesis. This method, in general, has two variants – the parametric and the nonparametric approaches. However, issues on reliability of this method always arise in many applications.

This paper addresses the issue on reliability by establishing reliability measure in terms of quantiles with respect to asymptotic distribution when this is approximately correct. The two bootstrapping variants are then investigated on their respective reliability measures. Whereas there are papers, for example [6], who claimed that parametric bootstrapping performs well in many applications, this paper shows that the claim is not true in all cases. Parallel to this, [1] highlights that "the performance of parametric and nonparametric bootstrapping are the same if the parameter of interest is the mean. Conversely, for the variance, the bootstrap estimation depends on the sample kurtosis of the data." Specifically, the bootstrapping reliability measures of both approaches on F-test, where the chi square random numbers came from N(0,1), vary depending on empirical biases, variances, extent of degrees of freedom and iterations.

In the case where the degrees of freedom are n=20 and m=20 corresponding to the chi square random numbers at numerator and denominator, respectively, the reliability measures are not satisfactory. For n=20 and m=40, both approaches give above 50 percent reliability measures. Among the two approaches, nonparametric bootstrapping performs better than parametric bootstrapping in terms of reliability. This result is also evident when n=40 and m=100. The spread of the empirical biases and variances, in this simulation, influences the reliability measures. The consistency result from empirical variances gives satisfactory results on reliability measures at all b. Relatively higher degrees of freedom improve the reliability measures which converge to 100 percent.

Using nonparametric bootstrapping in F-test gives better reliability, in this paper, than parametric bootstrapping with relatively higher degrees of freedom.

Furthermore, it is recommended to extend this study to other tests of hypotheses, include different nonparametric bootstrapping approaches in investigating reliability and explore reliability measures on dependent data.

REFERENCES

[1] S. Amiri, D. von Rosen, and S. Zwanzig, “On the comparison of parametric and nonparametric bootstrap,” Uppsala University Department of Mathematics Report 2008:15. Uppsala University, Uppsala, Sweden, 2008, unpublished.

[2] R. Barlow and F. Proschan, *Mathematical Theory of Reliability*, SIAM Classics edition, 1996, pp. 5-6.

[3] A. Davison and D. Hinkley, *Bootstrap Methods and their Applications.* Cambridge, United Kingdom: Cambridge University Press, 1997, ch. 1.

[4] P. Good and J. Hardin, *Common Errors in Statistics: How to Avoid Them*. New Jersey: John Wiley & Sons, 2005.

[5] J. MacKinnon, “Bootstrap hypothesis testing,” Queen’s Economics Department Working Paper No. 1127, Queen’s University, Ontario, Canada, 2007, unpublished.

[6] J. MacKinnon and R. Davidson, “Improving the reliability of bootstrap tests with the fast double bootstrap,” Queen’s Economics Department Working Paper No. 1044. Queen’s University, Ontario, Canada, 2006, unpublished.

[7] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, New York: Chapman & Hall, 1993, pp. 31-32.

[8] D. Politis, “The impact of bootstrap methods in time series,” *Statistical Science*, Vol. 18, No. 2, 2003, pp. 219–230.