International Journal of Advance Research, IJOAR .org
Volume 1, Issue 3, March 2013, Online: ISSN 2320-9143

CORRELATION ANALYSIS: THE BOOTSTRAP APPROACH

1Ikpotokin O. and Edokpa I. W.

Department of Mathematics and Statistics, Ambrose Alli University, Nigeria
1e-mail: ikpotokinosayomore@yahoo.co.uk

Abstract

For a general class of problems, the bootstrap method of resampling is one of the possible methods of constructing tests of significance. The sampling distribution of a test statistic for an experiment compiled by the bootstrap approach requires no reference to the population distribution and therefore no requirement that it should conform to a mathematically definable frequency distribution. Algorithms for the bootstrap distribution of correlation coefficients are presented and implemented. As an illustrative example, a critical value for Pearson's product moment correlation coefficients and Spearman's rank correlation coefficients are produced for a given set of data.

Keywords and phrases:

Bootstrap test, p-value, algorithm, paired observations, correlation

1.0 Introduction

There are several experimental situations in which there is only one set of n experimental subjects and two-observations are made on each subject. The data consists of n pairs, such as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. In an attempt to ensure that the probability of type I error is

approximately α in analyzing the linear relationship for paired observations, an algorithm for obtaining bootstrap distribution of paired observations is presented. A major problem of statistical inference is to obtain the test of significance when the form of the underlying probability distribution is unknown. The idea of a general method of dealing with this problem of obtaining the test of significance originated with Efron (1979, 1982). The essential feature of the

method is that a large number of bootstrap samples of the observations are considered, with the property that each bootstrap is equally likely under the hypothesis to be tested. A test on the level of significance is constructed by choosing a proportion, α, of the bootstrap as critical region. It is shown in Efron and Tibshirani (1993) that for a general class of problems, the bootstrap approach is one of the possible methods of constructing a test of significance. Several approaches which are computationally demanding such as permutation have been suggested as alternatives to the bootstrap approach because of it exactness; see Good (2000), Pesarin (2001), Bagui and Bagui (2005), Odiase and Ogbonmwan (2007).

Bootstrap tests are attractive because the distribution of the observations under the null hypothesis need not be known in order to obtain the p-value. Sideridis and Simos (2010) assert that the bootstrap test is as powerful as the best parametric test when based on the same statistic. Permutation procedures gives exact result most especially when it can carry out complete enumeration of all possible distinct rearrangements for small sample size. These procedures can sampling without replacement within a sample, but cannot avoid the impossibility of complete enumeration when the sample sizes are fairly large; thereby reducing the power of the permutation test, see Odiase and Ogbonmwan (2005). This paper therefore presents an algorithm that makes it possible to obtain large bootstrap configurations of an experiment without the problem of drawing a complete enumeration.

1.1 Correlation Analysis

Correlation coefficient has become the workhorse of quantitative research and analysis. Relationships among things are often examined in terms of whether they change together or separately. The world around us is understood through the multifold and interlaced correlations it manifests. The bootstrap method discussed in this paper is applied to measure linear association in paired, exchangeable observations, see Agresti (1992), Fahoome (2002). Exchangeability is a generalization of the concept of independent, identically distributed random variables. Bootstrap analysis of correlation assumes that in the null hypothesis, two variables X and Y (��$_i$ ∈ ��; ��$_i$ ∈ ��); are independent within each individual unit and pairs (�� , ��$_i$), $i = 1,2, \ldots, n$ are independent and identically distributed. Paired, exchangeable observations (��$_1$; ��$_1$), have the same distribution as (��$_1$; ��$_1$), and the marginal distributions of ��$_1$ and ��$_1$ are identical. A test of exchangeability of paired observations is given by Hollander (1971). Computational advances involving the use of bootstrap tests are well documented in Thompson (1993), Good (2000), Hesterberg et al (2003) and Berger (2006).

The two most commonly used correlation coefficients are the Pearson's correlation coefficient �� and the Spearman's rank correlation coefficient. Given the observations (��, ��$_i$), $i = 1, 2, \ldots$, the Pearson's correlation coefficient is defined as

$r =$ Σ x��−��� �� / �[(Σ��^{2}−����^{2})(Σ ��^{2}−����^{2})] (1)

When r is calculated from sample data, the obtained value is only an estimate of a corresponding population correlation coefficient, denoted by ��. To test the null hypothesis of no correlation, for example, ��$_0$: $\rho = 0$, we assume that both variables are measured on an interval or ratio scale. The calculation is based on the actual values and both variables (X and Y) have a normal distribution. If all the assumptions are met and ��$_0$: $\rho = 0$ is true, then, for n pairs of observations, $t =$ ��√��−1 has the t distribution with $n - 2$ degrees of freedom. A more general $\sqrt{1-��^2}$

way to test $��_0: \rho = ��_0$ or construct confidence intervals for ρ is based on Fisher

Z

transformation, $Z = \frac{1}{2}\ln\frac{1+��}{1-r}$. Z is approximately normal with $\varphi = (Z - \mu_2)\sqrt{n-3}$ having approximately the standard normal distribution, see Freund (1992).

To calculate the rank-correlation coefficient for n pairs of observations, find the sum of the squares of the differences, ��, between the ranks of the X's and Y's, and substitute into the formula

$$r = 1 - \frac{6\sum ��^2}{��(��^2-1)} \qquad (2)$$

When there are ties, assign to each of the tied observations the mean of the ranks which they jointly occupy. When using r to test the null hypothesis of no correlation between two variables X and Y, we do not have to make any assumptions about the nature of the populations sampled. To test the null hypothesis, the statistic, $Z = \frac{��_{�} - 0}{1/\sqrt{��-1}} = r_s/\sqrt{n-1}$, and this approximate the standard normal distribution.

2. Material and Methods

2.1 The Bootstrap Algorithm for Correlation

The p-value of a test statistic represents the probability of obtaining values of the test statistic that are equal to or more extreme than the observed value of the test statistic. In this paper, consideration is given to the bootstrap distribution of paired observations on which the correlation coefficient is to be computed, Bejerano (2003), Demiralp et al (2008). Given a bivariate sample $(��_1, ��_1), (��_2, ��_2), \ldots, (��_n, y_n)$ for which $(��_1, ��_2, \ldots, ��_n) \sim F_X$ and $(��_1, ��_2, \ldots, y_n) \sim F$ are simultaneously tested in an experiment with R as the test statistic. Let $��_0: F_X = F_Y$ against $��_1: F_X \neq F_Y$ or $��_1: F_X < F_Y$ or $��_1: F_X > F_Y$. For all $(n + ��)$

possible

bootstrap sample sizes, systematically develop a pattern required to generate bootstrap samples

as follows:

i. Combine the two sample sizes from the probability distribution (X and Y) as:

$$\theta = (x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n) \quad (3)$$

ii. Draw a bootstrap sample of size $m + n$ with replacement from step(i) to obtain

$$\theta^* = (x_1^*, x_2^*, \dots, x_n^*, y_1^*, y_2^*, \dots, y_n^*) \quad (4)$$

iii. Let B be the number of bootstrap samples. Independently repeat step (ii) a large B of times and obtain bootstrap replications.

iv. Separate the bootstrap sample in step (ii) into two parts according to the sample sizes of x and y respectively and compute the bootstrap Pearson correlation coefficient (r^*) as:

$$r^* = \frac{\sum x^* y^* - n \bar{x}^* \bar{y}^*}{\sqrt{[(\sum x^{*2} - n \bar{x}^{*2})(\sum y^{*2} - n \bar{y}^{*2})]}} \quad (5)$$

and the bootstrap rank-correlation coefficient

$$r_s^* = 1 - \frac{6 \sum d^{*2}}{n(n^2 - 1)} \quad (6)$$

The bootstrap replications obtained from step (iii) is represented as

$$r^{*1}, r^{*2}, \dots, r^{*B} \text{ and, } r_s^{*1}, r_s^{*2}, \dots, r_s^{*B} \quad (7)$$

v. Bootstrap p-value (P_{boot}) is calculated by the fraction of times the theoretical Pearson's correlation coefficient exceeds the correlation coefficient in the bootstrap replications and this is denoted by:

$$(P_{boot}) = \frac{1}{B} \#\{(r^*) \le (r)\} \text{ or } (P_{boot}) = \frac{1}{B} \#\{(r_s^*) \le (r_s)\} \quad (8)$$

A practical implementation of the bootstrap algorithms developed in section 3 shall be carried out in what follows.

Suppose a two-sample experiment is drawn from the same population distribution. If the first set denotes x_1, x_2, x_3, x_4, x_5 and the second set denotes y_1, y_2, y_3, y_4, y_5, and observed that $n = 5$ and $n = 5$, a pool of the observations into a single sample yield:

$$\theta = (x_1, x_2, x_3, x_4, x_5\ y_1, y_2, y_3, y_4, y_5) \quad (9)$$

The number of bootstrap samples with replacement from the original sample (9) is:

$$\theta^* = (x_1^*, x_2^*, x_3^*, x_5^*, y_2^*, y_3^*, x_2, y_2, y_4, y_1) \quad (10)$$

Note that the actual process involved in obtaining (10) is that, randomly assign them to an ordered $(x_1^*, \dots, x_n^*, y_1^*, \dots, y_n)$. So that an x_i^* can also be assigned the value of y_j in the original sample, and vice versa.

Separate (10) into two parts according to the size of x and y from the original sample, so that it becomes

$$x^* = (x_1^*, x_2^*, x_3^*, x_5, y_2)\ and\ y^* = (y_3^*, x_2^*, y_2^*, y_4, y_1) \quad (11)$$

If $B = 10$ for instance, repeat process (11) 10 times to obtain bootstrap configurations in Table 1, and Figure 1 illustrate the flow chart on how to estimate the statistic of interest from the set of bootstrap sample generated. To obtain the bootstrap p-value (P_{boot}), compute the bootstrap correlation coefficient (r^*) or (r_s^*) and count the number of r^* or r_s^* from the bootstrap data that is less than or equal to the Pearson's correlation coefficient (r) of the original data and then divide by the number of bootstraps sample performed, and vice versa for a right-tailed test.

$$\text{i.e.}(P_{boot}) = \frac{1}{10}\#\{r^* \leq r\}\ or\ \frac{1}{10}\#\{r_s^* \leq r_s\} \quad (12)$$

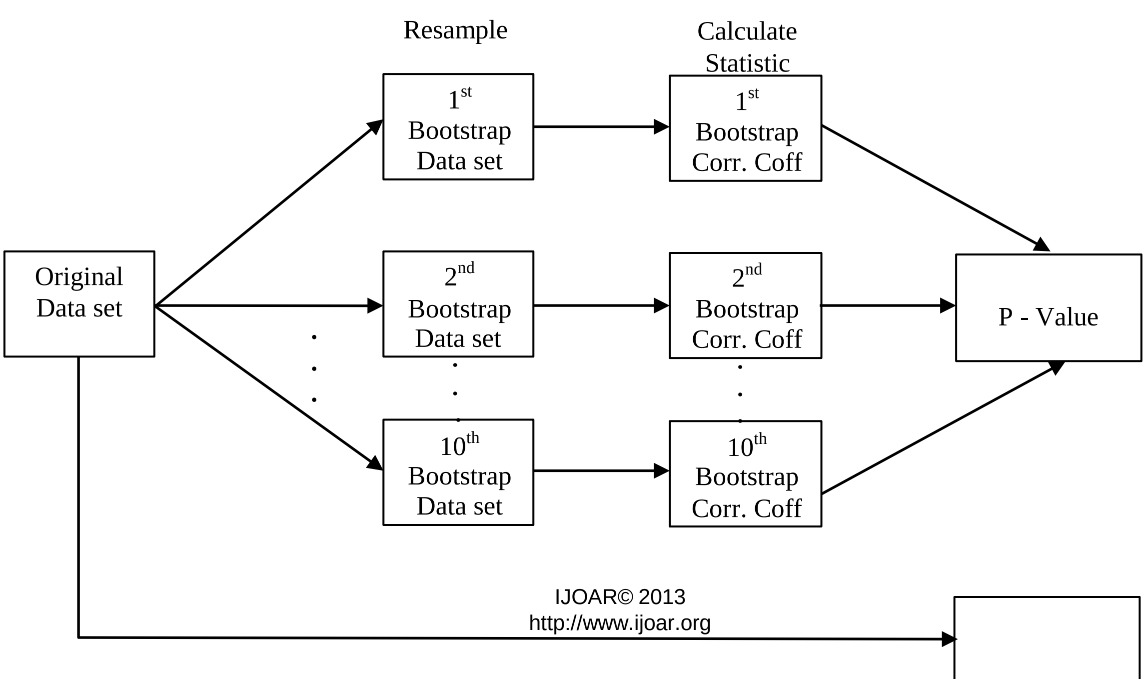
Note from Table 1 that each of the rows: 1 – 10 on the 1st column represent the number of bootstrap configurations. The 2nd and 3rd columns represented by x^* and y^* are the separated bootstrap samples in accordance with the sample sizes; n = 5 and n = 5 respectively. The 4th and the 5th columns denoted by (��*) and (��$^*_s$) represent the bootstrap Pearson and rank

correlation

coefficient.

Table 1: Bootstrap Replication for a Two Sample Experiment

S/No	x^*					y^*					r^*	��$^*_s$
1	��$^*_1$	��$^*_2$	��$^*_3$	��$^*_5$	��$^*_2$	��*	��$^*_2$	��$^*_2$	��$^*_1$	��$^*_2$	r^*_1	r^*_{s1}
2	��*	��$^*_1$	��$^*_3$	��$^*_1$	��$^2_1$	$^*_3$	��$^*_3$	��$^2_1$	��$^*_4$	��$^*_3$	r^*_2	r^*_{s2}
3	$_4^*$	��*	��$^*_5$	��$^*_2$	��$^*_1$	* ��	��$^*_1$	��$^*_5$	��$^*_2$	��$^*_4$	r^*_3	r^*_{s3}
4	��$^*_2$	$_5^*$	��*	��$^*_3$	��$^*_1$	*	��$^*_1$	��$^*_1$	��$^*_3$	��$^*_3$	r^*_4	r^*_{s4}
5	��$^*_3$	��$^*_1$	��$^*_4$	��$^*_5$	��$^*_3$	��$^*_2$	��$^*_4$	��$^*_3$	��$^*_3$	��$^*_5$	r^*_5	r^*_{s5}
6	��*	��$^*_2$	��$^*_4$	��$^*_4$	��$^*_2$	��*	��*	��$^*_4$	��$^*_2$	��$^*_1$	r^*_6	r^*_{s6}
7	��$^*_2$	��$^*_1$	��$^*_3$	��$^*_1$	��$^*_2$	��*	��$^*_2$	��$^*_2$	��$^*_2$	��$^*_1$	r^*_7	r^*_{s7}
8	��*	��*	��$^*_1$	��$^*_2$	��$^*_4$	��$^*_3$	��*	��$^*_4$	��$^*_5$	��$^*_5$	r^*_8	r^*_{s8}
9	��$^*_1$	��$^*_5$	��*	��$^*_4$	��$^*_5$	��$^*_5$	��*	��$^*_5$	��$^*_1$	��$^*_2$	r^*_9	r^*_{s9}
10	��*	��*	��*	��$^*_3$	��$^*_2$	��$^*_4$	��$^*_2$	��$^*_2$	��$^*_2$	��$^*_3$	r^*_{10}	r^*_{s10}



Corr. Coff.
Original D
t

Figure 1: Flow Chart Illustrating Bootstrap Method

When implementing the Spearman's rank correlation coefficient, the ranking of the two samples is done independently and the ranks so obtained retain the positions of their respective observations. Therefore, any exchange of observations in any pair will result in a fresh ranking of the two samples. When ties exist, the mean rank of the tied observations is assigned to each of the tied observations. Algorithm 1 depicts the procedure for generating ranks for the tied and untied observations as required by the Spearman's rank correlation coefficient. After independently sorting each sample in ascending order of magnitude, the algorithm ranks the observations and also takes care of tied observations. The algorithm presented in this paper can carry out a large replication of n-paired bootstrap samples by making the necessary adjustments to reflect the number of pairs.

3 Results and Discussion

The algorithms were implemented in Visual Basic code. The paired bootstrap p-values generated for the Pearson's and the Spearman's correlation coefficients are presented in Table 2 along with their theoretical results for the scores of 15 students in Statistics (X) and Computer Science (Y) as presented in Appendix 2. The algorithms can be applied to any sample size and the statistic of interest is computed each time a new bootstrap sample is generated. The scatter diagram and the bootstrap distribution of Pearson and Spearman's correlation coefficient for scores of paired students are displayed in Figures 2 and 3 respectively. Critical values for the bootstrap distribution of the Pearson's and Spearman's rank correlation coefficient for the scores of paired students are presented in Table 3 and Table 4.

Table 2: P-values for correlation coefficients (1-tailed)

Correlation	Coefficient	Theoretical p-value	Bootstrap p-value
Person	0.710040	0.00101	0.00100
Spearman rank	0.641964	0.00410	0.004667

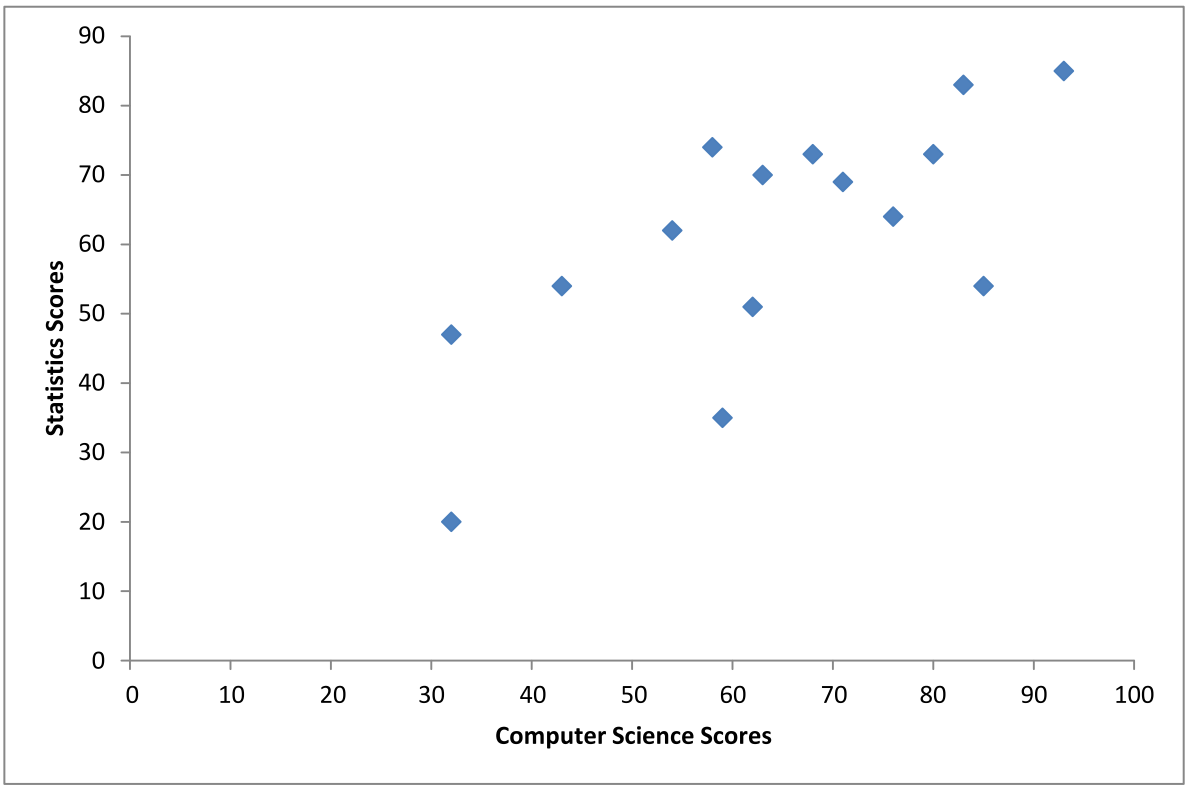


Fig.2. Scatter diagram of the scores of paired students

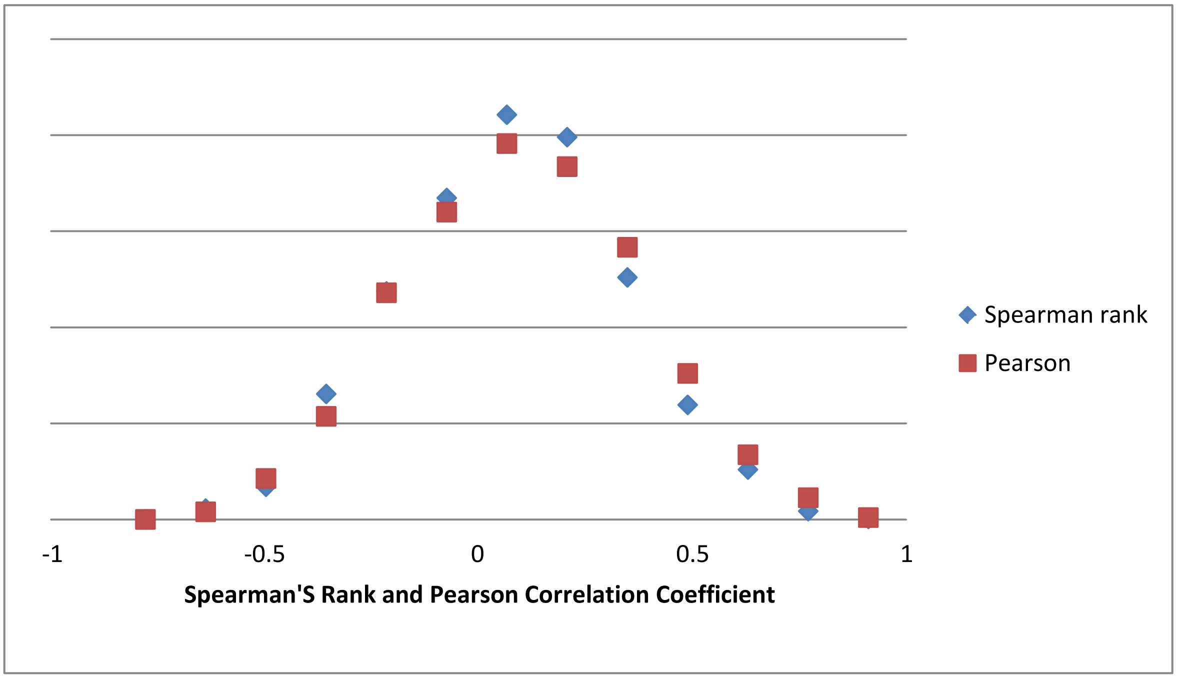


Fig.3. Distribution of Pearson and Spearman's correlation coefficient for scores of paired students

Table 3: Lower critical values C_α for r and r_s

(If ��$'$ $\leq$ ��, then �� $=_{\alpha} C^{\geq}$; $>$ ��, then �� $=_{\alpha} C^{>}$)

if ��$'$ Correlation Coefficient	α									
	0.001 ��$_\alpha$	��$'$	0.0025 ��$_\alpha$	��$'$	0.005 ��$_\alpha$	��$'$	0.01 ��$_\alpha$	��$'$	0.025 ��$_\alpha$	��$'$
Pearson (r)	-0.72	0.001	-0.67	0.0033	-0.61	0.0053	-0.58	0.01	-051	0.0253
Spearman (r)	-0.72	0.001	-0.71	0.0027	-0.64	0.0057	-0.59	0.01	-0.48	0.0283

Correlation Coefficient	α			
	0.05 ��$_\alpha$	��$'$	0.1 ��$_\alpha$	��$'$
Pearson (r)	-0.44	0.05	-0.35	0.1017
Spearman (r)	-0.43	0.0507	-0.33	0.1067

Table 4: Upper critical values C_α for r and r_s

(If ��′ ≤ ��, then �� $=_{\alpha} C^{\geq}$; > ��, then �� $=_{\alpha} C^{>}$)

if ��′ Correlation Coefficient	α									
	0.001		0.0025		0.005		0.01		0.025	
	��$_\alpha$	��′	��$_\alpha$	��′	��$_\alpha$	��′	��$_\alpha$	��′	��$_\alpha$	��′
Pearson (r)	0.69	0.001	0.67	0.0027	0.64	0.0043	0.60	0.008	0.51	0.0233
Spearman (r)	0.73	0.001	0.70	0.002	0.63	0.0047	0.57	0.0097	0.51	0.024

Correlation Coefficient	α			
	0.05		0.1	
	��$_\alpha$	��′	��$_\alpha$	��′
Pearson (r)	0.43	0.046	0.34	0.0937
Spearman (r)	0.42	0.049	0.33	0.0987

Statistical test is based on calculating the test statistic of interest, comparing the calculated test statistic with a critical value and accepting or rejecting the null hypothesis based on the outcome of the comparison. The critical values are usually determined by cutting off the most extreme 100 α % of the theoretical frequency distribution of the test statistic, where α is the level of significance, see Siegel and Castellan (1988). The p-values presented in Table 2 are approximately the same for the bootstrap approach both for the Pearson and for the Spearman, and theoretical approaches, indicating that the probability of a type I error is not more than ��. It is therefore advisable that the bootstrap test should be employed whenever possible. The critical values displayed in Tables 3 and 4 clearly reveal that correlation analysis can easily be handled by the bootstrap approach. Without difficulty, a nonparametric confidence interval can be constructed for the bootstrap distribution generated. This can be obtained exactly the same way permutation confidence intervals are obtained, see Odiase and Ogbomwan (2007).

4 Conclusion

As promising as the p-value obtained from the bootstrap method, yet it is computationally demanding. The intensive looping in computer programming required to generate a very large bootstrap configuration demands a good programming skill. In general, a straight forward but computer intensive method of computing the bootstrap p-value is given and the resultant value ensured that the probability of making a Type 1 error is approximately α.

REFERENCES

Agresti A. (1992): A Survey of Exact Inference for Contingency Tables. Statistical Science 7(1), 131-177.

Bagui S. and Bagui S (2005): An Algorithm and Code for Computing Exact Critical Values for Friedman's Nonparametric ANOVA (Visual Basic). Journal of Modern Applied Statistical Methods. 4(1): 312 – 318

Bejerano G. (2003): Efficient Exact P-value Computation and Applications to Biosequence Analysis. Proceedings of the seventh Annual conference on Research in Computational Molecular Biology, Berlin. pp 38-47

Berger (2006): Introduction to Resampling Techniques @ http://wfs.cgu.edu/bergerd/papers/resampling.xls

Demiralp S, Hoover K.D. and Perez S.J. (2008): A Bootstrap Method for Identifying and Evaluating a Structural Vector Auto-regression, Oxford Bulletin of Economics and Statistics 70, 509-533.

Efron B. (1979): Bootstrap methods; another look at the jackknife. The Annals of Statistics, 7, 1-26

Efron B. (1982): The jackknife, the bootstrap and other resampling plans. Society of Industrial and Applied Mathematics, CBMS-NSF Monographs, 38.

Efron B. and Tibshirani R.J.(1993): An introduction to the bootstrap. New York; Chapman and Hall.

Fahoome G. (2002): Twenty nonparametric statistics and their large sample applications. Journal of Modern Applied Statistical Methods 1(2), 248-268.

Freund T.E. (1992): Mathematical Statistics 5th edition. Prentice Hall, Englewood Cliffs, New Jersey.

Hesterberg T.C., Monaghan S., Moore D.S., Clipson A., and Epstein R. (2003): Bootstrap methods and Permutation tests. W.H. Freeman, New York.

Hollander, M. (1971): A nonparametric test for bivariate symmetry, Biometrika, 58, 203– 212.

Odiase J.I. and Ogbonmwan S.M.(2005): An algorithm for generating unconditional exact permutation distribution for a two-sample experiment. Journal of Modern Applied Statistical Methods. 4(1): 319-332.

Odiase J.I. and Ogbonmwan S.M. (2007): Correlation analysis: exact permutation paradigm. Mathematic Sieve, 59, 161-170.

Pesarin F. (2001): Multivariate Permutation Tests, Wiley, New York.

Sideridis G. D. and Simos P. (2010): Approximating the Sampling Distribution of the Correlation Coefficient using the Bootstrapping Method. The International Journal of Educational and Psychological Assessment, 5, 117 – 133.

Siegel, S. and Castellan, N. J. (1988): Nonparametric Statistics for the Behavioral Sciences (2nd edition), McGraw-Hill, New York.

Thompson, B. (1993). The use of statistical significance research: Bootstrap and other alternatives. The Journal of Experimental Education, 61, 361-377.

Appendix I: Bootstrap Algorithm for Spearman rank and Pearson Correlation

Step 1	Set value to X, N
Step 2	For I ← 1 to N do Step 3
Step 3	Set value to X(I)
Step 4	Set value to Y, N
Step 5	For J ← 1 to N do Step 6
Step 6	Set value to Y(I)
Step 7	[Set number of iteration T]
Step 8	For K ← 1 to T+1 do through Step 9 to Step 45
Step 9	Set x ← Round(Rnd(A(N+N)*N+N)
Step 10	For I ← 1 to N do through Step 11
Step 11	Set rank A(I) ← I
Step 12	For temp A ← X(J) do through Step 13 to Step 14
Step 13	Set X(J) ← X(J+1)
Step 14	Set X(J+1) ← temp A
Step 15	For trank A ← rank A(J) do through Step 16 to Step 17
Step 16	Set rank A(J) ← rank A(J+1)
Step 17	Set rank A(J+1) ← trank A
Step 18	For I ← 1 to N do Step 19
Step 19	Set rankB(I) ← I
Step 20	For temp B ← y(J) do through Step 21 to Step 22
Step 21	Set y(J) ← y(J+1)
Step 22	Set y(J+1) ← temp B
Step 23	For trank B ← rank B(J) do through Step 24 to Step 25
Step 24	Set rank B(J) ← rank B(J+1)
Step 25	Set rank B(J+1) ← trank B
Step 26	For I ← 1 to N do through Step 27 to Step 28
Step 27	Set diff (I) ← rank A (I) - rank B (I)
Step 28	Set diff Sqr (I) ← diff (I) * diff (I)
Step 29	For SumdiffSqr (I) ← 0 do step 30
Step 30	Set SumdiffSqr ← SumdiffSqr + diffSqr(I)
Step 31	For I ← 1 to N do Step 32
Step 32	Set rankcorrel ← 1 – ((6*SumdiffSqr)/(N*(N*N-1)))

Step 33	For I ← 1 to N do through Step 34 to Step 37
Step 34	Set Suma ← suma + x(I)
Step 35	Set Sumb ← sumb + y(I)
Step 36	Set avea ← suma /N
Step 37	Set aveb ← sumb/N + x(I)
Step 38	For xy (I) ← x(I) * y(I) do through step 39 to step 40
Step 39	Set X2 (I) ← x(I) * x(I)
Step 40	Set Y2 (I) ← y(I) * y(I)
Step 41	For Sumab ← sumab + xy(I) do through step 42 to step 43
Step 42	Set Suma2 ← suma2 + X2(I)
Step 43	Set Sumb2 ← sumb2 + Y2(I)
Step 44	For I ← 1 to N do through Step 45 to Step 46
Step 45	Set pcorrel ← (sumab – N*avea*aveb)/Sqr((suma2 – N*avea*avea)* (sumb2 – N*aveb*aveb))
Step 46	[Write out values: k, x(k), y(k), rankcorrel(k), pcorrel(k)]

Appendix 2:

The following data representing the Statistics grades, x, and Computer Science grade, y, of 15 students in a course in Statistics

x: 68, 54, 80, 62, 43, 32, 63, 71, 83, 59, 93, 85, 76, 32, 58

y: 73, 62, 73, 51, 54, 20, 70, 69, 83, 35, 85, 54, 64, 47, 74

Source: Department of Mathematics, Ambrose Alli University, Ekpoma, Edo State, Nigeria.