African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 5, Issue 1, 2022 (pp. 55-66)
www.abjournals.org

ESTIMATING THE RELIABILITY INDEX USING CONFIDENCE INTERVAL: A COMPERISM OF THE FISHER Z AND THE BOOTSTRAP CONFIDENCE INTERVAL METHOD

Imasuen Kennedy[1] and Dr. (Mrs.) U. Matilda Orheruata[2]

[1]Institute of Education, University of Benin, Nigeria
Email: kennedy.imasuen@uniben.edu; Tel: +234 0812 896 3837

[2]Department of Educational Evaluation and Counselling Psychology, University of Benin, Nigeria
Email: mati.orheruata@uniben.edu

Cite this article:

Imasuen K., Orheruata U. M. (2022), Estimating the Reliability Index Using Confidence Interval: A Comperism of the Fisher Z and the Bootstrap Confidence Interval Method. African Journal of Mathematics and Statistics Studies 5(1), 55-66. DOI: 10.52589/AJMSS-ZMRAGI1J.

Manuscript History

Received: 29 Dec 2021

Accepted: 25 Jan 2022

Published: 30 March 2022

ABSTRACT: *The study focused on estimating the reliability index using confidence interval by comparing the Fishers Z and the bootstrap confidence interval methods. The rational for the study was to examine the bootstrap and the Fisher –Z methods and finding out the better of the two. The population of the study consists of the senior secondary school students in Egor local government area, Edo State. There are a total of 12 school with 8,207 students. A sample size of 410 representing 5% of the total population of students were randomly selected from the 12 schools. The instrument for data collection was the Open Hemisphere Brain Dominance Scale 1.0 (OHBDS) a personality scale designed by Eric Jorgenson (2015). It was adapted for the study. The instrument was validated. The reliability was part of the issues raised in the study. The data were analyzed using the Pearson Product-Moment Correlation Coefficient to determine the reliability. The Fishers Z 95% and the Bootstrap (percentile and bias corrected and accelerated) confidence interval were also used. The findings revealed that as the sample size became large, the length of the interval became narrower; the three methods utilized in this study yielded the same length of the interval (width) when the same size was 100 and 150; and as the sample size increases, the bias corrected and accelerated bootstrap gave a shorter interval length, thereby becoming the best of the three method considered in the study. It was therefore recommended that reporting reliability should be based on interval estimation as against the point estimate, the sample size should be at least 100 and the bootstrap confidence interval should be adopted as it is not liable to the normality condition associated with the classical statistics.*

KEYWORD: Reliability, confidence interval, bootstrap, percentile.

Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 5, Issue 1, 2022 (pp. 55-66)

www.abjournals.org

INTRODUCTION

Reliability is one of the psychometric properties of any measuring instrument. It deals with how stable or consistent, the scores of any measuring instrument if used severally. The classical test theory provides a good explanation of reliability. It posits that there is an observed score (X) made up of true score (T) and an error score (E). This is represented as

$$X = T + E \tag{1.11}$$

The true score can be conceptualized as part of the observed score not affected by random error. Systematic errors are errors which are part of the use of any assessment; they are always there. They do not affect reliability and are seen as part of the stable component of a person's true score. The true and observed scores obtained by people would differ. Thus the variance of a test can be written as

$$S^2{}_x = S^2{}_t + S^2{}_e \tag{1.12}$$

where $S^2{}_x$ = variance of a group of individuals' observed score

$S^2{}_t$ = variance of a group of individuals' true score

$S^2{}_e$ = error variance in a group of individuals' score

Reliability(r_{xx}) is defined as the ratio of true score and observed score variances:

$$r_{xx} = \frac{S^2{}_t}{S^2{}_x} \tag{1.13}$$

where r_{xx} stands for reliability.

From (1.2) we have $S^2{}_t = S^2{}_x - S^2{}_e$ and substituting into (1.3) we obtain $r_{xx} = \frac{S^2{}_x - S^2{}_e}{S^2{}_x} = \frac{S^2{}_x}{S^2{}_x} - \frac{S^2{}_x}{S^2{}_e}$.

Thus, $r_{xx} = 1 - \frac{S^2{}_e}{S^2{}_x}$ (1.14)

From (1.4), when $S^2{}_x$ remains constant, reliability increases as error variance decreases. When error variance remains constant and we increase $S^2{}_x$, reliability increases also. When (1.4) is solved for S_e we have $1 - r_{xx} = \frac{S^2{}_e}{S^2{}_x}$ which implies that $S^2{}_e = S^2{}_x(1 - r_{xx})$. It follows that

$$S_e = S_x\sqrt{1 - r_{xx}} \tag{1.15}$$

(1.15) is called the standard error of measurement (SEM), a measure of intra-individual variability (Afemikhe, 2014)

Reliability as it applies to test, has two distinct meanings. One refers to stability over time, the second to internal consistency (Kline, 2000). Reliability is the degree to which a test consistently measures whatever it measures. It refers to the extent to which the same result can

Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J

be obtained using the same instruments more than one time. In simple terms, if our research is associated with high levels of reliability, then other researchers need to be able to generate the same results, using the same research methods under similar conditions. Reliability is an indicator of consistency, that is, an indicator of how stable a test score or data is across applications or time. A measure should produce similar or the same results consistently if it measures the same "thing." (Sawilowsky, 2000). A measure can be reliable without being valid but a measure cannot be valid without being reliable (Erfold, 2013).

Correlation coefficient is an important statistical procedure in the determination of the degree of reliability. The measure of the degree and direction of the relationship between two quantitative variables is known correction coefficient (Triola, 2008). This correlation coefficient can either be positive, negative or even neutral (zero). A correlation coefficient of + 1.0 is a perfect positive relationship, - 1.0 is a perfect negative relationship and that of 0.0 indicates no relationship. The nearer a correlation is to +1.0, the more reliable the results. If a measure is perfectly reliable, there is no error in measurement, that is, everything we observe is true score. Therefore, for a perfectly reliable measure, the reliability = 1. Now, if we have a perfectly unreliable measure, there is no true score, that is, the measure is entirely error. In this case, the reliability = 0. The value of a reliability estimate tells us the proportion of variability in the measure attributable to the true score. A reliability of 0.5 means that about half of the variance of the observed score is attributable to truth and half is attributable to error. A reliability of 0.8 means the variability is about 80% true ability and 20% error (National Council on Measurement in Education (NCME), American Educational Research Association (AERA), and American Psychological Association (APA), 2014). All measurement procedures involve error. However, it is the amount/degree of error that indicates how reliable a measurement is. When the amount of error is low, the reliability of the measurement is high. Conversely, when the amount of error is large, the reliability of the measurement is low (Erfold, 2013; Meyer, 2010).

Reliability of test scores may be affected by the variation in conditions under which the test is administered, errors of sampling, item difficulty, range of the group, ability level of those who took the test, length of the test, operations for estimating the reliability and faulty marking procedure (Kline, 2000). It is fundamental to note that reliability refers to the result and not the test itself. The samples from which the reliability coefficient are derived must be representative of the population for whom the test is designed and sufficiently large to be statistically reliable (Learn & Ken, 2012). According to Kline (2000), a reliability of 0.7 is a minimum for a good test. This is simply because the standard error of measurement (which is the estimated standard deviation of scores) of scores increases as the reliability decreases, and thus tests of low reliability are useless for practical application, where decision concerning individuals have to be made. Where repeated measurement is required, high parallel form reliability is useful.

Reliability estimates are usually obtained using various techniques or methods which includes the test –retest, to establish stability, split half and Cronbach alpha for internal consistence.

There are many methods/ techniques of obtaining the confidence interval for a reliability estimate, but for this study we shall examine the Fisher Z and the bootstrap method, and make a comparism of the two methods so as to ascertain the methods that will give a better interval.

Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 5, Issue 1, 2022 (pp. 55-66)

www.abjournals.org

Confidence intervals

In other to solve the problem associated with the point estimate of repointing reliability, that is the problem of not having a way of knowing how close a particular point estimate is to the population parameter, led to the call for an interval estimate (Bluman, 2008). An interval estimate of a parameter is an interval or a range of values used to estimate the parameter. The estimate may or may not contain the value of the parameter being estimated. In interval estimate, the parameter is specified as being between two values and a degree of confidence (usually in percentage) can be assigned between an interval is made.

The confidence level of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter, assuming that a large number of samples are selected and that the estimation process on the same parameter is repeated (Bluman, 2008).

A confidence interval is a specific interval estimate of a parameter determine by using data obtained from a sample and by using the specific confidence level estimate. Chernick and LaBudde, (2011), stated that one of the properties of the confidence interval is that if random sampling were repeated a number of times (that is infinitely), it is possible that 100(1-α)% of the points generated which represent the confidence interval will contain the true parameter.

Most researchers and authors usually reports reliability as point estimates not taking into cognizance the uncertainties of the estimate when assessing their test (Fan &Thompson, 2001). Most times, the reported point estimate in reliability may be lower or even higher when the confidence interval approach is employed (Imasuen & Omorogiuwa, 2019). Alluding to this, Kelly and Cheng (2012), stated that confidence interval is more important than the point estimate. Often times, researchers, tends to forget or ignore the sampling errors associated with the point estimate of reliability (Evers, Lucassen, Meijer & Sijtsma, 2010). On their part, Maxwell, Kelly and Rausch, (2008) opined that sample size plays an important role in the usage of the confidence interval estimate of reliability.

The Fisher Z- Transformation

Fisher (1958) devised a mathematical transformation Z of r that has an approximately normal sampling distribution irrespective of ρ or n. According to Glass and Hopkins (1995), this transformation known as the Fisher's Z transformation is a trigonometric function of r given as

$$Z = tanh^{-1}r \tag{1.16}$$

In natural logarithms, this can be written as

$$|Z| = \frac{1}{2}\left(\frac{1+|r|}{1-|r|}\right) \tag{1.17}$$

Which is approximately normally distributed with mean $\mu_{zr} = 0.5log_e(\frac{1+\rho}{1-\rho})$

and standard deviation $\sigma_{zr} = \frac{1}{\sqrt{n-3}}$

Fisher showed that the sampling distribution of Z for samples from a bivariate normal distribution approaches normality rapidly unlike the sampling distribution of r. Also, the

Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J

variance of Z is independent of the value of the parameter ρ even if samples are not large (Glass & Hopkins, 1995). This implies that sampling distribution of the Z would be nearly normal with mean equal to the parameter Z that corresponds to ρ, that is Z_ρ. The standard error of Fisher Z (that is the standard deviation of the sampling distribution) is determined only by n. The standard error of Fisher Z, σ_z is given as

$$\sigma_z = \frac{1}{\sqrt{n-3}}. \quad (1.18)$$

Fisher's Z- transformation provides what is needed for a solution to the problem of placing confidence interval (C.I) around r (Glass & Hopkins, 1995)

Steps in forming confidence interval using Fisher's –Z transformation

Step 1

Transform r to Z_r

Step 2

Compute σ_z, where $\sigma_z = \frac{1}{\sqrt{n-3}}$

Step 3

Obtain the required confidence interval for Z_ρ: $Z_r \pm (1-\alpha)\%\sigma_z$

Step 4

Transform lower and upper limits of $(1-\alpha)\%\sigma_z$ confidence interval for Z_ρ to r to obtain $(1-\alpha)\%\sigma_z$ confidence interval for ρ, where α is the level of significance.

Bootstrap Resampling Method

The bootstrap method is a non-parametric sampling techniques proposed by Efron and Tibshirani in 1985. It is a non-parametric technique for the estimation of the distribution by sampling with replacement from an original data set. According to Ogbonmwan and Imasuen (2004), the bootstrap is a Monte Carlo resampling method and a useful statistical tool for the estimation of sampling distribution of a random variable R(X, Y) and the unknown distribution F, by making use of the realized data set $X = (x_1, x_2, \ldots x_n)$

The bootstrap has come a long way to eliminate the constrain of traditional parametric statistics with its over-reliance on a small set of standard models from which theoretical solution are available (Ogbonmwan & Imasuen, 2004). The bootstrap as non-parametric statistics has been able to take care of the normality condition or assumptions which has constituted a major barrier to the classical statistics. Also it has taken care of the problem of generating smaller samples which cannot give a better result.

Efron (1985) gave the benefit accruing from the bootstrap method as: the bootstrap estimate is invariant under transformation and automatically produce accurate solutions; the bootstrap method does not depend on a lot of assumptions unlike the classical statistics; and the bootstrap method has the useful property that it tends to construct the bias in maximum likelihood

Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J

estimate based on data from a well-balanced but skewed data. Hence, the bootstrap has become an acceptable technique in almost every sphere of statistical endeavor.

The Bootstrap Confidence Interval

The percentile and the bias corrected and accelerated BC_a bootstrap method of Efron in 1985 can also be used to generate confidence interval for correlation coefficient and by inference, reliability estimates.

Percentile Bootstrap Method

Suppose we have a data (x_1, y_1) and bootstrap data set (x^*_1, y^*_1), (x^*_2, y^*_2) …, (x^*_n, y^*_n) generated with replications B. For each of these replication (x^{*1}_1, y^{*1}_1), …, (x^{*1}_n, y^{*1}_n) it is possible to compute the Pearson product-moment correlation coefficient $(r^*_1, r^*, \ldots r^*_B)$. Thus the 100(1-$\alpha$)% bootstrap percentile interval for the population correlation coefficient ρ is given as

$$r^*_{B(\frac{\alpha}{2})} < \rho < r^*_{B(1-(\frac{\alpha}{2}))} \tag{1.19}$$

where $r^*_{B(\frac{\alpha}{2})}$ is the $(B_{(\frac{\alpha}{2})})^{th}$ and $r^*_{B(1-(\frac{\alpha}{2}))}$ is the $(B_{1-(\frac{\alpha}{2})})^{th}$ value in the ordered list of the bootstrap distribution of r* (Efron & Tibshirani, 1993)

In practices, the percentile method of bootstrap confidence interval despite its importance, most times gives erratic results both in terms of the length of interval and of their skewness (Ogbonmwan & Imasuen,2004).

Bias- Corrected and Accelerated BC_a Bootstrap Method

The problems inherent in the percentile method gave rise to Bias- Corrected and Accelerated BC_a Bootstrap Method. The Bias- Corrected and Accelerated BC_a Bootstrap Method is an improvement on the percentile method. In the Bias- Corrected and Accelerated BC_a Bootstrap Method, the interval limit are given by percentiles that depends on both the accelerated $\hat{a}$ and bias correction $\hat{z}_0$

The 100(1-α)% Bias- Corrected and Accelerated confidence interval given by Efron and Tibshirani in John (2019) is given as

$$r^*_{B\alpha 1} < \rho < r^*_{B\alpha 2} \tag{1.20}$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\frac{\alpha}{2}}}{1-\hat{a}(\hat{z}_0 + z_{\frac{\alpha}{2}})}\right) \tag{1.21}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-(\frac{\alpha}{2})}}{1-\hat{a}(\hat{z}_0 + z_{1-(\frac{\alpha}{2})})}\right) \tag{1.22}$$

and $\hat{z}_0$ is usually gotten from the proportion of the correlation coefficient of the B bootstrap replication that is less than the correlation coefficient of the original sample data.

Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J

$$\hat{z}_0 = \Phi^{-1}\left(\#\left(\frac{r^*_B < r}{B}\right)\right) \quad (1.23)$$

and

$$\hat{a} = \frac{\sum_{i=1}^{n} (r_i - r_.)}{6\{\sum_{i=1}^{n} (r_i - r_.)^2\}^{\frac{3}{2}}} \quad (1.24)$$

where (r_i) is the (i)th Jacknife replication of the correlation coefficient r and

$$r_. = \frac{(\sum_{i=1}^{n} (r_i))}{n} \quad (1.25)$$

Statement of the problem

Most times, reporting reliability estimate is usually a difficult task as researchers have failed to agree on the optimal sample size for an acceptable reliability study. Some proposed 30 samples, (Bonnet & Wright, 2014) others at least 100 (Kline, 1986; Nunnally & Bernstein, 1994; Segal, 1994; Charter, 1999; Omorogiuwa & Imasuen, 2018). Moreover, there is the clarion call for reliability to be reported using confidence interval rather than as a single value. They are of the view that point estimate is bedeviled with a lot of problems hence the values should be on an interval having lower and upper bounds, (AERA, APA &NCME, 2014; Omorogiuwa & Imasuen, 2018). That is, reporting reliability should not be as a single value. The interval estimate builds on the concept of the point estimate and in addition conveys the degree of accuracy of the reliability estimate. The interval estimate is a range or band within which the parameter is presumed to lie, with a certain degree of confidence. Thus this study therefore, is to examine the bootstrap and the Fisher –Z confidence interval methods of reliability and compare both with the aim of finding out the better of the two.

Research Question

Is there a difference in test retest reliability estimate of an instrument using the fisher Z and the Bootstrap confidence interval for the sample sizes of 20, 30, 40, 50, 100, 200, 300 and 400?

Purpose of the Study

The purpose of this study is to analyze the sample sizes in test-retest using the Fisher Z and Bootstrap confidence interval methods.

Significance of the Study

The findings of the study will help psychometricians, educators and researcher to be aware of the minimum sample size in carrying out reliability studies. The finding will be an eye opener, to the use of confidence interval in reporting reliability index. Furthermore, this study will be beneficial for researchers, and other stake holders who may be having problems of choosing the appropriate sample size for reliability study.

Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 5, Issue 1, 2022 (pp. 55-66)

www.abjournals.org

METHODOLOGY

The study is correlational with a survey design. The population of the study consists of the senior secondary school students in Egor local government area, Edo State. There are a total of 12 school with 8,207 students. A sample size of 410 representing 5% of the total population of students were randomly selected from the 12 schools. The instrument for data collection was the Open Hemisphere Brain Dominance Scale 1.0 (OHBDS) a personality scale designed by Eric Jorgenson (2015). It was adapted for the study. It consists of a twenty (20) inventory items which was designed to hypothesized left-brain preferences among students, with a 4 point Likert scale. The items were under the options of response: SA= Strongly Agree, A = Agree, D = Disagree, SD = Strongly Disagree. SD was scored 1 point, D was scored 2 points, A was scored 3 points and SA scored 4 point.

The instrument was validated by Eric Jorgenson, but it was further validated by experts in measurement and evaluation. The content and face validity were used. The reliability of the instrument was part of the issues raised in the study. The instrument was administered to the senior secondary school students in the sampled schools. The mode of answering was explained to the students. The schools were visited twice to administer the questionnaire after an interval of two weeks. The reliability coefficient was estimated using the Pearson Product-Moment Correlation Coefficient. The Fishers Z 95% and the Bootstrap (percentile and bias corrected and accelerated) confidence interval were used. 1000 bootstrap samples were used. The width of the interval for the two methods for the various samples was determined. The sample size (s) with a shorter interval was adjudged to be the most suitable and stable.

RESULTS

Research question

Is there a difference in test retest reliability estimate of an instrument using the fisher Z and Bootstrap confidence interval for the sample sizes of 20, 30, 40, 50, 100, 200, 300 and 400?

Table 1: Fisher Z and Bootstrap Confidence Interval for Reliability Estimates for Sample sizes of 20,30,40,50,100,150, 200, 300 and 400

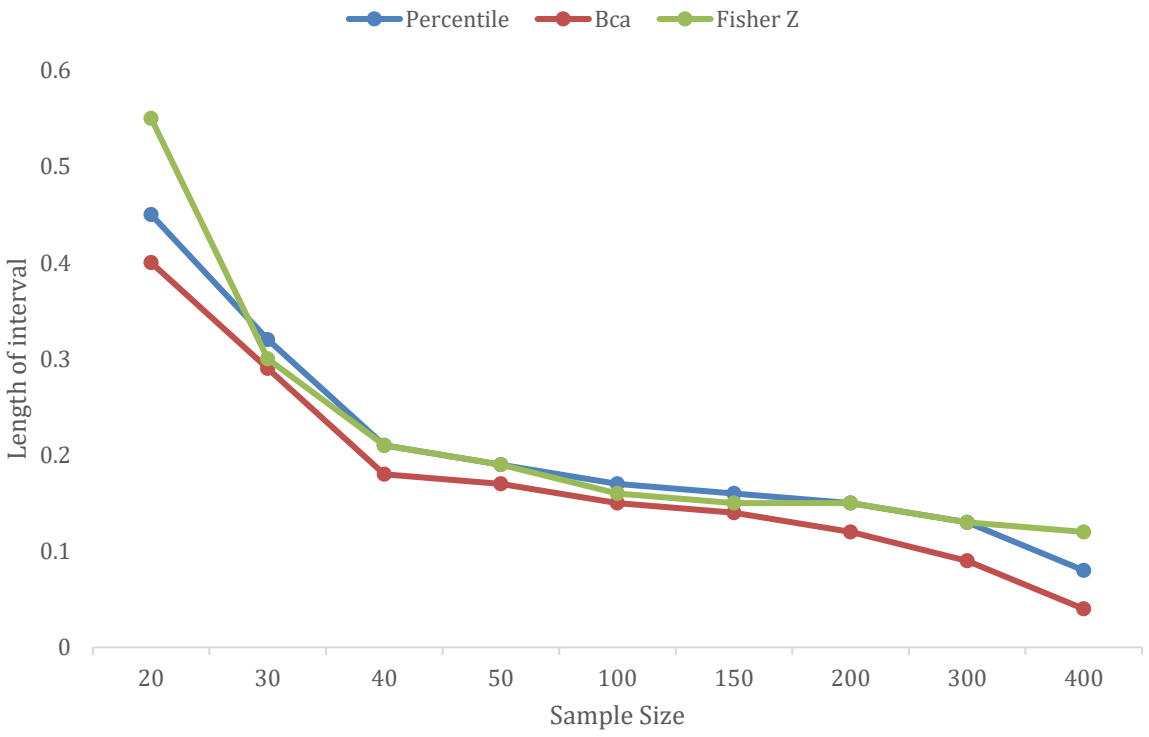
Sample size	Pearson r	Percentile Method			Bias corrected and accelerated			Fisher Z		
		LWB	UPB	WD	LWB	UPB	WD	LB	UB	WD
20	0.66	0.41	0.86	0.45	0.45	0.85	0.40	0.30	0.85	0.55
30	0.68	0.48	0.80	0.32	0.49	0.78	0.29	0.45	0.75	0.30
40	0.71	0.55	0.76	0.21	0.61	0.79	0.18	0.55	0.76	0.21
50	0.72	0.58	0.77	0.19	0.60	0.77	0.17	0.57	0.76	0.19
100	0.73	0.70	0.87	0.17	0.71	0.86	0.15	0.70	0.86	0.16
150	0.75	0.70	0.86	0.16	0.71	0.85	0.14	0.72	0.87	0.15
200	0.76	0.70	0.85	0.15	0.72	0.86	0.12	0.73	0.88	0.15
300	0.77	0.74	0.87	0.13	0.77	0.88	0.09	0.74	0.88	0.13
400	0.78	0.79	0.88	0.08	0.86	0.90	0.04	0.76	0.88	0.12

Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 5, Issue 1, 2022 (pp. 55-66)
www.abjournals.org

LWB = Lower Bound; UPB = Upper Bound; WD = Width or length of the interval

Table 1 shows the values of the lower bound, upper bound and width (length of the interval) for the bootstrap percentile, bootstrap bias corrected and accelerated, and the Fisher Z methods of the 95% confidence interval for the reliability index using the test retest approach. The sample size of 20 gave a Pearson r value of 0.66. 0.45, 0.40 and 0.55 were obtained as the length of the interval for the bootstrap percentile, bias corrected and the Fisher Z confidence interval. With a sample size of 30, the Pearson r value was 0.68. 0.19, 0.17 and 0.30 were obtained as the length of the interval for the bootstrap percentile, bias corrected and the Fisher Z confidence interval. The sample size of 40 gave a Pearson r value of 0.71; with 0.21, 0.18 and 0.21 as the length of the interval for the bootstrap percentile, bias corrected and the Fisher Z confidence interval. When the sample size was increased to 50, Pearson r gave a value of 0.72, with 0.19, 0.17 and 0.19 as the length of the interval for the bootstrap percentile, bias corrected and the Fisher Z confidence interval. The sample size of 100 gave a Pearson r value of 0.73 with 0.17, 0.15 and 0.16 as the length of the interval for the bootstrap percentile, bias corrected and the Fisher Z confidence interval. When the sample size became 150, we obtained Pearson r value of 0.75 with 0.16, 0.14 and 0.15 as the length of the interval for the bootstrap percentile, bias corrected and the Fisher Z confidence interval. The sample size of 200 gave a Pearson r value of 0.76 with 0.15, 0.12 and 0.15 as the length of the interval for the bootstrap percentile, bias corrected and the Fisher Z confidence interval. When sample size was increased to 300, Pearson r value became 0.77 with 0.13, 0.09 and 0.13 as the length of the interval for the bootstrap percentile, bias corrected and the Fisher Z confidence interval. The sample size of 400 gave a Pearson r value of 0.78 with 0.08, 0.04 and 0.12 as the length of the interval for the bootstrap percentile, bias corrected and the Fisher Z confidence interval. This is presented in figure 1



Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J

DISCUSSION OF FINDINGS

The study shows that the sample sizes of 20, 30, did not give the acceptable reliability index of ≥ 0.70 (Kline, 2000). However, the sample size of 40 and 50 did gave an acceptable reliability of ≥ 0.70. Interestingly, the lower bound for the three methods applied in this study fell outside the interval. Hence, the sample sizes of 20, 30, 40, and 50 were seen as not sufficient for a reliability study. But as the sample became 100, the reliability of the instrument became stronger. This agrees with the previous report of Kline (2000); Learn & Ken, (2010); and Imasuen & Omorogiuwa, (2018). However, this study disagreed with Bonnet & Wright (2014) who stated that samples as small as 30 could be used to establish reliability of any instrument with a clause that the scale items have strong inter-correlation. But what we see most times is that researchers do not bother to determine the inter-correlation of the items.

The study also revealed that the three methods utilized in this study yielded the same length of the interval (width) when the same size was 100 and 150. This implies that the optimal sample size for reliability study should be at least 100. Also, as the sample size increases, the bias corrected and accelerated bootstrap gave a shorter interval length, thereby becoming the best of the three method considered in the study. This is in tandem with Efron (1985), and Ogbonmwan and Imasuen (2004), who reported in their studies that the bootstrap method gives a better result when the sample size or replications are large. It was discovered that as the sample size became large, the length of the interval became narrower. This was obvious for the three methods considered. Hence the higher the sample size the smaller the sampling error.

CONCLUSIONS

For any instrument to be consider effective and efficient, it must satisfy the psychometric properties of which reliability is one. Therefore, the issue of the reliability of any measuring instrument should not be treated with levity. Many researchers and test developers uses the point estimate of reporting reliability using one value, without taking into cognizance the sampling errors associated with it. This study once again has demonstrated the need to always use a sample size of at least 100 samples. It was revealed that reporting reliability using the interval estimate gave a better result than the point estimate. Moreover, of the three method used in this study, the bootstrap bias corrected and accelerated gave a better result.

RECOMMENDATIONS

Reliability of measuring instrument plays a vital role, if the scores obtained from the instrument are to be trustworthy thereby minimizing the amount of errors. Based on this, we recommend as follows

- Reporting reliability should be based on interval estimation as against the point estimate.
- The sample size should be at least 100 so as the minimize the error in measurement
- The bootstrap confidence interval should be adopted as it is not bound by the normality condition associated with the classical statistics.

Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J

REFERENCES

Afemikhe, O.A. (2014). *Educational measurement and evaluation*. Ibadan: AMFITOP Books

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Bluman, A. G. (2009). *Elementary statistics: A step by step approach* (7th ed.). New York: McGraw-Hill Company

Bonnet, D., & Wright, D.A. (2014). Cronbach alpha reliability: Interval estimation, hypothesis testing and sample size planning. *Journal of Organizational Behaviour*, 36(1): 3 – 5.

Charter, R. A (1999). Sample size requirements for precise estimates of reliability coefficients. *The Journal of General Psychology*, 130 117 – 129.

Chernick M. R,. & LaBudde, R.A. (2011). *An introduction to bootstrap methods with application*. New Jersey: R. John Wiley & Sons Inc.,

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

Efron, B. (1985). Bootstrap confidence interval for a class of parametric problems. *Biometrik.* 72(1): 45 – 58 .

Erfold, B. T. (2013). *Assessment for counsellor* (2nd ed.). Belmont CA: Cengage Wadsworth.

Evers, A. V. M., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2010). *COTAN beoordeling system voor de kwaliteit van test* (COTAN assessment system for quality of tests). Retrieved from https://www.psynip.nl/wp-content/uploads/2016/07/COTAN-Beoorderlingssystem-2010-pdf

Fan, X., & Thompson, B. (2001). Confidence interval for effect sizes, confidence interval about score reliability coefficient please. An EPM guidelines editorial. *Educational and Psychological Measurement*. 61: 517 – 531.

Glass, V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology (*3rd ed.*)*. Boston: Allyn and Bacon

Imasuen, K.,& Omorogiuwa, O. K, (2018). Analysis of sample sizes in test – retest and Cronbach alpha reliability estimates. Unpublished M.Ed. Thesis, Department of Educational Evaluation and Counselling Psychology, University of Benin, Nigeria.

John, O. O. (2019). Confidence interval estimate of the correlation coefficient for age and systolic blood pressure of 20, 30 and 50 Individuals. *Journal of Advances in Mathematics and Computer Science*. 30(2): 1-8, 2019;

Kelly, K. & Cheng, Y. (2012). Estimation and confidence interval formulation for reliability coefficient of homogeneous measurement instruments. *Methodology*. 8: 39- 50.

Kline, P. (1986). *A hand book of test construction. Introduction to psychometric design* (2nd Ed). New York: Methune and Company.

Leann, T., & Ken, K. (2012). Sample size planning for composite reliability coefficient: Accuracy in parameter estimation via narrow confidence intervals. *British Journal of Mathematical and Statistical Psychology*, 65: 371 – 401.

Maxwell, S. E., Kelly, K., & Rausch, T. R. (2008). Sample size planning, for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*. 59: 537 -563

Meyer, P. (2010). *Reliability: Understanding statistics measurement*. New York NY: Oxford University Press.

Nunnally, J. C & Bernstein, I. H. (1994). *Psychometric theory*, (3rd ed.). New York: McGraw-Hill

Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J

Ogbonmwan, S. M., & Imasuen, K (2004). Bootstrap estimates for non-linear regression analysis. *Nigerian Annals of Natural Sciences*, 5(1): 101 – 118.

Sawilowsky, S.S. (2000). Psychometrics versus datametrics: Comments on Vacha-Haase's reliability generalization method and some EPM editorial policies. *Journal of Educational and Psychological Measurement*, 60: 157 – 173.

Triola M.F.(2008) *Elementary statistics with multimedia study guide*. Boston: Perason Education Inc.

Article DOI: 10.52589/AJMSS-ZMRAGI1J
DOI URL: https://doi.org/10.52589/AJMSS-ZMRAGI1J