IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 11, Issue 3 Ver. I (May - Jun. 2015), PP 69-72
www.iosrjournals.org

Effects of Transformations on Significance of the Pearson T-Test

Omokri, Peter
Department of Mathematics/Statistics, School of Applied Sciences
Delta State Polytechnic Ogwashi-Uku, Delta State

Abstract: *Techniques that improve on existing assumptions of normality and largeness of sample size would be of paramount interest as non normality increases with increasing complexity in data gathering techniques. An option of transformation could be either to carry out a normal transformation so as to improve on the normality of the data in other for the pearson and associated t-test to be applicable or to linearize the distribution so as to strengthen the linear relationship that exist between the data, the simulation study carried out on various transformations and sample sizes showed that the square root transformation fared consistently better than the other transformations and the sample size of n=80 proved to be the sample size that optimizes the pearson test.*

Keyword: *Pearson T-Test, Transformation, Discrete Data, Bivariate Correlation*.

I. Introduction

Bivariate data are data in which two variables are measured on an individual. When interest is in measuring the strength of association between such bivariate data, the correlation coefficient may be used. The most common correlation coefficient, called the Pearson product-moment correlation coefficient, measures the strength of the linear relation between variables. Pearson Correlation is employed to ratio and interval ratio data and most importantly when data is large or normally distributed. When the data is not normal and the sample size is small, the nonparametric Spearman rank correlation is useful.

Little work has been done for cases when the distribution of the data is relatively small or unknown. A test of the significance of Pearson's r may inflate Type I error rates and reduce power of the test when data are non-normally distributed. Several alternatives to the Pearson correlation are provided by literature, The relative performance and robustness of these alternatives has however been unclear (Bishara, and Hittner, 2012).

A comparison of the Pearson correlation to other approaches of testing significance of correlation, such as data transformation approaches and Spearman's rank-order correlation, suggests that the sampling distribution of Pearson’s r was insensitive to the effects of non-normality when testing the hypothesis that $\rho = 0$ (Duncan & Layard, 1973; Zeller & Levine (1974)).

Havlicek and Peterson (1977) extended these studies by examining the effects of non-normality and variations in scales of measurement on the sampling distribution of r, and accompanying Type I error rates, when testing $\rho = 0$. Their results indicated the robustness of the Pearson’s r to non-normality, non-equal interval measurement, and the combination of non-normality and non-equal interval measurement.

Edgell and Noon (1984) further expanded these studies by examining a variety of mixed and very non-normal distributions. They found that Pearson’s r was robust to nearly all non-normal and mixed-normal conditions when testing $\rho = 0$ at a nominal alpha of .05. The exceptions occurred with the very small sample size of n = 5, in which Type I error rates were slightly inflated for all distributions. Type I error was also inflated when one or both variables were extremely non-normal, such as with Cauchy distributions Zimmerman, Zumbo, & Williams (2003) for exceptions.

Although nonlinear transformations in many cases can induce normality and enhance statistical power, there are some distribution types for which optimal normalizing transformations are difficult to find.

As non-normality seems to be growing more common with complexity of data gathering techniques, techniques for improving on existing assumptions of normality and largeness of sample size would be of paramount interest. An option of transformation could be either to carry out a normal transformation so as to improve on the normality of the data in other for the pearson and associated t-test to be applicable or to linearize the distribution so as to strengthen the linear relationship that exist between the data. As Data transformation essentially entails the application of a mathematical function to change the measurement scale of a variable that optimizes the linear correlation between the data. The function is applied to each point in a data set — that is, each data point y_i is replaced with the transformed value

This study therefore seeks to determine the effect of transformation and sample size on the on the Pearson t-test.

II. Methodology

The study determines the effect of transformations of discrete distributions on the Pearson t-test. The distributions are: binomial, Poisson, geometric, hypergeometric, and negative binomial. The required transformations are: inverse, square-root, log, arcsine and the Box-Cox. The tests were conducted for various sample sizes of 10, 20, 80 and 160. For each distribution and sample size, 20 variables were simulated and paired in twos without replacement resulting in${}^{20}C_2$ (190) correlation tests. Five types of transformation (Box-Cox. inverse, square root, arcsine, and log) were carried out on each data set. After each test the proportion of significant (2-tailed test of ρ=0) were obtained. The statistical tools and techniques to be used are stated below.

1. **Pearson Moment Correlation**: The general Pearson moment correlation is given by

$$r = \frac{\sum(x-\bar{X})(y-\bar{Y})}{\sqrt{\left[\sum(x-\bar{X})^2\right]\left[\sum(y-\bar{Y})^2\right]}} \text{ or } r = \frac{\sum XY - (N\bar{X}\bar{Y})}{\sqrt{\left(\sum X^2 - N\bar{X}^2\right)\left(\sum Y^2 - N\bar{Y}^2\right)}} \quad (1)$$

2. Pearson - t-test. Oyeka (1990) gave the Pearson product-moment correlation test as

$$t_r^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (3)$$

Where the test is: H_0: ρ = 0. Under H_0, t_r^*follows the Student's t-distribution with (n−2) degrees of freedom, denoted t(n-2)

3. Inverse Transformation: The inverse transformation is given Eze, (2002) as

$$X' = \frac{1}{X} \quad (5)$$

4. Box-Cox Transformation: the Box-Cox transformation used as given by Osborne,(2010) is:

$$X^I = \frac{1}{\sqrt{x}}, \text{if } \lambda \neq 0$$
$$\log(x), \text{if } \lambda = 0 \quad (6)$$

Where λ if the parameter (mean of data)

5. Arcsine Transformation: The arcsine transformation effectively stretches the tails of data. The arcsine transform used is given by Eze (2002):

$$X^I = \sin^{-1}\sqrt{x_i} \quad (7)$$

8. Log Transformation: Eze (2002) gave the log transformation as X^I=log10^x (9)

9. Square Root Transformation: the square root transformation as given by Eze (2002) is
$X^I = \sqrt{x}$

III. Discussion Of Findings

Table 1: Pearson T-Test: Proportion Significant At □ =0.05

DISTRIBUTION	SAMPLE	RAW	INVERSE	SQUAREROOT	LOG	ARCSINE	BOXCOX
BINOMIAL	**10**	0.021053	0.026316	0.031579	0.031579	0.036842	**0.042105**
	20	0.047368	**0.052632**	0.047368	0.047368	**0.052632**	**0.052632**
	80	**0.068421**	0.063158	**0.068421**	0.063158	0.063158	0.047368
	160	0.052632	0.057895	**0.073684**	0.052632	0.047368	0.057895
POISSON DIST	**10**	0.057895	0.057895	**0.063158**	**0.063158**	**0.063158**	0.052632
	20	0.052632	0.057895	0.052632	0.057895	0.057895	**0.063158**
	80	0.073684	0.068421	**0.078947**	0.073684	0.073684	0.073684
	160	**0.084211**	0.068421	0.068421	0.068421	0.068421	0.063158
GEOMETRIC	**10**	**0.063158**	0.057895	0.057895	0.047368	0.047368	0.057895
	20	**0.042105**	0.036842	0.036842	0.036842	0.036842	0.031579
	80	0.047368	**0.068421**	0.047368	0.047368	0.047368	0.042105
	160	0.031579	0.036842	**0.052632**	0.036842	0.036842	0.031579
NEGATIVE BINOMIAL	**10**	0.047368	0.073684	0.052632	0.068421	0.068421	0.078947
	20	0.042105	0.047368	0.047368	0.047368	0.047368	0.047368
	80	0.036842	0.042105	0.042105	0.042105	0.042105	0.047368
	160	0.031579	0.026316	0.031579	0.031579	0.031579	0.031579
HYPERGEOMETRIC	**10**	0.010526	**0.047368**	0.021053	0.021053	0.015789	0.036842
	20	0.042105	0.042105	0.052632	0.057895	0.057895	**0.063158**
	80	0.047368	**0.063158**	0.052632	**0.063158**	0.063158	0.042105
	160	0.042105	0.042105	0.042105	0.031579	0.042105	0.026316

IV. Interpretation Of Result

Extracting the result for the Pearson test will give a true picture of the effects of transformations, distributions and sample size on Pearson t-test.

For binomial, the boxcox transformation has the highest proportion of 0.042105 at n=10 and the least proportion with the raw data. At n=20, the inverse, arcsine, and boxcox have the highest proportion at n=80, the square root transformation gave the highest proportion at n=160 and also doubled as the highest proportion for the binomial distribution.

In the case of Poisson, when n=10, the square root, logarithm and arcsine has the highest proportion of significance of 0.063158 and equal proportions of of 0.052632 in others. At n=20, the boxcox has the highest proportion as the least proportions were observed in the raw data and square root transformation. The square root has the highest proportion of 0.078947 at n=80 and least proportion of 0.0684 in the inverse transformation. The raw data at n=80 for Poisson has the highest proportion for both n=80 and in all Poisson cases.

The geometric distribution has the highest proportions at n=10&20 in the raw data; n=80 in inverse transformation and at n=160 in squareroot transformation. Sample size of n=80 and the inverse transformation was observed to have the highest proportion for the geometric distribution.

The boxcox transformation was observed to have the highest proportion of 0.078947 at n=10 and has the same proportion of 0.047368 in all transformation at n=20 except for the raw data. Similarly, n=160 had equal proportion of 0.031579 except for the raw data. Hence the boxcox transformation was observed to have the highest proportion across all sample sizes and transformation.

For hypergeometric, the inverse transformation has the highest proportion of 0.047368 at r=10 and the least proportion with raw data. At n=20, the boxcox has the highest proportion of 0.063158 and least proportion of 0.042105 with the raw data and inverse transformation. Hence the inverse and logarithmic transformation at n=80 recorded the highest proportion for the hypergeometric distribution.

In summary, the sample size of n=80 consistently performed better or at least equally with the rest sample sizes, and the squareroot transformation was observed to be consistently higher in proportion except for a very few instances. The proportion of significance was consistently higher in proportion when compared to other distribution.

V. Conclusion

In conclusion, the sample size of n=80 proved to be the sample size that optimizes the Pearson t-test since the proportion of significance at n=80 was virtually the highest in all cases followed by n=160 then n=20 The inverse transformation were greater than or equal to in performance when compared to the raw data except for n=80. Also the square root transformation did consistently better than the raw data and inverse transformation except for the case of geometric at n=80, negative binomial at n=10 and hyper geometric at n=10. Further comparison showed that the square root transformation fared consistently better than the raw data, inverse, logarithmic, arcsine and the Boxcox transformation except for few cases. The logarithmic transformation performed equally with the arcsine except for some few other cases.

VI. Recommendation For Further Study

The study was strictly on bivariate discrete distributions. It is therefore recommended that further research be carried out determine how these transformations and correlation tests will fare in marginal normal distributions (e.g. normal-binomial, normal-Poisson etc.) and marginal non-normal discrete distributions (as binomial-Poisson, Poisson-geometric etc.) as well as determine the effect of these transformations on the power of the bivariate correlation test. Better effort should be made to determine the actual value of λ for the boxcox transformation, also increasing the number of iterations may have helped in clearly breaking the ties in the observed proportions in this study.

References

[1]. Beversdorf, L. M and Sa, P. (2011) Tests for Correlation on Bivariate Non-Normal Data. Journal of Modern Applied Statistical Methods, Vol. 10, No. 2, 699-709

[2]. Bishara, A. J., & Hittner, J. B. (2012). Testing the Significance of a Correlation with Non-Normal Data: of Pearson, Spearman, Transformation, and Resampling Approaches. Psychological Methods, 17, 399-417. doi:10.1037/a0028087

[3]. Duncan, G. T., & Layard, M. W. J. (1973). A Monte-Carlo study of asymptotically robust

[4]. tests for correlation coefficients. Biometrika, 60, 551-558

[5]. Ebuh, G.U. and Oyeka, I.C.A. (2012) A Non Parametric Method for Estimating Partial

[6]. Correlation Coefficient. J. BiomBiostat 3(8)http://dx.doi.org/10.4172/2155-6180.1000156.

[7]. Edgell, S., & Noon, S. (1984). Effect of violation of normality on the t test of the correlation coefficient. Psychological Bulletin, 95(3), 576-583.

[8]. Eze, F.C (2002) An Introduction to Analysis of Variance; Lano Publishers, 51 Obiagu Road Enugu.

[9]. Hayes, A. (1996). Permutation Test is not Distribution-Free: Testing H0: ρ = 0. Psychological Methods, 1(2), 184-198.

[10]. Havlicek, L., & Peterson, N. (1977). Effect of the violation of assumptions upon significance levels of the Pearson r. Psychological Bulletin, 84(2), 373-377.

[11]. Osborne, J.O (2010) “Improving your data transformations: Applying the Box-Cox transformation’ Practical Assessment, Research & Evaluation, Vol 15, No 12 Page 4

[12]. Oyeka, I.C.A (2009), “An introduction to Applied Statistical Methods”. Nobern Avocation Publishing Company, Enugu

[13]. Rasmussen, J., & Dunlap, W. (1991). Dealing with nonnormal data: Parametric analysis of transformed data vs. nonparametric analysis. Educational and Psychological Measurement, 51(4), 809-820.

[14]. Zeller, R. A., & Levine, Z. H. (1974). The effects of violating the normality assumption underlying r. Sociological Methods & Research, 2, 511-519.

[15]. Zimmerman, D.W; Zumbo, B. D and Williams R. H (2003)Bias in Estimation and Hypothesis Testing of Correlation Psicológica, 24,133-158.