Thailand Statistician
July 2017; 15(2): 196-202
http://statassoc.or.th
Short communication

Testing for Zero Correlation between Two Uncorrelated Non-Linearly Dependent Random Variables: A Cautionary Note

Umberto Triacca*
Department of Information Engineering, Computer Science and Mathematics,
University of L'Aquila, via Vetoio, 1 (Coppito 1) I- 67100 L'Aquila, Italy.
*Corresponding author; e-mail: umberto.triacca@ec.univaq.it

Received: 29 September 2016
Accepted 10 January 2017

Abstract

A simulation study was performed to analyze the effects of violations of the normality assumption on the t-test of the Pearson correlation coefficient when the variables are not independent, even though the population correlation is zero. Large effects for violations of normality were found. The Type I error rate can be either inflated or deflated with respect to the assumed error rate. A recommendation is made that the use of the t-test be avoided where there are good reasons to believe that a nonlinear relationship exists between the variables.

Keywords: Student's t-test, Pearson correlation coefficient, normality assumption.

1. Introduction

Correlation analysis, expressed by correlation coefficients, is of paramount importance in statistics and psychometric research, and the associated literature is huge. It is widely known among researchers that uncorrelatedness does not imply independence, but the full importance of this fact is not quite as widely appreciated. The goal of this paper is to show that, when there is a nonlinear relationship between variables, the correlation analysis must be utilized more carefully. In particular, we will consider the performance of the traditional t-test for the null hypothesis that the population correlation is zero, when variables are uncorrelated but not independent. In this case the assumption of bivariate normality is necessarily violated. Using a Monte Carlo simulation study, we will show that the effect of non-normality is very serious. The null distribution theory is clearly non-robust.

It is well known that, often, the dependence between the variables inflates Type I error rates for tests of the Pearson correlation coefficient. In these cases, some authors have proposed to consider the t-test as a crude test of the hypothesis of independence (Edgell and Noon 1984). However, this practice can be incorrect since we find situations of nonindependent variables where Type I error rates for Pearson's r are substantially deflated. In our view, this is the major contribution of the article.

The rest of the paper is organized as follows. Section 2 briefly describes the t-test of zero correlation. Section 3 presents the Monte Carlo study. Section 4 is the conclusion.

2. The t-test of the Pearson Product-Moment Correlation Coefficient

For a wide class of problems, a matter of primary interest is whether or not two random variables X and Y are correlated. The most common measure of correlation between the random variables X and Y is Pearson's product-moment correlation coefficient,

$$\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]E[(Y-\mu_Y)^2]}},$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

Given a random sample $\{(X_i, Y_i),\ i = 1,...,n\}$ from a bivariate random variable (X,Y), ρ is customarily estimated by the sample correlation coefficient,

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

where $\bar{X} = n^{-1}\sum_{i=1}^{n}X_i$ and $\bar{Y} = n^{-1}\sum_{i=1}^{n}Y_i$. The sample correlation coefficient is the maximum likelihood estimate of the population correlation coefficient for bivariate normal data, and is asymptotically unbiased and efficient. Because of the variability of the correlation estimations, it is usually desirable to verify that a nonzero value of the sample correlation coefficient indeed reflects the existence of a statistically significant correlation between the variables of interest. This may be accomplished by testing the null hypothesis $H_0 : \rho = 0$, where a significant correlation is indicated if the hypothesis is rejected. A test statistic used for testing $\rho = 0$ is

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}.$$

It is well known that if (X,Y) is a bivariate Gaussian random variable and $\rho = 0$, then the test statistic t follows a t-distribution with $(n-2)$ degrees of freedom. For a proof of this result see Ravishanker and Dey (2002, p. 200).

Unfortunately, the assumption of bivariate normality is hardly verified in practice: real data are rarely normal (Micceri 1989). Nevertheless, the correlation coefficient is used by practitioners even though their data are nonnormal.

Many studies have investigated the robustness of the t-test of zero correlation to violations of the normality assumption. The results are controversial. Some studies (e.g., Pearson 1929, 1931, Rider 1932, Nair 1941, and Gayen 1951) claim the robustness of the distribution of r to nonnormal populations, while others claim the opposite (e.g., Baker 1930, Chesire et al. 1932, Kowalski 1972, Duncan and Layard 1973, and Zimmerman et al. 2003).

3. A Monte Carlo Study

In this section, we investigate the robustness of the t-test for the null hypothesis that the population correlation is zero, when variables are uncorrelated but not independent through Monte Carlo experiments. All the numerical calculations have been performed using the GNU

Regression, Econometric and Time-series Library (GRETL), a free, open-source software.

In order to obtain a pair of uncorrelated dependent random variables, first we consider a standard normal random variable Z and a uniform random variable, U, over the symmetric interval $[-\pi,\pi]$. Then the bivariate random vectors (X,Y) are constructed in the following manner:

DGP 1. Let $X = Z$ and $Y = Z^2$.
DGP 2. Let $X = Z$ and $Y = |Z|$.
DGP 3. Let $X = Z$ and $Y = \cos Z$.
DGP 4. Let $X = \sin U$ and $Y = U^2$.
DGP 5. Let $X = \sin U$ and $Y = |U|$.
DGP 6. Let $X = \sin U$ and $Y = \cos U$.

We note that in every case the components X and Y are not independent with zero correlation. The scatter plots of the different patterns of dependence are reported in Figure 1. For each DGP the sample correlation coefficient and the test statistic are computed. The t-test was repeated 100,000 times. The actual significance level of the test is estimated by the proportion of rejections with nominal significance level $\alpha = 0.05$ and $\alpha = 0.01$ and for sample sizes of $n = 30, 50, 70, 100$ and 200.

The results presented for $\alpha = 0.05$ in Table 1 and for $\alpha = 0.01$ in Table 2 show that there is a considerable size distortion (the difference between the nominal level of the test and its actual rejection probability). This seems not to depend on the sample size. Further, as it can be seen from the Tables 1 and 2, there are DGPs for which Type I error rates for Pearson's r are inflated and DGPs with deflated Type I error rates. In particular, focusing on the results for DGP1, DGP2 and DGP3, we observe that the test exhibits a severe oversizing with the error of rejection probability measure reaching values around 40% (DGP1, $\alpha = 0.05$). On the contrary, when the DGPs 4, 5 and 6 are considered, the t-test is very conservative. It is important to give some description about why this happens. Our speculation is that the sample data patterns from DPG1-3 tend to be more linear-like than the other three due to the type of nonlinear relationship. For instance, DPG1-3 have one Y values to every X whereas DGP4-6 have two.

We observe that in the DGPs 1-6 there is an exact nonlinear relationship between the variables. Of course we may wonder whether this is too trivial. We consider, for instance, the DGPs 1 and 6. In order to get more "realistic" patterns of dependence (see Figure 2), we replace them by considering the following modified DGPs:

DGP 7. Let $X = Z, Y = X^2 + 0.2V$, where V is a standard normal random variable, independent of Z.

DGP 8. Let $X = \cos U, Y = \sin U + 0.2W$, where W is a standard normal random variable, independent of U.

Basically, the results of the simulation experiment for DGPs 7 and 8, summarized in Table 3, are similar to those generated by DGPs 1 and 6.

Summarizing, in accordance with the literature (for example, Duncan and Layard 1973, and Edgell and Noon 1984), we find that there are many cases where the dependence between the variables inflates Type I error rates for tests of the Pearson correlation coefficient. However, interestingly, and in contrast to previous studies, we also find situations of nonindependent

variables where Type I error rates for Pearson's r are substantially deflated.

Thus, for the nonindependent variables case, the researcher must be concerned not only with the possibility of inflated Type I error rates but also with deflated Type I error rates. In conclusion, caution should be used in interpreting the results of the standard t -test when there is a nonlinear relationship between variables.

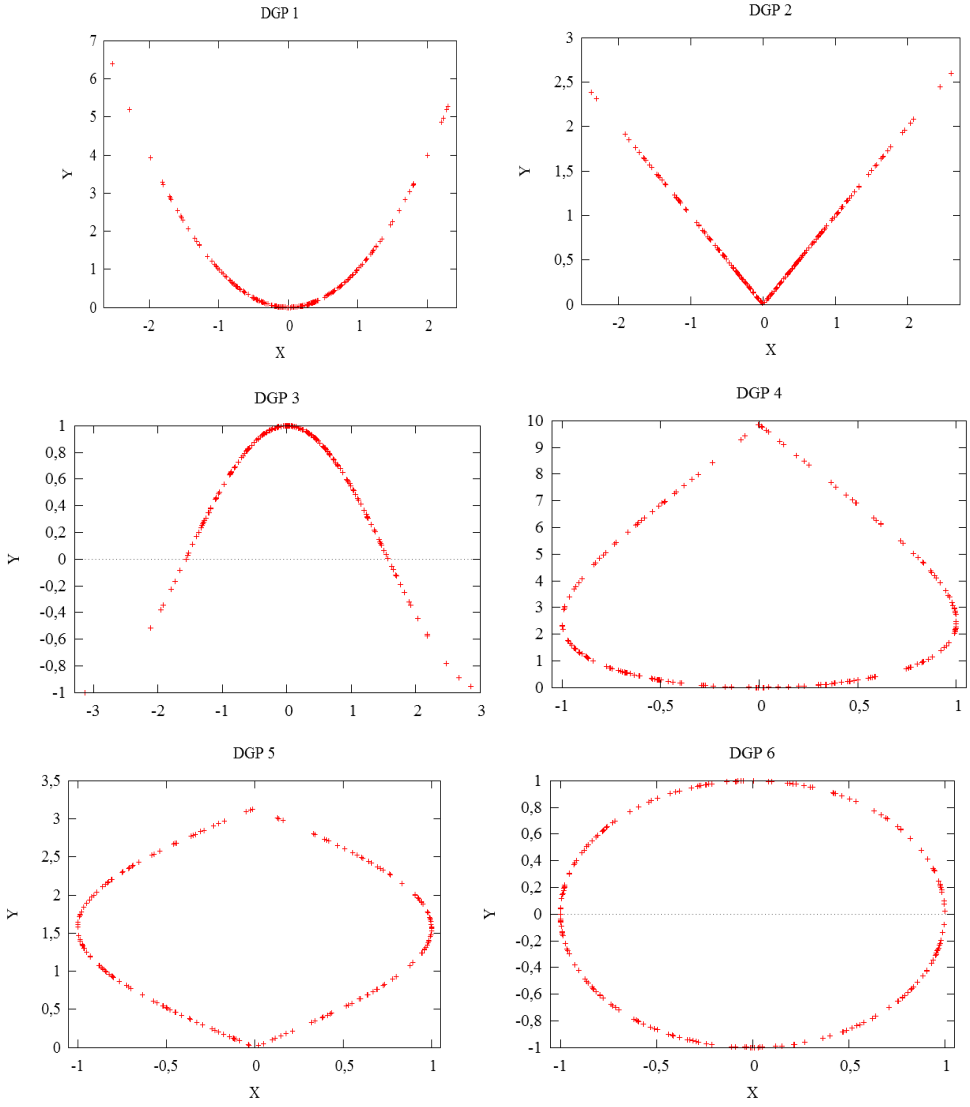

Figure 1 Scatter plots of different patterns of dependence between uncorrelated variables, DGP 1 - DGP 6

Table 1 Proportions of *t*-values which are significant at the 0.05 level

DGP	*n*				
	30	50	70	100	200
DGP 1	0.3797	0.3806	0.3823	0.3818	0.3812
DGP 2	0.2619	0.2567	0.2592	0.2589	0.2576
DGP 3	0.3017	0.2916	0.2908	0.2867	0.2833
DGP 4	0.0039	0.0029	0.0029	0.0028	0.0026
DGP 5	0.0033	0.0026	0.0020	0.0021	0.0020
DGP 6	0.0075	0.0063	0.0064	0.0063	0.0058

Table 2 Proportions of *t*-values which are significant at the 0.01 level

DGP	*n*				
	30	50	70	100	200
DGP 1	0.2347	0.2399	0.2453	0.2457	0.2483
DGP 2	0.1381	0.1375	0.1364	0.1382	0.1389
DGP 3	0.1745	0.1664	0.1644	0.1620	0.1609
DGP 4	0.0002	0.0001	0.0001	0.0001	0.0001
DGP 5	0.0001	0.0001	0.0001	0.0001	0.0001
DGP 6	0.0005	0.0003	0.0003	0.0004	0.0002

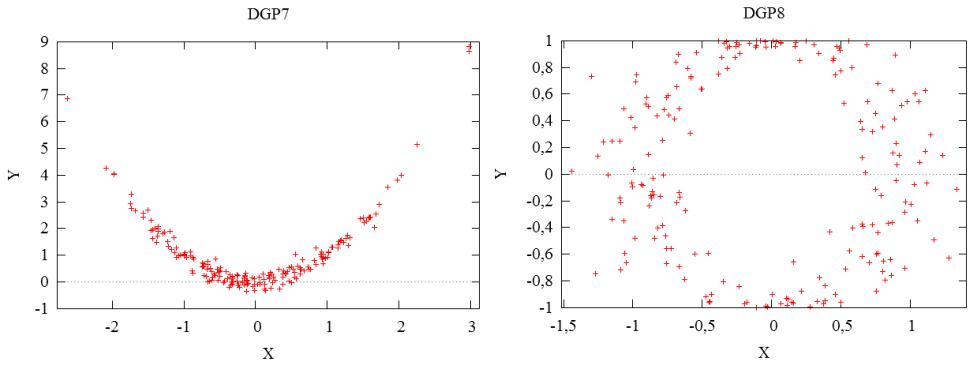

Figure 2 Scatter plots of different patterns of dependence between uncorrelated variables, DGP 7 and DGP 8

Table 3 Proportions of *t*-values which are significant at the α =0.05, 0.01 level

DGP	α	*n*				
		30	50	70	100	200
DGP 7	0.05	0.3722	0.3837	0.3801	0.3758	0.3769
DGP 7	0.01	0.2349	0.2432	0.2368	0.2414	0.2408
DGP 8	0.05	0.0078	0.0086	0.0079	0.0073	0.0077
DGP 6	0.01	0.0004	0.0006	0.0004	0.0002	0.0003

4. Conclusions and Recommendation

In this paper we investigated the robustness of the traditional t-test for testing the hypothesis that a population correlation equals zero. By a Monte Carlo study, we showed that when X and Y are uncorrelated but nonindependent, the null distribution theory is non-robust. The presence of a nonlinear relationship can lead to a significant value of r even if X and Y are uncorrelated. It happens even for large sample sizes. Thus, as suggested by Edgell and Noon (1984), the t-test could be considered a crude test of the hypothesis of independence. However, based on our results, it is important to note that this practice can lead to misleading inference. In fact, despite previous studies, we found situations of nonindependent variables (DGPs 4 - 6) where Type I error rates for Pearson's r are substantially deflated.

Another critical issue is that, when X and Y are uncorrelated but not independent, the t-test can be very conservative. Since a conservative test is less likely to find statistical significance (even when it does truly exist), it follows that the t-test can suffer from loss of power.

In conclusion, the message of this paper is that the practitioner needs to worry about the presence of a nonlinear relationship between the variables when using the t-test. If there is a nonlinear relationship, the correlation value r may be deceptive. Nonlinear associations can exist that are not revealed by this statistic. Thus, the use of the t-test should be avoided where there are good reasons to believe that a nonlinear relationship between variables exists. In this case, the Pearson correlation coefficient should be used only to establish the extent to which the existing relationship can be approximated by a linear relationship.

Acknowledgements

The author is grateful to anonymous referees of this journal for the review of the paper.

References

Baker GA. The significance of the product-moment coefficient of correlation with special reference to the character of the marginal distributions. J. Am. Stat. Assoc. 1930; 25: 387-396.

Chesire L, Oldis E, Pearson ES. Further experiments on the sampling distribution of the correlation coefficient. J. Am. Stat. Assoc. 1932; 27: 121-128.

Duncan GT, Layard MWJ. A Monte-Carlo study of asymptotically robust tests for correlation coefficients. Biometrika. 1973; 60: 551-558.

Edgell S, Noon S. Effect of violation of normality on the t test of the correlation coefficient. Psychol. Bull. 1984; 95: 576-583.

Gayen AK. The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes. Biometrika. 1951; 38: 219-247.

Kowalski CJ. On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. J. Roy. Stat. Soc. C Appl. Stat. 1972; 21:1-12.

Micceri T. The unicorn, the normal curve, and other improbable creatures. Psychol. Bull. 1989; 105: 156-166.

Nair ANK. Distribution of Student's t and the correlation coefficient in samples from non-normal populations. Sankhya. 1941; 5: 383-400.

Pearson ES. Some notes on sampling tests with two variables. Biometrika. 1929; 21: 337-360.

Pearson ES. The test of significance for the correlation coefficient. J. Am. Stat. Assoc. 1931; 26: 128-34.

Ravishanker N, Dey DKA. First course in linear model theory. Boca Raton: Chapman Hall/CRC; 2002.

Rider PR. The distribution of the correlation coefficient in small samples. Biometrika. 1932; 24: 382-403.

Zimmerman DW, Zumbo BD, Williams RH. Bias in estimation and hypothesis testing of correlation. Psicologica. 2003; 24: 133-158.