Nonparametric Test for Independence using Chatterjee's and Spearman's Correlation Coefficient

Debarshi Chakraborty
Email : debchak@iastate.edu
Course : STAT 546
Instructor : Dr. Kris De Brabanter

December 1, 2023

Abstract

This is a brief review of the paper *On relationships between Chatterjee's and Spearman's correlation coefficient* by Qingyang Zhang , Department of Mathematical Sciences, University of Arkansas. This paper deals with differences between the aforementioned coefficients of correlation and tries to develop a new test for independence between random variables using both of them. We will visit only the key ideas and concepts from the paper, all the proofs of the lemmas and theorems are available in the paper itself, the link to which will be attached in the references section. We also justify the theoretical findings with some supporting simulation results at the end. I found this paper quite simple yet interesting and fundamental to statistical literature, hopefully there will be much more proceedings regarding this in future.

Contents

1 Introduction

1.1 Some popular measures and limitations

Measuring the association or dependence between two random variables is a very common problem in Statistics, testing of hypothesis for the same is also an interesting and useful problem. Pearson's product moment correlation coefficient is one of the most popular such measure for continuous random variables, but it is designed only to capture linear dependence. Spearman's rank based correlation coefficient is a nonparametric alternative to he former which can capture monotonic dependence very well and is also robust to outliers due to the rank-based property. The vast applications of the aforementioned measures is due to the fact that under the null hypothesis of independence, they are asymptotically normal which facilitates easy calculation of p-values. However, none of them perform reasonably well in detecting non monotonic relationships between two random variables.

1.2 Further literature

In the past decades, several tests were developed which are consistent against all alternatives, which includes kernel based test, distance correlation test, tests based on copulas, tests based on graphs, etc. In practice, one of the most popular tests is the distance correlation test (Richards [Ric17]). But, since this test statistic lacks simple asymptotic theory, hence computing p-values becomes expensive since permutation tests are required. For example, under the null hypothesis of independence, deriving the distribution of this test statistic depends of the underlying distribution of the random variables and the standard approach is to approximate the distribution of distance correlation by permutation. This method is computationally very expensive. If R denotes the number of permutations and n denotes the sample size, it has a time complexity $O(Rn^2)$.

1.3 Recent work : Chatterjee's Correlation Coefficient

From the previous discussion, it is clear that a desirable measure of dependence is one which will be consistent against all alternatives and will also have nice asymptotic properties. Sourav Chatterjee developed one such measure in his seminal paper (Chatterjee [Cha20]) in the year 2021, which is rank based and satisfies the two properties mentioned above i.e. it is consistent against all alternatives and is asymptotically normal. We give a snapshot of it here.

Let X and Y be two continuous random variables and $(X_i, Y_i)_{i=1,2,...,n}$ be n i.i.d samples from the joint distribution of (X, Y). Assuming no ties, the data can be uniquely arranged as $(X_{(1)}, Y_{(1)}), ..., (X_{(n)}, Y_{(n)})$ such that $X_{(1)} < ... < X_{(n)}$. Let $R_i = \sum_{k=1}^{n} \mathbb{I}(Y_{(k)} < Y_{(i)})$. Chatterjee's correlation is defined as

$$\xi_n(X, Y) = 1 - \frac{3\sum_{i=1}^{n-1} |R_{i+1} - R_i|}{n^2 - 1}$$

Chatterjee showed that $\xi_n(X, Y) \to \xi(X, Y)$ in probability as $n \to \infty$ where $\xi(X, Y) = \frac{\int V(E(\mathbb{I}(Y \geqslant t|X)))dF_Y(t)}{\int V(\mathbb{I}(Y \geqslant t))dF_Y(t)}$. Here $0 \leqslant \xi(X, Y) \leqslant 1$, attaining the value 0 when X and Y are independent and attaining the value 1 if and only if Y is a measurable function of X. Note that, despite being named as "correlation", this coefficient measures any kind of dependence existing between two random variables. Also, the measure is not symmetric

in X, Y, which is not a desirable property, since statistical dependence is symmetric in nature. But this is not a big issue, it can be resolved easily, we will see that later.

Also, under independence, it can be proved that

$$\sqrt{n}\xi_n(X, Y) \to N(0, 2/5) \text{ in law as } n \to \infty.$$

This much theoretical knowledge about this measure is enough for our further discussion. Nevertheless, a lot of work has been done regarding this new measure such as it's CLT under dependence, power of the test, etc.

Moreover, it is empirically found that this test is much more powerful in detecting non-monotonic associations , especially sinusoidal and oscillating ones, compared to competiting measures of dependence. As we know, there is no free lunch out there, a natural question is does it have any drawback then? Yes. The only disadvantage is Chatterjee's test is less powerful in detecting linear or monotonic relationships compared to other popular tests.

1.4 Goal of this paper

We saw that Spearman's correlation coefficient is good at capturing monotonic relationships while Chatterjee's correlation does a pretty good job for non monotonic association between random variables i.e. they are in some sense complementary to each other. Motivated by these intuitions, this paper (Zhang [Zha23]) proceeds to develop a test for independence which incorporates both Spearman's and Chatterjee's correlation coefficients. Section 2 focuses on some key results that are used to develop the testing theory, section 3 demonstrates how the two coefficients can differ from one another, in other words how they capture different kinds of dependence structures. Section 4 illustrates the testing method concisely.

2 Some Asymptotic Results

2.1 Revisiting Spearman's Correlation Coefficient

If we follow the same notations of the previous section, Spearman's rank based correlation coefficient is defined as

$$S_n(X, Y) = 1 - \frac{6\sum_{i=1}^{n}(i - R_i)^2}{n(n^2-1)}$$

It is well known that $\sqrt{n}S_n(X, Y) \to N(0, 1)$ in distribution.

2.2 Joint distribution of S_n and ξ_n

Now we see the main results of the paper.

Theorem 1. *If X and Y are independent then we have*

- $Cov(S_n(X, Y), \xi_n(X, Y)) = 0 \forall n \geqslant 2$.
- $\sqrt{n}S_n(X, Y)$ *and* $\sqrt{n}\xi_n(X, Y)$ *are asymptotically joint normal.*

From the above theorem, our main theorem follows immediately.

Theorem 2. *If X and Y are independent then we have $(\sqrt{n}S_n(X,Y), \sqrt{n}\xi_n(X,Y))^T \to N(\mu, \Sigma)$ in distribution where $\mu = (0,0)^T, \Sigma = diag(1, 2/5)$.*

I did not include the proofs here since they are already done clearly in the paper. We will need this results later to justify the theoretical properties of the new test for independence.

3 When the two coefficients differ?

Here we demonstrate how can the two measures behave differently under different kinds of dependence structures. Since $0 \leqslant \xi_n \leqslant 1$ and $-1 \leqslant S_n \leqslant 1$, so we will compare the values of ξ_n with $|S_n|$.

- **Case 1** : For any $\epsilon > 0$, there exists ranks $\{R_1, ..., R_n\}$ such that $|S_n(X,Y)| < \epsilon$ and $\xi_n(X,Y) > 1 - \epsilon$. For n odd, consider the following ranks

$$R_i = (n - 2(i-1))\mathbb{I}(1 \leqslant i \leqslant (n+1)/2) + (2i - (n+1))\mathbb{I}((n+3)/2 \leqslant i \leqslant n)$$

  Some basic algebra shows that $\xi_n(X,Y) = 1 - \frac{6n-9}{n^2-1}$ and $|S_n(X,Y)| = \frac{3}{2n}$. For any fixed $\epsilon > 0$, we can find an odd number n such that $|S_n(X,Y)| < \epsilon$ and $\xi_n(X,Y) > 1 - \epsilon$.

- **Case 2** : For this case, a constructive proof is not straightforward, instead the author provides one particular example in the paper where ξ_n is relatively small but S_n is substantially large. For any $\epsilon > 0$, there exists rank s$\{R_1, ..., R_n\}$ such that

$$\xi_n(X,Y) = \epsilon + O(1/n) \text{ and } S_n(X,Y) = 1 - \sqrt{2/27}(1-\epsilon)^{3/2} + O(1/n).$$

The main objective of the above discussion was to point out that situations may arise where one of the coefficients take large values while the other takes smaller values, and vice versa. This supports our intuitions and empirical findings that Chatterjee's and Spearman's correlation are efficient in detecting different kinds of relationships between random variables, which motivates us to combine the two in order to develop a reasonably powerful test no matter what type of dependence exists is between X and Y.

4 A Novel Test for Independence

4.1 The Test Statistic

Motivated by the findings of previous sections, the author proposes the new test statistics as follows

$$I_n(X,Y) = max\{|S_n(X,Y)|, \sqrt{5/2}\,\xi_n(X,Y)\}$$

Clearly, the new test statistic I_n takes advantages of both S_n and ξ_n and therefore can be used as a versatile tool to test for both monotonic and non monotonic relationships. Also, by the virtue of **Theorem 2**, p-values for the tests can be calculated easily (in asymptotic sense of course).

4.2 Calculating p - values

We want to test

$$H_0 : X \text{ and } Y \text{ are independent vs } H_1 : \text{not } H_0$$

Under H_0, as $n \to \infty$, we have

$P(\sqrt{n}I_n(X,Y) > z) = P(max\{\sqrt{n}|S_n(X,Y)|, \sqrt{5n/2}\xi_n(X,Y)\} > z)$

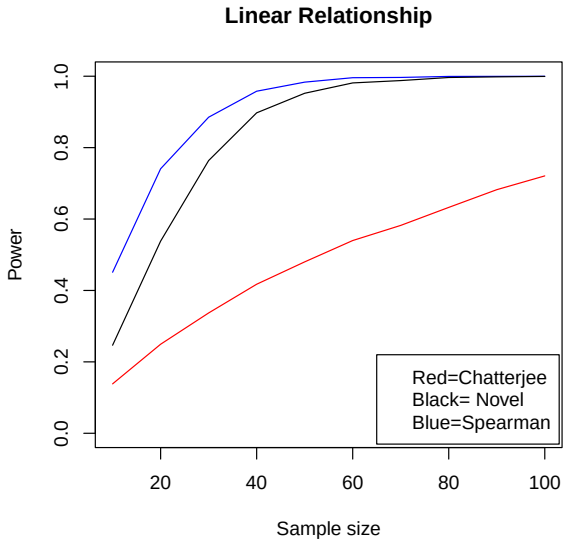
$= 1 - P(max\{\sqrt{n}|S_n(X,Y)|, \sqrt{5n/2}\xi_n(X,Y)\} < z) = 1 - P(\sqrt{n}|S_n(X,Y)| < z, \sqrt{5n/2}\xi_n(X,Y) < z)$

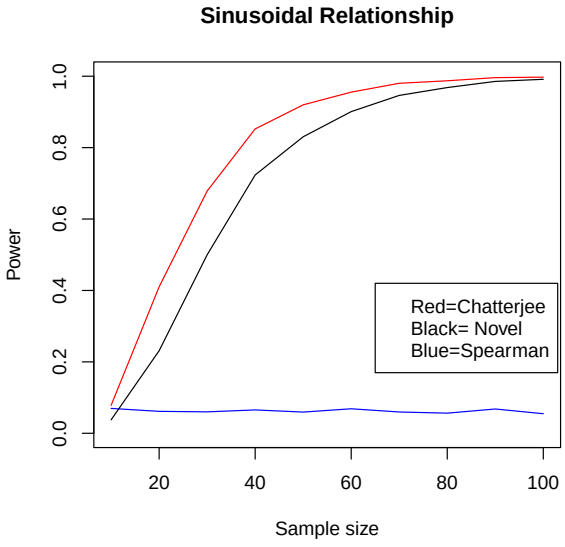
$= 1 - P(\sqrt{n}|S_n(X,Y)| < z)P(\sqrt{5n/2}\xi_n(X,Y) < z)$ (due to asymptotic independence)

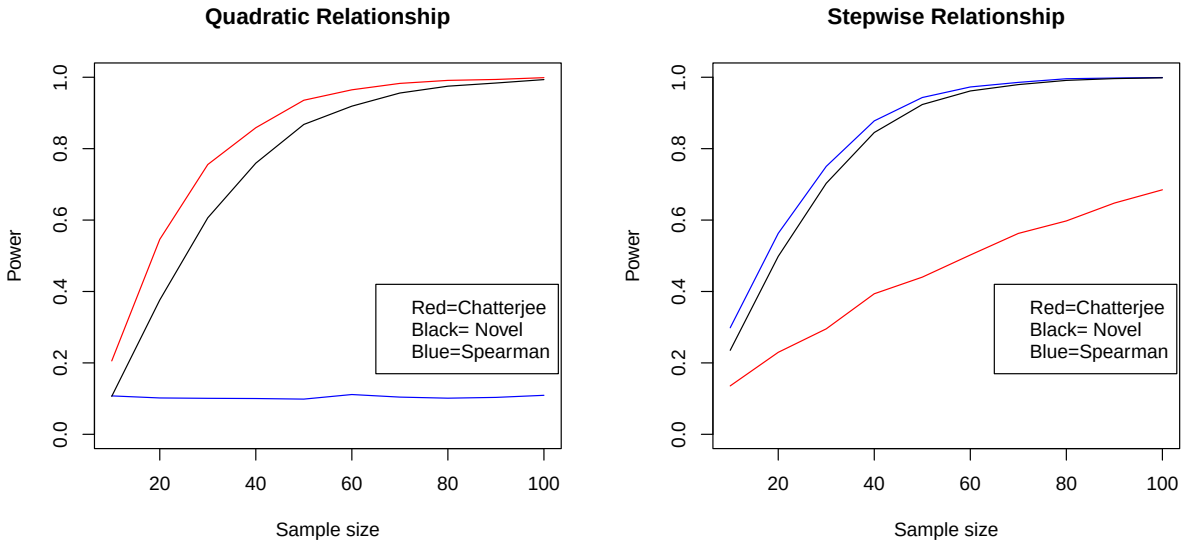
$= 1 - (2\Phi(z) - 1)\Phi(z)$ (due to asymptotic normality)

4.3 Simulation Studies

We consider four types of bivariate data on (X, Y) with different kinds of relationships and try to compare the empirical powers of S_n, ξ_n, I_n in each case.







Observations:

- When X and Y have a monotonous association between each other i.e. for the linear and stepwise case, Spearman clearly outperforms Chatterjee's correlation as expected.
- For non monotone relation between X and Y, exactly the opposite happens.
- Most importantly, the new test does a pretty good job in any case, irrespective of the type of association, that's wonderful.

Note : These plots are not taken from the paper, they were generated by me using the same relationship between X and Y mentioned in the paper.

5 Conclusion

Chatterjee's correlation attracted a lot of attention in the last two years and a lot of research has been done to boost it's power and many more things. Hence, this measure can probably be improved further. One basic limitation is as we mentioned before, Chatterjee's measure ξ_n is not symmetric in X and Y, as a consequence of which our new test statistic I_n becomes asymmetric too. This issue can be studied in detail, although some work has been done already.

References

[Cha20] Sourav Chatterjee. *A new coefficient of correlation*. 2020. arXiv: 1909.10140 [math.ST].

[Ric17] Donald St. P. Richards. *Distance Correlation: A New Tool for Detecting Association and Measuring Correlation Between Data Sets*. 2017. arXiv: 1709.06400 [stat.OT].

[Zha23] Qingyang Zhang. *On relationships between Chatterjee's and Spearman's correlation coefficients*. 2023. arXiv: 2302.10131 [stat.ME].

****