Journal of Modern Applied Statistical Methods

Volume 14 | Issue 2 Article 4

11-1-2015

Inferences About the Skipped Correlation Coefficient: Dealing with Heteroscedasticity and Non-Normality

Rand Wilcox
University of Southern California, rwilcox@usc.edu

DOI: 10.22237/jmasm/1446350580

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 2-8.

ISSN 1538 − 9472

Invited Article
Inferences About the Skipped Correlation Coefficient: Dealing with Heteroscedasticity and Non-Normality

Rand Wilcox
University of Southern California
Los Angeles, CA

A common goal is testing the hypothesis that Pearson's correlation is zero and typically this is done based on Student's T test. There are, however, several well- known concerns. First, Student's T is sensitive to heteroscedasticity. That is, when it rejects, it is reasonable to conclude that there is dependence, but in terms of making a decision about the strength of the association, it is unsatisfactory. Second, Pearson's correlation is not robust: it can poorly reflect the strength of the association. Even a single outlier can have a tremendous impact on the usual estimate of Pearson's correlation, which can result in a poor indication of the strength of the association among the bulk of the points. Numerous robust correlation coefficients have been proposed that deal with outliers among the marginal distributions, but these methods do not take into account the overall structure of the data in terms of dealing with outliers. A skipped correlation addresses this concern and methods for testing the hypothesis that this correlation is zero have been studied. However, there are serious limitations associated with one of these methods and extant studies regarding an alternative percentile bootstrap method do not address practical concerns reviewed in the paper. A minor goal is to report situations where this percentile bootstrap method can be unsatisfactory. The main result is that an alternative percentile bootstrap method performs well in simulations.

Keywords: Robust measures of association, level robust methods, non-normality, heteroscedasticity

Introduction

A basic goal is testing the hypothesis that the strength of the association between two random variables is zero. Certainly the best-known strategy is to test the hypothesis that Pearson's correlation is zero, using Student's T test.

Dr. Wilcox is Professor of Psychology at the University of Southern California. Email him at rwilcox@usc.edu.

$$H_0 : \rho = 0 \tag{1}$$

There are, however, well known concerns with this approach. First, Student's T assumes homoscedasticity. In practical terms, it provides a reasonable test of the hypothesis that two variables are independent, but in terms of making inferences about ρ, it can be unsatisfactory. For example, even when the null hypothesis is true, the probability of rejecting can increase as the sample size increases when there is heteroscedasticity (e.g., Wilcox, 2012). Roughly, the reason is that Student's T uses the wrong standard error when there is heteroscedasticity, given the goal of testing (1).

Another concern is that r, the usual estimate of ρ, is not robust. Even a single outlier can result in a poor reflection of the strength of the association among the bulk of the points. Numerous robust estimators have been proposed for dealing with outliers among the marginal distributions (e.g., Wilcox, 2012, chapter 9). Certainly the two best-known approaches are Kendall's tau and Spearman's rho. But a known concern with these measures of association is that they do not deal with outliers in a manner that takes into account the overall structure of the data. That is, based on the random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, situations are encountered where no outliers are detected among $X_1, \ldots, X_n$, ignoring Y, and no outliers are detected among $Y_1, \ldots, Y_n$, ignoring X, yet there are outliers that can have a substantial impact on Kendall's tau, Spearman's rho and other measures of association that do not deal with the overall structure of the data (e.g., Wilcox, 2012, chapter 9). A measure of the strength of an association that deals with this issue is the skipped correlation coefficient. The basic strategy is to use some outlier detection method that takes into account the overall structure of the data, remove any outliers that are found, and then compute Pearson's correlation using the remaining data.

There are many outlier detection methods that take into account the overall structure of the data. In the context of a skipped correlation, a projection type outlier detection method has been the focus of attention. No single outlier detection method dominates, but the projection-type method used here appears to perform relatively well in terms of avoiding masking and detecting truly unusual points (e.g., Wilcox, 2012). Masking refers to missing outliers due to their very presence. For example, in the univariate case, detecting outliers using the mean and standard deviation can result in masking. The basic problem is that outliers inflate the sample standard deviation, which in turn can result is missing even extreme outliers.

Based on the projection type method for detecting outliers, let ξ denote the population analog of the skipped correlation and consider the goal of testing

$$H_0 : \xi = 0 \tag{2}$$

A very simple approach is described in Wilcox (2012, Section 9.4.4). However, the method is limited to testing at the $\alpha = 0.05$ level and it assumes homoscedasticity. More recently, Pernet, Wilcox and Rousselet (2013) studied a bootstrap method when sampling from a bivariate normal distribution. But the impact of non-normality and heteroscedasticity was not addressed. A minor goal in this paper is to report results indicating situations where the Pernet et al. method can be unsatisfactory when dealing with non-normality and heteroscedasticity. The primary goal is to report simulation results on an alternative bootstrap method that provides good control over the Type I error probability for a broader range of situations.

Description of the methods to be compared

This section describes the projection outlier detection method followed by the two percentile bootstrap methods that were studied when testing (2). For brevity, just an outline of the method is provided. Complete computational details can be found in Wilcox (2012, section 6.4.9). Included is an R function called outpro for applying it, which is used here.

The projection method begins by estimating the center of the data cloud, say $\hat{\theta}$. Here this is done using the marginal medians. Then for fixed i, project all n points onto the line connecting $\hat{\theta}$ and (X_i, Y_i). Based on the projected points, let D_j ($j = 1, \ldots, n$) be the distance between the projection of (X_j, Y_j) and the center, $\hat{\theta}$. Next, check for outliers using the usual boxplot rule based on the D_j values. That is, if q_1 and q_2 are estimates of the lower and upper quartiles, respectively, based on $D_1, \ldots, D_n$, declare D_j an outlier if $D_j < 1.5(q_2 - q_1)$ or if $D_j > 1.5(q_2 - q_1)$, in which case (X_j, Y_j) is declared an outlier as well. This process is performed for each i ($i = 1, \ldots, n$) and (X_j, Y_j) is declared an outlier if its projected distance is flagged as an outlier for any i.

The percentile bootstrap method used by Pernet et al. (2013) is applied as follows:

1. Remove any points flagged as outliers using the projection method. Let m denote the sample size after outliers are removed.
2. Generate a bootstrap sample from the remaining data by resampling with replacement m points.
3. Compute Pearson's correlation based on this bootstrap sample yielding r^*.
4. Repeat steps 2-3 and B times yielding $r_1^*, \ldots, r_B^*$.
5. Put the values $r_1^*, \ldots, r_B^*$ in ascending order and label the results $r_{(1)}^* \leq \cdots \leq r_{(B)}^*$.
6. Let $l = \alpha B/2$, rounded to the nearest integer and $u = B - l$. Then the $1 - \alpha$ confidence interval for ξ is taken to be $(r_{(l+1)}, r_{(u)})$. This will be called method B1 henceforth.

An unusual feature of method B1 is that the process of generating bootstrap samples does not exactly mimic the manner in which the data are generated and the skipped correlation is computed. A percentile bootstrap method that does mimic the way data are generated, labeled method B2 here, begins by generating a bootstrap sample from all n points, removing any points flagged as outliers and then computing $\hat{\xi}^*$, Pearson's correlation based on the remaining data. That is, in the description of method B1, replace steps 1-3 with

1. Generate a bootstrap sample by resampling with replacement n points from the entire sample of size n.
2. Remove any points from the bootstrap sample in step 1 that are flagged as outliers using the projection method.
3. Compute Pearson's correlation using the points not flagged as outliers in step 2.

As done in step 4 of method B1, this process is repeated B times only now the results are labeled $\hat{\xi}_1^*, \cdots, \hat{\xi}_B^*$. The $1 - \alpha$ confidence interval for ξ is taken to be $\left(\hat{\xi}_{(l+1)}^*, \hat{\xi}_u^*\right)$.

It is noted that a p-value is readily computed when testing (2), which is motivated by general results in Liu and Singh (1997). Let Q^* be the proportion of $\hat{\xi}^*$ values that are less than zero. Then a p-value is $p = \min(2Q^*, (1 - 2Q^*)$.

Simulation results

Four types of distributions are considered: normal, symmetric and heavy-tailed (roughly meaning that outliers tend to be common), asymmetric and relatively light-tailed, and asymmetric and relatively heavy-tailed. More specifically, g-and-h distributions (Hoaglin, 1985) are used, which arise as follows. Let Z be a random variable having a standard normal distribution and let

$$W = \frac{\exp(gZ) - 1}{g} \exp(hZ^2/2)$$

If $g = 0$

$$W = Z \exp\left(h \frac{Z^2}{2}\right)$$

Then W has a g-and-h distribution, where g and h are parameters that determine the first four moments. The four distributions used here are the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = .2$, $g = 0$), an asymmetric distribution with relatively light tails ($h = 0$, $g = .2$), and an asymmetric distribution with heavy tails ($g = h = .2$). Table 1 summarizes the skewness (γ_1) and kurtosis (γ_2) of these distributions.

The number of bootstrap samples was taken to be $B = 1000$. Bradley (1978) suggests that as a general guide, when testing at the .05 level, the actual level should be between .025 and .075. Preliminary simulations based on $B = 500$ indicated that method B2 does not satisfy this criterion; increasing B to 1000 gave more satisfactory results.

Table 1. Some properties of the g-and-h distribution.

g	h	κ_2	κ_1
0.0	0.0	0.00	3.00
0.0	0.2	0.00	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

Observations were generated according to the model $Y = \lambda(X)\varepsilon$, where both X and ε have one of the g-and-h distributions in Table 1 and $\lambda(X)$ is used to model

heteroscedasticity. Three choices for $\lambda(X)$ were used: $\lambda(X) \equiv 1$ (homosecdasticity), $\lambda(X) = |X| + 1$ (so the conditional variance of Y, given X, is smallest when X is close to its mean), and $\lambda(X) = 1/(|X| + 1)$ (in which case the conditional variance of Y, given X, is largest when X is close to its mean. For convenience these three choices for λ will be called variance patterns (VP) 1, 2 and 3, respectively.

The simulation estimates of the actual Type I error probabilities were based on 2,000 replications. A common suggestion is that ideally, simulation estimates be based on 10,000 replications. However, when using method B2, a single replication takes a little over 14 seconds using the software R on a MacBook Pro with a 2.5 GHz processor. So 10,000 replications would require over 38 hours of execution time. To add perspective on the precision of the estimates, assuming Bradley's criterion is reasonable, consider the issue of whether the actual level is less than or equal .075. Using the method in Pratt (1968), it can be seen that based on a two-sided .95 confidence interval for the actual level, the confidence interval will not contain .075 if $\hat{\alpha} \leq .063$. In a similar manner, based on a two-sided .95 confidence interval, the confidence interval for the actual level does not contain .025 if $\hat{\alpha} \geq .0325$.

Table 2. Estimated Type I error probabilities, $n = 40$, $\alpha = .05$

g	*h*	VP	B2	B1
0.0	0.0	1	0.022	0.066
		2	0.022	0.071
		3	0.028	0.055
0.0	0.2	1	0.022	0.070
		2	0.024	0.080
		3	0.024	0.046
0.2	0.0	1	0.027	0.066
		2	0.024	0.072
		3	0.030	0.056
0.2	0.2	1	0.021	0.072
		2	0.024	0.080
		3	0.022	0.045

Table 2 shows the estimated Type I error probabilities when $n = 40$ and $\alpha = .05$. As can be seen, method B2 tends to be conservative, meaning that the estimated Type I error probability is always less than the nominal .05 level. The estimates are consistently close to .025 over all of the situations considered. So there is some possibility that the actual level drops below .025, but there is no strong indication that this is the case. In contrast, the estimates using method B1

are always greater than or equal to .05 with the two largest estimates equal to .08. So all indication are that in terms of avoiding a Type I error probability greater than the nominal level, B2 performs better than B1.

Concluding remarks

Some positive features of method B1 are that it reduces execution time compared to method B2 and it performs reasonably well in simulations when there is homoscedasticity and sampling is from a bivariate normal distribution. For most situations, it was estimated that the actual level using method B1 is less than .075, but for variance pattern VP 2 this is not the case when dealing with distributions with heavy-tails. In contrast, method B2 avoids Type I error probabilities greater than .05 among all of the situations considered, the only concern being that the actual level was estimated to be as low as .022 with a sample size of $n = 40$. That is, there is some possibility that B2 does not satisfy Bradley's criterion that the actual level should be at least .025. The main result for the goal of avoiding an actual level well above .05, all indications are that B2 is preferable to B1.

References

Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152.

Hoaglin, D. C. (1985) Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Exploring Data Tables, Trends, and Shapes.* (pp. 461-515). New York: Wiley.

Liu, R. G. & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association, 9*(2), 266–277.

Pernet, C. R., Wilcox, R. & Rousselet, G A. (2013). Robust correlation analyses: false positive and power validation using a new open source Matlab toolbox. *Frontiers in Quantitative Psychology and Measurement*. doi:10.3389/fpsyg.2012.00606

Pratt, J. W. (1968). A normal approximation for binomial, F, beta, and other common, related tail probabilities, I. *Journal of the American Statistical Association, 63*, 1457–1483.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. (3rd Ed.). San Diego, CA: Academic Press.