Wilcox *Journal of Statistical Distributions and Applications* (2015) 2:7
DOI 10.1186/s40488-015-0030-z

Journal of Statistical Distributions and Applications
a SpringerOpen Journal

METHODOLOGY Open Access

Comparing the variances of two dependent variables

Rand Wilcox

Correspondence: rwilcox@usc.edu
Department of psychology, University of Southern California, 3620 McClintock Ave, 90089-1061 Los Angeles, USA

Abstract

Various methods have been derived that are designed to test the hypothesis that two dependent variables have a common variance. Extant results indicate that all of these methods perform poorly in simulations. The paper provides a new perspective on why the Morgan-Pitman test does not control the probability of a Type I error when the marginal distributions have heavy tails. This new perspective suggests an alternative method for testing the hypothesis of equal variances and simulations indicate that it continues to perform well in situations where the Morgan-Pitman test performs poorly.

Keywords: Morgan-Pitman test; Heteroscedasticity; HC4 estimator; Well elderly 2 study

1 Introduction

A classic problem that arises in various situations is testing the hypothesis that two dependent variables have equal variances. For example, when measuring systolic and diastolic blood pressure, the quality of two different blood pressure gauges depends in part on whether one type of gauge has more variability than some other type. Rothstein et al. (1981) cite two examples in psychology in which a test of equality of variances is of interest. Other examples from psychology are described in Lord and Novick (1968); Games et al. (1972) and Levy (1976). Snedecor and Cochran (1967) also cite two examples, one of which deals with testing for differences in reliability between two laboratories.

Let σ_1^2 and σ_2^2 be the population variances associated with the random variables X and Y, respectively, where X and Y have some unknown bivariate distribution. Seemingly the best-known technique for testing

$$H_0 : \sigma_1^2 = \sigma_2^2 \tag{1}$$

is a method derived by Morgan (1939) and Pitman (1939). Letting

$$U = X + Y$$

and

$$V = X - Y,$$

if the null hypothesis is true, then ρ_{uv}, Pearson's correlation between U and V, is zero. So testing (1) can be accomplished by testing

$$H_0 : \rho_{uv} = 0. \tag{2}$$

Student's T test is the best-known method for testing (2). That is, use the test statistic

$$T_{uv} = r_{uv}\sqrt{\frac{n-2}{1-r_{uv}^2}},$$

where r_{uv} is the usual estimate of ρ_{uv} and T_{uv} is assumed to have a Student's T distribution with $n-2$ degrees of freedom, where n is the sample size. However, McCulloch (1987) as well as Mudholkar et al. (2003) establish that when sampling from heavy-tailed distributions, the actual Type I error probability exceeds the nominal level, sometimes substantially so. Roughly, they show that T_{uv} does not converge to a standard normal distribution as the sample size increases. More precisely, under normality, the variance of T_{uv} converges to one as the sample size increases. But for heavier-tailed distributions, this is no longer the case. The variance of T_{uv} converges to a value that is greater than one, which in turn results in Type I error probabilities greater than the nominal level when testing (2).

McCulloch (1987) suggests replacing Pearson's correlation with Spearman's correlation. But simulations in Wilcox (1990) indicate that again the actual level can be substantially higher than the nominal level when sampling from a heavy-tailed distribution. Wilcox (1990) reported simulation results on several other methods and found that all of them performed poorly under non-normality. They included methods derived by Tiku and Balakrishnan (1986) as well as a Box-Scheffé type test that has close connections to a method suggested by Levy (1976).

This paper provides a different perspective than the results reported by McCulloch (1987) and Mudholker et al. (2003) regarding why the Morgan-Pitman performs poorly when sampling from a heavy-tailed distribution. Details are given in Section 2. This alternative perspective suggests a general strategy for getting improved control over the Type I error probability. A particular variation of this strategy is described in Section 3. Simulation results based on the method in Section 3 are reported in Section 4.

2 The Morgan-Pitman test and heavy-tailed distributions

Consider the random variables X and Y, let r denote the usual estimate of Pearson's correlation, ρ, and consider the usual test statistic

$$T = r\sqrt{\frac{n-2}{1-r^2}} \tag{3}$$

for testing $H_0 : \rho = 0$. From basic principles, T has a Student's T distribution if either X or Y has a normal distribution and simultaneously, X and Y are independent. Note that independence implies homoscedasticity. That is, the conditional variance of Y, given X, does not depend on the value of X, which plays a fundamental role in the derivation of T (e.g., Hogg and Craig 1970). In the context of least squares regression, it is known that if there is heteroscedasticity, the usual test of the hypothesis of a zero slope uses the wrong standard error (e.g., Long and Ervin 2000). It follows that when testing $H_0 : \rho = 0$ using the test statistic T given by (3), again the wrong standard error is being used. As the sample size increases, the probability of rejecting can increase even when $\rho = 0$ but there is heteroscedasticity (e.g., Wilcox 2012).

Now consider the situation where X and Y have a bivariate normal distribution. For T_{uv} to provide reasonably good control over the Type I error probability when testing (1), it must be the case that there is homoscedasticity in terms of the association between

U and V. To provide a visual sense of the extent there is homoscedasticity, 5000 pairs of observations were generated from a bivariate normal distribution with $\rho = 0$ resulting in 5000 U and V values. Figure 1 shows a plot of the .2, .5 and .8 quantile regression lines using the running interval smoother (e.g. Wilcox 2012) where U is taken to be the independent variable. As can be seen the regression lines for the .2 and .8 quantiles suggest that there is very little if any heteroscedasticity.

Now look at Fig. 2 where again a sample size of $n = 5000$ was used, only now the marginal distributions for X and Y are g-and-h distributions with $g = 0$ and $h = .2$. This is a symmetric distribution with heavy tails. (More details about this distribution are given in Section 4.) Figure 2 shows the .2, .5 and .8 quantile regression lines for U and V. As can be seen, there is a clear indication of heteroscedasticity implying that even with a large sample size, the Morgan-Pitman test will perform poorly in terms of controlling the Type I error probability.

Of course, Fig. 1 does not establish that there is exact homoscedasticity regarding the association between U and V when dealing with normal distributions. The only point is that as we move from normal distributions toward heavy-tailed distributions, heteroscedasticity becomes more pronounced, so it is not surprising that the Morgan-Pitman test performs poorly for such situations.

Figure 2 also indicates why replacing Pearson's correlation with Spearman's correlation is unsatisfactory. Converting observations to ranks does not eliminate heteroscedasticity. Converting the data used in Fig. 2 to ranks, a plot of the .2, .5 and .8 quantile regression lines indicates that heteroscedasticity is less severe, which in turn suggests that using Spearman's correlation improves control over the Type I error probability compared to

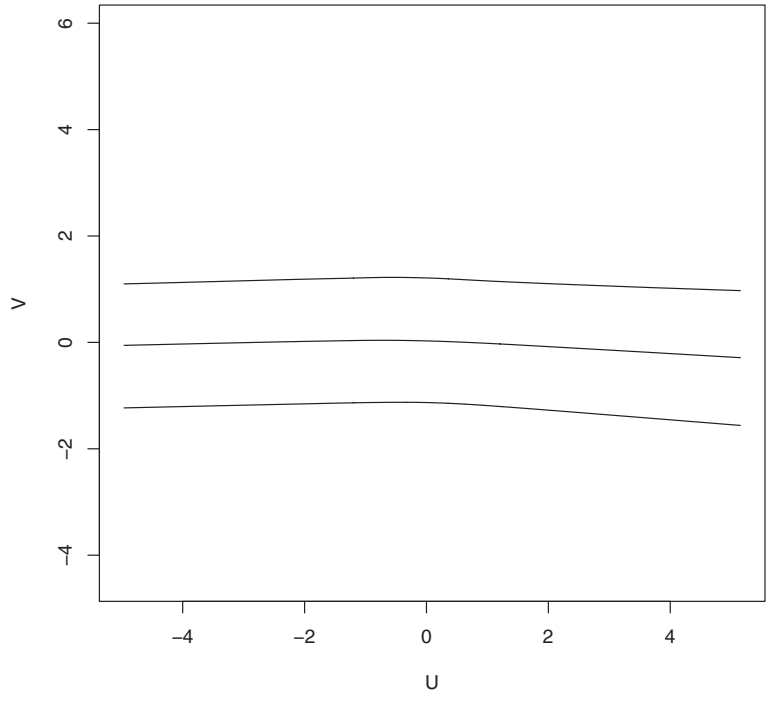

Fig. 1 Shown are the .2, .5 and .8 quantile regression lines for U and V based on a sample size of 5000 from a bivariate normal distribution. There is little or no indication of heteroscedasticity

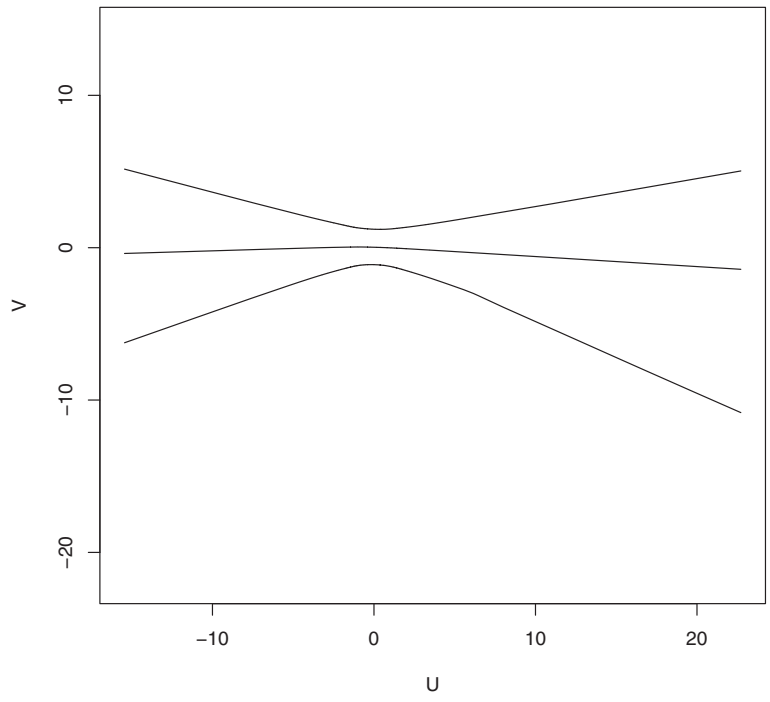

Fig. 2 Shown are the .2, .5 and .8 quantile regression lines for *U* and *V* based on a sample size of 5000 when *X* and *Y* have a symmetric, heavy-tailed distribution. Now there is a clear indication that there is heteroscedasticity

using T_{uv}, but that poor control over the Type I error probability will still be an issue. Simulation results in Section 4 demonstrate the extent to which this is the case.

3 Modification of the Morgan-Pitman test

Let b_0 and b_1 be the usual least squares estimate of β_0 and β_1, the intercept and slope of a regression line. In recent years, several heteroscedastic consistent (HC) methods have been derived for estimating the standard error of b_1 (e.g., Long and Ervin 2000, Godfrey 2006, Cribari-Neto et al. 2007). Cribari-Neto (2004) found that the so-called HC4 estimator performs relatively well. More recently, Cribari-Neto et al. (2004) derived the HC5 estimator and argued that it is better than HC4, particularly at handling outliers in the independent variable *X*. Ng and Wilcox (2009) compared several method for testing

$$H_0 : \beta_1 = 0$$

based on the HC4 and HC5 estimators. Included were several bootstrap methods. No single method dominated, but a non-bootstrap method, based in part on the HC4 estimator, performed relatively well. No advantage using the HC5 estimator was found.

For the random sample $(Y_1, X_1), \ldots, (Y_n, X_n)$, let

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}.$$

Let

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1},$$

$$\mathbf{H} = \text{diag}(\mathbf{X}\mathbf{C}\mathbf{X}')^{-1},$$

$$\bar{h} = \sum h_{ii}/n,$$

$$e_{ii} = h_{ii}/\bar{h}$$

and

$$d_{ii} = \min(4, e_{ii}).$$

Let $\mathbf{A}$ be the $n \times n$ diagonal matrix with the ith entry given by $r_i^2(1-h_{ii})^{-d_{ii}}$, where r_i is the ith residual based on the ordinary least squares estimator. The HC4 estimator is

$$\mathbf{S} = \mathbf{C}\mathbf{X}'\mathbf{A}\mathbf{X}\mathbf{C}.$$

The diagonal elements of $\mathbf{S}$ are the estimated squared standard errors of b_0 and b_1. For convenience, the HC4 estimates of the standard errors of b_0 and b_1 are denoted by S_0 and S_1.

Results in Ng and Wilcox (2009) indicate that a relatively good $1-\alpha$ confidence interval for the slope β_1 is

$$b_1 \pm tS_1, \tag{4}$$

where t is the $1-\alpha/2$ quantile of Student's t distribution with $\nu = n-2$ degrees of freedom. This suggests testing (1) by computing a confidence interval based on (4) but with X and Y replaced by U and V, respectively. This will be called method HC henceforth.

4 Simulation results

Simulations were used to assess the extent to which method HC controls the Type I error probability. The sample size was taken to be $n = 20$ and 100. Estimated Type I error probabilities, $\hat{\alpha}$, were based on 10000 replications.

Four types of distributions were used: normal, symmetric and heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed. More precisely, data were generated from bivariate distributions where the marginal distributions have one of four g-and-h distributions (Hoaglin 1985) that contain the standard normal distribution as a special case. If Z has a standard normal distribution, then by definition

$$V = \begin{cases} \frac{\exp(gZ)-1}{g}\exp(hZ^2/2), & \text{if } g > 0 \\ Z\exp(hZ^2/2), & \text{if } g = 0 \end{cases}$$

has a g-and-h distribution where g and h are parameters that determine the first four moments.

The four distributions used here were 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 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985). The correlation between X and Y was taken to be $\rho = 0$ and .5. To add perspective, results are also reported using the Morgan-Pitman (MP) test as well as the modification of the Morgan-Pitman suggested by McCulloch (1987) where Pearson's correlation is replaced by Spearman's correlation. Using Spearman's correlation is called method SP henceforth.

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

g	h	κ_1	κ_2
0.0	0.0	0.00	3.0
0.0	0.2	0.00	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

Table 2 shows the estimated probability of a Type I error when testing at the .05 level for sample sizes $n = 20$ and 100. As can be seen, method HC performs fairly well. Even for $n = 20$, the estimates range between .047 and .066. The Morgan-Pitman test is extremely unsatisfactory; the estimates exceed .20 when sampling from the heavy-tailed distributions considered here and $n = 20$. Increasing the sample size to 100, it deteriorates, particularly for the heavy-tailed distributions where the estimates range between .355 and .403. Spearman's correlation performs better than the Morgan-Pitman test based on Pearson's correlation, but for heavy-tailed distributions the estimates are greater than or equal to .08 even for $n = 100$.

Power comparisons seem meaningless for situations where the Type I error probability is not controlled reasonably well. Even for the skewed, light-tailed distribution considered here, where the kurtosis is only 3.68, the Morgan-Pitman test does not perform well, particularly as the sample size increases. But to provide at least some perspective, simulations were run again for the bivariate normal case where $\sigma_1 = 1$ and $\sigma_2 = 1.5$. With $\rho = 0$ and $n = 20$, power for methods HC, MP and SP was estimated to be .329, .391 and .346, respectively. Increasing $\sigma_2 = 2$, the estimates were .739, .842 and .775. With $\rho = .5$, $\sigma_2 = 1.5$ and $n = 20$, power for methods HC, MP and SP was estimated to be .421, .498 and .442, respectively. Increasing $\sigma_2 = 1.5$, the estimates were .828, .909 and .855. So for symmetric and sufficiently light-tailed distributions, the Morgan-Pitman test offers a power advantage that would seem to be of practical importance.

5 Illustrations

Rao (1948) reports data on the weight of cork borings from the north, east, west and south side of 28 trees. Comparing the variances of the east and south sides with the Morgan-Pitman test, the p-value is .043. Using the modification of the Morgan-Pitman

Table 2 Estimated probability of a Type I error

			$n = 20$			$n = 100$		
g	*h*	ρ	HC	MP	SP	HC	MP	SP
0.0	0.0	0.0	.047	.051	.052	.048	.050	.045
0.0	0.2	0.0	.065	.256	.088	.057	.355	.090
0.2	0.0	0.0	.053	.075	.052	.052	.090	.057
0.2	0.2	0.0	.066	.286	.087	.050	.403	.085
0.0	0.0	0.5	.052	.050	.050	.047	.051	.049
0.0	0.2	0.5	.059	.242	.080	.051	.354	.086
0.2	0.0	0.5	.052	.076	.056	.052	.087	.055
0.2	0.2	0.5	.061	.275	.086	.048	.398	.081

HC = Heteroscedastic Method
MP = Morgan-Pitman
SP = Spearman

test (method HC), the p-value is .186, the only point being that in practice, the choice of method can make a difference.

The next illustration is based on data from the Well Elderly 2 study Clark et al. (2011). A general goal in this study was to assess the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. A portion of the study was aimed at understanding the impact of intervention on a measure of depressive symptoms based on the Center for Epidemiologic Studies Depressive Scale (CESD). The CESD (Radloff 1977) is sensitive to change in depressive status over time and has been successfully used to assess ethnically diverse older people (Lewinsohn et al. 1988). Higher scores indicate a higher level of depressive symptoms. The sample size is 328.

Figure 3 shows a kernel density estimate of the distribution of CESD scores before and after intervention. Note that for the central portion of the distributions, as well as the left tails, there appears to be little or no difference. The median for both marginal distributions is 10. But Fig. 3 also suggests that more extreme measures of depressive symptoms are less likely after intervention. One way of providing a partial check on this possibility, but certainly not the only way, is to test the hypothesis that the marginal distributions have equal variances. The Morgan-Pitman test was applied yielding a p-value equal to .004. But boxplots suggest that sampling is from skewed, heavy-tailed distributions. So there is some doubt about whether the Morgan-Pitman test provides an adequate test of the hypothesis of equal variances. Using instead method HC, the p-value is .013, which provides more convincing evidence that there is less variation after intervention.

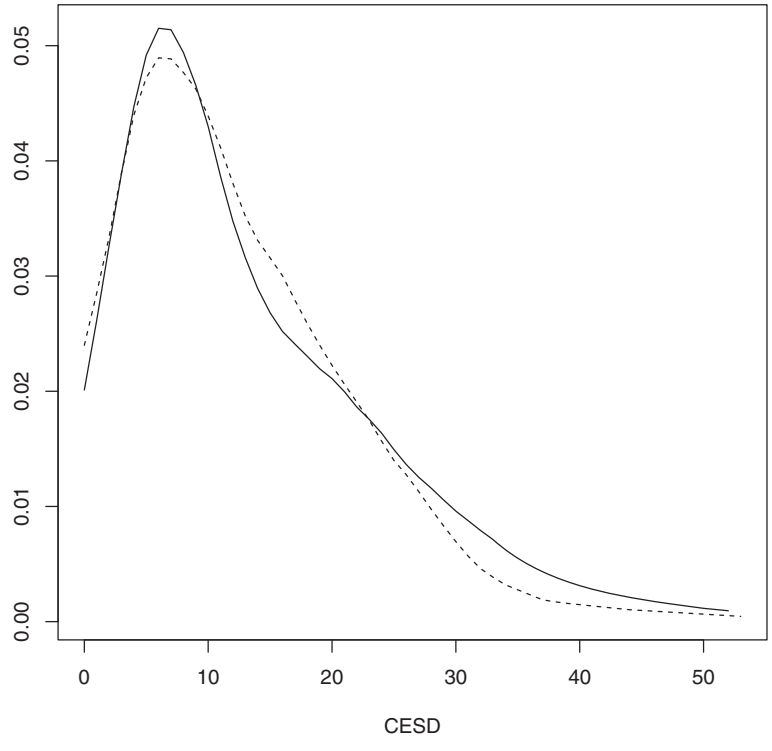

Fig. 3 Kernel density estimates of the distribution of CESD scores before intervention (solid line) and after intervention

In particular, this result suggests that after intervention, relatively high CESD scores are less likely to occur.

6 Concluding remarks

Of course, simulations do not prove that a method provides adequate control over the Type I error probability among all situations that might be encountered. The main result is that method HC continues to perform well in situations where the Morgan-Pitman test, and the variation based on Spearman's correlation, perform poorly.

Heterosceasticity can be addressed using a variety of other methods as noted in the introduction. Evidently, in terms of controlling the probability of a Type I error, alternative methods would provide at best a slight improvement over method HC among the situations considered in the simulations simply because HC performs well. Perhaps situations can be found where some other method for dealing with heteroscedasticity offers a practical advantage, but this remains to be determined.

Competing interests
The author declares that he has no competing interests.

Authors' contributions
RW carried all out all of the analyses and the writing of the paper. The author read and approved the final manuscript.

Received: 7 April 2015 Accepted: 6 August 2015
Published online: 15 August 2015

References

Clark, F, Jackson, J, Carlson, M, Chou, CP, Cherry, BJ, Jordan-Marsh, M, Knight, BG, Mandel, D, Blanchard, J, Granger, DA, Wilcox, RR, Lai, MY, White, B, Hay, J, Lam, C, Marterella, A, Azen, SP: Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomise Controlled Trial. J. Epidemiol. Community Health. **66**, 782–790 (2011). doi:10.1136/jech.2009.099754

Cribari-Neto, F: Asymptotic inference under heteroskedasticity of unknown form. Comput. Stat. Data Anal. **45**, 215–233 (2004)

Cribari-Neto, F, Souza, TC, Vasconcellos, KLP: Inference under heteroskedasticity and leveraged data. Commun Stat - Theory Methods. **36**, 1877–1888 (2007)

Games, PA, Winkler, HB, Probert, DA: Robust tests for homogeneity of variance. Educ. Psychol. Meas. **32**, 887–909 (1972)

Godfrey, LG: Tests for regression models with heteroskedasticity of unknown form. Comput. Stat. Data Anal. **50**, 2715–2733 (2006)

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

Hogg, RV, Craig, AT: Introduction to Mathematical Statistics. 3rd Ed. Macmillan, New York (1970)

Lewinsohn, PM, Hoberman, HM, Rosenbaum, M: A prospective study of risk factors for unipolar depression. J. Abnorm. Psychol. **97**, 251–64 (1988)

Levy, KJ: A procedure for testing the equality of p correlated variances. Br. J. Math. Stat. Psychol. **29**, 89–93 (1976)

Long, JS, Ervin, LH: Using heteroscedasticity consistent standard errors in the linear regression model. Am. Stat. **54**, 217–224 (2000)

Lord, F, Novick, M: Statistical theories of mental test scores. Addison-Wesley, Reading, MA (1968)

McCulloch, CE: Tests for equality of variance for paired data. Commun. Stat. Theory Methods. **16**, 1377–1391 (1987)

Morgan, WA: A test for the significance of the difference between two variances in a sample from a normal bivariate population. Biometrika. **31**, 13–19 (1939)

Mudholkar, GS, Wilding, GE, Mietlowski, WL: Robustness Properties of the Pitman-Morgan Test. Commun. Stat. Theory Methods. **32**, 1801–1816 (2003)

Ng, M, Wilcox, RR: Level robust methods based on the least squares regression line. J. Mod. Appl. Stat. Methods. **8**, 384–395 (2009)

Pitman, EJG: A Note on Normal Correlation. Biometrika. **31**, 9–12 (1939)

Radloff, L: The CESD scale: a self report depression scale for research in the general population. Appl. Psychol. Meas. **1**, 385–401 (1977)

Rao, CR: Tests of significance in multivariate analysis. Biometrika. **35**, 58–79 (1948)

Rothstein, SM, Bell, WD, Patrick, JA, Miller, H: A jackknife test of homogeneity of variance with paired replicates of data. Psychometrika. **46**, 35–40 (1981)

Snedecor, GW, Cochran, W: Statistical Methods. 6th Ed. University Press, Ames, IA (1967)

Tiku, ML, Balakrishna, N: A robust test for testing the correlation coefficient. Commun. Stat. Simul. Comput. **15**, 945–971 (1986)

Wilcox, RR: Comparing the variances of two dependent groups. J. Educ. Stat. **15**, 237–247 (1990)

Wilcox, RR: Introduction to Robust Estimation and Hypothesis Testing. 3rd Edition. Academic Press, San Diego, CA (2012)