Tests for equality of variances between two samples which contain both paired observations and independent observations

Abstract

Tests for equality of variances between two samples which contain both paired observations and independent observations are explored using simulation. New solutions which make use of all of the available data are put forward. These new approaches are compared against standard approaches that discard either the paired observations or the independent observations. The approaches are assessed under equal variances and unequal variances, for two samples taken from the same distribution. The results show that the newly proposed solutions offer Type I error robust alternatives for the comparison of variances, when both samples are taken from the same distribution.

Keywords Brown-Forsythe test; Equal variances; Partially overlapping samples; Pitman-Morgan test; Simulation; Robustness

1. Introduction

An equality of variances test is often performed as a preliminary test to inform the most appropriate statistical test for a comparison of means Mirtagioğlu *et al.* (2017). The pitfalls of this process are well documented (Zimmerman, 2004; Zimmerman and Zumbo, 2009; Rasch *et al.*, 2011; Rochon *et al.*, 2012). This paper considers tests for equality of variances where it is the equality of variances that is of importance in their own right. Examples include a comparison of two treatments that have a similar mean efficacy, or a comparison of products in quality control, or a comparison of variances in human populations. Tests for equal variances have wide ranging applications including areas in archaeology, environmental science, business and medical research (Gastwirth *et al.*, 2009).

Numerous tests for the comparisons of variances for two independent samples have been documented (Conover, *et al.,* 1981). The Pitman-Morgan test is widely regarded as the optimum test of equal variances with two paired samples under normality (Mudholkar *et al.*, 2003). However, situations may arise where there are two samples which contain both independent observations and paired observations (Derrick *et al.*, 2015). For example, when some experimental data in a paired samples design is missing due to an error or accident.

This paper is concerned with the direct comparison of variances between two samples, which contain both paired observations and independent observations. For simplicity, these scenarios are referred to as partially overlapping samples (Martinez-Camblor *et al.,* 2013; Derrick *et al.*, 2017). The conditions of Missing Completely at Random (MCAR) are assumed.

In the two partially overlapping samples scenario, if the number of paired observations is relatively large and the number of independent observations is relatively small, a solution may be to discard independent observations and perform a test for equal variances on the paired observations. The standard F-test is not appropriate for paired samples (Kenny, 1953). For the comparison of variances for paired data, the Pitman-Morgan test can be performed (Pitman 1938; Morgan 1939). However, the Pitman-Morgan test is not robust to violations of the assumption of normality (Mudholkar *et al.*, 2003; Grambsch, 2015). For heavy tailed distributions the Type I error rate of the Pitman-Morgan test is larger than nominal Type I error rate (McCulloch, 1987; Wilcox, 2015).

Alternatively, if the number of independent observations is relatively large and the number of paired observations is relatively small, a solution may be to discard paired observations and perform one of numerous established tests for the comparison of variances with independent observations.

When the normality assumption is met, the standard F-test is the uniformly most powerful test for two independent samples. However, the standard F-test is not robust to deviations from normality (Marozzi, 2011).

Levene (1960) proposed that for two independent groups, the differences between the absolute deviations from the group means could be used to assess equality of variances. In the two sample case, this test is equivalent to Student's t-test applied to absolute deviations from the group means. This version of Levene's test, fails to control the Type I error rate when the population distribution is skewed (Carroll and Schneider, 1985; Nordstokke and Zumbo, 2007).

Brown and Forsythe (1974) proposed alternatives to Levene's test when data are not normally distributed. These alternatives use deviations from the median or trimmed mean. These variations are also often referred to as "Levene's test" (Carroll and Schneider, 1985; Gastwirth *et al.*, 2009). For the avoidance of doubt, in this paper the convention followed is that assessing equality of variances using deviations from the mean is referred to as Levene's test. Assessing equality of variances using deviations from the median is referred to as the Brown-Forsythe test.

Conover *et al.* (1981) explored 56 tests for equal variances for two independent groups and noted that the five tests that are Type I error robust use deviations from the median rather than deviations from the mean. Conover *et al.* (1981) found that the only test that consistently meets Bradley's (1978) liberal Type I error robustness criteria is the Brown-Forsythe test, using absolute deviations from the median. There is no uniformly robust and most powerful test applicable for all distributions and sample sizes. The general consensus is praise of the Brown-Forsythe test using deviations from the median (Carroll and Schneider, 1985; Nordstokke and Zumbo, 2007; Mirtagioğlu *et al.*, 2017). However, it should be noted that this test can be conservative with small sample sizes (Loh, 1987; Lim and Loh, 1995). The use of absolute deviations rather than squared deviations better maintains Type I error robustness (Cody and Smith, 1997).

Performing a test using either only the independent observations or only the paired observations may result in loss of power. The discarding of data is particularly problematic if the overall total sample size is small. In addition, if the assumption of MCAR is not reasonable, the discarding of data is likely to cause bias.

Bhoj (1979, 1984) and Ekbohm (1981, 1982) debated methods using all of the available data for testing the equality of variances in scenarios that they refer to as "incomplete data". In this debate the authors do not recognise that a combination of independent observations and paired observations may occur by design and not only by accident. Bhoj (1979) and Ekbohm (1981, 1982) independently considered a weighted combination of existing independent sum of squares techniques to create a new test statistic. Other solutions such as ignoring the pairing and performing the F-test on all of the available data were considered by Ekbohm (1982). Bhoj (1984) concluded that his test statistic is the most powerful if the correlation is negative or small. Otherwise, performing the F-test on all of the available data is more powerful than the solutions put forward by either of the authors (Ekbolm, 1982; Bhoj 1984).

The simulations performed by these authors were on a relatively small scale, with only 1,000 replicates at each point in their design space. No solution was comprehensively agreed upon for all scenarios, and this is likely to contribute to them not being well established. Furthermore the non-robustness of the Pitman-Morgan test has a detrimental impact on their weighted tests. A solution that uses all available data without a complex weighting structure, or the discarding of valuable information about the pairing, may therefore be advantageous.

For the comparison of means when both independent observations and paired observations are present, partially overlapping samples t-tests are given by Derrick, *et al.* (2017). These solutions are generalised forms of the t-test and are Type I error robust under normality. These solutions are also robust in the comparison of two ordinal samples where the scale represents interval data (Derrick and White, 2018)

We propose that as an alternative test of equal variances when there is a combination of paired observations and independent observations, the partially overlapping samples t-test can be performed, using deviations from the group medians, as outlined below.

Let X_{ji} denote the *i*-th observation in group *j* for *j* = {Sample1, Sample 2}, and $\tilde{X}_j$ denote the sample median, so that $Y_{ji} = \left| X_{ji} - \tilde{X}_j \right|$, then

$$T_{\text{var1}} = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{p\_y}\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2r\left(\frac{n_c}{n_1 n_2}\right)}} \text{ and } S_{p\_y} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}}$$

The test statistic T_{var1} is referenced against the t-distribution with degrees of freedom:

$$v_1 = (n_c - 1) + \left(\frac{n_a + n_b + n_c - 1}{n_a + n_b + 2n_c}\right)(n_a + n_b).$$

where $n_a =$ number of unpaired observations exclusive to Sample 1, $n_b =$ number of unpaired observations exclusive to Sample 2, $n_c =$ number of pairs, $n_j =$ total number of observations in Sample *j*, $S_j^2 =$ variance of Sample *j* based on the Y_{ji} observations.

For the comparison of variances, Loh (1987) suggested adapting the unequal variances t-test using deviations from the medians. For the comparison of means, Student's t-test is sensitive to deviations from the equal variances assumption (Ruxton, 2006; Derrick, Toher and White, 2016). As a result of this Derrick *et al*. (2017) additionally proposed the partially overlapping samples t-test for unequal variances. We propose that the partially overlapping samples test statistic unconstrained to equal variances can be similarly modified to provide a test for equality of variances so that:

$$T_{\text{var2}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2r\left(\frac{S_1 S_2 n_c}{n_1 n_2}\right)}}$$

The test statistic T_{var2} is referenced against the t-distribution with degrees of freedom:

$$v_2 = (n_c - 1) + \left(\frac{\gamma - n_c + 1}{n_a + n_b + 2n_c}\right)(n_a + n_b) \text{ where } \gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2 / n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2 / n_2\right)^2}{n_2 - 1}}$$

Methodology for assessing the Type I error rate of these proposals is given in Section 2, with an example application given in Section 3.

2. Methodology

For two samples containing both independent observations and paired observations, approaches for the comparison of variances are assessed using simulation. The approaches considered are the Brown-Forsythe test, the Pitman-Morgan test, and the proposed T_{var1} and T_{var2}. Type I error robustness is assessed using Bradley's (1978) liberal robustness criteria. Power is assessed for test statistics that do not violate Bradley's liberal criteria.

Within the simulation design, the sizes of n_a, n_b, n_c are {5, 10, 30, 50}. The correlation coefficients ρ are {0.00, 0.25, 0.50, 0.75}. Simulations for each possible parameter combination of n_a, n_b, n_c, ρ are performed in a factorial design. Standard Normal deviates are calculated using the Box-Muller (1958) transformation. For the n_c observations, correlated Standard Normal deviates are obtained as per Kenney and Keeping (1951)

In Section 4.1, the comparison of variances is performed for normally distributed data. Under the null hypothesis, X_1 ~ N(0,1) and X_2 ~ N(0,1). Under the alternative hypothesis, the observations in Sample 2 are multiplied by two, thus X_1 ~ N(0,1) and X_2 ~ N(0,4).

In Section 4.2, the comparison of variances is performed for skewed distributions. Under the null hypothesis, Normal deviates are first generated as above, and then the exponential of each value is calculated. Under the alternative hypothesis this process is repeated, and each of the observations in Sample 2 are multiplied by two to create unequal variances.

For each parameter combination, the data generating process is repeated 10,000 times, and each of the statistical tests to be evaluated is performed on each replicate. Under the null hypothesis, the proportion of the replicates where the null hypothesis is rejected represents the Type I error rate. Under the alternative hypothesis, the proportion of the replicates where the null hypothesis is rejected, represents the power of the test, assuming Type I error rates can be reasonably compared. The simulations and tests are performed in R, at the 5% significance level, two-sided.

The simulation design allows that the conditions of MCAR can be assumed.

3. Example

In the assessment of an undergraduate university module, two lecturers share the marking of 32 student submissions. As part of the marking regulations, at random six of the submissions are independently assessed by both lecturers. The remaining submissions are randomly split between the two lecturers, ensuring that both have an equal number to assess. Thus Lecturer 1 has one sample comprising of six paired observations and 13 independent observations. Likewise, Lecturer 2 has a sample of equal size. The samples are partially overlapping by design, thus MCAR can be reasonably assumed.

There is concern that the lecturers do not allocate marks at the top end and the bottom end of the marking scale in the same way. Tests for equal variances are performed on the independent observations (Table 1), the paired observations (Table 2), and all observations.

Table 1. Marks awarded to the 26 students randomly allocated to the lecturers.

Lecturer 1	55	56	58	60	60	60	61	61	62	62	64	65	67
Lecturer 2	40	50	51	60	60	60	60	60	61	66	69	72	82

Table 2. Marks awarded by each lecturer for the six students that are marked by both.

Student	A	B	C	D	E	F
Lecturer 1	54	55	60	63	65	70
Lecturer 2	50	56	60	61	67	73

The Brown-Forsythe test is performed on the data in Table 1 using the R package “lawstat” (Gastwirth *et al.*, 2015). This shows no evidence to reject the null hypothesis of equal variances (t = -1.9673, v = 24, p = 0.061).

The Pitman-Morgan test is performed on the data in Table 2 using the R package “PairedData” (Champely, 2013). This shows no evidence to reject the null hypothesis of equal variances (t = -2.352, v = 4, p = 0.078).

In order to perform the tests for equal variances using all of the available data, for each submission marked my Lecturer 1 the absolute deviation from the median mark given by Lecturer 1 is calculated. Similarly, the absolute deviations for Lecturer 2 are calculated.

The partially overlapping samples t-test is performed on the absolute deviations using the R package “Partiallyoverlapping” (Derrick, 2017). The null hypothesis of equal variances is rejected at the 5% significance level for both the equal variances assumed variant (t_{var1} = -2.324, v_1 = 26.211, p = 0.028) and the equal variances not assumed variant (t_{var2} = -2.183, v_2 = 17.488, p = 0.043). It would appear that Lecturer 2 is making greater use of the full range of potential marks relative to Lecturer 1.

4.1 Comparison of variances for two samples from the Normal distribution

Type I error rates and power are summarised for each of; the Brown-Forsythe test, BF, the Pitman-Morgan test, PM, and the partially overlapping samples tests, T_{var1} and T_{var2}.

Each of the test statistics are assessed under the null hypothesis where X_1 ~ N (0,1) and X_2 ~ N (0,1). The Type I error robustness for each of the parameter combinations within the simulation design are summarised in Figure 1.

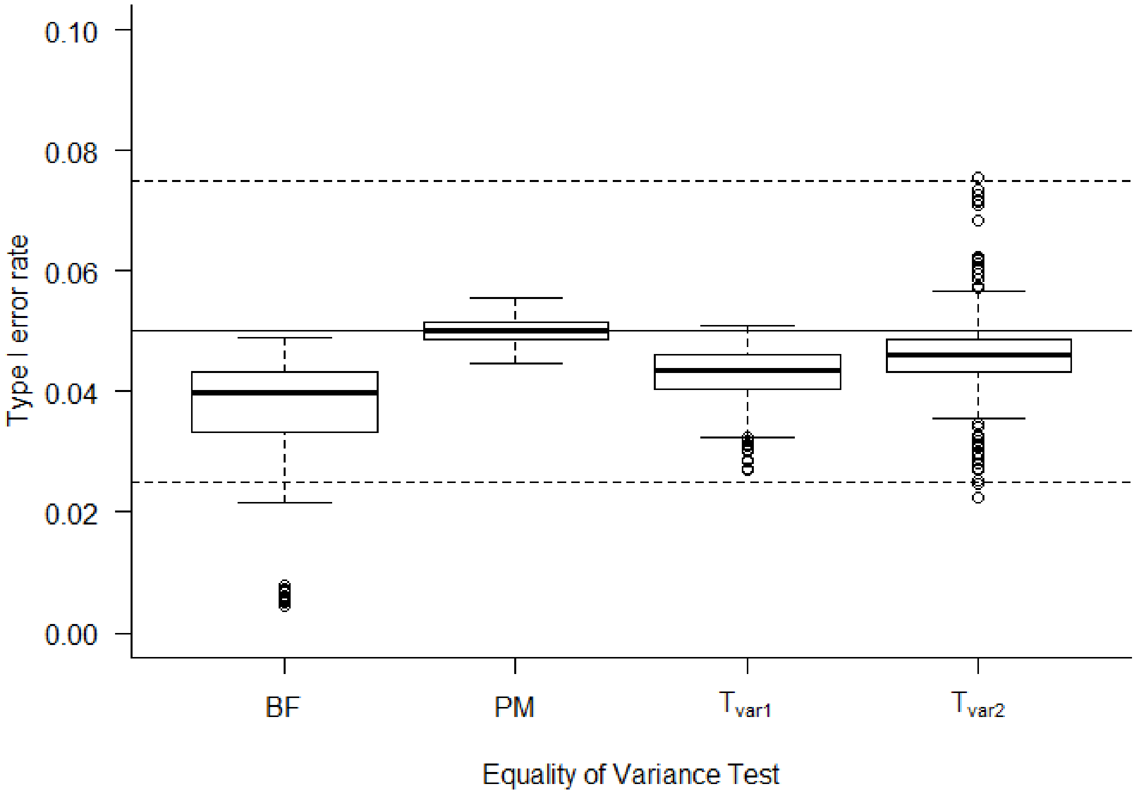

Figure 1. Type I error robustness for each parameter combination, assessed against Bradley's liberal criteria, samples from Standard Normal distribution

Figure 1 shows that the Pitman-Morgan test and the proposed test statistics are Type I error robust throughout the simulation design, with T_{var1} being more conservative than T_{var2}. For the smallest sample sizes within the design, the Brown-Forsyth test is very conservative.

Relative power comparisons for each of the test statistics are assessed where X_1 ~ N (0,1) and X_2 ~ N (0,4). The power averaged across the simulation design for increasing ρ is given in Figure 2.

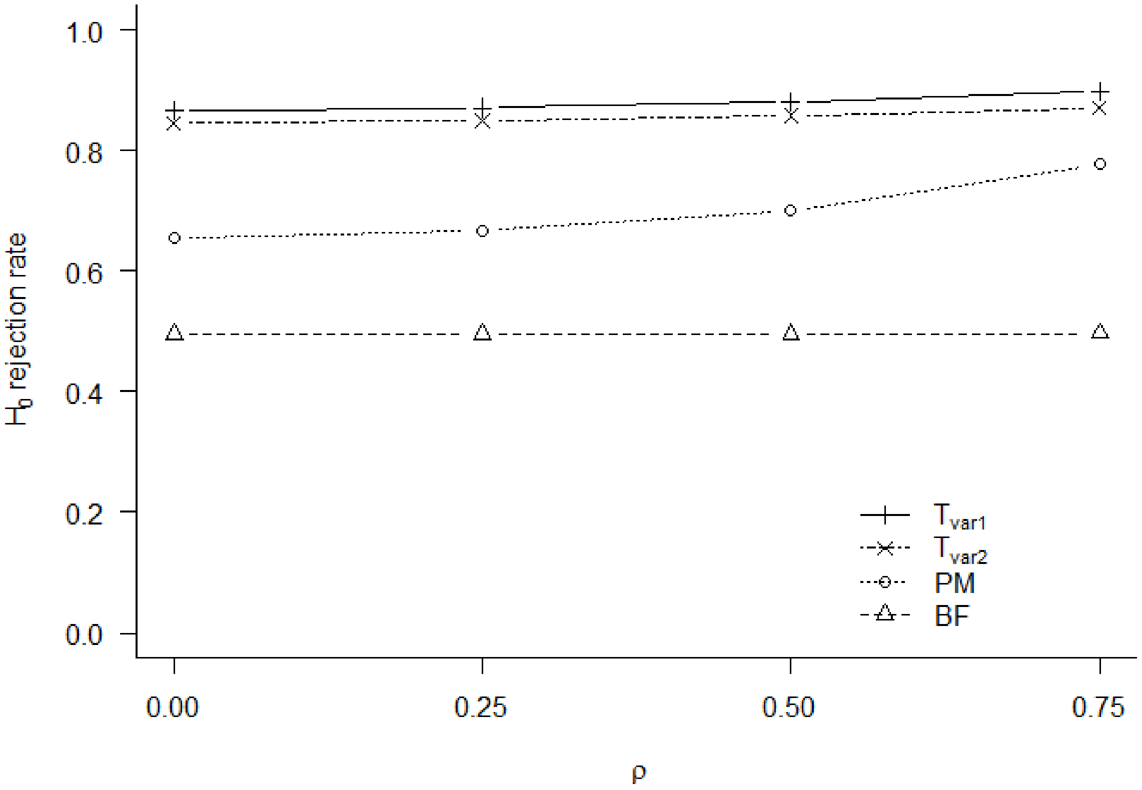

Figure 2. Relative power, averaged across the simulation design for increasing ρ, samples from Normal distributions.

Figure 2 shows that the proposed test statistics T_{var1} and T_{var2} perform similarly to each other under normality, and they have superior power qualities to the standard tests which discard data.

4.2 Comparison of variances for two samples from skewed distributions

Each of the test statistics are assessed when both samples are taken from skewed but identical distributions. The Type I error robustness for each of the parameter combinations within the simulation design are summarised in Figure 3.

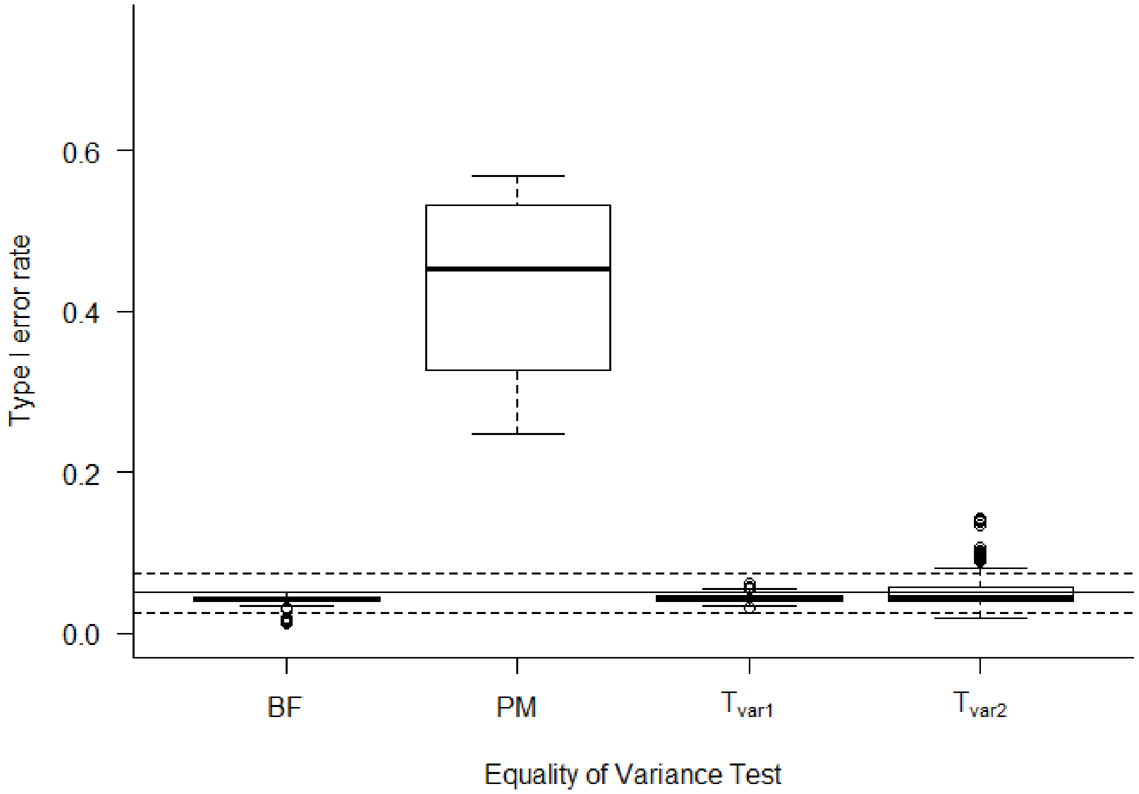

Figure 3. Type I error robustness for each parameter combination, assessed against Bradley's liberal criteria, samples from skewed distribution.

Figure 3 shows that the Pitman-Morgan test is not Type I error robust when the samples are taken from identical heavy tailed distributions. This supports the findings by McCulloch (1987) and Wilcox (2015). In addition it can be seen that T_{var2} does not fully maintain Type I error robustness. Further investigation shows that T_{var2} is liberal when one of the samples is more dominant in terms of size, and when there is a large imbalance between the number of independent observations and the number of pairs.

Relative power comparisons for each of the test statistics are assessed where the samples are taken from different skewed distributions. Due to the poor Type I error robustness of the Pitman-Morgan test and T_{var2}, this comparison is done only for the Brown-Forsythe test and T_{var1}. The power averaged across the simulation design for increasing ρ is given in Figure 4.

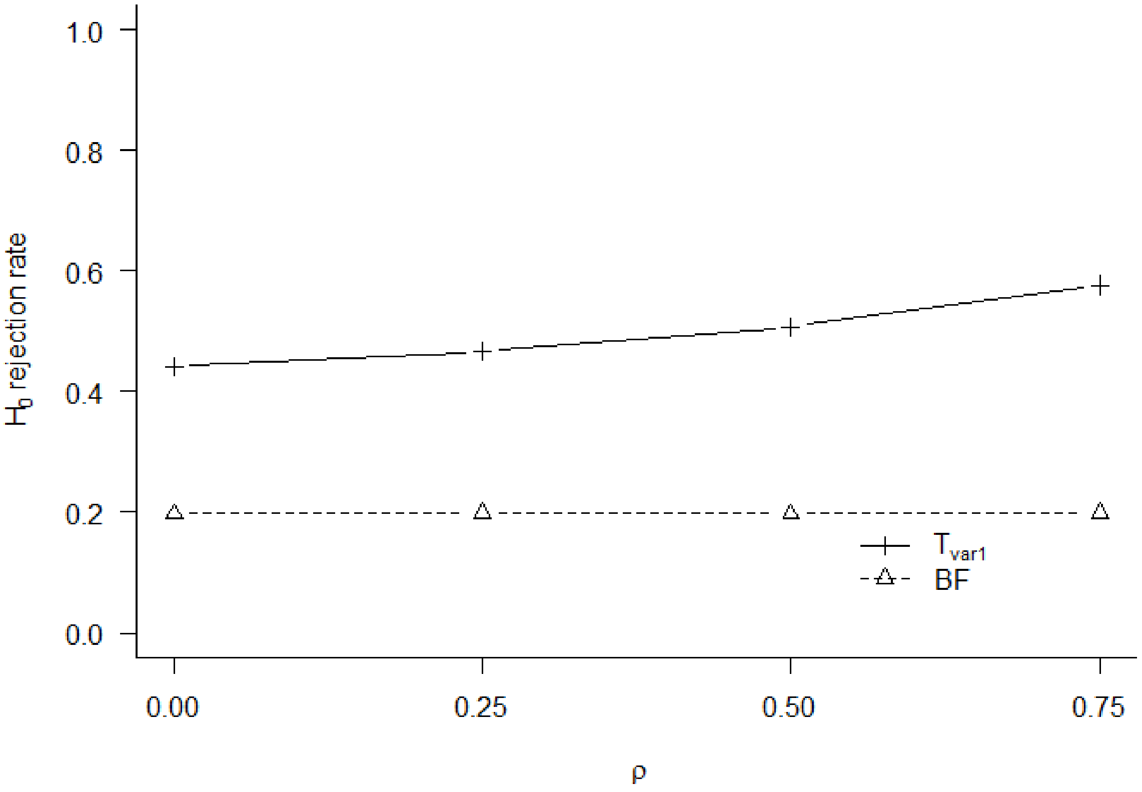

Figure 4. Relative power, averaged across the simulation design for increasing ρ, samples from skewed distributions.

Figure 4 shows that the proposed solution, T_{var1}, is more powerful than the Brown-Forsythe test. A comparison of Figure 4 against Figure 2 also indicates that both the Brown-Forsythe test and the newly proposed test, T_{var1}, are less powerful when samples are taken from a heavy-tailed distribution.

5. Conclusion

A common research question in psychology, education, medical sciences, business and manufacturing, is whether or not the variances are equal (Gastwirth, Gel and Miao, 2009).

There has been little research into techniques for the comparison of variances for samples that contain both independent observations and paired observations. Standard solutions that involve discarding data are less than desirable. Two solutions that make use of the tests statistics by Derrick *et al.* (2017) are proposed in this paper. Simulations across a range of sample sizes show that these solutions are Type I error robust under normality and the assumption of MCAR. These solutions are more powerful than established solutions that discard data, namely the Pitman-Morgan test and the Brown-Forsythe test.

The equal variances form of the partially overlapping samples variances test, T_{var1}, is marginally more powerful than the unconstrained form of the test T_{var2}.

The proposed test statistic T_{var1} further maintains Type I error robustness for skewed distributions where T_{var2} does not. T_{var1} is therefore recommended as a powerful alternative to test for the equality of variances between two samples when there is a combination of paired observations and independent observations in two samples.

References

Bhoj, D. S. (1979). Testing equality of variances of correlated variates with incomplete data on both responses. *Biometrika. 66*(3), 681–683. doi: 10.1093/biomet/66.3.681

Bhoj, D. S. (1984). On testing equality of variances of correlated variates with incomplete data. *Biometrika. 71*(3), 639–641. doi: 10.1093/biomet/71.3.639

Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Brown, M. B., and Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association, 69*(346), 364-367. doi: 10.1080/01621459.1974.10482955

Carroll, R. J., and Schneider, H. (1985). A note on Levene's tests for equality of variances. *Statistics and probability letters, 3*(4), 191-194. doi: 10.1016/0167-7152(85)90016-1

Champely, S. (2013). PairedData: Paired Data Analysis. R package version 1.0.1.

Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics, 23*(4), 351-361. doi: 10.1080/00401706.1981.10487680

Cody, R. P., and Smith, J. K. (1986). *Applied statistics and the SAS programming language*. Elsevier North-Holland Inc.

Derrick, B. (2017). Partiallyoverlapping: Partially Overlapping Samples t-Tests. R package version 1.0

Derrick, B., Dobson-McKittrick, A., Toher, D., and White P. (2015). Test statistics for comparing two proportions with partially overlapping samples. *Journal of Applied Quantitative Methods, 10*(3), 1-14.

Derrick, B., Russ, B., Toher, D., and White P. (2017). Test statistics for the comparison of means for two samples which include both paired observations and independent observations. *Journal of Modern Applied Statistical Methods, 16*(1), 137-157. doi: 10.22237/jmasm/1493597280

Derrick, B., Toher, D., and White, P. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods for Psychology, 12*(1), 30-38. doi: 10.20982/tqmp.12.1.p030

Derrick, B., White, P. (2018) Methods for comparing the responses from a Likert question, with paired observations and independent observations in each of two samples. *International Journal of Mathematics and Statistics [in press]*

Ekbohm, G. (1981). A test for the equality of variances in the paired case with incomplete data. *Biometrical Journal. 23*(3), 261–265. doi: 10.1002/bimj.4710230306

Ekbohm, G. (1982). On comparing variances in the paired case with incomplete data. *Biometrika. 69*(3), 670–673. doi: 10.1093/biomet/69.3.670

Gatwirth, J. L., Gel, Y. R., and Miao, W. (2009). The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. *Statistical Science, 24*(3), 343-360.

Gatwirth, J. L., Gel, Y. R., Wallace-Hui, W. L., Lyubchich, V., Miao, W., and Noguchi, K. (2015). lawstat: Tools for Biostatistics, Public Policy, and Law. R package version 3.0.

Grambsch, P. M. (1994). Simple robust tests for scale differences in paired data. *Biometrika, 81*(2), 359-372.

Kenney, J. F., and Keeping, E. S. (1951). Mathematics of statistics, Princeton, NJ: Van Nostrand, Part. 2, 2nd edition.

Kenny, D. T. (1953). Testing of differences between variances based on correlated variates. *Canadian Journal of Experimental Psychology 7*(1), 25. doi: 10.1037/h0083569

Levene, H. (1960). Robust tests for equality of variances. *Contributions to Probability and Statistics, 1,* 278-292.

Lim, T. S., and Loh, W. Y. (1996). A comparison of tests of equality of variances. *Computational Statistics and Data Analysis, 22*(3), 287-301.

Loh, W. Y. (1987). Some modifications of Levene's test of variance homogeneity. *Journal of Statistical Computation and Simulation, 28*(3), 213-226.

Marozzi, M. (2011). Levene type tests for the ratio of two scales. *Journal of Statistical Computation and Simulation, 81*(7), 815-826. doi: 10.1080/00949650903499321

Martínez-Camblor, P., Corral, N., and María de la Hera, J. (2013). Hypothesis test for paired samples in the presence of missing data. *Journal of Applied Statistics, 40*(1), 76-87. doi: 10.1080/02664763.2012.734795

McCulloch, C. E. (1987). Tests for equality of variances with paired data. *Communications in Sstatistics - Theory and Methods, 16*(5), 1377-1391. doi: 10.1080/03610928708829445

Mirtagioğlu, H., Yiğit, S., Mendeş, E., and Mendeş, M. (2017). A Monte Carlo Simulation Study for Comparing Performances of Some Homogeneity of Variances Tests. Journal of Applied Quantitative Methods, 12(1), 1-11.

Morgan, W. A. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika, 31*(1/2), 13-19.

Mudholkar, G. S., Wilding, G. E., and Mietlowski, W. L. (2003). Robustness properties of the Pitman–Morgan test. *Communications in Statistics - Theory and Methods, 32*(9), 1801-1816. doi: 10.1081/STA-120022710

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, *29*(3/4), 322-335.

Rasch, D., Kubinger, K. D., and Moder, K. (2011). The two sample t-test: pre-testing its assumptions does not pay off. *Statistical Papers, 52*(1), 219-231. doi: 10.1007/s00362-009-0224-x

Rochon, J., Gondan, M., and Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology, 12*, 81 doi: 10.1186/1471-2288-12-81

Wilcox, R. (2015). Comparing the variances of two dependent variables. *Journal of Statistical Distributions and Applications*, *2*(1), 1-8. doi: 10.1186/s40488-015-0030-z

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*(1), 173-181. doi: 10.1348/000711004849222

Zimmerman, D. W., and Zumbo, B. D. (2009). Hazards in choosing between pooled and separate-variances t tests. *Psicológica: Revista de metodología y psicología experimental, 30*(2), 371-390.

Zimmerman, D. W., and Zumbo, B. D. (1993). *The relative power of parametric and nonparametric statistical methods.* In G. Keren and C. Lewis (Eds.) A handbook for data analysis in the behavioral sciences: Methodological issues (481-517). Hillsdale, NJ: Erlbaum.