SORT 42 (1) January-June 2018, 59-72 DOI: 10.2436/20.8080.02.69

Heteroscedasticity irrelevance when testing means difference

Pablo Flores M.[1] and Jordi Ocaña[2]

Abstract

Heteroscedasticity produces a lack of type I error control in Student's t test for difference between means. Pretesting for it (e.g., by means of Levene's test) should be avoided as this also induces type I error. These pretests are inadequate for their objective: not rejecting the null hypotheses is not a proof of homoscedasticity; and rejecting it may simply suggest an irrelevant heteroscedasticity. We propose a method to establish irrelevance limits for the ratio of variances. In conjunction with a test for dispersion equivalence, this appears to be a more affordable pretesting strategy.

MSC: 62F03.

Keywords: Homoscedasticity, equivalence test, indifference zone, pretest, Student's t test

1. Introduction

Student's t test for determining possible inequalities between two population means is subject to normality and homoscedasticity assumptions. (Readers not familiar with basic statistical techniques, such as Student's or Welch's test, may refer to sources like the SPSS tutorial at https://libguides.library.kent.edu/SPSS/IndependentTTest.) Presumably, these assumptions are or are not confirmed by means of other (pre)tests on the same data. The pretests (or the order in which they are applied) may vary. When the null hypothesis of a normality test (Shapiro-Wilk, Kolmogorov-Smirnov, etc.) is rejected, the traditional procedure is to assume that the sample does not come from a normal distribution. In such cases, a non-parametric approach is adopted, for example the Wilcoxon's test to compare the location parameters of two independent samples – possibly under the additional yet false assumption of its supposedly higher robustness to dispersion differences. Otherwise, when the null hypothesis of normality is not rejected, this assumption

[1] Grupo de Investigación en Ciencia de Datos CIDED, Escuela Superior Politécnica de Chimborazo, Facultad de Ciencias, Panamericana Sur km 1 1/2, EC060155 Riobamba, Ecuador; email: p_flores@espoch.edu.ec

[2] Departament de Genètica, Microbiologia i Estadística, Secció d'Estadística, Universitat de Barcelona, Facultat de Biologia, Diagonal 643, 08028 Barcelona, Spain; email: jocana@ub.edu

Received: December 2017
Accepted: May 2018

is taken as true; and pretesting then proceeds to the next step by means of a test with perfect homoscedasticity as the null hypothesis (F, Levene, Bartlett, Cochran, etc.). If its null hypothesis is not rejected, then homoscedasticity is taken as true. This leads to the use of Student's t test as an adequate procedure for comparing means. Otherwise, heteroscedasticity is assumed and a procedure like Welch's test (Welch, 1947) is adopted.

Although it is not unusual to find such pretesting recommendations, several studies (Hsu, 1938; Overall, Atlas and Gibson, 1995; Scheffé, 1970) show that these strategies alter the overall type I error probability (TIEP) especially when sample sizes are unequal. Zimmerman (2004) performed a simulation study using different sample sizes, levels of heteroscedasticity and levels of significance to estimate the overall TIEP. The results showed that when Student's test is performed without any homoscedasticity pretesting, and when Levene's pretest is used to decide between Student's or Welch's test, the overall TIEP is severely inflated. On the other hand, the TIEP for Welch's test remains close to the significance level for all heteroscedasticity levels. In strategies that alter the TIEP, the largest variance associated with the largest sample size deflates the TIEP while it is inflated when the largest variance is associated with the smallest sample size. The severity of this distortion increases with the heteroscedasticity level. In addition, overall TIEP distortion increases as the significance level of the pretest decreases, the overall TIEP ceases to be affected at high levels of significance in the preliminary test, e.g., at the non-usual value $\alpha = 0.20$.

Rasch, Kubinger and Moder (2011), state that pretesting to validate the assumptions in the comparison of means test leads to alterations in the type I and type II error probabilities. These authors show that using a pretest for normality (Kolmogoroff-Smirnov) and a pretest for equality of variances (Levene) causes an increase in the overall TIEP. In contrast, when Welch's test is used directly (without a pretest), these overall TIEP distortions largely disappear. They conclude that pretesting does not pay off. Instead, applying Welch's test directly without pretesting is best, and it should be recommended in textbooks as well as implemented in statistical software as the standard option for comparing means. In addition, the authors advise that Wilcoxon's and Student's t-test should never be used.

The next section introduces some concepts and notation in equivalence testing. Section three describes the algorithm that we have used to determine these irrelevance limits. In the fourth section, a simulation study comparing the previously cited pretesting strategies is presented. In the fifth section, two illustrative examples are presented. Finally, in the last section the main conclusions are discussed.

2. Equivalence testing concepts and some additional notation

The above results contribute to other evidence indicating that pretesting in order to fulfill validity conditions (not only in the problem of means comparison) is not a reliable strategy. However, one may ask if this inadequacy is (fully or partially) due to the fact that these pretests are intrinsically inappropriate for their goal: Note that their null hypothesis states complete fulfilment of the normality or homoscedasticity assumptions. As is well known, not rejecting the null hypothesis is not a proof of its correctness, while rejecting it may simply indicate an irrelevant departure from perfect normality or homoscedasticity. In other words, asserting that there is a non-significant difference between variances should not be confused with there being homogeneity. In the words of Altman and Bland (1995) "*Absence of evidence is not evidence of absence*".

Figure 1 schematically shows these ideas in the specific case of the homoscedasticity assumption, which will constitute the focus of the present paper.

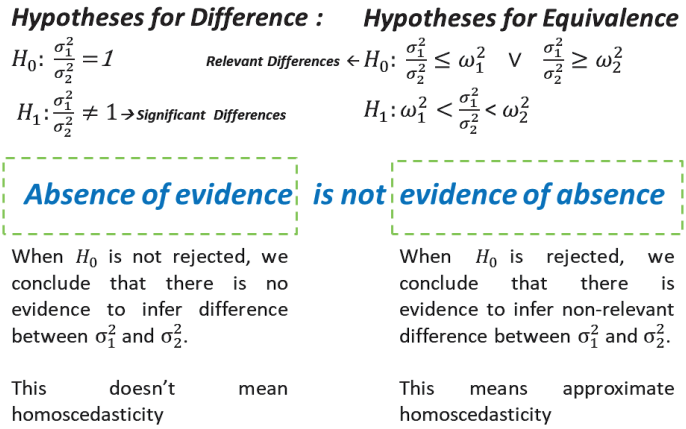

Figure 1: *Traditional and equivalence approach.*

Wellek (2010) (p. 164), proposes an approach that is based on equivalence testing. In this class of tests, the alternative hypothesis states equivalence, i.e., perfect fit (to normal) or equality (of variances) **except for irrelevant deviations** while the null hypothesis states relevant ones. In this approach, the relevant differences between variances are stated in the null hypothesis; thus the assumption of near homoscedasticity is reinforced if the null is rejected.

In brief, Wellek's test may be described as follows: For the hypotheses

$$\begin{aligned} H_0 &: \frac{\sigma_1^2}{\sigma_2^2} \leq \omega_1^2 \wedge \frac{\sigma_1^2}{\sigma_2^2} \geq \omega_2^2 && \text{No equivalence (relevant difference of variances)} \\ H_1 &: \omega_1^2 < \frac{\sigma_1^2}{\sigma_2^2} < \omega_2^2 && \text{Equivalence (non-relevant difference)} \end{aligned} \tag{1}$$

with $\omega_1^2 < 1 < \omega_2^2$, a uniformly more powerful invariant test is one whose critical region is given by:

$$\{\widetilde{C}^{(1)}_{\alpha,n_1-1,n_2-1}(\omega_1^2,\omega_2^2) < Q < \widetilde{C}^{(2)}_{\alpha,n_1-1,n_2-1}(\omega_1^2,\omega_2^2)\},$$

where Q stands for the test statistic:

$$Q = \frac{S_X^2}{S_Y^2} = \frac{(n_2-1)}{(n_1-1)} \frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2}{\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2} \tag{2}$$

and the critical constants $\widetilde{C}^{(1)}_{\alpha,n_1-1,n_2-1}(\omega_1^2,\omega_2^2)$, $\widetilde{C}^{(2)}_{\alpha,n_1-1,n_2-1}(\omega_1^2,\omega_2^2)$, are determined by the equations:

$$F_{n_1-1,n_2-1}(\widetilde{C}_2/\omega_1^2) - F_{n_1-1,n_2-1}(\widetilde{C}_1/\omega_1^2) = \alpha = F_{n_1-1,n_2-1}(\widetilde{C}_2/\omega_2^2) - F_{n_1-1,n_2-1}(\widetilde{C}_1/\omega_2^2).$$

$F_{n_1-1,n_2-1}()$ corresponds to the cumulative distribution function of a centred F distribution, with n_1-1 degrees of freedom in the numerator and n_2-1 in the denominator.

One of the most important aspects of equivalence testing is to establish the equivalence limits. Wellek does not propose a technical criterion to determine them, instead he provides some hints based on what he calls "a common statistical sense", which may not be enough in many applications. For this reason, here we develop a procedure that allows us to calculate these limits in the specific problem of determining (enough) homoscedasticity when the end objective is to perform a comparison of means and assuming that normality is fulfilled. As its input, the procedure requires objective information on the experimental design and (admittedly, less objective) information on the tolerable possible distortion in the TIEP (perhaps with the help of "common statistical sense").

3. Irrelevance limits for the ratio of dispersions of two Gaussian distributions

As mentioned above, using an equivalence dispersion test for two Gaussian distributions as an homoscedasticity pretest overcomes the logical difficulty of approaches like the F test when it is used for the same purpose. However, the equivalence approach has a notable ambiguity: The values of the equivalence or irrelevance limits (ω_1^2, ω_2^2) that define the hypotheses to be tested must be specified. Criteria such as common statistical sense or the researchers prior knowledge on their subject of interest may be subjective and insufficient.

If the equivalence test refers to a parameter involved in a validity requirement for another test – for example, the ratio of variances for Student's t test – then one possibility is to define an irrelevance limit $\delta > 0$ for the difference between the true TIEP and the significance level α. Obviously $\alpha \pm \delta$ must be inside the $(0,1)$ interval. This irrelevance (or permissiveness or indifference) parameter δ is the maximum distance above and

below α that is acceptable as an irrelevant affectation of the TIEP. This approach may seem to be imprecise and prone to arbitrariness but it follows the line of thought (fairly correct in our opinion) that these validity conditions are just idealizations. Possibly, perfect normality and perfect homoscedasticity are never present in nature. As Box (1979) states about normality, a normal distribution does not exist in the real world, but models known to be false often derive in useful approximate results; what is really important is not whether the populations "are normal" but knowing if the approximate model is good enough to be useful. In our approach the approximation will or will not be considered good based on how close the true TIEP is to the nominal significance level. In this same sense, Cochran (1942) suggested that a distance of 20% of the true TIEP from the nominal significance level is an acceptable approximation. This authoritative criterion, known sometimes as "Cochran's Criterion", could be used as the default in algorithms implementing the method proposed here.

In Student's t test, the true TIEP is a continuous function of the population ratio of variances $\omega^2 = \sigma_1^2/\sigma_2^2$ and its value equals the nominal significance level α at $\omega^2 = 1$. From this point of complete homoscedasticity (and depending also on the sample sizes), this TIEP function may be of an increasing or decreasing nature. As a consequence, ω_1^2 and ω_2^2 define an interval around 1 and they correspond to the ratio ω^2 values where the TIEP equals $\alpha - \delta$ or $\alpha + \delta$.

Given a nominal significance level α, a degree of permissiveness δ and sample sizes n_1, n_2, the procedure for obtaining the pair (ω_1^2, ω_2^2) is based on a simulation iterative process. More precisely, starting from a ratio in the neighbourhood of $\omega^2 = 1$, the true TIEP of Student's t test is obtained by simulation, as the proportion of null hypothesis rejections. This process is iterated by progressively decrementing or incrementing this ratio until crossing the threshold $\alpha \pm \delta$ and until the TIEP reaches these limits with a given precision. The following additional safeguard is included: Provided that the resulting TIEP in each simulation iteration is just an estimation of the true TIEP, the algorithms implementing the method may require that a confidence interval for the true TIEP must be fully included inside $\alpha \pm \delta$.

The simulation process is fast because, to repeatedly generate Student's t statistic values, it is not necessary to simulate pairs of independent Gaussian full data samples of sizes n_1 and n_2, respectively, and then compute the t statistic from them. Instead, provided that we are simulating under a Student's t test scenario of true null hypothesis, the difference of the sample means (the numerator of the t statistic) can be directly generated from a Gaussian distribution with zero mean and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. In addition, the sum of squares necessary for computing the pooled variance estimate, $\sum_{i=1}^{n_1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_{i=1}^{n_2}\left(X_{2i} - \overline{X}_2\right)^2$ can be directly generated as the sum of two independent values (and also independently from the difference between sample means) that are generated from a distribution $\sigma_i^2 \chi_{n_i-1}^2, i = 1,2$, where χ_ν^2 stands for a chi-square distribution with ν degrees of freedom. A further simplification comes from the fact that the only relevant parameter is the ratio of variances and not the variances themselves; thus, one of the variances to be simulated can be fixed at one. What is more,

because complete symmetry exists between the equivalence limits in balanced cases, it is sufficient to obtain only one of them, e.g., the second one, ω_2^2, and then compute $\omega_1^2 = 1/\omega_2^2$. Finally, a variance reduction technique based on the method of"control variates" is applied to avoid the need for very large numbers of simulation replicates to deliver acceptable precision. This technique is also applied in the simulations described in the next chapter, and it is explained in the Appendix.

Table 1: *Indifference zone* (ω_1^2, ω_2^2) *with* $\delta = 0.2\alpha$.

	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$n = (5,5)$	(0.130 - 7.691)	(0.225 - 4.428)	(0.397 - 2.519)
$n = (3,7)$	(0.709 - 1.412)	(0.779 - 1.289)	(0.819 - 1.163)
$n = (7,3)$	(0.711 - 1.410)	(0.776 - 1.325)	(0.832 - 1.166)
$n = (10,10)$	(0.002 - 501.0)	(0.097 - 10.325)	(0.282 - 3.542)
$n = (6,14)$	(0.727 - 1.408)	(0.783 - 1.292)	(0.846 - 1.157)
$n = (14,6)$	(0.716 - 1.362)	(0.787 - 1.264)	(0.859 - 1.148)
$n = (5,10)$	(0.679 - 1.387)	(0.741 - 1.286)	(0.819 - 1.196)
$n = (10,5)$	(0.716 - 1.452)	(0.786 - 1.331)	(0.862 - 1.256)

For illustrative purposes, Table 1 displays the irrelevance limits for some sample sizes (balanced and unbalanced) and significance level scenarios. These values were obtained from 100000 simulation replicates. The results show that, first, there is more heteroscedasticity permissiveness (wider irrelevance intervals) in the balanced scenarios than in the unbalanced ones and, second, that larger sample sizes correspond to wider irrelevance intervals in the balanced cases.

4. Results on pretesting homoscedasticity

4.1. Overall TIEP affectation when the F pretest is used to verify the homoscedasticity assumption

Many tests have been developed for the hypotheses

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$
$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1, \tag{3}$$

to eventually prove heteroscedasticity – and not homoscedasticity. As has been previously stated, some studies use Levene's test as their pretesting option. Provided that the test for heteroscedasticity irrelevance considered in this paper is based on the ratio

Q of sample variances and the Fisher-Snedecor **F** distribution, for the sake of comparison we consider here the traditional F test that is based on the Q statistic to prove heteroscedasticity and we then use it as a reference for comparison with the equivalence approach. However, very similar results were delivered by complementary simulations using Levene's test and other tests for heteroscedasticity (not presented here). At this point, it would be fair to advise against the widespread use of the F test given its lack of robustness in front of departures from normality (see, for example, point 4.3 in Rasch and Guiard, 2004). These drawbacks do not invalidate the results in the present paper because we assume and simulate under perfect normality of data conditions. However, these considerations may be of obvious practical interest.

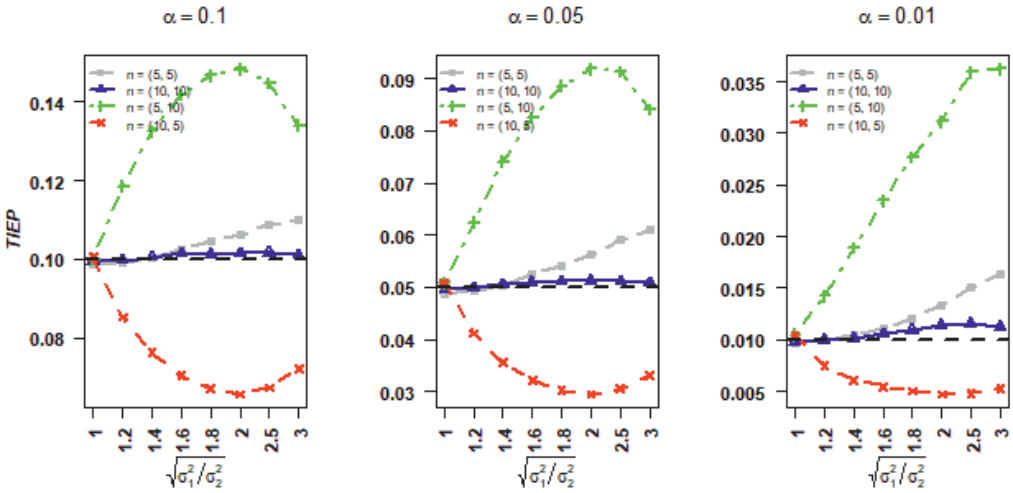

Figure 2: *Overall TIEP estimation when Student's t or Welch's test are conditioned to the result of the F test: If the null hypothesis of variances equality is not rejected, then Student's t test is applied; otherwise Welch's test is applied. The scale of the TIEP axis differs in accordance with the different significance levels under consideration. The relative distance from the nominal significance level is of importance here.*

Figure 2 illustrates similar results to those obtained in the references cited in this paper. They were obtained from 100000 simulation replicates and correspond to scenarios defined by crossing significance levels of $\alpha = 0.1$, 0.05 and 0.01; several heteroscedasticity degrees given by the ratio $\omega = \sqrt{\sigma_1^2/\sigma_2^2}$; and sample sizes that are balanced ($n = (5,5)$ and $n = (10,10)$), and unbalanced ($n = (5,10)$ and $n = (10,5)$), always under equality of population means. Independently of the significance level for comparison of means, all F pretests were performed at a fixed 0.05 significance level.

These results agree with those obtained in the previous studies: There is inflation or deflation in the overall TIEP when the decision to use Student's t or Welch's test is conditioned to the result of a pretest (here, the F test) to (supposedly) verify homoscedasticity. This affectation is clearly less concerning in the case of balanced sample sizes as well as with growing sample sizes. However, when there are few observations and/or unbal-

ancing, the affectation increases considerably as the level of heteroscedasticity grows; so once again we verify that performing this type of pretest is a bad strategy.

4.2. Overall TIEP affectation when the equivalence dispersion pretest is used to verify the homoscedasticity assumption

Figure 3 shows comparable simulation results when the Wellek's equivalence pretest is used, and once the zone of indifference (ω_1^2, ω_2^2) has been determined for each significance level (of the comparison of means test) and sample sizes scenario. The δ values correspond to those suggested by Cochran's criterion, with a tolerance limit for the TIEP equal to 20% of the significance level. We observe much greater control of the TIEP (not perfect, but in any case within the irrelevance limits) with values much closer to the significance level than when pretesting was entrusted to the F test. Independently of the comparison of means significance level, all of the equivalence pretests were performed at a fixed 0.05 significance level.

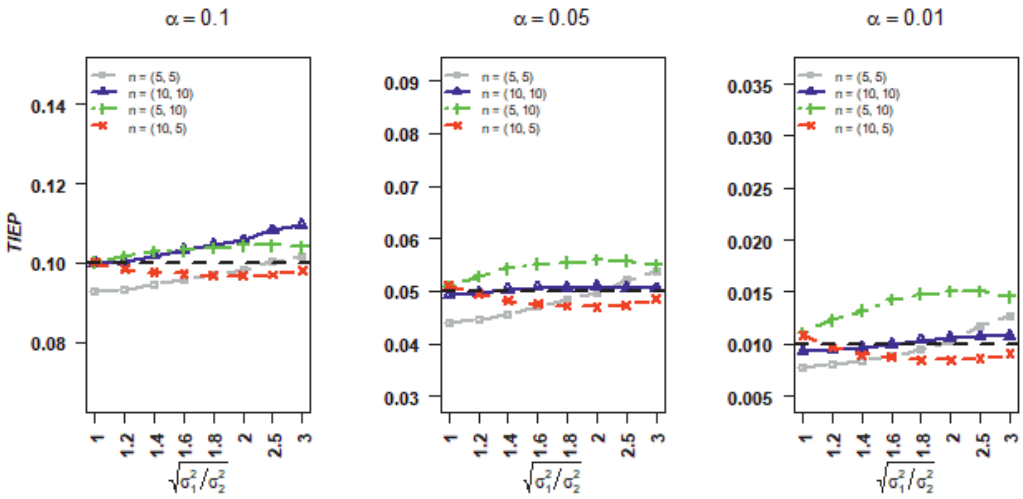

Figure 3: *Overall TIEP estimation when Student's t test or Welch's test are conditioned to the result of the equivalence Wellek's test: If the null hypothesis of relevant ratio of variances is rejected, then Student's t test is applied; otherwise Welch's test is applied.*

4.3. Pretesting vs non-pretesting strategies

Figure 4 shows that, for all sample sizes under consideration, performing Student's t test directly without prior verification of the homoscedasticity assumption greatly inflates or deflates the TIEP as the heteroscedasticity increases. The inflation/deflation of TIEP depends on sample size and especially on balancing/unbalancing; so, for unbalanced

cases, this TIEP's affectation is much greater. For unbalanced cases, the TIEP is below the significance level when the largest sample corresponds to greater variance; whereas, when the smallest sample corresponds to greater variance, the estimated TIEP is above the significance level.

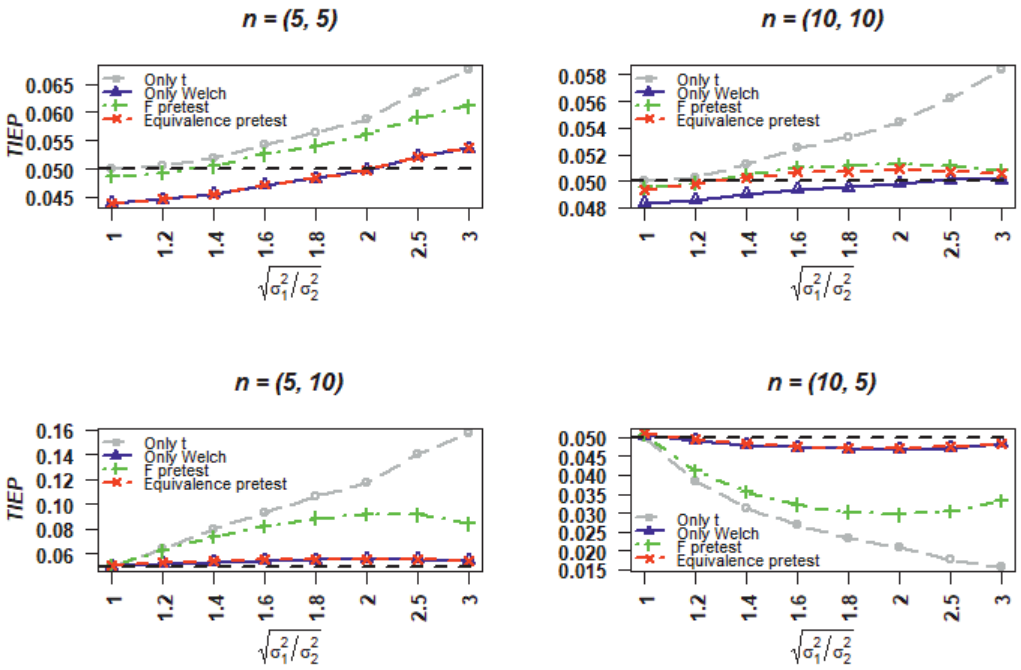

Figure 4: *TIEP affectation using* $\alpha = 0.05$ *and* $\delta = 0.01$ *jointly comparing the pretesting and non-pretesting strategies. Note again that the TIEP scales differ.*

When a traditional pretest such as the F test is used to verify homoscedasticity before deciding on Student's t test or Welch's test should be used, the TIEP also inflates/deflates in the same way as the previous case, although with less intensity, and it becomes less concerning for increasing balanced sample sizes.

Very similar behaviour occurs when Welch's test is used directly, without pretesting, and when pretesting is based on the Wellek's equivalence test. Both strategies are quite stable, with true TIEP values close to the nominal significance level. For low and unbalanced sample sizes, the equivalence test has low power; the null hypothesis stating a disturbing level of heteroscedasticity is rarely rejected; due to there being not enough evidence to prove a non-disturbing level of heteroscedasticity, the cautionary approach of using Welch's test is taken (which seems a more reliable strategy than assuming homoscedasticity based on being unable to prove heteroscedasticity by means of Student's test).

The strategy of using exclusively Welch's test and the strategy based on Wellek's pretest only slightly differ for large and preferably balanced sample sizes. Then, as

the heteroscedasticity irrelevance test reaches enough power, more often there is some evidence to think on a non-disturbing heteroscedasticity and to use Student's t test instead of Welch's test. Although both strategies are tenable (while the other two should be advised against), it is difficult to say which strategy is best. Equivalence pretesting translates into a less conservative strategy, but both have true TIEP values that are very close to the nominal significance level, i.e., always within the $\alpha \pm \delta$ limits.

5. Illustrative examples

To illustrate these methods, we will use two datasets available at the website of the University of Sheffield. The data files and the *R* scripts with the functions implementing the methods described above are available on request to the authors. In all these examples, tests were performed at a nominal significance level of 0.05 and irrelevance in the TIEP distortion was fixed in a 20% level and, therefore, the Cochran's criterion was applied.

The first dataset is available at: https://www.sheffield.ac.uk/mash/statistics2/data.

These data are part of a study trying to relate margarine (or more precisely, its active ingredient, stanol ester) as part of a low fat, low cholesterol diet, with the reduction on cholesterol levels. Here we compare this response on 18 subjects, which are assigned in a balanced way to two margarine types, A and B.

From the algorithm described in Section 3, and provided that both sample sizes (A and B) are $n_1 = n_2 = 9$, all values of the true ratio of variances ranging from 0.1076 to 9.2928 are acceptable to keep the true TIEP of Student's t test inside the limits 0.05 ± 0.01. We feed these equivalence limits (0.1076 and 9.2928) into Wellek's algorithm to determine the critical region of the equivalence test (Section 2). The resulting critical region is $0.3420 < Q < 2.9236$. Provided that the sample variances are 1.7090 and 0.6820, and thus the resulting test statistic is $Q = 2.5059$, then the null hypothesis stating the existence of a relevant heteroscedasticity is rejected. Therefore, applying Student's t test may be considered acceptable. Its resulting p-value is 0.2771 and, therefore, it is impossible to reject the null hypothesis of means equality. Under the "always Welch - no pretests" strategy the resulting p-value is very similar, 0.2801, obviously with the same conclusion.

The second dataset considered here is available at: https://www.sheffield.ac.uk/polopoly_fs/1.570199!/file/stcp-Rdataset-Diet.csv.

These data correspond to a study relating loss in body weight with three diets. We will consider only two groups: diets 1 and 3, and the loss in body weight after 6 weeks of treatment will be used as the observed variable. The respective sample sizes are higher than in the previous example and they are unbalanced: $n_1 = 24$ and $n_3 = 27$. Given these sample sizes and the previously fixed tolerance in the TIEP, 0.05 ± 0.01, the resulting equivalence or heteroscedasticity irrelevance limits are 0.0008 and 3.5068. These equivalence limits conduct to the critical region of the equivalence test defined

by $0.0015 < Q < 1.7900$. For the sample variances 5.0183 and 5.7387, for diet 1 and diet 3, respectively, and then for the ratio $Q = 0.8744$, the hypothesis of a relevant heteroscedasticity is rejected. Consequently, applying Student's *t* test may be considered acceptable. It provides a p-value of 0.0066, which conducts to the rejection of the null hypothesis of equality of means in favor of the two-sided alternative of difference. Again, Welch's test would come to the same conclusion, with a 0.0065 p-value.

Additional examples are available in the *R* scripts mentioned at the beginning of this section.

6. Conclusions and discussion

This paper reinforces the arguments against traditional pretests, such as the *F* test (or Levene's, Bartlett's, Cochran's, etc.) for testing the homoscedasticity assumption prior to Student's *t* test for comparison of means. It seems to support the categorical statement of Rasch et al. (2011) that directly advises against using Student's *t* test and instead promotes making routine use of Welch's test. Our results only qualify this conclusion slightly. Since there is only a small difference between directly applying Welch's test without any previous homoscedasticity verification and pretesting by means of an equivalence/irrelevance dispersion test, and because also both strategies seem to be reliable, it is difficult to recommend any one of them over the others. In any case, the decision should be made on the basis of balancing what is preferable: on the one hand, we have an always small difference in type I error control, which is slightly less conservative in equivalence pretesting; and, on the other, we have the opposite situation when applying only Welch's test without pretesting – which in any case is a simpler procedure.

When choosing between an equivalence pretesting approach or a more robust test against the failure to fulfil validity conditions, all doubts will disappear in situations lacking this second option. For example, it is our opinion that generalizing to more than 2 groups in Welch's test (Welch, 1951) leads to poor control of the TIEP. This could spark interest in continuing this study by expanding it to more general situations. An obvious first step would be to study the suitability of Wellek's test for heteroscedasticity irrelevance for more than two groups (Wellek, 2010, p. 227) as a pretest for the one-way ANOVA.

Acknowledgements

This research is partially supported by Grant MTM2015-64465-C2-1-R (MINECO/FEDER) from the Ministerio de Economía y Competitividad (Spain) and by grant 2014 SGR 464, Generalitat de Catalunya.

The authors are also very grateful to all three referees of this paper, for their very valuable and constructive comments.

A. Appendix: A variance reduction technique when the simulation output is a proportion

In the simulations described in this paper, the parameter to be estimated was a probability. Vegas and Ocaña (1992) and Ocaña and Vegas (1995) developed a simulation variance reduction technique based on the "control variates" method, specifically devoted to this situation. To implement control variates, the simulation output of each simulation replicate, say Y, (here it is an "indicator" variable: 1 if in the end the null hypothesis of equality of means has been rejected, 0 otherwise) should be paired with a correlated "control variate", say C, with known expectation, $E(C)$. In the present study, C was the outcome of Student's t test under the same simulated data but adapted to come from a perfect homoscedasticity scenario, with known $E(C) = \alpha$. In fact, the generation process was the inverse. First, a scenario of perfect homoscedasticity was simulated to obtain C; then, these (homoscedastic) simulated values were subsequently transformed to represent each desired degree of heteroscedasticity in order to obtain Y.

Assume that, after performing m simulation replicates, the simulation output (absolute frequencies) and the associated probabilities (here with $p_{.1} = \alpha$) can be summarized as shown in the following table:

	$C=0$	$C=1$	
$Y=0$	m_{00}	m_{01}	$m_{0.}$
$Y=1$	m_{10}	m_{11}	$m_{1.}$
	$m_{.0}$	$m_{.1}$	m

	$C=0$	$C=1$	
$Y=0$	p_{00}	p_{01}	$p_{0.}$
$Y=1$	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	1

Ocaña and Vegas (1995) showed that

$$\tilde{p}_{1.} = p_{.0}\tilde{p}_{10} + p_{.1}\tilde{p}_{11} = p_{.0}\frac{m_{10}}{m_{00}+m_{10}} + p_{.1}\frac{m_{11}}{m_{01}+m_{11}}$$

is an unbiased estimator of $p_{1.}$, which is more efficient than the raw relative frequency, $m_{1.}/m$. Its variance can be estimated by means of:

$$\tilde{\sigma}^2_{\tilde{p}_{1.}} = \frac{\tilde{p}_{00}\tilde{p}_{10}}{np_{.0}+p_{.0}-2} + \frac{\tilde{p}_{01}\tilde{p}_{11}}{np_{.1}+p_{.1}-2}.$$

References

Altman, D.G. and J.M. Bland (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311, 485.

Box, G.E. (1979). Robustness in the strategy of scientific model building. *Robustness in Statistics*, 1, 201–236.

Cochran, W.G. (1942). The χ^2 correction for continuity. *Iowa State College Journal of Science*, 16, 421–436.

Hsu, P. (1938). Contribution to the theory of "student's" t-test as applied to the problem of two samples. *Statistical Research Memoirs*.

Ocaña, J. and E. Vegas (1995). Variance reduction for bernoulli response variables in simulation. *Computational Statistics and Data Analysis*, 19, 631–640.

Overall, J.E., R.S. Atlas and J.M. Gibson (1995). Tests that are robust against variance heterogeneity in $k \times 2$ designs with unequal cell frequencies. *Psychological Reports*, 76, 1011–1017.

Rasch, D. and V. Guiard (2004). The robustness of parametric statistical methods. *Psychology Science*, 46, 175–208.

Rasch, D., K.D. Kubinger and K. Moder (2011). The two-sample t test: pre-testing its assumptions does not pay off. *Statistical Papers*, 52, 219–231.

Scheffé, H. (1970). Practical solutions of the behrens-fisher problem. *Journal of the American Statistical Association*, 65, 1501–1508.

Vegas, E. and J. Ocaña (1992). Variance reduction for bernouilli response variables. In *Computational Statistics*, pp. 103–107. Springer.

Welch, B. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, 38, 330–336.

Welch, B.L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika*, 34, 28–35.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. USA: CRC Press.

Zimmerman, D.W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57, 173–81.