



OPEN ACCESS



Operations Research and Decisions

www.ord.pwr.edu.pl

OPERATIONS
RESEARCH
AND DECISIONS
QUARTERLY



Goodness and lack of fit tests to pretest normality when comparing means

Pablo Flores^{1*}  María de Lourdes Palacios² 

¹Grupo de Investigación en Ciencia de Datos CISED, Facultad de Ciencias, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador

²Carrera de Matemáticas, Facultad de Ciencias, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador

*Corresponding author, email address: p_flores@esPOCH.edu.ec

Abstract

Previous studies show that processes related to traditional pretests to prove the perfect fulfillment of assumptions in comparison means tests lead to severe alterations in the overall Type I error probability and power. These problems seem to be overcome when pretests based on an equivalence approach are used. The paper proposes a lack of fit tests based on equivalence to pretest normality on homoscedastic samples with measurable departures from normality. The Type I error probability and power produced by this equivalence pretest are compared with two traditional goodness of fit pretests and with the direct use of the t-Student and Wilcoxon test of means comparison. Furthermore, since the irrelevance limit for the lack of fit test is an arbitrary value, we propose a non-subjective methodology to find it. Results show that this proposed equivalence test controls the overall Type I Error Probability and produces adequate power; therefore, its use is recommended.

Keywords: normality, lack of fit, goodness of fit, equivalence, assumptions, Type I error probability

1. Introduction

Traditionally, the t-Student [22] or the Welch test [24] (if homoscedasticity is not proved) to find significant differences between means is conditioned to prove normality assumption using goodness of fit pretests such as Pearson's chi-square [16] or Shapiro–Wilk [21]. When normality cannot be established, the null hypothesis of perfect normality is rejected, and a non-parametric test (Wilcoxon–Mann Whitney test [14, 26]) to compare means is used. Although these procedures are common, several studies [12, 20] show that pretesting normality and homoscedasticity assumptions alter the Type I error probability (TIEP), so performing this procedure could lead to severe mistakes in the inference process.

Zimmerman [27] proved that, especially for low and unbalanced sample sizes, the overall TIEP of a means comparison test is highly inflated when homoscedasticity is pretested. This inflation is inversely

proportional to the significance level but TIEP begins to deflate from the non-practical $\alpha = 0.20$ value; the author recommends not pretesting homoscedasticity, instead using Welch's test directly (without pretesting) seems to be a good idea since it keeps the TIEP close to the significance level.

Rasch et al. [18] pretested normality and homoscedasticity assumptions simultaneously for the two-sample t-test, concluding that this procedure "does not pay off", for it carries unknown risks in the overall Types I and II error probabilities, instead using Welch's test without pretesting is a suitable procedure since it produces acceptable TIEP and power. Authors recommend using the Welch test as a standard option to be implemented in books and statistical software, avoiding using Wilcoxon's and t-Student's tests.

The traditional pretests establish in their null hypothesis the perfect fulfillment of the assumption (i.e., perfect normality or homoscedasticity). Regarding this, we would like to mention George Box's: *the statistician knows, for example, that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false, he can often derive results which match, to a useful approximation, those found in the real world* [4]. So, there will always be uncertainty about whether it is appropriate or not to prove perfect normality (or homoscedasticity). In addition, the following inconsistency seems to reinforce the theory that the traditional approach is not adequate to test the assumptions described. Note that the fulfillment of assumptions is always in the null hypothesis, and as we know, it is not possible to make a decision related to H_0 , for example, the phrase "we accept H_0 " is statistical nonsense, so when the null hypothesis is not rejected, we only can conclude that there is no evidence to prove non-normality (or heteroscedasticity), which is no necessary evidence to prove assumption fulfillment, or as argue by Altman and Bland: [1] *Absence of evidence is not evidence of absence*. It follows that this kind of test is only used to prove significant differences and not equalities. Henceforth for practical purposes, we will call these inconsistencies mentioned in this paragraph logical difficulties.

Likely, an equivalence approach instead of the traditional one explained above solves these logical difficulties, understanding equivalence as equality except for an irrelevant deviation given by a threshold or equivalence/irrelevance limit [25]. In a previous study, it was shown that when an equivalence approach (specifically the dispersion test of two Gaussian distributions) is used to pretest homoscedasticity when testing means difference, the results are as well as those obtained for the Welch test directly applied without pretest, for the TIEP is adequately controlled around the significance level [9]. In this previous study, using appropriate irrelevance limits (ω_1^2, ω_2^2), the equivalence hypothesis test $H_0 : \frac{\sigma_1^2}{\sigma_2^2} \leq \omega_1^2 \vee \frac{\sigma_1^2}{\sigma_2^2} \geq \omega_2^2$ vs. $H_1 : \omega_1^2 < \frac{\sigma_1^2}{\sigma_2^2} < \omega_2^2$ (irrelevant homoscedasticity) instead $H_0^* : \frac{\sigma_1^2}{\sigma_2^2} = 0$ (perfect homoscedasticity) vs. $H_1^* : \frac{\sigma_1^2}{\sigma_2^2} \neq 0$ was used. In this case, if null hypothesis H_0 is rejected, it is evidence to accept alternative hypothesis H_1 and conclude irrelevant (not perfect) homoscedasticity, so when the objective is not to prove significant differences but to find a certain similarity between compared parameters, an equivalence approach is better than the traditional one. Some researchers have oriented their results by preferring equivalence instead of traditional tests [11, 19].

Regarding the effectiveness of hypothesis tests, another point to consider is the robustness and sensitivity of the means comparison tests against alterations or contamination of normality and homoscedasticity

assumptions since this could also influence the TIEP alterations above mentioned. Referring to normality, studies prove that t-Student and Welch's test remain robust against normality contamination. In contrast, its non-parametric alternative (the Wilcoxon test) is sensitive against these same deviations [17, 23]. On the other hand, the t-Student is a very sensitive test against homoscedasticity deviations, while the Welch test seems to remain robust against these same contaminations [10, 15].

As a continuation of the work shown in the above paragraph [9], from normal and non-normal homoscedastic samples, we propose an equivalence approach (lack of fit tests) to pretest normality when testing means difference; this process is compared with two traditional goodness of fit pretests in terms of their TIEP and their Power. We have chosen the Chi-Square traditional pretest since, theoretically, it is easy to contrast it with the lack of equivalence pretest, and the Shapiro–Wilk pretest was selected since it is a suitable alternative to prove normality when there are deviations of this distribution [8]. Section 2 introduces some essential concepts about lack and goodness of fit tests. Section 3 describes the algorithm built to determine a non-arbitrary irrelevance limit in the lack of fit test. Section 4 describes the simulation methodology to achieve the proposed goal, Section 5 shows the obtained results, and finally, in Section 6, the main conclusions are discussed.

2. Goodness and lack of fit tests for normality

2.1. Goodness of fit tests

Both goodness-of-fit tests (Chi-square and Shapiro–Wilk) used in this manuscript to prove normality have the same statistical logic. Still, with the aim that the reader can relate and compare a goodness-of-fit test with a lack of fit, in this section, we will exclusively develop the Chi-square test.

Given a sample from a continuous variable X grouped in k intervals, its true distribution can be represented by the probability vector $\pi = (\pi_1, \pi_2, \dots, \pi_k)$, while its reference distribution to which the true one is asserted to fit (target distribution) is given by the theoretical probabilities $\pi^0 = (\pi_1^0, \pi_2^0, \dots, \pi_k^0)$ obtained from a normal distribution. Using this notation, a goodness of fit test to prove normality is represented by equation

$$\begin{aligned} H_0^* : \pi &= \pi^0 && \text{perfect normality} \\ H_1^* : \pi &\neq \pi^0 && \text{non-normality} \end{aligned} \tag{1}$$

Specifically, the chi-square goodness of fit test, establishes the statistic

$$X^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

where o_i are the observed frequencies obtained from the sample. Note that, relative frequencies (estimation of probabilities π_i) are obtained by $\hat{\pi}_i = \frac{o_i}{n}$. e_i are the expected frequencies obtained from normal

probabilities π_i . n is the sample size $n = \sum_{i=1}^k o_i = n \sum_{i=1}^k \hat{\pi}_i$.

X^2 follows a χ^2 distribution with $k - c - 1$ degrees of freedom, c is the number of parameters for the target distribution. In the case of normal, it has two parameters ($c = 2$) but if parameters μ, σ^2 are known, degrees of freedom for chi-square is reduced to $k - 1$, thus $X^2 \sim \chi_{k-1}^2$ d.f.

As mentioned in Section 1, the goodness of fit test proposed in equation 1 seems to be inappropriate for testing normality. Not rejecting H_0^* is not proof of the assumption fulfillment; it only proves that there is no evidence to conclude that distribution is different from a normal, which does not necessarily implies normality. On the other hand, rejecting H_0^* might simply indicate an irrelevant departure from perfect normality. Precisely the same thing happens if we consider the Shapiro–Wilk test.

2.2. Lack of fit test

This test is based on an equivalence approach where the null hypothesis H_0 states a lack of fit instead of goodness of fit, while the alternative hypothesis H_1 states equivalence, i.e., perfect fit to normal except irrelevant deviations [25, p. 265]. In brief, this test may be described as follows:

Using the square Euclidean distance $d^2(\pi, \pi^0) = \sum_{i=1}^k (\pi_i - \pi_i^0)^2$ as a measure of the dissimilarity degree between the true and target distribution, for the hypothesis:

$$\begin{aligned} H_0 : d^2(\pi, \pi^0) &\geq \epsilon^2 && \text{lack of fit to normality} \\ H_1 : d^2(\pi, \pi^0) &< \epsilon^2 && \text{irrelevant normality} \end{aligned} \quad (2)$$

By Theorem 14.3-4 of the book *Discrete Multivariate Analysis. Theory and Practice*" [2, p. 470], we know that normalized vector $\sqrt{n}(\hat{\pi}_1 - \pi_1, \dots, \hat{\pi}_k - \pi_k)$ of relative frequencies converges asymptotically in law to a random variable with a k -dimensional normal distribution, with mean 0 and covariance matrix

$$\Sigma = D_\pi - \pi' \pi = \begin{pmatrix} \pi_1 & 0 & \dots & 0 \\ 0 & \pi_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \pi_k \end{pmatrix} = \begin{pmatrix} \pi_1^2 & \dots & \pi_1 \pi_k \\ \pi_2 \pi_1 & \dots & \pi_2 \pi_k \\ \vdots & & \vdots \\ \pi_k \pi_1 & \dots & \pi_k^2 \end{pmatrix}$$

Given that $\pi \mapsto d^2(\pi, \pi^0)$ is differentiable in the parameter space Π with gradient vector $\nabla d^2(\pi, \pi^0) = 2(\pi, \pi^0)$, it implies that we can use the δ -method to find the asymptotic distribution of the estimated square Euclidean distance $d^2(\hat{\pi}, \pi^0)$ [6].

Thus based on

$$\begin{aligned} \sigma_a^2[\sqrt{n}d^2(\hat{\pi}, \pi^0)] &= 4(\pi - \pi^0)(D_\pi - \pi' \pi)(\pi - \pi^0)' \\ &= 4 \left[\sum_{i=1}^k (\pi_i - \pi_i^0)^2 \pi_i - \sum_{i_1=1}^k \sum_{i_2=1}^k (\pi_{i_1} - \pi_{i_1}^0)(\pi_{i_2} - \pi_{i_2}^0) \pi_{i_1} \pi_{i_2} \right] \end{aligned}$$

we may conclude that

$$\sqrt{n}(d^2(\hat{\pi}, \pi^0) - d^2(\pi, \pi^0)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_a^2[\sqrt{n}d^2(\hat{\pi}, \pi^0)]) \text{ as } n \rightarrow \infty$$

Since $\hat{\pi}_i$ is a consistent estimator for π_i ($i = 1, \dots, k$), the asymptotic variance is consistently estimated by replacing each probability π_i with the homologous relative frequency $\hat{\pi}_i$. Denoting this estimated variance by $v_n^2(\hat{\pi}, \pi^0)$, we have:

$$v_n^2(\hat{\pi}, \pi^0) = 4 \left[\sum_{i=1}^k (\hat{\pi}_i - \pi_i^0)^2 \hat{\pi}_i - \sum_{i_1=1}^k \sum_{i_2=1}^k (\hat{\pi}_{i_1} - \pi_{i_1}^0)(\hat{\pi}_{i_2} - \pi_{i_2}^0) \hat{\pi}_{i_1} \hat{\pi}_{i_2} \right]$$

Given this consistency of $v_n^2(\hat{\pi}, \pi^0)$ for $\sigma_a^2[\sqrt{n}d^2(\hat{\pi}, \pi^0)]$, we can conclude that $d^2(\hat{\pi}, \pi^0)$ follows a normal distribution with known variance $\frac{v_n^2(\hat{\pi}, \pi^0)}{n}$ and unknown expected value θ . Thus:

$$d^2(\hat{\pi}, \pi^0) \xrightarrow{\mathcal{L}} \mathcal{N} \left(\theta, \frac{v_n^2(\hat{\pi}, \pi^0)}{n} \right) \tag{3}$$

From these results, we can obtain the upper limit bound for a confidence interval of $d^2(\hat{\pi}, \pi^0)$ as $l_{upp} = d^2(\hat{\pi}, \pi^0) + Z_\alpha v_n^2(\hat{\pi}, \pi^0)$ with $Z_\alpha = \Phi^{-1}$ the correspondence quantile for a normal distribution using a confidence level α .

Finally, we can take the following decision in the hypothesis test of equation (2). Reject null hypothesis H_0 and conclude normality if $l_{upp} < \epsilon^2$. Results for this equivalence test to detect normality can be computed with the `normequiv` function from `equivNorm` R-Package available as supplementary material.

3. Irrelevance limits for the lack of fit test

In the lack of fit test proposed in equation (2), ϵ^2 is a fundamental value to detect normality; however, there are no technical criteria to establish this irrelevance limit adequately. The author of this approach says that this value can be obtained based on the researcher’s experience or using common statistical sense. To find a non-subjective or non-arbitrary irrelevance limit, we propose an algorithm where samples far from normality are simulated using the Fleishman method [7], which is explained as follows: From a normal standard $Z \sim \mathcal{N}(\mu = 0, \sigma = 1, \gamma_1 = 0, \gamma_2 = 0)$, where μ is the mean, σ is the standard deviation, γ_1 is the skewness and γ_2 is the kurtosis. Fleishman method contaminates this normality by altering its third γ_1 and fourth γ_2 moments through the linear combination $Y = a + bZ + cZ^2 + dZ^3$, for instance, $(a = 0, b = 1, c = 0, d = 0)$ returns again the same perfect standard normal Z , and this becomes contaminated as the coefficients (a, c, d) and b move away from zero and one respectively; so, variable $X = \mu + \sigma Y$ follows an unknown distribution with parameters $(\mu, \sigma, \gamma_1, \gamma_2)$. `fleishman.coef` function from `BinNonNor` R-Package [13] computes the Fleishman coeficientes (a, b, c, d) from γ_1, γ_2 values given by the user. In addition, `rnorm` function from `equivNorm` R-Package available as supplementary material allows to one obtain random non-normal numbers with departures from normality depending on Fleishman coefficients.

Besides, as discussed in Section 1, there is no perfect normality, but some models known to be false often produce useful approximate results, so what matters is not if populations are normal but if the approximation of a model is good enough to be useful. In this work, the approximation will be considered good or not based on how close the TIEP is to the significance level α . In this regard, Cochran suggested that a distance of 20% of the true TIEP from the nominal significance level α is an acceptable approxi-

mation [5], i.e., the TIEP should be within the interval $\alpha \pm 0.2\alpha$. This authoritative criterion known as Cochran's criterion, is the default in algorithms implementing the proposed method in this manuscript.

From these considerations, here is the procedure to obtain irrelevance limit ϵ building an iterative algorithm of simulation: With Fleishman's method, we simulate $nSim$ pairs of independent non-normal samples (from the lowest possible contamination) of size n_1, n_2 with the same theoretical mean μ and variance σ^2 ; these samples are tested with a t-Student to obtain an estimate of the TIEP, which is computed as the proportion of null hypothesis ($H_0 : \mu_1 - \mu_2 = 0$) rejections when it is true. Next, a confidence interval for this TIEP is compared with the interval suggested by the Cochran criteria to consider a model acceptable. This process will be repeated iteratively until the confidence interval for the TIEP falls outside the Cochran criteria or until the highest possible normal contamination is reached to obtain the smallest upper bound above which one can be sure to declare irrelevant normality in really normal samples. Finally, the square root of this upper limit (which grows as the normal is contaminated) is used as the epsilon value we seek.

`epsilon` function from `equivNorm` R-Package available as supplementary material computes this irrelevance limit ϵ for a given number of simulations, sample sizes n_1, n_2 , and confidence level α . Table 1 shows the results of ϵ values found with `epsilon` using $nSim = 100,000$ simulation replicates, different n_1, n_2 and α . We can observe that large and unbalanced samples have higher irrelevance limits, i.e., a wider irrelevance. In addition, with some exceptions, it seems that the irrelevance limit increases as the level of significance increases.

Table 1. Irrelevance limits for a lack of fit test

n_1, n_2	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5, 5	0.4120446	0.4915836	0.5289253
3, 7	0.4739120	0.5424760	0.5289253
7, 3	0.4739120	0.6113572	0.5289253
10, 10	0.3497503	0.4210272	0.4670088
6, 14	0.2737383	0.4924864	0.4670088
14, 6	0.2737383	0.4210272	0.4670088
5, 10	0.3099487	0.4592030	0.4399581
10, 5	0.2775543	0.4592030	0.4399581

4. Simulation methodology

Using samples from normal and unknown non-normal distributions (different measurable departures with Fleishman), overall TIEP and power for means comparison tests are estimated as rejections proportion of null hypothesis of means equality in five cases:

- 1) applying directly (without pretest) t-Student,
- 2) applying its non-parametric alternative directly, i.e., Wilcoxon,
- 3) pretesting normality with chi-square goodness of fit test before deciding t-Student or Wilcoxon as the best alternative,
- 4) the same last process but using the Shapiro–Wilk goodness of fit test,
- 5) pretesting normality by equivalence/lack of fit approach to decide t-Student or Wilcoxon as the best alternative.

`rejectH0` function from `equivNorm` R-package computes this TIEP estimation depending on the simulation replicates number, sample size n_1, n_2 , irrelevance limit ϵ , theoretical means difference, Fleishman coefficients, and significance level α given by the user. Note that when the theoretical means difference is zero, the function computes rejections proportions of $H_0 : \mu_1 - \mu_2 = 0$ when true, i.e., TIEP. In contrast, if theoretical means is different from zero function, computes rejections proportions of $H_0 : \mu_1 - \mu_2 = 0$ when it is false, i.e., power.

Table 2. Skewness, kurtosis and Fleishman's coefficients for different degrees of normality contamination

Contamination degree	Skewness γ_1	Kurtosis γ_2	Fleishman coefficients			
			a	b	c	d
No contamination (0)	0	0	0	1	0	0
Low (1)	0.25	0.70	-0.0368	0.9334	0.0368	0.0213
Mild (2)	0.75	1	-0.1191	0.9559	0.1191	0.0098
High (3)	1.3	2	-0.2491	0.9843	0.2491	-0.0164
Severe (4)	2	6	-0.3137	0.8263	0.3137	0.0227
Extreme (5)	3	15	-0.3457	0.5883	0.3457	0.0861

Skewness γ_1 and kurtosis γ_2 defining the departure levels from normality were chosen based on a study [3] which shows that in terms of absolute value, the cut-off of skewness and kurtosis are between 0.25 to 0.75 for low contamination, between 0.75 to 1.25 for mild contamination, between 1.25 to 1.75 for high contamination, between 1.75 to 2.25 for very high contamination, and although it is not very common greater than 2.25 for severe and extreme contamination. Fleishman coefficients for these γ_1 and γ_2 values were computed by `BinNonNor` R-package, and these results are shown in Table 2. In addition, density plots for the different departure from normality are in Figure 1; this graphic shows the skewness and kurtosis alterations concerning normal density.

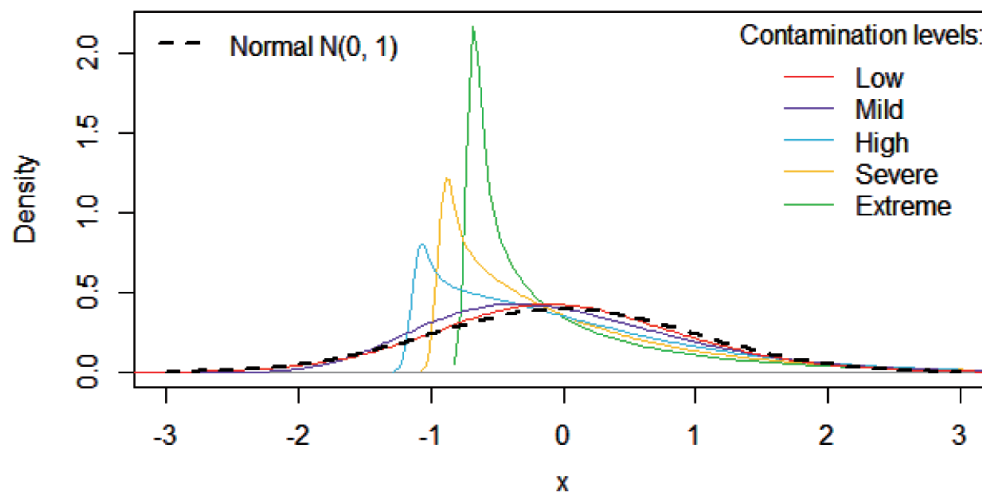


Figure 1. Density curve corresponding to different departures from normality, compared with the non-contaminated, perfect, normal distribution

5. Results

Figure 2 shows the Type I error probability (TIEP) estimation for five different contamination degrees given in Table 2 using a significance level of $\alpha = 0.05$, and the irrelevance limits stated in Table 1 for the

equivalence pretest. Applying Wilcoxon directly without a pretest is the worst option of all since its TIEP is outside the Cochran criterion represented by the shaded area (in this case, for $\alpha = 0.05$, the Cochran criterion is $\alpha \pm 0.2\alpha = [0.04, 0.06]$). The other four options seem to be adequate since they remain inside this Cochran acceptance interval, especially for large sample sizes and low normal contamination. To apply directly t-Student is the best option since its TIEP is closer to the significance level than the other options. This reinforces the conclusions of works presented in Section 1 about the robustness of t-Student facing non-normality. On the contrary, the poor results in the Wilcoxon test confirm the sensitivity of this test against departures from normality.

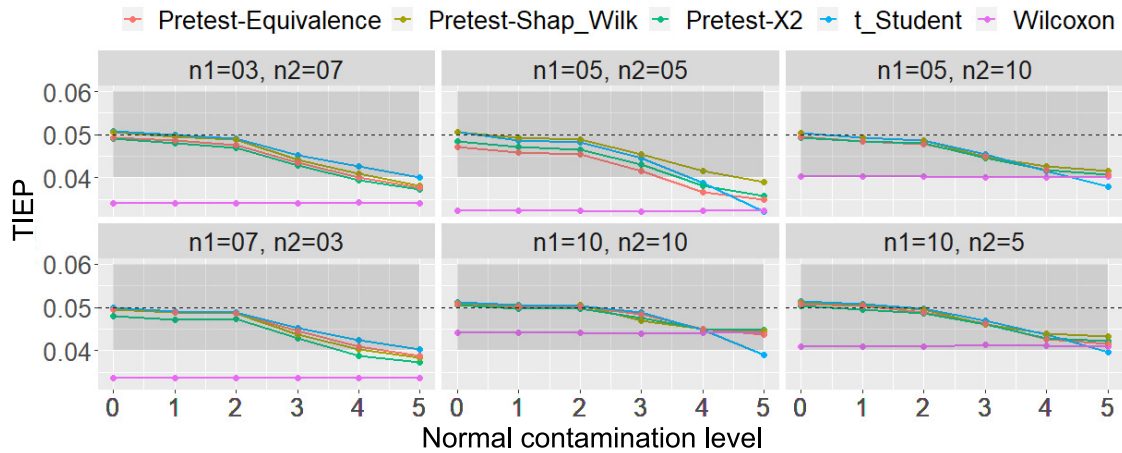


Figure 2. Overall TIEP estimation when t-Student and Wilcoxon test are applied directly and when they are used depending on the result of chi-square, Shapiro–Wilk and equivalence pretests to detect normality; $\alpha = 0.05$

Regarding pretesting strategies, both approaches (traditional and equivalence) generate a very similar TIEP, the three pretests (equivalence, Chi-square, and Shapiro–Wilk) produce a TIEP inside the Cochran criteria, but with slight exceptions for extreme normal contamination, which we know is not very common in the real world. These strategies have similar behavior for the different sample sizes and contamination degrees.

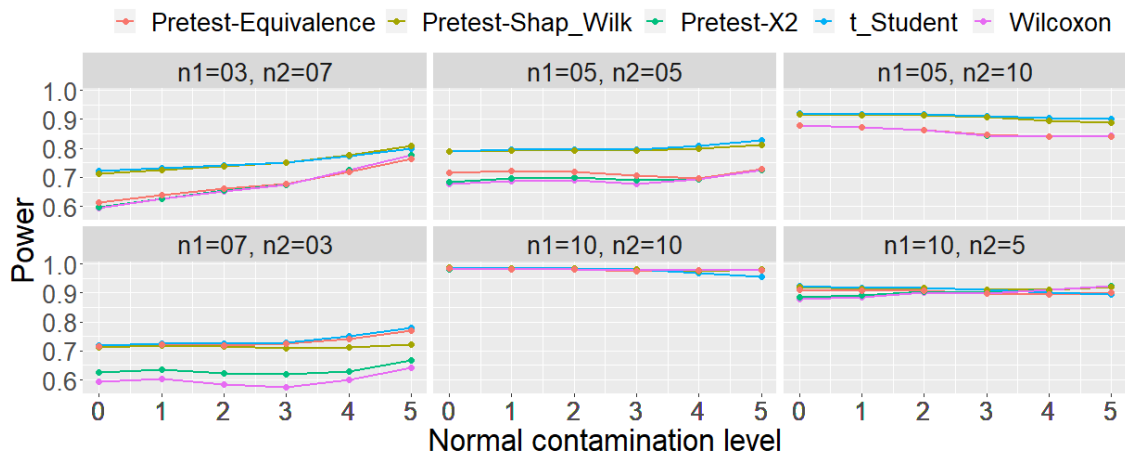


Figure 3. Overall Power estimation when t-Student and Wilcoxon test are applied directly and when they are used depending on the result of chi-square, Shapiro–Wilk, and equivalence pretests to detect normality; $\alpha = 0.05$

Figure 3 shows high estimated power values for large and balanced sample sizes; normality contamination does not influence much except for the smallest and unbalanced samples. If we compare only the pretesting processes, the equivalence approach (lack of fit test) produces higher power than the chi-square

(goodness of fit test). Still, in some cases, the Shapiro–Wilk pretest produces the highest power. On the other hand, applying the t-Student test directly leads to the highest power of all, and using the Wilcoxon test shows the lowest power of all, which is principally evident for low sample sizes.

6. Conclusions and discussion

Analyzing the lack of fit test independently, it is concluded that pretesting the normality assumption under this equivalence approach for a mean difference test is a good alternative since the overall TIEP remains stable around the significance level (inside the Cochran criterion) and the power is relatively large. Reminding that this lack of fit test rejects its null hypothesis to prove normality when the upper bound of the confidence interval for $d^2(\pi, \pi^0)$ is less than irrelevance limit ϵ^2 , choosing an inappropriate ϵ could lead to wrong conclusions, for instance, an epsilon smaller than it should be would make the test unable of testing for normality, even for samples from a normal distribution, or conversely, an epsilon larger than it should be would declare any sample as coming from a normal distribution, including samples that do not. So, our proposed methodology in Section 3 to find non-arbitrary irrelevance ϵ value is valuable to obtain good results in this paper.

On the other hand, although we would like to reinforce the works in Section 1 concluding that the pretesting process is the cause of TIEP alterations when differences of means are tested, this paper shows that both pretesting methods (traditional and equivalence approaches) produce acceptable results. Pretesting with the chi-square and Shapiro–Wilk goodness of fit tests appears to have almost the same results as the lack of fit pretest for normality. We do not say that logical difficulties discussed in Sections 1 and 2, such as the misstatement of traditional tests or trying to prove perfect normality are false; in fact, we believe that they exist and that equivalence tests overcome them; however, it seems that these logical difficulties are not the only cause of the TIEP alterations in the pretesting process, but the robustness or sensitivity of comparison means tests influence too in these alterations.

Although the studies presented in Section 1 show that pretesting the assumptions of normality and homoscedasticity alters the TIEP and that performing this procedure could lead to severe errors in the inference process [12, 20], the present study provides a broader perspective of what is happening in the pretesting process, demonstrating that not only this process, including the logical difficulties of using traditional tests, are causing alterations, but also the robustness of the different means comparison tests against deviations of their assumptions have an essential role in controlling the TIEP and obtaining high power. Thus, in those previous studies where the assumptions of normality and homoscedasticity were pretested, it was never taken into account that in addition to the traditional process of pretesting, the sensitivity of the t-Student to deviations from homoscedasticity and the robustness of the Welch test to these same deviations are factors that determined the alterations or control of the TIEP. Furthermore, the study that determines that the equivalence test to prove homoscedasticity controls the TIEP [9] does not take into account that this control could be because, unlike the traditional test, it detects non-homoscedasticity (when it exists) and leads most cases to the Welch test (robust to deviations from homoscedasticity) instead to the t-Student (sensitive to departures from homoscedasticity). In the same way, in the case of normality presented in this manuscript, although the equivalence test proposed here produces good results, this is since, in most cases, irrelevant normality is detected, which leads to using a Student's t-test,

which is robust to deviations from normality, instead of a Wilcoxon test that is sensitive to these same deviations, as shown in Figure 2.

Based on these arguments, we finally recommend that when testing means difference: i) if homoscedasticity can be assumed, one must pretest normality using any of both approaches, although if we focus on the logical difficulties presented in this paper, the equivalence approach would be more recommendable, ii) when normality can be assumed, one must pretest homoscedasticity exclusively with an equivalence approach and iii) when it is necessary to pretest both assumptions (normality and homoscedasticity), one exclusively must use the equivalence approach to pretest both cases. In this last scenario, one might think that for the case of normality, it is possible to replace the traditional approach with the equivalence, since both produce similar results. However, the logical difficulties discussed in this paper should not be left aside, which, as we saw, are overcome with the equivalence approach. In addition, it is well known that, in reality, normality or homoscedasticity cannot be assumed in a data set, and it is necessary to pretest these two assumptions simultaneously and unlikely the two pretests work well if they have different approaches.

Supplementary remark

All the results contained in this paper were computed with informatics functions built by authors in R software. To make these functions available to any user who wants to replicate the results of this paper or wants to take some advantage of them, we have compiled these functions in the `equivNorm` R-package available on GitHub <https://github.com/pablof1988/equivNorm>. Although all the open source, help, examples, and other details about the use of the `equivNorm` package are available on GitHub, a brief guide explaining the overall use of the package is available in Appendix A

Acknowledgement

The authors are grateful to two anonymous reviewers for their valuable comments and suggestions made on the previous draft of this manuscript.

References

- [1] ALTMAN, D. G., AND BLAND, J. M. Statistics notes: Absence of evidence is not evidence of absence. *BMJ* 311, 7003 (1995), 485.
- [2] BISHOP, Y. M. M., FIENBERG, S. E., HOLLAND, P. W. *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, 1977.
- [3] BLANCA, M. J., ARNAU, J., LÓPEZ-MONTIEL, D., BONO, R., AND BENDAYAN, R. Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 9, 2 (2013), 78–84.
- [4] BOX, G. E. P. Science and statistics. *Journal of the American Statistical Association* 71, 356 (1976), 791–799.
- [5] COCHRAN, W. G. The χ^2 correction for continuity. *Iowa State College Journal of Science* 16, 1 (1942), 421–436.
- [6] DOOB, J. L. The limiting distributions of certain statistics. *The Annals of Mathematical Statistics* 6, 3 (1935), 160–169.
- [7] FLEISHMAN, A. I. A method for simulating non-normal distributions. *Psychometrika* 43, 4 (1978), 521–532.
- [8] FLORES MUÑOZ, P., MUÑOZ ESCOBAR, L., AND SÁNCHEZ ACALO, T. Study of the power of test for normality using unknown distributions with different levels of non normality. *Revista Perfiles* 21, 1 (2019), 4–11.
- [9] FLORES, P., AND OCAÑA, J. Heteroscedasticity irrelevance when testing means difference. *SORT-Statistics and Operations Research Transactions* 42, 1 (2018), 59–72.
- [10] FLORES, P., AND OCAÑA, J. Pretesting strategies for homoscedasticity when comparing means their robustness facing non-normality. *Communications in Statistics - Simulation and Computation* 51, 1 (2022), 280–292.
- [11] FLORES, P., SALICRÚ, M., SÁNCHEZ-PLA, A., AND OCAÑA, J. An equivalence test between features lists, based on the Sorensen–Dice index and the joint frequencies of GO term enrichment. *BMC Bioinformatics* 23, 207 (2022), 1–21.
- [12] HSU, P. Contribution to the theory of “student’s” t-test as applied to the problem of two samples. *Statistical Research Memoirs* 2 (1938), 1–24.

- [13] INAN, G., DEMIRTAS, H., AND GAO, R. *BinNonNor: Data Generation with Binary and Continuous Non-Normal Components*, 2021. R package version 1.5.3.
- [14] MANN, H. B., AND WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50–60.
- [15] MONTILLA, J.-M., AND KROMREY, J. Robustness of the t tests in comparison of means, under violation of normality and homoscedasticity assumptions. *Ciencia e Ingeniería* 31, 2 (2010), 101–107, (in Spanish).
- [16] PEARSON, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50, 302 (1900), 157–175.
- [17] RASCH, D., AND GUIARD, V. The robustness of parametric statistical methods. *Psychology Science* 46, 2 (2004), 175–208.
- [18] RASCH, D., KUBINGER, K. D., AND MODER, K. The two-sample t test: pre-testing its assumptions does not pay off. *Statistical Papers* 52, 1 (2011), 219–231.
- [19] SÁNCHEZ-PLA, A., SALICRÚ, M., AND OCAÑA, J. An equivalence approach to the integrative analysis of feature lists. *BMC Bioinformatics* 20, (2019), 441.
- [20] SCHEFFÉ, H. Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association* 65, 332 (1970), 1501–1508.
- [21] SHAPIRO, S. S., AND WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3/4 (1965), 591–611.
- [22] STUDENT. The probable error of a mean. *Biometrika* 6, 1 (1908), 1–25.
- [23] SULLIVAN, L. M., AND D’AGOSTINO, R. B. Robustness of the t test applied to data distorted from normality by floor effects. *Journal of Dental Research* 71, 12 (1992), 1938–1943.
- [24] WELCH, B. L. On the comparison of several mean values: an alternative approach. *Biometrika* 38, 3/4 (1951), 330–336.
- [25] WELLEK, S. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press, 2010.
- [26] WILCOXON, F. Individual comparisons by ranking methods. *Biometrics Bulletin* 1, 6 (1945), 80–83.
- [27] ZIMMERMAN, D. W. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology* 57, 1 (2004), 173–181.

A. Appendix

A brief introduction about equivnorm R-Package

This R package implements methods to prove normality (or even another distribution) using an equivalence approach. Simulation functions can be used to measure the effectiveness (in terms of Type 1 error probability and power) of the pretest to prove normality when testing means difference, in addition, the robustness of t-Student and Wilcoxon test can be obtained.

equivNorm must be installed with a working R version ($\geq 4.2.0$). Installation could take a few minutes on a regular desktop or laptop. Package can be installed from devtools package, then it needs to be loaded using `library(equivNorm)`, or from Github using the code:
`devtools::install_github("pablof1988/equivNorm")`.

equivNorm package provides the following functions

- LackFitTest proves if a sample comes from a target distribution through an equivalence approach.
- Normequiv prove if a sample comes from a normal distribution through an equivalence approach.
- Rnonorm generates random numbers from non-normal distributions using Fleishman’s coefficients.
- Epsilon computes a non-arbitrary epsilon value of irrelevance for an equivalence normality test of lack of fit.
- RejectH0 is a simulation function to estimate the Type I error probability and power of hypothesis tests.
- Both Tests and tiepT are complementary simulation functions whose main purpose is to help to the rejectH0 function.