Abstracts of the Ninth Workshop on Simulation (2018) 1-2

Pretesting Assumptions for the validity of two sample Mean Tests

Pablo Flores M. [1] Jordi Ocaña [2] Tania Sánchez A. [3]

The Student's t-test [12] to prove possible inequalities of two population means is subject to normality and homoscedasticity assumptions which traditionally are verified through pretests. This strategy is sometimes recommended in statistical books and courses. When in a normality pretest (Shapiro Wilk, Kolmogorov Smirnov, Anderson darling, etc.) the null hypothesis of perfect normality is rejected, a non-parametric test (Wilcoxon - Mann - Whitney test) is usually used ([15], [6]) -sometimes wrongly assuming its robustness in front of heteroscedasticity. Otherwise, normality is assumed and a parametric procedure like the t-test or the Welch's test [13] is used. Which of these procedures is chosen depends on an additional pretest to compare variances (F, Levene, Bartlet ...). If the null hypothesis of perfect variances equality is not rejected, homoscedasticity is assumed and the t-test is used, otherwise the final decision is based on the Welch's test.

There are several studies ([5], [8], [11], [16], [10]) showing that pretesting alters the overall Type I Error Probability (TIEP). These authors agree in advising against pretesting, and sometimes recommending the direct use (without pretest) of the Welch's test as the standard option, since it keeps the overall TIEP stable around the nominal significance level α.

These drawbacks associated to pretesting may be due to pretesting itself, but they also may be due to the way pretesting is performed. In fact, the above mentioned pretests are inadequate for reaching its goal. Rejecting its null hypothesis may indicate an irrelevant departure from normality or homoscedasticity, and not rejecting it is not a proof of the corresponding assumption. The problem may be not pretesting but the way it is done. An equivalence approach, e.g. [14], may be more appropriate to solve the problem. In equivalence pretests the alternative hypothesis states perfect homoscedasticity or fit to normal, except for irrelevant deviations, while relevant deviations are stated in the null hypothesis, so the assumptions are assumed if it is rejected. In equivalence testing, to state the adequate irrelevance limits is very important. [4] established a numerical algorithm that calculates these limits as a function of sample size and significance

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

[2] Departament de Genética, Microbiologia i Estadística, Secció d' Estadística, Universitat de Barcelona, Facultat de Biologia, Diagonal 643, 08028 Barcelona, Spain, E-mail: `jocana@ub.edu`

[3] Escuela Superior Politécnica de Chimborazo, Facultad de Ciencias, Panamericana Sur km 1 1/2, EC060155 Riobamba, Ecuador

level α for the homoscedasticity problem, assuming normality.
With respect to the normality assumption, Box G. mentioned "*in the real world a normal distribution does not exist but, from models known to be false, often useful approximate results can be derived* [1]", this leads us once again to the question about the possible inadequate formulation of traditional hypotheses used for normality pretesting.

[4], through simulations, showed that assuming normality and pretesting homoscedasticity under an equivalence approach, the TIEP is always close to the significance level α. The estimated values of these probabilities are very similar in all cases to those obtained when the Welch's test is applied directly, without pretest. It seems that the concerns about the inadequate formulation of the traditional pretests to prove the assumptions are correct. In addition, from these results it is demonstrated that any researcher who can guarantee normality can pretest homoscedasticity prior to choosing between Welch or t-test, without fear of increasing the probability of Type I error, as long as the pretest is performed under an equivalence approach. However, it is well known that in practice the only way to know if a set of data comes from a normal distribution or not is pretesting this assumption.

The aim of this paper is to extend the previous results, comparing through simulation the overall TIEP affectation when normality and homoscedasticity are pretested either under an equivalence approach or under the classical pretesting approach. Following [1], we will focus on determining if any of the indicated procedures or models is good enough to be useful. We considerate a good approximation when the global TIEP is close to the nominal significance level α. The Cochran's criterion establishes that a maximum distance of 20% of TIEP around α is considered a good approximation [2]. In addition [9] defining the robustness of a test, suggest a similar criterion. To reduce the variability and to improve the precision of the simulation estimates, a variance reduction technique for Bernoulli variables named "Control Variates" will be used [7].

To generate non normal samples, we will use the Fleishman's system of distributions [3]. This system considers that every distribution for which the first four moments exist can be obtained through the transformation $Z = a + bX + cX^2 + dX^3$, where X is a standard normal random variable and Z is a variable with unknown distribution and parameters $(\mu = 0, \sigma^2 = 1, \gamma_1, \gamma_2)$ where γ_1 and γ_2 represent the skewness and kurtosis respectively (a standard normal distribution is obtained when $b = 1$ and $a = c = d = 0$). Finally, with $y = \mu + \sigma Z$ unknown distributions with parameters $(\mu, \sigma^2, \gamma_1, \gamma_2)$ can be generated. In our case, μ will be always the same and σ^2 will be stablished from the ratio of variances σ_1^2/σ_2^2.

References

[1] George EP Box. Robustness in the strategy of scientific model building. *Robustness in statistics*, 1:201–236, 1979.

[2] William G Cochran. The χ^2 correction for continuity. *Iowa State College Journal of Science*, 16(1):421–436, 1942.

[3] Allen I Fleishman. A method for simulating non-normal distributions. *Psychometrika*, 43(4):521–532, 1978.

[4] Pablo Flores Muñoz. Un pre test de irrelevancia de la diferencia de varianzas en la comparación de medias. Master's thesis, Universitat Politècnica de Catalunya, 2017.

[5] PL Hsu. Contribution to the theory of "student's" t-test as applied to the problem of two samples. *Statistical Research Memoirs*, 1938.

[6] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[7] Jordi Ocaña and Esteban Vegas. Variance reduction for bernoulli response variables in simulation. *Computational statistics and data analysis*, 19(6):631–640, 1995.

[8] John E Overall, Robert S Atlas, and Janet M Gibson. Tests that are robust against variance heterogeneity in k× 2 designs with unequal cell frequencies. *Psychological reports*, 76(3):1011–1017, 1995.

[9] Dieter Rasch and Volker Guiard. The robustness of parametric statistical methods. *Psychology Science*, 46:175–208, 2004.

[10] Dieter Rasch, Klaus D Kubinger, and Karl Moder. The two-sample t test: pre-testing its assumptions does not pay off. *Statistical Papers*, 52(1):219–231, 2011.

[11] Henry Scheffé. Practical solutions of the behrens-fisher problem. *Journal of the American Statistical Association*, 65(332):1501–1508, 1970.

[12] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[13] BL Welch. On the comparison of several mean values: an alternative approach. *Biometrika*, 38(3/4):330–336, 1951.

[14] Stefan Wellek. *Testing statistical hypotheses of equivalence and noninferiority*. CRC Press, 2010.

[15] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[16] Donald W Zimmerman. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1):173–181, 2004.