LETTER TO THE EDITOR

Goodness-of-fit and concordance analysis

Bondad de ajuste y análisis de concordancia

Christian R. Fau[1], Solange Nabzo[1] and Veronica Nasabun[2]*

[1]Ophthalmological Foundation 2020, Iberoamerican Cochrane Network; [2]School of Nursing, Andrés Bello University. Santiago, Chile

Our life is wasted on details...
We must simplify. Simplify.
Henry David Thoreau

Dear Dr. Manuel Garza León, we have read in the May-June 2019 issue of the *Revista Mexicana de Oftalmología*, an article published by Saucedo-Urdapilleta, et al., "Comparative analysis and repeatability assessment of IOL Master 500 versus IOL Master 700 biometry in cataract patients"[1]. This study compares the IOL Master 500 and the IOL Master 700 in a population of 55 eyes of 55 patients, where it was intended to establish a concordance analysis between both equipment, what they call a "repeatability analysis", according to the measurements of axial length, keratometry, anterior chamber depth (ACD) and white-to-white distance. After critically analyzing the article, we decided to express some thoughts related to the statistics used.

In the study, in the statistical analysis section, it is established that: "The database was reviewed using the Kolmogorov-Smirnov (KS) test", which is not cited in the article. The test result is not included in the results, like if it was never done. Regardless of this fact, it is important to reflect on what these goodness-of-fit tests are and which should be used in this case.

Many investigations use parametric statistical tests in their analysis. In this case, Pearson's correlation coefficient and Student's t-test were used for paired data; both assume a normal distribution in the sample. Violating this assumption makes the interpretation of results complex, even when there are studies that indicate that these tests are robust when both the assumption of normality and the homoscedasticity assumption are violated[2]. In general, in cases where the sample is not normally distributed, the use of non-parametric tests is recommended, and in this case it was proposed to use "the Wilcoxon signed-rank test", and it is not clear if it was used or not. In practice, parametric tests are used in many investigations, assuming normality and without any verification of assumption. This step that should be performed prior to data analysis, is usually not performed due to lack of awareness of the authors.

There are currently several statistical tests that allow us to verify the assumption of normality. These are the K-S test, the K-S test with the Lilliefors correction (K-S-L), the Shapiro-Wilk test (S-W), the Jarque-Bera test (J-B) and the Anderson Darling test (A-D)[3]. The K-S test is one of the most classic for the study of normality. It was developed by two Russian mathematicians, A. Kolmogorov and N.V. Smirnov, who presented two similar tests in the 1930s. This test compares a theoretical distribution function with an empirical one, and provides a p-value, the probability that the analyzed sample differs from a random sample of size "n", obtained from a normal distribution. Therefore, in this test and others, the null or H0 hypothesis is that there are no differences between the samples and, therefore, the null hypothesis is not rejected, that is, $p > 0.05$. This is an excessively conservative test, which accepts H0

Correspondence:
*Christian R. Fau
Avda. Presidente Riesco 5157, Dep 212,
Las Condes Santiago, 7560854, Chile
E-mail: cfau@fundacion2020.org; chfauf@gmail.com

Date of reception: 04-11-2019
Date of acceptance: 07-11-2019
DOI: 10.24875/RMOE.M20000105

Available online: 01-03-2020
Rev Mex Oftalmol (Eng). 2020;94(2):87-89
www.rmo.com.mx

in an excessively high number of occasions, so, despite its wide use and easy access, it is the least suitable test to verify data normality. Lilliefors, in 1967, with the intention of improving the K-S test, proposed a modification that is used when the mean and variance are unknown. Although at the time it was proposed as an improvement, it is still very conservative, and although it rejects H0 in some cases, it requires sample sizes of over 500 participants to have an adequate performance. The S-W test (1965) is one of the most consolidated tests with the greatest statistical power among the current ones, especially when used with short-tailed distributions and with small sample sizes. Its best performance is with sample sizes greater than 50 participants, and it even improves when sample size increases, and is the best classic method to use with sample sizes smaller than 50 participants, even when it loses performance. The J-B test (1987) has shown high consistency, especially when working with large samples with symmetric distributions and long tails. For this there is an Urzúa correction (1996), which has not been shown to significantly improve the classical test. This test shows good performance with sample sizes of more than 200 participants, and in smaller sample sizes it has worse performance than the K-S-L test. Finally, the A-D test involves a modification of the Crammer-Von Mises test. It is based on the difference of squares between the distributions. This test is the best when analyzing small and symmetric distributions, smaller than 30 participants. It is one of the most powerful statistical tests in most cases, except for excessively large samples, where it behaves more conservatively, such as classical tests. Therefore, the main tests used to establish normality in the studies should be A-D for small samples of less than 30 participants, S-W for samples of more than 50 participants, and both for the intermediate segment. In the study cited here, they should have used one of these two tests.

Regarding the repeatability analysis mentioned in the study, in which the Pearson correlation coefficient and the Student's t-test were used for paired data, it is important to clarify that repeatability is understood as performing more than one measurement in the same person with the same instrument, but under identical conditions. In the case of the study this is not true, since it is actually a concordance analysis between two measurement methods to evaluate how equivalent they are.

Concordance analyses between variables are widely used in the clinical practice, concordance between measurements is altered due to intraobserver and interobserver variability and by the measuring instrument itself, which is what was under evaluation in this case. In the case of continuous quantitative variables, it is common that inappropriate statistical analysis techniques are used, in this case the Pearson correlation coefficient. This is not an adequate method to assess the degree of agreement between two variables, since a r = 1 can be obtained, that is, a perfect correlation, although one of the measurement methods is proportionally biased, so therefore, in all the measurements it marks an X value + a constant (c), in spite of this perfect correlation there is a null concordance between the measurements. That is to say an equipment shows X and the other X + c, they are totally different, therefore, Pearson's correlation coefficient does not provide information on the agreement between two methods, but only measures the linear association between two variables. Also, the Student's t-test for paired data is not a suitable technique for this type of analysis. It involves only a comparison of means, without comparing distribution. In this type of analysis, analyzing the variance is preferred over the mean, so it is better to use an ANOVA, and based on this ANOVA, an intraclass correlation coefficient (ICC) is calculated, which is a parametric test. This coefficient estimates the average of the correlations between all possible pairs of observations available, like Pearson's r, this ICC ranges between 0 and 1, so that the maximum concordance between the two methods would be 1, and in that case, the variability observed would be explained by the subjects and not by the differences between the measurement methods or the observers. When the ICC value is 0, the concordance observed is only a product of chance. Regarding this, the ICC can be assessed as follows: > 0.90 very good; 0.90-0.71 good; 0.70-0.51 moderate; 0.50-1.31 mediocre; <0.30 very low or zero. There are other methods for assessing concordance, such as Lin's coefficient of concordance, Deming's orthogonal regression method, the Passing-Bablock regression model, etc., but they are rarely used.

Returning to the study, the author should not have used Pearson's correlation coefficient and Student's t-test for paired data, but rather an ANOVA analysis and an ICC. As a result of this bad choice, there was a problem that the author avoided explaining in the text, this was: an ACD and a significantly longer axial length was obtained with the IOL Master 700 than with the IOL Master 500 ($p < 0.038$ and $p < 0.0003$), but Pearson's r was r = 0.959 and r = 0.997, respectively, a very high correlation. This is, are the equipment correlated or not when measuring? Do they measure or not the same? In this case, as noted earlier, the author encountered a proportional bias where one equipment systematically measures more than the other and, therefore, Pearson's r has no value. The ICC had to be calculated, which would have shown that the concordance was not high.

The important effort that is made in the development of an investigation should not be destroyed by an error of analysis, when in fact the data should be used to obtain valid conclusions. It is important that authors who do not have advanced knowledge of statistical analysis and software management such as Stata, SAS or SPSS, seek the advice of biostatistics experts, since the risk is that when they send an article for evaluation, it can be rejected, or if accepted, when it is read it will be quickly discarded, endangering personal and journal prestige, and ultimately, not generating any impact as a publication.

References

1. Saucedo-Urdapilleta R, González-Godínez S, Mayorquín-Ruiz M, Moragrega-Adame E, Velasco-Barona C, González-Salinas R. Estudio comparativo entre los biómetros ópticos IOL Master 500 versus IOL Master 700 en pacientes con catarata y análisis de repetibilidad. Rev Mex Oftalmol. 2019;93(3):130-6.
2. Finch H. Comparison of the performance of nonparametric and parametric MANOVA test statistics when assumptions are violated. Methodology. 2005;1(1):27-38.
3. Pedrosa IG, Juarros-Basterretxea J, Basteiro J. Pruebas de bondad de ajuste en distribuciones simétricas, ¿qué estadístico utilizar? Universitas Psychologica. 2015;14(1): 245-54.