Revista Mexicana de Oftalmología

Permanyer

CARTA AL EDITOR

Bondad de ajuste y análisis de concordancia

Goodness-of-fit and concordance analysis

Christian Fau[1], Solange Nabzo[1] y Veronica Nasabun[2]*

[1]Fundación Oftalmológica 2020, Iberoamerican Cochrane Network; [2]Escuela de Enfermería, Universidad Andrés Bello. Santiago, Chile

Nuestra vida se desperdicia en los detalles…. Hay que simplificar. Simplificar.
Henry David Thoreau

Estimado Dr. Manuel Garza León, hemos leído en la *Revista Mexicana de Oftalmología*, del mes de mayo-junio de 2019, un artículo publicado por Saucedo-Urdapilleta, et al., «Estudio comparativo entre los biómetros ópticos IOL Master 500 versus IOL Master 700 en pacientes con catarata y análisis de repetibilidad»[1]. En este estudio se realiza una comparación entre el IOL Master 500 y el IOL Master 700, en una población de 55 ojos de 55 pacientes, donde se pretendía establecer un análisis de concordancia entre ambos equipos, a lo que llaman «analizar la repetibilidad», según las mediciones de longitud axial, queratometrías, profundidad de cámara anterior (ACD) y distancia blanco-blanco. Luego de que analizamos críticamente el artículo, decidimos presentarle algunas reflexiones en relación con la estadística empleada.

En el estudio, en la sección de análisis estadístico, se establece que: «la base de datos se revisó usando el test de Kolmogorov-Smirnov (K-S)", el cual nunca más se cita en el artículo. El resultado del test no se encuentra en los resultados, como que nunca se hubiese realizado. Independientemente de este hecho, es importante hacer una reflexión de para qué son estos test de bondad de ajuste y cuál debe ser utilizado en este caso.

Muchas investigaciones utilizan pruebas estadísticas paramétricas en su análisis. En este caso se usó el coeficiente de correlación de Pearson y la prueba t de Student para datos pareados, ambos presuponen una distribución normal en la muestra. El violar este supuesto hace que las interpretaciones de los resultados sean complejas, aun cuando hay estudios que señalan que estas pruebas son robustas cuando se viola tanto el supuesto de normalidad como el de homocedasticidad[2]. En general, en los casos en que la muestra no se distribuye normalmente, se recomienda el uso de pruebas no paramétricas, y en este caso se planteó usar «la prueba de los rangos con signo de Wilcoxon», la cual tampoco se aclara si se usó o no. En la práctica, en muchas investigaciones se emplean pruebas paramétricas, suponiendo la normalidad y sin ningún tipo de comprobación del supuesto. Este paso que se requiere realizar previo al análisis de los datos, muchas veces no se realiza por desconocimiento de los autores.

Actualmente existen varias pruebas estadísticas que permiten comprobar el supuesto de normalidad. Estas son: la prueba de K-S, la prueba de K-S con la corrección de Lilliefors (K-S-L), la prueba de Shapiro-Wilk (S-W), la prueba de Jarque-Bera (J-B) y la prueba de Anderson Darling (A-D)[3]. La prueba de K-S es una de las más clásicas en el estudio de la normalidad. Fue desarrollada por dos matemáticos rusos, A. Kolmogorov

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

Fecha de recepción: 04-11-2019
Fecha de aceptación: 07-11-2019
DOI: 10.24875/RMO.M20000107

Disponible en internet: 01-03-2020
Rev Mex Oftalmol. 2020;94(2):100-102
www.rmo.com.mx

y N.V. Smirnov, quienes presentaron dos pruebas similares en la década de 1930. Esta prueba compara una función de distribución teórica con la empírica y proporciona un valor de p, probabilidad de que la muestra analizada difiera de una muestra aleatoria de tamaño n, obtenida de una distribución normal, por lo tanto, en esta prueba y otras, la hipótesis nula o H_0 es que no hay diferencias entre las muestras y, por lo tanto, se busca no rechazar la hipótesis, o sea un $p > 0.05$. Esta es una prueba excesivamente conservadora, la cual acepta la H_0 en un número excesivamente elevado de ocasiones, por lo que, a pesar de su amplio uso y fácil acceso, es la prueba menos adecuada para comprobar la normalidad. Lilliefors, en 1967, con la intención de mejorar la prueba de K-S, propuso una modificación que se utiliza cuando la media y la varianza son desconocidas. Si bien en su momento se planteó como una mejora, sigue siendo muy conservadora, y si bien rechaza la H_0 en algunos casos, requiere de tamaños de muestra sobre 500 participantes para tener un mejor desempeño. La prueba de S-W (1965) es una de las pruebas más consolidadas y con mayor potencia estadística entre las actuales, especialmente se utiliza cuando se trabaja con distribuciones de colas cortas y con tamaños muestrales pequeños. Su mejor desempeño es con tamaños de muestra mayores a 50 participantes, e incluso mejora con el aumento del tamaño muestral, y es el mejor método clásico a usar con tamaños muestrales menores de 50 participantes, aun cuando pierde desempeño. La prueba de J-B (1987) ha demostrado una alta consistencia, especialmente cuando se trabaja con muestras grandes de distribuciones simétricas y de colas largas. Para esta existe una corrección de Urzúa (1996), la cual no ha demostrado mejorar de forma significativa la prueba clásica. Esta prueba muestra un buen desempeño con tamaños muestrales de más de 200 participantes, y en tamaños muestrales menores tiene peor desempeño que la prueba de K-S-L. Finalmente, la prueba de A-D supone una modificación del test de Crammer-Von Mises. Se basa en la diferencia de cuadrados entre las distribuciones. Esta prueba es la mejor cuando se analizan distribuciones simétricas y de tamaño pequeño, menor de 30 participantes. Es una de las pruebas estadísticas más potente en la mayoría de los casos, exceptuando las muestras excesivamente grandes, en donde se comporta más conservadoramente, como las pruebas clásicas. Por lo tanto, las principales pruebas para establecer normalidad en los estudios debieran ser A-D en muestras pequeñas, de menos de 30 participantes, S-W en muestras de más de 50 participantes, y en el segmento intermedio ambas. En el estudio citado al principio se debió utilizar una de estas dos pruebas.

En relación con el análisis de la repetibilidad, planteado en el estudio, en el cual se utilizó el coeficiente de correlación de Pearson y la prueba t de Student para datos pareados, cabe hacer una aclaración, se entiende por repetibilidad el realizar en una misma persona más de una medición con un mismo instrumento, pero en condiciones idénticas. En el caso del estudio esto no aplica, ya que en este caso se trata en realidad de un análisis de concordancia entre dos métodos de medición y se busca evaluar que tan equivalentes son.

Los análisis de concordancia entre variables son muy utilizados en la práctica clínica, la concordancia entre mediciones se ve alterada por la variabilidad intraobservador, interobservador y por el propio instrumento de medición, que es lo que en este caso se busca evaluar. En el caso de variables cuantitativas continuas es frecuente que se utilicen técnicas estadísticas inapropiadas de análisis, en este caso el coeficiente de correlación de Pearson. Este no es un método adecuado para evaluar el grado de acuerdo entre dos variables, ya que se puede obtener un $r = 1$, o sea una correlación perfecta, a pesar de que uno de los métodos de medición está sesgado de forma proporcional, por lo tanto, en todas las mediciones marca un valor X + una constante (c), a pesar de esta correlación perfecta hay una nula concordancia entre las mediciones, o sea un equipo marca X y el otro X+ c, son totalmente diferentes, por lo tanto, el coeficiente de correlación de Pearson no proporciona información sobre el acuerdo entre dos métodos, sino que solo mide la asociación lineal entre dos variables. Tampoco la prueba t de Student para datos pareados es una técnica adecuada para este tipo de análisis. En esta se realiza solo una comparación de medias, sin comparar su distribución. En este tipo de análisis se prefiere el análisis de la varianza por sobre la media, por lo que se prefiere usar un análisis de ANOVA, luego basado en este análisis de ANOVA se calcula un coeficiente de correlación intraclase (CCI), que es una prueba paramétrica. Este coeficiente estima el promedio de las correlaciones entre todos los posibles pares de observaciones disponibles, al igual que el r de Pearson, este CCI oscila entre 0 y 1, de modo que la máxima concordancia entre los dos métodos sería 1, y en ese caso, la variabilidad observada estaría explicada por los sujetos y no por las diferencias entre los métodos de medición o los observadores. Cuando el valor del CCI es 0, la concordancia observada es solo

producto del azar. En relación con esto, el CCI se puede valorar de la siguiente manera: > 0.90 muy buena; 0.90-0.71 buena; 0.70-0.51 moderada; 0.50-0.31 mediocre; < 0.30 mala o nula. Existen otros métodos para evaluar la concordancia, como son el coeficiente de concordancia de Lin, el método de regresión ortogonal de Deming, el modelo de regresión de Passing-Bablock, etc., pero son muy poco utilizados.

Volviendo al estudio, el autor no debió utilizar el coeficiente de correlación de Pearson y la prueba t de Student para datos pareados, sino un análisis de ANOVA y un CCI. Producto de esta mala elección, se produjo un problema que el autor eludió explicar en el texto, este fue: se obtuvo una ACD y una longitud axial significativamente más larga con el IOL Master 700 que con el IOL Master 500 ($p < 0.038$ y $p < 0.0003$), pero el r de Pearson fue $r = 0.959$ y $r = 0.997$, respectivamente, altísima correlación, o sea ¿los equipos están o no correlacionados al medir?, ¿miden o no miden lo mismo? En este caso, como se señaló antes, el autor se topó con un sesgo proporcional donde un equipo mide sistemáticamente más que el otro y, por lo tanto, la r de Pearson no tiene valor alguno. Se debió calcular el CCI, el cual habría mostrado que la concordancia no era alta.

El importante esfuerzo que se realiza en el desarrollo de una investigación no puede quedar destruido por un error de análisis, cuando en realidad los datos están para poder realizar algo mejor. Es importante que los autores que no posean conocimientos avanzados de análisis estadísticos y manejos de *software* como Stata, SAS o SPSS, se asesoren con bioestadísticos, ya que se arriesgan a que al enviar un artículo para su evaluación en una revista, este sea rechazado, o de ser aceptado, cuando este sea leído será descartado rápidamente, lo que arriesga su prestigio y el de la revista que lo publica, y finalmente no genera ningún impacto como publicación.

Bibliografía

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.