Temas especiales

35 Años 1981-2016 Dirección General de Investigación

Una primera mirada a las pruebas estadísticas de hipótesis

Ver boletín completo

"Es necesario construir mentalmente una hipótesis teórica y someterla a la prueba de las mediciones experimentales"

Max Planck

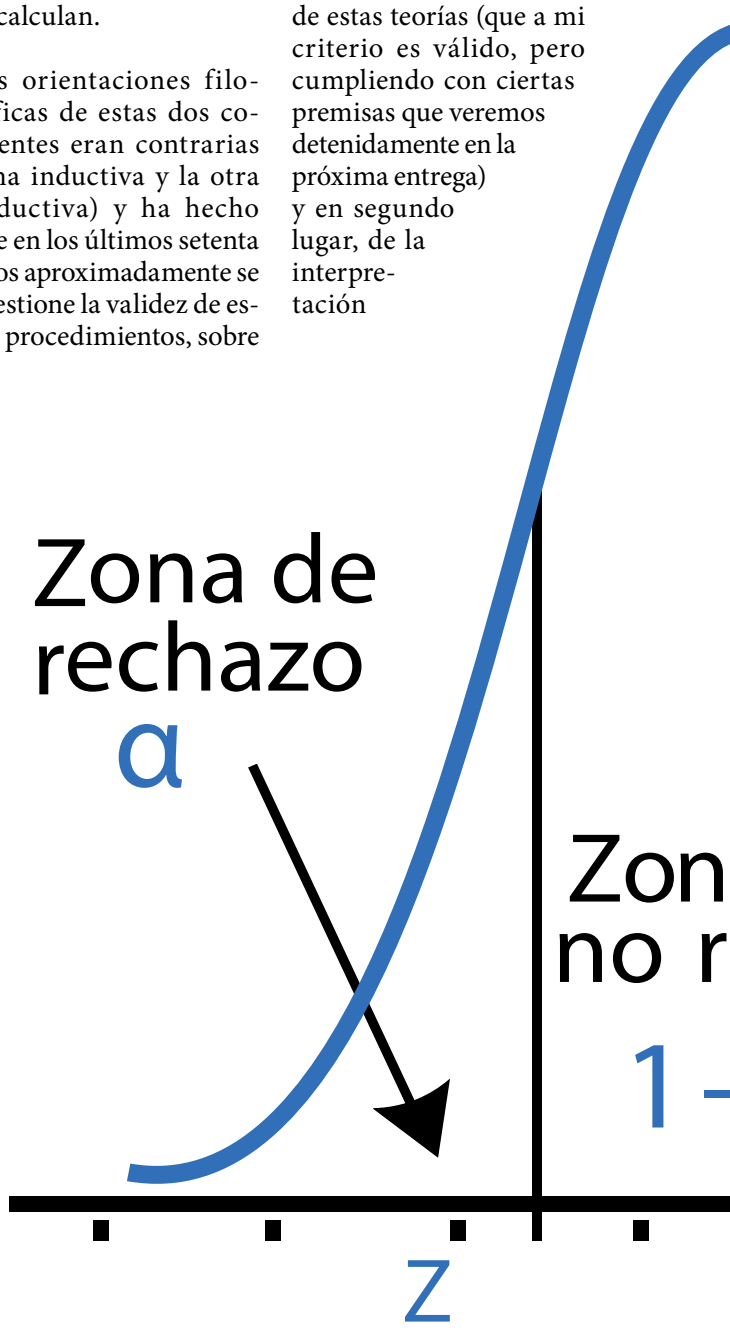
Oscar Federico Nave Herrera
Programa de Asesoría Estadística para Investigación
Dirección General de Investigación
fnave@digi.usac.edu.gt

Esta frase del gran físico y matemático Max Planck, resume el objetivo fundamental del proceso de investigación en el que se busca la explicación de los fenómenos y de paso, el tema del presente artículo: probar las hipótesis por medio de la experimentación.

Este tema ha caminado un largo trecho que inició a principios del siglo XX con los trabajos de Ronald A. Fisher y su propuesta de "dócima o prueba de significación", estableciendo que, para que un resultado fuera significativo, debía calcularse la probabilidad bajo la premisa de una hipótesis (que llamó hipótesis nula) de obtener valores del estadístico iguales o más extremos que los observados en el experimento, lo que en resumen y en forma más práctica, se interpretó como una medida de discrepancia entre los datos y la hipótesis nula y se le llamó "valor p"; Fisher indicó que se debía tener un valor umbral o límite de tolerancia para poder rechazar la hipótesis nula, llamado nivel de significancia o valor alfa (α).

Pocos años después, alrededor de finales de los años 20, dos matemáticos y estadísticos muy críticos a la propuesta de Fisher, Jerzy Neyman y Egon Pearson (hijo del gran Karl Pearson), plantearon una propuesta diferente que llamaron "dócima o prueba de hipótesis", con la cual buscaban reglas que fueran aplicables a las hipótesis planteadas (nula y alterna), de manera que se pudieran reducir los errores que se podrían cometer si se rechazaba una hipótesis verdadera (error tipo I) o si se aceptaba una hipótesis falsa (error tipo II); propusieron que la magnitud de estos errores debe ser calculada para cada experimento e interpretarse en función de las posibles consecuencias si se comete uno de ellos, creando los conceptos de regiones de rechazo y aceptación de hipótesis, dentro de las distribuciones de probabilidad de los estadísticos de prueba que se calculan.

Las orientaciones filosóficas de estas dos corrientes eran contrarias (una inductiva y la otra deductiva) y ha hecho que en los últimos setenta años aproximadamente se cuestione la validez de estos procedimientos, sobre la base en primer lugar, de la tendencia de los cursos y libros de estadística a brindar al estudiante una especie de híbrido de estas teorías (que a mi criterio es válido, pero cumpliendo con ciertas premisas que veremos detenidamente en la próxima entrega) y en segundo lugar, de la interpretación



que por muchos años se ha dado a los resultados que la literatura científica publica, la discusión sobre lo que es significativo o no significativo, si el valor p es suficiente, los criterios para determinar el umbral de rechazo a la hipótesis nula (¿por qué se usa 0.05 "de cajón"? por ejemplo), así como la discrepancia que se da a veces entre la significación estadística y la significación práctica, la falta de homogeneidad en el uso de los valores p en la literatura científica y el sesgo editorial de solo publicar lo significativo.

Todo lo anteriormente indicado, motivó que en el 2001, el Comité Internacional de Directores de Revistas Médicas sentenciara: "Se evitará la dependencia exclusiva de las pruebas estadísticas de verificación de hipótesis, tal como el uso de los valores p, que no aportan ninguna información cuantitativa importante". De manera más tajante, en febrero de 2015, los editores de la revista Basic and Applied Social Psychology (revista que se encuentra en el cuartil intermedio superior de la especialidad según el portal web SCImago Journal & Country Rank) anunciaron que desde esa fecha, se prohíbe la publicación de artículos que "desarrollen procedimientos de pruebas de significancia de hipótesis nula" (un verdadero engendro de lo que Fisher o Neyman-Pearson propusieron en su momento); el asunto venía gestándose desde su Editorial publicado en enero de 2014, en el que se indica: "Se ha demostrado que el procedimiento de prueba de significación de hipótesis nula es lógicamente inválido y provee poca información acerca de la probabilidad real de cualquiera de las hipótesis nula o experimental".

Todo esto deriva de un estudio publicado en dicha revista por el Dr. David Trafimow de la Universidad Estatal de Nuevo México, en el 2003, en el cual al aplicar análisis bayesiano (opuesto a la teoría estadística tradicional), concluye que la probabilidad de obtener un resultado experimental dado que la hipótesis nula es verdadera no es igual a la probabilidad que la hipótesis nula sea verdadera dada la presencia de un hallazgo experimental, llamándolo "inferencia inversa", discutiendo sobre la debilidad de las pruebas de hipótesis en determinar una verdadera significancia. Aunque el Dr. Trafimow indica que la única forma de evitar el mal uso del valor p es evitar sacar conclusiones de este, no hay que perder de vista que existe en mucha de la literatura científica, lo que algunos autores han dado en llamar "p-hacking", que es una mala aplicación de las pruebas de hipótesis por el sesgo que se da cuando los investigadores recogen o seleccionan los datos o aplican el análisis estadístico hasta que los resultados no significativos se convierten en significativos.
Como se ve, esta situación viene ya de varios años atrás, pero aún no ha tenido eco en otras publicaciones, ni ha despertado comentarios de matemáticos, estadísticos o epidemiólogos, aunque Nature en su volumen 519 de marzo de 2015 lo destacara como noticia diciendo: "Una prueba estadística polémica ha llegado a su fin, al menos en una revista". De acuerdo al índice y tipo de citaciones del artículo del Dr. Trafimow, parece ser que su teoría se ha quedado dentro del ámbito de su revista, de la cual es Editor, existiendo algunos artículos que lo citan en otras revistas dentro de la misma área de conocimiento. La mayoría de publicaciones sigue considerando válido el procedimiento tradicional, pero recomiendan que los resultados no se centren solamente en el valor p, sino que se acompañe el mismo con otros elementos que den mayor información, tales como el efecto de tamaño e intervalos de confianza. Considerando lo anterior, una prueba estadística de hipótesis, bien diseñada, bien aplicada y sobre todo, bien interpretada no deja de tener validez, los otros elementos que se mencionan pueden complementarla y hacer que el análisis sea más robusto. Continuaremos en la próxima entrega con este apasionante tema.

a de
echazo

–α