Revista FABICIB, Año 1999, Volumen 3, págs. 163-166

Algo de la estadística de todos los días: el valor P o P- value

Carrera, Elena F. de; Contini, Liliana E.; Vaira, Stella M.

Departamento de Matemática. Area Estadística. Facultad de Bioquímica y Cs. Biológicas - UNL.
Paraje el Pozo. C.C. 242 -3000- Santa Fe. Argentina. Tel. (0342)-4575210. E-mail: ecarrera@fbcb.unl.edu.ar

RESUMEN: El objetivo de este artículo es aclarar algunos conceptos de uso muy frecuente tales como valor P y α que están permanentemente confundidos. Está dirigido a los usuarios de la estadística, es decir, aquellos investigadores no estadísticos que en su quehacer necesitan permanentemente contrastar hipótesis.

SUMMARY: ***SOMETHING OF STATISTICS FOR EVERY DAY: THE P-VALUE.*** **Carrera, Elena F. de; Contini, Liliana E.; Vaira, Stella M.** The goal of this work is to clarify some concepts frecuently used and mistaken like P-value and α. It is no aimed at statistic researchers but at the ones who use statistics at their jobs and need to constrast hypothesis permanently.

Introducción

Los lectores actuales de revistas bioquímicas, médicas, químicas y biológicas, por enumerar sólo algunas, así como los científicos biomédicos se enfrentan a una gran variedad de citas estadísticas; los editores a la hora de recibir una publicación requieren tratamiento estadístico preciso de los autores (1,2,3). Entre ellas o ellos desde hace unos años ha aparecido la letra P o en algunas publicaciones p, así con minúscula. En esta nota se usará la mayúscula para identificarlo porque es el sujeto de la misma.

La presencia de $P < 0.001$, $P > 0.02$, $P < 0.0001$, $P = 0.00123$ es común en todas las publicaciones científicas. Además, siempre se presenta la discusión si el P debe ser exacto como se señala en la última expresión anterior o bien como figura en las tres primeras. En síntesis ¿qué es el P? ¿Se debe expresar con igualdades o con desigualdades? Este artículo pretende arrojar un poco de luz en la discusión señalada. Además pretende aclarar la confusión que existe entre este novel valor P y el ya tradicional $\alpha < 0.05$ ó $\alpha < 0.01$. La expresión de novel valor P es por el uso generalizado que se da actualmente al mismo, aunque aparece asociado al modelo estadístico de contraste de hipótesis que propusieron Neyman y Pearson en 1928 (4).

El contraste de hipótesis

Uno de los recursos estadísticos más utilizados es el de contraste de hipótesis, prueba (o testeo) de hipótesis. Esto no es nada más que un procedimiento de decisión basado en datos experimentales para probar una hipótesis acerca de algún sistema científico (5,6). Debe tenerse en cuenta que una hipótesis estadística es una conjetura o afirmación acerca de una o más poblaciones (6).

Por ejemplo, si se está interesado en contrastar la absorción media de un medicamento bajo la presencia de un alimento G y en ausencia del mismo; se formulará entonces una hipótesis inicial: la absorción media del medicamento es la misma bajo la presencia o ausencia del alimento G; ésta naturalmente tiene implícita estas otras hipótesis: la absorción es distinta en ambas circunstancias, o una absorción es mejor que la otra. La estructura de este procedimiento es bastante simple. Se formula una hipótesis nula y simultáneamente una hipótesis alternativa (6,7). En esta formulación, cuando se rechaza la hipótesis nula se acepta una hipótesis complementaria alternativa, lo cual en un ensayo clínico puede ser que los dos tratamientos sean igualmente efectivos o bien que no lo sean.

El valor P aparece siempre que se encuentre un contrastes de hipótesis, ya sea de manera implícita o explícita. Casos como aquellos en que se quiere saber: si la correlación entre las variables existe, si hay homogeneidad de variancias, si las variables se comportan de forma normal o log-normal, sólo para señalar algunos pocos hechos, también constituyen contraste de hipótesis, o sea no sólo se contrastan las medias. En cada uno de estos casos se calculará el estadístico de prueba que corresponda y que depende de la comparación a realizar. Esto implica que, basado en los datos obtenidos, se realicen

unos cálculos propios en cada caso dando lugar a una cantidad que se denomina estadístico. Sobre este estadístico se basará la decisión a tomar en la investigación que se está realizando. (5-11)

El nivel de significancia

En un contraste de hipótesis es probable que se puedan cometer dos tipos de errores, uno llamado de Tipo I, que se produciría cuando nuestros datos nos conducen a rechazar la hipótesis nula en favor de la alternativa, cuando de hecho la hipótesis nula es verdadera y otro llamado de Tipo II. Este, se produce cuando a partir de nuestros datos, concluimos que la hipótesis nula es verdadera cuando en realidad es falsa (5-11). En otra nota hablaremos sobre estos tipos de errores y sus implicancias en las investigaciones que se realizan.

La probabilidad de cometer un error de Tipo I, se conoce con el nombre de nivel de significancia y se representa con la letra griega α. Esto delimita lo que comúnmente se conoce como zona de rechazo (5-11). Estamos dando un tamaño al error de Tipo I, es decir qué probabilidad queremos asignar al decidir que las cosas son diferentes cuando en realidad no existe ninguna diferencia (7). La elección del valor de α es arbitraria (3). Los valores más usados son 0,05 y 0,01 en la práctica, en la mayoría de los casos se usa el valor 0.05 con mayor frecuencia que el 0.01 ya señalado. Este valor del 0.05, fue una estrategia propuesta por Sir Ronald Fisher, quien consideró el 95% de los valores interiores de una distribución cualquiera como comunes y el 5% remanente como infrecuentes. Algunos autores consideran este nivel como un nivel de rareza (3). El significado de ese 0.05 es que 5 de cada 100 veces se comete una equivocación, por eso la designación de error, al decidir que las cosas a comparar son distintas cuando en realidad no lo son. Es lo mismo decir que una vez en 20 ocurre ese suceso, esto significa que es una probabilidad muy baja y ello conduce a rechazar la hipótesis nula. Si se desea no equivocarse tanto al rechazar la hipótesis nula se puede elegir un nivel de significancia del 1% o sea de 0.01. En este último caso el α es menor y por lo tanto se está diciendo que la posibilidad que se admita el error al decir que las cosas que se comparan son diferentes, cuando en realidad no lo son, será menor y por ende el trabajo tendrá mayor precisión.

De lo señalado se desprende que es necesario establecer un estándar de "rareza" al comenzar a realizar el trabajo.

P y alfa

Habiendo establecido la hipótesis nula y el nivel de significancia a emplear, se evalúa entonces la probabilidad de haber obtenido el dato observado si la hipótesis nula fuera verdadera (5-11).

Esta probabilidad es usualmente llamada el valor P o P-value. Se observa que todavía no se ha comparado el estándar de rareza, que seleccionó el investigador, con el valor P que se ha calculado.

La interpretación del valor P es problemática (7). Si se realiza un ensayo para comparar dos tratamientos y se obtiene un valor grande de P, por ejemplo más grande que 0.2, entonces se puede decir que datos tales como los que se obtuvieron deberían ocurrir en 20 de cada 100 veces que la hipótesis nula es verdadera. La posibilidad de obtener esos datos es algo elevada. No se puede rechazar la posibilidad que la hipótesis nula es verdadera; esto es que los dos tratamientos son igualmente efectivos. Más aún, se puede tener la intuición, casi la certeza, que la hipótesis nula no es verdadera y uno de los tratamientos es superior. Pero no se tienen, a partir de la experiencia realizada, elementos suficientes que permitan asegurarlo. Además se encuentra un área gris donde cualquier afirmación que se haga resulta conflictiva. Para solucionar este problema es necesario que se compare el valor de P con α, el nivel de significancia o estándar de rareza que se ha adoptado.

Si el valor P excede el valor α no se rechaza la hipótesis nula. Más aún no puede decirse que haya evidencias para creer que la hipótesis nula es verdadera, pero sí que no hay demasiada evidencia para rechazarla. Esto es una importante distinción. Cuando P es menor que el α el resultado se dice estadísticamente significativo (mucho menor que éste nivel se dice altamente significativo) y cuando supera el valor de α se dice no significativo o que no hay evidencias para rechazar la hipótesis nula.

La expresión no significativa quiere decir que es probable que la diferencia obtenida sólo se deba al azar y el término significativo sólo quiere decir que es improbable que la diferencia obtenida se deba al azar.

Lo significativo estadísticamente y lo significativo clínicamente

El uso de la expresión significativo o no significativo puede llevar a error porque algo puede ser estadísticamente significativo pero no clínicamente significativo(2,14). Similarmente no es razonable tomar un resultado no significativo como indicando que no se produce efecto, solamente porque no puede no rechazarse la hipótesis nula.

La significatividad estadística no es más que un dato de la probabilidad de que haya una diferencia cualquiera, de cualquier magnitud entre las cosas que se están comparando o sea si es probable que las conclusiones que se extrajeron sean verdaderas (3,7,12).

Es importante señalar la diferencia entre significatividad estadística y significatividad clínica. La significatividad estadística no dice nada acerca de la magnitud o importancia de la diferencia observada. Esa magnitud o importancia es la que atañe a significatividad o importancia clínica. Esta se refiere a la importancia de una diferencia en los resultados clínicos entre los pacientes tratados y los controles.(7,13). Pero ambos conceptos están relacionados profundamente. La significatividad estadística es una condición previa indispensable para la significatividad clínica pero no indica nada acerca de la magnitud real del efecto evaluado (3)

P exacto o P menor que

Existe una confusión generalizada entre el valor P y el nivel de significancia α adoptado, o sea el error de tipo I, la preselección de un nivel de confianza α tiene sus raíces en la filosofía de que debe controlarse el riesgo máximo de cometer un error de tipo I.(6)

Es necesario recordar que P es la probabilidad de haber obtenido el valor observado cuando la hipótesis nula es verdadera y por lo tanto surge después de trabajar y efectuar los cálculos con los datos obtenidos en la investigación. Estos valores de P deberían ser exactos, salvo en aquellos que resultan muy pequeños como por ejemplo 8.10^{-6} y que se indica entonces como $p<0,0001$, ya que los programas actuales de estadística están preparados para calcularlos (14,15). Además permitirían a los otros lectores del artículo, en base a los datos comprobar el resultado. Es más, un gran número de revistas científicas tales como Archivos Argentinos de Pediatría y Clinical Chemistry, entre otros, exigen que estos valores de P sean exactos (14,16).

En paralelismo con el desarrollo de nuevas metodologías estadísticas lo ha sido el asombroso desarrollo en la potencia y disponibilidad de facilidades en computación. Por ello Henderson (1993) dijo también que en el pasado era necesario calcular el valor P con tablas estadísticas, pero en la actualidad los programas estadísticos disponibles para las microcomputadoras pueden calcular el P exacto. ¿ Entonces por qué estos valores de P siguen expresándose en forma no exacta?(15).

Además el conocer el valor P exacto permitirá visualizar cuán cerca está el estadístico obtenido a partir de los datos del valor que corresponde al nivel α de significancia establecido. Si está el valor de P muy cerca de α nos está diciendo que la aceptación o rechazo está muy cerca del nivel de rareza seleccionado y por lo tanto es necesario seguir investigando, aumentando el tamaño de la muestra o cambiando el diseño del experimento. Si se encuentra lejos del nivel de rareza elegido está diciendo cuán fuerte es la evidencia que se obtuvo para aceptar o no la hipótesis nula.

Estos valores de P deben ir acompañados en los casos que se comparan cantidades como medias, de los correspondientes intervalos de confianza de la diferencia, así se demuestra explícitamente la magnitud de la incertidumbre y su dirección tal como Neyman (1934) lo propuso como una alternativa válida para su modelo de contraste de hipótesis (10). Esta afirmación es compartida en la actualidad por editores de revistas químicas, bioquímicas y médicas en general, quienes aconsejan además a los autores de publicaciones no confiar únicamente en el uso del valor P(1,2,12,15).

En realidad el valor P como lo demostró Schervish (1996) es una función continua de la hipótesis para datos fijos y sostuvo que el uso sólo de este valor como medida del soporte o evidencia de una hipótesis tiene serias fallas lógicas.(17)

Conclusión

Es importante que cada investigador establezca el nivel de significancia con el que trabajó, para luego realizar la comparación del mismo con el valor P que obtenga a partir de los datos experimenta-

les y se remarca que es conveniente que este valor sea exacto, salvo cuando es muy pequeño, en cuyo caso se da una cota para el mismo. Además de preocuparse por elegir un buen test de hipótesis para cada experiencia, la tendencia actual es que el investigador elija un buen método de conclusión basado en P.

También se remarca la importancia de acompañar, cuando esto es posible, los valores estimados con sus intervalos de confianza para mostrar explícitamente la magnitud y dirección de la incertidumbre.

Agradecimientos

A la Dra. Ma. Cristina Lurá por sus agudas observaciones que condujeron a la corrección de este manuscrito, a la Bioq. Beatriz Abramovich por su apoyo y aportes y a la Dra. Marta Zanelli, (Ph. D. en Biometría) por sus comentarios y los artículos específicos que nos ha hecho llegar.

Bibliografía

1- Glantz, S. A. 1994. Todo está en los cifras. JACC (J. Am. College Cardiology). **3**, 6: 308-310.

2- International Committee of Medical Journal Editors. 1997. Uniform Requirements for Manuscripts Submitted to Biomedical Journals. New English Journal of Medicine. **336**, 4:309-316.

3- Norman y Streiner 1996. «Bioestadística». Mosby/Doyma Libros. División Iberoamericana. 41-45.

4- Ludbrook, J., Dudley, H. 1998. Why permutation tests are superior to t and F tests in Biomedical Research. The American Statistician. **52**,2: 127-132.

5- Steel, R.G.D., Torrie, J.H. 1995. «Bioestadística: Principios y Procedimientos». ed McGraw Hill. 83 -117.

6- Walpole,R.E.,Myers, R.H. 1992. «Probabilidad y estadística». Ed. Mc. Graw Hill. 299 - 315.

7- Altman, D. G. 1997. «Practical Statistics for Medical Research». Chapman & Hall. (London). 8ª ed.167 –170.

8- Bickel, P.J., Doksum, K.A. 1977. «Mathematical Statistics. Basic Ideas and Selected Topics». Prentice-Hall, Inc.

9- Lindgren, B. W. 1976. «Statistical Theory». Macmillan Publishing Co., Inc. (New York). 3rd ed.

10- Hogg, R.V., Craig, A.T. 1967. «Introduction to Mathematical Statistics». The Macmillan Company, (New York). 2da ed, 6ta reimpresión.

11- Mood, A.M., Graybill, F.A., Boes, D.C. 1974. «Introduction to the Theory of Statistics». McGraw-Hill Series in Probabilty and Statistics. 3rd ed.

12- Altman, D.G. & Machin, D. 1993. Current statistical issues in clinical cancer research Br. J.Cancer, **68**: 455-456.

13- Sackett, D.L., Haynes, R.B., Guyatti, G.H., Tugwell, P. 1991. «Epidemiología clínica. Ciencia básica para la medicina clínica». 2da ed. Ed. Panamericana.191-248.

14- Harris, E. 1993. On P value and Confidence Intervals (Why Can't We P with More Confidence?). Clin. Chem. **39**, 6: 927-928.

15- Henderson, R. 1993. Chemistry with Confidence: Shoul Clinical Chemistry Require Confidence Intervals for Analytical and Other Data?. Clin. Chem. **39**, 6: 929-935 .

16- Reglamento de Publicaciones. Archivos Argentinos de Pediatría.

17- Schervish, M.J. 1996. P values: What they are and What are not. The American Statistician. **50**, 3: 203-206.