V.10 N.1: 45-52 ISSN-2215 2741

Revista Clinica -HSJD-

EM Escuela de Medicina

TEMA 7-2020:

Valor P.
Correcta e incorrecta interpretación.

Recibido: 16/11/2018
Aceptado: 13/01/2020

[1] Lorenzo Marin Barboza
[2] Danny Paredes Rodríguez

[1] Médico General. lmb.eip@gmail.com
[2] Médico especialista en medicina crítica y terapia intensiva. Hospital Dr. Escalante Pradilla. dparedes03@hotmail.com

Resumen

Actualmente, gracias al predominio de la escuela frecuentista, para realizar inferencia a partir de los resultados de ensayos clínicos, se utiliza principalmente el valor P y la probabilidad de error alfa. Estos a pesar de ser herramientas estadísticas útiles e importantes, tienen limitaciones, las cuales no son ampliamente comprendidas, lo que lleva a confusión de su significado y de la utilidad de estas herramientas. A la hora de realizar investigación clínica, se debe conocer bien el problema en cuestión, los resultados y las herramientas estadísticas óptimas a utilizar, para obtener conclusiones lo más correctas posible. La preferencia de unos métodos estadísticos sobre otros, genera una visión e interpretación limitada de la evidencia.

Palabras claves

Medicina basada evidencia; estadística; investigación; valor p; intervalo de confianza; Bayes; Fisher; Neyman; inferencia.

Abstract

Thanks to the predominance of the frequentist school, nowadays, inference from the data obtained in clinical trials, is mainly done by using the P value and the alpha probability of error. These two, despite being useful and important statistical tools, have limitations that are not widely understood, leading to confusion about its meaning and utility. When doing clinical research, there must be good comprehension of the problem at hand, the outcomes and the optimal statistical tools to employ, in order to get conclusions as true as possible. The preference of some statistical methods over another, generates a limited vision and interpretation of the evidence.

Key words

Evidence based medicine; statistics; research; p value; confidence interval; Bayes; Fisher; Neyman; inference.

Introducción

La ciencia médica al igual que otras disciplinas, ha evolucionando de la mano de ciencias más elementales y básicas, dentro de estas ciencias básicas, una particularmente importante es la estadística, ya que brinda a los profesionales de la salud, herramientas para analizar de forma crítica y obtener conocimiento de la amplia y creciente literatura médica, y en concordancia con esto, tomar las mejores decisiones en su práctica profesional. Pero la estadística, no es una ciencia "absoluta"; dentro de su comunidad existe controversia entre los Frecuentistas y Bayesianos [1], escuelas de pensamiento que difieren en cuál método estadístico es mejor para inferir conocimiento. El valor P y la comprobación de hipótesis son elementos centrales en esta controversia; estos son ampliamente utilizados para realizar inferencia estadística en investigación clínica, pero lamentablemente son frecuentemente mal utilizados, mal comprendidos y sobrevalorados por la comunidad médica [2-6]; además de ser incompatibles. Irónicamente no son métodos basados en "evidencia" [3]

Discusión

El razonamiento inferencial, es el proceso por el cual ligamos las observaciones de la naturaleza con el conocimiento subyacente, hay de dos tipos: deductivo e inductivo. En la deductiva se formula una hipótesis y se predice las observaciones que se realizarían si esta hipótesis fuera cierta; en la inductiva se realiza el proceso inverso, es decir, a partir de nuestras observaciones evaluamos cuál hipótesis es más válida. La deducción nos ofrece más seguridad, ya que las predicciones serán ciertas siempre que las hipótesis sean correctas. La inducción permite que las conclusiones sean más amplias, pero no nos permite con seguridad saber si las conclusiones sean realmente ciertas. Para entenderlo mejor, podemos pensar en enumerar los síntomas que observaremos en un paciente con determinada enfermedad, siendo esto un proceso deductivo, pero es más difícil dar las probabilidades de diferentes enfermedades basándose en los signos y síntomas observados en un paciente, lo cual es inferencia. Esta inferencia médica se puede comparar con la inferencia estadística que se realiza a partir de los resultados u observaciones obtenidas de ensayos clínicos. El objetivo de la inferencia estadística, es estimar probabilidades sobre la relación entre los datos observados de una muestra y los valores (desconocidos) de la población total [4]. Filósofos como Karl Popper y Rudolf Carnap han abordado el problema de la inducción pero fracasaron, y esto mostró que no hay solución metodológica al problema del conocimiento científico falible [3]. Los Bayesianos afirman que este problema se había solucionado cuantitativamente hace más de 200 años con el teorema de Bayes [3], por otro lado, los frecuentistas realizan inferencia estadística con la prueba de significancia y de hipótesis.

En la medicina moderna, las intervenciones y medidas terapéuticas van en función del beneficio del paciente; la eficacia y efecto terapéutico de estas intervenciones deberían haber sido demostradas con evidencia, previo a la utilización en pacientes.

Una de las formas en que podemos obtener evidencia son los ensayos clínicos. En el contexto de un ensayo clínico, podemos pensar en el efecto terapéutico como una fuerza que "empuja" el grupo intervención (una nueva intervención o tratamiento) lejos del grupo control (un placebo o un tratamiento estándar ya comprobado), respecto a un resultado de interés [2]. Para medir esta fuerza terapéutica, se usan dos conceptos claves: magnitud y precisión. La magnitud es el tamaño del efecto producido, el cual se mide con herramientas cuantitativas como probabilidades de riesgo, riesgo relativo, riesgo absoluto, etc.

Figura 1

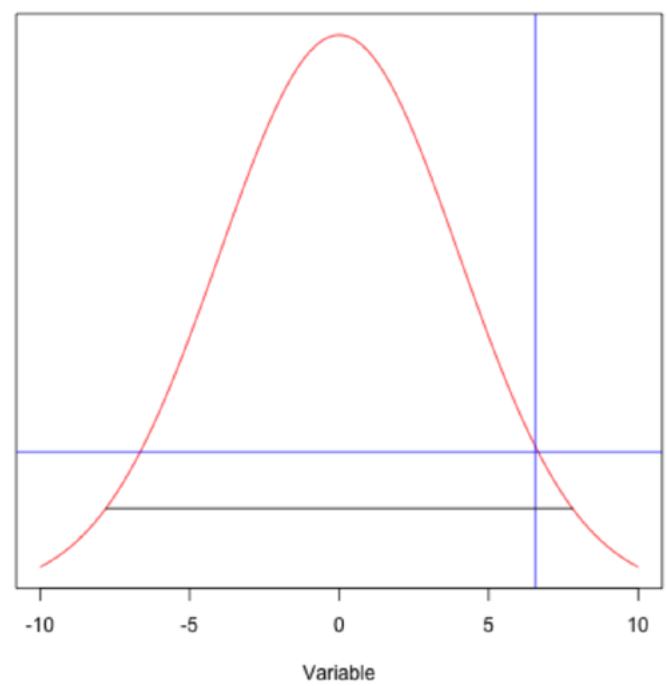

Figura 1.
Se ilustra la distribución normal (montaña simétrica) de una población. El eje X representa el valores de una variable hipotética; el eje Y representa la frecuencia de individuos en la población para cada punto en X; esta frecuencia se puede interpretar como probabilidad. La población tiene un promedio de 0 y pocos individuos con resultados a los extremos (poca probabilidad de que un individuo de la población tenga ese valor). La línea azul interseca un valor de X, donde la probabilidad de obtener ese valor es 0.05 en el eje Y, indicado por la línea azul. El valor en X con esta probabilidad de 0.05, depende de la distribución de la población. Entre más se aleje a la derecha esta línea en X, menor será esta probabilidad. La línea negra representa el intervalo de confianza de 95%, nos da una medida de la variabilidad (ancho de la montaña) y del tamaño del efecto. Los valores extremos en X del intervalo de confianza, indican en Y una probabilidad de 0.025. Esos extremos sumados son el 5% que se excluye cuando el intervalo de confianza es de 95%.
Fuente: diseño por el autor con RStudio. Version 1.0.143 – © 2009-2016 RStudio, Inc.

¿Que tan grande tiene que ser el efecto para ser relevante? Es un asunto de juicio clínico. La precisión es la cantidad de variabilidad o distribución de los datos. Entre menos variabilidad, más precisa se consideran las tomas. Sin embargo, en ensayos clínicos, esta variabilidad en un mismo individuo, tiene menos influencia cuantitativa, en la precisión de los resultados estimados, que la variabilidad paciente-paciente.

En investigación clínica cuando se habla de precisión del tamaño del efecto, se usa un concepto que mezcla ambas incertidumbres ya mencionadas (variabilidad "intra" y "entre" pacientes). La medida de precisión más utilizada en investigación clínica son los intervalos de confianza(IC) [2].

Además de esto existen dos tipos de incertidumbre; estas reflejan falta de información, una se puede cuantificar mediante matemática de probabilidades con información disponible, y para atender esta incertidumbre se derivaron la prueba de significancia y comprobación de hipótesis. Otra incertidumbre no cuantificable se presenta cuando no hay información previa relevante [2].

Técnicamente, el "valor P" (la "P" viene de Probabilidad [4], es el producto de la "prueba de significancia" [2], se define como la probabilidad de obtener un resultado igual o mayor, asumiendo que la hipótesis nula es cierta [3,4,7]. En los años 1920's [3], Sir Ronald Fisher, científico británico, en pequeños experimentos de agricultura, utilizó la idea de una hipótesis nula, la cual es la hipótesis "inversa" a la hipótesis que se busca comprobar (efectividad de la intervención); su propósito es reflejar cómo se distribuirían los datos resultantes si el tratamiento o intervención no tuvieran un efecto real [2].

En un mundo sin incertidumbre, el "no efecto" se registraría en los resultados como un efecto de magnitud 0 (grupo control con resultado idéntico al grupo intervención). Debido a que la incertidumbre es inherente al mundo natural, casi siempre hay una medición de efecto no-cero, incluso si la intervención no tuviera efectos biológicos del todo [2]; debido a que el azar impide que el resultado del grupo control sea idéntico al resultado del grupo intervención. Por lo tanto, el tamaño del efecto, tiene que ser "suficientemente grande" para que sobrepase, lo que se podría esperar del ruido estadístico subyacente de los datos [2]; a esto se le llama "estadísticamente significativo" (2,4), y se logra con un valor P menor o igual a 0.05 (Figura 1).

Para Fisher, el valor P reflejaba el grado en que los datos observados eran incompatibles con la hipótesis nula [7]. Si el valor P calculado, era menor que 0.05, Fisher propuso que, 1) un evento infrecuente había ocurrido o 2) la hipótesis nula (no efecto y no diferencia) era falsa [2]. Así pues, el valor P de Fisher, es una forma de medir que tan inesperado sería el resultado del grupo intervención, asumiendo que la hipótesis nula es correcta (representada por el grupo control) (Figura 2).

El criterio de menor a .05, fue una propuesta informal en sus primeros escritos, como un ejemplo y no como un estándar [2], lo que significa que fue una selección arbitraria, y "si hubiéramos evolucionado con 6 dedos tal vez sería 0.06" [4]; además Fisher pretendía que el valor P, se utilizara en un proceso fluido no cuantificable, de derivar conclusiones de los datos, en combinación con conocimiento previo [3].

Fisher, no postuló un gran número de hipotéticas repeticiones del experimento, para interpretar el valor P (importante para diferenciarlo de sus posteriores críticos); no obstante, Fisher afirmó que para aprender con certeza cómo utilizar la intervención (medida terapéutica) y lograr los resultados desea-

dos, se debían repetir, no de forma hipotética, los experimentos [2]. Otra condición implícita era que las muestras iban a ser tomadas al azar de una población infinita, lo que es una limitación del valor P, por la casi imposibilidad de tomar una muestra realmente aleatoria [2,4]. Lamentablemente en ciertas ocasiones, incluso la repetición de enormes ensayos clínicos, no aportan lo necesario para entender una terapia [2].

Como todo parámetro estadístico, el valor P puede resultar engañoso. Esto por dos problemas [2-4, 7] (Figura 3): 1)no nos dice nada sobre qué tan diferentes son los cohortes, ya que se enfoca en "la cola" de la distribución del grupo control, pero no nos habla sobre variabilidad de las poblaciones o magnitud del efecto (efectos grandes o pequeños pueden tener un valor P <0.05) y 2) es sensible al tamaño de la población, debido a que con una población suficientemente grande un efecto minúsculo puede arrojar un valor P "significativo", y con una población pequeña requiere un efecto muy grande para lograr un valor P <0.05.

Figura 2

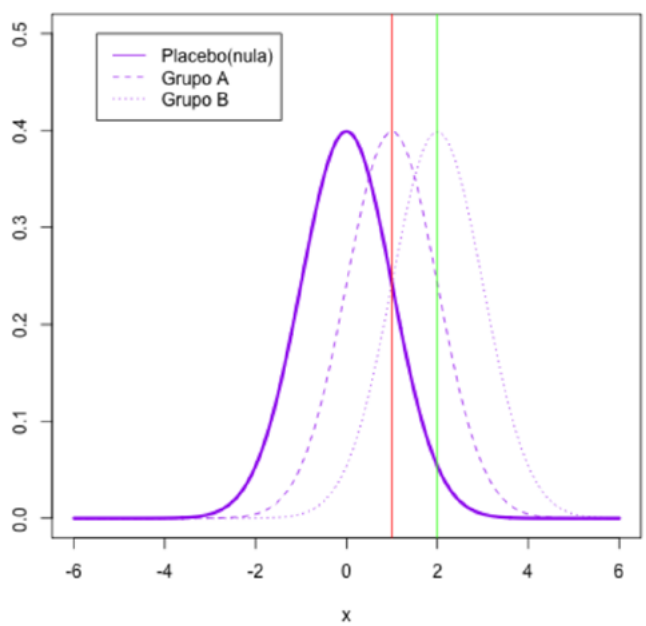

Figura 2.
Representación gráfica de la distribución de tres grupos de un ensayo clínico hipotético. El eje X representa la diferencia de los resultados de la variable medida (ejemplo: diferencia de mortalidad). El promedio del grupo placebo (hipótesis nula) se ajusta como cero, esto porque representa el control. La intervención A vemos que tuvo una diferencia de +1 y la intervención B de +2. Las líneas verticales nos ayudan a indicar el punto donde el resultado (promedio) de cada grupo, interseca la distribución normal de la hipótesis nula. Es evidente que el resultado del grupo A es más "compatible" con la hipótesis nula que el grupo B. El resultado del grupo B se presenta al extremo del grupo placebo, lo que indica que es, probabilísticamente, un resultado inesperado (asumiendo que la hipótesis nula es cierta). Nótese que el intervalo de confianza 95% de A, muy probablemente incluye valores iguales o menores que el promedio del placebo.
Tx: tratamiento.

Fuente: diseño por el autor con RStudio. Version 1.0.143 – © 2009-2016 RStudio, Inc.

Figura 3

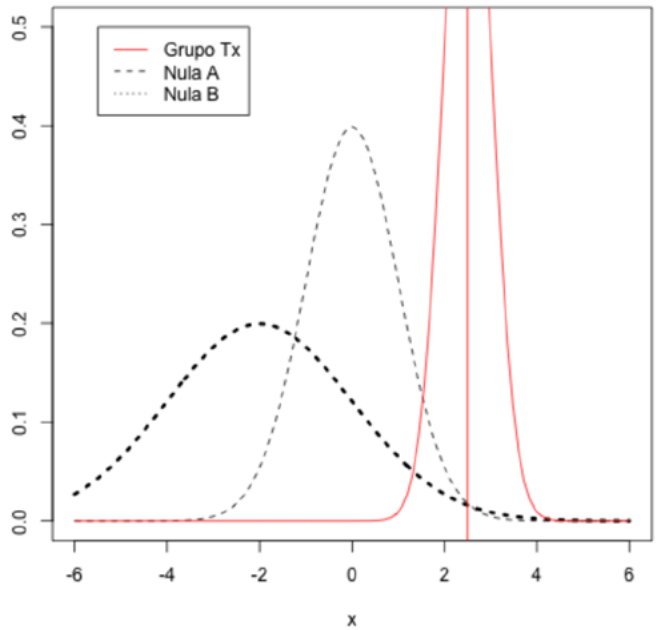

Figura 3.
Otro ejemplo hipotético de dos ensayos distintos, donde en uno se estudia Nula A vs Grupo Tx, y el otro se estudia Nula B vs Grupo Tx. Asumiendo que la línea vertical (promedio de Tx) interseca ambas nulas, en exactamente una probabilidad igual o menor a 0.05, podemos decir que el resultado del Tx es "significativo" para ambos grupos control. Pero ese 0.05, no nos dice nada sobre la variabilidad o precisión (ancho de la montaña) de ningún grupo, ni sobre el tamaño del efecto (diferencia de resultados). Si vemos los promedios de ambas nulas, podemos concluir que la diferencia o tamaño del efecto es distinto, siendo mayor en Nula A vs Tx.
Tx: tratamiento.

Fuente: diseño por el autor con RStudio. Version 1.0.143 – © 2009-2016 RStudio, Inc.

Es importante saber que el valor P reportado, es ajustado y no real; para obtener un valor P real se deben presentar las siguientes condiciones [4]: que el muestreo sea realmente aleatorio de una población infinita, que la distribución de las poblaciones sea gaussiana (normal) y que el promedio y la desviación estándar, sea como se indica; sin embargo, estas condiciones se pueden asumir o estimar y se usa aproximación con un modelo estadístico, esto aumenta validez al valor P, pero los modelos estadísticos son necesariamente imperfectos; es decir, el valor P es solo una estimación y no una probabilidad precisa[4].

Posteriormente, en los años 1930's [3], el matemático Jerzy Neyman y el estadístico Egon Pearson, se propusieron mejorar el trabajo de Fisher, de esto se origina la hipótesis alterna (cualquier hipótesis distinta a la nula) que junto con la hipótesis nula, se utilizan para definir los errores tipo I (Alfa, falsos positivos) y tipo II (beta, falsos negativos) [2,3]; prueba de hipótesis. Fisher creía que con su prueba de significancia se podía extrapolar, a partir de los experimentos, al mundo fuera de estos: inferencia inductiva; pero Neyman rechazó esta idea, y propuso que la "prueba de hipótesis" permitiría tener un "proceder inductivo", en lugar de hacer inferencia inductiva, lo que implica, que se podía aceptar la hipótesis nula o la alterna; es decir, no importaba lo que creía el científico acerca de la evidencia, el experimento debía decirle al científico que hacer [2,3]; esta decisión pone al investigador en riesgo de dos tipos de errores: aceptar la hipótesis alterna y que ambos tratamientos no son iguales cuando si son iguales (falso positivo), o aceptar la hipótesis nula y que ambos tratamientos son iguales cuando no son iguales (falso negativo) [3]. Neyman para explicar el concepto de prueba de hipótesis y tasas de error postularon que un experimento se repetiría (hipotéticamente) un número de veces cercano a infinito. Este concepto de repeticiones casi infinitas, no es un problema para los matemáticos, pero sí para los científicos al tratar de entender empíricamente qué significa. Neyman y Pearson nunca especificaron qué significaba esto en el contexto científico [2]. Dos factores se desprenden del concepto de repeticiones hipotéticas a largo plazo: 1) no se puede concluir nada respecto a si el experimento realizado fue correcto, y 2) el objetivo es mantener la tasa de errores en un mínimo aceptable [3]. La probabilidad de errores tipo I aceptables se define en la fase de diseño del ensayo, es es casi universalmente 0.05, pero no es lo mismo que el valor P [2,3]. Esta tasa de error alfa permite decidir cuál hipótesis aceptar como correcta, dependiendo de si las distancias, entre los resultados de ambos grupos, son suficientemente grandes: mayor o igual a 0.06 aceptar nula, menor o igual a 0.04 aceptar alterna [2]. Es decir, con 0.05, si realizamos 20 experimentos iguales, 1 va a ser un falso positivo, y no hay forma de identificarlo; los clínicos aceptaron esta idea sin objeción [2]. Este abordaje, es atractivo por la noción de medir, de forma objetiva las probabilidades de error, y se tenía la intención de usar estos, junto a elementos de juicio, en relación con la gravedad que implica un potencial error; pero este elemento de juicio ha desaparecido [3].

En años recientes, con la dominancia de la escuela frecuentista (la cual se ha descrito como "basada en error") [3], se ha utilizado el híbrido que mezcla ambos métodos, la prueba de hipótesis de Neyman y Pearson en la fase de diseño, y la prueba de significancia de Fisher en la fase de análisis; lo que ha llevado erróneamente a unificar, el valor P como la probabilidad de error alfa [2,3], para lograr con este "0.05" dos propósitos: obtener evidencia de cuanto los datos contradicen la hipótesis nula y obtener un parámetro objetivo de tasa de errores [3]. Entonces, a pesar de ser estos dos métodos incompatibles, se genera la idea equivocada de que un simple número puede determinar el valor de la evidencia y los resultados a largo plazo de un experimento, sin considerar la plausibilidad biológica o evidencia previa [3]. Su incompatibilidad lógica se concreta de la siguiente forma: la prueba de significancia (valor P), trata de evaluar las implicaciones de lo observado en un solo experimento, es a corto plazo y es inductivo; en la prueba de hipótesis (probabilidad de error alfa), agrupamos el resultado de nuestro experimento real, junto a los resultados de experimentos hipotéticos, es a largo plazo y es deductivo. Si pudiéramos combinarlos, implicaría que se podría llegar a fines inductivos (extraer conclusiones científicas) con medios deductivos (cálculo objetivo de probabilidades) [3].

Siendo esto así, se generan dudas de porqué sigue siendo tan popular [2,3,7,10]. Además de sus consecuencias socioeconómicas en la industria farmacéutica y en la carrera de profesionales, se han planteado varias explicaciones, por ejemplo que ofrece una herramienta de triage objetivo para decidir lo que es seguro de ignorar y a lo que hay que poner atención [2]. La academia médica, revistas y el gobierno lo encon-

traron útil como parámetro "objetivo", para obtener conclusiones "científicas", tomar decisiones y cambiar normas [3,7]. Otros candidatos son [10]: la editorial de una revista que requiera aseveraciones de resultados estadísticamente significativos, ansiedad de los investigadores en el cuidado de no mostrar asociaciones no significativas, la creencia de que un resultado significativo es de más peso, facilidad de describir los resultados con un lenguaje estadístico de significancia, ya que sin este lenguaje el investigador tendría que describir los resultados en sus propias palabras, la investigación frecuentemente arroja varios resultados y se decide reportar y señalar aquellos que son significativos, conveniencia en un meta-análisis de clasificar los estudios según significativos o no, desconocimiento de otros métodos estadísticos aparte de prueba de significancia y falta de entrenamiento estadístico.

Malas Interpretaciones

En la práctica, la forma más frecuente para pasar de evidencia a inferencia, es etiquetar el valor P como esperado o inesperado, en base a las expectativas y conocimiento previo: con un valor P=0.12 inesperado, se infiere que no hay diferencia entre tratamientos [3,5]; con un valor P=0.12 esperado, se infiere que hay una tendencia o se buscan explicaciones alternas como población pequeña; con un valor P=0.01 inesperado, se infiere un golpe de suerte por distractores no controlados o "dragado de datos"(obtención fortuita de resultados significativos); pero el peor y más frecuente es aceptar el veredicto de significancia como un indicador binario de la existencia de una relación o no [3]. Es pensamiento mágico creer que "significativo", se puede interpretar como "verdadero" y "no significativo", como "falso" [4]. Significativo, debería interpretarse, solo como válido de atención, en la forma de más experimentación [5].

Creer que un valor P de 0.05, implica que la hipótesis nula tiene una probabilidad de 5% de ser correcta es un error categórico, ya que el valor P es calculado bajo la suposición de que la hipótesis nula es correcta [3-5].

Otras interpretaciones incorrectas [3-4]: un valor P<0.05 es "estadísticamente significativo"; el valor P es la probabilidad de que los resultados del estudio sean debido al azar; el valor P es la probabilidad de que la hipótesis nula es cierta; estudios con valores P en lados opuestos del 0.05 están en conflicto; estudios con el mismo valor P proporcionan la misma evidencia contra la hipótesis nula; un valor P de 0.05 significa que si rechazo la hipótesis nula, solo hay 5% de probabilidad de cometer un error alfa; una conclusión científica o una terapia debe ser basada en si un valor P es significativo o no.

Soluciones Propuestas

Para practicar efectivamente medicina basada en evidencia, no debemos analizarlo todo con una misma herramienta o modelo; incluso el Comité Internacional de Editores de Revistas Médicas (ICMJE por su siglas en inglés), en sus recomendaciones de reporte, indica no depender solo en prueba de hipótesis y valores P, ya que no transmiten información importante sobre tamaño del efecto y estimados de precisión [13]. Se ha propuesto, por ejemplo, disminuir el valor P umbral de 0.05 a 0.005, lo que cambiaría, de significativo a "sugestivo", un tercio de la literatura previa [6]; otros, involucran darle atención al tamaño del efecto con los intervalos de confianza, los cuales ayudan a desprenderse de la automaticidad del valor P, estos ahora son más citados que en el pasado, frecuentemente son usados sólo como un sustituto de la prueba de hipótesis; pero estos no son la panacea y tampoco proveen un mecanismo para unir evidencia externa [3].

Cuando se reporta un resultado de un experimento, este por lo general es el promedio obtenido de los resultados de la población; el intervalo de confianza de 95%, me permite conocer el rango de resultados que se podrían obtener un 95% de las veces; si este rango es muy amplio, implica que hay mucha variabilidad en los resultados de la población, y si es angosto implica más precisión; a la hora de analizar un intervalo de confianza es importante evaluar si alguno de los extremos de este, cruza el promedio de la hipótesis nula, es decir, si el resultado del grupo control es un valor que se incluye dentro del rango del intervalo de confianza. El resultado que se reportó (un promedio), podría ser "engañoso", porque a pesar de que ese promedio es distinto al promedio de la hipótesis nula, la distribución de los resultados evidencia que hay individuos con igual o menor resultado que el de la hipótesis nula (Figura 2)[12].

Análisis Bayesiano [3,4,8,9,14]

El cálculo objetivo de la probabilidad de que un paciente tenga una enfermedad habiendo resultado positivo para un examen, a partir de la sensibilidad, especificidad y prevalencia de la enfermedad, se realiza por medio del cálculo de probabilidades condicionales utilizando el teorema de Bayes:

$$P(A \mid B) = P(B \mid A) \times (P(A)/P(B))$$

Donde P indica probabilidad, "|" indica unión condicional (ejemplo: "dado que"), A y B representan posibles condiciones; la fórmula se podría leer: la probabilidad de que A suceda dado que B sucedió, es igual a la probabilidad de que B suceda dado que A sucedió, multiplicado por la probabilidad de A dividido entre la probabilidad de B.

En términos simples, si un paciente dio positivo para una prueba altamente sensible, no podemos asegurar que el paciente tiene altas probabilidades de tener la enfermedad, debemos obtener el valor predictivo positivo, el cual se calcula con la prevalencia de la enfermedad; si la prevalencia de la enfermedad en la población es muy baja, las probabilidades de que el paciente tenga la enfermedad son bajas (probable falso positivo), ocurre lo contrario en el caso que la prevalencia sea alta (probable verdadero positivo). Vemos entonces como la evidencia y probabilidades previas influyen en las probabilidades posteriores.

Este análisis es frecuente verlo en interpretación de pruebas diagnósticas, pero no así en el análisis e interpretación de resultados de ensayos clínicos. Esto por varias razones, incluyendo la dificultad de incluir evidencia previa de forma objetiva, y se piensa que si se hace de forma subjetiva (de acuerdo a lo que el investigador "cree") no sería científico, además también la dificultad real de realizar y comprender este análisis bayesiano.

Debe notarse en la definición del valor P, su naturaleza condicional, "si la hipótesis nula es correcta"; esta connotación condicional hace que difiera enormemente de probabilidades ordinarias. Además se vuelve importante, la inclusión de evidencia previa, para que influya en la posibilidades posteriores, de que la evidencia que se obtuvo implica una relación real; un ejemplo es que un ensayo clínico que demuestra un beneficio de la homeopatía sobre un placebo con un valor P de 0.04; es evidente que si se incluyera la plausibilidad previa mediante evidencia de un mecanismo biológico o otros estudios, y bajo una análisis bayesiano, este resultado "estadísticamente significativo", no nos parecería tan "significativo". Por lo tanto un análisis bayesiano resulta una herramienta estadística útil.

Conclusiones

Es importante recordar que los médicos, no somos estadísticos de carrera, y el análisis estadístico es una tarea difícil de realizar, pero el avance de la medicina nos ha exigido una comprensión al menos básica para lograr navegar por la extensa literatura y poder practicar medicina basada en evidencia; no es de sorprenderse, que los comités de los ensayos clínicos, se forme de un equipo multidisciplinario, incluyendo estadísticos, para verificar los procesos [11]. El valor P y la prueba de hipótesis son herramientas importantes en la medicina basada en evidencia, pero tienen que usarse con cuidado y con un apropiado entendimiento [4]; los investigadores deben conocer las limitaciones de las estadísticas [14]. La apropiada inferencia requiere total transparencia respecto al reporte de los métodos y los resultados [7] y ningún simple índice debe debería sustituir el apropiado razonamiento científico [7].

Bibliografía

1. Bland JM Altman DG. Bayesians and frequentists. BMJ : British Medical Journal.1998;317(7166):1151-1160.
2. Mark DB, Lee KL, Harrell FE. Understanding the Role of P Values and Hypothesis Tests in Clinical Research. JAMA Cardiol. 2016;1(9):1048–1054.
3. Goodman S. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. Ann Intern Med. 1999;130:995-1004.
4. Cohen HW. P values: use and misuse in medical literature. Am J Hypertens. 2011 Jan;24(1):18-23.
5. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008 Jul;45(3):135-40.
6. Ioannidis JPA. The Proposal to Lower P Value Thresholds to .005. JAMA. 2018 Apr 10;319(14):1429-1430.
7. Wasserstein RL Lazar NA. The ASA's statement on P-values: context, process, and purpose. Am Stat. 2016;70(2):129-133.
8. Goodman S MD PhD. Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. Ann Intern Med. 1999;130:1005-1013.
9. Davidoff F. Standing statistics right side up. Ann Intern Med. 1999 Jun 15;130(12):1019-21.
10. Ahlbom A. Significance testing: Why does it prevail? Eur J Epidemiol. 2017 Jan;32(1):1-2.
11. Vaux DL. Research methods: Know when your numbers are significant. Nature. 2012 Dec 13;492(7428):180-1.
12. Gardner MJ Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing Br Med J (Clin Res Ed). 1986 Mar 15;292(6522) 746-50.
13. International Committee of Medical Journal Editors [www.icmje.org]. Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals – updated December 2016. http://www.ICMJE.org. Accesado en Julio 16, 2018.
14. Nuzzo R. Scientific Method: statistical errors. Nature. 2014; 506, 150-152.

Declaración de conflicto de intereses

Se declara que no hay conflicto de intereses en el presente reporte.

Diseño y diagramación
Bach. Evelyn Naranjo Madrigal
Diseño Gráfico - UCR
evelynnaranjo2@gmail.com