El elusivo valor de *p*: una aproximación intuitiva para el no-estadístico

Rufino Menchaca-Díaz*

Resumen

La inferencia estadística clásica se establece obteniendo el valor de *p* mediante el método del contraste de hipótesis, o bien estimando los intervalos de confianza de los resultados puntuales. El contraste de hipótesis permite comparar la probabilidad de ocurrencia esperada de un evento con la ocurrencia observada tomando en cuenta un margen de variabilidad debida al azar. De esta manera se acepta o se rechaza la hipótesis nula en base al resultado observado con cierto grado de confianza. En el presente ensayo se ofrece una aproximación intuitiva a la inferencia estadística basada en el método del contraste de hipótesis, usando para esto un ejemplo de un problema sencillo de probabilidad con distribución binomial, una analogía del contraste de hipótesis con un sistema de justicia y una explicación breve sobre la significancia estadística.

Palabras clave: Análisis de datos, técnicas de estimación, incertidumbre.

Summary

Classical statistical inference is mainly based in *p* value calculation using the hypothesis contrast method, or estimating confidence intervals for punctual results. Hypothesis contrast permits the comparability between observed probability and predicted probability considering some grade of random variability. The null hypothesis can be accepted or rejected using the observed result with a certain grade of confidence. In this assay, an intuitive approximation of the hypothesis contrast method for statistical inference is presented, using a simple binomial problem for probability calculation, an analogy between the hypothesis contrast and a justice system, and a brief explanation about statistical significance.

Key words: Data analysis, estimation techniques, uncertainty.

INTRODUCCIÓN

Sir William Osler (1849-1919) señaló que *"la medicina es la ciencia de la incertidumbre y el arte de la probabilidad"*.[1] No obstante, casi un siglo después, es posible considerar que el estudio sistemático de la probabilidad ha acrecentado la parte científica de la medicina reduciendo la incertidumbre. Por consiguiente, aceptemos como una consideración más adecuada para nuestra época, la frase contemporánea de Salvador Pita: *"la medicina es una ciencia de probabilidad y el arte de manejar la incertidumbre"*.[2] La medicina no es una ciencia exacta, es una ciencia fáctica o de hechos que se basa en el conocimiento de las probabilidades de ocurrencia de los mismos. Los médicos, casi sin darnos cuenta, enfrentamos el reto de integrar las probabilidades de que haya sucedido o pueda suceder un evento. Así, empleamos las maniobras clínicas que pudieran identificar con mayor probabilidad la presencia de una disfunción determinada; seleccionamos las pruebas diagnósticas que más probablemente pudieran ayudarnos a establecer el diagnóstico; integramos (con los dos anteriores) el diagnóstico más probable; aconsejamos el método de tratamiento que pueda tener más probabilidad de éxito; establecemos el tiempo más probable de sobrevida ante una enfermedad; o proponemos las medidas de prevención que pudieran (probablemente)

* Neurólogo Hospital Ángeles Tijuana. Profesor de Neurociencias, Epidemiología y Bioestadística Facultad de Medicina y Psicología Universidad Autónoma de Baja California.

Correspondencia:
Rufino Menchaca-Díaz, M.S.P; D.C.
Correo electrónico: rufino@uabc.edu.mx

Aceptado: 24-01-2012.

ser las más efectivas. No podemos evitar, ante cualquiera de estas situaciones, un cierto margen de error. La forma como enfrentamos esa incertidumbre restante, esa posibilidad de errar, constituye el arte de ser médico, y es particular para cada clínico. Está forjado a su vez por el conocimiento, la empatía, el interés humanitario y por el aprendizaje de los éxitos y de los errores, que sólo se adquiere a través de la experiencia.

Dado que nuestra práctica profesional se basa, eminentemente, en el conocimiento de las probabilidades de ocurrencia de los eventos de salud y enfermedad, ya sea en las personas o en las poblaciones, estamos obligados a entender, al menos de manera intuitiva, cómo se estudian los eventos probabilísticos. Dos formas utilizadas muy frecuentemente en los artículos de investigación para el estudio de la probabilidad de ocurrencia de los eventos de interés son: 1) utilizando el método del contraste de hipótesis para establecer el valor de *p* y 2) calculando los intervalos de confianza de los estimadores. En este ensayo trataremos de hacer una aproximación al primero de estos dos métodos de la inferencia estadística clásica.

EL CONTRASTE DE HIPÓTESIS Y EL VALOR DE *P*

En casi todos los artículos de investigación original encontramos mención del valor de *p*. Sabemos que un valor de *p* menor a 0.05 nos indica que el resultado es estadísticamente significativo, es decir, que los resultados observados tienen una probabilidad muy baja de ser producto del azar. El valor de *p* es, por lo tanto, una manera de controlar el efecto del azar. Llevemos esto a un terreno más conocido para nosotros, simples médicos *anaritmetas* parciales y completos *aritmofóbicos*. El ejemplo más conocido de un experimento aleatorio es lanzar la moneda al aire. Cuando lanzamos la moneda al aire, esperamos que el resultado observado se ajuste a las leyes de predicción compatibles con el experimento de la moneda, esto es, que nos aparezca una cara o una cruz (águila o sello, águila o sol, o como prefieran llamarles) de la moneda, ambas con una probabilidad de ocurrencia del 50% (0.5). El comportamiento de este experimento debe ajustarse siempre a la probabilidad predicha del 50%. Es decir, si lanzamos la moneda varias veces, esperamos observar una proporción de caras o de cruces cercanos al 50%. Pero no siempre se observará exactamente 50% de caras y 50% de cruces. Si lanzo la moneda 10 veces, no espero observar siempre 5 caras y 5 cruces como único resultado. Sé que por azar puede haber diferentes combinaciones de caras y cruces. Quizás 6 de una y 4 de la otra, o incluso 7 y 3, 8 y 2; o eventualmente obtener 9 caras de una y sólo 1 de la otra; también puede suceder que las 10 veces se obtenga una misma cara de la moneda y 0 de la otra; pero claro, esperamos que estos resultados más "extremos" ocurran de una forma mucho menos probable.

Si queremos probar que la moneda está en realidad "cargada" y que, por ende, favorece más a una cara que a la otra, deberíamos probar que el comportamiento de la moneda es diferente a lo esperado según nuestra regla de predicción compatible al 50%. Pero, ¿qué tan diferente?, ¿qué tanto debemos esperar como variación aleatoria razonable o permitida? y ¿cuándo debemos considerar el resultado observado como un comportamiento extremo que orienta a que la moneda está en realidad "cargada"? Para responder la pregunta anterior empleamos los siguientes pasos:

1) Planteamos la hipótesis general sobre el resultado que esperamos observar, si la moneda no ha sido alterada; es decir, esperamos que se comporte de acuerdo a una probabilidad de 50% de caras y 50% de cruces (aceptando cierta variabilidad permitida) y consideramos simultáneamente la otra posibilidad que interesa probar, a saber, que la moneda en realidad está "cargada" y, por lo tanto, favorecerá más a un desenlace determinado. En estadística esto corresponde a plantear la hipótesis nula (H0) y la hipótesis alterna (H1).
2) Establecemos el monto de variabilidad que vamos a aceptar como normal y cuando vamos a considerar un resultado como extremo, el cual nos indicaría que la moneda en realidad está "cargada". En estadística esto corresponde a establecer el nivel de significancia estadística o alfa (α), y por lo general son valores de 0.05 (5%) o 0.01 (1%).
3) Realizamos el experimento, aplicando al resultado una prueba estadística que permita establecer el nivel de probabilidad (valor de *p*), usando las reglas de predicción del 50-50% de la hipótesis general. En estadística se emplean diferentes pruebas para establecer el valor de *p* de acuerdo al tipo de datos que se analizan, usando de referencia la probabilidad inherente a la hipótesis nula.
4) Ocupando los criterios previamente mencionados, determinamos si el monto de la evidencia es compatible con un resultado esperado según la hipótesis general o por el contrario nos lleva a rechazar la hipótesis general y a aceptar la hipótesis alterna. En estadística un valor de *p* menor a alfa rechaza la hipótesis nula y permite aceptar la hipótesis alterna.

Continuando con el ejemplo de la moneda, si lanzamos la moneda 12 veces y observamos que aparecen 10 caras y sólo 2 cruces: ¿es la diferencia observada compatible con lo esperado por el azar?, o bien, ¿es la diferencia observada

suficiente para que concluyamos que la moneda está "cargada" y que, por ende, favorece más la aparición de caras? Hagamos nuestro análisis estadístico siguiendo los pasos arriba mencionados.

1) Establecemos H0 y H1.
   H0: La probabilidad ocurrencia de caras es igual a la probabilidad de ocurrencia cruces (Probabilidad 1 = Probabilidad 2; la moneda está "normal").
   H1: La probabilidad de ocurrencia de caras es diferente a la probabilidad de ocurrencia de cruces (Probabilidad 1 ≠ Probabilidad 2; la moneda está "cargada").
2) Establecemos el nivel de significancia que aceptaremos como evidencia suficiente para aceptar o rechazar la hipótesis nula. Usualmente un valor de alfa de 0.05 (5%).
3) Calculamos la probabilidad de haber observado específicamente 10 caras y sólo 2 cruces, si la probabilidad de ocurrencia de ambas fuese del 50%. Para lograr esto, usamos las reglas de probabilidad binomial, mediante la fórmula:[3]

$$P = \frac{n!}{(e!)(f!)}(p1)^e(p2)^f$$

   Donde *P* es la probabilidad que queremos calcular; *n* es el número de experimentos o veces que lanzamos la moneda, en este caso 12; *!* es el símbolo para factorial, en este caso 12*!* = 12 x 11 x 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1; *e* es la frecuencia de éxitos, en este caso 10 caras; *f* la frecuencia de fracasos, en este caso 2 cruces (la denominación éxito o fracaso es arbitraria y aplica para el estudio de la probabilidad binomial); *p1* es la probabilidad de éxitos, en este caso 0.5 (50% de caras) y *p2* es la probabilidad de fracasos, en este caso 0.5 (50% de cruces). Por lo tanto:

$$P = \frac{12!}{(10!)(2!)}(0.5)^{10}(0.5)^2$$

$$P = \frac{479{,}001{,}600}{(3{,}628{,}800)(2)}(0.00097)(0.25)$$

$$P = 0.016$$

4) En base a los criterios de significancia preestablecidos, el valor de *p* observado en este resultado es de 0.016, que es menor al 0.05 preestablecido como crítico. Nos permite rechazar la hipótesis nula y aceptar la alterna como más probable. La probabilidad de haber observado este resultado particular, en base a lo esperado por el azar, es de sólo 1.6%. Por tanto, el resultado es más compatible con la hipótesis alterna de que la moneda está "cargada".

Llevemos nuestro ejemplo a un terreno más clínico. Supongamos que se ha establecido internacionalmente que la prevalencia de demencia en adultos mayores de 80 años es de 30%. Sin embargo, en una muestra aleatoria de 15 sujetos mexicanos de este grupo de edad encontramos datos de demencia en 7 de ellos. ¿Es este resultado observado en sujetos mexicanos compatible con lo reportado a nivel internacional (H0)?, o bien, ¿es el resultado observado en sujetos mexicanos suficientemente distinto a lo reportado internacionalmente (H1)? Trata de calcular el valor de *p* para este resultado observado y usando un alfa menor a 0.05 como evidencia, acepta o rechaza la hipótesis nula según sea el caso. La respuesta correcta se muestra al final.

En estos ejemplos usamos la distribución de probabilidad binomial. Otras formas de analizar la probabilidad de un resultado específico es utilizando la distribución de probabilidades de la curva normal (prueba de z); la distribución de probabilidades de t (prueba t de Student); la distribución de probabilidades de F (prueba de ANOVA); la distribución de probabilidades de ji cuadrada (prueba de χ^2); o la distribución Poisson. El uso de cada una de ellas depende principalmente del tipo de datos que se analizan.[4,5]

EL CONTRASTE DE HIPÓTESIS: UNA ANALOGÍA

Podemos incurrir en un error al aceptar o rechazar la hipótesis nula, favoreciendo un resultado distinto a la realidad. Comparemos esto con lo que se observa en el sistema de justicia de muchos países:

1) Se parte de una premisa: el sujeto es inocente hasta que no se demuestre lo contrario (el equivalente a la hipótesis nula). El fiscal quiere demostrar que el sujeto es culpable (el equivalente a la hipótesis alterna).
2) El sistema judicial establece de antemano cuáles pruebas pueden ser catalogadas como evidencia (el equivalente a establecer el nivel de significancia o α).
3) Para poder inculpar al sospechoso, la evidencia que aporte el fiscal debe ser tan sólida que deje poco espacio a la duda (el equivalente a realizar una prueba estadística y calcular el valor de *p*).
4) Se establece el veredicto de inocencia o culpabilidad (el equivalente a aceptar o rechazar la hipótesis nula).

En el sistema de justicia mencionado se puede incurrir en dos tipos de error: cuando se establece un veredicto de culpabilidad en un inocente, o cuando se establece

un veredicto de inocencia en un culpable. En el *cuadro I* se puede observar esquemáticamente este escenario.

De los dos tipos de error en los que podemos incurrir, se establece como peor situación el encontrar culpable a un inocente, ya que se está generando un daño al castigar a quien no lo merece. En investigación también podemos incurrir en errores al analizar los resultados. En el *cuadro II* se señalan los posibles errores en los que se puede incurrir al analizar los resultados de un ensayo clínico para establecer la eficacia de un tratamiento.

En ambos casos incurrimos en un error más grave, si aceptamos la hipótesis alterna cuando la hipótesis nula es la verdadera: en el juicio, donde el veredicto de la persona es culpable cuando en realidad es inocente; o en el caso del ensayo, donde el tratamiento es efectivo cuando en realidad no sirve. Este tipo de error se denomina error tipo 1. En estadística se permite por lo general una probabilidad menor a 5%, o en ocasiones menor a 1% de incurrir en este error, esto es, un valor de *p* menor a 0.05 o menor a 0.01.

El otro tipo de error, conocido como error tipo 2, se comete al aceptar la hipótesis nula cuando en realidad la hipótesis alterna es la verdadera. En nuestros ejemplos, cuando encontramos inocente a un culpable, o si hallamos en un ensayo que un tratamiento no es efectivo cuando en realidad sí lo es.

Para controlar el error tipo 1 en un juicio, se considera sólo la evidencia más sólida; en un ensayo clínico, reduciendo la probabilidad del azar con un valor crítico de alfa menor a 5% o menor a 1%. El error tipo 2 se controla en el juicio aportando más evidencia; en el ensayo clínico, aumentando el tamaño de la muestra.

Cuadro I. Tipos de error en un juicio.

		En realidad es: Culpable	En realidad es: Inocente
Evidencia en el juicio:	Culpable	**Acierto**	**Error (I)**
	Inocente	**Error (II)**	**Acierto**

Cuadro II. Tipos de error en un ensayo.

		En realidad el tratamiento es: Efectivo	En realidad el tratamiento es: No efectivo
Evidencia en el ensayo:	Efectivo	**Acierto**	**Error (I)**
	No efectivo	**Error (II)**	**Acierto**

EL NIVEL DE SIGNIFICANCIA ESTADÍSTICA O ALFA

Sir Ronald Fisher (1890-1962), uno de los más grandes estadísticos de todos los tiempos, empieza su libro *Diseño de Experimentos* diciendo: *"Una mujer declara que tan sólo probando una taza de té con leche, ella puede decir cuál de éstos fue puesto primero en la taza"*.[6] Esta es una anécdota que en realidad le sucedió a Fisher cuando invitó a Miss Buriel Bristol una taza de té con leche y ella la rechazó pues la leche había sido agregada al final. En su libro, Fisher propone cómo validar la opinión de la experta en té y saber si realmente ella logra identificar, con sólo probarlo, qué se sirvió primero, el té o la leche. Para convencer a Fisher, la experta debería identificar sin error, ocho tazas de té, en las que, en 4 se sirvió primero el té y en otras 4 primero la leche. Existen 70 posibles combinaciones diferentes si se mezclan estas 8 tazas. Si la experta lograba identificar las 8 tazas correctamente, la probabilidad de que hubiese acertado sólo por azar era de apenas 1.4% (según análisis de probabilidad χ^2 usando la prueba exacta de Fisher), evidencia que el autor consideraba como suficiente para demostrar que el resultado observado no era producto del azar, sino que la experta lograba realmente diferenciar acertadamente las tazas de té. De hecho, Fisher consideraba ya como nivel crítico aceptable para establecer la significancia estadística, una probabilidad menor al 5% (*p* menor a 0.05). En este ejemplo observamos ya los principios del contraste de hipótesis, análisis que fue refinado posteriormente por otros autores. Actualmente se acepta internacionalmente como evidencia significativa un valor de *p* menor a 0.05. Los valores de *p* menores a 0.01 también son usados con frecuencia.

El concepto de significancia estadística que deriva del contraste de hipótesis pone de manifiesto la necesidad de rechazar o aceptar una hipótesis general de la cual se parte para contrastar los resultados observados en un estudio o experimento. Rechazar la hipótesis nula para aceptar la hipótesis alterna recae en un concepto puramente probabilístico, pues aún los eventos raros pueden

ocurrir por azar. En consecuencia, una aceptación de una hipótesis (nula o alterna) no significa verificación concluyente de la realidad, sino una aproximación provisional, que podrá ser refutada o verificada en experimentos posteriores.

El valor de *p* es una herramienta que facilita la comprensión de los fenómenos probabilísticos, disminuyendo la incertidumbre y permitiendo una interpretación más apropiada de los resultados observados en un estudio. Los médicos, en general, mantenemos poca relación con operaciones matemáticas o probabilísticas y por ello sufrimos en ocasiones de cierta dificultad para la interpretación de los conceptos de la estadística inferencial. Esperamos que esta aproximación, muy intuitiva, al contraste de hipótesis sirva para aclarar algunos conceptos básicos y ayude a los profesionales de la salud a comprender mejor el elusivo valor de *p*.

RESPUESTA AL EJERCICIO:

$$P = \frac{15!}{(7!)(8!)}(0.3)^7(0.7)^8$$

$p = 0.0811$, mayor a 0.05; por tanto, se acepta la hipótesis nula.

REFERENCIAS

1. Silverman ME, Murray TJ, Bryan CS, eds. *The quotable osler*. Philadelphia Pa, USA: American College of Physicians 2002.
2. Pita-Fernández S, Pértega-Díaz S. Pruebas diagnósticas: sensibilidad y especificidad. *Cad Aten Primaria* 2003; 10: 120-124.
3. Jaisingh L. *Statistics for the utterly confused*. 2 ed. New York, N.Y.: McGraw-Hill; 2006.
4. Altman DG. *Practical statistics for medical research*. London, UK: Chapman & Hall; 1991.
5. Pagano M, Gauvreau K. *Fundamentos de bioestadística*. México: Thomson Learning, Inc.; 2001.
6. Fisher RA. Design of experiments. London, UK: Macmillan Publishers Co; 1935.

Reconocimiento a revisores

Agradecemos a los revisores de los trabajos enviados a Acta Médica que, además de los miembros del Comité Editorial, nos favorecieron con su labor durante 2011.

Dra. Ma. de Lourdes Basurto Acevedo
Unidad de Investigación Médica en Endocrinología
IMSS

Dr. Tomás Barrientos Fortes
Ángeles Lomas

Dr. Raúl Caltenco
Ángeles Lomas

Dr. Javier Vilchis Licona
Ángeles Mocel