

Notas metodológicas.

¿Estadísticamente significativo o clínicamente relevante?

05/09/2006

Martín Caicoya Gómez-Morán

Jefe de Servicio de Prevención de Riesgos Laborales del Principado de Asturias. Hospital Monte Naranco

El apellido Fisher está unido a la genialidad y a la extravagancia, una extravagancia repulsiva. Bobby Fisher (o Fischer como también se escribe) es un brillante ajedrecista y un desagradable ciudadano. Pero el Fisher que más influencia ha tenido en la historia del saber es un inglés: Ronald A Fisher (1890-1962).

Muy pronto se interesó por la genética desde la perspectiva estadística, abandonando la idea de convertirse en un biólogo. Le habían deslumbrado las obras de Francis Galton, cuya teoría saltacionista, que postula que la evolución ocurre a saltos entre periodos de larga quiescencia, tiene alguna verosimilitud a la luz del saber actual. Nuestro Fisher contribuyó muy pronto al saber con su "The Genetical Theory of Natural Selection" publicado por primera vez en 1930. Es hoy, todavía, un texto imprescindible para los que se quieren adentrar en esta disciplina.

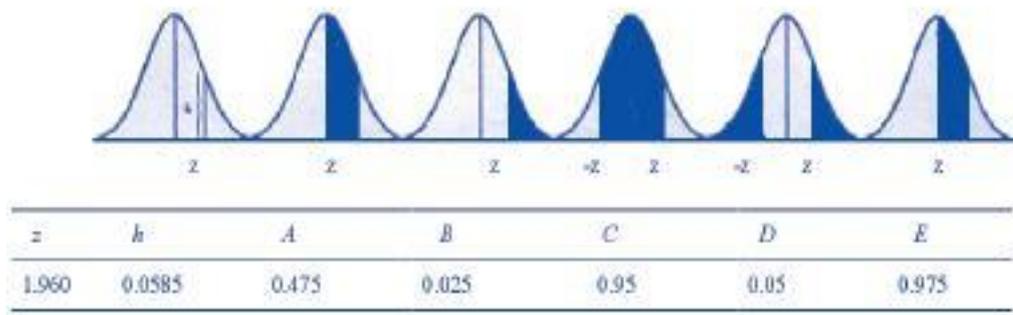
Pero sus teorías, la interpretación de sus sofisticadas matemáticas, le llevaron a ser uno de los líderes más sólidos del movimiento eugenésico que impregnó el mundo occidental antes de la 2ª gran guerra y que sentó las bases para el exterminio de las llamadas razas inferiores. Baste saber que en los muy admirados y liberales países escandinavos sobrevivieron las leyes eugenésicas hasta mediados de siglo pasado. En ese afán desmedido por encontrar en los genes todas las razones y causas, Fisher murió en un exilio dorado en Australia afirmando que el cáncer de pulmón no estaba producido por el tabaco; sino que ciertos genes predisponían tanto a fumar como al cáncer.

Ya cuando Richard Doll y Austin Bradford Hill publicaron su artículo seminal sobre tabaco y cáncer de pulmón en 1951, Fisher criticó con dureza los resultados descalificando todo estudio que no fuera experimental y bien controlado. Fue precisamente Fisher el que ayudó a establecer el criterio de la $p \leq 0,05$ que tanta trascendencia tiene. Recordemos sus palabras: *es conveniente trazar una línea al nivel aproximado en el que uno puede decir: "Bien hay algo en el tratamiento, o una coincidencia como esta no ocurre más de una cada veinte veces.."*

Si una en veinte no parece suficiente rareza, se puede trazar la línea en una en 50 (2%) o una en 100 (el punto 1 por ciento). Personalmente, el escritor prefiere sentar un estándar bajo de significación al 5 por ciento, e ignorar enteramente todos los resultados que no alcanzan ese nivel

Se está refiriendo a un estudio experimental (él sentó su diseño) en el que un grupo de sujetos (en su caso, semillas) reciben un tratamiento y otro no reciben nada. Midió el valor de cada planta al final del estudio y comparó la media entre los dos grupos. A Fisher también debemos ese estadístico que llamamos **z** que nos sirve para calcular, con facilidad, la probabilidad de la que habla en el texto transcrito. También fue el primero que hizo tablas de probabilidad que utilizamos para, una vez calculada la **z**, saber que probabilidad hay de que ocurra una diferencia como la encontrada, ver figura 1.

Figura 1



La idea de la z es brillante por su utilidad, claridad y simplicidad. Veamos en que se basa. Si la media experimental es M y la control μ la z es:

$$z = \frac{M - \mu}{\sigma} \{1\}.$$

Es decir la diferencia de medias ponderada por σ , la desviación típica (DT).

La idea es brillante porque ocurre que si las medias son idénticas ($M = \mu$), z cobra el valor de 0, porque $M - \mu = 0$ y $0 / \sigma = 0$. Si M vale tanto como μ más la DT, z será 1 ($\mu + \sigma - \mu / \sigma = 1$). Y si M vale $\mu + 2 \sigma$ z será 2. Y así sucesivamente.

Pero ocurre que la media más 3,80 veces la DT y la media menos esa cantidad abarcan el 99,99% de todos los valores, en una distribución normal. Traducido a nuestro estadístico z, valores de z entre +/-3,80

abarcarán casi todos los valores que puede tomar. Este es el gran hallazgo. Antes el investigador tenía que manejar una fórmula engorrosa donde hay logaritmos neperianos y raíces cuadradas. Imagínese lo tedioso, y falible, que era hacer el cálculo. Fisher puso a disposición de los estudiosos unas tablas que caben en unas hojas que puede Vd. encontrar en cualquier libro de estadística.

A la hora de calcular el valor de la z hay una salvedad. No se puede utilizar la desviación típica que tiene la muestra: esa nos habla de la variabilidad de los datos de esa muestra. Hay que distinguir entre DT de la media de la muestra, que es la que conocemos, y DT de la media de las medias, que es sobre la que trabajamos. Recuerde que comparamos medias, no valores individuales. Esto es importante, y muchas veces confuso.

Efectivamente, en un experimento no se comparan valores individuales, la tensión de tal o cual sujeto, sino medias: la media de la tensión arterial del grupo control con la del grupo experimento. Nuestra hipótesis es que son diferentes (si es que los son) gracias al tratamiento. Evidentemente si una vale 155 y otra 170, lo son. Pero la probabilidad introduce la incertidumbre y su ponderación.

La media encontrada (170 en hipertensos), decimos que es el valor medio encontrado en una muestra al medir en cada individuo la tensión arterial. Si tomáramos, a esa misma población, que llamamos de hipertensos, otra muestra, con idéntica técnica de muestreo, fácilmente obtendríamos otra media. ¿Cómo sé yo que la media no iba a ser 155, o incluso inferior? La respuesta es la probabilidad.

Pero, ¿cómo se va a calcular una probabilidad con un solo valor? Se puede. Esta es la potencia y belleza de la estadística. Ocurre, y esto es un teorema, que con una media de una muestra, y su desviación típica, yo puedo saber la distribución de probabilidad de todas las posibles medias de las infinitas muestras que se tomaran a esa población. Esta es una de las bases de la estadística.

El teorema central del límite fue por primera vez postulado por Laplace, basado en los estudios de su compatriota de Moivre. Lo que dice es que si una variable X toma numerosos valores, independientes entre sí, y todos ellos siguen el mismo modelo de distribución (cualquiera que éste sea), se puede elaborar una variable z que se distribuye normalmente con media la encontrada y desviación típica la de la muestra dividida por la raíz cuadrada del tamaño de la muestra. En concreto es la siguiente:

$$Z = \frac{(X_1 + X_2 + \dots + X_n) - n\mu}{\sigma\sqrt{n}} = \frac{(\sum_{i=1}^n X_i) - n\mu}{\sigma\sqrt{n}}$$

{2}

Si en la fórmula 1 colocamos la DT de la media (es decir, la DT de la muestra dividida por la raíz cuadrada del tamaño de la muestra, que se denomina error estándar (EE), también se tendrá el estadístico z.

$$z = \frac{M - \mu}{\sigma/\sqrt{n}} \quad \{3\}$$

Este teorema se llama también el de los grandes números, porque necesita una muestra grande para que se verifique. Con carácter general, o al menos en los modelos de probabilidad clásicos, se admite una aproximación aceptable siempre que n sea mayor o igual que 30. En tamaños inferiores a 30 se usa la Distribución de Student.

Bien, ya tenemos una forma de hallar la probabilidad de un resultado así, de que ocurra una diferencia como la encontrada ¿Por qué si la probabilidad es baja, por ejemplo una $p=0,0001$, decimos que el tratamiento antihipertensivo tuvo un efecto? Aquí viene otro concepto estadístico. Una población se define por su media y su desviación típica (y su función de probabilidad, que en este caso es la normal). De manera que la población de hipertensos se definía por su media, 170 y la DT, la que sea. Ahora resulta que, después de tratar, la media es 155: ¿es por que ésta es una nueva población, precisamente definida por su media 155 y DT? Pues decimos que sí, con una probabilidad de equivocarnos del 1 por 10000. Y como, en este experimento, el resto de las variables permanecieron inalteradas, creemos que lo que modificó esa población para crear una nueva fue el tratamiento.

Gracias al Teorema Central del Límite y al estadístico z se ha podido hacer una evaluación estadística de manera sencilla.

Volvamos a la fórmula 1. Suponga que la z vale 1,96. En la figura 1, podrá ver que hay 2,5% probabilidades de encontrar valores así o mayores y 2,5% de encontrarlos así o menores (curva D). Si usted está seguro, antes de experimentar, que la intervención nunca hará que la media sea menor (cosa que el investigador de tratamientos antihipertensivos nunca podría decir), le basta el valor de la probabilidad superior. Podremos decir, con un error del 2,5%, que el tratamiento ha surtido efecto porque la muestra experimental no pertenece a la población control (curva B) Pero la mayoría de las veces uno no sabe si la intervención

hará mayor o menor la media, por eso se elige la p de 5% para el valor de z de 1,96. Se dice: una p de dos colas de 0,05.

Observe que sumando a la media la cantidad 1,96 y restándole esa misma cantidad se abarcan el 95% de todos los valores, curva C. En esa propiedad de la distribución normal se funda el intervalo de confianza: Tenemos una confianza (seguridad) del 95% de que el valor de la media de la población se encuentra entre esos dos valores. De ahí el nombre de intervalo de confianza. En definitiva, aunque no es de todo correcto, se puede decir que cuando el intervalo de confianza de la media experimental no incluye el valor de la media control es que son diferentes.

Fíjese en la fórmula {3} el valor tan importante que tiene el tamaño de la muestra. Cuanto mayor sea, más pequeño será el EE. Y si el divisor es pequeño, será mayor la z , que es lo que busca: cuanto mayor la z menos probabilidades hay de que las dos muestras sean iguales, como se puede comprobar en las tablas.

También la variabilidad de la muestra tiene importancia. A más dispersión de valores, más variabilidad, mayor será la DT. Cuan dispersa es una muestra depende de dos factores: de la propia variabilidad del carácter en la población y del tamaño de la muestra. Este segundo efecto se debe al fenómeno de regresión a la media: A medida que se incrementa la muestra cada vez hay más valores que se agrupan más en torno a la media, por tanto disminuye la variabilidad, es decir la DT.

De manera que para encontrar una diferencia significativa basta incrementar el tamaño de la muestra. A eso se llama incrementar el poder: la capacidad para encontrar diferencias.

Supongamos que ha tratado a un grupo de 20 hipertensos con el fármaco experimental A y a un grupo control de 20 hipertensos con el estándar B. El primer grupo tiene una tensión arterial media de 138 y el segundo de 139. La DT fue de 2,59, luego el EE (DT/\sqrt{n}) será de 0,82. Usted hace su test de hipótesis y encuentra que la p es de 0,22 pues la t vale de 1,22. Se emplean en este caso las tablas de Student porque la muestra es pequeña. Supongamos que la empresa farmacéutica, que paga el ensayo, dice que hay que demostrar que su fármaco es mejor: Doctor, tome una muestras de 400. Ahora la DT es menor, porque muchos más valores están próximos a la media: es de 0,7. Y el EE es de 0,06. La z , fue de 14,28, la $p=0,000$. Altamente significativo, pero, ¿clínicamente relevante?

La enseñanza de esto es que un experimento debe ser diseñado para encontrar una diferencia que sea clínicamente relevante. Si no puede encontrarla es que le falta poder, pero si encuentra diferencias irrelevantes le sobra, como es el caso descrito. Por eso, para calcular el tamaño de la muestra hay que partir de varios

supuestos y criterios. Estos son, el valor de la p de rechazo, generalmente 0,05, el valor del poder para encontrar diferencias, como las supuestas, si las hubiere, generalmente 80% o más, la diferencia que se busca (a partir de cuántos milímetros de mercurio A será mejor que B) y el error estándar.

Bibliografía:

1. Fisher RA. *Statistical Methods for Research Workers* 1925. London. Oliver & Boyd (p 504)

Palabras clave:

de la evidencia científica a las decisiones clínicas
teorema central del límite

Autores:

Martín Caicoya Gómez-Morán

Nº:6 de 2006