La significación es irrelevante y los P-values engañosos. ¿Qué hacer?

A. Blasco

Departamento de Ciencia Animal. Universidad Politécnica de Valencia.
P.O. Box 22012. Valencia 46071. Spain
ablasco@dca.upv.es

1. La significación es irrelevante

Con arreglo a la estadística clásica, un experimento debe estar diseñado para que aparezcan diferencias significativas a partir de una cierta diferencia que se considera relevante. Si el experimento está bien diseñado, se calcula además la potencia del test que se va a aplicar, y cuando no aparecen diferencias significativas es porque estas son menores que esa diferencia relevante con una determinada probabilidad. En experimentos bien diseñados "significativo" significa que hay diferencias y "n.s." (no significativo) que no hay diferencias entre tratamientos; si se ha diseñado el experimento sin calcular la potencia del test, como es frecuente, n.s, significa "no sé", no sé si hay o no diferencias. Esta forma de proceder presenta varios problemas:

1. ¿Qué ocurre con todos los caracteres medidos que no son el que se usó para diseñar el experimento? Aquí la significación puede aparecer cuando las diferencias son irrelevantes (lo que no sería un problema grave) o puede aparecer el temido n.s. cuando las diferencias sí lo son, con lo que nos vemos obligados a decir "no sé" cuando pudiera ser importante detectar diferencias.

2. Puede ocurrir algo peor, puede aparecer una diferencia pequeña, irrelevante, junto a un n.s., dando la falsa seguridad de que no hay en realidad diferencias entre tratamientos. Por ejemplo, una diferencia entre tratamientos de 0.1 lechones en tamaño de camada es obviamente irrelevante, pero puede ir acompañada de un intervalo de confianza [-1.5, 1.7], lo que implica que es posible que haya una diferencia relevante entre tratamientos y que no sabemos cuál de los dos tratamientos es el que provocaría esa importante diferencia en tamaño de camada.

3. Puede ocurrir algo mucho peor todavía, que sí que haya una diferencia significativa entre tratamientos y que esta sea relevante, pero que el intervalo de confianza incluya valores irrelevantes. Por ejemplo, podríamos encontrar una diferencia significativa de 1.1 lechones con un intervalo de confianza de [0.3, 1.9], lo que quiere decir que es perfectamente posible que esa diferencia sea irrelevante. Sin embargo, toda la discusión del artículo se basa usualmente en la diferencia de 1.1 y en que es significativa, e incluso se puede recomendar tomar en base a ello alguna decisión que puede ser catastrófica dado que no sabemos en realidad si la diferencia entre tratamientos es en realidad de 0.3 lechones.

Frecuentemente se tiene la falsa impresión de que es más probable que sea de 1.5 que de 0.1 lechones, pero esto no es así. Si repetimos infinitas veces un experimento tendremos infinitos intervalos de confianza, de los que el 95% contendrán al valor verdadero no sabemos dónde, a veces por el centro y a veces en un extremo. Nosotros afirmamos que nuestro intervalo es uno de los buenos esperando equivocarnos a lo largo de nuestra carrera un 5% de ocasiones como máximo, pero no sabemos si nuestro intervalo es de los que contiene el valor verdadero hacia el centro o no.

4. Finalmente pueden aparecer diferencias significativas meramente por azar. Cuando medimos muchos caracteres o muchos efectos, podrían aparecer diferencias significativas que en realidad no se corresponden con diferencias reales, puesto que se corre siempre un riesgo (habitualmente como máximo un 5% de las veces; esto es, una de cada veinte) de que esto ocurra. A veces aparecen en la literatura conmovedores intentos de explicar tal o cual interacción de las muchas que se han estimado, cuando esta interacción apareció como significativa meramente por azar.

Hasta aquí hemos hablado de experimentos bien diseñados, que son los menos. La realidad habitual es que:

1. No se presenta la potencia del test porque el experimento se diseñó sólo para que aparecieran diferencias significativas a partir de cierta cantidad. En ese caso la potencia es del 50%, lo que implica que si hubiera un valor relevante en la frontera de la significación, lo detectaríamos sólo la mitad de las veces (figura 1),

2. No se ha diseñado el experimento por ignorancia, por falta de medios (porque simplemente se utilizan los medios de los que se dispone) o porque no se tiene idea de qué diferencia se quiere detectar. Esto último ocurre cuando lo que se mide son caracteres cuya cuantificación no tiene una significación biológica o económica clara; por ejemplo los resultados de un panel de pruebas de calidad de carne o de una actividad enzimática. No resulta claro a partir de qué valor las diferencias son relevantes.

3. Con datos de campo, en los que con frecuencia hay muchos datos, aparecen diferencias significativas por todas partes sin que en realidad tengan estas en muchos casos relevancia alguna.

La pregunta está frecuentemente mal planteada. No es ¿Hay diferencias entre tratamientos? Para responder a esa pregunta no hace falta hacer el experimento; la respuesta es invariablemente: "sí, hay diferencias". Hay que recordar que todo es diferente en esta vida[1]. Dos razas de cerdos diferirán en 0.001 lechones de tamaño de camada o en 0.01 gramos de peso adulto, pero diferirán. El problema es si esas

[1] "Dicho sea de paso, decir de dos cosas que son idénticas es un sisnsentido, y decir de una que es idéntica consigo misma es no decir nada en absoluto". (L. Wittgenstein, *Tractatus* 5.5303).

diferencias son relevantes. Lo importante es conocer el intervalo de confianza de la diferencia entre tratamientos, puesto que podrían aparecer diferencias significativas con intervalos que contuvieran valores irrelevantes o diferencias no significativas con intervalos que incluyeran valores relevantes. Pero si esto es así, ¿para qué queremos la significación? De momento lo único que hace es contribuir a la confusión del lector poco avisado, y no añade nada a la información que proporciona un intervalo de confianza. ¿Quieres que aparezcan diferencias significativas? ¡Aumenta el tamaño de la muestra! ¿Quieres que no haya diferencias significativas? ¡Reduce el tamaño de muestra! La significación no parece muy útil para la discusión de gran parte de los resultados, francamente.

Otro de los inconvenientes de los test de hipótesis es que no cuantifican, dan como respuesta sólo SI o NO, y esto pueden dar lugar a paradojas. Por ejemplo, la diferencia en tamaño de camada entre las razas A y B puede ser n.s., entre las razas B y C puede ser también n.s., pero entre las razas A y C puede ser significativa.

Queda un uso perverso de la significación: la inclusión de efectos en un modelo de acuerdo a si son significativos o no. En numerosas ocasiones nos encontramos con la sentencia "Tras un análisis preliminar, se excluyeron del modelo los efectos que no fueron significativos". Pero un efecto puede tener una influencia notable sobre los datos y ser no significativo debido al tamaño muestral. Si el efecto existe realmente, convendría incluirlo sea o no significativo. El problema es que no sabemos, basándonos en la muestra, si hay o no hay efectos, sólo disponemos de nuestras estimaciones. Incluir efectos reduce la varianza del error y disminuye los grados de libertad. Si hay suficientes datos se pueden incluir efectos sin muchos problemas de ajuste, aunque habría que examinar cada caso viendo qué sucede al incluirlos o no. En la práctica lo mejor es incluir los efectos sobre los que haya motivos biológicos u otras razones para incluirlos, sean o no significativos. Decidir si se está sobreparametrizando un modelo no es fácil, aunque hay varios criterios que pueden ayudar (AIC, BIC, DIC, TIC, etc.), pero en la mayor parte de casos es irrelevante, particularmente si no se está interesado en el efecto sino en quitar ruidos de fondo.

Todos estos errores de interpretación están relacionados con la impresión de que el nivel de significación tiene algo que ver con las probabilidades de que la hipótesis nula sea cierta. De hecho el nivel de significación no tiene nada que ver con esta probabilidad. Como se pone antes de iniciar el experimento y es independiente del tamaño de la muestra, no puede indicar la probabilidad de rechazar la hipótesis nula si fuera cierta. Podemos poner un nivel de significación del 5%, y obtener mucha más evidencia de que la hipótesis nula es falsa. Lo que ocurre es que *no disponemos de ninguna "regla de medida" que nos indique la evidencia proporcionada por nuestra muestra. Cuando rechazamos la hipótesis nula lo hacemos siempre con una probabilidad del 100%*. Aceptar o rechazar una hipótesis nula se parece a una sentencia de un tribunal; culpable o inocente, no "inocente pero sólo un poco". Como este resultado es bastante pobre, si obtenemos una evidencia mucho mayor que el 5% es muy irritante conservar este nivel de significación, por lo que cuando nadie mira lo cambiamos al 1% pretendiendo que siempre pensamos que ese era el nivel máximo de equivocaciones a lo largo de nuestra carrera que estábamos dispuestos a tolerar.

Aunque presentar niveles de significación en función de los resultados obtenidos es estrictamente incorrecto, las revistas científicas no sólo no lo prohíben sino que lo fomentan recomendando, como el Journal of Animal Science, el uso de términos carentes de sentido como "muy significativo".

2. Los P-values son engañosos

Un P-value es la probabilidad de que aparezca un valor igual o superior a la diferencia que hemos encontrado, en el caso de que realmente no haya diferencias entre tratamientos. El uso que hace Fisher del P-value es bien claro: si este es bajo, digamos un 2%, o bien no es cierto que los tratamientos sean iguales o bien son iguales pero hemos obtenido una muestra excepcional en la que parece que difieran. Hasta aquí estamos todos de acuerdo. El problema es *cuánto* de excepcional es la muestra si nos sale un P-value del 2%.

El P-value tiene al menos dos interpretaciones incorrectas:

1. Es interpretado como la probabilidad de que no haya diferencias entre tratamientos. Esto es obviamente incorrecto, el P-value es la probabilidad de la muestra, no la probabilidad de la diferencia entre tratamientos. Lo que pasa es que lo que nos interesa realmente es la probabilidad de que los tratamientos no difieran. Como esto no es posible saberlo en el marco de la estadística clásica nos conformamos con la probabilidad de obtener otras muestras si repitiéramos el experimento (muestras por cierto que no hemos tomado ni vamos a tomar). Ya que no podemos ir al Amazonas, nos conformaremos viendo un documental.

2. Es interpretado como el nivel de significación para aceptar o rechazar la hipótesis nula. Esto es incorrecto, puesto que los niveles de significación se ponen *antes de realizar el experimento*, no dependiendo de cómo salgan las cosas: El P-value no puede dar el nivel de significación porque si repetimos el experimento el P-value cambia, y en estadística clásica las conclusiones se sacan no sólo de la muestra sino de las posibles repeticiones del experimento.

El problema no es sólo que el P-value no sea un indicador de la probabilidad de que la hipótesis nula sea cierta, sino que no está claro ni siquiera *cuánta evidencia* en contra de la hipótesis nula muestra. Supongamos que obtenemos un P-value del 5% y que el valor verdadero coincide con el de nuestra muestra, ¿qué ocurriría si repitiéramos el experimento? Como los valores muestrales se distribuirían en torno al valor verdadero, en la mitad de las ocasiones nos saldrían resultados "no significativos" (figura 1). Por supuesto que P-values de 0.00001 indican mucha evidencia de que la hipótesis nula es falsa, pero la estadística no se creó para cuando se tienen muchos datos sino para distinguir los efectos reales existentes de los procesos de mero azar, y es en la frontera de la significación donde la estadística es particularmente útil, en la mayoría de los otros casos los problemas no son de estadística sino de cálculo numérico.

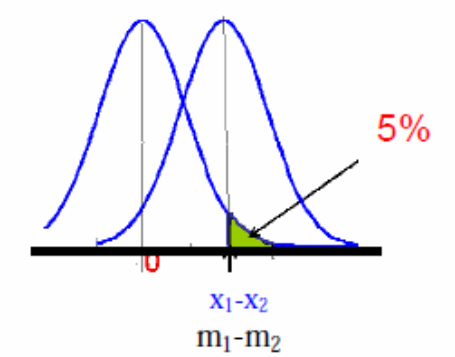

Figura 1. P-value de 0.05 cuando la muestra coincide con el valor verdadero

3. La comparación de modelos está mal resuelta

Los test de hipótesis son una forma particular de comparación de modelos. Los tests frecuentistas tienen la propiedad de que cuando la muestra es grande favorecen invariablemente al modelo más complejo (el test de razón de verosimilitudes, por ejemplo). En el caso bayesiano, las probabilidades posteriores de los modelos dependen fuertemente de las distribuciones de probabilidad *a priori* de los parámetros de los modelos, además de depender de las probabilidades *a priori* de los modelos en sí. Si estas últimas se toman iguales (lo que puede ser simplemente incorrecto), tenemos los factores de Bayes [2], que siguen dependiendo fuertemente de las distribuciones a priori de los parámetros de los modelos.

Descartados los test frecuentistas por irrelevantes y los bayesianos porque no sabemos cómo definir con precisión la probabilidad a priori, queda la fontanería. Esta consiste en un conjunto de métodos que no son ni frecuentistas ni bayesianos y que utilizan frecuentemente mecanismos de ambas escuelas simultáneamente. Por ejemplo, un mecanismo consiste en obtener una estima del valor predictivo de un modelo minimizando la distancia del modelo a la distribución verdadera de los datos. Como todo esto depende de los valores de los parámetros y no los conocemos, los sustituimos por su estima máximo verosímil y obtenemos el AIC[3], o por su media posterior y obtenemos el DIC. El problema es que, además de no saber exactamente qué estamos haciendo (el modelo elegido puede, por ejemplo, no ser el más probable), no sabemos qué quiere decir el resultado del criterio elegido, no sabemos qué son tres puntos de AIC o de DIC y tenemos que fiarnos de simulaciones, de la opinión de estadísticos conspicuos o de la intuición. James O. Berger me dijo en una ocasión que

[2] Como casi toda la estadística bayesiana, los factores de Bayes fueron propuestos por Laplace, no por Bayes.

[3] La distancia que se minimiza se conoce como distancia de Kullback (que no es en realidad una distancia), y tiene una justificación bayesiana más o menos traída por los pelos, pero el AIC usa estimas de máxima verosimilitud frecuentistas. No son auténticos métodos de contraste de hipótesis, con propiedades de inferencia claramente establecidas.

él sólo utilizaba el DIC como análisis exploratorio, para seleccionar tres o cuatro modelos de entre treinta o cuarenta; sin embargo el DIC (que no es un método bayesiano, por cierto), se usa porque es un subproducto de MCMC sencillo de calcular, como aquél que buscaba bajo la luz de una farola sus llaves perdidas no porque las perdiera allí, sino porque allí había luz.

3. Qué hacer

1. ***No hacer test de hipótesis***. Convendría dejar de publicar test de hipótesis cuando no fueran necesarios; es decir, en la mayor parte de las ocasiones. Para las inferencias que se realizan comparando tratamientos, estimando parámetros genéticos o estimando respuestas a la selección o efectos de genes, los intervalos de confianza son más relevantes que los test de hipótesis.

2. ***Publicar intervalos de confianza y no errores estándar***, a menos que sea imprescindible. En el caso de correlaciones y heredabilidades puede que las distribuciones de muestreo no sean normales si los valores están cerca de la frontera del espacio paramétrico, pero en ese caso lo errores estándar son también de poca utilidad, ¿qué quiere decir una correlación de 0.8 ± 0.3?

3. ***Discutir los asuntos importantes con los valores críticos de los intervalos de confianza***. Por ejemplo si se está recomendando seleccionar un carácter, no dar un valor de la heredabilidad de 0.20 ± 0.07 y recomendar la selección, sino fijarse en el límite inferior del intervalo de confianza y decir que la heredabilidad podría ser en realidad 0.06, por lo que es a riesgo del lector el seleccionar o no.

4. ***Comparar con otros autores considerando los intervalos de confianza respectivos***; si Morgan tiene una heredabilidad de 0.10 ± 0.07, la nuestra de antes NO es superior a la suya, simplemente podría serlo (o ser inferior, vete a saber).

5. ***Usar intervalos de credibilidad bayesianos***. El uso de MCMC hace facilísimo crear intervalos de credibilidad que satisfagan las preguntas del lector. Ejemplo:

Cómo sustituir los test de hipótesis por algo más interesante
$P(A-B|\mathbf{y})>0$ da la probabilidad de que la diferencia entre los tratamientos A y B sea mayor que cero (**y** son los datos). Esto no es un test de hipótesis, pero los sustituye con ventaja. Si esta probabilidad es del 93% no quiere decir que las diferencias sean n.s., porque aquí no hay significaciones, aquí se trabaja con la verdadera probabilidad de que los tratamientos sean diferentes, por lo que ese 93% puede ser suficiente para preferir el tratamiento A (depende de las necesidades del investigador).

La probabilidad de Relevancia
Para analizar *cualquier* resultado es importante conocer qué cantidad es relevante R para la variable que se está analizando. La probabilidad de que la diferencia entre tratamientos sea mayor que un valor relevante R es muy útil para decidir (figura 3a). *Podríamos usar cocientes en lugar de diferencias* $P(A/B >R|\mathbf{y})$, así representaríamos la probabilidad de que un tratamiento sea, p. ej., un 10% superior a otro (R=1.1).

La probabilidad de similitud

En variables continuas "distinto de cero" significa mayor o menor que una cierta cantidad que se considera *relevante* a efectos económicos o biológicos, puesto que las medias de la población control y la seleccionada nunca van a ser *exactamente* iguales. En la figura 2a se observa que se puede afirmar con una probabilidad del 96% que la diferencia entre tratamientos no ha sido distinta de una cierta cantidad relevante (no ha sido "distinta de cero"), mientras que la figura 2b muestra que no hay datos suficientes como para llegar a una conclusión. Esto es interesante, porque permite distinguir cuándo no aparecen diferencias entre las poblaciones y cuándo simplemente no se dispone de datos suficientes como para afirmar que hay diferencias.

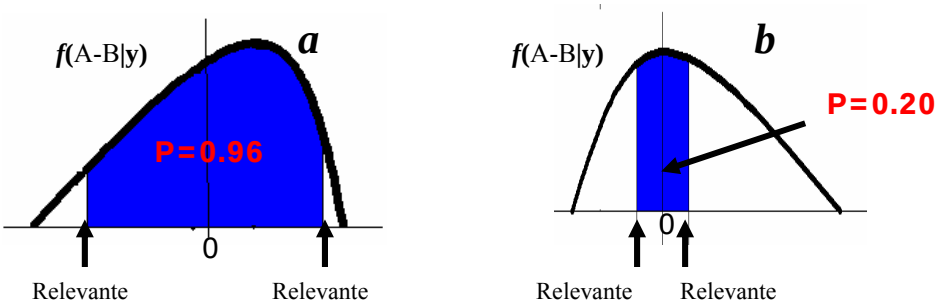

Figura 2. Probabilidad de similitud entre las poblaciones A y B. ***a.*** Las poblaciones son similares. ***b.*** No tenemos datos suficientes como para precisar si son similares.

El valor mínimo garantizado

Otra inferencia interesante es conocer el mínimo valor de un parámetro con una probabilidad determinada. Frecuentemente se afirma, que la heredabilidad de un carácter es relevante, por ejemplo 0.20, cuando su intervalo de confianza puede ir de 0.01 a 0.39, con lo que en realidad podría ser irrelevante. Una inferencia interesante puede ser conocer el valor que *al menos* puede tener un parámetro (o una diferencia de medias, o el efecto de un QTL) con una probabilidad determinada. En la figura 3b se representa el valor mínimo k que debe tener la diferencia entre dos poblaciones A y B con una probabilidad del 95%.

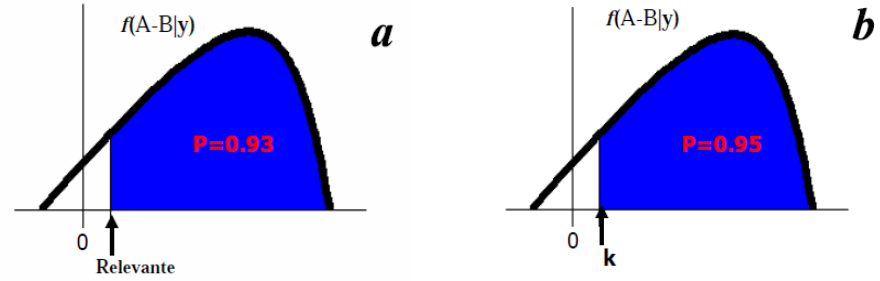

Figura 3. a. Probabilidad de que la diferencia entre tratamientos sea relevante. **b.** Intervalo [k, +∞) indicando que la diferencia entre tratamientos tiene un valor k o superior con una probabilidad del 95%.