A. Blasco

LA SIGNIFICACIÓN ES IRRELEVANTE Y LOS P-VALORES ENGAÑOSOS. ¿QUÉ HACER?

Separata ITEA

INFORMACIÓN TÉCNICA ECONÓMICA AGRARIA, VOL. **107** N.º 1 (48-58), 2011

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

A. Blasco

Departamento de Ciencia Animal. Universidad Politécnica de Valencia. P.O. Box 22012. Valencia 46071. Spain. Tel. 963 877 433. E-mail: ablasco@dca.upv.es

Resumen

En este artículo se revisan los métodos estadísticos más habituales usados en el análisis de datos en agricultura, tanto para comparar tratamientos como para expresar la incertidumbre respecto a un parámetro estimado. Se subraya que son muy frecuentes las interpretaciones erróneas de estos métodos, se indica su interpretación correcta y sus limitaciones, y se propone una metodología más informativa para presentar los resultados. Se discute el concepto de "valor relevante", que se propone como fundamental para el diseño de experimentos y la interpretación de los resultados. Finalmente se propone una alternativa bayesiana al análisis de datos.

Palabras clave: análisis de datos, significación, P-valor, test de hipótesis, análisis bayesiano.

Summary
Significance is irrelevant and P-values are misleading. What can we do?

The main statistical methods used in agricultural research for treatment comparison and for describing uncertainty about estimated parameters are reviewed. Since wrong interpretations of these methods are common, more informative procedures for presenting results are proposed. The concept of "relevant valor" is discussed and it is assumed as a fundamental concept for designing experiments and results interpretation. Finally, a Bayesian alternative of analyzing data is presented.

Key words: data analysis, significance, P-valor, hypothesis test, bayesian analysis.

Introducción

No hace falta leer mucha literatura científica para darse cuenta de que las contradicciones entre resultados de experimentos son frecuentes; o lo que es peor, que las conclusiones del artículo no están apoyadas en los resultados del experimento, sino en los prejuicios de los autores o en información proveniente de la literatura. En este artículo expondré primero las confusiones que se generan en la estadística clásica con los errores estándar, la significación y los P-valores, luego propondré una forma de presentar los resultados que contribuye a no concluir más de lo que un experimento permite (y que también ayuda en la discusión de los artículos), y terminaré exponiendo una forma de presentar los resultados con metodología Bayesiana, que me parece la más recomendable para ayudar a comprender los resultados de un experimento. Me limitaré en este artículo a la mera comparación de dos tratamientos; por ejemplo la comparación entre un grupo seleccionado y uno control o la comparación entre los resultados de dos piensos, aunque la generalización de lo que propongo es inmediata en la mayor parte de los casos.

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. En experimentos bien diseñados "significativo" significa que hay diferencias y "n.s." (no significativo) que no hay diferencias entre tratamientos (corriendo un riesgo determinado de equivocarnos); si se ha diseñado el experimento sin definir la potencia del test, como es frecuente, n.s, significa "no sé", no sé si hay o no diferencias. *Con una diferencia entre tratamientos n.s. no puedo decir en ningún caso que se observa una "tendencia"*. Si la diferencia es n.s. esto quiere decir que los resultados que se observan se deben a muestreo aleatorio, y que repitiendo el experimento pueden tener perfectamente signo contrario.

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 "n.s." 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 además no sabemos cuál de los dos tratamientos es el que provocaría esa importante diferencia en tamaño de camada. El investigador puede tener la falsa impresión de que una diferencia de 0.1 lechones que es n.s. implica que no hay diferencias relevantes entre tratamientos, lo que no se puede afirmar sin ver el intervalo de confianza.

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 el verdadero valor está por el centro del intervalo de confianza, pero esto no tiene por qué ser 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. Como en realidad no hacemos infinitas repeticiones sino que sólo estimamos el intervalo de confianza una vez, 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. Lo mismo ocurre cuando se establecen muchas comparaciones entre niveles de un tratamiento; el primer nivel con cada uno de los otros, el segundo nivel con los restantes, etc., aquí también aparecen diferencias significativas 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).

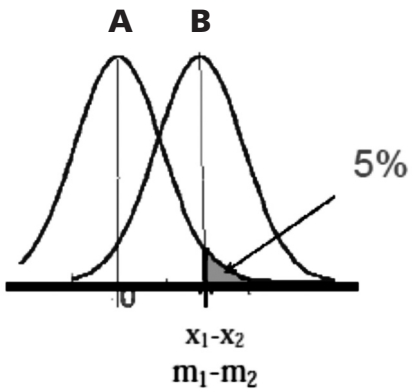

Figura 1. A sería la distribución de las muestras repitiendo infinitas veces el experimento, si el valor verdadero de la diferencia entre tratamientos fuera 0; esto es, si realmente no hubiera diferencia entre tratamientos. B es la distribución de las muestras cuando el valor verdadero ($m_1 - m_2$) coincide con el de la muestra concreta de nuestro experimento ($x_1 - x_2$). En este ejemplo el valor verdadero está situado exactamente en el umbral de significación. El área sombreada es el P-valor de 0.05 y la flecha indica el umbral de significación. Aunque el P-valor parezca pequeño, si se repite el experimento, la mitad de las veces saldrá n.s.

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 los resultados de una actividad enzimática. En esos casos 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 relevancia alguna.
4. Con datos de laboratorio, frecuentemente escasos y costosos de obtener, se obtienen resultados "n.s." en abundancia, pero que el experimentador quisiera interpretar dado el trabajo que costó obtenerlos. Entonces se recurre con frecuencia expresiones absurdas como "se observa una tendencia…". Esto es absurdo por dos razones; la primera ya la hemos comentado, "n.s." significa "no sé", quiere decir esto que los valores observados se debe al azar y que una hipotética repetición del experimento daría un resultado diferente. La segunda razón es que en ocasiones la diferencia "n.s." que aparece puede ser muy grande, con lo que ni siquiera es defendible la expresión "tendencia". En ocasiones se indica un nivel de significación del 10% y se vuelve a hablar de "tendencias" aunque las diferencias observadas sean espectaculares. Aquí se confunde nuestro nivel de incertidumbre con lo que los datos dicen: nuestra incertidumbre puede ser elevada, pero los datos dicen que es posible que las diferencias entre tratamientos sean grandes, no que se observe una tendencia.

La pregunta está frecuentemente mal planteada. La pregunta correcta 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 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.

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., la diferencia entre las razas B y C puede ser también n.s., pero la diferencia 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". Sin embargo, 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, ba-

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

sándonos 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.) y que veremos luego, 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 el nivel de significación 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 en nuestro experimento 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 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". La alternativa de dar el P-valor exacto suele confundir en lugar de clarificar la situación, como veremos a continuación.

Los P-valores son engañosos

Un P-valor 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-valor 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[2]. El problema es *cuán excepcional* es la muestra si nos sale un P-valor del 2%, ¿es extraordinariamente excepcional, o menos de lo que parece? ¿Qué quiere decir un 2%?

El P-valor tiene al menos dos interpretaciones incorrectas:

2. Bueno, no todos. Una crítica al modelo de Fisher es que tan excepcional es la cola superior del 2% como el área de probabilidad del 2% alrededor de cero, o cualquier otra área de probabilidad del 2%, pero no entraremos en digresiones filosóficas que incomoden al lector.

1. El P-valor es interpretado como la probabilidad de que no haya diferencias entre tratamientos. Esto es obviamente incorrecto, el P-valor es la probabilidad de la diferencia entre las muestras, no la probabilidad de la diferencia entre los tratamientos. Lo que pasa es que lo que nos interesa realmente es conocer 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 que, por cierto, no hemos tomado ni vamos a tomar). Ya que no podemos ir al Amazonas, nos conformaremos viendo un documental.

2. El P-valor 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-valor no puede dar el nivel de significación porque si repetimos el experimento el P-valor 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-valor 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, sólo sabemos que un P-valor pequeño presenta más evidencia que uno grande. Supongamos que obtenemos un P-valor 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-valores de 0.00001 indican evidencias mayores 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.

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, y si hay muchos datos todos los efectos acaban por ser significativos. En el caso bayesiano, en el que esto no ocurre, el problema reside en que 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 incorrecto), tenemos los factores de Bayes[3], 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 fre-

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

cuentemente 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[4], 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.

Una solución que últimamente se está imponiendo en la comparación de modelos, dadas las dificultades anteriores, es la validación cruzada, en la que parte de los datos se usan para elegir los modelos y parte de los datos para comprobar si los modelos predicen bien nuevos datos. Esto se puede hacer de muchas formas y con mayor o menor sofisticación, pero en conjunto da la tranquilidad psicológica de que si un modelo predice adecuadamente nuevos datos, es un buen modelo. Esta solución no está exenta de críticas; por ejemplo, un modelo podría predecir bien un tipo de datos pero no otros; por ejemplo, los datos extremos podrían predecirse bien con un modelo y mal con otro y viceversa, con lo que no está claro cuál se debería elegir. Determinar cuándo una predicción es suficientemente buena en conjunto se convierte en algo bastante arbitrario; hace falta además decidir cómo se mide la discrepancia entre la predicción y el dato observado. Finalmente, estos métodos suelen ser complicados de computar, lo que no los hace aconsejable necesariamente para problemas sencillos como los de comparación de medias, sobre todo cuando hay muchas medias a comparar, lo que es frecuente en problemas de tipo biológico.

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 o estimando parámetros, los intervalos de confianza son más relevantes que los test de hipótesis.
2. *No publicar LS-means sino diferencias entre tratamientos.* Cuando se comparan tratamientos la pregunta debería ser si uno es superior a otro y, sobre todo, cuán superior es un tratamiento a otro. Para cuantificarlo hay que hacer la diferencia entre tratamientos *y dar la precisión (el intervalo de confianza) de esta diferencia*. Si se dan las LS-means (las medias calculadas por mínimos cuadrados) se puede calcular fácilmente la diferencia entre tratamientos, pero el error típico de la diferencia entre tratamientos no está explícito y no se calcula inmediatamente. Presentar las medias generales y las diferencias entre tratamientos junto a sus intervalos de confianza es más informativo, aunque puede no ser práctico cuando el número de tratamientos (esto es, de niveles dentro de un factor) es muy elevado.

4. 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.

3. *Referir las diferencias entre tratamientos a su "valor relevante".* Para analizar *cualquier* resultado es importante conocer qué cantidad es relevante para la variable que se está analizando. No es posible realizar un diseño experimental adecuado sin conocer qué diferencia se quiere detectar; esta sería el valor relevante para ese carácter. No es posible tampoco analizar adecuadamente una diferencia entre tratamientos si no se sabe qué diferencia es relevante para el problema que se está tratando. En muchas ocasiones una diferencia será relevante por motivos económicos; por ejemplo, un 0.1 en índice de conversión ó 0.5 lechones de diferencia en tamaño de camada. En otros caracteres es más difícil precisar un valor relevante; por ejemplo, si los resultados de un panel de pruebas de calidad de carne dan tres puntos más a un tratamiento en "sabor a anis" ¿es esto mucho o poco? O bien, ¿cuál es el valor relevante de una actividad enzimática? En esos casos una solución puede ser referirse a una fracción de la desviación típica del carácter[5]. Por ejemplo, en los caracteres cuyo valor relevante se deduce de datos económicos (índice de conversión, contenido en carne, producción de leche, etc.) se puede comprobar que el valor relevante está entre 1/2 y 1/3 de la desviación típica del carácter.

4. *Publicar intervalos de confianza y no errores estándar,* a menos que sea imprescindible. Los intervalos de confianza de las diferencias entre tratamientos están en torno al doble del error estándar de esta diferencia. En el caso de correlaciones y heredabilidades puede que si repetiéramos infinitas veces el experimento las distribuciones que obtuviéramos no fueran normales (ocurre cuando los valores están cerca de los límites), pero en ese caso lo errores estándar son también de poca utilidad, ¿qué quiere decir una correlación de 0.8 ± 0.3, si sabemos que no puede ser mayor que 1.0?

5. *Discutir los asuntos importantes con los valores críticos de los intervalos de confianza*. Por ejemplo si una diferencia entre tratamientos es de 0.20 ± 0.07 hay que fijarse en el límite inferior del intervalo de confianza (que está en torno al doble del error estándar) y decir que la diferencia entre tratamientos podría ser en realidad 0.06.

6. *Comparar con otros autores considerando los intervalos de confianza respectivos;* si un autor tiene una diferencia entre tratamientos de 0.10 ± 0.07, la nuestra de 0.20 ± 0.07 NO es superior a la suya, simplemente podría serlo (o ser inferior), debido a que los intervalos de confianza son grandes.

7. *Usar intervalos de credibilidad bayesianos.* El uso de modernas técnicas de integración (conocidas como MCMC, Markov Chain Monte Carlo) en los programas de ordenador hace facilísimo crear intervalos de credibilidad que satisfagan las preguntas del lector, como veremos en el apartado siguiente.

Una alternativa bayesiana

La estadística bayesiana fue de uso común a lo largo del siglo XIX y principios del siglo XX hasta que fue sustituida por la que hoy

5. En ocasiones se habla de valor relevante como un porcentaje de la media del carácter. Todo esto es bastante arbitrario, pero a mí me parece más importante definir un valor relevante a partir de la variabilidad del carácter. Tener seis dedos es importante no porque sea un 20% más que la media de una mano (o el 10% de las dos), sino porque es un carácter muy poco variable. Lo mismo ocurre con el rendimiento a la canal y otros caracteres.

llamamos "estadística clásica" o frecuentista a partir de los años 30 del siglo XX. El principal problema de la estadística bayesiana era operativo: sus resultados daban lugar a integrales múltiples difíciles de resolver. El desarrollo de los ordenadores permitió que estas integrales se estimaran con precisión usando los métodos a los que nos hemos referido antes (MCMC), y desde los años 90 la estadística bayesiana está volviendo a ser usada ampliamente y, en algunos campos como el de la mejora genética animal, de forma habitual. En la actualidad se ha usado principalmente para problemas complejos para los que la estadística frecuentista no tenía una solución, o si la tenía no era fácil de implementar. Sin embargo creo que es posible usar la estadística bayesiana para problemas sencillos, ofreciendo más información en la interpretación de resultados que la que ofrece la estadística clásica. En las revisiones de Blasco (2001, 2005) se pueden encontrar ejemplos sencillos de su aplicación. Otros ejemplos sencillos de comparación de medias en análisis de calidad de carne, en reproducción, y en genética se encuentran en Zomeño *et al*. (2010), Mocé *et al*. (2010) y Peiró *et al*. (2010) respectivamente.

La estadística bayesiana fue criticada cuando apareció la alternativa "clásica" porque sus resultados parecían depender de los valores *a priori* necesarios para aplicar este tipo de inferencias. Aquí hay que distinguir dos aplicaciones de la estadística: la estimación y los test de hipótesis. No hay problemas con el uso de información *a priori* en el caso de la estimación, porque se usan habitualmente "*aprioris*" muy poco informativos que prácticamente no afectan a los resultados; los llamados "*aprioirs* planos" son un ejemplo muy usado, y es común que los programas de ordenador los asuman por defecto. Los test de hipótesis (factores de Bayes) ya vimos que sí son afectados por la información *a priori*, y no se puede recomendar su uso de forma general, pero también vimos que la solución "clásica" es tan insatisfactoria como la bayesiana.

1. *Cómo sustituir los test de hipótesis por algo más interesante:* P(A-B|*y*)>0 (donde *y* son los datos) da la probabilidad de que la diferencia entre los tratamientos A y B sea mayor que cero. 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).
2. *La probabilidad de Relevancia:* La probabilidad de que la diferencia entre tratamientos sea mayor que un valor relevante R es muy útil para tomar decisiones (figura 2a). En muchas ocasiones el cambiar o no un tratamiento depende de que haya una elevada probabilidad de que las diferencias entre tratamientos sean relevantes. No depende tanto de que la diferencia entre tratamientos estimada sea grande, porque al ir esta diferencia acompañada de un error de estimación, podría ser que en realidad la diferencia no fuera tan grande como aparenta.
3. *El uso de cocientes en lugar de diferencias:* Es frecuente que sea más interesante el valor relativo entre tratamientos que su diferencia en valor absoluto. Por ejemplo, cuando no está claro cuál es el valor relevante, el cociente da una idea de la importancia relativa de los tratamientos; es más expresivo decir que el sabor a anís de la carne de un tratamiento es un 20% más acusado que el otro que decir que ambos tratamientos difieren en tres puntos. En la estadística clásica tenemos un problema con la precisión; el error están-

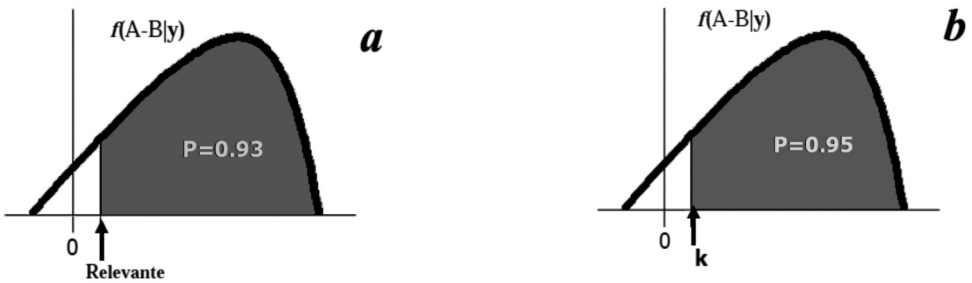

Figura 2. a. Probabilidad de que la diferencia entre tratamientos sea relevante. b. Intervalo [k, +∞) indicando que la diferencia entre tratamientos tiene un valor k o superior con una probabilidad del 95%.

dar de un cociente de LS-means no es fácil de calcular y hay que acudir a aproximaciones complicadas. En estadística bayesiana con MCMC es elemental calcular cualquier intervalo de probabilidad. Podríamos representar, por ejemplo, la probabilidad de que un tratamiento sea un 20% superior a otro.

4. *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 3a 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 3b 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.
5. *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 2b se representa el valor mínimo k que debe tener la diferencia entre dos poblaciones A y B con una probabilidad del 95%.
6. *El intervalo más corto con una probabilidad del 95%:* Como hemos dicho antes, a veces las distribuciones de las muestras al repetir infinitas veces un experimento no son simétricas. En ese caso el intervalo más corto (el más preciso) con el 95% de la probabilidad de incluir al valor verdadero no es simétrico en torno al valor estimado. El intervalo de confianza para una correlación de 0.8 podría ser [0.9, 0.5], nunca tendríamos el caso absurdo de 0.8 ± 0.3.
7. *El software:* No hay todavía un desarrollo de software bayesiano comparable al frecuentista, pero ya hay programas suficientemente amigables para resolver casi cualquier problema. El programa gratuito WinBugs *(http://www.mrc-bsu.cam.ac.uk/bugs/)* cubre un amplísimo rango de pro-

blemas, y su dificultad de uso es similar a la que puede tener el SAS. Para los genetistas, que necesitan utilizar la correlación entre efectos aleatorios y que usan modelos poco estándar en el mundo de la estadística, hay programas públicos específicos. En la página web de la red ACTEON (*http://acteon.webs.upv.es/*) se puede encontrar software gratuito para problemas genéticos específicos. En particular, el programa TM puede estimar componentes de varianza y efectos en modelos multicaracteres mixtos con varios efectos aleatorios correlacionados, con datos multinormales, discretos (con varios umbrales) y con datos censurados.

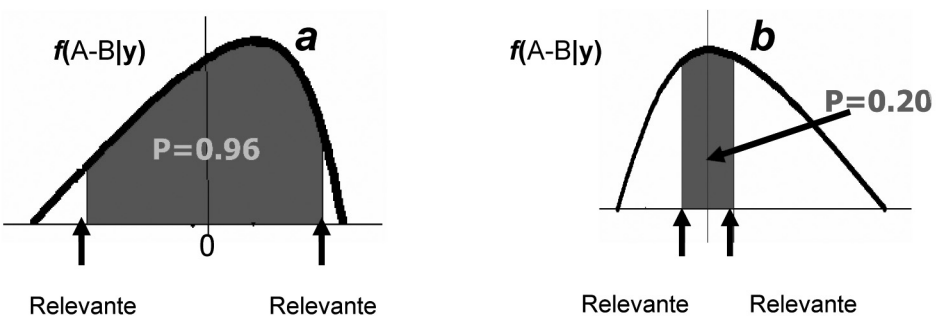

Figura 3. Probabilidad de similitud entre las poblaciones A y B. a. Las poblaciones son similares. b. No tenemos datos suficientes como para precisar si son similares.

Bibliografía

Blasco A. 2001. The Bayesian controversy in Animal Breeding. *J. Anim. Sci.* 79: 2023-2046.

Blasco A. 2005. The use of Bayesian statistics in meat quality analyses. *Meat Sci.* 69: 115-122.

Moce L, Blasco A, Santacreu MA. 2010. In vivo development of vitrified rabbit embryos: effects on prenatal survival and placental development. *Theriogenology*. 73: 704-710.

Peiró R, Herrler A, Santacreu MA, Merchán M, Argente MJ, García ML, Folch JM, Blasco A. 2010. Expression of progesterone receptor related to the polymorphism in the PGR gene in the rabbit reproductive tract. *J. Anim. Sci*. 88: 421-427.

Zomeño, Blasco A, Hernández P. 2010. Influence of genetic line on lipid metabolism traits of rabbit muscle. *J. Anim. Sci.* 88: 3419-3427.

(Aceptado para publicación el 17 de diciembre de 2010)