SOBRE EL SESGO DE ESTIMACION DE COMPONENTES DE VARIANZA DE UN CARACTER DICOTOMICO

C.Moreno, L.Varona, L.A.García Cortés[1], J.Altarriba
Unidad de Genética Cuantitativa y Mejora Animal
Facultad de Veterinaria. Universidad de Zaragoza
Miguel Servet 177. 50013 Zaragoza

[1]National Institute of Animal Science. Dept.for Research in Pigs and Horses
P.O.Box 39.DK-8830. Tjele. Dinamarca

La estimación de componentes de varianza en caracteres binarios es un tema que ha recibido escasa atención en el ámbito de la genética cuantitativa. Las posibles causas de este hecho podrían ser la lógica atención por los problemas existentes en caracteres continuos, la carencia de bases de datos con información discreta y la falta de paquetes informáticos con los que llevar a cabo el análisis de dicha información. Ultimamente, con el "descubrimiento" de nuevas técnicas de computación, tradicionales en otras esferas científicas, se ha logrado un incremento del interés por los problemas todavía presentes en este campo. Sin embargo, este entusiasmo inicial solo se ha traducido en una exposición causística del problema, dando lugar a conclusiones erróneas o, en el mejor de los casos, a resultados sobradamente conocidos en otras áreas.

Si hubiera que citar algún trabajo de referencia, dentro del contexto genético, éste sería el artículo de Hoschele et al. (1987). En este trabajo se realiza una estimación de componentes en caracteres categóricos mediante un planteamiento bayesiano, calculando las integrales obtenidas a partir de una aproximación normal. Como resultado se afirma que el método propuesto, denominado por los autores como MML (*marginal maximum likelihood*), da lugar a heredabilidades infraestimadas cuando se analizan casos con incidencias superiores al 90%, obteniéndose una sobreestimación en situaciones de incidencia inferior al 90% y con un tamaño de subclase (número de observaciones por cada combinación de niveles fijos y aleatorios) inferior a 2. Se concluye que el procedimiento ofrece buenos resultados excepto cuando el tamaño de subclase es inferior a 2, destacando que esto es cierto incluso cuando el número de observaciones por cada nivel aleatorio es elevado. La justificación de estos resultados y conclusión se basa en el mal comportamiento de la aproximación normal. Sin embargo, estas afirmaciones ofrecen una visión equivocada del problema. Es cierto que la aproximación normal produce sesgo cuando el número de observaciones por nivel fijo o aleatorio es bajo, lo que es erróneo es sostener que se produce sesgo siempre que el tamaño de subclase es inferior a 2. La aproximación normal funciona correctamente siempre y cuando la cantidad de información por nivel sea elevada, independientemente del tamaño de subclase (McCullagh y Nelder, 1989, pag.120). Además, la incidencia nunca determina la "dirección" del sesgo (infraestimación o sobreestimación). Es bien sabido que incidencias extremas

provocan una disminución de la cantidad de información observada y, en consecuencia, generan mayor sesgo. La obtención de infraestimación o sobreestimación con MML depende de si el modelo es aleatorio o fijo dominante, respectivamente. Hay que tener en cuenta que la inversion de matrices de información con un elevado número de valores próximos a cero en la diagonal de los fijos, situación común en modelos fijo dominantes, da lugar a coeficientes excesivamente elevados en la diagonal de los aleatorios, hecho que justifica la obtención de componentes sobreestimados.

La utilización de técnicas de integración tan poderosas como el muestreo de Gibbs (MG) ha permitido ofrecer una nueva perspectiva de la estimación de componentes en categóricos. El sesgo observado empleando MML no es exclusivamente consecuencia del mal comportamiento de la aproximación normal ya que, mediante un método que no está basado en ninguna aproximación, sigue obteniéndose sesgo. Concretamente, los resultados que hemos obtenido, trabajando con modelos macho, muestran que con MG se elimina el sesgo de infraestimación, presentado por MML, en modelos aleatorio dominantes donde cada nivel aleatorio está estimado con un reducido número de observaciones. Sin embargo, el sesgo no desaparece en aquellos casos donde la cantidad de información por nivel fijo es escasa. Con respecto a la utilización de MG hay que destacar, además, los siguientes puntos. Primero, la evidencia de sesgo está supeditada a modelos mixtos. Segundo, el problema no radica exclusivamente en la existencia de niveles fijos con todas las observaciones en una única categoría (este tipo de variables son especialmente problemáticas ya que su estimador tiende a infinito). Esta afirmación surge como consecuencia tanto de la utilización de modelos donde se ha eliminado este tipo de inconveniente y sigue observándose sesgo, como de modelos donde hay una gran número de esta clase de variables y no aparece sesgo. Tercero, el sesgo se reduce conforme se incrementa la cantidad de información del componente. Sin embargo, esta afirmación no implica que el problema deba entenderse como consecuencia de una falta de información del componente. En los casos estudiados, la distribución marginal del componente es simétrica, pero se encuentra desplazada del valor simulado.

Como conclusión de las anteriores observaciones puede destacarse que, la obtención de sesgo en la estimación de componentes en caracteres categóricos mediante MG, es un proceso provocado por la estimación de niveles fijos con escasa cantidad de información; sesgo que se reduce según se incrementa la información del componente.

Con respecto a posibles soluciones de este problema comentar que, debido a que los modelos aleatorios no presentan sesgo, se ha propuesto como alternativa intentar estimar a los factores fijos como si fueran aleatorios. Este tipo de estrategias consigue reducir el sesgo, aunque no lo elimina totalmente. Otro tipo de solución, frecuentemente empleada en estudios sociológicos (Park y Brown, 1994), se

basa en la obtención de distribuciones a priori que logren disminuir el efecto de la falta de información en los niveles fijos. Por último, señalar que casi todos los autores están de acuerdo en que parece difícil hallar una solución general para todos los casos.

Bibliografía

Hoschele,I.; Gianola,D.; Foulley,J.L. 1987, Estimation of variance components with quasi-continuous data using bayesian methods. *J.Anim.Breedg.Genet*.104,334-349.

McCullagh,P.; Nelder,J.A. 1989, *Generalizaed Linear Models*. London: Chapman and Hall.

Park,T.; Brown,M.B. 1994, Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association* 89,44-52.