ESTIMACIÓN DEL NÚMERO DE DATOS FALTANTES EN POBLACIONES SELECCIONADAS

G. Yagüe[1], C. Moreno[1], L.A. García-Cortés[2] y J. Altarriba[1]
[1] Genética cuantitativa y Mejora animal. Facultad de Veterinaria. Universidad de Zaragoza.
[2] Departamento de Mejora Genética y Biotecnología. INIA. Madrid.

INTRODUCCIÓN

El tratamiento de poblaciones sometidas a un proceso de pérdida de información no ignorable, como consecuencia de la aplicación de una estrategia de selección, requiere la inclusión de dicho proceso en el modelo para garantizar la obtención de resultados insesgados (Im et al.,1989; Gianola et al., 1989).

Sin embargo, la resolución analítica de casos donde no sólo se pierden observaciones sino que también se desconoce su número no es inmediata (Yagüe et al., 2000).

El objetivo de este trabajo es presentar un procedimiento "ad hoc" que permita realizar estimaciones insesgadas de los parámetros de poblaciones donde se desconoce el número de datos perdidos.

MATERIALES Y MÉTODOS

Para presentar este procedimiento se considera un caso sencillo, una distribución normal con media μ y varianza σ^2, donde sólo se observan n_o valores que superan el punto de truncamiento t (conocido), siendo n_m el número de datos menores que t desconocido. La proporción de población desconocida queda definida según $p = \Phi\left(\frac{t-\mu}{\sigma}\right)$, cumpliéndose, además, la siguiente relación:

$$\frac{n_m}{n_o} = \frac{p}{1-p} \qquad [1]$$

de tal manera que si se conocen n_o y p, es posible calcular el número de datos faltantes, n_m, a partir de [1]

Supongamos ahora que no se dispone de una sola distribución, sino de una mezcla de ellas, dado que puede asumirse que cada dato y_i procede de una distribución normal i, con media μ_i y varianza σ^2. Consideramos el mismo punto de truncamiento t y los mismos n_o y n_m en el conjunto de la población, siendo p la proporción de valores perdidos referidos al total de la población. Si definimos p como:

$$p = E\left[\Phi\left(\frac{t-\mu_i}{\sigma}\right)\right] \approx \frac{\sum_{i=1,n}\left[\Phi\left(\frac{t-\mu_i}{\sigma}\right)\right]}{n}$$

entonces n_m puede obtenerse utilizando [1]. A partir de estas expresiones es posible realizar inferencias en casos donde se desconoce el número de datos perdidos llevando a cabo los dos pasos siguientes hasta obtener convergencia:

- estimar las variables del modelo (mediante muestreo de Gibbs) condicionado al conocimiento del número de observaciones perdidas.
- "estimar" el número de observaciones perdidas fuera del proceso de Gibbs, a partir de la expresión anterior, usando los valores de las estimaciones obtenidas en el primer caso.

Sea ahora el modelo animal mixto *y=XB+Zu+Ie*, donde *y* es el vector de fenotipos, B es el vector de un efecto fijo de 20 niveles, *u* el vector de valores genéticos aditivos y *e* el vector de residuos del modelo. Se simulan dos generaciones, con 2000 individuos en la población base (G_0) y 3000 en la generación filial (G_1), realizándose selección fenotípica por truncamiento en ambas generaciones (los puntos de truncamiento son conocidos) y perdiéndose toda la información referente a los individuos no seleccionados, incluyendo su cantidad. Como consecuencia de la inclusión del proceso de pérdida de información en el modelo se obtiene un modelo mezcla (*mixture model*) donde el número de componentes es desconocido (Yagüe et al., 2000).

Estudiamos dos casos, donde tanto en G_0 como en G_1 se selecciona a un 50% y un 30% de los individuos (*p*=0.5 para el primer caso y *p*=0.7 para el segundo). El número de observaciones perdidas se estima independientemente para cada generación, siendo $n_{m(G0)}$ este número para la población base y $n_{m(G1)}$ para la generación filial.

RESULTADOS Y DISCUSIÓN

A continuación se muestran los resultados para las dos proporciones de datos perdidos (Tabla 1: *p*=0.5; Tabla 2: *p*=0.7) y para dos intervalos de valores del efecto fijo de 20 niveles (200-240; 200-216), siendo *v(a)* la varianza genética aditiva, *v(e)* la varianza residual, h^2 la heredabilidad, *ntot* el tamaño global de la población y *b(i)* el nivel *i* del efecto fijo. Entre paréntesis se encuentran las desviaciones típicas.

Tabla 1: Estimas de las variables en una población con selección fenotípica del 50%.

	Proporción de datos perdidos (p) = 50% (250 repeticiones)					
Variable	V. simulado	Valor estimado		V. simulado	Valor estimado	
v(a)	20.0	21.216	(3.674)	20.0	21.619	(5.148)
v(e)	80.0	82.300	(5.010)	80.0	86.394	(8.386)
h^2	0.20	0.205	(0.033)	0.20	0.200	(0.045)
$n_{m(G0)}$	1000	1055.1	(100.3)	1000	1161.4	(170.3)
$n_{m(G1)}$	1500	1555.2	(112.6)	1500	1690.1	(205.8)
ntot	5000	5110.3	(190.0)	5000	5351.6	(358.3)
b(1)	200.00	198.94	(2.304)	200.00	199.10	(1.343)
b(10)	218.95	218.59	(1.028)	207.58	206.62	(1.348)
b(20)	240.00	239.87	(0.708)	216.00	215.09	(1.272)

Tabla 2: Estimas de las variables en una población con selección fenotípica del 30%.

	Proporción de datos perdidos (p) = 70% (950 repeticiones)					
Variable	V. simulado	Valor estimado		V. simulado	Valor estimado	
v(a)	20.0	22.290	(6.060)	20.0	23.014	(5.890)
v(e)	80.0	86.636	(8.712)	80.0	96.584	(13.418)
h^2	0.20	0.204	(0.052)	0.20	0.194	(0.048)
$n_{m(G0)}$	1400	1423.7	(171.2)	1400	2044.0	(507.0)
$n_{m(G1)}$	2100	2089.3	(229.9)	2100	2836.6	(615.3)
ntot	5000	5012.9	(364.4)	5000	6380.6	(1082.4)
b(1)	200.00	200.77	(1.813)	200.00	197.00	(2.357)
b(10)	218.95	217.85	(1.418)	207.58	204.79	(2.104)
b(20)	240.00	239.19	(1.080)	216.00	213.09	(2.248)

Se observa que la estrategia presentada es capaz de recuperar gran parte de la información perdida, obteniéndose en general unas estimaciones aceptables para los valores simulados. Por otra parte, se consiguen mejores estimas con el mayor de los intervalos de valores de los niveles del efecto fijo (200 a 240), hecho justificado por la mayor capacidad informativa de dichos niveles. Asimismo, es necesario

resaltar que incluso en casos extremos (30% de datos conocidos) los resultados siguen siendo satisfactorios, si se comparan con otras alternativas (modelo Tobit, por ejemplo).

Esta estrategia también ha sido aplicada a un modelo mixto con un factor fijo de 400 niveles y con un intervalo regular entre ellos de 200 a 240. En la Tabla 3 se muestran los resultados, para un 50% de población seleccionada. Las variables se estiman, por un lado, considerando al efecto fijo como tal (Modelo A) y, por otro, considerándolo como un factor aleatorio (Modelo B), donde μ es la media general y *v(f)* es la varianza del factor aleatorio.

Tabla 3: Estimas de las variables de un modelo mixto con un factor fijo de 400 niveles y una selección del 50%.(200 repeticiones)

Efecto estimado como fijo (Modelo A)				Efecto estimado como aleatorio (Modelo B)			
Variable	V. simulado	V. estimado		Variable	V. simulado	V. estimado	
				μ	220.0	223.097	(0.726)
v(a)	20.0	37.017	(10.454)	*v(a)*	20.0	20.789	(4.352)
v(e)	80.0	208.790	(37.998)	*v(e)*	80.0	66.101	(4.421)
h^2	0.20	0.153	(0.043)	h^2	0.20	0.238	(0.047)
				v(f)	134.0	75.354	(6.002)
$n_{m(G0)}$	1000	2362.3	(472.7)	$n_{m(G0)}$	1000	478.9	(62.5)
$n_{m(G1)}$	1500	3165.6	(561.9)	$n_{m(G1)}$	1500	719.9	(80.3)
ntot	5000	8028.0	(1011.0)	*ntot*	5000	3698.9	(132.3)
b(1)	200.00	191.83	(4.365)				
b(200)	223.96	210.55	(5.327)				
b(400)	240.00	230.18	(6.805)				

Los resultados obtenidos permiten afirmar que, utilizando la estrategia de la estimación del número de datos faltantes, las estimas son mejores cuando se considera al factor fijo de 400 niveles como un efecto aleatorio. Comparando el Modelo A con las Tablas 1 y 2 se manifiesta el empeoramiento de los resultados debido al aumento de niveles y, por tanto, a la disminución del número de observaciones por nivel. Sin embargo, el Modelo B, si bien la mayoría de las variables están considerablemente infraestimadas, supone una alternativa para este tipo de situaciones.

Como conclusión podemos decir que el procedimiento presentado aparece como una herramienta eficaz en el tratamiento de poblaciones donde la selección ha producido una pérdida de información no ignorable y de magnitud desconocida, situaciones en las que utilizando otros procedimientos no se obtienen resultados aceptables.

REFERENCIAS

- Gianola, D., Fernando, R.L., Im, S. et Foulley, J.L. 1989. "Likelihood estimation of quantitative genetic parameters when selection occurs: models and problems." Genome, **31**:768-777.
- Im, S., Fernando, R.L., y Gianola, D. 1989. “Likelihood inferences in animal breeding under selection: a missing-data theory view point”. Genet. Sel. Evol., **21**:399-414.
- Yagüe, G., Moreno, C., García-Cortés, L.A. y Altarriba, J. (Caldes de Montbui, 2000). “Inferencia en poblaciones donde se desconoce la cantidad de información perdida”. X Reunión Nacional de Mejora Genética Animal.