ESTIMACION DE LOS COMPONENTES DE VARIANZA EN DISEÑOS DESBALANCEADOS

CARLOS G. VÁSQUEZ P.[1]

Resumen

En este trabajo, se presenta en forma de ejemplos la estimación de los componentes de varianza para los diseños experimentales: una categoría de clasificación, p × q factorial método directo de Henderson, y jerárquico o anidado, suponiendo en todos ellos efectos aleatorios.

Introducción

La teoría estadística se ha enfocado principalmente a aquellos diseños experimentales en los cuales todos los niveles a estudiar de uno o más factores presentan igual número de observaciones y se les conoce como modelos balanceados; sin embargo, la investigación realizada en el estudio de procesos biológicos donde el material de trabajo son animales (tal es el caso de nutrición, reproducción, genética, etc.) se plantean problemas ya que se presentan generalmente modelos desbalanceados; es decir, hay diferente número de observaciones en las celdas, debido a que algunos animales durante el transcurso del experimento pueden morir o presentar algún factor como enfermedades que los deje fuera del experimento, y por ende el análisis de la información será basado en modelos desbalanceados.

En el área de genética al trabajar con características cuantitativas es importante conocer cuáles son los componentes de la variación atribuidas a las diferentes causas que constituyen el modelo genético. Cuando se trabaja con modelos desbalanceados la tarea de obtener estos componentes no es directa, como es el caso de los modelos balanceados. Herdenson (1953), Searle (1968) y Harvey (1975) son algunos de los autores que más han contribuido al estudio de la teoría de los modelos desbalanceados, siendo Harvey el que más se ha enfocado a aspectos biológicos.

El propósito de este trabajo es presentar en forma de ejemplos el método para la estimación de los coeficientes k y la estimación de los componentes de varianza para los diseños desbalanceados; una categoría de clasificación, un diseño factorial p × q utilizando el método uno de Henderson, y jerárquico o anidado suponiendo todos los factores aleatorios.

Ejemplos

A. Una categoría de clasificación.

Supóngase que existe un factor A (ya sean padres, tratamientos, etc.) y que el factor presenta cinco niveles tomados al azar del total de la población (padres: a, b, c, d, e) donde cada nivel presenta diferentes números de observaciones (n_i) esto es:

FACTOR A

Niveles (padres (i))	Número de observaciones ($n_{i.}$) (Progenie)
a	1
b	4
c	5
d	3
e	2
Total 5	$n.. = N = 15$

El modelo que presenta el total de la variación en este ejemplo es: $Y_{ij} = \mu + S_i + \epsilon_{(i)j}$; donde Y_{ij} es la respuesta de la

Recibido para su publicación el 13 de diciembre de 1982.

[1] Departamento de Genética Animal, Instituto Nacional de Investigaciones Pecuarias, SARH, Km 15.5 Carretera México-Toluca, México, D.F. C.P. 05110.

j-ésima progenie del i-ésimo padre; μ es la media poblacional; S_i es el efecto del i-ésimo padre; y $\varepsilon_{(i)j}$ es el error aleatorio NID $(0, \sigma^2)$.

El cuadro de análisis de varianza será pues:

Origen de la variación (O.V.)	Grados de libertad (GL)	Cuadrados medios (CM)	Esperanza de los cuadrados medios (ECM)
Debido a padres	S − 1	CM_s	$\sigma^2 + K\,\sigma_s^2$
Error	N − S	CM_e	σ^2
TOTAL	N − 1		

La estimación del coeficiente k sería entonces:

$$K = [1/(S-1)]\left(N - \frac{\Sigma n_i^2}{N}\right)$$

$$= \frac{1}{4}\left[15 - \frac{1}{15}(1^2 + \ldots + 2^2)\right] = 2.83$$

Substituyendo el valor de 2.83 por el de K en la columna ECM la estimación de los componentes de varianza puede obtenerse por la solución simultánea de las ecuaciones.

$$CM_s = \sigma^2 + 2.83\,\sigma_s^2$$

$$CM_e = \sigma^2$$

Así que los componentes de varianza para este diseño serán:

$$\sigma^2 = CM_e$$

y

$$\sigma_s^2 = \frac{1}{2.83}\left[CM_s - CM_e\right]$$

o presentado en forma matricial, los componentes de varianza se obtendrían resolviendo el siguiente sistema:

$$\begin{bmatrix} CM_s \\ CM_e \end{bmatrix} = \begin{bmatrix} 1 & 2.83 \\ 1 & 0.00 \end{bmatrix} \begin{bmatrix} \sigma^2 \\ \sigma_s^2 \end{bmatrix}$$

B. Diseño Factorial p × q.

Supónganse que existen dos factores: factor A con tres niveles y factor B con dos niveles, esto es un 3 × 2 factorial, que los factores (A y B) son aleatorios, y que cada combinación (ij) presenta diferentes números de observaciones, es decir:

		FACTOR A (i)			
		1	2	3	$n_{.j}$
FACTOR B (j)	1	1(n_{ij})	2	4	7
	2	3	1	2	6
	$n_{i.}$	4	3	6	13 (n..)

El modelo en este diseño sería:

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + \epsilon_{(ij)k}$$

donde Y_{ijk} es la k-ésima observación del i-ésimo factor A y el j-ésimo factor B; A_i es el efecto del i-ésimo nivel del factor A; B_j es el efecto del j-ésimo nivel del factor B; AB_{ij} es el efecto de la interacción entre el i-ésimo nivel del factor A y el j-ésimo nivel del factor B; $\varepsilon_{(ij)k}$ es el error aleatorio NID $(0, \sigma^2)$. El análisis de varianza para este modelo será:

O.V.	GL	CM	ECM
Factor A	(i − 1)	CM_A	$\sigma^2 + K_4\,\sigma^2_{AB} + K_5\,\sigma^2_A$
Factor B	(j − 1)	CM_B	$\sigma^2 + K_2\,\sigma^2_{AB} + K_3\,\sigma^2_B$
AXB	(i − 1) (j − 1)	CM_{AB}	$\sigma^2 + K_1\,\sigma^2_{AB}$
Error	$\sum_i \sum_j (n_{ij}-1)$	CM_e	σ^2
TOTAL	N − 1		

Para estimar los coeficientes K de acuerdo al método directo presentado por Henderson y conocido como Henderson 1, es necesario contar con la siguiente información: i) Debido al diferente número de observaciones en cada subclase de los factores A y B,

para el factor A:

$$\sum_{i=1}^{3} \frac{\sum_{j=1}^{2} n_{ij}^2}{n_{i.}} = \frac{1}{4}(1^2 + 3^2) + \frac{1}{3}(2^2 + 1^2) + \frac{1}{6}(4^2 + 2^2) = 7.5$$

para el factor B:

$$\sum_{j=1}^{2} \frac{\sum_{i=1}^{3} n_{ij}^2}{n_{j}} = \frac{1}{7}(1^2 + 2^2 + 4^2) + \frac{1}{6}(3^2 + 1^2 + 2^2) = 5.33$$

ii) Debido a las diferentes observaciones en cada nivel de los factores A y B

para el factor A.

$$\sum_{i=1}^{3} n_{i.}^2/n_{..} = \frac{1}{13}(4^2 + 3^2 + 6^2) = 4.69$$

y para el factor B

$$\sum_{j=1}^{2} n_{j}^2/n_{..} = \frac{1}{13}(7^2 + 6^2) = 6.54$$

iii) Debido a las diferentes observaciones en cada una de las celdas (N_{ij})

$$\frac{\sum_{i=1}^{3}\sum_{j=1}^{2} n_{ij}^2}{n_{..}} = \frac{1}{13}(1^2 + 2^2 + \ldots + 2^2) = 2.69$$

Con esta información los valores de K serán:

$$K_1 = \frac{1}{gl\,(AB)}\left[n_{..} - \sum_i \frac{\sum_j n_{ij}^2}{n_{i.}} - \sum_j \frac{\sum_i n_{ij}^2}{n_{j}} + \frac{\sum_i \sum_j n_{ij}^2}{n_{..}}\right] = \frac{1}{2}[13 - 7.5 - 5.33 + 2.69] = 1.43$$

$$K_2 = \frac{1}{gl\,(B)}\left[\sum_j \frac{\sum_i n_{ij}^2}{n_{j}} - \frac{\sum_i \sum_j n_{ij}^2}{n_{..}}\right] = \frac{1}{1}(5.33 - 2.69) = 2.64$$

$$K_3 = \frac{1}{gl\,(B)}\left[n_{..} - \frac{\sum_j n_{j}^2}{n_{..}}\right] = \frac{1}{1}(13 - 6.54) = 6.46$$

$$K_4 = \frac{1}{gl\,(A)}\left[\sum_i \frac{\sum_j n_{ij}^2}{n_{i.}} - \frac{\sum_i \sum_j n_{ij}^2}{n_{..}}\right] = \frac{1}{2}(7.5 - 2.69) = 2.41$$

$$K_5 = \frac{1}{gl\,(A)}\left[n_{..} - \frac{\sum_i n_{i.}^2}{n_{..}}\right] = \frac{1}{2}(13 - 4.69) = 4.16$$

Por lo tanto, los estimadores de los componentes de varianza pueden obtenerse resolviendo el siguiente sistema de ecuaciones lineales.

$$\begin{bmatrix} CM_A \\ CM_B \\ CM_{AB} \\ CM_e \end{bmatrix} = \begin{bmatrix} 1 & 2.41 & 0 & 4.16 \\ 1 & 2.64 & 6.46 & 0 \\ 1 & 1.43 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \widehat{\sigma}^2 \\ \widehat{\sigma}^2_{AB} \\ \widehat{\sigma}^2_B \\ \widehat{\sigma}^2_A \end{bmatrix}$$

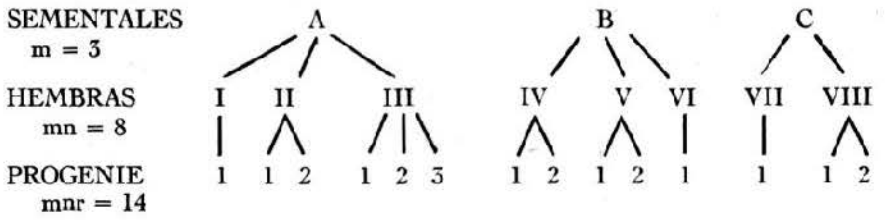
Hay que hacer notar que en los diseños balanceados $K_1 = K_2 = K_4$.

C. Diseño Jerárquico o Anidado.

En algunos experimentos, todos los niveles de un cierto factor son diferentes a través de los niveles del otro factor. Frecuentemente los niveles son escogidos al azar para cada factor; sin embargo, esto no es una condición necesaria para el llamado "diseño jerárquico"; de hecho, este diseño puede ser aleatorio, mixto o fijo. Se debe recalcar que este diseño es otro tipo de diseño completamente aleatorio.

Ejemplo:

Si mn hembras son escogidas al azar de la población, n hembras son cruzadas con cada uno de los m sementales, suponiendo que todos los individuos involucrados en los mn cruzamientos son escogidos al azar; y de cada mn cruzas, r individuos son productos donde n (número de hembras en cada semental) y r (número de progenie en cada cruza) pueden ser diferentes en una o más subclases, por lo tanto, esto nos lleva a un diseño desbalanceado.



Nótese que el semental A fue cruzado con las hembras I, II y III y que cada cruzamiento dio una, dos y tres progenies respectivamente. Asimismo, el semental B fue cruzado con las hembras IV, V y VI obteniendo dos, dos y una progenie, respectivamente, etc.

El modelo involucrado en este diseño es:

$$Y_{mnr} = \mu + S_m + D_{(m)n} + O_{(mn)r}$$

Donde: Y_{mnr} es el valor fenotípico observado en la r-ésima progenie en la n-ésima hembra y el m-ésimo semental; μ es la media poblacional; S_m es el efecto del m-ésimo semental; $D_{(m)n}$ es el efecto de la n-ésima hembra anidado en el m-ésimo semental; $O_{(mn)r}$ es el efecto de la r-ésima progenie anidada en la n-ésima madre y el m-ésimo semental.

El análisis de varianza para este modelo es:

O.V.	GL	CM	ECM
Semental (S)	m − 1	CM_S	$\sigma^2 + K_2\,\sigma^2_D + K_3\,\sigma^2_S$
Madre/Semental (D)	m (n − 1)	CM_D	$\sigma^2 + K_1\,\sigma^2_D$
Progenie/D/S (O)	mn (r − 1)	CM_O	σ^2
TOTAL	n − 1		

Los coeficientes K pueden ser estimados como sigue:

$$K_1 = \frac{1}{gl(D)}\left[n.. - \sum_{i=1}^{a}\left(\sum_{j=1}^{n_i} n_{ij}^2/n_{i.}\right)\right]$$

$$= \frac{1}{5}\left[14 - \left(\frac{1}{6}(1^2+2^2+3^2)+\frac{1}{5}(2^2+2^2+1^2)+\frac{1}{3}(1^2+2^2)\right)\right] = 1.64$$

$$K_2 = \frac{1}{gl\,(S)}\left[\sum_{i=1}^{a}\left(\sum_{j=1}^{n_i} n_{ij}^2/n_{i.}\right) - \left(\frac{\sum_{ij} n_{ij}^2}{n_{..}}\right)\right] = \frac{1}{2}\left[\frac{1}{6}(1^2+2^2+3^2)\right.$$

$$\left. + \frac{1}{5}(2^2+2^2+1^2)+\frac{1}{3}(1^2+2^2) - \frac{1}{14}(1^2+..+2^2)\right] = 1.90$$

$$K_3 = \frac{1}{gl\,(S)}\left[n.. - \sum_i n_{i.}^2/n..\right] = \frac{1}{2}\left[14 - \frac{1}{4}(6^2+5^2+3^2)\right] = 4.50$$

Los componentes de varianza pueden ser estimados por la solución del siguiente sistema de ecuaciones lineales:

$$\begin{bmatrix} CM_S \\ CM_D \\ CM_O \end{bmatrix} = \begin{bmatrix} 1 & 1.90 & 4.50 \\ 1 & 1.64 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\sigma}^2 \\ \hat{\sigma}_D^2 \\ \hat{\sigma}_S^2 \end{bmatrix}$$

Hay que hacer notar que para los diseños balanceados las constantes K_1 y K_2 son iguales.

Summary

This paper presents in the form of examples, the methods for estimating variance components for the one way classification, two way classification (p × q factorial), Henderson 1 and hierarchical designs with unequal subclass numbers, assuming random effects in each of the models.

Literatura citada

ANDERSON, V.L. and R.A. MCLEAN, 1974, Design of Experiments. A Realistic Approach. *Marcel Dekker, Inc.*, N.Y.

BENNETT, C.A. and N.L. FRANKLIN, 1954, Statistical Analysis in Chemistry and the Chemical Industry. Wiley, N.Y.

HARVEY, R.W., 1975, Least-Squares Analysis of Data with Unequal Subclass Numbers. AHS H-4 Agriculture Research Service. *U.S. Department of Agriculture.*

HENDERSON, C.R., 1953, Estimation of Variance and Covariance Components. *Biometrics*, 9, 226-252.

SEARLE, S.R., 1968, Another Look at Henderson's Methods of Estimating Variance Components *Biometrics*, 24, 749-488.

SEARLE, S.R., 1971, Topics in Variance Component Estimation. *Biometrics*, 27, 1-76.

SEARLE, S.R. and C.R. HENDERSON, 1961, Computing Procedures for Estimating Components of Variance in the Two-way Clasification, Mixed Model. *Biometrics*, 17, 6-70616.

STEELE, G.D.R. and J.H. TORRIE, 1980, Principles and Procedures of Statistics. A Biometrical Approach. 2nd Ed. *McGraw-Hill Book, Co.*