

Contraste de hipótesis en diseños de bloques incompletos desequilibrados

Guillermo Vallejo y José Ramón Escudero
Universidad de Oviedo

El diseño de bloques al azar capacita al investigador para reducir la varianza del error y obtener estimaciones más precisas de los efectos de los tratamientos. En la mayoría de los experimentos el tamaño de los bloques debe ser reducido para lograr la homogeneidad de las unidades experimentales e incrementar la precisión del diseño. En estos casos es posible seguir utilizando diseños de bloques en los cuales una parte de los tratamientos es asignada a los bloques conforme algún patrón especificado de antemano. Estos diseños son conocidos como diseños de bloques incompletos equilibrados. Sin embargo, también puede ocurrir que por accidente o cualquier otra razón los investigadores se vean en la obligación de emplear diseños de bloques incompletos, pero carentes de equilibrio. En el presente trabajo se muestra el proceso a seguir en este último caso, para determinar qué funciones y contrastes son estimables y cuáles de las hipótesis son de interés por resultar interpretables.

Testing of hypotheses in unbalanced incomplete block designs. The randomized complete block design enables an experimenter to reduce the error variance and obtain more precise estimates of treatment effects. In many experiments the size block must be small to obtain homogeneity of the experimental units, and increase the precision of the design. In this case it is possible to use randomized block designs in which every treatment is not present in every block, although if are presented according to some specified pattern. These designs are known as balanced or partially balanced incomplete block designs. However, it may happen by accident or some other reason, that the investigators is forced to use incomplete block designs, but unbalanced. The purpose of this work is to show a method that allows determining the functions and contrasts that are estimable and the associated hypotheses that are of general interest.

El método más eficaz para soslayar la heterogeneidad de las unidades experimentales en todo tipo de experimentación, a la vez que pilar fundamental en la que se asienta el diseño experimental moderno, se conoce con el nombre de modelo de bloques al azar. Como resulta conocido, dicho diseño implica replicar el experimento, al menos, tantas veces como niveles tenga la variable de bloqueo. Sin embargo, en ciertas situaciones los investigadores son incapaces de asignar todos los tratamientos a cada uno de los bloques. En principio, esta situación puede acontecer por razones de índole muy diversa, no obstante, los problemas más frecuentes surgen por limitaciones en los aparatos y, sobre manera, por no disponer del adecuado número de unidades experimentales para lograr que la homogeneidad dentro de los bloques sea la mayor posible. Para resolver estas cuestiones, y otras de carácter similar, es factible usar diseños de bloques al azar en los cuales cada tratamiento no se encuentre presente a lo largo de todas las dimensiones de la variable de bloqueo. Estos diseños son conocidos como diseños de bloques incompletos al azar. Y si bien es cierto que en

la investigación psicológica su empleo es prácticamente nulo, no es menos cierto que en ocasiones constituirían la opción más adecuada para abordar la cuestión reseñada. Trabajos con un agradable sabor aplicado donde se pueden encontrar estos diseños son los de Lentner y Bishop (1986) y John y Quenouille (1977). Por su parte, el libro de Das y Giri (1979) da buena cuenta de los principales diseños de bloques incompletos, de su análisis y de los métodos más relevantes utilizados en su construcción.

Por regla general, la mayoría de los textos especializados recomiendan reducir el tamaño de los bloques utilizando los diseños de bloques incompletos equilibrados introducidos por Yates (1936) y los diseños de bloques incompletos parcialmente equilibrados introducidos por Bose y Nair (1939). Tanto unos como otros, requieren que las unidades experimentales sean asignadas a los bloques conforme algún patrón convenientemente especificado con anterioridad a la realización del experimento. En concreto, el equilibrio propio del primer grupo de diseños consiste en que cada par de tratamientos ocurra el mismo número de veces a lo largo de los diferentes niveles de la variable de bloqueo (λ). Supongamos que la variable de tratamiento consta de p niveles y que en cada bloque sólo podemos manejar exactamente k ($k < p$) tratamientos. La condición necesaria y suficiente para construir un diseño de bloques incompleto equilibrado consiste en seleccionar un número de bloques que coincida con el total de combinaciones posibles de p tratamientos tomados de k en k . Sin embargo, frecuentemente esta

clase de diseños pueden obtenerse con un menor número de bloques con tal de que se satisfaga: (a) $pr = bk$, (b) $\lambda = r(k-1) / p-1$ y (c) $b \geq p$. Donde b denota el número de bloques y r el número de veces que cada nivel de tratamientos aparece a lo largo de los diferentes bloques. A su vez, el patrón de balanceo de los diseños de bloques incompletos parcialmente equilibrados consiste en que algunos pares de tratamientos aparezcan juntos en cada bloque λ_1 veces, otros pares aparezcan juntos λ_2 veces, ..., y los pares restantes aparezcan juntos λ_m veces. Como consecuencia, estos diseños exigen un menor número de bloques que los diseños de bloques incompletos equilibrados, aunque la eficiencia de estos últimos es la misma para cada par de tratamientos, mientras que la de los diseños parcialmente equilibrados es distinta para cada par de asociados.

Sin restar importancia a lo expuesto, en la fase de experimentación propiamente dicha también puede ocurrir, bien sea por accidente o bien sea por cualquier otra razón, que determinadas observaciones se pierdan. Consecuentemente, quiebran los requisitos combinatorios requeridos por estos diseños para alcanzar el equilibrio. Por ejemplo, un paciente puede abandonar el tratamiento, un aparato puede fallar en el registro de determinados datos, uno o más animales mueren durante su manipulación, algunos sujetos no siguen adecuadamente las instrucciones, etc. Lo cierto es que situaciones como las descritas no les son en absoluto extrañas a los investigadores, aunque no suelen aludir a ellas. La pérdida de observaciones de un modo arbitrario conlleva que no todas las usuales funciones paramétricas pueden ser observadas. Tales situaciones, a menos que sean descubiertas conducen a una incorrecta asignación de los grados de libertad correspondientes a determinadas fuentes de variación, a la obtención de matrices singulares que, por cierto, muchos programas de ordenador invierten sin advertir de tal anomalía al usuario y, sobre todo, a que efectuemos contrastes con sumas de cuadrados que no son adecuadamente conocidas. Para intentar abordar un panorama como el descrito al investigador se le ofrecen varias salidas. Una de las más novedosas, y también de las más fáciles de implementar, es la que se muestra en el presente manuscrito siguiendo los trabajos de Dodge y Shah (1977) y de Dodge (1985).

Estimación de funciones y contrastes

Para estimar la función de respuestas de cada uno de los tratamientos implicados en el diseño de bloques incompleto, un apropiado modelo lineal estándar puede ser escrito como sigue:

$$y_{ijk} = \mu + \alpha_j + \beta_k + \epsilon_{ijk}, \quad (i = 1, \dots, n_{jk}; j = 1, \dots, p, k = 1, \dots, q) \tag{1}$$

donde, μ , α_j y β_k son parámetros desconocidos y los errores ϵ_{ijk} son variables aleatorias con media cero e igual varianza. También se asume que $n_j = \sum_{k=1}^q n_{jk} \neq 0$ para todo $k = 1, \dots, q$, que $n_k = \sum_{j=1}^p n_{jk} \neq 0$ para todo $j = 1, \dots, p$ y que cada fila y columna de la matriz de datos contiene al menos una observación.

El modelo de la ecuación (1) también puede ser especificado en forma matricial por

$$E(\mathbf{y}) = \mathbf{j}\mu + \mathbf{X}_1\alpha + \mathbf{X}_2\beta = \mathbf{X}\theta \tag{2}$$

Aquí el vector de observaciones \mathbf{y} es de dimensión $N \times 1$ y la matriz de diseño \mathbf{X} ($\mathbf{j}, \mathbf{X}_1, \mathbf{X}_2$) es de dimensión $N \times (1+p+q)$. Es-

trechamente relacionada con la matriz de diseño \mathbf{X} se halla la matriz de incidencia \mathbf{N} , la cual especifica el número de observaciones efectuadas bajo los niveles j -ésimo y k -ésimo de las variables de tratamiento y bloqueo, respectivamente. En concreto, $\mathbf{N} = \mathbf{X}'_1\mathbf{X}_2$.

Dicho lo anterior, nos centraremos, inicialmente, en mostrar qué funciones $\mu + \alpha_j + \beta_k$ o valores esperados de y_{ijk} son estimables; posteriormente, haremos lo propio con los contrastes o diferencias entre pares de medias. En el caso que nos ocupa, una función paramétrica puede ser definida como una combinación lineal de parámetros $\mu, \alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_q$. Tal función es estimable si se puede expresar como una combinación lineal de las expectativas $\mu + \alpha_j + \beta_k$ de las celdas de la matriz de incidencia para las cuales existen observaciones disponibles. Actualmente existen distintos procedimientos para saber que expectativas son realmente estimables (ver, por ejemplo, Milliken y Johnson, 1984; Searle, 1997), no obstante, de los procedimientos examinados por nosotros, el algoritmo o *proceso R* introducido por Birkes, Dogd-ge y Seely (1976) es el más completo.

El *proceso R* es un bucle iterativo que se aplica sobre la matriz \mathbf{N} para obtener una matriz final, \mathbf{M} , que informa de las celdas que contienen funciones paramétricas estimables. Dicho proceso, además de proporcionar una base para las funciones estimables de cada uno de los efectos implicados en el diseño, también permite determinar los grados de libertad asociados con las hipótesis de interés y las porciones del diseño que están conectadas. De acuerdo con los autores citados, el algoritmo *R*, cuyo funcionamiento es muy similar al que rige en el popular juego de las cuatro esquinas, implica los cuatro pasos que siguen: En primer lugar, se parte de una matriz de ceros \mathbf{M} de dimensión $p \times q$. En segundo lugar, cada entrada de la matriz \mathbf{M} toma el valor uno si la correspondiente entrada de la matriz de incidencia \mathbf{N} es distinta de cero. Esto es, $m_{jk} = 1$ si $n_{jk} \neq 0$. En tercer lugar, cada entrada de la matriz \mathbf{M} toma el valor uno si existe alguna fila, columna e interacción de ambas que valgan uno; esto es, para cada par de valores (j, k) si existen r y s , tal que $m_{js} = m_{rs} = m_{rk} = 1$, entonces $m_{jk} = 1$. Por último, se repite sucesivamente el paso 3, usando los nuevos m_{jk} distintos de cero como ángulos del rectángulo hasta que ninguna entrada pueda ser modificada. A partir de la matriz final \mathbf{M} , se colige rápidamente que una función paramétrica $\mu + \alpha_j + \beta_k$ es estimable sí y sólo sí $m_{jk} = 1$. Además de los citados Birke *et al.* (1976), Dodge (1985) también presenta una demostración de lo dicho. Para ilustrar lo dicho, a continuación, se muestra la matriz final \mathbf{M} obtenida a partir de una hipotética matriz de incidencia \mathbf{N} en la cual el número de niveles de la variable de tratamiento y de la variable de bloqueo son los mismos; en concreto, $p = 6$ y $q = 6$.

$$N = \begin{bmatrix} 0 & 2 & 0 & 0 & 3 & 0 \\ 2 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 2 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 \\ 3 & 0 & 2 & 0 & 0 & 0 \end{bmatrix} \quad M = \begin{bmatrix} 1^* & 1 & 1^* & 0 & 1 & 1^* \\ 1 & 1 & 1^* & 0 & 1^* & 1^* \\ 1^* & 1^* & 1 & 0 & 1^* & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1^* & 1 & 1^* & 0 & 1^* & 1^* \\ 1 & 1^* & 1 & 0 & 1^* & 1^* \end{bmatrix} \tag{3}$$

Desde \mathbf{M} resulta fácil verificar las funciones que son realmente estimables. Por ejemplo, se observa que la función $\mu + \alpha_1 + \beta_1$ es actualmente estimable, mientras que la función $\mu + \alpha_4 + \beta_2$ no lo es.

A su vez, mediante sencillas operaciones con la matriz \mathbf{M} también se puede determinar qué contrastes son estimables. En

concreto, una diferencia entre cualquier par de niveles de la variable de interés resulta estimable si el elemento jk de la matriz S ($S = MM'$) es distinto de cero. Lo dicho es una consecuencia de lo anterior y se puede verificar teniendo en cuenta que si $s_{jk} \neq 0$, debe existir algún l tal que $m_{jk} = m_{lk} = 1$. De este modo, las funciones paramétricas $\mu + \alpha_j + \beta_k$ y $\mu + \alpha_l + \beta_k$ tienen que ser estimables y, por extensión, las diferencias entre ambas expectativas. Desde la matriz S se puede descubrir rápidamente cuales de los $p \times (p-1)/2$ pares de tratamientos existentes son estimables. Para ello tan sólo hay que reemplazar por unos todos los elementos distintos de cero que se hallen por debajo de la diagonal principal de S y colocar en las filas de la matriz triangular inferior resultante los valores $\alpha_j = \alpha_2, \alpha_3, \dots, \alpha_p$ y en las columnas los valores $\alpha_j = \alpha_1, \alpha_2, \dots, \alpha_{p-1}$. A continuación, utilizamos esta última matriz con apariencia de triángulo contador para determinar, por un lado, los contrastes que son estimables y, por otro lado, cuántos pueden ser simultáneamente estimados de forma independiente. Lo primero puede ser hecho verificando la intersección de filas y columnas del triángulo que contienen unos, mientras que lo segundo implica contar las filas cuyos elementos no sean todos ceros. Dicha base coincide con los grados de libertad de la variable de interés, en nuestro caso con f_a . En la Figura 1 se muestra el triángulo contador para las diferencias de α del ejemplo citado más arriba.

	α_1	α_2	α_3	α_4	α_5
α_2	1				
α_3	1	1			
α_4	0	0	0		
α_5	1	1	1	0	
α_6	1	1	1	0	1

Figura 1. Triángulo contador para determinar los contrastes α estimables

De acuerdo con dicha Figura 1, se observa que $\alpha_1 - \alpha_2$ es estimable, mientras que $\alpha_1 - \alpha_4$ no lo es. También se aprecia que el número de filas cuyos elementos no son todos nulos son cuatro, por tanto, $f_a = 4$. Se excusa decir que un procedimiento similar se puede utilizar para calcular f_b , pero sustituyendo S por $M'M$.

Determinación de las sumas de cuadrados e hipótesis asociadas

En esta sección se genera una solución g-inversa para resolver el sistema de ecuaciones normales $X'X\theta = X'y$ correspondiente al modelo de efectos de la ecuación (2). También se presentan las sumas de cuadrados requeridas para el análisis de la varianza y las hipótesis que cada fuente tiene asociada.

Para encontrar una solución a los parámetros del vector θ vamos a escribir en forma matricial el sistema de ecuaciones normales

$$\begin{bmatrix} J'J & J'X_1 & J'X_2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} J'y \\ X'_1y \\ X'_2y \end{bmatrix} \tag{4}$$

donde $J'J$ es la suma de las entradas de la matriz N , los vectores $J'X_1$ y $J'X_2$ contienen el número de observaciones de las filas y de las columnas, respectivamente, de la matriz de incidencia N y X'_1X_1 y X'_2X_2 son matrices diagonales cuyas entradas están configuradas con las sumas de las filas y de las columnas de la matriz N .

En el modelo sobreparametrizado que viene siendo considerado el orden de X es mayor que su rango, en concreto $R(X) = f_a + q$, con lo que $X'X$ no es invertible. De acuerdo con Searle (1997), esto se puede solucionar de una manera relativamente fácil particionando la matriz $X'X$ y construyendo una submatriz cuyo rango sea igual a $f_a + q$. Para ello se redefine N_q como N , X'_qX_q como X'_2X_2 y $J'X_q$, $\hat{\beta}_q$ y X'_qy como $J'X_2$, B y X'_2y con sus j -ésimas filas y/o k -ésimas columnas eliminadas. Teniendo en cuenta las modificaciones reseñadas el sistema de ecuaciones normales se puede reescribir

$$\begin{bmatrix} J'J & J'X_1 & J'X_q & J'X_0 \\ X'_1J & X'_1X_1 & N_q & C_q \\ X'_qJ & N'_q & X'_qX_q & O \\ X'_0J & C'_q & O & X'_0X_0 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha} \\ \hat{\beta}_q \\ \hat{\beta}_0 \end{bmatrix} = \begin{bmatrix} J'y \\ X'_1y \\ X'_qy \\ X'_0y \end{bmatrix} \tag{5}$$

donde C_q son las k -ésimas columnas eliminadas de N . Las ecuaciones son resueltas utilizando la matriz

$$G = \begin{bmatrix} O & O & O & O \\ O & \begin{pmatrix} X'_1X_1 & N_q \end{pmatrix}^{-1} & O & O \\ O & \begin{pmatrix} N'_q & X'_qX_q \end{pmatrix} & O & O \\ O & O & O & O \end{bmatrix} \tag{6}$$

como una inversa generalizada. Sustituyendo la matriz de la parte izquierda de la ecuación (5) por la inversa de G , es inmediato que $\hat{\mu} = 0$ y $\hat{\beta}_0 = 0$ y

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} X'_1X_1 & N_q \\ N'_q & X'_qX_q \end{bmatrix}^{-1} \begin{bmatrix} X'_1y \\ X'_qy \end{bmatrix}$$

$$\tilde{\theta} = (\tilde{X}'\tilde{X}')^{-1} \tilde{X}'y \tag{7}$$

con

$$\hat{\alpha} = (X'_1X_1)^{-1} (X'_1y - N_q\hat{\beta}_q)$$

$$y \hat{\beta}_q = (X'_qX_q - N'_q(X'_1X_1)^{-1}N'_q)^{-1} (X'_qy - N'_q(X'_1X_1)^{-1}X'_1y)$$

Por último, en la Tabla 1 presentamos la partición de la suma de cuadrados total e hipótesis asociadas con las fuentes del modelo. Siguiendo el trabajo de Speed y Hocking (1978), para simplificar los aspectos computacionales utilizaremos la notación $R(\cdot)$. Dicha cantidad implica la reducción que se produce en la suma de cuadrados como resultado de ajustar diferentes modelos.

Observando detenidamente la Tabla 1 se aprecia rápidamente las hipótesis que tienen realmente interés. En nuestro caso las asociadas con los tratamientos y bloques ajustados, ya que ambas tienen una interpretación clara; las hipótesis restantes al no ofrecer ninguna interpretación razonable carecerían de interés.

Tabla 1
Partición de la suma de cuadrados e hipótesis asociadas con el diseño

Fuentes de variación	Suma de cuadrados	G.L.	Hipótesis asociadas
R(μ, α, β)	$((\tilde{X}'\tilde{X})^{-1} \tilde{X}'\tilde{y})' \tilde{X}'\tilde{y}$	R(X)	
R(μ)	$(j'j)^{-1} j'y$	1	$H_0: \mu + \sum_j n_j \alpha_j + \sum_k n_k \beta_k = 0$
R(β/μ)	$((X'_{22} X_2)^{-1} X'_2 y) - R(\mu)$	q-1	$H_0: \beta_k + \sum_j n_{jk} \alpha_j / n_k$ igual $\forall k$
R($\alpha/\mu, \beta$)	$R(\mu, \alpha, \beta) - ((X'_{22} X_2)^{-1} X'_2 y)' X'_2 y$	f_α	$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_p$
R(μ, α, β)	$((\tilde{X}'\tilde{X})^{-1} \tilde{X}'\tilde{y})' \tilde{X}'\tilde{y}$	R(X)	
R(μ)	$(j'j)^{-1} j'y$	1	$H_0: \mu + \sum_j n_j \alpha_j + \sum_k n_k \beta_k = 0$
R(α/μ)	$((X'_{11} X_1)^{-1} X'_1 y) - R(\mu)$	p-1	$H_0: \alpha_j + \sum_k n_{jk} \beta_k / n_j$ igual $\forall j$
R($\beta/\mu, \alpha$)	$R(\mu, \alpha, \beta) - ((X'_{11} X_1)^{-1} X'_1 y)' X'_1 y$	f_β	$H_0: \beta_1 = \beta_2 = \dots = \beta_q$
Residual	$y'y - ((\tilde{X}'\tilde{X})^{-1} \tilde{X}'\tilde{y})' \tilde{X}'\tilde{y}$	N-R(X)	
Error	$y'y - ((\tilde{X}'\tilde{X})^{-1} \tilde{X}'\tilde{y})' \tilde{X}'\tilde{y}$	N-(p+1)	
No-aditividad	Residual - Error	(p+q)-R(X)	
Total	$y'y$	N	

Agradecimientos

Este trabajo ha sido financiado mediante un proyecto de investigación concedido por el MCYT (Ref.: B05-2000-0410) y FICYT (PR-01-GE-2).

Referencias

Birkes, D., Dodge, Y. y Seely, J. (1976). Spanning sets for estimable contrasts in classification models. *Annals of Statistics*, 4, 86-107.

Bose, R. C. y Nair, K. R. (1939). Partially balanced incomplete block designs. *Sankhya*, 4, 337.

Das, M. N. y Giri, N. C. (1979). *Design and Analysis of Experiments*. New York: Halstead Press.

Dodge, Y. (1985). *Analysis of Experiments with Missing Data*. New York: John Wiley & Sons, Inc.

Dodge, Y. y Shah, K. R. (1977). Estimation of parameters in latin squares and greco-latin squares with missing observations. *Communications in Statistics-Theory and Methods*, 6, 1.465-1.472.

John, J. A. y Quenouille, M. H. *Experiments: Design and Analysis*. London: Mc Millan.

Lenter, M. y Bishop, T. (1986). *Experimental Design and Analysis*. Blacksbrg: Walley Book Co.

Milliken, G. A. y Johnson, D. E. (1984). *Analysis of Messy Data: Designed Experiments*. New York: Van Nostrand Reinhold.

Searle, S. R., (1997). *Linear Models*. Wiley Classics Library. New York: John Wiley & Sons, Inc.

Speed, F. M. y Hocking, R. R. (1976). The use of the R(.) notation with unbalanced data. *The American Statistician*, 30, 30-34.

Yates, F. (1936). Balanced incomplete randomized blocks. *Annals of Eugenics*, 7, 121-140.

Aceptado el 16 de octubre de 2001