

Diseño por bloques aleatorizados completos

Introducción

Al estudiar la influencia de un factor sobre una variable cuantitativa es frecuente que aparezcan otras variables o factores que también influyen y que deben ser controladas.

A estas variables se las denomina variables *bloque*, y se caracterizan por

- (i) No son el motivo del estudio sino que aparecen de forma natural y obligada en el mismo.
- (ii) Se asume que no tienen interacción con el factor en estudio.

Ejemplo. Se quieren determinar las necesidades energéticas de una persona cuando anda, come o hace deporte. Supongamos que se tienen 10 personas para realizar el experimento y se considera como variable respuesta o cuantitativa, el número de calorías consumidas por segundo. Los resultados varían según el individuo considerado. Aquí, el factor es la actividad realizada, con 3 posibles niveles: andar, comer o hacer gimnasia.

Si a cada una de las personas se le asigna una actividad distinta puede ser que la variabilidad observada entre las distintas actividades sea debida a las diferencias entre los propios individuos. Una posible solución es que cada uno de los individuos realice las tres actividades. De este modo, la variable bloque es el *tipo de persona* y cada uno de los bloques es cada persona.

A cada bloque (persona) se le aplican los 3 niveles del factor por orden aleatorio:

Persona 1	Persona 2	...	Persona 10
C	G	...	A
A	A	...	G
G	C	...	C

Observaciones

Una variable bloque no presenta interacción con los factores en estudio.

El modelo se dice que es de *bloques aleatorizados completos* cuando en cada bloque se presentan todos los posibles tratamientos (o un múltiplo de ese número) y dentro de cada bloque se asignan los tratamientos de forma aleatoria.

En ocasiones no se pueden asignar todos los tratamientos sobre cada bloque, de modo que se tienen los diseños por *bloques aleatorizados incompletos*.

Modelo

Suponemos que el número de unidades experimentales para cada bloque coincide con el número de tratamientos, esto es, hay una observación para cada cruce de los niveles del factor y del bloque. La variable respuesta Y puede depender de un primer factor de interés (A) y de la variable bloque (B). El modelo es:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

para $i = 1, \dots, a$ y $j = 1, \dots, b$, siendo:

- μ el efecto medio global
- α_i el efecto incremental sobre la media causado por el nivel i del factor A
- β_j el efecto incremental sobre la media causado por el nivel j del bloque B
- ε_{ij} el término de error.

Vamos a suponer que

$$\sum_{i=1}^a \alpha_i = \sum_{j=1}^b \beta_j = 0$$

El problema consiste en comparar las medias de los tratamientos, esto es

$$H_0 \equiv \mu_1 = \mu_2 = \dots = \mu_a$$

$$H_1 \equiv \mu_i \neq \mu_j \quad i \neq j$$

lo cual es equivalente a

$$H_0 \equiv \alpha_i = 0$$

$$H_1 \equiv \alpha_j \neq 0 \quad \exists j$$

Se consideran las siguientes hipótesis sobre el modelo:

- **Normalidad:** ε_{ij} sigue una distribución normal. Esto es equivalente a que Y_{ij} sigue una distribución normal.
- **Linealidad:** $E(\varepsilon_{ij}) = 0$. Esto es equivalente a que $E(Y_{ij}) = \mu + \alpha_i + \beta_j$.
- **Homocedasticidad:** $Var(\varepsilon_{ij}) = \sigma^2$. Esto es equivalente a que $Var(Y_{ij}) = \sigma^2$.
- **Independencia:** ε_{ij} son independientes entre sí. Esto es equivalente a que Y_{ij} son independientes entre sí.

Estimación de los parámetros

Se estiman por el método de los mínimos cuadrados.

Se calcula

$$\min_{\mu, \alpha_i, \beta_j} \phi = \min \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \mu - \alpha_i - \beta_j)^2$$

sujeto a

$$\sum_{i=1}^a \alpha_i = \sum_{j=1}^b \beta_j = 0$$

Se tiene

$$\begin{aligned}\frac{\partial \phi}{\partial \mu} &= -2 \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \mu - \alpha_i - \beta_j) = 0 \implies \\ \sum_{i=1}^a \sum_{j=1}^b y_{ij} - ab\mu &= 0 \implies \\ \hat{\mu} &= \frac{1}{ab} \sum_{i=1}^a \sum_{j=1}^b y_{ijk} = \bar{y}_{..}\end{aligned}$$

Para i fijado

$$\begin{aligned}\frac{\partial \phi}{\partial \alpha_i} &= -2 \sum_{j=1}^b (y_{ij} - \mu - \alpha_i - \beta_j) = 0 \implies \\ \sum_{j=1}^b y_{ij} - b\mu - b\alpha_i &= 0 \implies \\ \hat{\alpha}_i &= \frac{1}{b} \sum_{j=1}^b y_{ij} - \bar{y}_{..} = \bar{y}_{i.} - \bar{y}_{..}\end{aligned}$$

Análogamente, para j fijado,

$$\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$$

Así,

$$\begin{aligned}\hat{y}_{ij} &= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) = \\ &= \bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..}\end{aligned}$$

El número de parámetros a estimar en total es

$$1 + (a - 1) + (b - 1) = a + b - 1$$

Como el número total de observaciones es $N = ab$, entonces el número de grados de libertad es

$$ab - a - b + 1 = (a - 1)(b - 1)$$

De este modo, como

$$\begin{aligned} SCE &= \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - (\bar{y}_{i\cdot} + \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}))^2 = \\ &= \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2, \end{aligned}$$

entonces la estima de la varianza total es

$$\hat{\sigma}^2 = \frac{1}{(a-1)(b-1)} \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2.$$

Descomposición de la suma de cuadrados

Descomponemos la variabilidad total de la siguiente forma:

$$SCT = SC_A + SC_B + SCE$$

siendo

$$\begin{aligned} SCT &= \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^a \sum_{j=1}^b y_{ij}^2 - N\bar{y}_{\cdot\cdot}^2 = \\ &= \sum_{i=1}^a \sum_{j=1}^b y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{N} \equiv \text{“Suma de cuadrados total”} \end{aligned}$$

$$\begin{aligned} SC_A &= b \sum_{i=1}^a (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = b \sum_{i=1}^a \bar{y}_{i\cdot}^2 - N\bar{y}_{\cdot\cdot}^2 = \\ &= \frac{1}{b} \sum_{i=1}^a y_{i\cdot}^2 - \frac{y_{\cdot\cdot}^2}{N} \equiv \text{“Suma de cuadrados explicada debido al factor A”} \end{aligned}$$

$$\begin{aligned} SC_B &= a \sum_{j=1}^b (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 = a \sum_{j=1}^b \bar{y}_{\cdot j}^2 - N\bar{y}_{\cdot\cdot}^2 = \\ &= \frac{1}{a} \sum_{j=1}^b y_{\cdot j}^2 - \frac{y_{\cdot\cdot}^2}{N} \equiv \text{“Suma de cuadrados explicada debido al bloque”} \end{aligned}$$

$$\begin{aligned} SCE &= \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2 = SCT - SC_A - SC_B \\ &\equiv \text{“Suma de cuadrados residual”} \end{aligned}$$

Se puede demostrar que

$$E \left[\frac{SC_A}{a-1} \right] = \sigma^2 + \frac{b \sum_{i=1}^a \alpha_i^2}{a-1}$$

$$E \left[\frac{SC_B}{b-1} \right] = \sigma^2 + \frac{a \sum_{j=1}^b \beta_j^2}{b-1}$$

$$E \left[\frac{SCE}{(a-1)(b-1)} \right] = \sigma^2$$

De este modo, se puede considerar el estadístico

$$F_0 = \frac{MC_A}{MCE} = \frac{\frac{SC_A}{a-1}}{\frac{SCE}{(a-1)(b-1)}}$$

de modo que si H_0 es cierta, F_0 se distribuye como una F de Snedecor, $F_{a-1, (a-1)(b-1), \alpha}$, de manera que se rechaza la H_0 si

$$F_0 > F_{a-1, (a-1)(b-1), \alpha}$$

Se podría plantear contrastar la igualdad entre los bloques, aunque hacer esto está en contra de la suposición previa: si se ha considerado un diseño por bloques es porque estos influyen en el experimento.

Si se considerase un estadístico de la forma

$$F_0 = \frac{MC_B}{MCE} = \frac{\frac{SC_B}{b-1}}{\frac{SCE}{(a-1)(b-1)}}$$

aparecerían dificultades teóricas dado que la aleatorización sólo se realiza dentro de cada bloque.

La tabla de análisis de la varianza queda como:

F. V.	S. C.	G. L.	M. C.	F
Factor A	SC_A	$a - 1$	$MC_A = \frac{SC_A}{a-1}$	$F_0 = \frac{MC_A}{MCE}$
Factor Bloque	SC_B	$b - 1$	$MC_B = \frac{SC_B}{b-1}$	
Residual	SCE	$(a - 1)(b - 1)$	$MCE = \frac{SCE}{(a-1)(b-1)}$	
Total	SCT	$ab - 1$		

Ejemplo. Se realiza un experimento para determinar el efecto de cuatro sustancias químicas diferentes sobre la resistencia de una tela. Las sustancias se emplean como parte del proceso terminal de planchado permanente. Para ello, se escogen cinco muestras de tela y se aplica un diseño aleatorizado por bloques completos mediante la prueba de cada sustancia en un orden aleatorio sobre cada una de las muestras de tela. Se probará la diferencia en las medias utilizando para ello el análisis de la varianza con $\alpha = 0,01$. Los datos aparecen a continuación.

Sustancia\Muestra	1	2	3	4	5	Media
1	1,3	1,6	0,5	1,2	1,1	$\bar{y}_1. = 1,14$
2	2,2	2,4	0,4	2,0	1,8	$\bar{y}_2. = 1,76$
3	1,8	1,7	0,6	1,5	1,3	$\bar{y}_3. = 1,38$
4	3,9	4,4	2,0	4,1	3,4	$\bar{y}_4. = 3,56$
Media	$\bar{y}_{.1} = 2,3$	$\bar{y}_{.2} = 2,53$	$\bar{y}_{.3} = 0,88$	$\bar{y}_{.4} = 2,2$	$\bar{y}_{.5} = 1,9$	$\bar{y}_{..} = 1,96$

El factor de interés es la sustancia química, con cuatro niveles y el factor bloque es la muestra de tela, con cinco niveles. Entonces $a = 4$, $b = 5$ y $n = 20$.

Las sumas de cuadrados son:

$$SCT = \sum_{i=1}^a \sum_{j=1}^b y_{ij}^2 - n\bar{y}_{..}^2 = 1,3^2 + 1,6^2 + \dots + 4,1^2 + 3,4^2 - 20 \times 1,96^2 = 25,69$$

$$SC_A = b \sum_{i=1}^a \bar{y}_i.^2 - n\bar{y}_{..}^2 = 5(1,14^2 + 1,76^2 + 1,38^2 + 3,56^2) - 20 \times 1,96^2 = 18,04$$

$$SC_B = a \sum_{j=1}^b \bar{y}_{.j}^2 - n\bar{y}_{..}^2 = 4(2,3^2 + 2,53^2 + 0,88^2 + 2,2^2 + 1,9^2) - 20 \times 1,96^2 = 6,69$$

$$SCE = SCT - SC_A - SC_B = 25,69 - 18,04 - 6,69 = 0,96$$

La tabla ANOVA es:

F.V.	S.C.	G.L.	M.C.	F
Sustancia	18.04	3	6.01	75.13
Muestra	6.69	4	1.67	
Residual	0.96	12	0.08	
Total	25.69	19		

Como $F_{3,12;0,01} = 5,9526$, existe una diferencia significativa en las sustancias químicas en cuanto al efecto que tienen sobre la resistencia promedio de la tela.

Observación

Si las medias de los tratamientos son diferentes entre sí se pueden considerar los tests de comparaciones múltiples y de rangos estudentizados, que se vieron para el modelo unifactorial general. Se ha de reemplazar el número de réplicas por nivel del factor (n) por el número de bloques (b).

A su vez, los grados de libertad del error han de cambiarse de $(N - a)$ en el caso general a $(a - 1)(b - 1)$.

Cuadrados latinos

En principio, un modelo completo con tres factores y todas las posibles interacciones sería:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijk}$$

para $i = 1, \dots, a$, $j = 1, \dots, b$ y $k = 1, \dots, K$; siendo $\varepsilon_{ijk} \sim N(0, \sigma^2)$ independientes.

El problema reside en que sólo se dispone de $(a \cdot b \cdot K)$ observaciones para estimar $(a \cdot b \cdot K + 1)$ parámetros.

Las posibles soluciones son:

1. Prescindir de alguna interacción, es decir, simplificar el modelo.
2. Replicar el experimento aunque, a veces, no es posible obtener tantas observaciones.
3. Construir un diseño de *cuadrados latinos*.

Podemos aplicar un diseño de cuadrados latinos cuando:

- Hay tres factores
- El número de niveles es el mismo para cada factor, es decir $a = b = k$
- No se esperan interacciones entre los factores

El diseño por cuadrados latinos trata de sacar el máximo de información con el mínimo de observaciones. La idea es que cada nivel del factor aparezca una vez con cada uno de los niveles de las otras variables. Además, han de cumplirse las condiciones de normalidad, homocedasticidad, independencia y falta de interacción.

Diseño y construcción de un cuadrado latino

Se parte de un factor y dos variables bloque, de modo que el número de niveles del factor y el de las variables bloque sea igual.

A cada combinación de los niveles de los bloques se le asigna un solo nivel del factor, de modo que cada uno de estos niveles aparece una y sólo una vez en cada fila y en cada columna. De este modo,

- (i) Se toma un cuadrado latino al azar de una tabla, o se genera por medio de un programa.
- (ii) Se toman tres permutaciones al azar de los números $1, 2, \dots, I$
- (iii) Con la primera permutación se aleatorizan las columnas, con la segunda las filas y con la tercera los tratamientos.

Ejemplo.

Se realiza un experimento para comparar 4 procesos de fabricación. Se supone que influye tanto el momento del día como el día de la semana en que se realiza. Entonces, se consideran las siguientes variables.

1. Cada día se divide en 4 periodos distintos de 6 horas cada uno (de 0–6, de 6–12, de 12–18, 18–24).
2. Los día que se consideran son L, M, X y J.
3. Como factor se toma el proceso de fabricación con cuatro niveles: P_1, P_2, P_3 y P_4 .

Elegimos al azar un cuadrado latino

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>B</i>	<i>C</i>	<i>D</i>	<i>A</i>
<i>C</i>	<i>D</i>	<i>A</i>	<i>B</i>
<i>D</i>	<i>A</i>	<i>B</i>	<i>C</i>

De las $4! = 24$ posibles permutaciones, se eligen las siguientes:

(1, 3, 4, 2)

(3, 2, 1, 4)

(4, 3, 2, 1)

asignando, entonces,

col.1	L	fila 1	12-18	Letra A	P_4
col.2	X	fila 2	6-12	Letra B	P_3
col.3	J	fila 3	0-6	Letra C	P_2
col.4	M	fila 4	18-24	Letra D	P_1

Queda el siguiente cuadrado latino

	L	X	J	M
12-18	P_4	P_3	P_2	P_1
6-12	P_3	P_2	P_1	P_4
0-6	P_2	P_1	P_4	P_3
18-24	P_1	P_4	P_3	P_2

que, reordenando, queda como

	L	M	X	J
0-6	P_2	P_3	P_1	P_4
6-12	P_3	P_4	P_2	P_1
12-18	P_4	P_1	P_3	P_2
18-24	P_1	P_2	P_4	P_3

Las ventajas que presenta el diseño latino es que permite estudiar el efecto de un factor y dos variables bloque con I niveles cada uno de ellos y con I^2 unidades experimentales, en lugar de I^3 unidades experimentales. Aunque se impone siempre la limitación de que todos ellos tengan siempre el mismo número de niveles.

Modelo

Este modelo sería:

$$y_{ij(k)} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ij(k)}$$

para $i = 1, \dots, I$, $j = 1, \dots, I$ y $k = 1, \dots, I$; de modo que k viene fijado con el par (i, j) , donde:

- μ el efecto medio global.
- α_i el efecto incremental sobre la media causado por el nivel i de la variable α (efecto fila).
- β_j el efecto incremental sobre la media causado por el nivel j de la variable β (efecto columna).
- γ_k el efecto incremental sobre la media causado por el nivel k del factor γ (efecto letra).
- $\varepsilon_{ij(k)}$ el término de error, siendo $\varepsilon_{ij(k)} \sim N(0, \sigma^2)$ independientes.

Supondremos además que:

$$\sum_{i=1}^I \alpha_i = \sum_{j=1}^I \beta_j = \sum_{k=1}^I \gamma_k = 0$$

Así, las observaciones serían, por ejemplo para tres niveles:

	I	II	III
1	$y_{11(1)}$ A	$y_{12(2)}$ B	$y_{13(3)}$ C
2	$y_{21(3)}$ C	$y_{22(1)}$ A	$y_{23(2)}$ B
3	$y_{31(2)}$ B	$y_{32(3)}$ C	$y_{33(1)}$ A

El índice i indica el nivel de la variable α (la fila), el índice j indica el nivel de la variable β (la columna) y el índice k indica el nivel del factor γ (la letra), que aparece entre paréntesis porque para cada par (i, j) sólo hay un k posible.

Se denota por D el conjunto de índices que aparecen en el cuadrado latino (i, j, k) .

En el ejemplo de los procesos de fabricación, el cuadrado latino está formado por 16 elementos, así,

$$D = \{(1, 1, 2) (1, 2, 3) (1, 3, 1) (1, 4, 4) (2, 1, 3) (2, 2, 4) (2, 3, 2) (2, 4, 1) (3, 1, 4) (3, 2, 1) (3, 3, 3) (3, 4, 2) (4, 1, 1) (4, 2, 2) (4, 3, 4) (4, 4, 3)\}.$$

Se pueden definir, también, los siguientes conjuntos:

$$\begin{aligned} \text{Para } i \text{ fijo: } D_{(i)} &= \{(j, k), (i, j, k) \in D\} \\ \text{Para } j \text{ fijo: } D_{(j)} &= \{(i, k), (i, j, k) \in D\} \\ \text{Para } k \text{ fijo: } D_{(k)} &= \{(i, j), (i, j, k) \in D\} \end{aligned}$$

En el ejemplo:

$$D_{(i=1)} = \{(1, 2) (2, 3) (3, 1) (4, 4)\}$$

$$D_{(j=3)} = \{(1, 1) (2, 2) (3, 3) (4, 4)\}$$

$$D_{(k=2)} = \{(1, 1) (2, 3) (3, 4) (4, 2)\}$$

El número de elementos de estos conjuntos es 4.

Estimación de los parámetros

Para estimar los parámetros del modelo, hay que minimizar la siguiente función

$$\text{mín } \Phi = \text{mín}_{\mu, \alpha_i, \beta_j, \gamma_k} \sum_{(i, j, k) \in D} (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_k)^2$$

sujeto a

$$\sum_{D_{(i)}} \alpha_i = \sum_{D_{(j)}} \beta_j = \sum_{D_{(k)}} \gamma_k = 0$$

Derivando respecto de los parámetros e igualando a 0 se obtienen los siguientes estimadores:

$$\begin{aligned}\hat{\mu} &= \bar{y}_{...} = \frac{1}{I^2} \sum_{(i,j,k) \in D} y_{ij(k)} \\ \hat{\alpha}_i &= \bar{y}_{i..} - \bar{y}_{...} = \frac{1}{I} \sum_{(j,k) \in D(i)} y_{ij(k)} - \bar{y}_{...} \\ \hat{\beta}_j &= \bar{y}_{.j.} - \bar{y}_{...} = \frac{1}{I} \sum_{(i,k) \in D(j)} y_{ij(k)} - \bar{y}_{...} \\ \hat{\gamma}_k &= \bar{y}_{..k} - \bar{y}_{...} = \frac{1}{I} \sum_{(i,j) \in D(k)} y_{ij(k)} - \bar{y}_{...}\end{aligned}$$

Por otro lado, como el número de parámetros a estimar es

$$(I - 1) + (I - 1) + (I - 1) + 1 = 3I - 2,$$

el número de grados de libertad de la varianza es

$$I^2 - 3I + 2 = (I - 1)(I - 2)$$

con lo cual necesitamos que $I \geq 3$:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{(I - 1)(I - 2)} \sum_{(i,j,k) \in D} (y_{ij(k)} - \hat{y}_{ij(k)})^2 = \\ &= \frac{1}{(I - 1)(I - 2)} \sum_{(i,j,k) \in D} (y_{ij(k)} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2\end{aligned}$$

Análisis de la varianza

Se utilizará el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \gamma_1 = \dots = \gamma_I = 0 & (\text{el factor } \gamma \text{ no influye}) \\ H_1 : \text{algún } \gamma_i \neq 0 & (\text{el factor } \gamma \text{ influye}) \end{cases}$$

Para contrastar esta hipótesis, descomponemos la suma de cuadrados total en la siguiente suma de cuadrados:

$$\begin{aligned}SCT &= \sum_{(i,j,k) \in D} (y_{ij(k)} - \bar{y}_{...})^2 = \sum_{(i,j,k) \in D} y_{ij(k)}^2 - I^2 \bar{y}_{...}^2 = \\ &= SC_\alpha + SC_\beta + SC_\gamma + SCE\end{aligned}$$

donde

$$SC_\alpha = I \sum_{(j,k) \in D(i)} (\bar{y}_{i..} - \bar{y}_{...})^2 = I \sum_{(j,k) \in D(i)} \bar{y}_{i..}^2 - I^2 \bar{y}_{...}^2 \equiv$$

“Suma de cuadrados explicada debido a la variable α ”

$$SC_\beta = I \sum_{(i,k) \in D(j)} (\bar{y}_{.j.} - \bar{y}_{...})^2 = I \sum_{(i,k) \in D(j)} \bar{y}_{.j.}^2 - I^2 \bar{y}_{...}^2 \equiv$$

“Suma de cuadrados explicada debido a la variable β ”

$$SC_\gamma = I \sum_{(i,j) \in D(k)} (\bar{y}_{..k} - \bar{y}_{...})^2 = I \sum_{(i,j) \in D(k)} \bar{y}_{..k}^2 - I^2 \bar{y}_{...}^2 \equiv$$

“Suma de cuadrados explicada debido al factor γ ”

$$SCE = \sum_{(i,j,k) \in D} (y_{ij(k)} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2 = (I-1)(I-2)\hat{\sigma}^2 \equiv$$

“Suma de cuadrados residual”

La tabla de análisis de la varianza es:

F. V.	S. C.	G. L.	M. C.	F
Factor α	SC_α	$I - 1$	$MC_\alpha = \frac{SC_\alpha}{I-1}$	
Factor β	SC_β	$I - 1$	$MC_\beta = \frac{SC_\beta}{I-1}$	
Factor γ	SC_γ	$I - 1$	$MC_\gamma = \frac{SC_\gamma}{I-1}$	$F_\gamma = \frac{MC_\gamma}{MCE}$
Residual	SCE	$(I-1)(I-2)$	$MCE = \frac{SCE}{(I-1)(I-2)}$	
Total	SCT	$I^2 - 1$		

Entonces, se rechaza $H_0 : \gamma_1 = \dots = \gamma_I = 0$ al nivel α cuando

$$F_\gamma > F_{I-1, (I-1)(I-2); \alpha}$$

Ejemplo. Supongamos que un experimentador está estudiando el efecto de cinco fórmulas diferentes de la mezcla de dinamita sobre la fuerza explosiva observada. Cada fórmula se

prepara usando un lote de materia prima, lo suficientemente grande para que sólo se hagan cinco mezclas. Mas aún, las mezclas las preparan cinco operarios, pudiendo existir una diferencia sustancial en la habilidad y experiencia entre ellos. El diseño apropiado para este problema consiste en probar cada fórmula exactamente una vez, utilizando cada lote de materia prima, y en que cada fórmula sea preparada exactamente una vez por cada uno de los cinco operarios.

El diseño resultante es un cuadrado latino. Las cinco fórmulas se representan mediante las letras latinas A, B, C, D y E. Los datos aparecen a continuación:

Materia	1	2	3	4	5
1	A 24	B 20	C 19	D 24	E 24
2	B 17	C 24	D 30	E 27	A 36
3	C 18	D 38	E 26	A 27	B 21
4	D 26	E 31	A 26	B 23	C 22
5	E 22	A 30	B 20	C 29	D 31

Para simplificar los datos se resta 25 unidades y se obtiene:

Materia	1	2	3	4	5	$\bar{y}_{i.}$
1	A -1	B -5	C -6	D -1	E -1	-2.8
2	B -8	C -1	D 5	E 2	A 11	1.8
3	C -7	D 13	E 1	A 2	B -4	1
4	D 1	E 6	A 1	B -2	C -3	0.6
5	E -3	A 5	B -5	C 4	D 6	1.4
$\bar{y}_{.j.}$	-3.6	3.6	-0.8	1	1.8	0.4

Las medias sobre el factor γ son:

letra latina	$\bar{y}_{..k}$
A	$\bar{y}_{..1} = (-1 + 11 + 2 + 1 + 5)/5 = 18/5 = 3,6$
B	$\bar{y}_{..2} = (-5 - 8 - 4 - 2 - 5)/5 = -24/5 = -4,8$
C	$\bar{y}_{..3} = (-6 - 1 - 7 - 3 + 4)/5 = -13/5 = -2,6$
D	$\bar{y}_{..4} = (-1 + 5 + 13 + 1 + 6)/5 = 24/5 = 4,8$
E	$\bar{y}_{..5} = (-1 + 2 + 1 + 6 - 3)/5 = 5/5 = 1$

Las sumas de cuadrados son:

$$SCT = \sum_{i=1}^I \sum_{j=1}^I y_{ij}^2 - n\bar{y}_{..}^2 = 680 - 25 \times 0,4^2 = 676$$

$$SC_{\alpha} = I \sum_{i=1}^I \bar{y}_{i..}^2 - n\bar{y}_{..}^2 = 5((-2,8)^2 + 1,8^2 + 1^2 + 0,6^2 + 1,4^2) - 25 \times 0,4^2 = 68$$

$$SC_{\beta} = I \sum_{j=1}^J \bar{y}_{.j}^2 - n\bar{y}_{..}^2 = 5((-3,6)^2 + 3,6^2 + (-0,8)^2 + 1^2 + 1,8^2) - 25 \times 0,4^2 = 150$$

$$SC_{\gamma} = I \sum_{k=1}^I \bar{y}_{..k}^2 - n\bar{y}_{..}^2 = 5(3,6^2 + (-4,8)^2 + (-2,6)^2 + 4,8^2 + 1^2) - 25 \times 0,4^2 = 330$$

$$SCE = SCT - SC_{\alpha} - SC_{\beta} - SC_{\gamma} = 676 - 68 - 150 - 330 = 128$$

La tabla ANOVA es:

F.V.	S.C.	G.L.	M.C.	F
Material (A)	68	4	17	
Operario (B)	150	4	37.5	
Fórmula (C)	330	4	82.5	7.73
Error	128	12	10.67	
Total	676	24		

Como

$$F_{4,12;0,1} = 2,48,$$

existen diferencias significativas en la fuerza explosiva media debido a las cinco fórmulas.

Cuadrados greco–latinos

Se puede considerar una extensión del diseño de cuadrados latinos que permite estudiar un factor y 3 variables bloque con sólo I^2 observaciones (siempre que el factor y las variables bloque tengan todos I niveles).

Se considera un cuadrado latino de dimensión $(I \times I)$ y se superpone sobre él otro cuadrado con los tratamientos denotados por letras griegas.

Se dice que son ortogonales cuando cada letra griega aparece combinada con una letra latina una y sólo una vez en cada fila y columna.

Por ejemplo,

– *Cuadrado latino 1:*

A	B	C	D
D	C	B	A
B	A	D	C
C	D	A	B

– *Cuadrado latino 2:*

α	β	γ	δ
γ	δ	α	β
δ	γ	β	α
β	α	δ	γ

con lo que el cuadrado greco–latino queda como

$A\alpha$	$B\beta$	$C\gamma$	$D\delta$
$D\gamma$	$C\delta$	$B\alpha$	$A\beta$
$B\delta$	$A\gamma$	$D\beta$	$C\alpha$
$C\beta$	$D\alpha$	$A\delta$	$B\gamma$

Se pueden construir cuadrados greco–latinos en diseños completamente cruzados para $I > 3$, salvo para $I = 6$, porque en este caso sólo existe un cuadrado latino.

El modelo es

$$y_{ij(kh)} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_h + \varepsilon_{ij(kh)}$$

donde

$$\sum_{i=1}^I \alpha_i = \sum_{j=1}^I \beta_j = \sum_{k=1}^I \gamma_k = \sum_{h=1}^I \delta_h = 0.$$

Se obtiene una tabla ANOVA semejante al modelo de cuadrados latinos, considerando también el término δ_h . Los grados de libertad de la SCE son igual a $(I - 1)(I - 3)$. El contraste de la F es el mismo que en el caso de los cuadrados latinos.

Aplicación con R

```
# Diseño por bloques
# Cargamos datos de la libreria de Faraday
library(faraway)
data(penicillin); attach(penicillin)

penicillin
# Se muestra el numero de replicas
replications(yield ~ blend + treat,penicillin)

par(pty="s")
par(mfrow=c(2,2))
plot.design(yield ~ blend + treat)
plot.design(yield ~ blend + treat, fun=median)
plot(yield ~ blend + treat)

interaction.plot(treat,blend,yield,type="l",col=1:3)

modelo <- aov(yield ~ blend + treat)
summary(modelo)
par(pty="s")
par(mfrow=c(2,2))
plot(modelo)

# medias de tratamientos
model.tables(modelo,"mean")

#.....

# Cuadrado latino
# Cargamos datos de la libreria de Faraday
library(faraway)
data(breaking); attach(breaking)

breaking
# Se muestra el numero de replicas
replications(y ~ operator + day + supplier, breaking)

matrix(supplier,4,4)

par(pty="s")
par(mfrow=c(2,2))
plot.design(y ~ operator + day + supplier, breaking)
plot(y ~ operator + day + supplier, breaking)

replications(y ~ operator + day + supplier)

modelo <- aov(y ~ operator + day + supplier, breaking)
summary(modelo)

# medias de tratamientos
model.tables(modelo,"mean")

# Estimacion de los efectos
modelo$coef
# o bien
coef <- lm(y ~ operator + day + supplier, breaking)
summary(coef)
```

Aplicación con SAS

```
options ls=70 nodate nonumber;
title 'ANOVA bloques aleatorizados completos';
data ano;
input bloque trata $ medida;
cards;
1 1 21
1 2 26
1 3 16
1 4 28
2 1 36
2 2 38
2 3 25
2 4 35
3 1 25
3 2 27
3 3 22
3 4 27
4 1 18
4 2 17
4 3 18
4 4 20
5 1 22
5 2 26
5 3 21
5 4 24
proc anova;
class bloque trata;
model medida=bloque trata;
means trata /duncan tukey;
means trata;
run;
```

ANOVA bloques aleatorizados completos

The ANOVA Procedure

Class Level Information

Class	Levels	Values
bloque	5	1 2 3 4 5
trata	4	1 2 3 4

Number of observations 20

Dependent Variable: medida

Source	DF	Sum of Squares	Mean Square	F Value
Model	7	637.6000000	91.0857143	12.53
Error	12	87.2000000	7.2666667	
Corrected Total	19	724.8000000		

Source	Pr > F
Model	0.0001
Error	

Corrected Total

R-Square	Coeff Var	Root MSE	medida Mean
0.879691	10.95803	2.695676	24.60000

Source	DF	Anova SS	Mean Square	F Value
bloque	4	500.8000000	125.2000000	17.23
trata	3	136.8000000	45.6000000	6.28

Source	Pr > F
bloque	<.0001
trata	0.0083

Duncan's Multiple Range Test for medida

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	12
Error Mean Square	7.266667

Number of Means	2	3	4
Critical Range	3.715	3.888	3.993

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	trata
A	26.800	5	2
A			
A	26.800	5	4
A			
A	24.400	5	1
B	20.400	5	3

Tukey's Studentized Range (HSD) Test for medida

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	12
Error Mean Square	7.266667
Critical Value of Studentized Range	4.19852
Minimum Significant Difference	5.0615

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	trata
A	26.800	5	2
A			
A	26.800	5	4
A			
B A	24.400	5	1
B			
B	20.400	5	3

Level of trata	N	-----medida-----	
		Mean	Std Dev
1	5	24.400000	6.94982014
2	5	26.800000	7.46324326
3	5	20.400000	3.50713558
4	5	26.800000	5.54075807

```

options ls=70 nodate nonumber;
title 'ANOVA cuadrados latinos';
data ano;
input bloque1 bloque2 trata $ medida;
cards;
1 1 B 810
1 2 C 1100
1 3 D 840
1 4 A 650
2 1 C 1080
2 2 D 880
2 3 A 540
2 4 B 740
3 1 A 700
3 2 B 780
3 3 C 1055
3 4 D 1025
4 1 D 910
4 2 A 600
4 3 B 830
4 4 C 900
proc anova;
class bloque1 bloque2 trata;
model medida=bloque1 bloque2 trata;
means trata /duncan tukey;
means trata;
run;

```

ANOVA cuadrados latinos

The ANOVA Procedure

Class Level Information

Class	Levels	Values
bloque1	4	1 2 3 4
bloque2	4	1 2 3 4
trata	4	A B C D

Number of observations 16

Dependent Variable: medida

Source	DF	Sum of Squares	Mean Square	F Value
Model	9	396400.0000	44044.4444	7.09
Error	6	37250.0000	6208.3333	
Corrected Total	15	433650.0000		

Source	Pr > F
Model	0.0135
Error	

Corrected Total

R-Square	Coeff Var	Root MSE	medida Mean
0.914101	9.380116	78.79298	840.0000

Source	DF	Anova SS	Mean Square	F Value
bloque1	3	17600.0000	5866.6667	0.94
bloque2	3	7662.5000	2554.1667	0.41
trata	3	371137.5000	123712.5000	19.93

Source	Pr > F
bloque1	0.4759
bloque2	0.7510
trata	0.0016

Duncan's Multiple Range Test for medida

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	6
Error Mean Square	6208.333

Number of Means	2	3	4
Critical Range	136.3	141.3	143.8

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	trata
A	1033.75	4	C
A			
B A	913.75	4	D
B			
B	790.00	4	B
C	622.50	4	A

Tukey's Studentized Range (HSD) Test for medida

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	6
Error Mean Square	6208.333
Critical Value of Studentized Range	4.89559
Minimum Significant Difference	192.87

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	trata
A	1033.75	4	C
A			
B A	913.75	4	D
B			
B C	790.00	4	B
C			
C	622.50	4	A

Level of trata	N	-----medida----- Mean	Std Dev
A	4	622.50000	68.4957420
B	4	790.00000	39.1578004
C	4	1033.75000	91.0471490
D	4	913.75000	79.5167697