



por mínimos cuadrados mediante el procedimiento general del modelo lineal, donde ahora

$$X'X = \begin{pmatrix} n_1 & & & & \\ & n_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & n_I \end{pmatrix} \quad X'Y = \begin{pmatrix} \sum_{j=1}^{n_1} Y'_{1j} \\ \sum_{j=1}^{n_2} Y'_{2j} \\ \vdots \\ \sum_{j=1}^{n_I} Y'_{Ij} \end{pmatrix}$$

Y entonces

$$(X'X)^{-1} X'Y = \begin{pmatrix} \bar{Y}'_{1\bullet} \\ \bar{Y}'_{2\bullet} \\ \vdots \\ \bar{Y}'_{I\bullet} \end{pmatrix}$$

por lo que las medias muestrales resultan ser los estimadores bajo este modelo.

Ahora planteamos el contraste de la igualdad de todas las medias, esto es, el contraste de la hipótesis nula:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$

Esta hipótesis se puede formular de modo equivalente así

$$\mu_1 - \mu_I = \mu_2 - \mu_I = \cdots = \mu_{I-1} - \mu_I = 0$$

lo cual admite esta forma matricial

$$AB = \begin{pmatrix} 1 & & & & -1 \\ & 1 & & & -1 \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & & 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} \mu'_1 \\ \mu'_2 \\ \vdots \\ \mu'_I \end{pmatrix} = 0 \quad (4.1)$$

Bajo esta restricción lineal, que reduce los parámetros a una única media común  $\mu$ , podemos considerar el nuevo modelo con ese único parámetro que se estimaría mediante la media global:

$$\hat{\mu} = \frac{1}{\sum_{i=1}^I n_i} \sum_{i=1}^I \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^I n_i \bar{Y}_{i\bullet} = \bar{Y}_{\bullet\bullet}$$

El contraste de la hipótesis de igualdad se puede llevar a cabo por el método basado en la  $\Lambda$  de Wilks, que compara la matriz de covarianzas de los residuos bajo el modelo general y bajo la hipótesis nula. Así, si la hipótesis nula es cierta, el estadístico de contraste verifica

$$\frac{|E|}{|E + H|} \in \Lambda(d, q, n - p)$$

siendo  $E$  la matriz de covarianzas de los residuos bajo el modelo general y  $E_H = E + H$  la matriz de covarianzas de los residuos bajo la hipótesis nula,  $p$  el número de parámetros independientes bajo el modelo general y  $p - q$  el número de parámetros independientes bajo la hipótesis nula.

Esto da lugar a una descomposición de la matriz de covarianzas, que en este caso se suele representar mediante la llamada tabla del análisis multivariante de la varianza o tabla MANOVA:

Fuente de variación	Matriz de covarianzas	Grados de libertad
Entre poblaciones	$H = \sum_{i=1}^I \sum_{j=1}^{n_i} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})'$	$I - 1$
Error	$E = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet}) (Y_{ij} - \bar{Y}_{i\bullet})'$	$\sum_{i=1}^I (n_i - 1)$
Total	$E_H = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet}) (Y_{ij} - \bar{Y}_{\bullet\bullet})'$	$\sum_{i=1}^I n_i - 1$

Observamos que esta tabla no es más que una extensión de la tabla ANOVA al caso multivariante.

Así, como

$$H \in \text{Wishart}_d(\Sigma, I - 1)$$

$$E \in \text{Wishart}_d(\Sigma, n - I)$$

y además son independientes, tenemos que

$$\frac{|E|}{|E + H|} \in \Lambda(d, I - 1, n - I)$$

El sentido común nos invita a rechazar la hipótesis nula cuando la variabilidad proveniente de las diferencias entre poblaciones (que medimos mediante la matriz  $H$ ) sea grande comparada con la proveniente del error (medida por  $E$ ).

Por tanto, rechazaremos la hipótesis nula cuando el estadístico  $|E|/|E + H|$  tome un valor menor que el cuantil  $\alpha$  de la distribución  $\Lambda(d, I - 1, n - I)$ , siendo  $\alpha$  el nivel de significación fijado de antemano.

Al igual que en el modelo lineal general multivariante, aquí también podemos plantear el procedimiento de unión–intersección para el contraste de la hipótesis nula de igualdad de todas las medias. En ese caso, el estadístico de contraste sería

$$\phi_{\max} = \text{máximo autovalor de } HE^{-1}$$

y rechazaremos la hipótesis nula cuando  $\phi_{\max} > \phi_{\max, \alpha}$  siendo  $\phi_{\max, \alpha}$  el cuantil  $1 - \alpha$  de la distribución de  $\phi_{\max}$ .

**Ejemplo 4.1** *Sobre los datos de los lirios de Fisher, vamos a efectuar el contraste de igualdad del vector de medias para las tres especies. Lo haremos suponiendo que cada individuo tiene distribución normal multivariante, las muestras son independientes y tienen la misma matriz de covarianzas.*

### 4.3. Comparaciones múltiples.

Para un modelo lineal general tenemos

$$P \left\{ a'ABb \in \left( a'A\hat{B}b - \sqrt{\phi_{\max,\alpha} a'A(X'X)^{-1}A'ab'Eb}, \right. \right. \\ \left. \left. a'A\hat{B}b + \sqrt{\phi_{\max,\alpha} a'A(X'X)^{-1}A'ab'Eb} \right) \forall a, b \right\} = 1 - \alpha$$

obteniendo así un conjunto de intervalos de confianza simultáneos para  $a'ABb$ , con nivel de confianza  $1 - \alpha$ .

En nuestro caso, las matrices  $A$  y  $B$  se encuentran en la expresión (4.1) y entonces

$$a'AB = (a_1, a_2, \dots, a_{I-1}) \cdot \begin{pmatrix} (\mu_1 - \mu_I)' \\ (\mu_2 - \mu_I)' \\ \vdots \\ \vdots \\ (\mu_{I-1} - \mu_I)' \end{pmatrix} \\ = a_1 (\mu_1 - \mu_I)' + a_2 (\mu_2 - \mu_I)' + \dots + a_{I-1} (\mu_{I-1} - \mu_I)' = \sum_{i=1}^I c_i \mu_i'$$

siendo  $c_i = a_i \ \forall i \in \{1, \dots, I-1\}$  y  $c_I = -\sum_{i=1}^{I-1} a_i$ . De este modo, pasamos de considerar cualquier vector  $a$  a considerar cualquier vector  $(c_1, \dots, c_I)$  que verifique  $\sum_{i=1}^I c_i = 0$ .

De igual modo,

$$a'A\hat{B} = \sum_{i=1}^I c_i \bar{Y}'_{i\bullet}$$

A continuación observamos que

$$a'A = \left( a_1, \dots, a_{I-1}, -\sum_{i=1}^{I-1} a_i \right) = (c_1, \dots, c_I)$$

de modo que

$$a'A(X'X)^{-1}A'a = (c_1, \dots, c_I) \cdot \begin{pmatrix} 1/n_1 & & \\ & \ddots & \\ & & 1/n_I \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ \vdots \\ c_I \end{pmatrix} = \sum_{i=1}^I \frac{c_i^2}{n_i}$$

Luego, nos queda el siguiente conjunto de intervalos de confianza simultáneos

$$\left( \sum_{i=1}^I c_i \bar{Y}'_{i\bullet} b - \sqrt{\phi_{\max,\alpha} \left( \sum_{i=1}^I \frac{c_i^2}{n_i} \right) b'Eb}, \sum_{i=1}^I c_i \bar{Y}'_{i\bullet} b + \sqrt{\phi_{\max,\alpha} \left( \sum_{i=1}^I \frac{c_i^2}{n_i} \right) b'Eb} \right)$$

para  $\sum_{i=1}^I c_i \mu'_i b$  con nivel de confianza  $1 - \alpha$ . El vector  $(c_1, \dots, c_I)$  indica qué poblaciones vamos a comparar, mientras que el vector  $b$  indica qué componentes del vector respuesta  $Y$  vamos a utilizar para la comparación. De nuevo,  $\phi_{\max, \alpha}$  es el cuantil  $1 - \alpha$  de la distribución de  $\phi_{\max}$ .

Así, si tomamos  $c_r = 1$ ,  $c_s = -1$  y  $c_i = 0$  si  $i \neq r$ ,  $i \neq s$ , quedando el vector  $(0, \dots, 0, 1, 0, \dots, 0, -1, 0, \dots, 0)$ , y  $b_j = 1$ ,  $b_k = 0$  si  $k \neq j$ , para obtener  $b = (0, \dots, 0, 1, 0, \dots, 0)$ , entonces

$$\left( (\bar{Y}_{r\bullet} - \bar{Y}_{s\bullet})' b \pm \sqrt{\phi_{\max, \alpha} \left( \frac{1}{n_r} + \frac{1}{n_s} \right) b' E b} \right) \quad r, s \in \{1, \dots, I\} \quad j \in \{1, \dots, d\}$$

es un conjunto de intervalos de confianza simultáneos para la comparación de las poblaciones  $r$ -ésima y  $s$ -ésima en cada una de las componentes del vector  $Y$ .

Nótese que en este caso  $b' E b$  es la suma de cuadrados de los residuos relativos a la componente  $j$ -ésima. En general,  $b' E b$  es la evaluación de la forma cuadrática  $E$  en el vector  $b$ , y contiene la suma de cuadrados de las combinaciones lineales de los residuos según el vector  $b$ .

**Ejemplo 4.2** *Sobre el ejemplo de los lirios, vamos a efectuar las comparaciones múltiples de los vectores de medias de cada especie.*

#### 4.4. MANOVA con dos factores de variación.

Hasta aquí hemos considerado la comparación de poblaciones clasificadas según un único criterio. En esta sección suponemos que hay dos factores: A y B. Del factor A podemos distinguir I niveles, mientras que en el factor B podemos encontrar J niveles. En cada una de las  $I \cdot J$  posibilidades realizamos K observaciones de un vector aleatorio  $Y$ . El objetivo será estudiar la influencia de los factores A y B, o de su interacción, en la media del vector  $Y$ .

Así, planteamos el siguiente modelo:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + U_{ijk} \quad k \in \{1, \dots, K\} \quad i \in \{1, \dots, I\} \quad j \in \{1, \dots, J\}$$

siendo  $U_{ijk} \in N_d(0, \Sigma)$ . El parámetro  $\mu$  representa la media global, los parámetros  $\alpha_i$  representan el efecto principal del factor A, los parámetros  $\beta_j$  representan el efecto principal del factor B y los parámetros  $\gamma_{ij}$  representan la interacción de los factores A y B. Además verifican

$$\sum_{i=1}^I \alpha_i = \sum_{j=1}^J \beta_j = \sum_{i=1}^I \gamma_{ij} = \sum_{j=1}^J \gamma_{ij} = 0$$

Este modelo también se puede ver como un caso particular del modelo lineal general multivariante. Así, aplicaremos los resultados conocidos para el modelo lineal general, tanto en lo relativo a la estimación por mínimos cuadrados de los parámetros como en lo concerniente a contrastes de hipótesis referidas al modelo.

En este sentido planteamos hipótesis del tipo:

$$\begin{aligned} H_A & : \alpha_i = 0 \quad \forall i \\ H_B & : \beta_j = 0 \quad \forall j \\ H_{AB} & : \gamma_{ij} = 0 \quad \forall i, j \end{aligned}$$

que se estudian en base a la siguiente descomposición de la variabilidad, mediante lo que llamaremos tabla MANOVA II:

Fuente de variación	Matriz de covarianzas	Grados de libertad
Factor A	$H_A = JK \sum_{i=1}^I (\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet}) (\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})'$	$I - 1$
Factor B	$H_B = IK \sum_{j=1}^J (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet}) (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet})'$	$J - 1$
Interacción	$H_{AB} = K \sum_{i=1}^I \sum_{j=1}^J (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet}) \times (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet})'$	$(I - 1)(J - 1)$
Error	$E = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (Y_{ijk} - \bar{Y}_{ij\bullet}) (Y_{ijk} - \bar{Y}_{ij\bullet})'$	$IJ(K - 1)$
Total	$\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (Y_{ijk} - \bar{Y}_{\bullet\bullet\bullet}) (Y_{ijk} - \bar{Y}_{\bullet\bullet\bullet})'$	$IJK - 1$

De este modo la hipótesis  $H_A : \alpha_i = 0 \forall i$  se contrasta en base al estadístico

$$\frac{|E|}{|E + H_A|} \in \Lambda(d, I - 1, IJ(K - 1))$$

la hipótesis  $H_B : \beta_j = 0 \forall j$  se contrasta en base al estadístico

$$\frac{|E|}{|E + H_B|} \in \Lambda(d, J - 1, IJ(K - 1))$$

y la hipótesis  $H_{AB} : \gamma_{ij} = 0 \forall i, j$  se contrasta en base al estadístico

$$\frac{|E|}{|E + H_{AB}|} \in \Lambda(d, (I - 1)(J - 1), IJ(K - 1))$$

**Ejemplo 4.3** *En una especie de césped, denominada Paspalum, se está investigando el efecto que experimenta al ser infectada con un hongo. Al mismo tiempo se tiene en cuenta la temperatura, dentro de un diseño con cuatro valores diferentes: 14, 18, 22, 26°C. En cada realización del experimento se miden tres variables:*

- $Y_1 =$  El peso fresco de las raíces (medido en gramos)
- $Y_2 =$  La longitud máxima de las raíces (medida en milímetros)
- $Y_3 =$  El peso fresco de las hojas (medido en gramos)

Los datos se encuentran en el fichero "tema4ejemplo3.txt". Vamos a contrastar el efecto del tratamiento con hongos, el efecto de la temperatura y la posible interacción entre ambos efectos.

### Comparaciones múltiples

Con la misma forma de proceder que en el MANOVA con un factor de variación, vamos a obtener intervalos de confianza simultáneos para combinaciones lineales de los parámetros de este modelo. Lo haremos con los efectos principales del factor A.

La hipótesis  $H_A$  se puede expresar en cualquiera de estas formas equivalentes:

$$\begin{aligned} \alpha_1 = \alpha_2 = \dots = \alpha_I = 0 &\iff \bar{\mu}_{1\bullet} = \bar{\mu}_{2\bullet} = \dots = \bar{\mu}_{I\bullet} \\ &\iff \bar{\mu}_{1\bullet} - \bar{\mu}_{I\bullet} = \bar{\mu}_{2\bullet} - \bar{\mu}_{I\bullet} = \dots = \bar{\mu}_{(I-1)\bullet} - \bar{\mu}_{I\bullet} = 0 \\ &\iff AB = \begin{pmatrix} \bar{\mu}'_{1\bullet} - \bar{\mu}'_{I\bullet} \\ \bar{\mu}'_{2\bullet} - \bar{\mu}'_{I\bullet} \\ \vdots \\ \bar{\mu}'_{(I-1)\bullet} - \bar{\mu}'_{I\bullet} \end{pmatrix} = 0 \end{aligned}$$

Razonando igual que en el MANOVA I, llegamos a

$$a'AB = \sum_{i=1}^I c_i \bar{\mu}'_{i\bullet}$$

con  $\sum_{i=1}^I c_i = 0$ . De igual modo,  $a'\hat{A}\hat{B} = \sum_{i=1}^I c_i \bar{Y}'_{i\bullet\bullet}$ . Finalmente, obtenemos el siguiente conjunto de intervalos de confianza simultáneos

$$\left( \sum_{i=1}^I c_i \bar{Y}'_{i\bullet\bullet} b - \sqrt{\phi_{\max, \alpha} \left( \sum_{i=1}^I \frac{c_i^2}{JK} \right) b'Eb}, \sum_{i=1}^I c_i \bar{Y}'_{i\bullet\bullet} b + \sqrt{\phi_{\max, \alpha} \left( \sum_{i=1}^I \frac{c_i^2}{JK} \right) b'Eb} \right)$$

para  $\sum_{i=1}^I c_i \bar{\mu}'_{i\bullet} b$  con nivel de confianza  $1 - \alpha$ , siendo  $\phi_{\max, \alpha}$  el cuantil  $1 - \alpha$  de la distribución del autovalor más grande de  $H_A E^{-1}$ .

### 4.5. Diseño por bloques aleatorizados.

Queremos estudiar el efecto de ciertos tratamientos sobre un vector aleatorio, pero en la experimentación debemos tener en cuenta la presencia de otro efecto debido a una variable de tipo bloque. Para ello, elaboramos un diseño experimental en el que cada tipo de tratamiento se observará en cada nivel de la variable bloque. No consideramos replicación. Como resultado obtenemos los vectores aleatorios:  $Y_{ij}$  que representa el vector aleatorio observado bajo el tratamiento  $i$ -ésimo y en el bloque  $j$ -ésimo. Adoptamos el modelo siguiente:

$$Y_{ij} = \mu + \alpha_i + \beta_j + U_{ij} \quad i \in \{1, \dots, I\} \quad j \in \{1, \dots, J\}$$

siendo  $U_{ij} \in N_d(0, \Sigma)$ . El parámetro  $\mu$  representa la media global, los parámetros  $\alpha_i$  representan el efecto del tratamiento y los parámetros  $\beta_j$  representan el efecto bloque. Observamos la gran semejanza con el modelo MANOVA II anterior, del que se diferencia en la ausencia de replicación, lo cual impide la estimación de interacciones. Además, en el presente modelo distinguimos entre los tratamientos, que constituyen el objetivo primordial del estudio, y la variable bloque, que se considera únicamente para controlar su efecto sobre la variable respuesta.

Se verifica

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

Este modelo también se puede ver como un caso particular del modelo lineal general multivariante. Así, aplicaremos los resultados conocidos para el modelo lineal general, tanto en lo relativo a la estimación por mínimos cuadrados de los parámetros como en lo concerniente a contrastes de hipótesis referidas al modelo.

En este sentido planteamos hipótesis del tipo:

$$\begin{aligned} H_T &: \alpha_i = 0 \quad \forall i \\ H_B &: \beta_j = 0 \quad \forall j \end{aligned}$$

que se estudian en base a la siguiente descomposición de la variabilidad:

Fuente de variación	Matriz de covarianzas	Grados de libertad
Tratamientos	$H_T = J \sum_{i=1}^I (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})'$	$I - 1$
Efecto Bloque	$H_B = I \sum_{j=1}^J (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}) (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})'$	$J - 1$
Error	$E = \sum_{i=1}^I \sum_{j=1}^J (Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet}) \times (Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet})'$	$(I - 1)(J - 1)$
Total	$\sum_{i=1}^I \sum_{j=1}^J (Y_{ij} - \bar{Y}_{\bullet\bullet}) (Y_{ij} - \bar{Y}_{\bullet\bullet})'$	$IJ - 1$

De este modo la hipótesis  $H_T : \alpha_i = 0 \forall i$  se contrasta en base al estadístico

$$\frac{|E|}{|E + H_T|} \in \Lambda(d, I - 1, (I - 1)(J - 1))$$

y la hipótesis  $H_B : \beta_j = 0 \forall j$  se contrasta en base al estadístico

$$\frac{|E|}{|E + H_B|} \in \Lambda(d, J - 1, (I - 1)(J - 1))$$

**Ejemplo 4.4** *En las islas Cook se realizó un experimento en bloques aleatorizados para estudiar el efecto de seis tratamientos para combatir un parásito de las plantas de las alubias. Se midieron tres variables*

$Y_1 =$  El número de parásitos por hoja)

$Y_2 =$  El peso de las alubias por planta (medido en kilogramos)

$Y_3 = \text{sen}^{-1}(\sqrt{p})$ , donde  $p$  es la proporción de hojas infestadas con el parásito

Los datos se encuentran en el fichero "tema4ejemplo4.txt". Vamos a contrastar si hay diferencias entre los tratamientos.

### Comparaciones múltiples

También podemos considerar intervalos de confianza simultáneos para combinaciones lineales de los parámetros de este modelo. Así, podemos obtener el siguiente conjunto de intervalos de confianza simultáneos

$$\left( \sum_{i=1}^I c_i \bar{Y}'_{i\bullet} b - \sqrt{\phi_{\max, \alpha} \left( \sum_{i=1}^I \frac{c_i^2}{J} \right) b' E b}, \sum_{i=1}^I c_i \bar{Y}'_{i\bullet} b + \sqrt{\phi_{\max, \alpha} \left( \sum_{i=1}^I \frac{c_i^2}{J} \right) b' E b} \right)$$

para  $\sum_{i=1}^I c_i \bar{\mu}'_{i\bullet} b$  con nivel de confianza  $1 - \alpha$ , siendo  $\phi_{\max, \alpha}$  el cuantil  $1 - \alpha$  de la distribución del autovalor más grande de  $H_T E^{-1}$ .

### Bibliografía.

- Johnson, R.A. y Wichern, D.W. (1982). *Applied multivariate statistical analysis*. Prentice-Hall.
- Mardia, K.V., Kent, J.T. y Bibby, J.M. (1979). *Multivariate analysis*. Academic Press.
- Seber, G.A.F. (1984). *Multivariate observations*. Wiley.