

### Distribución Normal Multivariante

Sea  $V$  una matriz  $n \times n$  real, simétrica y definido positiva; y sea  $\mu = (\mu_1, \dots, \mu_n)^t \in \mathbb{R}^n$

DEFINICIÓN. Diremos que el vector aleatorio (v.a.)  $n$ -dimensional  $Y = (Y_1, \dots, Y_n)^t$  sigue distribución **Normal Multivariante** de parámetros  $\mu$  y  $V$  si tiene la función de densidad conjunta:

$$f(y_1, \dots, y_n) = \frac{1}{(2\pi)^{n/2} |V|^{1/2}} \exp\left\{-\frac{1}{2}(y - \mu)^t V^{-1}(y - \mu)\right\} \quad y = (y_1, \dots, y_n)^t \in \mathbb{R}^n.$$

En tal caso lo denotaremos  $Y \sim N_n(\mu, V)$ .

OBSERVACIONES:

- La función característica de  $Y$  es  $\varphi_Y(t) = \exp\{it^t \mu - \frac{1}{2}t^t V t\}$ ,  $t \in \mathbb{R}^n$
- A partir de la función característica se puede definir la distribución Normal Multivariante para el caso en que  $V$  sea una matriz singular.

TEOREMA. Sea  $Y \sim N_n(\mu, V)$  y  $X = AY + \beta$ , donde  $A$  es una matriz  $k \times n$  de rango  $k$  ( $k \leq n$ ) y  $\beta$  un vector  $k \times 1$ . Entonces  $X = (X_1, \dots, X_k)^t \sim N_k(A\mu + \beta, AVA^t)$ .

COROLARIO. Las distribuciones marginales de una distribución Normal Multivariante son Normales.

COROLARIO. Las componentes de un vector aleatorio normalmente distribuido son independientes si y sólo si son incorreladas dos a dos.

COROLARIO. Sea  $Y \sim N_n(\mu, \sigma^2 I_n)$  y  $A$  una matriz ortogonal  $n \times n$ . Entonces,  $X = AY \sim N_n(A\mu, \sigma^2 I_n)$ .

TEOREMA. La distribución Normal Multivariante es reproductiva respecto del vector de medias y la matriz de covarianzas.

TEOREMA.  $X = (X_1, \dots, X_n)^t \sim N_n(\mu, V)$  si y sólo si toda combinación lineal de  $X_1, \dots, X_n$  sigue distribución Normal (i.e.  $\forall \lambda \in \mathbb{R}^n \lambda^t X \sim N(\lambda^t \mu, \lambda^t V \lambda)$ ).

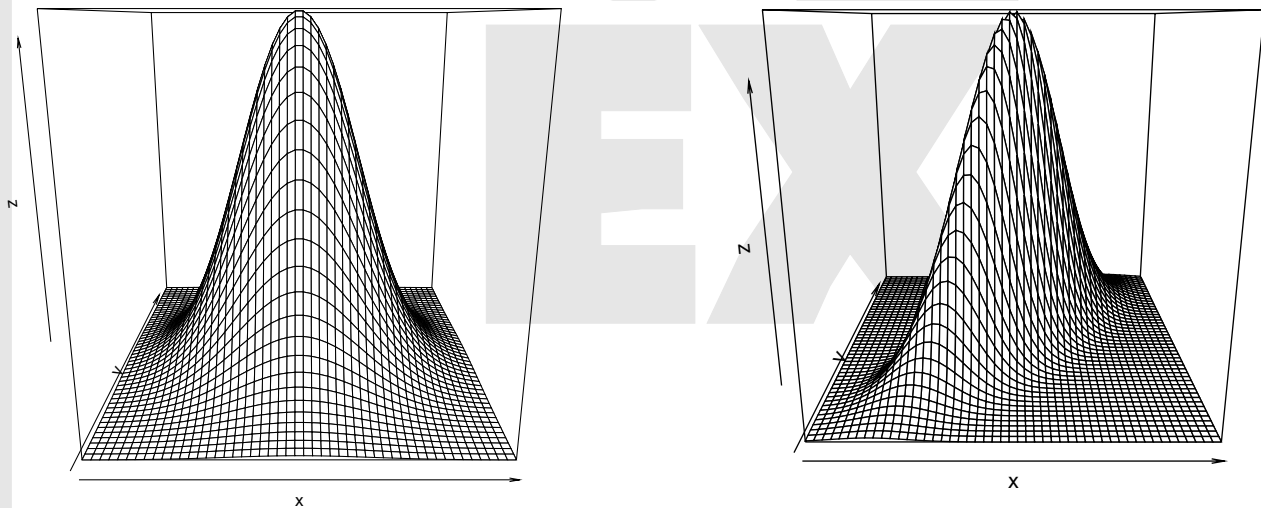


Fig. Densidades de distribuciones Normales Bivariantes

### Distribución Chi-cuadrado no central

TEOREMA. Sea  $X \sim N_n(\mu, I_n)$ , entonces la variable aleatoria  $Y = X^t X$  tiene función de densidad

$$f(x) = \sum_{k=0}^{\infty} \exp\{-\mu^* x\} \frac{\mu^{*k}}{k!} \frac{x^{\frac{n+2k}{2}-1} \exp\{-\frac{x}{2}\}}{\Gamma(\frac{n+2k}{2}) 2^{\frac{n+2k}{2}}} \quad \text{si } x > 0; \quad 0 \quad \text{si } x \leq 0,$$

siendo  $\mu^* = \frac{1}{2} \mu^t \mu$ . Diremos que  $Y$  se distribuye según una **Chi-cuadrado no central** de parámetros  $n$  y  $\mu^*$  y escribiremos  $Y \sim \chi^2(n, \mu^*)$ .

OBSERVACIONES:

- La función generatriz de momentos de una distribución  $\chi^2(n, \lambda)$  es

$$(1 - 2t)^{-n/2} \exp\left\{-\lambda \left(1 - \frac{1}{1 - 2t}\right)\right\} \quad t \text{ en un entorno de } 0$$

- $\chi^2(n, 0) \equiv \chi^2(n)$ .

COROLARIO. Si  $X \sim N_n(\mu, \sigma^2 I_n)$ , entonces  $X^t X / \sigma^2 \sim \chi^2(n, \mu^*)$ , siendo  $\mu^* = \frac{1}{2\sigma^2} \mu^t \mu$ .

COROLARIO. Sean  $Y_i \sim \chi^2(n_i, \mu_i)$ ,  $i = 1, \dots, k$ , variables aleatorias independientes. Entonces

$$\sum_{i=1}^k Y_i \sim \chi^2\left(\sum_{i=1}^k n_i, \sum_{i=1}^k \mu_i\right).$$

COROLARIO. Sea  $X \sim N_n(\mu, D)$ , siendo  $D$  una matriz diagonal definido positiva. Entonces  $X^t D^{-1} X \sim \chi^2(n, \mu^*)$ , siendo  $\mu^* = \frac{1}{2} \mu^t D^{-1} \mu$ .

### Distribución F-Snedecor no central

TEOREMA. Sean  $X \sim \chi^2(n_1, \mu)$  e  $Y \sim \chi^2(n_2, 0)$  dos variables aleatorias independientes. Entonces la variable aleatoria  $Z = \frac{X/n_1}{Y/n_2}$  tiene función de densidad

$$f(x) = \sum_{k=0}^{\infty} \exp\{-\mu\} \frac{\mu^k}{k!} \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1+2k}{2}} \Gamma\left(\frac{n_1+n_2+2k}{2}\right)}{\Gamma\left(\frac{n_1+2k}{2}\right) \Gamma\left(\frac{n_2}{2}\right)} x^{\frac{n_1+2k}{2}-1} \left(1 + \frac{n_1}{n_2} x\right)^{-\frac{n_1+n_2+2k}{2}} \quad \text{si } x > 0; \quad 0 \quad \text{si } x \leq 0$$

Diremos que  $Z$  se distribuye según una **F-Snedecor no central** de parámetros  $n_1$ ,  $n_2$  y  $\mu$  y escribiremos  $Z \sim F(n_1, n_2, \mu)$ .

OBSERVACIÓN:

- $F(n_1, n_2, 0) \equiv F(n_1, n_2)$ .



**Definición de forma cuadrática.**

DEFINICIÓN. Sea  $A$  una matriz  $n \times n$  y  $x \in \mathbb{R}^n$ . Llamaremos **forma cuadrática** a cualquier expresión del tipo:

$$x^t Ax = \sum_{i,j=1}^n a_{ij} x_i x_j$$

OBSERVACIÓN: Existe una única matriz simétrica,  $B = (A + A^t)/2$ , tal que  $x^t Ax = x^t Bx$ ,  $\forall x \in \mathbb{R}^n$  (de ahora en adelante supondremos que la matriz que define la forma cuadrática es simétrica).

DEFINICIÓN.

1. Diremos que una forma cuadrática  $x^t Ax$  (la matriz  $A$ ) es **definido positiva** si  $x^t Ax \geq 0 \forall x \in \mathbb{R}^n$ , siendo  $x^t Ax = 0$  si y sólo si  $x = 0$ .
2. Diremos que una forma cuadrática  $x^t Ax$  (la matriz  $A$ ) es **semidefinido positiva** si  $x^t Ax \geq 0 \forall x \in \mathbb{R}^n$

**Principales momentos de formas cuadráticas.**

LEMA. Sea  $X$  un vector aleatorio  $n$ -dimensional con vector de medias  $\mu$  y matriz de covarianzas  $V$ . Sea  $A$  una matriz simétrica  $n \times n$ . Se verifica:

1.  $E[X^t AX] = tr(AV) + \mu^t A \mu$

Si además  $X \sim N_n(\mu, V)$ , entonces

2.  $Var[X^t AX] = 2tr((AV)^2) + 4\mu^t AV A \mu.$

3.  $Cov[X, X^t AX] = 2VA \mu.$

**Distribuciones de probabilidad asociadas a formas cuadráticas.**

DEFINICIÓN. Diremos que una matriz  $A$   $n \times n$  es **idempotente** si  $AA = A$ .

LEMA.

1. Si  $A$  es una matriz idempotente, entonces  $I - A$  es idempotente.
2. Si  $A$  es una matriz simétrica e idempotente, entonces  $r(A) = tr(A)$ .

TEOREMA 1. Sea  $X \sim N_n(\mu, V)$ . Sea  $A$  una matriz  $n \times n$  simétrica. Entonces se verifica:

$$X^t AX \sim \chi^2(r(A), \frac{1}{2} \mu^t A \mu) \text{ cualquiera que sea } \mu \iff AV \text{ es idempotente.}$$

TEOREMA 2. Sea  $X \sim N_n(\mu, V)$ . Sea  $A$  una matriz  $n \times n$  simétrica y semidefinido positiva y  $B$  una matriz  $s \times n$  de rango  $s \leq n$ . Entonces se verifica:

$$X^t AX \text{ y } BX \text{ son independientes cualquiera que sea } \mu \iff BVA = 0.$$

TEOREMA 3. Sea  $X \sim N_n(\mu, V)$ . Sean  $A$  y  $B$  dos matrices  $n \times n$  simétricas y semidefinido positivas. Entonces se verifica:

$$X^t AX \text{ y } X^t BX \text{ son independientes cualquiera que sea } \mu \iff AVB = 0 (\iff BVA = 0).$$

## Modelo Lineal General.

DEFINICIÓN. Sea  $Y = (Y_1, \dots, Y_n)^t$  un vector aleatorio  $n$ -dimensional y  $X$  una matriz  $n \times p$  ( $p \leq n$ ) de constantes conocidas. Diremos que  $Y$  satisface un **modelo lineal** si

$$E[Y] = X\beta,$$

donde  $\beta = (\beta_1, \dots, \beta_p)^t$  es un vector de parámetros desconocidos.

Es conveniente escribir

$$Y = X\beta + \mathcal{E}, \quad (1)$$

donde  $\mathcal{E} = (\mathcal{E}_1, \dots, \mathcal{E}_n)^t$  es un vector aleatorio no observable con  $E[\mathcal{E}] = 0$ . La relación (1) se conoce como **Modelo Lineal General**.

### OBSERVACIONES:

- $p < n$  significa que tenemos más observaciones que parámetros.
- Rango de la matriz  $X$ :
  - Modelos de rango completo:  $r(X) = p$ .
  - Modelos de rango no completo:  $r(X) < p$ .
- Distribución del vector  $\mathcal{E}$ :
  - Modelo básico:  $E[\mathcal{E}] = 0$ ,  $Cov[\mathcal{E}] = E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ .
  - Modelo Normal:  $\mathcal{E} \sim N_n(0, \sigma^2 I_n)$ .



### Modelo Lineal General: Ejemplos.

EJEMPLO 1. Supongamos que un físico realiza un experimento en el que desea establecer una ecuación que relacione la distancia recorrida por una partícula con el tiempo que lleva en movimiento, esto es

$$s = \beta_0 + \beta_1 t$$

Ahora bien, si por alguna razón no puede medir dicha distancia de manera precisa, en realidad estará midiendo  $y = s + \mathcal{E}$ . El error de medida  $\mathcal{E}$  puede ser considerado una variable aleatoria no observable con  $E[\mathcal{E}] = 0$ , en tanto que  $y$  es una variable aleatoria observable. La ecuación resultante es

$$y = \beta_0 + \beta_1 t + \mathcal{E}$$

A partir de esta ecuación no vamos a poder determinar con exactitud  $\beta_0$  ó  $\beta_1$ , pero si repetimos el experimento obtendremos varias observaciones de  $t$  y de  $y$ , en base a las cuales podremos estimar  $\beta_0$  y  $\beta_1$ . Esto a su vez nos permitirá también dar estimaciones para el valor de  $s$ .

Sean  $Y = (Y_1, \dots, Y_n)^t$  las observaciones de  $y$  en los tiempos  $T = (t_1, \dots, t_n)^t$ . El experimento se ajusta al modelo lineal

$$Y = (\mathbf{1}_n, T) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \mathcal{E}$$

siendo  $\mathbf{1}_n = (1, \dots, 1)^t \in \mathbb{R}^n$  y  $\mathcal{E}$  el vector formado por los errores cometidos en cada observación. Este tipo de modelo es un *modelo de relación funcional* con error de medida.

EJEMPLO 2. Consideremos la relación entre la altura  $h$  y el peso  $w$  de los habitantes de una ciudad. Sabemos que dicha relación existe pero no es funcional como en el ejemplo anterior. Si asumimos que  $(h, w)$  tienen distribución conjunta normal bivalente, entonces el valor esperado para la altura de un individuo dado que su peso es  $w$  viene dado por

$$E[h|w] = \beta_0 + \beta_1 w$$

donde  $\beta_0$  y  $\beta_1$  son funciones de los parámetros de la distribución normal. Si  $\mathcal{E} = h - E[h|w]$  podemos escribir

$$h = \beta_0 + \beta_1 w + \mathcal{E}$$

Para conocer exactamente los valores de  $\beta_0$  y  $\beta_1$ , se procederá igual que en el ejemplo anterior, midiendo las alturas  $H = (H_1, \dots, H_n)^t$  de  $n$  individuos y sus pesos respectivos  $W = (W_1, \dots, W_n)^t$ . Estos datos se ajustarán al modelo lineal

$$H = (\mathbf{1}_n, W) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \mathcal{E}$$

lo cual nos permitirá estimar los valores de  $\beta_0$  y  $\beta_1$ . Este es un *modelo de regresión* y también es un caso especial de modelo lineal.

EJEMPLO 3. Supongamos que un agricultor quiere contrastar el efecto de  $p$  tipos de fertilizantes en el rendimiento de una variedad de trigo. Para ello se utilizará en fertilizante  $i$  en  $n_i$  parcelas y se observará la producción en cada parcela (en Kg/hta.) de dicha variedad de trigo,  $i = 1, \dots, p$ . Así, el rendimiento obtenido en la  $j$ -ésima parcela entre las que se utilizó el fertilizante  $i$ ,  $Y_{ij}$ , puede escribirse como:

$$Y_{ij} = \alpha + \beta_i + \mathcal{E}_{ij}$$

y los datos obtenidos en esta experiencia se ajustan al modelo lineal:

$$Y = X \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \mathcal{E}$$

donde  $Y = (Y_{11}, \dots, Y_{1n_1}, \dots, Y_{p1}, \dots, Y_{pn_p})^t$ ,  $\mathcal{E} = (\mathcal{E}_{11}, \dots, \mathcal{E}_{1n_1}, \dots, \mathcal{E}_{p1}, \dots, \mathcal{E}_{pn_p})^t$  y

$$X = \begin{pmatrix} 1 & 1 & \dots & 0 \\ \vdots & \vdots & n_1 & \vdots \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 1 \\ \vdots & \vdots & n_p & \vdots \\ 1 & 0 & \dots & 1 \end{pmatrix}$$

Este caso especial de modelo lineal es un *modelo de diseño de experimentos*.



### Estimación en el modelo lineal básico.

TEOREMA 1. (Gauss-Markov). Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango completo tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Entonces el mejor estimador lineal insesgado de  $\beta$  es  $\hat{\beta} = S^{-1}X^tY$ , donde  $S = X^tX$  (en el sentido de ser  $\hat{\beta}_i$  el estimador lineal insesgado de mínima varianza de  $\beta_i$ ,  $i = 1, \dots, p$ ).

TEOREMA 2. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango completo tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Sea  $\lambda$  un vector  $p \times 1$  de constantes conocidas. Entonces el estimador lineal insesgado de mínima varianza de  $\lambda^t\beta$  es  $\lambda^t\hat{\beta} = \lambda^tS^{-1}X^tY$ , donde  $S = X^tX$

### Estimación en el modelo lineal Normal.

TEOREMA 3. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango completo tal que  $\mathcal{E} \sim N_n(0, \sigma^2 I_n)$ . Entonces los estimadores:

$$\hat{\beta} = S^{-1}X^tY \quad y \quad \tilde{\sigma}^2 = \frac{1}{n-p}(Y - X\hat{\beta})^t(Y - X\hat{\beta}) = \frac{1}{n-p}Y^t(I_n - XS^{-1}X^t)Y,$$

donde  $S = X^tX$ , verifican las siguientes propiedades:

1. Son insesgados.
2.  $\hat{\beta} \sim N_p(\beta, \sigma^2 S^{-1})$ .
3.  $(n-p)\tilde{\sigma}^2/\sigma^2 \sim \chi^2(n-p)$ .
4.  $\hat{\beta}$  y  $\tilde{\sigma}^2$  son independientes.

COROLARIO 1. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango completo tal que  $\mathcal{E} \sim N_n(0, \sigma^2 I_n)$ . Entonces:

$$I = \left[ \frac{(n-p)\tilde{\sigma}^2}{\alpha_1}, \frac{(n-p)\tilde{\sigma}^2}{\alpha_0} \right],$$

es un intervalo de confianza para  $\sigma^2$  a un nivel de confianza  $1 - \alpha$ , siendo  $\alpha_0$  y  $\alpha_1$  tales que  $\int_{\alpha_0}^{\alpha_1} g(x)dx = 1 - \alpha$ , donde  $g(x)$  es la función de densidad de una distribución  $\chi^2(n-p)$ .

COROLARIO 2. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango completo tal que  $\mathcal{E} \sim N_n(0, \sigma^2 I_n)$ . Sea  $S^{-1} = (X^tX)^{-1} = (c_{ij})_{i,j=1,\dots,p}$ . Entonces:

$$I = \left[ \hat{\beta}_i - t_{\alpha/2}\sqrt{c_{ii}\tilde{\sigma}^2}, \hat{\beta}_i + t_{\alpha/2}\sqrt{c_{ii}\tilde{\sigma}^2} \right],$$

es un intervalo de confianza para  $\beta_i$  a un nivel de confianza  $1 - \alpha$ , siendo  $t_{\alpha/2}$  tal que  $\int_{t_{\alpha/2}}^{\infty} g(x)dx = \alpha/2$ , donde  $g(x)$  es la función de densidad de una distribución  $t(n-p)$ ,  $i = 1, \dots, p$ .

COROLARIO 3. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango completo tal que  $\mathcal{E} \sim N_n(0, \sigma^2 I_n)$ . Sea  $\lambda$  un vector  $p \times 1$  de constantes conocidas. Entonces el estimador de máxima verosimilitud de  $\lambda^t\beta$  es  $\lambda^t\hat{\beta}$ ,  $\lambda^t\hat{\beta} \sim N(\lambda^t\beta, \sigma^2\lambda^tS^{-1}\lambda)$ , donde  $S = X^tX$ , e

$$I = \left[ \lambda^t\hat{\beta} - t_{\alpha/2}\sqrt{\tilde{\sigma}^2\lambda^tS^{-1}\lambda}, \lambda^t\hat{\beta} + t_{\alpha/2}\sqrt{\tilde{\sigma}^2\lambda^tS^{-1}\lambda} \right],$$

es un intervalo de confianza para  $\lambda^t\beta$  a un nivel de confianza  $1 - \alpha$ , siendo  $t_{\alpha/2}$  tal que  $\int_{t_{\alpha/2}}^{\infty} g(x)dx = \alpha/2$ , donde  $g(x)$  es la función de densidad de una distribución  $t(n-p)$ .

**Contraste de la hipótesis  $H_0 : \beta = \beta^*$  ( $\beta^*$  vector de constantes conocidas).**

TEOREMA 4. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango completo,  $r(X) = p$ , tal que  $\mathcal{E} \sim N_n(0, \sigma^2 I_n)$ . El método de razón de verosimilitud para contrastar la hipótesis  $H_0 : \beta = \beta^*$  ( $\beta^*$  vector de constantes conocidas) nos conduce a rechazar  $H_0$  para el nivel de significación  $\alpha$  si  $F \geq F_\alpha(p, n - p)$ , donde

$$F = \frac{n - p}{p} \frac{Q_1}{Q_0},$$

siendo

- $Q = Q_0 + Q_1 = (Y - X\beta^*)^t(Y - X\beta^*)$
- $Q_0 = (Y - X\hat{\beta})^t(Y - X\hat{\beta}) = Y^t(I_n - XS^{-1}X^t)Y$

(con  $\hat{\beta} = S^{-1}X^tY$ ,  $S = X^tX$ ), y  $F_\alpha(p, n - p)$  un valor tal que  $\int_{F_\alpha(p, n - p)}^\infty g(x)dx = \alpha$ , con  $g(x)$  la función de densidad de una distribución  $F(p, n - p)$ . Además  $F \sim F(p, n - p, \lambda)$ ,  $\lambda = \frac{1}{2\sigma^2}(\beta - \beta^*)^tS(\beta - \beta^*)$ .

**Contraste de la hipótesis  $H_0 : r^t\beta = r_0$  ( $r$  vector de constantes conocidas,  $r_0$  valor conocido).**

TEOREMA 5. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango completo,  $r(X) = p$ , tal que  $\mathcal{E} \sim N_n(0, \sigma^2 I_n)$ . El método de razón de verosimilitud para contrastar la hipótesis  $H_0 : r^t\beta = r_0$  ( $r$  vector de constantes conocidas,  $r_0$  valor conocido) nos conduce a rechazar  $H_0$  para el nivel de significación  $\alpha$  si

$$\frac{|r^t\hat{\beta} - r_0|}{\sqrt{\hat{\sigma}^2 r^t S^{-1} r}} \geq t_{\alpha/2}(n - p),$$

siendo  $S = X^tX$  y  $t_{\alpha/2}(n - p)$  un valor tal que  $\int_{t_{\alpha/2}(n - p)}^\infty g(x)dx = \alpha/2$ , con  $g(x)$  la función de densidad de una distribución  $t(n - p)$ .

**Contraste de la hipótesis  $H_0 : \beta_1 = \beta_1^*, \dots, \beta_r = \beta_r^*$  ( $r < p$ ) ( $\beta_i^*$  valor conocido,  $i = 1, \dots, r$ ).**

TEOREMA 6. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango completo,  $r(X) = p$ , con la siguiente partición  $Y = X_1\gamma_1 + X_2\gamma_2 + \mathcal{E}$ , donde  $\gamma_1$  es un vector  $r \times 1$  y  $X_1$  una matriz  $n \times r$  ( $r < p$ ); y tal que  $\mathcal{E} \sim N_n(0, \sigma^2 I_n)$ . El método de razón de verosimilitud para contrastar la hipótesis  $H_0 : \gamma_1 = \gamma_1^*$  ( $\gamma_1^*$  vector de constantes conocidas) nos conduce a rechazar  $H_0$  para el nivel de significación  $\alpha$  si  $F \geq F_\alpha(r, n - p)$ , donde

$$F = \frac{n - p}{r} \frac{Q_1}{Q_0},$$

siendo

- $Q_0 = (Y - X\hat{\beta})^t(Y - X\hat{\beta}) = Y^t(I_n - XS^{-1}X^t)Y$
- $Q_1 = (Y - X_1\gamma_1^*)^t(XS^{-1}X^t - X_2(X_2^tX_2)^{-1}X_2^t)(Y - X_1\gamma_1^*)$

(con  $\hat{\beta} = S^{-1}X^tY$ ,  $S = X^tX$ ), y  $F_\alpha(r, n - p)$  un valor tal que  $\int_{F_\alpha(r, n - p)}^\infty g(x)dx = \alpha$ , con  $g(x)$  la función de densidad de una distribución  $F(r, n - p)$ . Además  $F \sim F(r, n - p, \lambda)$ ,  $\lambda = \frac{1}{2\sigma^2}(\gamma_1 - \gamma_1^*)^tB(\gamma_1 - \gamma_1^*)$ ,  $B = X_1^tX_1 - X_1^tX_2(X_2^tX_2)^{-1}X_2^tX_1$ .



### Estimación en el modelo lineal básico.

DEFINICIÓN. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Diremos que una transformación lineal de los parámetros  $\beta_i, (i = 1, \dots, p), \psi = \lambda^t \beta$ , es una **función lineal estimable** (f.l.e.) si existe un vector  $n \times 1, c$ , tal que

$$E[c^t Y] = \lambda^t \beta \quad \forall \beta$$

i.e.  $c^t Y$  es un estimador insesgado de  $\psi$ .

#### OBSERVACIÓN:

$\psi = \lambda^t \beta$  es f.l.e. si y sólo si  $\lambda$  es una combinación lineal de las filas de  $X$ .

PROPOSICIÓN 1. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Entonces se verifica que  $\psi = \lambda^t \beta$  es f.l.e. si y sólo si existe solución del sistema  $X^t X r = \lambda$ .

PROPOSICIÓN 2. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Sea  $\psi = \lambda^t \beta$  una f.l.e. y  $\hat{\beta}$  una solución de las ecuaciones normales,  $X^t X \beta = X^t Y$ . Sea  $\hat{\psi} = \lambda^t \hat{\beta}$ . Entonces se verifica:

1.  $\hat{\psi} = r^t X^t Y$  siendo  $r$  cualquier solución del sistema  $X^t X r = \lambda$ .
2.  $\hat{\psi}$  es única, no dependiendo de la solución elegida,  $\hat{\beta}$ , de las ecuaciones normales.
3.  $E[\hat{\psi}] = \psi$  y  $Var[\hat{\psi}] = \sigma^2 r^t X^t X r$ , siendo  $r$  cualquier solución del sistema  $X^t X r = \lambda$ .
4.  $\hat{\psi}$  es el estimador lineal insesgado de mínima varianza de  $\psi$ .

DEFINICIÓN. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Sea  $A$  una matriz  $p \times m$ ,  $A = (A_1, \dots, A_m)$ ,  $A_i$  un vector  $p \times 1, i = 1, \dots, m$ . Diremos que  $A^t \beta$  es una **función matricial lineal estimable** si  $A_i^t \beta$  es una f.l.e.,  $i = 1, \dots, m$ .

PROPOSICIÓN 3. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Entonces se verifica que  $X\beta$  y  $X^t X \beta$  son funciones matriciales lineales estimables.

PROPOSICIÓN 4. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Entonces se verifica que el estimador lineal insesgado de mínima varianza de una combinación lineal de f.l.e. es el dado por la misma combinación lineal de los estimadores lineales insesgados de mínima varianza de dichas f.l.e.

### Estimación en el modelo lineal básico.

DEFINICIÓN. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Llamaremos **reparametrización** a cualquier transformación de  $\beta$ ,  $\alpha = U\beta$ , donde  $U$  es una matriz  $k \times p$  de rango  $k$  y  $\alpha$  es una función matricial lineal estimable.

TEOREMA 1. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Entonces existe una matriz  $Z$ ,  $n \times k$  de rango  $k$ , y una reparametrización  $\alpha$  que lo reducen a un modelo de rango completo  $Y = Z\alpha + \mathcal{E}$ .

TEOREMA 2. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Dada una reparametrización  $\alpha^*$ , existe una matriz  $Z^*$ ,  $n \times k$  de rango  $k$ , tal que  $Y = Z^*\alpha^* + \mathcal{E}$ .

### Estimación en el modelo lineal Normal.

TEOREMA 3. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $\mathcal{E} \sim N_n(0, \sigma^2 I_n)$ . Sea  $\alpha$  una reparametrización e  $Y = Z\alpha + \mathcal{E}$  el correspondiente modelo de rango completo. Entonces los estimadores:

$$\hat{\alpha} = (Z^t Z)^{-1} Z^t Y \quad \text{y} \quad \hat{\sigma}^2 = \frac{1}{n-k} (Y - Z\hat{\alpha})^t (Y - Z\hat{\alpha})$$

verifican las siguientes propiedades:

1.  $\hat{\alpha}$  y  $\hat{\sigma}^2$  son estimadores insesgados de  $\alpha$  y  $\sigma^2$ , respectivamente.
2.  $\hat{\alpha} \sim N_k(\alpha, \sigma^2(Z^t Z)^{-1})$ .
3.  $\hat{\sigma}^2 = \frac{1}{n-k} (Y - X\hat{\beta})^t (Y - X\hat{\beta})$ , siendo  $\hat{\beta}$  cualquier solución de las ecuaciones normales,  $X^t X\beta = X^t Y$ . Además  $(n-k)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-k)$ .
4.  $\hat{\alpha}$  y  $\hat{\sigma}^2$  son independientes.

COROLARIO 1. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $\mathcal{E} \sim N_n(0, \sigma^2 I_n)$ . Sea  $\alpha$  una reparametrización e  $Y = Z\alpha + \mathcal{E}$  el correspondiente modelo de rango completo. Entonces:

$$I = \left[ \frac{(n-k)\hat{\sigma}^2}{\alpha_1}, \frac{(n-k)\hat{\sigma}^2}{\alpha_0} \right],$$

es un intervalo de confianza para  $\sigma^2$  a un nivel de confianza  $1 - \alpha$ , siendo  $\alpha_0$  y  $\alpha_1$  tales que  $\int_{\alpha_0}^{\alpha_1} g(x)dx = 1 - \alpha$ , donde  $g(x)$  es la función de densidad de una distribución  $\chi^2(n-k)$ .

COROLARIO 2. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $\mathcal{E} \sim N_n(0, \sigma^2 I_n)$ . Sea  $\alpha$  una reparametrización e  $Y = Z\alpha + \mathcal{E}$  el correspondiente modelo de rango completo. Sea  $\psi = \lambda^t \beta$  una f.l.e. Entonces el estimador de máxima verosimilitud de  $\psi$  es  $\hat{\psi} = \lambda^t \hat{\beta}$ , siendo  $\hat{\beta}$  cualquier solución de las ecuaciones normales,  $X^t X\beta = X^t Y$ .  $\hat{\psi} \sim N(\psi, \sigma^2 a^t (Z^t Z)^{-1} a)$ , donde  $a$  es un vector  $k \times 1$  tal que  $\psi = a^t \alpha$ . Además

$$I = \left[ \hat{\psi} - t_{\alpha/2} \sqrt{\hat{\sigma}^2 a^t (Z^t Z)^{-1} a}, \hat{\psi} + t_{\alpha/2} \sqrt{\hat{\sigma}^2 a^t (Z^t Z)^{-1} a} \right],$$

es un intervalo de confianza para  $\psi$  a un nivel de confianza  $1 - \alpha$ , siendo  $t_{\alpha/2}$  tal que  $\int_{t_{\alpha/2}}^{\infty} g(x)dx = \alpha/2$ , donde  $g(x)$  es la función de densidad de una distribución  $t(n-k)$ .

**Contraste de hipótesis en el modelo lineal de rango no completo.**

DEFINICIÓN. Sea  $Y = X\beta + \mathcal{E}$  un modelo lineal de rango no completo,  $r(X) = k < p$ , tal que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_n$ . Una hipótesis  $H_0$  se dice **estimable** si existe un conjunto de f.l.e. linealmente independientes  $\lambda_1^t \beta, \dots, \lambda_s^t \beta$ ,  $s \leq k$ , tal que  $H_0$  es cierto si y sólo si  $\lambda_1^t \beta = \dots = \lambda_s^t \beta = 0$ .

## Introducción al Diseño de Experimentos

**Diseño estadístico de experimentos:** Proceso de planificación de un experimento del cual se obtendrán datos apropiados, los cuales pueden ser analizados mediante métodos estadísticos dando lugar a conclusiones válidas y objetivas.

### Principios básicos del diseño de experimentos:

- **REPLICACIÓN:** Repetición de un experimento básico.
  - Estimación del error experimental.
- **ALEATORIZACIÓN:** La asignación del material experimental y el orden en la cual las unidades experimentales son elegidas, son determinados aleatoriamente.

### Procedimiento experimental general:

1. Reconocimiento y establecimiento del problema.
2. Elección de factores o niveles:
  - Selección de las variables independientes a ser investigadas en el experimento.
  - Factores: cualitativos, cuantitativos.
3. Selección de la variable respuesta.
4. Elección del diseño de experimentos:
  - Tamaño muestral elegido (replicaciones).
  - Orden en el que se tomarán los datos.
  - Método de aleatorización empleado.
5. Realización del experimento.
6. Análisis de datos.
7. Conclusiones y recomendaciones.



**Diseño completamente aleatorizado: Modelo de efectos fijos.**

EJEMPLO.

Supongamos que un agricultor quiere comparar el rendimiento de 5 variedades de trigo. Un procedimiento posible sería marcar 20 parcelas de tierra y sembrar en ellas las variedades, escogiendo la variedad A para cuatro de ellas, la B para otras cuatro, y así sucesivamente. La asignación de las variedades a las parcelas se haría mediante un procedimiento aleatorio y de cada una de ellas se obtendría una única observación, la producción obtenida en cada parcela.

DEFINICIÓN. Llamaremos **diseño completamente aleatorizado** a aquel en el que la asignación de los niveles del factor o factores, que pueden influir en la variable respuesta, a cada una de las unidades experimentales consideradas se hace totalmente al azar.

UN FACTOR  $\implies$  MODELO DE CLASIFICACIÓN DE UNA VÍA

MODELO:

$$Y_{ij} = \mu + \tau_i + \mathcal{E}_{ij}, \quad i = 1, \dots, r; \quad j = 1, \dots, n_i \quad (N = \sum_{i=1}^r n_i)$$

OBSERVACIÓN:

- El modelo se dice de *efectos fijos* si  $\tau_i, i = 1, \dots, r$ , son parámetros (constantes) desconocidos.
- Si  $n_i = n \forall i$ , el modelo se dice *balanceado*.
- Si  $n_i \neq n$  para algún  $i$ , el modelo se dice *no balanceado*.
- Asumiremos, en general, que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_N$ , con  $\mathcal{E} = (\mathcal{E}_{11}, \dots, \mathcal{E}_{rn_r})^t$ , y para determinadas cuestiones inferenciales que  $\mathcal{E} \sim N_N(0, \sigma^2 I_N)$ .

ECUACIONES NORMALES:

$$\begin{aligned} N\mu + \sum_{i=1}^r n_i \tau_i &= Y_{..} \\ n_i \mu + n_i \tau_i &= Y_{i.}, \quad i = 1, \dots, r \end{aligned}$$

donde  $Y_{i.} = \sum_{j=1}^{n_i} Y_{ij}, i = 1, \dots, r$ , e  $Y_{..} = \sum_{i=1}^r \sum_{j=1}^{n_i} Y_{ij}$

**Estimación en un modelo con un factor, completamente aleatorizado y de efectos fijos**

FUNCIONES LINEALES ESTIMABLES:  $\{\mu + \tau_i, i = 1, \dots, r\}$  es una base del espacio de f.l.e. Además el estimador lineal insesgado de mínima varianza de  $\mu + \tau_i$  es  $\bar{Y}_{i.} = Y_{i.}/n_i, i = 1, \dots, r$

DEFINICIÓN. En un modelo con un factor, completamente aleatorizado y de efectos fijos,  $Y_{ij} = \mu + \tau_i + \mathcal{E}_{ij}, i = 1, \dots, r; j = 1, \dots, n_i (N = \sum_{i=1}^r n_i)$ , llamaremos **contraste** a una función  $\psi = \sum_{i=1}^r c_i \tau_i$  tal que  $\sum_{i=1}^r c_i = 0$ .

OBSERVACIONES:

- $\psi = \sum_{i=1}^r c_i \tau_i$  es una f.l.e. si y sólo si es un contraste.
- Si  $\psi = \sum_{i=1}^r c_i \tau_i$  es un contraste, entonces  $\hat{\psi} = \sum_{i=1}^r c_i \bar{Y}_{i.}, E[\hat{\psi}] = \psi$  y  $Var[\hat{\psi}] = \sigma^2 \sum_{i=1}^r c_i^2/n_i$

**Contraste de hipótesis en un modelo con un factor, completamente aleatorizado y de efectos fijos**

CONTRASTE DE LA HIPÓTESIS  $H_0 : \tau_1 = \dots = \tau_r$

TABLA DE ANÁLISIS DE LA VARIANZA DE UNA VÍA (ANOVA DE 1 VÍA)

Fuentes de variación	g.l.	Suma de cuadrados	Media Cuadrática	F
Debida a la media	1	$R(\mu) = N\bar{Y}_{..}^2$		
Debida a $\tau$ (ajustado)	$r - 1$	$Q_1 = \sum_{i=1}^r n_i \bar{Y}_{i.}^2 - N\bar{Y}_{..}^2$	$MCT = Q_1 / (r - 1)$	$F = \frac{MCT}{MCE}$
Error	$N - r$	$Q_0 = \sum_{i=1}^r \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^r n_i \bar{Y}_{i.}^2$	$MCE = Q_0 / (N - r)$	
Total	$N$	$\sum_{i=1}^r \sum_{j=1}^{n_i} Y_{ij}^2$		

OBSERVACIÓN:

- $F \sim F(r - 1, N - r, \lambda)$ , siendo  $\lambda = \frac{1}{2\sigma^2} \sum_{i=1}^r n_i (\tau_i - \bar{\tau})^2$ , con  $\bar{\tau} = \frac{1}{N} \sum_{i=1}^r n_i \tau_i$



Contraste de hipótesis en un modelo con un factor, completamente aleatorizado y de efectos fijos

COMPARACIONES MÚLTIPLES.

- MÉTODO LSD DE FISHER.

TEOREMA 1. Sea  $Y_{ij} = \mu + \tau_i + \mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, n_i$  ( $N = \sum_{i=1}^r n_i$ ). Sea  $\psi = \sum_{i=1}^r c_i \tau_i$  un contraste y  $\hat{\psi} = \sum_{i=1}^r c_i \bar{Y}_i$ . su estimador lineal insesgado de mínima varianza. Entonces, se verifica:

$$P \left( \psi \in \hat{\psi} \pm t_{\alpha/2}(N-r) \tilde{\sigma} \sqrt{\sum_{i=1}^r \frac{c_i^2}{n_i}} \right) = 1 - \alpha,$$

siendo  $t_{\alpha/2}(N-r)$  tal que si  $T \sim t(N-r)$ ,  $P(T > t_{\alpha/2}(N-r)) = \alpha/2$ .

- MÉTODO DE BONFERRONI.
- MÉTODO DE TUKEY.

DEFINICIÓN. Sean  $Z_1, \dots, Z_k$  y  $U$  v.a. independientes con distribuciones:  $Z_i \sim N(0,1)$ ,  $i = 1, \dots, k$  y  $U \sim \chi^2(n)$ . Definimos la variable aleatoria

$$q = \max_{i \neq j} \frac{|Z_i - Z_j|}{\sqrt{U/n}}.$$

A la distribución de probabilidad asociada a esta variable aleatoria la llamaremos **distribución del rango (recorrido) estudentizado** y la denotaremos por  $q(k, n)$ .

TEOREMA 2. (Intervalos de confianza simultáneos de Tukey). Sea  $Y_{ij} = \mu + \tau_i + \mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, n$ . Entonces, se verifica:

$$P \left( \tau_i - \tau_j \in (\bar{Y}_i. - \bar{Y}_j.) \pm \frac{\tilde{\sigma}}{\sqrt{n}} q_{\alpha}(r, r(n-1)), \text{ para todo } i \neq j \right) = 1 - \alpha,$$

siendo  $q_{\alpha}(r, r(n-1))$  tal que si  $q \sim q(r, r(n-1))$ ,  $P(q > q_{\alpha}(r, r(n-1))) = \alpha$ .

- MÉTODO DE SCHEFFÉ.

TEOREMA 3. (Intervalos de confianza simultáneos de Scheffé). Sea  $Y_{ij} = \mu + \tau_i + \mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, n_i$  ( $N = \sum_{i=1}^r n_i$ ). Entonces, se verifica:

$$P \left( \sum_{i=1}^r c_i \tau_i \in \left( \sum_{i=1}^r c_i \bar{Y}_i. \right) \pm \tilde{\sigma} \sqrt{(r-1) F_{\alpha}(r-1, N-r) \sum_{i=1}^r \frac{c_i^2}{n_i}}, \text{ para todo } c_i \text{ tal que } \sum_{i=1}^r c_i = 0 \right) = 1 - \alpha,$$

siendo  $F_{\alpha}(r-1, N-r)$  tal que si  $F \sim F(r-1, N-r)$ ,  $P(F > F_{\alpha}(r-1, N-r)) = \alpha$ .

### Diseño de bloques completamente aleatorizados.

#### EJEMPLO.

Supongamos que queremos comparar  $r$  catalizadores para el petróleo crudo. Una forma de diseñar el experimento es tomar al azar  $b$  barriles de petróleo crudo, cada uno de ellos dividirlo en  $r$  partes y asignar un catalizador a cada una de esas partes para cada barril. La asignación de los catalizadores a las partes de cada barril la haremos al azar, utilizándose en cada barril los  $r$  catalizadores.

DEFINICIÓN. Llamaremos **diseño de bloques completamente aleatorizados** a aquel en el que la asignación de los niveles del factor o factores, que pueden influir en la variable respuesta, a cada de las partes en que se divide cada bloque se hace totalmente al azar. Además, dentro de cada bloque se aplicarán todos los niveles del factor o factores posibles.

#### MODELO:

$$Y_{ij} = \mu + \tau_i + \alpha_j + \mathcal{E}_{ij}, \quad i = 1, \dots, r; \quad j = 1, \dots, b,$$

donde  $\mu$ ,  $\tau_i$ ,  $i = 1, \dots, r$ , y  $\alpha_j$ ,  $j = 1, \dots, b$  son constantes desconocidas y  $\mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, b$ , son variables aleatorias incorreladas con media 0 y  $\text{Var}[\mathcal{E}_{ij}] = \sigma^2$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, b$ .





### Diseño con dos factores completamente aleatorizado.

EJEMPLO.

El voltaje de salida máximo de un tipo particular de baterías se piensa que está influido por el material utilizado en la placa (tres posibles materiales) y la temperatura del lugar donde se instala la batería (que puede ser 50, 65 u 80°F). Para contrastar este hecho se realizan cuatro replicas del experimento consistente en obtener el voltaje de salida de las correspondientes baterías para cada combinación posible de material y temperatura.

DEFINICIÓN. Llamaremos **experimento factorial** a aquel en el que en cada replicación del mismo se investigan todas las combinaciones posibles de niveles de los factores, (por ejemplo, si el factor A tiene r niveles y el factor B tiene b niveles, entonces cada replicación contiene rb combinaciones de tratamientos).

DOS FACTORES  $\implies$  MODELO DE CLASIFICACIÓN CRUZADA DE DOS VÍAS

### Diseño con dos factores completamente aleatorizado: Modelo de efectos fijos.

#### Diseño con dos factores sin interacción y una observación por celda.

MODELO:

$$Y_{ij} = \mu + \tau_i + \alpha_j + \mathcal{E}_{ij}, \quad i = 1, \dots, r; \quad j = 1, \dots, b$$

OBSERVACIÓN:

- El modelo se dice de *efectos fijos* si  $\tau_i$ ,  $i = 1, \dots, r$ , y  $\alpha_j$ ,  $j = 1, \dots, b$ , son parámetros (constantes) desconocidos.
- Asumiremos, en general, que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_{rb}$ , con  $\mathcal{E} = (\mathcal{E}_{11}, \dots, \mathcal{E}_{rb})^t$ , y para determinadas cuestiones inferenciales que  $\mathcal{E} \sim N_{rb}(0, \sigma^2 I_{rb})$ .

ECUACIONES NORMALES:

$$\begin{aligned} rb\mu + \sum_{i=1}^r b\tau_i + \sum_{j=1}^b r\alpha_j &= Y_{..} \\ b\mu + b\tau_i + \sum_{j=1}^b \alpha_j &= Y_{i.}, \quad i = 1, \dots, r \\ r\mu + \sum_{i=1}^r \tau_i + r\alpha_j &= Y_{.j}, \quad j = 1, \dots, b \end{aligned}$$

donde  $Y_{i.} = \sum_{j=1}^b Y_{ij}$ ,  $i = 1, \dots, r$ ,  $Y_{.j} = \sum_{i=1}^r Y_{ij}$ ,  $j = 1, \dots, b$ , e  $Y_{..} = \sum_{i=1}^r \sum_{j=1}^b Y_{ij}$

### Estimación en un modelo bifactorial completamente aleatorizado y de efectos fijos.

#### Caso sin interacción y una observación por celda.

FUNCIONES LINEALES ESTIMABLES: Una base del espacio de f.l.e. viene dada por  $\mu + \tau_i + \alpha_j$ ,  $i = 1, \dots, r$  y  $\mu + \tau_1 + \alpha_j$ ,  $j = 1, \dots, b$  Además el estimador lineal insesgado de mínima varianza de  $\mu + \tau_i + \alpha_j$  es  $\bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..}$ , siendo  $\bar{Y}_{i.} = Y_{i.}/b$ ,  $\bar{Y}_{.j} = Y_{.j}/r$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$  e  $\bar{Y}_{..} = Y_{..}/rb$

DEFINICIÓN. En un modelo bifactorial de efectos fijos:  $Y_{ij} = \mu + \tau_i + \alpha_j + \mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, b$ , llamaremos **contrastos** a funciones del tipo  $\psi = \sum_{i=1}^r c_i \tau_i$  tal que  $\sum_{i=1}^r c_i = 0$  ó  $\psi' = \sum_{j=1}^b c_j \alpha_j$  tal que  $\sum_{j=1}^b c_j = 0$ .

OBSERVACIONES:

- Un contraste es una f.l.e.
- $\psi = \sum_{i=1}^r c_i \tau_i$  ( $\psi' = \sum_{j=1}^b c_j \alpha_j$ ) es una f.l.e. si y sólo si es un contraste.
- Si  $\psi = \sum_{i=1}^r c_i \tau_i$  ( $\psi' = \sum_{j=1}^b c_j \alpha_j$ ) es un contraste, entonces  $\hat{\psi} = \sum_{i=1}^r c_i \bar{Y}_{i.}$ ,  $E[\hat{\psi}] = \psi$  y  $Var[\hat{\psi}] = \sigma^2 \sum_{i=1}^r c_i^2/b$  ( $\hat{\psi}' = \sum_{j=1}^b c_j \bar{Y}_{.j}$ ,  $E[\hat{\psi}'] = \psi'$  y  $Var[\hat{\psi}'] = \sigma^2 \sum_{j=1}^b c_j^2/r$ )

**Contraste de hipótesis en un modelo bifactorial completamente aleatorizado y de efectos fijos.  
Caso sin interacción y una observación por celda.**

CONTRASTE DE LAS HIPÓTESIS  $H_0 : \tau_1 = \dots = \tau_r$  Y  $H'_0 : \alpha_1 = \dots = \alpha_b$

TABLA DE ANÁLISIS DE LA VARIANZA.

F. Variación	g.l.	Suma de Cuadrados	Media Cuadrática	F	EMC
Factor A	$r - 1$	$SC_A = \sum_{i,j} (\bar{Y}_{i.} - \bar{Y}_{..})^2$	$MCT_A = SC_A / (r - 1)$	$F_A = \frac{MCT_A}{MCE}$	$\sigma^2 + \frac{b}{r-1} \sum_i (\tau_i - \bar{\tau})^2$
Factor B	$b - 1$	$SC_B = \sum_{i,j} (\bar{Y}_{.j} - \bar{Y}_{..})^2$	$MCT_B = SC_B / (b - 1)$	$F_B = \frac{MCT_B}{MCE}$	$\sigma^2 + \frac{r}{b-1} \sum_j (\alpha_j - \bar{\alpha})^2$
Error	$(r - 1)(b - 1)$	$SCE = \sum_{i,j} (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$	$MCE = \frac{SCE}{(r-1)(b-1)}$		$\sigma^2$
Total	$rb - 1$	$SC_T = \sum_{i,j} (Y_{ij} - \bar{Y}_{..})^2$			

OBSERVACIONES:

- $F_A \sim F(r - 1, (r - 1)(b - 1), \lambda_\tau)$ , siendo  $\lambda_\tau = \frac{b}{2\sigma^2} \sum_{i=1}^r (\tau_i - \bar{\tau})^2$ , con  $\bar{\tau} = \frac{1}{r} \sum_{i=1}^r \tau_i$
- $F_B \sim F(b - 1, (r - 1)(b - 1), \lambda_\alpha)$ , siendo  $\lambda_\alpha = \frac{r}{2\sigma^2} \sum_{j=1}^b (\alpha_j - \bar{\alpha})^2$ , con  $\bar{\alpha} = \frac{1}{b} \sum_{j=1}^b \alpha_j$
- Los procedimientos de comparaciones múltiples dados en el caso de un factor son válidos en esta situación, para  $\tau_i - \tau_j$  y  $\alpha_i - \alpha_j$ .

**Diseño con dos factores sin interacción balanceado (con  $n$  observaciones por celda).**

MODELO:

$$Y_{ijk} = \mu + \tau_i + \alpha_j + \mathcal{E}_{ijk}, \quad i = 1, \dots, r; \quad j = 1, \dots, b; \quad k = 1, \dots, n$$

OBSERVACIÓN:

- Asumiremos, en general, que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_{rbn}$ , con  $\mathcal{E} = (\mathcal{E}_{111}, \dots, \mathcal{E}_{rbn})^t$ , y para determinadas cuestiones inferenciales que  $\mathcal{E} \sim N_{rbn}(0, \sigma^2 I_{rbn})$ .



**Diseño con dos factores balanceado con interacción.**

DEFINICIÓN. Un modelo de clasificación cruzada de dos vías,  $Y_{ijk} = \mu_{ij} + \mathcal{E}_{ijk}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ ,  $k = 1, \dots, n$ , se dice que es un **modelo aditivo o sin interacción** si  $(\mu_{ij} - \mu_{kj}) - (\mu_{il} - \mu_{kl}) = 0$ ,  $i, k = 1, \dots, r$ ;  $j, l = 1, \dots, b$ . En otro caso el modelo se dice **con interacción o no aditivo**.

MODELO:

$$Y_{ijk} = \mu + \tau_i + \alpha_j + (\tau\alpha)_{ij} + \mathcal{E}_{ijk}, \quad i = 1, \dots, r; \quad j = 1, \dots, b; \quad k = 1, \dots, n$$

OBSERVACIÓN:

- El modelo se dice de *efectos fijos* si  $\tau_i$ ,  $i = 1, \dots, r$ ,  $\alpha_j$ ,  $j = 1, \dots, b$  y  $(\tau\alpha)_{ij}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ , son parámetros (constantes) desconocidos.
- Asumiremos, en general, que  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_{rbn}$ , con  $\mathcal{E} = (\mathcal{E}_{111}, \dots, \mathcal{E}_{rbn})^t$ , y para determinadas cuestiones inferenciales que  $\mathcal{E} \sim N_{rbn}(0, \sigma^2 I_{rbn})$ .
- El modelo anterior puede escribirse:

$$Y_{ijk} = \mu^* + \tau_i^* + \alpha_j^* + (\tau\alpha)_{ij}^* + \mathcal{E}_{ijk}, \quad i = 1, \dots, r; \quad j = 1, \dots, b; \quad k = 1, \dots, n,$$

siendo  $\mu^* = \mu + \bar{\tau} + \bar{\alpha} + \overline{(\tau\alpha)}$ ;  $\tau_i^* = \tau_i - \bar{\tau} + \overline{(\tau\alpha)}_{i.} - \overline{(\tau\alpha)}$ ;  $\alpha_j^* = \alpha_j - \bar{\alpha} + \overline{(\tau\alpha)}_{.j} - \overline{(\tau\alpha)}$ ,  $j = 1, \dots, b$ ;  $(\tau\alpha)_{ij}^* = (\tau\alpha)_{ij} - \overline{(\tau\alpha)}_{i.} - \overline{(\tau\alpha)}_{.j} + \overline{(\tau\alpha)}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ ; con  $\bar{\tau} = \sum_i \tau_i / r$ ,  $\bar{\alpha} = \sum_j \alpha_j / b$ ,  $\overline{(\tau\alpha)}_{i.} = \sum_j (\tau\alpha)_{ij} / b$ ,  $\overline{(\tau\alpha)}_{.j} = \sum_i (\tau\alpha)_{ij} / r$  y  $\overline{(\tau\alpha)} = \sum_{i,j} (\tau\alpha)_{ij} / rb$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ . Se verifica que  $\mu^* = \sum_{i,j,k} E[Y_{ijk}] / rbn$  y  $\sum_i \tau_i^* = \sum_j \alpha_j^* = 0$ ,  $\sum_i (\tau\alpha)_{ij}^* = 0$ ,  $\forall j$  y  $\sum_j (\tau\alpha)_{ij}^* = 0$ ,  $\forall i$ .

PROPOSICIÓN.  $(\tau\alpha)_{ij}^* = 0$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$  si y sólo si  $(\mu_{ij} - \mu_{kj}) - (\mu_{il} - \mu_{kl}) = 0$ ,  $i, k = 1, \dots, r$ ;  $j, l = 1, \dots, b$ ; siendo  $\mu_{ij} = \mu + \tau_i + \alpha_j + (\tau\alpha)_{ij}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ .

ECUACIONES NORMALES:

$$\begin{aligned} rbn\mu + \sum_{i=1}^r bn\tau_i + \sum_{j=1}^b rn\alpha_j + \sum_{i=1}^r \sum_{j=1}^b n(\tau\alpha)_{ij} &= Y_{...} \\ bn\mu + bn\tau_i + \sum_{j=1}^b n\alpha_j + \sum_{j=1}^b n(\tau\alpha)_{ij} &= Y_{i..}, \quad i = 1, \dots, r \\ rn\mu + \sum_{i=1}^r n\tau_i + rn\alpha_j + \sum_{i=1}^r n(\tau\alpha)_{ij} &= Y_{.j.}, \quad j = 1, \dots, b \\ n\mu + n\tau_i + n\alpha_j + n(\tau\alpha)_{ij} &= Y_{ij.}, \quad i = 1, \dots, r, \quad j = 1, \dots, b \end{aligned}$$

donde  $Y_{ij.} = \sum_{k=1}^n Y_{ijk}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ ,  $Y_{i..} = \sum_{k=1}^n \sum_{j=1}^b Y_{ijk}$ ,  $i = 1, \dots, r$ ,  $Y_{.j.} = \sum_{k=1}^n \sum_{i=1}^r Y_{ijk}$ ,  $j = 1, \dots, b$ , e  $Y_{...} = \sum_{k=1}^n \sum_{i=1}^r \sum_{j=1}^b Y_{ijk}$

**Estimación en un modelo bifactorial completamente aleatorizado y de efectos fijos. Caso con interacción y  $n$  observaciones por celda.**

FUNCIONES LINEALES ESTIMABLES: Una base del espacio de f.l.e. viene dada por  $\psi_{ij} = \mu + \tau_i + \alpha_j + (\tau\alpha)_{ij}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$  Además el estimador lineal insesgado de mínima varianza de  $\psi_{ij}$  es  $\bar{Y}_{ij.}$ , siendo  $\bar{Y}_{ij.} = Y_{ij.} / n$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ .

OBSERVACIÓN:

- Las funciones  $\tau_i^*$ ,  $i = 1, \dots, r$ ,  $\alpha_j^*$ ,  $j = 1, \dots, b$ , y  $(\tau\alpha)_{ij}^*$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ , son f.l.e., siendo sus estimadores lineales insesgados de mínima varianza:  $\hat{\tau}_i^* = \bar{Y}_{i..} - \bar{Y}_{...}$ ,  $i = 1, \dots, r$ ,  $\hat{\alpha}_j^* = \bar{Y}_{.j.} - \bar{Y}_{...}$ ,  $j = 1, \dots, b$ , y  $(\widehat{\tau\alpha})_{ij}^* = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ ; con  $\bar{Y}_{ij.} = Y_{ij.} / n$ ,  $\bar{Y}_{i..} = Y_{i..} / bn$ ,  $\bar{Y}_{.j.} = Y_{.j.} / rn$  e  $\bar{Y}_{...} = Y_{...} / rbn$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ .

**Contraste de hipótesis en un modelo bifactorial completamente aleatorizado y de efectos fijos.  
Caso con interacción y  $n$  observaciones por celda.**

CONTRASTE DE LA HIPÓTESIS  $H_0 : (\tau\alpha)_{ij}^* = 0, i = 1 \dots, r, j = 1 \dots, b.$

TABLA DE ANÁLISIS DE LA VARIANZA.

F. Variación	g.l.	Suma de Cuadrados	Media Cuadrática	F	EMC
Interacción	$(r-1)(b-1)$	$SC_I = \sum_{i,j,k} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$	$MCT_I = \frac{SC_I}{(r-1)(b-1)}$	$F_I = \frac{MCT_I}{MCE}$	$\sigma^2 + \frac{n}{(r-1)(b-1)} \sum_{ij} (\tau\alpha)_{ij}^{*2}$
Error	$rb(n-1)$	$SCE = \sum_{i,j,k} (Y_{ijk} - \bar{Y}_{ij.})^2$	$MCE = \frac{SCE}{rb(n-1)}$		$\sigma^2$
Total	$rbn - 1$	$SC_T = \sum_{i,j,k} (Y_{ijk} - \bar{Y}_{...})^2$			

OBSERVACIONES:

- $F_I \sim F((r-1)(b-1), rb(n-1), \lambda_{(\tau\alpha)})$ , siendo  $\lambda_{(\tau\alpha)} = \frac{n}{2\sigma^2} \sum_{i=1}^r \sum_{j=1}^b (\tau\alpha)_{ij}^{*2}$ .
- Desde un punto de vista práctico para estudiar el modelo bifactorial:

$$Y_{ijk} = \mu + \tau_i + \alpha_j + (\tau\alpha)_{ij} + \mathcal{E}_{ijk}, \quad i = 1, \dots, r; \quad j = 1, \dots, b; \quad k = 1, \dots, n,$$

el primer contraste de hipótesis que debemos realizar es

$$H_0 : (\tau\alpha)_{ij}^* = 0, \quad i = 1 \dots, r, \quad j = 1 \dots, b.$$

Si concluimos que no hay interacción, entonces debemos ajustar un modelo aditivo

$$Y_{ijk} = \mu + \tau_i + \alpha_j + \mathcal{E}_{ijk}, \quad i = 1, \dots, r; \quad j = 1, \dots, b; \quad k = 1, \dots, n.$$

Si probamos la existencia de interacción entre ambos factores, entonces el efecto del factor A en la respuesta depende del nivel del factor B en el que nos encontremos, y viceversa. Así, parece lógico que fijemos cada nivel del factor B (A) y estudiemos el efecto del factor A (B) en la respuesta, por ejemplo mediante un procedimiento de comparaciones múltiples.



### Familia Exponencial de Distribuciones

Sea  $\Theta$  un intervalo de  $\mathbb{R}$ , y sea  $\{f_\theta: \theta \in \Theta\}$  una familia de funciones de densidad (o funciones masa de probabilidad) definidas sobre  $\mathbb{R}^n$ . Supondremos que el conjunto  $\{x \in \mathbb{R}^n: f_\theta(x) > 0\}$  es independiente de  $\theta$ .

DEFINICIÓN: Si existen funciones reales  $Q(\theta)$  y  $D(\theta)$  sobre  $\Theta$  y funciones reales medibles  $T(x_1, \dots, x_n)$  y  $S(x_1, \dots, x_n)$  sobre  $\mathbb{R}^n$  tales que:

$$f_\theta(x_1, \dots, x_n) = \exp\{Q(\theta)T(x_1, \dots, x_n) + D(\theta) + S(x_1, \dots, x_n)\},$$

entonces diremos que la familia  $\{f_\theta: \theta \in \Theta\}$  es una familia exponencial uniparamétrica.

OBSERVACIÓN:

1. Sean  $X_1, \dots, X_m$  vectores aleatorios  $n$ -dimensionales i.i.d. con funciones de densidad (o funciones masa de probabilidad)  $f_\theta(y_1, \dots, y_n)$  pertenecientes a una familia exponencial uniparamétrica. Entonces el vector aleatorio  $X = (X_1, \dots, X_m)^t$  tiene una distribución conjunta con función de densidad (o función masa de probabilidad) perteneciente a una familia exponencial uniparamétrica.
2. Para el caso  $n = 1$ , si  $T(x) = x$  diremos que la familia viene expresada en forma canónica.
3. A  $Q(\theta)$  a veces se le llama parámetro natural de la familia.
4. Para el caso  $n = 1$ , se verifica que:

$$\begin{aligned} E[T(X)] &= -D'(\theta)/Q'(\theta) \\ \text{Var}[T(X)] &= [Q''(\theta)D'(\theta) - D''(\theta)Q'(\theta)]/[Q'(\theta)]^3 \end{aligned}$$

Sea  $\Theta$  un intervalo  $k$ -dimensional de  $\mathbb{R}^k$ , y sea  $\{f_\theta: \theta \in \Theta\}$  una familia de funciones de densidad (o funciones masa de probabilidad) definidas sobre  $\mathbb{R}^n$ . Supondremos que el conjunto  $\{x \in \mathbb{R}^n: f_\theta(x) > 0\}$  es independiente de  $\theta$ .

DEFINICIÓN: Si existen funciones reales  $Q_1(\theta), Q_2(\theta), \dots, Q_k(\theta)$  y  $D(\theta)$  sobre  $\Theta$  y funciones reales medibles  $T_1(x_1, \dots, x_n), T_2(x_1, \dots, x_n), \dots, T_k(x_1, \dots, x_n)$  y  $S(x_1, \dots, x_n)$  sobre  $\mathbb{R}^n$  tales que:

$$f_\theta(x_1, \dots, x_n) = \exp\left\{\sum_{i=1}^k Q_i(\theta)T_i(x_1, \dots, x_n) + D(\theta) + S(x_1, \dots, x_n)\right\},$$

entonces diremos que la familia  $\{f_\theta: \theta \in \Theta\}$  es una familia exponencial  $k$ -paramétrica.

## Modelo Lineal Generalizado

DEFINICIÓN (NELDER AND WEDDERBURN (1972)): Sea  $Y = (Y_1, \dots, Y_n)^t$  un vector aleatorio  $n$ -dimensional,  $X$  una matriz  $n \times p$  ( $p < n$ ) de constantes conocidas:

$$X = \begin{pmatrix} x_1^t \\ \vdots \\ x_n^t \end{pmatrix},$$

$g: \mathbb{R} \rightarrow \mathbb{R}$  una función monótona y derivable (llamada función de enlace ó link function) y  $\beta = (\beta_1, \dots, \beta_p)^t$  un vector de parámetros desconocidos. Diremos que  $Y$  satisface un **modelo lineal generalizado** si

1.  $Y_1, \dots, Y_n$  son v.a. independientes con funciones de densidad (o funciones masa de probabilidad) pertenecientes a una familia exponencial uniparamétrica expresada en forma canónica, i.e.

$$f_{\theta_i}(y) = \exp\{Q(\theta_i)y + D(\theta_i) + S(y)\}, \quad y \in \mathbb{R}; \quad i = 1, \dots, n.$$

2.  $g(E[Y_i]) = x_i^t \beta$ ,  $i = 1, \dots, n$ .

OBSERVACIÓN: Si definimos  $G: \mathbb{R}^n \rightarrow \mathbb{R}^n$   $(x_1, \dots, x_n)^t \rightarrow G(x_1, \dots, x_n) = (g(x_1), \dots, g(x_n))^t$ , entonces la segunda condición se puede escribir:

$$G(E[Y]) = X\beta$$



### Modelo Lineal Generalizado

EJEMPLOS:

1. Modelo lineal normal.
2. Tendencia de la mortalidad: Para poblaciones grandes la probabilidad de que un individuo elegido al azar muera de una enfermedad dada en un instante determinado es pequeña. Si suponemos que la muerte de diferentes individuos son sucesos independientes, entonces el número de muertos,  $Y$ , durante un período de tiempo fijo puede ser modelizado por una distribución de Poisson

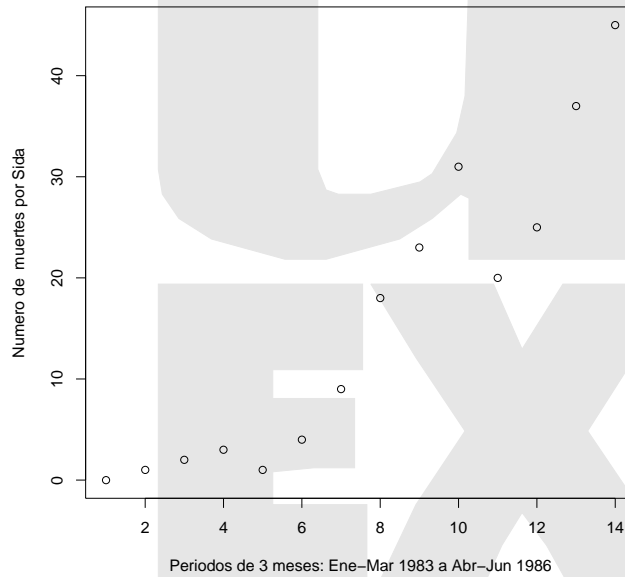
$$f_{\lambda}(y) = \frac{\lambda^y e^{-\lambda}}{y!} \quad y = 0, 1, \dots$$

siendo  $\lambda$  el número medio de muertos por período de tiempo.

La tendencia de la mortalidad puede ser modelizada tomando v.a. independientes  $Y_1, \dots, Y_n$  que sean el número de muertes ocurridas en intervalos de tiempos sucesivos numerados por  $i = 1, \dots, n$ . Sea  $\lambda_i = E[Y_i]$ ,  $i = 1, \dots, n$ .

El número de muertes por Sida en Australia en períodos de tres meses entre 1983 y 1986 aparecen en la siguiente tabla y el siguiente gráfico:

$i$	1	2	3	4	5	6	7
$y_i$	0	1	2	3	1	4	9
$i$	8	9	10	11	12	13	14
$y_i$	18	23	31	20	25	37	45



Claramente el número de muertos crece con  $i$ . Para estos datos un modelo posible es el basado en la distribución de Poisson con  $\lambda_i = i^\theta$ , donde  $\theta$  es un parámetro a ser estimado. Este modelo puede ser descrito como un modelo lineal generalizado en el cual la función de enlace es  $g(\lambda_i) = \log \lambda_i = \theta \log i$  y por tanto  $x_i^t = (\log i)$ ,  $i = 1, \dots, n$ , y  $\beta = (\theta)$ .

### Estimación en Modelos Lineales Generalizados

Consideremos un modelo lineal generalizado como el dado en la definición y utilicemos el método de máxima verosimilitud para estimar el vector de parámetros  $\beta$ . Así la log-verosimilitud será:

$$\ell(\theta; y) = \sum_{i=1}^n y_i Q(\theta_i) + \sum_{i=1}^n D(\theta_i) + \sum_{i=1}^n S(y_i) \quad \theta = (\theta_1, \dots, \theta_n)^t, \quad y = (y_1, \dots, y_n)^t$$

donde  $\mu_i = E[Y_i] = -D'(\theta_i)/Q'(\theta_i)$  y  $\eta_i = g(\mu_i) = x_i^t \beta, i = 1, \dots, n$ .

Una propiedad de la familia exponencial de distribuciones es que satisface suficientes condiciones de regularidad para asegurar que el máximo global de la función de log-verosimilitud se obtiene de manera única como solución de las ecuaciones

$$\frac{\partial \ell}{\partial \theta} = 0,$$

o equivalentemente

$$\frac{\partial \ell}{\partial \beta} = 0.$$

Se prueba que para  $j = 1, \dots, p$

$$U_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n \frac{(y_i - \mu_i) x_{ij}}{Var[Y_i]} \frac{\partial \mu_i}{\partial \eta_i}$$

donde  $x_{ij}$  es el  $j$ -ésimo elemento de  $x_i^t$ .

En general las ecuaciones  $U_j = 0, j = 1, \dots, p$  son no lineales y tienen que ser resueltas por métodos numéricos. Si utilizamos el Método de Newton-Raphson la  $m$ -ésima aproximación viene dada por

$$b^{(m)} = b^{(m-1)} - \left( \frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} \right)_{\beta=b^{(m-1)}}^{-1} U^{(m-1)}$$

donde  $U^{(m-1)} = (U_1, \dots, U_p)_{\beta=b^{(m-1)}}^t$ .

Un procedimiento alternativo, a veces más sencillo que el método de Newton-Raphson, es el método de tanteo (method of scoring), consistente en reemplazar en la ecuación anterior  $\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}$  por  $E[\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}]$ .

Para obtener la distribución asociada al estimador máximo verosímil de  $\beta, \hat{\beta}$ , juegan un papel muy importante las cantidades  $U_j, j = 1, \dots, p$ . Para modelos lineales generalizados se verifica que  $E[U_j] = 0, j = 1, \dots, p$ . Además denotaremos su matriz de covarianzas por  $\mathcal{I} = E[UU^t]$ , siendo  $U = (U_1, \dots, U_p)^t$ . Supondremos que  $\mathcal{I}$  es no singular.

Se prueba que, asintóticamente,  $U \sim N_p(0, \mathcal{I})$  y por tanto  $U^t \mathcal{I}^{-1} U \sim \chi^2(p)$ . A partir de estos resultados se demuestra que, asintóticamente,

$$\hat{\beta} \sim N_p(\beta, \mathcal{I}^{-1})$$

y por tanto

$$(\hat{\beta} - \beta)^t \mathcal{I} (\hat{\beta} - \beta) \sim \chi^2(p)$$

Al estadístico  $(\hat{\beta} - \beta)^t \mathcal{I} (\hat{\beta} - \beta)$  se le suele llamar estadístico de Wald.

#### OBSERVACIONES:

1. Para modelos lineales normales los resultados anteriores son exactos.
2.  $\mathcal{I}$  dependerá en general de  $\beta$  con lo cual, en aplicaciones prácticas, la consideraremos evaluada en  $\hat{\beta}$ .
3. El resultado anterior nos permite obtener intervalos de confianza aproximados para  $\beta_i, i = 1, \dots, p$ .

Si  $\mathcal{I}^{-1} = (v_{ij})_{i,j=1,\dots,p}$ , entonces  $\hat{\beta}_i \pm z_{\alpha/2} \sqrt{v_{ii}}$  es un intervalo de confianza aproximado al nivel  $1 - \alpha$  para el parámetro  $\beta_i$ , siendo  $z_{\alpha/2}$  un valor tal que si  $Z \sim N(0, 1)$ , entonces  $P(Z > z_{\alpha/2}) = \alpha/2$ .



## Contraste de Hipótesis en Modelos Lineales Generalizados

BONDAD DE AJUSTE. Valoración de lo adecuado que es un modelo para describir un conjunto de datos.

Método: comparar la verosimilitud del modelo propuesto con la verosimilitud del modelo maximal o saturado, siendo éste el que tiene la misma distribución y función de enlace que el modelo de interés, y además el número de parámetros es igual al total de observaciones (se considera que este modelo proporciona una descripción completa de los datos, al menos para la distribución considerada).

- $L(\hat{\beta}_{max}; y)$ ,  $L(\hat{\beta}; y)$
- Estadístico de la razón de verosimilitud generalizado:

$$\lambda = \frac{L(\hat{\beta}_{max}; y)}{L(\hat{\beta}; y)}$$

ó

$$\log \lambda = \ell(\hat{\beta}_{max}; y) - \ell(\hat{\beta}; y)$$

- Criterio: valores grandes de  $\log \lambda$  indican que el modelo propuesto proporciona una pobre descripción de los datos.
- Distribución muestral de  $\log \lambda$ : se verifica que asintóticamente  $2(\ell(\hat{\beta}; y) - \ell(\beta; y)) \sim \chi^2(p)$ , ( $\beta$  es  $p \times 1$ ).
- Deviance (Nelder and Wedderburn (1972)):  $D = 2 \log \lambda = 2(\ell(\hat{\beta}_{max}; y) - \ell(\hat{\beta}; y))$

$$D = 2((\ell(\hat{\beta}_{max}; y) - \ell(\beta_{max}; y)) - (\ell(\hat{\beta}; y) - \ell(\beta; y)) + (\ell(\beta_{max}; y) - \ell(\beta; y)))$$

- Si el modelo propuesto es adecuado, entonces  $D \sim \chi^2(n - p)$  (asintóticamente).

TEST DE HIPÓTESIS SOBRE  $\beta$ .

$$\begin{aligned} H_0 &: \beta = \beta_0 \quad (\beta_0 = (\beta_1, \dots, \beta_q)^t) \\ H_1 &: \beta = \beta_1 \quad (\beta_1 = (\beta_1, \dots, \beta_p)^t; \quad q < p < n) \end{aligned}$$

Método:

- Cada hipótesis indica un modelo.
- Los modelos a ser comparados deben tener la misma distribución y función de enlace, diferenciándose exclusivamente en el número de parámetros.
- Se comparan los estadísticos de bondad de ajuste para ambos modelos:
  - $D_0$ : deviance asociada al modelo dado en  $H_0$ .  $D_1$ : deviance asociada al modelo dado en  $H_1$ .
  - $\Delta D = D_0 - D_1 = 2(\ell(\hat{\beta}_1; y) - \ell(\hat{\beta}_0; y))$ .
  - Si los dos modelos ( $H_0$  y  $H_1$ ) describen bien los datos, entonces, asintóticamente,  $D_0 \sim \chi^2(n - q)$  y  $D_1 \sim \chi^2(n - p)$ . Bajo ciertas condiciones de independencia, de lo anterior se obtiene que, asintóticamente,  $\Delta D \sim \chi^2(p - q)$ .



### Estimación y Contraste de Hipótesis en Modelos Lineales Generalizados

EJEMPLO. Se pretende modelizar, para enfermos de Leucemia, la relación existente entre el número de semanas transcurridas hasta la muerte de un paciente desde su diagnóstico y el  $\log_{10}$  del número inicial de leucocitos que presentaba. Para ello se han tomado 17 pacientes con Leucemia, cuyos datos están en la siguiente tabla:

$x_i$	3.36	2.88	3.63	3.41	3.78	4.02	4	4.23	3.73	3.85
$y_i$	65	156	100	134	16	108	121	4	39	143
$x_i$	3.97	4.51	4.54	5	5	4.72	5			
$y_i$	56	26	22	1	1	5	65			

MODELO.

- La distribución exponencial es utilizada habitualmente para describir tiempos de supervivencia:

$$Y_i \sim \text{Exp}(\theta_i), \quad f_{\theta_i}(y_i) = \theta_i \exp\{-y_i\theta_i\}, \quad i = 1, \dots, n.$$

- Una posible especificación para  $E[Y]$  es  $E[Y_i] = \exp\{\beta_1 + \beta_2 x_i\}$ ,  $i = 1, \dots, n$ , que asegura que  $E[Y]$  es no negativa para todos los valores de los parámetros y todos los valores de  $x$ . En este caso la función de enlace es  $g(x) = \log(x)$ .

RESULTADOS.

- Estimación de  $\beta$

```
> regr<-glm(y~x,family=Gamma(link="log"))
> summary(regr)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.4766	1.6003	5.297	8.96e-05 ***
x	-1.1091	0.3865	-2.870	0.0117 *

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Null deviance: 26.282 on 16 degrees of freedom  
Residual deviance: 19.457 on 15 degrees of freedom

$$\hat{\beta}_1 = 8.4766, \hat{\beta}_2 = -1.1091$$

–  $\beta_1 \in 8.4766 \pm 1.96 * 1.6003 = [5.34, 11.61]$  con una confianza del 95%.

–  $\beta_2 \in -1.1091 \pm 1.96 * 0.3865 = [-1.87, -0.35]$  con una confianza del 95%.

- Adecuación del modelo:  $D = 19.457$  (15 g.l.),  $p = 0.1938$ .
- Contraste  $H_0 : \beta_2 = 0$  vs.  $H_1 : \beta_2 \neq 0$

```
> regr1<-glm(y~1,family=Gamma(link="log"))
> anova(regr1,regr,test="Chisq")
```

Analysis of Deviance Table

Model 1: y ~ 1

Model 2: y ~ x

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi )
1	16	26.2821			
2	15	19.4565	1	6.8256	0.0069



## Diseño completamente aleatorizado: Modelo de efectos aleatorios.

## EJEMPLO.

Una compañía textil produce un tipo de fibra en un gran número de telares. La compañía quiere que los telares sean homogéneos, de manera que se obtenga de ellos una fibra de resistencia uniforme. El ingeniero que controla el proceso de producción sospecha que, además de la variación de resistencias dentro de muestras producidas por el mismo telar, también puede haber variaciones significativas entre los telares. Para investigar esto se seleccionan 4 telares aleatoriamente y se toman 6 determinaciones de dureza para la fibra manufacturada por cada telar. El experimento se realiza en orden aleatorio. La pregunta que se pretende contestar es si difieren significativamente los telares de la compañía.

## MODELO:

$$Y_{ij} = \mu + \tau_i + \mathcal{E}_{ij}, \quad i = 1, \dots, r; \quad j = 1, \dots, n_i \quad (N = \sum_{i=1}^r n_i),$$

donde  $\mu$  es una constante desconocida y  $\tau_i$  y  $\mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, n_i$ , son variables aleatorias.

## OBSERVACIONES:

- El modelo se dice de *efectos aleatorios* pues  $\tau_i$ ,  $i = 1, \dots, r$ , son variables aleatorias.
- Asumiremos, en general, que:
  1.  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_N$ , con  $\mathcal{E} = (\mathcal{E}_{11}, \dots, \mathcal{E}_{rn_r})^t$
  2.  $E[\tau] = 0$  y  $E[\tau\tau^t] = \sigma_\tau^2 I_r$ , con  $\tau = (\tau_1, \dots, \tau_r)^t$
  3. Las v.a.  $\tau_i$  y  $\mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, n_i$  son incorreladas.

Para determinadas cuestiones inferenciales supondremos, además, que:

1.  $\mathcal{E} \sim N_N(0, \sigma^2 I_N)$ .
  2.  $\tau \sim N_r(0, \sigma_\tau^2 I_r)$ .
- $Var[Y_{ij}] = \sigma_\tau^2 + \sigma^2 \forall i, j$ . A las varianzas  $\sigma_\tau^2$  y  $\sigma^2$  se las llama *componentes de la varianza*. Así el modelo general también se conoce como *modelo de las componentes de la varianza*.
  - Si  $n_i = n \forall i$ , el modelo se dice *balanceado*.
  - Si  $n_i \neq n$  para algún  $i$ , el modelo se dice *no balanceado*.

PROPOSICIÓN 1. Sea  $Y_{ij} = \mu + \tau_i + \mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, n_i$  ( $N = \sum_{i=1}^r n_i$ ), donde  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_N$ , con  $\mathcal{E} = (\mathcal{E}_{11}, \dots, \mathcal{E}_{rn_r})^t$ ;  $E[\tau] = 0$  y  $E[\tau\tau^t] = \sigma_\tau^2 I_r$ , con  $\tau = (\tau_1, \dots, \tau_r)^t$ ; y las v.a.  $\tau_i$  y  $\mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, n_i$  son incorreladas. Se verifica que  $Cov[Y_{ij}, Y_{kl}] = 0$  si  $i \neq k$  y

$$Cov[Y_{ij}, Y_{il}] = \begin{cases} \sigma_\tau^2 & \text{si } j \neq l \\ \sigma_\tau^2 + \sigma^2 & \text{si } j = l \end{cases}$$

### Inferencia en un modelo con un factor, completamente aleatorizado y de efectos aleatorios

LEMA. Sean  $Z_1, \dots, Z_k$  variables aleatorias incorreladas de media  $\mu$  y varianza  $\sigma^2$ . Entonces se verifica que:

$$E \left[ \sum_{i=1}^k (Z_i - \bar{Z})^2 \right] = (k-1)\sigma^2.$$

PROPOSICIÓN 2. Sea  $Y_{ij} = \mu + \tau_i + \mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, n_i$  ( $N = \sum_{i=1}^r n_i$ ), donde  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_N$ , con  $\mathcal{E} = (\mathcal{E}_{11}, \dots, \mathcal{E}_{rn_r})^t$ ;  $E[\tau] = 0$  y  $E[\tau\tau^t] = \sigma_\tau^2 I_r$ , con  $\tau = (\tau_1, \dots, \tau_r)^t$ ; y las v.a.  $\tau_i$  y  $\mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, n_i$  son incorreladas. Sean

$$MCE = \frac{1}{N-r} \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \quad \text{y} \quad MCT = \frac{1}{r-1} \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{Y}_i - \bar{Y}_{..})^2.$$

Entonces se verifica:

1.  $E[MCE] = \sigma^2$ .
2. Si el modelo es balanceado con  $n_i = n$ ,  $i = 1, \dots, r$ , entonces  $E[MCT] = \sigma^2 + k_0 \sigma_\tau^2$ , siendo  $k_0 = n$ .

OBSERVACIONES:

- Si el modelo es no balanceado, entonces  $k_0 = \frac{N^2 - \sum_{i=1}^r n_i^2}{N(r-1)} (> 0)$ , siendo  $N = \sum_{i=1}^r n_i$ .
- El resultado anterior nos indica que  $MCE$  sigue siendo en esta situación un estimador insesgado de  $\sigma^2$ . Además nos permite obtener un estimador insesgado de  $\sigma_\tau^2$

$$\hat{\sigma}_\tau^2 = \left( \frac{1}{r-1} \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{Y}_i - \bar{Y}_{..})^2 - \frac{1}{N-r} \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \right) \frac{1}{k_0}$$

- En el caso balanceado estos estimadores tienen propiedades óptimas.

LEMA. Sea  $Z = (Z_1, \dots, Z_n)^t \sim N_n(\mu, \sigma^2 I_n)$ . Sea  $U = \frac{1}{\sigma^2} \sum_{i=1}^n (Z_i - \bar{Z})^2$ . Entonces  $U \sim \chi^2(n-1)$

PROPOSICIÓN 3. Sea  $Y_{ij} = \mu + \tau_i + \mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, n$ , donde  $\mathcal{E} \sim N_{rn}(0, \sigma^2 I_{rn})$ , con  $\mathcal{E} = (\mathcal{E}_{11}, \dots, \mathcal{E}_{rn})^t$ ;  $\tau \sim N_r(0, \sigma_\tau^2 I_r)$ , con  $\tau = (\tau_1, \dots, \tau_r)^t$ ; y las v.a.  $\tau_i$  y  $\mathcal{E}_{ij}$ ,  $i = 1, \dots, r$ ;  $j = 1, \dots, n$  son independientes. Entonces se verifica:

1.  $\frac{1}{\sigma^2} \sum_{i=1}^r \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2 \sim \chi^2(r(n-1))$ .
2.  $\frac{1}{\sigma^2 + n\sigma_\tau^2} \sum_{i=1}^r \sum_{j=1}^n (\bar{Y}_i - \bar{Y}_{..})^2 \sim \chi^2(r-1)$ .
3. Los estadísticos anteriores son independientes.

OBSERVACIÓN:

- Para el modelo balanceado, del resultado anterior se deduce que

$$F = \frac{\sigma^2}{\sigma^2 + n\sigma_\tau^2} \frac{MCT}{MCE} \sim F(r-1, r(n-1))$$

- Para el modelo balanceado, podemos realizar el contraste de hipótesis:

$$H_0 : \sigma_\tau^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_\tau^2 > 0$$

TABLA DE ANÁLISIS DE LA VARIANZA DE UNA VÍA: EFECTOS ALEATORIOS

Fuentes de variación	g.l.	Suma de cuadrados	Media Cuadrática	EMC	F
Debida a la media	1	$N\bar{Y}_{..}^2$			
Entre tratamientos	$r - 1$	$\sum_{i=1}^r \sum_{j=1}^n (\bar{Y}_{i.} - \bar{Y}_{..})^2$	$MCT = \frac{1}{r-1} \sum_{i=1}^r \sum_{j=1}^n (\bar{Y}_{i.} - \bar{Y}_{..})^2$	$\sigma^2 + n\sigma_\tau^2$	$F = \frac{MCT}{MCE}$
Dentro tratamientos	$r(n - 1)$	$\sum_{i=1}^r \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2$	$MCE = \frac{1}{r(n-1)} \sum_{i=1}^r \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2$	$\sigma^2$	
Total	$rn$	$\sum_{i=1}^r \sum_{j=1}^n Y_{ij}^2$			



**Diseño con dos factores completamente aleatorizado: Modelo de efectos aleatorios.**

**Diseño con dos factores sin interacción y  $n$  observaciones por celda.**

MODELO:

$$Y_{ijk} = \mu + \tau_i + \alpha_j + \mathcal{E}_{ijk}, \quad i = 1, \dots, r; \quad j = 1, \dots, b \quad k = 1, \dots, n,$$

donde  $\mu$  es una constante desconocida y  $\tau_i, \alpha_j$  y  $\mathcal{E}_{ijk}, i = 1, \dots, r, j = 1, \dots, b, k = 1, \dots, n$ , son variables aleatorias.

OBSERVACIONES:

- El modelo se dice de efectos aleatorios pues  $\tau_i, i = 1, \dots, r$ , y  $\alpha_j, j = 1, \dots, b$  son variables aleatorias.
- Asumiremos, en general, que:

1.  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_{rbn}$ , con  $\mathcal{E} = (\mathcal{E}_{111}, \dots, \mathcal{E}_{rbn})^t$
2.  $E[\tau] = 0$  y  $E[\tau\tau^t] = \sigma_\tau^2 I_r$ , con  $\tau = (\tau_1, \dots, \tau_r)^t$
3.  $E[\alpha] = 0$  y  $E[\alpha\alpha^t] = \sigma_\alpha^2 I_b$ , con  $\alpha = (\alpha_1, \dots, \alpha_b)^t$
4. Las v.a.  $\tau_i, \alpha_j$  y  $\mathcal{E}_{ijk}, i = 1, \dots, r, j = 1, \dots, b, k = 1, \dots, n$  son incorreladas.

Para determinadas cuestiones inferenciales supondremos, además, que:

1.  $\mathcal{E} \sim N_{rbn}(0, \sigma^2 I_{rbn})$ .
  2.  $\tau \sim N_r(0, \sigma_\tau^2 I_r)$ .
  3.  $\alpha \sim N_b(0, \sigma_\alpha^2 I_b)$ .
- $Var[Y_{ijk}] = \sigma_\tau^2 + \sigma_\alpha^2 + \sigma^2 \forall i, j, k$ . A las varianzas  $\sigma_\tau^2, \sigma_\alpha^2$  y  $\sigma^2$  se las llama componentes de la varianza. Así el modelo general también se conoce como modelo de las componentes de la varianza.

CONTRASTE DE LAS HIPÓTESIS  $H_0 : \sigma_\tau^2 = 0$  vs.  $H_1 : \sigma_\tau^2 > 0$  Y  $H'_0 : \sigma_\alpha^2 = 0$  vs.  $H'_1 : \sigma_\alpha^2 > 0$

TABLA DE ANÁLISIS DE LA VARIANZA.

F. Variación	g.l.	Suma de Cuadrados	Media Cuadrática	F	EMC
Factor A	$r - 1$	$SC_A = \sum_{i,j,k} (\bar{Y}_{i..} - \bar{Y}_{...})^2$	$MCT_A = SC_A / (r - 1)$	$F_A = \frac{MCT_A}{MCE}$	$\sigma^2 + b\sigma_\tau^2$
Factor B	$b - 1$	$SC_B = \sum_{i,j,k} (\bar{Y}_{.j.} - \bar{Y}_{...})^2$	$MCT_B = SC_B / (b - 1)$	$F_B = \frac{MCT_B}{MCE}$	$\sigma^2 + r\sigma_\alpha^2$
Error	$(r - 1)(b - 1)$	$SCE = \sum_{i,j,k} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$	$MCE = \frac{SCE}{(r-1)(b-1)}$		$\sigma^2$

OBSERVACIONES:

- Si  $H_0$  es cierto, entonces  $F_A \sim F(r - 1, (r - 1)(b - 1))$ .
- Si  $H'_0$  es cierto, entonces  $F_B \sim F(b - 1, (r - 1)(b - 1))$ .



**Diseño con dos factores completamente aleatorizado: Modelo de efectos aleatorios.**

**Diseño con dos factores con interacción y  $n$  observaciones por celda.**

MODELO:

$$Y_{ijk} = \mu + \tau_i + \alpha_j + (\tau\alpha)_{ij} + \mathcal{E}_{ijk}, \quad i = 1, \dots, r; \quad j = 1, \dots, b \quad k = 1, \dots, n,$$

donde  $\mu$  es una constante desconocida y  $\tau_i, \alpha_j, (\tau\alpha)_{ij}$  y  $\mathcal{E}_{ijk}, i = 1, \dots, r, j = 1, \dots, b, k = 1, \dots, n$ , son variables aleatorias.

OBSERVACIONES:

- El modelo se dice de efectos aleatorios pues  $\tau_i, \alpha_j$  y  $(\tau\alpha)_{ij}, i = 1, \dots, r, j = 1, \dots, b$ , son variables aleatorias.

- Asumiremos, en general, que:

1.  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_{rbn}$ , con  $\mathcal{E} = (\mathcal{E}_{111}, \dots, \mathcal{E}_{rbn})^t$
2.  $E[\tau] = 0$  y  $E[\tau\tau^t] = \sigma_\tau^2 I_r$ , con  $\tau = (\tau_1, \dots, \tau_r)^t$
3.  $E[\alpha] = 0$  y  $E[\alpha\alpha^t] = \sigma_\alpha^2 I_b$ , con  $\alpha = (\alpha_1, \dots, \alpha_b)^t$
4.  $E[(\tau\alpha)] = 0$  y  $E[(\tau\alpha)(\tau\alpha)^t] = \sigma_{\tau\alpha}^2 I_{rb}$ , con  $(\tau\alpha) = ((\tau\alpha)_{11}, \dots, (\tau\alpha)_{rb})^t$
5. Las v.a.  $\tau_i, \alpha_j, (\tau\alpha)_{ij}$  y  $\mathcal{E}_{ijk}, i = 1, \dots, r, j = 1, \dots, b, k = 1, \dots, n$  son incorreladas.

Para determinadas cuestiones inferenciales supondremos, además, que:

1.  $\mathcal{E} \sim N_{rbn}(0, \sigma^2 I_{rbn})$ .
  2.  $\tau \sim N_r(0, \sigma_\tau^2 I_r)$ .
  3.  $\alpha \sim N_b(0, \sigma_\alpha^2 I_b)$ .
  4.  $(\tau\alpha) \sim N_{rb}(0, \sigma_{\tau\alpha}^2 I_{rb})$ .
- $Var[Y_{ijk}] = \sigma_\tau^2 + \sigma_\alpha^2 + \sigma_{\tau\alpha}^2 + \sigma^2 \forall i, j, k$ . A las varianzas  $\sigma_\tau^2, \sigma_\alpha^2, \sigma_{\tau\alpha}^2$  y  $\sigma^2$  se las llama componentes de la varianza. Así el modelo general también se conoce como modelo de las componentes de la varianza.

CONTRASTE DE LAS HIPÓTESIS:

1.  $H_0 : \sigma_\tau^2 = 0 \quad vs. \quad H_1 : \sigma_\tau^2 > 0$ .
2.  $H'_0 : \sigma_\alpha^2 = 0 \quad vs. \quad H'_1 : \sigma_\alpha^2 > 0$ .
3.  $H''_0 : \sigma_{\tau\alpha}^2 = 0 \quad vs. \quad H''_1 : \sigma_{\tau\alpha}^2 > 0$ .

TABLA DE ANÁLISIS DE LA VARIANZA.

F. Variación	g.l.	Suma de Cuadrados	Media Cuadrática	F	EMC
Factor A	$r - 1$	$SC_A = \sum_{i,j,k} (\bar{Y}_{i..} - \bar{Y}_{...})^2$	$MCT_A = SC_A / (r - 1)$	$F_A = \frac{MCT_A}{MCT_I}$	$\sigma^2 + n\sigma_{\tau\alpha}^2 + bn\sigma_\tau^2$
Factor B	$b - 1$	$SC_B = \sum_{i,j,k} (\bar{Y}_{.j.} - \bar{Y}_{...})^2$	$MCT_B = SC_B / (b - 1)$	$F_B = \frac{MCT_B}{MCT_I}$	$\sigma^2 + n\sigma_{\tau\alpha}^2 + rn\sigma_\alpha^2$
Interacción	$(r - 1)(b - 1)$	$SC_I = \sum_{i,j,k} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$	$MCT_I = \frac{SC_I}{(r-1)(b-1)}$	$F_I = \frac{MCT_I}{MCE}$	$\sigma^2 + n\sigma_{\tau\alpha}^2$
Error	$rb(n - 1)$	$SCE = \sum_{i,j,k} (Y_{ijk} - \bar{Y}_{ij.})^2$	$MCE = SCE / rb(n - 1)$		$\sigma^2$

OBSERVACIONES:

- Si  $H_0$  es cierto, entonces  $F_A \sim F(r - 1, (r - 1)(b - 1))$ .
- Si  $H'_0$  es cierto, entonces  $F_B \sim F(b - 1, (r - 1)(b - 1))$ .
- Si  $H''_0$  es cierto, entonces  $F_I \sim F((r - 1)(b - 1), rb(n - 1))$ .

**Diseño con dos factores completamente aleatorizado: Modelo Mixto.**

**Diseño con dos factores con interacción y  $n$  observaciones por celda.**

MODELO:

$$Y_{ijk} = \mu + \tau_i + \alpha_j + (\tau\alpha)_{ij} + \mathcal{E}_{ijk}, \quad i = 1, \dots, r; \quad j = 1, \dots, b \quad k = 1, \dots, n,$$

donde  $\mu$  y  $\tau_i$ ,  $i = 1, \dots, r$ , son constantes desconocidas y  $\alpha_j$ ,  $(\tau\alpha)_{ij}$  y  $\mathcal{E}_{ijk}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ ,  $k = 1, \dots, n$ , son variables aleatorias.

OBSERVACIONES:

- El modelo se dice de mixto pues  $\tau_i$ ,  $i = 1, \dots, r$ , son parámetros y  $\alpha_j$  y  $(\tau\alpha)_{ij}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ , son variables aleatorias.
- Asumiremos, en general, que:
  1.  $E[\mathcal{E}] = 0$  y  $E[\mathcal{E}\mathcal{E}^t] = \sigma^2 I_{rbn}$ , con  $\mathcal{E} = (\mathcal{E}_{111}, \dots, \mathcal{E}_{rbn})^t$
  2.  $\sum_{i=1}^r \tau_i = 0$ .
  3.  $E[\alpha] = 0$  y  $E[\alpha\alpha^t] = \sigma_\alpha^2 I_b$ , con  $\alpha = (\alpha_1, \dots, \alpha_b)^t$
  4.  $E[(\tau\alpha)_{ij}] = 0$ ,  $Var[(\tau\alpha)_{ij}] = \sigma_{\tau\alpha}^2 (r-1)/r$ ,  $Cov[(\tau\alpha)_{ij}, (\tau\alpha)_{kj}] = -\sigma_{\tau\alpha}^2/r$  (supondremos  $\sum_i (\tau\alpha)_{ij} = 0$ ,  $j = 1, \dots, b$ ) y  $Cov[(\tau\alpha)_{ij}, (\tau\alpha)_{kj'}] = 0$ ,  $j \neq j'$ .
  5. Las v.a.  $\alpha_j$ ,  $(\tau\alpha)_{ij}$  y  $\mathcal{E}_{ijk}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ ,  $k = 1, \dots, n$  son incorreladas.

Para determinadas cuestiones inferenciales supondremos, además, que:

1.  $\mathcal{E} \sim N_{rbn}(0, \sigma^2 I_{rbn})$ .
2.  $\alpha \sim N_b(0, \sigma_\alpha^2 I_b)$ .
3.  $(\tau\alpha)_{ij} \sim N(0, \frac{r-1}{r} \sigma_{\tau\alpha}^2)$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, b$ .

CONTRASTE DE LAS HIPÓTESIS:

1.  $H_0 : \tau_1 = \dots = \tau_r = 0$ .
2.  $H'_0 : \sigma_\alpha^2 = 0$  vs.  $H'_1 : \sigma_\alpha^2 > 0$ .
3.  $H''_0 : \sigma_{\tau\alpha}^2 = 0$  vs.  $H''_1 : \sigma_{\tau\alpha}^2 > 0$ .

TABLA DE ANÁLISIS DE LA VARIANZA.

F. Variación	g.l.	Suma de Cuadrados	Media Cuadrática	F	EMC
Factor A	$r - 1$	$SC_A = \sum_{i,j,k} (\bar{Y}_{i..} - \bar{Y}_{...})^2$	$MCT_A = SC_A / (r - 1)$	$F_A = \frac{MCT_A}{MCT_I}$	$\sigma^2 + n\sigma_{\tau\alpha}^2 + \frac{bn}{r-1} \sum_i \tau_i^2$
Factor B	$b - 1$	$SC_B = \sum_{i,j,k} (\bar{Y}_{.j.} - \bar{Y}_{...})^2$	$MCT_B = SC_B / (b - 1)$	$F_B = \frac{MCT_B}{MCE}$	$\sigma^2 + rn\sigma_\alpha^2$
Interacción	$(r - 1)(b - 1)$	$SC_I = \sum_{i,j,k} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$	$MCT_I = \frac{SC_I}{(r-1)(b-1)}$	$F_I = \frac{MCT_I}{MCE}$	$\sigma^2 + n\sigma_{\tau\alpha}^2$
Error	$rb(n - 1)$	$SCE = \sum_{i,j,k} (Y_{ijk} - \bar{Y}_{ij.})^2$	$MCE = SCE / rb(n - 1)$		$\sigma^2$

OBSERVACIONES:

- Si  $H_0$  es cierto, entonces  $F_A \sim F(r - 1, (r - 1)(b - 1))$ .
- Si  $H'_0$  es cierto, entonces  $F_B \sim F(b - 1, rb(n - 1))$ .
- Si  $H''_0$  es cierto, entonces  $F_I \sim F((r - 1)(b - 1), rb(n - 1))$ .