# Práctica 4: Experimentos con uno y dos factores

## 1ª parte

Modelos Lineales

Grados en Estadística y Matemáticas

# Índice

## Modelos de Diseño de Experimentos

Recordemos que en los modelos de Diseño de Experimentos aparecen dos elementos comunes:

- La **variable respuesta** ($Y$): la característica que puede ser observada, medida y analizada. Supondremos que es cuantitativa continua y con distribución normal.
- El **factor o factores**: un factor es una variable cuyo posible efecto sobre la respuesta se quiere estudiar; los niveles de un factor son los valores que puede tomar el factor. El factor será tratado como una variable cualitativa.

El objetivo del análisis estadístico que vamos a realizar a través de estos modelos es determinar la posible influencia del factor o factores en la variable respuesta $Y$. Sólo consideraremos modelos de efectos fijos.

# Diseño Completamente Aleatorizado

# Diseño Completamente Aleatorizado

## Toma de datos

La asignación de los $k$ 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.

Como resultado obtendremos $k$ muestras independientes, no necesariamente del mismo tamaño:

- Si todas las muestras tienen el mismo tamaño se dice que el diseño es balanceado
- Si las muestras tienen distinto tamaño se dice que el diseño es no balanceado

# Diseño Completamente Aleatorizado

## Ejercicio 1

Una operación de llenado tiene tres máquinas idénticas que se ajustan para vaciar una cantidad especifica de un producto en recipientes de igual tamaño. Con el propósito de identificar diferencias entre las cantidades (en litros) vaciadas por cada máquina, se toman muestras aleatorias, en forma periódica, de cada una. Para un periodo particular se observaron los datos que aparecen en el fichero `maquina.dat`.

Se trata de determinar si existen diferencias estadísticamente significativas en las cantidades promedio vaciadas por las tres máquinas.

# Diseño Completamente Aleatorizado

## Modelo matemático

El modelo matemático que se utiliza para analizar estos datos es el siguiente:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, \ldots, k \quad j = 1, \ldots, n_i \quad N = \sum_{i=1}^{k} n_i$$

donde

- $Y_{ij}$: representa el valor de $Y$ para la observación $j$ del nivel $i$ del factor
- $\mu$: es la media poblacional de $Y$ si el factor no estuviera presente.
- $\tau_i$: efecto que sobre la media poblacional de $Y$ tiene el nivel $i$ del factor.
- $\varepsilon_{ij}$: componente aleatoria en la medición de $Y_{ij}$ (no observable y con media 0).

Si denotamos $\boldsymbol{\mathcal{E}} = (\mathcal{E}_{11}, \ldots, \mathcal{E}_{kn_k})^t$, para aplicar toda la metodología vista en clase asumiremos:

$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \mathbf{I_N})$$

## Tabla ANOVA

El contraste de hipótesis más importante para este modelo es el siguiente:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_k$$
$$H_1 : \tau_i \neq \tau_j \text{ para algún par de niveles}(i,j)$$

Para resolverlo se utiliza la tabla Análisis de la Varianza vista en clase, donde se comparan las fuentes de variabilidad del experimento:

| F.V. | g.l. | Suma de cuadrados | M.C. | F |
|---|---|---|---|---|
| Debida a $H_0$ | $k-1$ | $Q_1 = \sum_{i=1}^{k} n_i \bar{Y}_{i.}^2 - N\bar{Y}_{..}^2$ | $\frac{Q_1}{k-1}$ | $F = \frac{(N-k)Q_1}{(k-1)Q_0}$ |
| Error | $N-k$ | $Q_0 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k} n_i \bar{Y}_{i.}^2$ | $\frac{Q_0}{N-k}$ | |

# Diseño Completamente Aleatorizado

## Comparaciones múltiples

Si aceptamos $H_1$, habremos detectado diferencias entre los tratamientos en cuanto al comportamiento de la variable respuesta $Y$. Para saber en qué tratamientos residen dichas diferencias hay que aplicar un procedimiento de comparaciones múltiples o comparaciones dos a dos donde se realizan, de modo simultáneo, todos los contrastes posibles.

$$H_0 : \tau_i = \tau_j$$
$$H_1 : \tau_i \neq \tau_j$$

Para resolverlos hay gran cantidad de métodos: LSD, Scheffé, Tukey,...

# Diseño de Bloques al Azar

# Diseño de Bloques al Azar

## Toma de datos

Supongamos que estamos estudiando la posible influencia que un factor (con $k$ niveles) puede ejercer sobre una variable cuantitativa continua $Y$.

- Además del factor, existe otra fuente de variabilidad en el experimento con influencia sobre el mismo.
- De acuerdo a esta fuente de variabilidad, se divide a las unidades experimentales en $s$ grupos denominados bloques.
- Dentro de cada bloque se elige, de forma aleatoria, una unidad experimental para cada uno de los niveles del factor.

Las observaciones aparecen agrupadas por bloques, obteniendo como resultado $k$ muestras relacionadas, una para cada nivel del factor.

# Diseño de Bloques al Azar

## Ejercicio 2

Un fabricante desea emplear un nuevo tipo de aleación en la producción de filtros y quiere determinar la velocidad de agitación adecuada ya que piensa que ésta puede influir en la resistencia que alcanza el producto final.

En el local de la fabrica se cuenta con 4 hornos, cada uno de los cuales tiene sus propias características de operación, lo que los convierte en posibles fuentes de variabilidad.

Cada horno puede operarse a las tres velocidades propuestas por un experto (10 rpm, 15 rpm y 20 rpm). Los datos de resistencia (en $Kg/cm^2$) se encuentran en el archivo `filtros.dat`

# Diseño de Bloques al Azar

## Ejercicio 2 (continuación)

Los datos podrían disponerse de este modo

| Velocidad de Agitación | H1 | H2 | H3 | H4 |
|---|---|---|---|---|
| 10 rpm | 8.3 | 9.4 | 8.5 | 7.5 |
| 15 rpm | 14.4 | 17.8 | 15.8 | 13.4 |
| 20 rpm | 11.1 | 13.2 | 11.8 | 10.1 |

Las unidades experimentales (aleación para fabricar un filtro) se han agrupado en bloques: los hornos donde se produce la aleación.

De esta forma, para cada tratamiento (velocidad de agitación) tenemos 4 datos de resistencia, uno correspondiente a cada horno. Son tres muestras relacionadas.

# Diseño de Bloques al Azar

## Modelo matemático

El modelo matemático que se utiliza para analizar estos datos es el siguiente:

$$Y_{ij} = \mu + \tau_i + \delta_j + \varepsilon_{ij} \quad i = 1, \ldots, r \quad j = 1, \ldots, s$$

donde

- $Y_{ij}$: representa el valor de $Y$ para el nivel $i$ y el bloque $j$.
- $\mu$: es la media poblacional de $Y$ sin ninguna fuente de variabilidad presente.
- $\tau_i$: efecto que sobre la media poblacional de $Y$ tiene el nivel $i$ del factor.
- $\delta_j$: efecto sobre la media poblacional de $Y$ por pertenecer al bloque $j$.
- $\varepsilon_{ij}$: componente aleatoria en la medición de $Y_{ij}$ (no observable y con media 0).

Si denotamos $\boldsymbol{\mathcal{E}} = (\mathcal{E}_{11}, \ldots, \mathcal{E}_{rs})^t$, para aplicar toda la metodología vista en clase asumiremos:

$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}_{rs}(\mathbf{0}, \sigma^2 \mathbf{I_{rs}})$$

## Tabla ANOVA

El contraste de hipótesis que se plantea es el siguiente:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_r$$

$$H_1 : \text{los efectos de los tratamientos no son iguales}$$

Para resolverlo se utiliza la tabla Análisis de la Varianza vista en clase, donde se comparan las fuentes de variabilidad del experimento:

| F.V. | g.l. | Suma de cuadrados | M.C. | F |
|---|---|---|---|---|
| Tr. ($\tau$) | $r-1$ | $Q_1 = \sum_{i=1}^{r} s\bar{Y}_{i\cdot}^2 - ks\bar{Y}_{\cdot\cdot}^2$ | $\frac{Q_1}{r-1}$ | $F_T = \frac{(s-1)Q_1}{Q_0}$ |
| Bl. ($\delta$) | $s-1$ | $Q_1' = \sum_{j=1}^{s} r\bar{Y}_{\cdot j}^2 - rs\bar{Y}_{\cdot\cdot}^2$ | $\frac{Q_1'}{s-1}$ | $F_B = \frac{(r-1)Q_1'}{Q_0}$ |
| Error | $(r-1)(s-1)$ | $Q_0 = \sum_{i=1}^{r}\sum_{j=1}^{s} Y_{ij}^2 - \sum_{i=1}^{r} s\bar{Y}_{i\cdot}^2 - \sum_{j=1}^{s} r\bar{Y}_{\cdot j}^2 + rs\bar{Y}_{\cdot\cdot}^2$ | $\frac{Q_0}{(r-1)(s-1)}$ | |

# Diseño de Bloques al Azar

## Comparaciones múltiples

Si aceptamos $H_1$, habremos detectado diferencias entre los tratamientos en cuanto al comportamiento de la variable $Y$. Para saber en qué tratamientos residen dichas diferencias hay que aplicar un procedimiento de comparaciones múltiples o comparaciones dos a dos, donde se realizan, de modo simultáneo, todos los contrastes posibles

$$H_0 : \tau_i = \tau_j$$
$$H_1 : \tau_i \neq \tau_j$$

Para resolverlos hay gran cantidad de métodos: LSD, Scheffé, Tukey,...

# Diseño de Cuadrados Latinos

# Diseño de Cuadrados Latinos

## Toma de datos

Supongamos que estamos estudiando la posible influencia que un factor (con $k$ niveles) puede ejercer sobre una variable cuantitativa continua $Y$.

- Además del factor, existen otras dos fuentes de variabilidad en el experimento con influencia sobre el mismo.
- Si dividimos la población en bloques de acuerdo a ambas fuentes de variabilidad, el número de observaciones necesarias para abarcar todas las combinaciones posibles podría ser demasiado grande.
- Cada fuente de variabilidad se divide en $k$ niveles con los cuales se forma un cuadrado $k \times k$.
- Cada tratamiento del factor aparece una sola vez en cada fila y cada columna.

# Diseño de Cuadrados Latinos

## Ejercicio 3

Un investigador quiere evaluar la productividad de cuatro variedades de aguacate, Reed, Hass, Pinkerton y Bacon. Para ello decide realizar el ensayo en un terreno que posee un gradiente de pendiente de oriente a occidente y además, diferencias en la disponibilidad de Nitrógeno de norte a sur.

Para controlar los efectos de la pendiente y la disponibilidad de Nitrógeno, estableció 4 zonas de acuerdo a la pendiente y otras 4 en base a la disponibilidad de Nitrógeno, fijando así 16 parcelas. Posteriormente realizó un plantación de cada variedad en cada una de las pendientes y para cada una de las diferentes concentraciones de Nitrógeno.

Los datos correspondientes a la producción (en kg/parcela) se encuentran en el archivo `aguacate.txt`.

# Diseño de Cuadrados Latinos

## Ejercicio 3 (continuación)

Veamos como se disponen los datos en el cuadrado latino que ha sido elegido por el experimentador.

| | | Pendiente | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| Nitrógeno | 1 | Reed<br>785 | Hass<br>730 | Pinkerton<br>700 | Bacon<br>595 |
| | 2 | Hass<br>855 | Bacon<br>775 | Reed<br>760 | Pinkerton<br>710 |
| | 3 | Pinkerton<br>950 | Reed<br>885 | Bacon<br>795 | Hass<br>780 |
| | 4 | Bacon<br>945 | Pinkerton<br>950 | Hass<br>880 | Reed<br>835 |

# Diseño de Cuadrados Latinos

## Modelo matemático

El modelo matemático que se utiliza para analizar estos datos es el siguiente:

$$Y_{ijh} = \mu + \tau_i + \delta_j + \gamma_h + \varepsilon_{ijh}\ , \quad i,j = 1,\ldots,k$$

y $h$ varía según marca el cuadrado latino elegido. En este modelo:

- $Y_{ijh}$: representa el valor de $Y$ en el nivel $i$, fila $j$, columna $h$.
- $\mu$: es la media poblacional de $Y$ sin ninguna fuente de variabilidad presente.
- $\tau_i$: efecto que sobre la media poblacional de $Y$ tiene el nivel $i$ del factor.
- $\delta_j$: efecto sobre la media poblacional de $Y$ por estar en la fila $j$.
- $\gamma_h$: efecto sobre la media poblacional de $Y$ por estar en la columna $h$.
- $\varepsilon_{ijh}$: componente aleatoria en la medición de $Y_{ijh}$ (no observable y con media 0).

Si denotamos $\boldsymbol{\mathcal{E}} = (\mathcal{E}_{11},\ldots,\mathcal{E}_{k^2})^t$, para aplicar toda la metodología vista en clase asumiremos:

$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}_{k^2}(\mathbf{0}, \sigma^2 \mathbf{I}_{\mathbf{k^2}})$$

## Tabla ANOVA

El contraste de hipótesis que se plantea es el siguiente:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_k$$

$$H_1 : \text{los efectos de los tratamientos no son iguales}$$

Tabla de Análisis de la Varianza

| F.V. | g.l. | Suma de cuadrados | M.C. | F |
|---|---|---|---|---|
| Tr. ($\tau$) | $k-1$ | $Q_1 = \sum_{i=1}^{k} k\bar{Y}_{i\cdot\cdot}^2 - k^2\bar{Y}_{\cdots}^2$ | $\frac{Q_1}{k-1}$ | $F_T = \frac{(k-2)Q_1}{Q_0}$ |
| Fil. ($\delta$) | $k-1$ | $Q_1' = \sum_{j=1}^{k} k\bar{Y}_{\cdot j\cdot}^2 - k^2\bar{Y}_{\cdots}^2$ | $\frac{Q_1'}{k-1}$ | $F_A = \frac{(k-2)Q_1'}{Q_0}$ |
| Col. ($\gamma$) | $k-1$ | $Q_1'' = \sum_{j=1}^{k} k\bar{Y}_{\cdot\cdot j}^2 - k^2\bar{Y}_{\cdots}^2$ | $\frac{Q_1''}{k-1}$ | $F_B = \frac{(k-2)Q_1''}{Q_0}$ |
| Error | $(k-2)(k-1)$ | $Q_0 = \sum_{i,j,h=1}^{k} Y_{ijh}^2 - \sum_{i=1}^{k} k\bar{Y}_{i\cdot\cdot}^2 - \sum_{j=1}^{s} k\bar{Y}_{\cdot j\cdot}^2 - \sum_{j=1}^{s} k\bar{Y}_{\cdot\cdot j}^2 + 2k^2\bar{Y}_{\cdots}^2$ | $\frac{Q_0}{(k-2)(k-1)}$ | |

# Diseño de Cuadrados Latinos

## Hipótesis teóricas

Para utilizar este modelo hemos de asumir normalidad, homocedasticidad y no autocorrelación, que se comprobarán a través del análisis de los residuos.

## Comparaciones múltiples

Si aceptamos $H_1$, habremos detectado diferencias entre los tratamientos en cuanto al comportamiento de la variable $Y$. Para saber en qué tratamientos residen dichas diferencias hay que aplicar un procedimiento de comparaciones múltiples o comparaciones dos a dos, es decir realizaremos de modo simultáneo, todos los contrastes posibles

$$H_0 : \tau_i = \tau_j$$
$$H_1 : \tau_i \neq \tau_j$$

.

# Otros ejercicios

## Ejercicio 4

Desde el incremento en los precios de la gasolina, se han desarrollado varios dispositivos que se colocan en los carburadores de los automóviles. Una empresa selecciona tres de estos dispositivos para someterlos a prueba. La empresa desea compararlos con los carburadores estándar, con el propósito de determinar si existe un incremento apreciable de kilómetros por litro de gasolina con el uso de estos dispositivos.

La compañía selecciona cinco automóviles para el experimento. Para controlar la variación, se planea utilizar el mismo conductor para todo el experimento. Se observan los datos del archivo carbu.dat.

Estudia las posibles diferencias de consumo entre los carburadores.

## Ejercicio 5

El conjunto de datos `iris` proporciona las medidas en centímetros de las variable *longitud del sépalo*, *anchura del sépalo*, *longitud del pétalo* y *anchura del pétalo*, respectivamente, para 50 flores de tres especies of iris: setosa, versicolor y virginica.

Se trata de detectar si hay diferencias significativas entre las tres especies en cada una de las cuatro medidas. Los datos se encuentran en el archivo `iris.dat`

## Ejercicio 6

Una empresa de bebidas está interesada en evaluar el impacto de diferentes métodos de carbonatación en la efervescencia de sus refrescos. Se han identificado cuatro métodos de carbonatación que se aplicarán a un lote de refrescos:

- A: Carbonatación por inyección de $CO_2$
- B: Carbonatación natural (fermentación)
- C: Carbonatación por absorción
- D: Carbonatación por mezcla de gases

El objetivo del estudio es determinar si hay diferencias significativas en la efervescencia de los refrescos, medida a través de la cantidad de dióxido de carbono ($CO_2$) disuelto, al aplicar los diferentes métodos de carbonatación.
Para controlar la variabilidad entre los diferentes lotes de refrescos y el tiempo de almacenamiento, cada método de carbonatación se aplicará una vez para cada lote de refresco y en cada tiempo de almacenamiento. Los datos se encuentran en el archivo `carbonatacion.txt`. Analiza estos datos para determinar las posibles diferencias en contenido de dióxido de carbono de refrescos carbonatados con los 4 métodos.