Análisis de la varianza (ANOVA)

Mª Isabel Aguilar, Eugenia Cruces y Bárbara Díaz

UNIVERSIDAD DE MÁLAGA
Departamento de Economía Aplicada (Estadística y Econometría)
Parcialmente financiado a través del PIE13-024 (UMA)

Análisis de la Varianza (ANOVA)

UNIVERSIDAD DE MÁLAGA

- **Introducción**
- **ANOVA de un factor**
- **Hipótesis básicas**
- **ANOVA multifactorial**
- **Diseño por bloques**
- **ANOVA no paramétrico**

UNIVERSIDAD DE MÁLAGA

Introducción

Introducción

- Técnica multivariante desarrollada por Fisher en la década de los 20 y 30. Inicialmente se aplicó para analizar el efecto de distintos niveles de fertilizante sobre las cosechas

- El **ANOVA** estudia la influencia de uno o más **factores** (variables categóricas) sobre una **variable de respuesta** (variable continua)

- El factor se presenta en diferentes **tratamientos o niveles** (ej. el factor "abono" tiene tres tratamientos: A, B y C)

Introducción

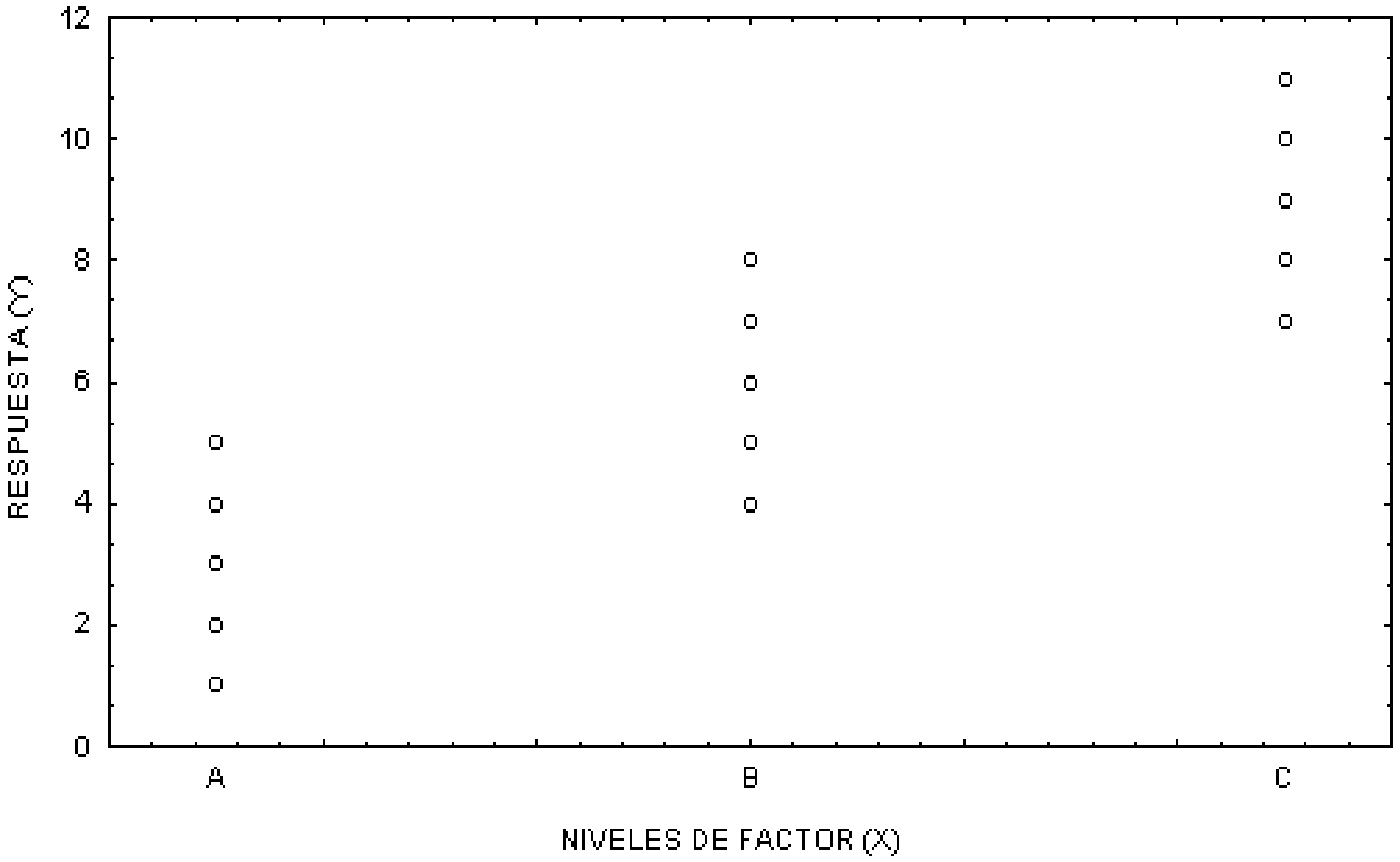
- Generalización del contraste de la diferencia de medias en dos poblaciones normales a k poblaciones.
- Considera el efecto de factores no controlables por el experimento, intentando aislar los errores atribuibles al mismo.
- Puede enfocarse también como un caso especial de regresión: análisis de la influencia de una serie de variables explicativas (factores) sobre una variable dependiente (variable respuesta).
- **ANOVA unifactorial** (un único factor: por ejemplo, abono) y **multifactorial** (varios factores: abono, riego,…).

Introducción

Ejemplo:

- Se tienen 15 parcelas en las que se prueba el efecto de tres niveles de abono en la cosecha (A, B y C).
- Se asignan de forma aleatoria 5 parcelas a cada nivel de abono.
- En el siguiente gráfico, puede observarse que la respuesta media es diferente para cada nivel de factor. En concreto, se registra una mayor cosecha media para el nivel o tratamiento C que para el A o el B.

Introducción



Análisis de la varianza

Unidades experimentales (n)	• • • • • • • • • • (n_1)	• • • • • • • • • • • • • • • • (n_2)	…	• • • • • • • • • • • • (n_k)
Tratamientos (var. explicativa, X)	X_1 Nivel 1	X_2 Nivel 2	…	X_k Nivel k
Variable Respuesta (Y) Hipótesis sobre Y	Y_1 $N(\mu_1, \sigma)$	Y_2 $N(\mu_2, \sigma)$	…	Y_k $N(\mu_k, \sigma)$
Respuesta observada en la muestra	y_{11} y_{21} .. y_{i1} $y_{n_1 1}$	y_{12} y_{22} .. y_{i2} $y_{n_2 2}$	…	y_{1k} y_{2k} .. y_{ik} $y_{n_k k}$
Totales muestrales	$T_{.1}$	$T_{.2}$	…	$T_{.k}$
Medias muestrales	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	…	$\bar{Y}_{.k}$

Introducción

- En ANOVA se distingue entre el **modelo de efectos fijos** y el **modelo de efectos aleatorios**.
  - **Modelo de efectos fijos:** los niveles del factor (o factores) están fijados de antemano. Las conclusiones únicamente pueden aplicarse a los niveles estudiados.
  - **Modelo de efectos aleatorios:** los niveles del factor (o factores) se extraen de forma aleatoria de un conjunto más amplio de niveles. Los resultados son válidos para el conjunto de niveles que se tuvo en cuenta en el diseño inicial.

Estudiaremos tan sólo el modelo de efectos fijos.

Introducción

- **Tipos de diseños del experimento:**
  - **Diseño completamente aleatorizado:** las unidades experimentales son homogéneas y la asignación de los distintos tratamientos se hace de forma aleatoria.
  - **Diseño en bloques completamente aleatorizado:** no todas las unidades experimentales son homogéneas. Se forman grupos homogéneos y se les asignan a cada uno de ellos todos los tratamientos.

UNIVERSIDAD DE MÁLAGA

ANOVA de un factor

ANOVA de un factor

- El **ANOVA de un factor completamente aleatorizado** permite decidir si los distintos niveles de factor (variable explicativa) producen diferentes efectos en la variable de respuesta o, por el contrario, el comportamiento de ésta es el mismo para todos los niveles.
- Se trata de comprobar si existen diferencias significativas en la respuesta media para los distintos niveles del factor

ANOVA de un factor

- La hipótesis nula establece que las medias poblaciones son iguales, y, por tanto, igual a la media global, frente a la alternativa de que al menos una es diferente:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_j = \cdots \mu_k = \mu$$
$$H_1: \textbf{no todas las } \mu_j \textbf{ son iguales}$$

  - Si se acepta H_0 la respuesta no se ve afectada por los distintos niveles de factor. Las diferencias se deben al azar, al error experimental.
  - Si se rechaza H_0, se pueden distinguir los efectos que producen los distintos tratamientos.

ANOVA de un factor

- El **objetivo** es analizar si los k niveles del factor X influyen de igual manera en la variable respuesta Y.
- Para ello el modelo se basa en **descomponer la variabilidad de la respuesta** en dos partes:
  - La originada por el factor objeto de estudio.
  - La producida por el error experimental.

ANOVA de un factor

- y_{ij}: Respuesta de una unidad experimental *i* ante el tratamiento *j*.

Es la suma de la respuesta media del grupo de unidades experimentales sometidas al tratamiento *j* (μ_j) y el error experimental o efecto aleatorio asociado (ε_{ij}):

$$y_{ij} = \mu_j + \varepsilon_{ij} \quad \longrightarrow \quad \varepsilon_{ij} = y_{ij} - \mu_j$$

- τ_j: Efecto diferencial del j-ésimo tratamiento en relación al efecto medio global μ:

$$\tau_j = \mu_j - \mu \quad \longrightarrow \quad \mu_j = \mu + \tau_j$$

- Modelo final:

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

ANOVA de un factor

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

- Dado que $\tau_j = \mu_j - \mu$ y $\varepsilon_{ij} = y_{ij} - \mu_j$, la expresión anterior puede escribirse:

$$y_{ij} = \mu + (\mu_j - \mu) + (y_{ij} - \mu_j)$$

- Reordenando, la diferencia respecto a la media global de la observación de la unidad experimental *i* sometida al tratamiento *j*, se descompone en dos sumandos:

$$(y_{ij} - \mu) = (\mu_j - \mu) + (y_{ij} - \mu_j)$$

ANOVA de un factor

- $(y_{ij} - \mu)$: **Diferencia respecto a la media global** de la observación de la unidad experimental *i* sometida al tratamiento *j*. Se descompone en dos partes:
  - $(\mu_j - \mu)$: **Diferencia debida al tratamiento**. Es la diferencia entre la media del grupo de unidades experimentales sometidas al tratamiento *j* y la media global del experimento.
  - $(y_{ij} - \mu_j)$: **Diferencia aleatoria.** Es la diferencia entre la respuesta de la unidad experimental *i* sometida al tratamiento *j* y la media del grupo sometido al tratamiento *j*. Por tanto, no viene explicada por el factor y recoge el error experimental.

ANOVA de un factor

- Al trabajar con las medias muestrales como estimadores de las poblacionales, la descomposición anterior queda:

$$\left(y_{ij} - \bar{Y}\right) = \left(\bar{Y}_{.j} - \bar{Y}\right) + \left(y_{ij} - \bar{Y}_{.j}\right)$$

  - $\left(y_{ij} - \bar{Y}\right)$: Refleja la desviación de la respuesta de la unidad experimental *i* sometida al tratamiento *j* respecto a la media muestral global.
  - $\left(\bar{Y}_{.j} - \bar{Y}\right)$: Refleja la desviación de la media muestral del tratamiento *j* respecto a la media muestral global.
  - $\left(y_{ij} - \bar{Y}_{.j}\right)$: Refleja la desviación no explicada por el tratamiento *j* (se denomina desviación residual).

ANOVA de un factor

- Elevando al cuadrado ambos miembros de la igualdad y sumando para todos los individuos de cada grupo de la muestra, se llega a la ecuación fundamental del ANOVA :

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\overline{Y})^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{Y}_{.j}-\overline{Y})^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\overline{Y}_{.j})^2$$

- De forma abreviada, esto significa que las variaciones totales en la respuesta **(SCT: suma de cuadrados total)** se explican por los efectos de las variaciones inducidas por los distintos tratamientos **(SCTR: suma de cuadrados de los tratamientos),** más una componente residual que recoge las variaciones debidas al error experimental **(SCE: suma de cuadrado de los errores)**.

$$\mathrm{SCT}=\mathrm{SCTR}+\mathrm{SCE}$$

ANOVA de un factor

- No es excesivamente complicado demostrar que:

$$\frac{SCTR}{\sigma^2} \sim \chi^2_{k-1} \qquad \frac{SCE}{\sigma^2} \sim \chi^2_{n-k}$$

- Además, SCTR y SCE son independientes, por lo que:

$$\frac{(SCTR/\sigma^2)/k-1}{(SCE/\sigma^2)/n-k} = \frac{CMTR}{CME} \rightarrow \frac{\chi^2_{k-1}/k-1}{\chi^2_{n-k}/n-k} \sim F_{k-1,n-k}$$

CMTR: cuadrados medios de los tratamientos

CME: cuadrados medios de los errores

ANOVA de un factor

Resumen del ANÁLISIS DE LA VARIANZA de un factor

Hipótesis nula y alternativa del contraste

H_0: $\mu_1 = \mu_2 = \cdots = \mu_j = \cdots \mu_k = \mu$

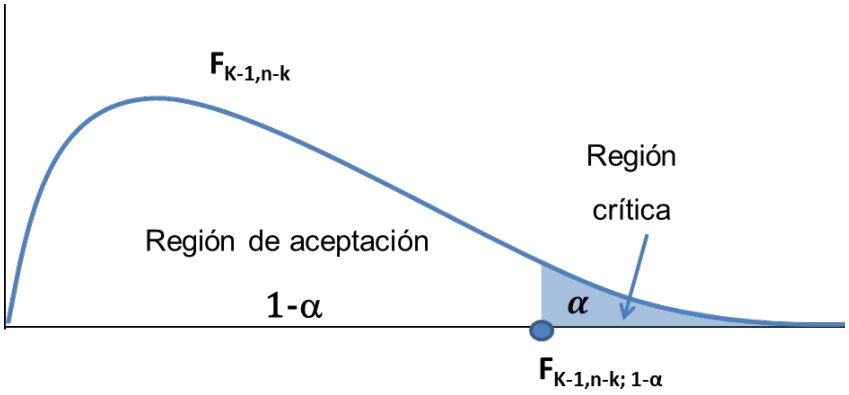
H_1: *no todas las* μ_j *son iguales*

Estadístico de Prueba

$$\frac{CMTR}{CME} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_{.j} - \bar{Y})^2 / k-1}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{Y}_{.j})^2 / n-k} \sim F_{k-1,n-k}$$

Región Crítica

La región crítica viene dada por la cola derecha de la distribución F (valores altos de *CMTR* frente a los de *CME* llevan a rechazar H_0)



ANOVA de un factor

Tabla ANOVA				
Fuente de variación	**Grados de libertad**	**Suma de cuadrados**	**Cuadrados medios**	**F observada**
- Tratamientos (Entre grupos)	k-1	SCTR $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_{.j}-\bar{Y})^2$	SCTR/k-1	$F_{obs}=\frac{CMTR}{CME}$
- Error (Dentro grupos)	n-k	SCE $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{Y}_{.j})^2$	SCE/n-k	
Total	n-1	SCT $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{Y}_{..})^2$	$\Rightarrow \eta^2 = R^2 = (SCTR/STC)$	

El significado de η^2 (eta cuadrado), similar al de R^2 en el análisis de regresión, da el porcentaje de variabilidad en la respuesta que viene explicado por el factor o variable explicativa

ANOVA de un factor: comparaciones múltiples

- Si se rechaza la H_0 de igualdad de medias, han de localizarse el/los tratamientos con medias diferentes en la variable de respuesta.
- Para ello existen distintos contrastes que realizan comparaciones de medias dos a dos.
  - LSD (Mínima Diferencia Significativa) de Fisher.
  - Bonferroni
  - Tuckey
  - Scheffé

ANOVA de un factor: comparaciones múltiples

- **LSD:** No garantiza el mantenimiento del nivel de significación para el conjunto del experimento.

- **Bonferroni:** Garantiza el nivel de significación para el conjunto del experimento pero no debe haber muchos tratamientos distintos. Si los hubiera, el nivel de significación individual de cada contraste sería muy bajo por lo que rechazar la igualdad de las dos medias se hace bastante más difícil que en las comparaciones individuales del método LSD.

ANOVA de un factor: comparaciones múltiples

- **Tuckey:** Es más conservador que el test LSD pero necesita que el experimento sea equilibrado (mismo número de unidades experimentales en cada tratamiento).
- **Scheffé:** Es el más conservador de todos. Garantiza el del nivel de significación para el conjunto del experimento, en todos los posibles contrastes que se puedan realizar entre medias, no sólo por parejas, sino también entre subconjuntos de medias. Más indicado para contrastes planificados, diseñados *ex-ante* por el investigador.

Hipótesis básicas

Hipótesis básicas

- **Normalidad.** Necesario para construir el estadístico de prueba que se distribuye como una *F*. En cualquier caso, el test es bastante robusto a la violación de esta hipótesis de normalidad.
- **Homoscedasticidad.** Necesario para rechazar hipótesis nula por diferencias de medias entre grupos y no por diferenicias de variabilidad.
- **Aleatoriedad.** Necesario para todo el proceso de inferencia.

Efecto muy importante de los **outliers** y de la **heteroscedasticidad** (variancias muy diferentes entre los grupos). El test basado en la F tiende a tomar valores grandes y rechazar la Hipótesis nula.

Hipótesis básicas

UNIVERSIDAD DE MÁLAGA

Soluciones:

- Transformar la variable para evitar relación entre la desviación típica y la media.
- Uso de un test no paramétrico de localización de varias poblaciones (Kruskall-Wallis).

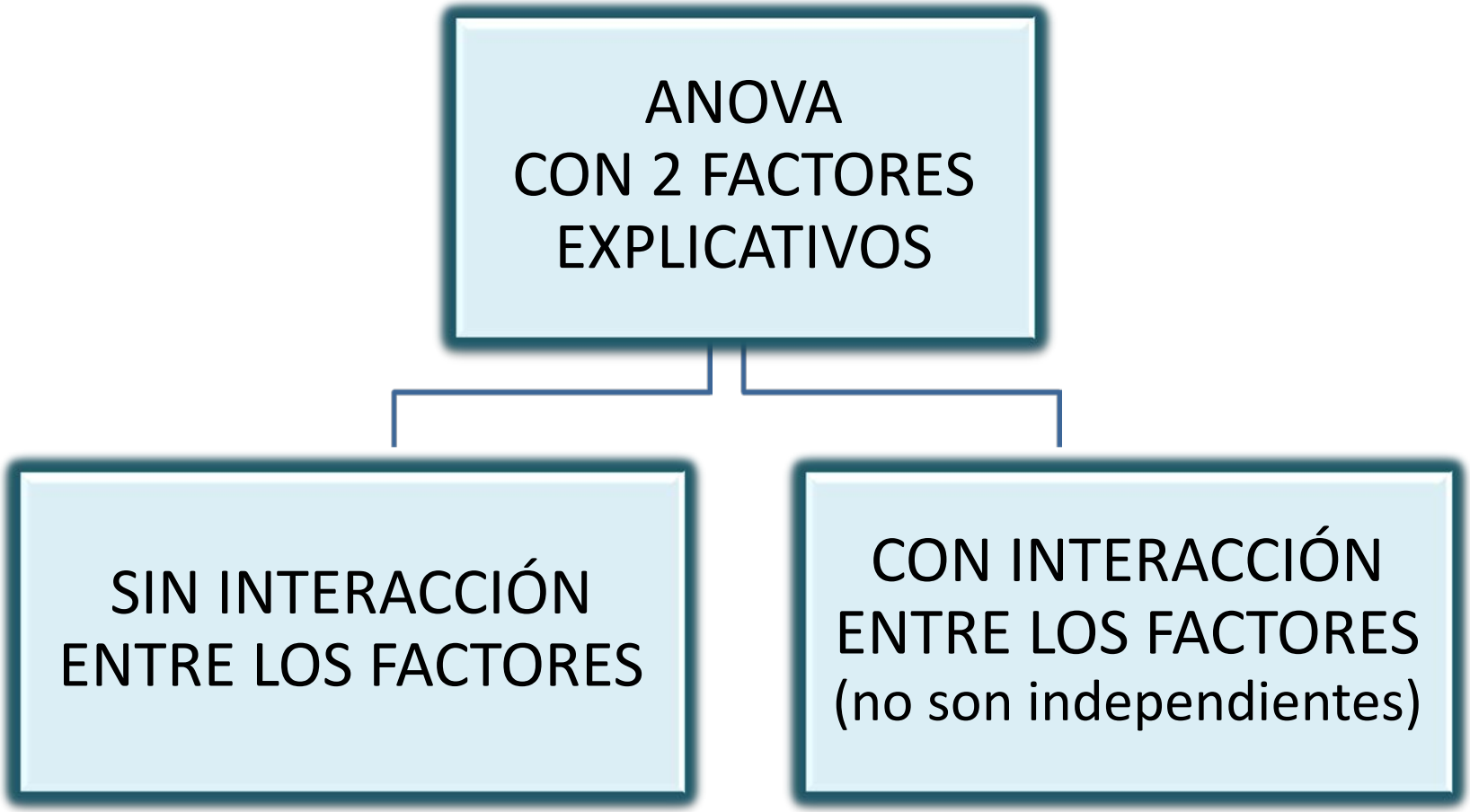
UNIVERSIDAD DE MÁLAGA

ANOVA multifactorial

ANOVA multifactorial

- **ANOVA multifactorial:** Análisis de la Varianza donde se analiza la influencia de más de un factor o variable explicativa.
- Para simplificar, se parte de una variable respuesta *Y*, y dos factores (F_1 *y* F_2), aunque el test se generaliza fácilmente para incluir más de dos factores.
- Y_{ijk} hace referencia a la observación que ocupa el lugar *k*, con un nivel *i* del factor 1 y *j* del factor 2.

ANOVA multifactorial



ANOVA de 2 factores

- La **interacción entre factores** se produce cuando el efecto de un factor está influenciado por los niveles de otros factores (no son independientes). Este efecto puede estudiarse de forma aislada.
- Procedimientos de **comparaciones múltiples**:
  - Si en el modelo la interacción es significativa, no pueden realizarse pruebas de comparación múltiple, ya que el comportamiento de cada uno de los factores está distorsionado por el otro.
  - Si no existe interacción las pruebas de comparaciones múltiples estudiadas para el ANOVA simple siguen siendo válidas.

ANOVA de 2 factores sin interacción

UNIVERSIDAD DE MÁLAGA

- **Especificación del modelo:**

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

- Donde:
  - α_i es el efecto diferencial del tratamiento *i* del factor 1
  - β_j es el efecto diferencial del tratamiento *j* del factor 2
  - ε_{ijk} es el residuo no explicado por los factores, debido al azar

- La descomposición de la suma de cuadrados sería:

$$SCT = SCTR1 + SCTR2 + SCE$$

ANOVA de 2 factores con interacción

UNIVERSIDAD DE MÁLAGA

- **Especificación del modelo:**

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

- Donde:
  - α_i es el efecto diferencial del tratamiento *i* del factor 1
  - β_j es el efecto diferencial del tratamiento *j* del factor 2
  - $(\alpha\beta)_{ij}$ es el efecto de interacción entre el nivel i del factor 1 y el nivel j del factor 2
  - ε_{ijk} es el residuo no explicado por los factores, debido al azar
- La descomposición de la suma de cuadrados sería:

$$SCT = SCTR1 + SCTR2 + SCInt + SCE$$

UNIVERSIDAD DE MÁLAGA

Diseño por bloques

Diseño por bloques

- A veces, **las unidades experimentales no son homogéneas** (en el ejemplo de las parcelas, pueden existir diferencias en función del tipo de terreno, la localización, el tamaño, etc.).
- El **diseño en bloques aleatorios** es apropiado y eficiente en estas situaciones porque elimina las diferencias iniciales entre las unidades experimentales, permitiendo verificar las diferencias entre los promedios de *k* tratamientos en condiciones homogéneas.
- Este diseño adjudica todos los tratamientos a cada uno de los bloques de unidades experimentales homogéneas. Así, se obtiene un efecto real de los tratamientos, no distorsionado o encubierto por la variación entre bloques de diferentes características.

Diseño por bloques

Bloques	Niveles de factor				
	1	2		k	$T_{i.}$
1	y_{11}	y_{12}	...	y_{1k}	$T_{1.}$
2	y_{21}	y_{22}	...	y_{2k}	$T_{2.}$
...	...	...	...	...	...
...	...	...	...	...	...
n	y_{n1}	y_{n2}	...	y_{nk}	$T_{n.}$
$T_{.j}$	$T_{.1}$	$T_{.2}$	...	$T_{.k}$	T
$\bar{Y}_{.j}$	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	...	$\bar{Y}_{.k}$	$\bar{Y}$

$$SCT = SCB + SCTR + SCE$$

Diseño por bloques

UNIVERSIDAD DE MÁLAGA

Tabla ANOVA para un diseño en bloques				
Fuente de variación	**Grados de libertad**	**Suma de cuadrados**	**Cuadrados medios**	**F observada**
Bloques	n-1	SCB= $\sum_{i=1}^{n}\sum_{j=1}^{k}(\bar{Y}_{i.}-\bar{Y})^2$		
Tratamientos	k-1	SCTR= $\sum_{i=1}^{n}\sum_{j=1}^{k}(\bar{Y}_{.j}-\bar{Y})^2$	CMTR=SCTR/k-1	$F_{obs}=\frac{CMTR}{CME}$
Error	(n-1)(k-1)	SCE= $\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij}-\bar{Y}_{i.}-\bar{Y}_{.j}+\bar{Y})^2$	CME=SCE/(n-1)(k-1)	
Total	nk-1	STC= $\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij}-\bar{Y})^2$	$\eta^2=R^2=(SCTR/STC)$ $\eta^2_{ajustado}=R^2_{ajustado}=(CMTR/CMT)$	

Diseño por bloques

- En el estudio de este modelo debe de tenerse en cuenta que nunca existe interacción entre el factor-tratamiento y el factor-bloque.
- La SCB (suma de cuadrados del bloque) se usa con el fin de aislar el efecto provocado por la falta de homogeneidad de las unidades experimentales.
- El tratamiento estadístico para el modelo de diseño de experimentos con un factor tratamiento y un factor bloque es exactamente igual que el diseño de experimentos con dos factores tratamiento sin interacción. Sin embargo, su planteamiento y objetivos son diferentes.

UNIVERSIDAD DE MÁLAGA

ANOVA no paramétrico

ANOVA no paramétrico

- Test no paramétrico de localización (mediana) para k muestras aleatorias independientes.

$$H_0: Me_1 = Me_2 = \cdots = Me_j = \cdots Me_k = Me$$
$$H_1: no\ todas\ las\ Me_j\ son\ iguales$$

- **Test de Kruskal-Wallis**. Diseños de un factor completamente aleatorizado. Aplicable a variables cuantitativas y a ordinales.
- Generalización del test de suma de rangos de Wilcoxon de comparación de medianas de dos poblaciones con muestras independientes.