Análisis de la varianza (ANOVA)

Miguel Ángel Tarancón Morán & Consolación Quintana Rojo

Área de Estadística Económica y Empresarial

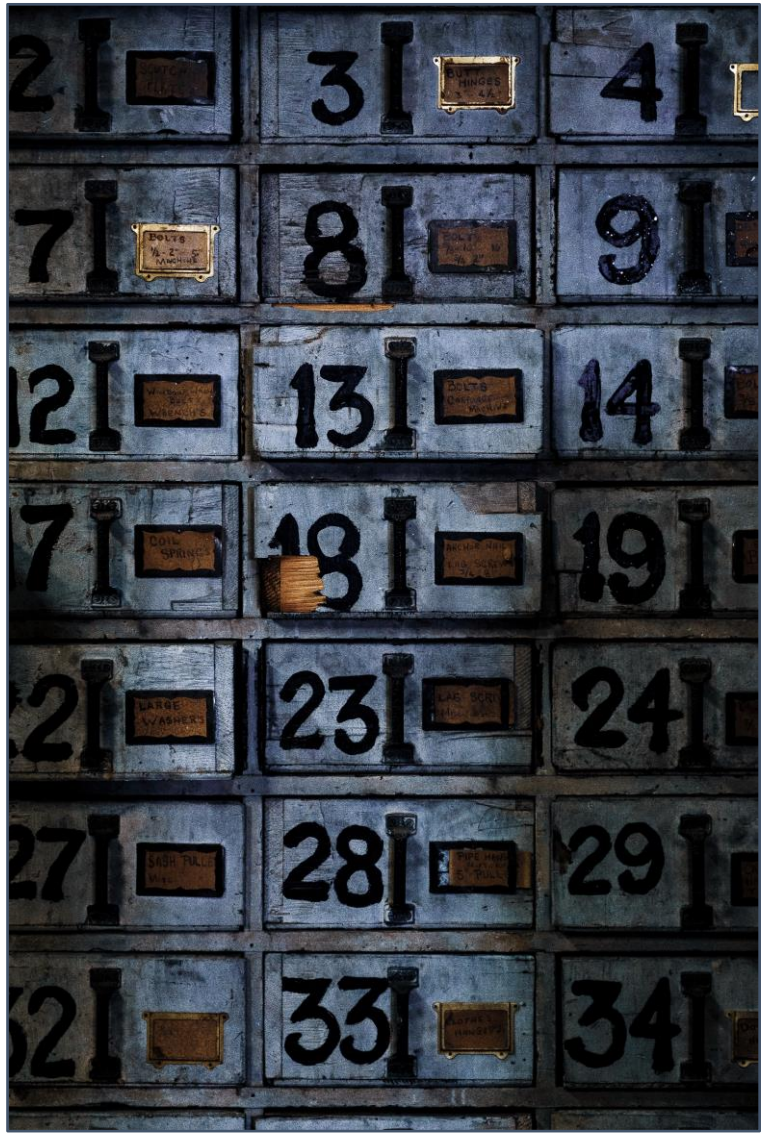
Departamento de Economía Aplicada 1

Universidad de Castilla – La Mancha

Contenidos.

1. Introducción.
2. ANOVA de un factor.
3. ANOVA multifactorial.
4. Análisis multivariante de la varianza.

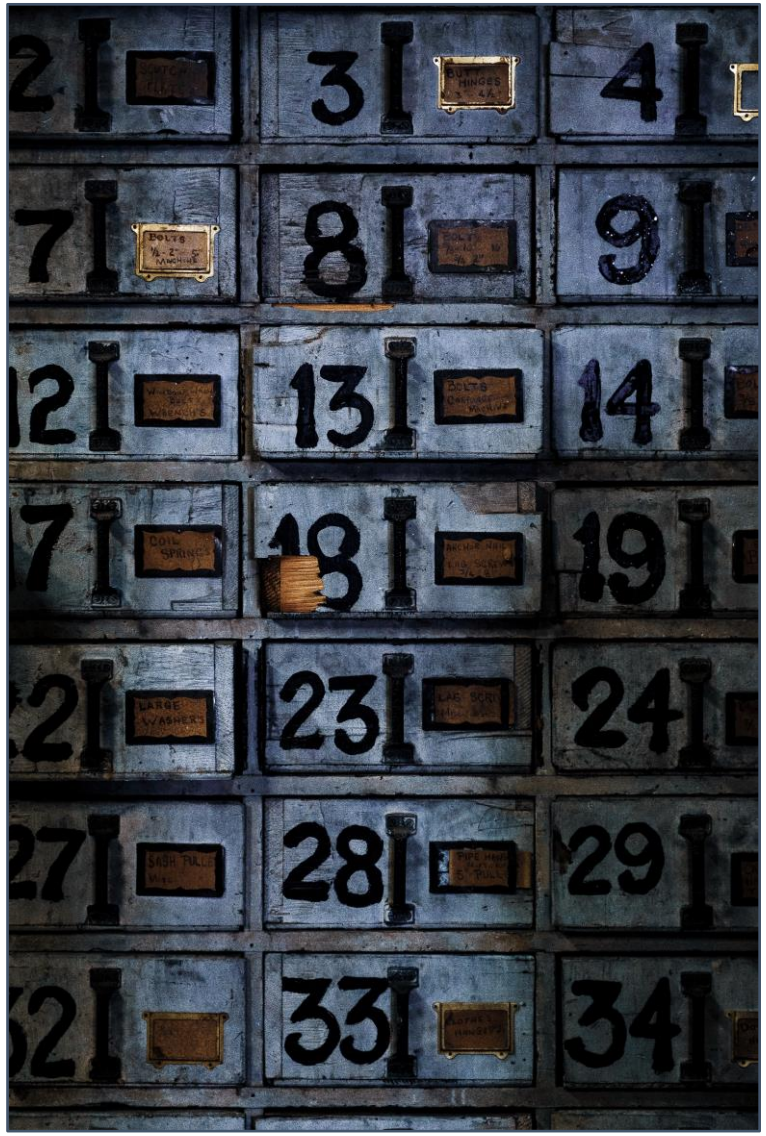
Introducción.



- En el análisis de la varianza (**ANOVA**) existen una o varias variables **dependientes cuantitativas** (escala métrica) cuyo **valor medio** se hace depender de los **niveles o categorías** que presentan **uno o varios atributos** (variables independientes cualitativas, categóricas, atributos o factores).

- El análisis ANOVA es una **generalización** del contraste de hipótesis sobre la diferencia de medias nula para 2 poblaciones normales con varianza poblacional desconocida.

ANOVA de un factor.



- **Hipótesis nula: las medias** de la variable dependiente para cada nivel o categoría del factor considerado **son iguales**.

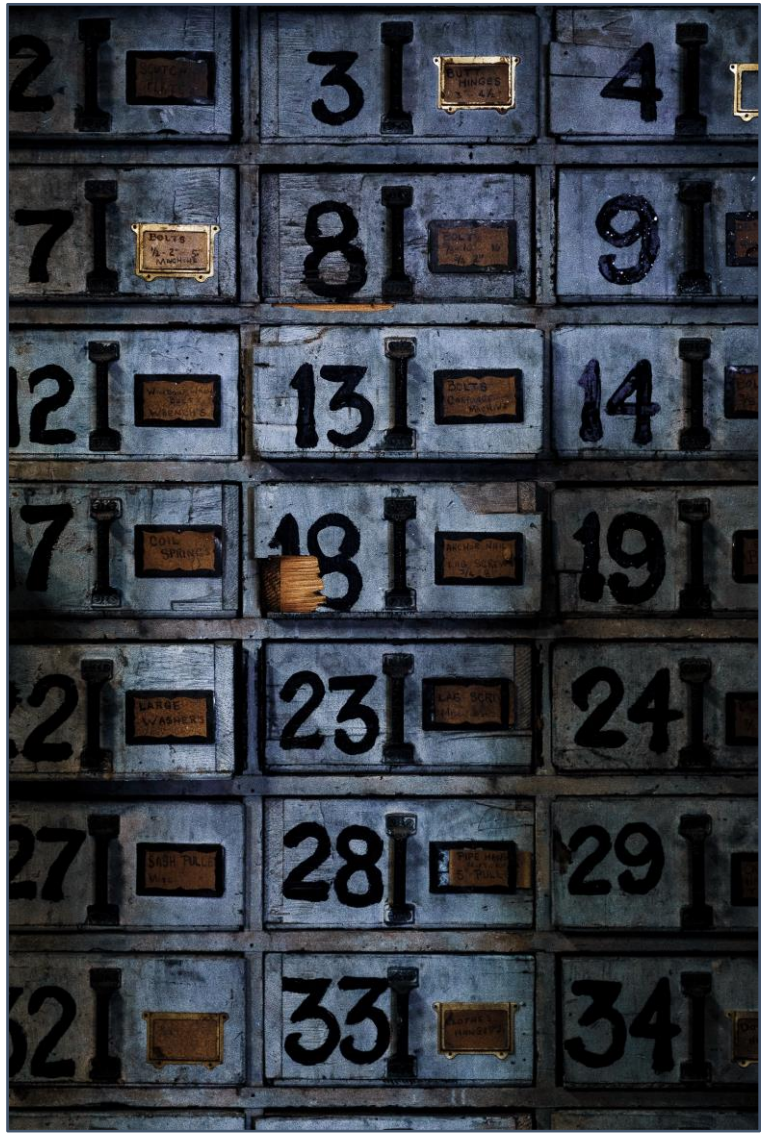
- **Si se cumple H_0** es que **el nivel o categoría que toma el factor no influye sobre el valor medio de la variable dependiente**, por lo que el comportamiento de la variable métrica será **independiente** del grupo de pertenencia de los casos, según el factor.

ANOVA de un factor.

Niveles del factor	Media	Tamaño muestra	Valores de la variable dependiente para los casos de la muestra					
1	μ_1	m_1	y_{11}	y_{12}	...	y_{1j}	...	y_{1m_1}
2	μ_2	m_2	y_{21}	y_{22}	...	y_{2j}	...	y_{2m_2}
...	...	...	...	...	...	...	...	...
i	μ_i	m_i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{im_i}
...	...	...	...	...	...	...	...	...
N	μ_N	m_N	y_{N1}	y_{N2}	...	y_{Nj}	...	y_{Nm_N}

$$M = \sum_{N}^{i=1} m_i$$

Hipótesis nula: $H_0: \mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu_N$



- También puede plantearse el problema como que el valor que toma la variable métrica para el caso j del nivel i del factor, y_{ij}, puede expresarse como:

$$y_{ij} = \mu_i + \varepsilon_{ij} = \mu + a_i + \varepsilon_{ij}$$

Con:

μ_i media de la variable métrica en el nivel i del factor

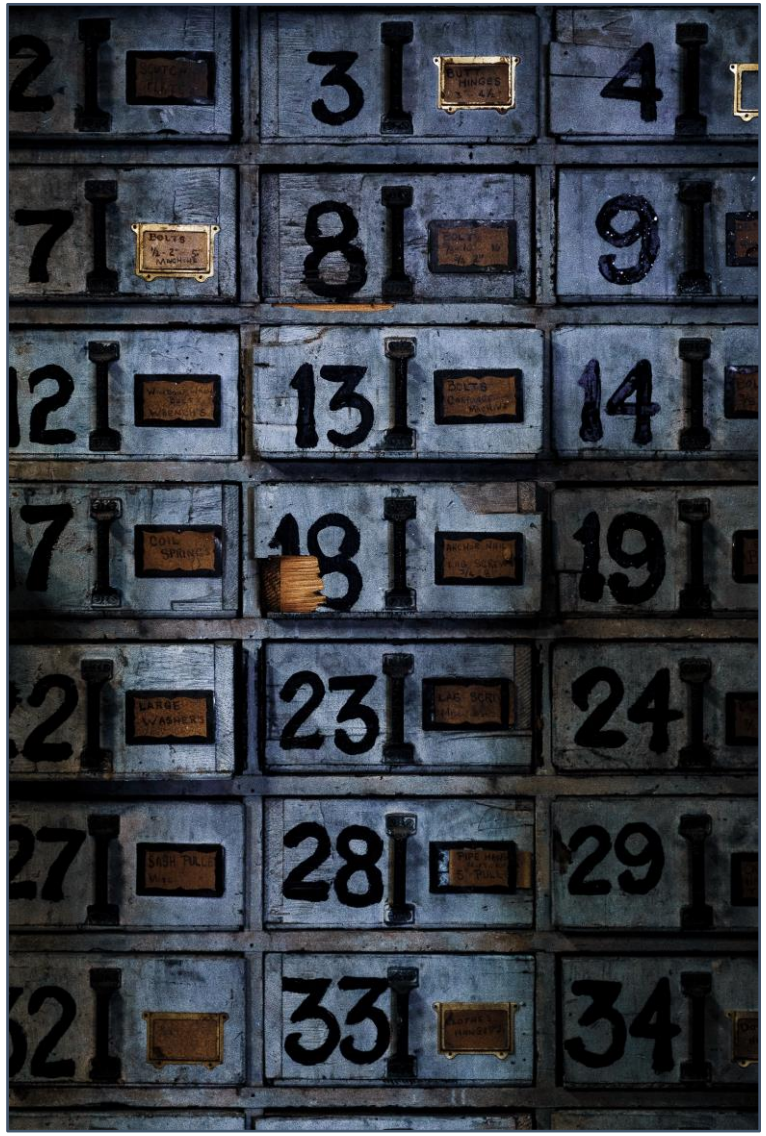
ε_{ij} comportamiento específico de la variable métrica en el individuo j del nivel i

μ media de la variable métrica

a_i comportamiento específico de la variable métrica del nivel i del factor

- En este caso, $H_0: a_i = 0$ para todo i.

ANOVA de un factor.



- Modelo:

$$y_{ij} = \mu_i + \varepsilon_{ij} = \mu + a_i + \varepsilon_{ij}$$

$i = 1, 2, \dots, N$; $j = 1, 2, \dots, m_i$ para el nivel i ; $M = \sum_{N}^{i=1} m_i$

- Sustitución por medias:

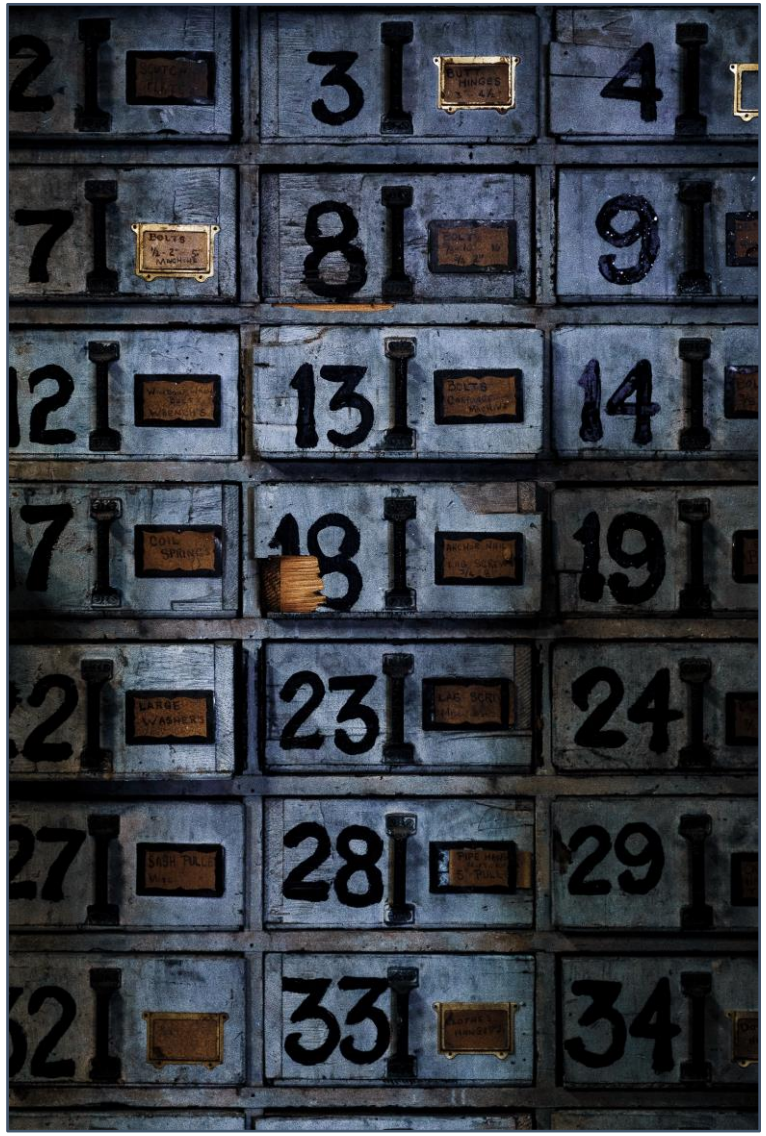
$$y_{ij} = \mu + (\mu_i - \mu) + (y_{ij} - \mu_i)$$

$a_i = \mu_i - \mu$; $\varepsilon_{ij} = y_{ij} - \mu_i$

- Estimación:

$$y_{ij} = \bar{y}_{\bullet\bullet} + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) + (y_{ij} - \bar{y}_{i\bullet})$$

$$\hat{\mu} \to \bar{y}_{\bullet\bullet} = \frac{\sum_{ij} y_{ij}}{\sum_i m_i} \; ; \; \hat{\mu}_i \to \bar{y}_{i\bullet} = \frac{\sum_j y_{ij}}{m_i}$$



- Estimación:

$$y_{ij} = \bar{y}_{\bullet\bullet} + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) + (y_{ij} - \bar{y}_{i\bullet})$$

- Recolocando, elevando al cuadrado y sumando las ecuaciones para todos los casos de la muestra (es decir, para todos los i, j:

$$\sum_i \sum_j (y_{ij} - \bar{y}_{\bullet\bullet})^2 = \sum_i m_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i\bullet})^2$$

Variación total (VT) | Variación entre-grupos (VE) | Variación intra-grupos (VI)

- Bajo H_0 (**no** existen diferencias entre medias) se cumplirá que el estadístico :

$$F = \frac{\frac{VE}{N-1}}{\frac{VI}{M-N}} = \frac{\frac{\sum_i m_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}{N-1}}{\frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i\bullet})^2}{M-N}} \rightarrow F_{(N-1;M-N)}$$

ANOVA de un factor.

- **Ejemplo:** contrastar si la variedad de olivo plantado en una finca tiene una influencia significativa o no sobre el rendimiento obtenido (kilogramos por árbol, para árboles de 6 a 10 años).

- **Muestra:** 48 fincas.
  - 15 de variedad Arbequina.
  - 20 de variedad Cornicabra.
  - 13 de variedad Picual.

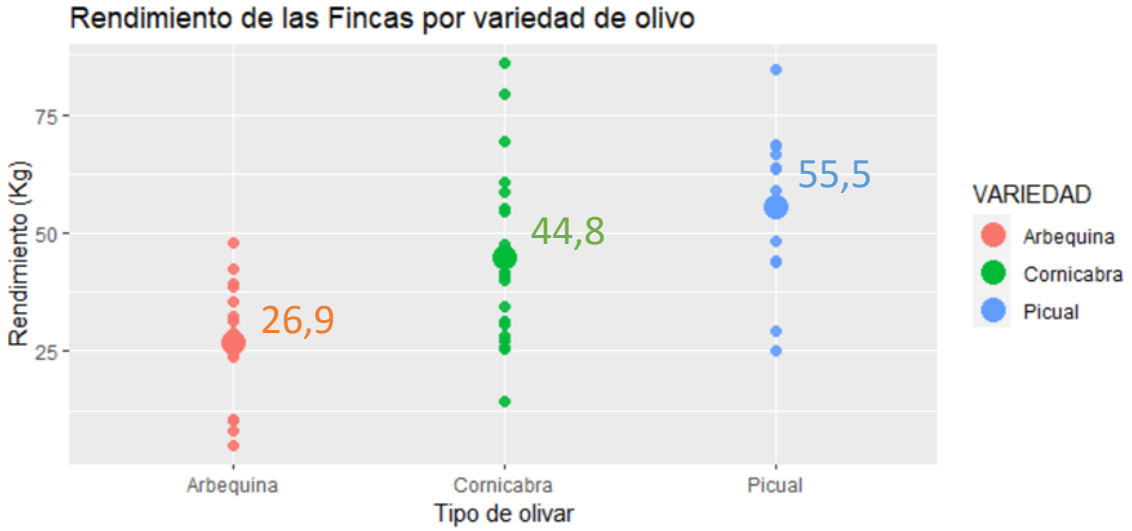
- **Variables:**
  - RENDIMIENTO (dependiente, métrica)
  - VARIEDAD (atributo o factor, con las categorías “Arbequina”, “Cornicabra” y “Picual”.

ANOVA de un factor.

- **Ejemplo:** datos.

VARIEDAD: ARBEQUINA		VARIEDAD: CORNICABRA		VARIEDAD: PICUAL	
FINCA	RENDIMIENTO	FINCA	RENDIMIENTO	FINCA	RENDIMIENTO
F1	24,36	F16	69,28	F36	43,78
F2	39,41	F17	14,24	F37	58,83
F3	38,47	F18	44,81	F38	84,58
F4	10,47	F19	55,25	F39	68,34
F5	23,70	F20	54,59	F40	44,06
F6	32,36	F21	30,78	F41	56,18
F7	42,50	F22	31,21	F42	48,30
F8	24,68	F23	34,33	F43	25,19
F9	35,55	F24	85,77	F44	68,63
F10	31,40	F25	25,80	F45	63,89
F11	28,78	F26	41,70	F46	29,36
F12	10,32	F27	25,60	F47	63,48
F13	8,29	F28	39,81	F48	66,66
F14	5,09	F29	40,88		
F15	48,03	F30	58,48		
		F31	47,65		
		F32	28,21		
		F33	27,11		
		F34	79,37		
		F35	60,57		

- **Ejemplo:** los puntos grandes son el rendimiento medio de cada grupo de la muestra de fincas por tipo de olivo cultivado.



- **Cuestión a resolver por ANOVA:** las diferencias entre los rendimientos medios observados en la muestra, ¿indican rendimientos medios significativamente distintos considerando las poblaciones completas?

- **Ejemplo:** cálculo del estadístico del contraste:

```
                  Df Sum Sq Mean Sq F value   Pr(>F)
olivares$VARIEDAD  2   5940  2970.0   10.45 0.000187 ***
Residuals         45  12786   284.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\frac{\sum_i m_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}{N-1}$$

$$\frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i\bullet})^2}{M-N}$$

$$F = \frac{\frac{VE}{N-1}}{\frac{VI}{M-N}} = \frac{\frac{\sum_i m_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}{N-1}}{\frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i\bullet})^2}{M-N}}$$

P-valor = 0,000187 < 0,05: Se rechaza la hipótesis nula de medias de los grupos iguales, para una significación de 0,05. Por tanto, la muestra apoya el hecho de que existen diferencias relevantes entre los rendimientos medios de las fincas, según la variedad cultivada.

- **Hipótesis básicas del contraste F de ANOVA.** Para poder confiar en los resultados de la prueba, se han de confirmar dos hipótesis de partida:
  - **Normalidad:** cada una de las poblaciones deben seguir una distribución Normal en la variable métrica analizada.
  - **Homogeneidad en las varianzas:** las poblaciones deben poseer varianzas suficientemente parecidas (homogéneas) entre sí en la variable métrica analizada.
  - Si alguna de las hipótesis no se cumplen, hay que recurrir a **pruebas alternativas robustas**, como la prueba de *Kruskal-Wallis*.

- **Comparaciones múltiples:**
  - ANOVA nos dice si **existen o no diferencias significativas en las medias poblacionales** de cada grupo definido por cada nivel o categoría del factor.
  - **Si se rechaza H0**, presumiblemente habrá diferencias significativas entre las medias. ¿Entre todas? No necesariamente. Puede ser solo entre algunas.
  - Así, será preciso analizar **qué medias concretas difieren del resto** (es decir, qué niveles o categorías del factor hacen diferir a la media poblacional de la variable dependiente con respecto al resto).
  - Pruebas de comparaciones múltiples: HSD de

ANOVA multifactorial.

- El **ANOVA multifactorial** es una generalización del caso anterior.
- Imaginemos que tenemos una variable dependiente (Y, por ejemplo el **salario**) y dos factores (**sexo** y **nivel de estudios**).
- Ahora podremos estudiar si el nivel o categoría que presenta cada factor influye sobre el valor medio de la variable dependiente; pero también **si la interacción** entre los niveles o categorías de los factores influyen.
- Es decir, podemos estudiar el **efecto individual** de cada factor sobre Y; pero también si existen **efectos conjuntos**.

ANOVA multifactorial.

- Caso de **2 factores**:

$$y_{ij} = \mu + a_i + b_j + (a\&b)_{ij} + \varepsilon_{ij}$$

- y_{ij}: Valor observación ij de variable dependiente
- μ: Efecto medio común
- a_i: Efecto específico del nivel i del factor A
- b_j: Efecto específico del nivel j del factor B
- $(a\&b)_{ij}$: Efecto de la interacción entre los niveles ij de ambos factores
- ε_{ij}: Error

- En este caso, se realizarán tres contrastes F:
  - Uno en el que H_0: $a_i = 0\ \forall i$.
  - Uno en el que H_0: $b_j = 0\ \forall j$.
  - Uno en el que H_0: $(a\&b)_{ij} = 0\ \forall i, j$.

Análisis multivariante de la varianza

- Es una generalización de ANOVA, tanto para uno como para varios factores.
- En este análisis, se estudia el efecto de los niveles o categorías de uno o varios factores sobre **varias variables dependientes métricas que están correlacionadas** entre sí.
- Se conoce como análisis **MANOVA**.

¡Muchas gracias!