

EVALUACIÓN DE FUENTES DE VARIACIÓN DE DATOS. ANOVA

- **Introducción**
- **Análisis de varianza de una vía**
 - o Fuentes de varianza y significación
 - o La tabla del ANOVA
 - o Suposiciones implícitas
 - o Modelos de efectos fijos y aleatorios
- **Análisis de varianza de dos vías**
 - o Fuentes de varianza y significación
 - o La tabla del ANOVA
 - o Interacción y su estimación
- **ANOVA encajado (nested)**

EVALUACIÓN DE FUENTES DE VARIACIÓN DE DATOS. ANOVA

Introducción

- Las pruebas de significación (test de hipótesis) sirven para comparar dos situaciones diferentes o, dicho de otra manera, el efecto de un factor a dos niveles.
- En ocasiones interesa comparar más de dos situaciones diferentes o el efecto que tienen un factor que puede tomar varios valores o niveles.
- Su pongamos que determinamos el contenido en Cu de una muestra por medio de siete procedimientos de mineralización diferentes:

	MÉTODO						
	1	2	3	4	5	6	7
	5,59	5,67	5,75	4,74	5,52	5,52	5,43
	5,59	5,67	5,47	4,45	5,47	5,62	5,52
	5,37	5,55	5,43	4,65	5,66	5,47	5,43
	5,54	5,5	5,45	4,94	5,52	5,18	5,43
	5,37	5,43	5,24	4,95	5,62	5,43	5,52
	5,42	5,57	5,47	5,06	5,76	5,33	5,52
Media	5,48	5,57	5,47	4,80	5,59	5,43	5,48
Desviación	0,11	0,093	0,16	0,23	0,11	0,15	0,05

- Cada columna difiere de las demás en que los datos se obtienen después de aplicar un procedimiento de mineralización diferente.
- El factor que hace que las columnas difieran ¿tiene algún efecto sobre las medias de esas columnas?. Dicho de otra forma, ¿todos los procedimientos de mineralización originan el mismo resultado?

EVALUACIÓN DE FUENTES DE VARIACIÓN DE DATOS. ANOVA

Introducción

	MÉTODO						
	1	2	3	4	5	6	7
	5,59	5,67	5,75	4,74	5,52	5,52	5,43
	5,59	5,67	5,47	4,45	5,47	5,62	5,52
	5,37	5,55	5,43	4,65	5,66	5,47	5,43
	5,54	5,57	5,45	4,94	5,52	5,18	5,43
	5,37	5,43	5,24	4,95	5,62	5,43	5,52
	5,42	5,57	5,47	5,06	5,76	5,33	5,52
Media	5,48	5,57	5,47	4,80	5,59	5,43	5,48
Desviación	0,11	0,093	0,16	0,23	0,11	0,15	0,05

- Si no hubiera efecto cada valor individual seguiría el modelo:

$$x_{ij} = \mu + e_{ij}$$

siendo los e_{ij} puramente aleatorios. La varianza total de la tabla únicamente depende de la precisión de los procedimientos y se estimaría mediante σ_e^2 .

- Si el procedimiento de mineralización tiene efecto, el modelo sería:

$$x_{ij} = \mu + a_j + e_{ij} ; \sum_j a_j = 0$$

siendo a_j el efecto del pre-tratamiento j sobre la media global.

- El término a_j introduce una varianza adicional en los datos experimentales, de manera que ésta es mayor que σ_e^2 .

ANÁLISIS DE VARIANZA DE UNA VÍA (one-way ANOVA)

Fuentes de varianza y significación

	FACTOR (MUESTRA)					
	1	2	...	j	...	k
	x_{11}	x_{12}	...	x_{1j}	...	x_{1k}
	x_{21}	x_{22}	...	x_{2j}	...	x_{2k}

	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ik}

	x_{n1}	x_{n2}	...	x_{nj}	...	x_{nk}
Media	\bar{x}_1	\bar{x}_2	...	\bar{x}_j	...	\bar{x}_k
Varianza	s_1^2	s_2^2	...	s_j^2	...	s_k^2

- Si el factor no tuviera efecto, la única fuente de variación serían las incertidumbres de las medidas. La varianza de la tabla, σ^2 , podría ser estimada a partir de la primera columna:

$$s_1^2 = \sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 / (n_1 - 1)$$

o bien a partir de cualquiera de las k columnas y todas las columnas deberían dar el mismo resultado:

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_j^2 = \dots = \sigma_k^2 = \sigma^2$$

- En este caso, la varianza promediada es una mejor estimación de σ^2

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + \dots + (n_k - 1)s_k^2}{n_1 + \dots + n_k - k} = \frac{\sum_{j=1}^k (n_j - 1)s_j^2}{\sum_{j=1}^k (n_j - 1)}$$

siendo $n_1 = n_2 = \dots = n_k$

ANÁLISIS DE VARIANZA DE UNA VÍA (one-way ANOVA)

Fuentes de varianza y significación

	FACTOR (MUESTRA)						
	1	2	j	k	
	X ₁₁	X ₁₂	X _{1j}	X _{1k}	
	X ₂₁	X ₂₂	X _{2j}	X _{2k}	
	
	X _{i1}	X _{i2}	X _{ij}	X _{ik}	
	
	X _{n1.1}	X _{n2.2}	X _{nj.j}	X _{nk.k}	
Media	\bar{X}_1	\bar{X}_2	\bar{X}_j	\bar{X}_k	\bar{x}
Varianza	S ₁ ²	S ₂ ²	S _j ²	S _k ²	

- La varianza de la tabla también podría ser estimada también a partir de la varianza de las medias de las columnas

$$s_{\bar{x}}^2 = \sum_{j=1}^k (\bar{X}_j - \bar{x})^2 / (k - 1) \quad \bar{x} = \sum_{j=1}^k \bar{X}_j / k$$

- Por el Teorema del límite central $s_{\bar{x}}^2$ estima $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n_j}$; $\sigma^2 = n_j \sigma_{\bar{x}}^2$

- Por tanto, σ^2 se puede estimar mediante $s^2 = n_j s_{\bar{x}}^2 = n_j \sum_{j=1}^k (\bar{X}_j - \bar{x})^2 / (k - 1)$

ANÁLISIS DE VARIANZA DE UNA VÍA (one-way ANOVA)

Fuentes de varianza y significación

	FACTOR (MUESTRA)						
	1	2	j	k	
	X ₁₁	X ₁₂	X _{1j}	X _{1k}	
	X ₂₁	X ₂₂	X _{2j}	X _{2k}	
	
	X _{i1}	X _{i2}	X _{ij}	X _{ik}	
	
	X _{n1.1}	X _{n2.2}	X _{nj.j}	X _{nk.k}	
Media	\bar{X}_1	\bar{X}_2	\bar{X}_j	\bar{X}_k	\bar{x}
Varianza	S ₁ ²	S ₂ ²	S _j ²	S _k ²	

- Si el factor no tiene efecto, ambas estimaciones de σ^2 serán idénticas.

$$s_p^2 = \sum_{j=1}^k (n_j - 1) s_j^2 / \sum_{j=1}^k (n_j - 1)$$

$$s^2 = n_j s_{\bar{x}}^2 = n_j \sum_{j=1}^k (\bar{X}_j - \bar{x})^2 / (k - 1)$$

- Si el factor tiene efecto:

s_p^2 no es afectada pues sus componentes se determinan dentro de cada columna: varianza dentro de las columnas (within column).

$s_{\bar{x}}^2$ sí que será afectada pues sus componentes son las medias de cada columna que se verán afectadas (varianza entre columnas o between column). Por tanto $s_{\bar{x}}^2$ estima

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n_j} + \sigma_a^2$$

término adicional
(efecto del factor)

ANÁLISIS DE VARIANZA DE UNA VÍA (one-way ANOVA)

Fuentes de varianza y significación

- Se realiza ahora una prueba de significación formulando las dos hipótesis:

- Si el factor no tiene efecto s_p^2 y $n_j s_x^2$ estiman σ^2

$$\Rightarrow H_0: \sigma_p^2 = n_j \sigma_x^2 \quad \text{ó} \quad \sigma_a^2 = 0$$

- Si el factor tiene efecto $n_j s_x^2$ estima una varianza más grande que s_p^2

$$\Rightarrow H_1: \sigma_p^2 < n_j \sigma_x^2 \quad \text{ó} \quad \sigma_a^2 > 0$$

σ^2 estimada dentro de las columnas

σ^2 estimada entre las columnas

- La comparación se hace mediante una prueba F (o de Fischer) de comparación de varianzas.

Comparación de dos varianzas: Prueba F de Fischer			
$H_0: \sigma_1^2 = \sigma_2^2$	$H_1: \sigma_1^2 > \sigma_2^2$	(test de 1 cola)	
$s_1^2 \geq s_2^2$	$F = \frac{S_1^2}{S_2^2}$	$F_{\text{crit}(1-\alpha, v_1, v_2)}$	$v_1 = n_1 - 1$ $v_2 = n_2 - 1$

siendo s_1 la varianza entre columnas y s_2 la varianza dentro de las columnas.

ANÁLISIS DE VARIANZA DE UNA VÍA (one-way ANOVA)

La tabla del ANOVA

- El ANOVA puede entenderse y llevarse a cabo como una separación de la varianza total de la tabla en sus componentes.

$$s_T^2 = \sum_{j=1}^k \sum_{i=1}^{n_k} (x_{ij} - \bar{x})^2 / (n-1) ; \quad n = \sum_{j=1}^n n_j ; \quad \bar{x} = \frac{\sum_{j=1}^k \sum_{i=1}^{n_k} x_{ij}}{n}$$

- Para ello se deben calcular las varianzas que siempre se determinan según:

$$\text{varianza} = \frac{\text{Suma de Cuadrados}}{\text{grados de libertad}}$$

- SUMAS DE CUADRADOS

$$SS_T = \sum_j \sum_i (x_{ij} - \bar{x})^2$$

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

$$(x_{ij} - \bar{x})^2 = (x_{ij} - \bar{x}_j)^2 + (\bar{x}_j - \bar{x})^2 + 2(x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{x})$$

Al hacer los sumatorios, primero sobre las filas (i) y luego sobre las columnas (j), el último término se anula:

$$SS_T = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2 + \sum_j n_j (\bar{x}_j - \bar{x})^2$$

dentro de las columnas

entre las columnas

$$SS_T = SS_R + SS_A$$

ANÁLISIS DE VARIANZA DE UNA VÍA (one-way ANOVA)

La tabla del ANOVA

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

$$SS_T = SS_R + SS_A$$

GRADOS DE LIBERTAD

- Para SS_T son $(n-1)$ ← **n puntos menos el centroide**
- Para SS_A se gastan $(k-1)$ ← **k columnas menos el centroide**
- SS_R quedan $(n-1)-(k-1) = (n-k)$

MEDIAS CUADRADAS O CUADRADOS MEDIOS

- $MS_A = SS_A/(k-1)$ ← estimación de σ^2 **entre** las columnas
- $MS_R = SS_R/(n-k)$ ← estimación de σ^2 **dentro de** las columnas

TEST DE HIPÓTESIS

- Consiste en calcular

$$F = \frac{MS_A}{MS_R} = \frac{SS_A / (k - 1)}{SS_R / (n - k)}$$

que se compara con el F crítico de una cola con $k-1$ y $n-k$ grados de libertad.

ANÁLISIS DE VARIANZA DE UNA VÍA (one-way ANOVA)

La tabla del ANOVA

- Los resultados se presentan en forma de tabla

Fuente	g.d.l	SS	MS	F	p
Entre columnas (A)	k-1	SS_A	$SS_A/(k-1)$	MS_A/MS_R	p
Dentro de columnas (residual)	n-k	SS_R	$SS_R/(n-k)$		
Total	n-1	SS_T			

$F_{crit} (0,05, k-1, n-k) = \dots$

DISTR.F($MS_A/MS_R; k-1; n-k$)

- Para el ejemplo

Fuente	g.d.l	SS	MS	F	p
Entre columnas (A)	6	2,6834	0,4472	23,1529	$8 \cdot 10^{-11}$
Dentro de columnas (residual)	35	0,6761	0,0193		
Total	41	3,3595			

$F_{crit} (0,05, 6, 35) = 2,37$

DISTR.F.INV(0.05, k-1, n-k)

**$F_{cal} > F_{crit}$
P < 0.05**

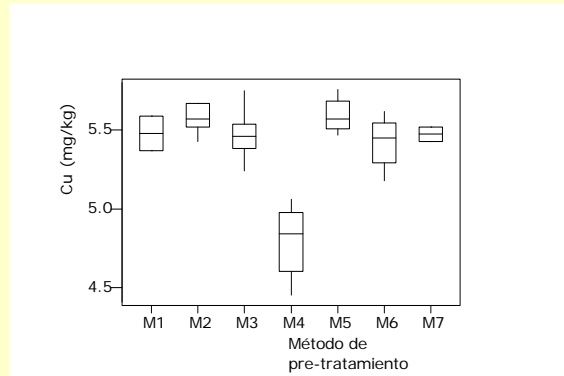
- En el caso del ejemplo, el factor (pre-tratamiento de la muestra) sí que tiene un efecto significativo sobre los resultados analíticos.

ANÁLISIS DE VARIANZA DE UNA VÍA (one-way ANOVA)

Suposiciones implícitas

- LA MS_R se estima a partir de una varianza promediada, por tanto se supone que todas las columnas tienen varianzas homogéneas: homocedasticidad. A veces eso no es cierto.

- Una forma de comprobarla es hacer una inspección visual de los datos antes del ANOVA, por medio de procedimientos tales como los boxplots



- Existen otras pruebas para comprobar que las varianzas son similares, como la F de Fischer.

- Si se comprueba la heterocedasticidad cabe varias opciones: Eliminar las columnas con varianza muy grande o transformar las variables.

ANÁLISIS DE VARIANZA DE UNA VÍA (one-way ANOVA)

Modelos de efectos fijos y efectos aleatorios

- El modelo general del ANOVA es: $x_{ij} = \mu + a_j + e_{ij}$

donde cada resultado se descompone en varios componentes, uno de los cuales es el efecto del factor a_j .

Modelos de efectos fijos (fixed effect models) o ANOVA modelo I

El efecto del factor hace desviarse de forma fija la media de cada grupo j de la gran media:

$$x_{ij} = (\mu + a_j) + e_{ij}$$

MS_A estima $\sigma^2 + \sum_j n_j a_j^2 / (k - 1)$ y la prueba F se hace de la siguiente manera

$$H_0 : a_1 = a_2 = \dots = a_k = 0$$

$$H_1 : a_j \neq 0 \text{ para al menos un } j$$

- Si se rechaza H_0 , al menos una columna difiere significativamente del resto (al menos un método de pretratamiento es diferente de los otros).
- Para saber qué columna es la diferente y cuál es la cuantía de la diferencia, se utiliza la diferencia menos significativa (**LSD**) (análogo a una prueba de significación)

$$LSD = t_{crit} \sqrt{MS_R (2/n_j)}$$

- 1) Se calcula con t_{crit} al nivel de significación que se desee y con un n° g.d.l. igual al que se utilizó para calcular MS_R .
- 2) Se compara cada par $|\bar{x}_1 - \bar{x}_2|$ con LSD y se elige aquel j que participe en más parejas que sean superiores a la LSD

ANÁLISIS DE VARIANZA DE UNA VÍA (one-way ANOVA)

Modelos de efectos fijos y efectos aleatorios

Modelo de efectos aleatorios (random effect models) ANOVA modelo II

- No estamos interesados en un efecto específico debido a una cierta columna, sino en un efecto general sobre todas las columnas y que dicho efecto esté normalmente distribuido.
- El modelo es como el general: $x_{ij} = \mu + a_j + e_{ij}$
- MS_A estima $\sigma^2 + \sigma_a^2 \frac{n - (\sum n_j^2 / n)}{k - 1}$ o para iguales n_j estima $\sigma^2 + n_j \sigma_a^2$
- Puesto que el efecto es aleatorio y en realidad afecta al material sobre el que se opera y no al factor que se está estudiando, no tiene sentido conocer qué columna es significativamente diferente de las otras. No obstante, si se rechaza H_0 , el efecto existe y se puede conocer su cuantía σ_a^2
- En el caso de que los n_j sean idénticos (todas las columnas tienen la misma longitud) la varianza debida a la heterogeneidad de la muestra es simplemente:

$$s_a^2 = (MS_A - MS_R) / n_j$$

ANÁLISIS DE VARIANZA DE UNA VÍA (one-way ANOVA)

EXCEL

Modelos de efectos fijos y efectos aleatorios

- La diferencia entre ambos modelos no es siempre evidente sino más bien de **tipo filosófico**, ya que los cálculos se hacen de la misma manera.
- Sea un ejercicio de intercomparación: Se envía la misma muestra a varios laboratorios.
 - a) Si estamos interesados en ver si los laboratorios trabajan bien (proficiency testing) se trataría de un modelo de **efectos fijos**. (El material analizado debería ser perfectamente homogéneo: MRC)
 - b) Si estamos interesados en saber si el material analizado es perfectamente homogéneo, (material testing) sería un modelo de **efectos aleatorios**. Evidentemente, todos los laboratorios trabajarían bien, por lo que si $MS_A > MS_R$, la "culpa" no será de ninguno.
 - c) Si se sabe que todos los laboratorios trabajan bien, y que el material analizado es homogéneo, el modelo es también de **efectos aleatorios**:
 - la varianza dentro de las columnas describiría la repetitividad
 - la varianza global la reproducibilidad
 - y la varianza entre columnas el componente debido a la varianza entre laboratorios (bias de los laboratorios).

ANOVA DE DOS VÍAS Y MULTIVÍA

- A veces se estudian dos o más factores. P. e. efecto que tienen sobre la determinación de Cu en suelos varios procedimientos de mineralización de la muestra y simultáneamente varias formas de desecarla.

		MÉTODO		
		1	2	3
Seca al Aire		5,59	5,67	5,75
		5,59	5,67	5,47
		5,37	5,55	5,43
		5,54	5,57	5,45
		5,37	5,43	5,24
		5,42	5,57	5,47
Seca en estufa		4,74	5,52	5,52
		4,45	5,47	5,62
		4,65	5,66	5,47
		4,94	5,52	5,18
		4,95	5,62	5,43
		5,06	5,76	5,33

- Estas tablas se llaman tablas de dos vías (two-way) y su análisis ANOVA de dos vías (two-way ANOVA) ya que los datos están sujetos a una **dobles clasificación**.
- Cada intersección es una **celda** y puede haber uno o más datos en ella. Si hay **replicación** cada celda contienen más de un dato (en el ejemplo, cada celda contiene 6 datos).

ANOVA DE DOS VÍAS Y MULTIVÍA

La tabla del ANOVA

FACTOR A	1	2	...	j	...	k	Medias Factor A
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1k}	$\bar{x}_{1.}$
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2k}	$\bar{x}_{2.}$
...
h	x_{h1}	x_{h2}	...	x_{hj}	...	x_{hk}	$\bar{x}_{h.}$
...
l	x_{l1}	x_{l2}	...	x_{lj}	...	x_{lk}	$\bar{x}_{l.}$
Medias Factor B	$\bar{x}_{.1}$	$\bar{x}_{.2}$...	$\bar{x}_{.j}$...	$\bar{x}_{.k}$	Gran media $\bar{x}_{..}$

- Los datos siguen un modelo lineal: $x_{hj} = \mu + a_h + b_j + e_{hj}$

Cada valor viene afectado por el efecto del factor a, el efecto del factor b y un residual que debería ser aleatorio..

- Si no hay replicación (un resultado por celda), la gran media de los datos es:

$$\bar{x} = \sum_h \sum_j x_{hj} / lk ; \quad h = 1 \text{ a } l ; \quad j = 1 \text{ a } k$$

- Hay **l niveles** del factor A y la media de cada nivel se calcula según $\bar{x}_{h.} = \sum_j x_{hj} / k$
- Hay **k niveles** del factor B y la media de cada nivel viene dada por $\bar{x}_{.j} = \sum_h x_{hj} / l$

ANOVA DE DOS VÍAS Y MULTIVÍA

La tabla del ANOVA

- La SS_T puede dividirse en componentes debidos a los diferentes factores y a los residuales

$$SS_T = \sum_h \sum_j (x_{hj} - \bar{x})^2 = SS_A + SS_B + SS_R$$

g.d.l. = t = (kl-1 si no hay replicación)

- Los diferentes componentes, g.d.l. y MS son:

- Para el factor A

$$SS_A = \sum_h \sum_j (\bar{x}_h - \bar{x})^2 = k \sum_h (\bar{x}_h - \bar{x})^2 ; \text{ g.d.l.} = l - 1$$

$$MS_A = SS_A / (l - 1)$$

- Para el factor B

$$SS_B = \sum_h \sum_j (\bar{x}_j - \bar{x})^2 = l \sum_j (\bar{x}_j - \bar{x})^2 ; \text{ g.d.l.} = k - 1 ;$$

$$MS_B = SS_B / (k - 1)$$

- Para los residuales

$$SS_R = SS_T - SS_A - SS_B ; \text{ g.d.l.} = r = t - (k-1) - (l-1)$$

[(kl-1) - (k-1) - (l-1) si no hay replicación]

$$MS_R = SS_R / r$$

- Cuando se hacen réplicas, en cada celda existe más de un valor. No es preciso que todas las celdas contengan el mismo número de réplicas, pero debería evitarse el que haya grandes diferencias.

ANOVA DE DOS VÍAS Y MULTIVÍA

La tabla del ANOVA

- La tabla es similar a la del ANOVA de una vía

Fuente	g.d.l	SS	MS	F
Efectos globales	(l-1) + (k-1)	$SS_A + SS_B$		
Factor A	l-1	SS_A	$SS_A / (l-1)$	MS_A / MS_R
Factor B	k-1	SS_B	$SS_B / (k-1)$	MS_B / MS_R
Residual	$t - (k-1) - (l-1) = r$	SS_R	SS_R / r	replicación con n_j datos por celda
Total	$t = n_j k l - 1$	SS_T		

- Se hacen pruebas F de 1 cola con los correspondientes g.d.l. de numerador y denominador para comprobar si los efectos de los factores A y B son estadísticamente significativos frente a la varianza residual

- En el caso del ejemplo

Fuente	g.d.l	SS	MS	F	p
Efectos globales	3	1,7501			
Factor A	1	0,5041	0,5041	10,34	0,003
Factor B	2	1,2460	0,6230	12,78	$8 \cdot 10^{-5}$
Residual	32	1,5601	0,0488		
Total	35	3,3102			

DISTR.F.INV (0,05; n_1 ; n_2) DISTR.F(F; n_1 ; n_2)

- Los F críticos son:

Factor A (1 y 32 g.d.l.) 4,15
Factor B (2 y 32 g.d.l.) 3,30



ambos factores son significativos

ANOVA DE DOS VÍAS Y MULTIVÍA

Interacciones

- A veces el efecto de uno de los factores depende del nivel del otro factor. Esto es la **interacción**.
- Hay otro término adicional como fuente de varianza, y el modelo lineal subyacente es:

$$x_{hj} = \mu + a_h + b_j + (ab)_{hj} + e_{hj}$$
- El nº de g.d.l. necesarios es el producto $(k-1)(l-1)$. Ese es el nº de g.d.l. que queda para los residuales cuando no hay replicaciones, por tanto **si se desea estudiar la interacción deben hacerse réplicas**.
- La Tabla del ANOVA que resulta es:

Fuente	g.d.l	SS	MS	F
Efectos globales	$(l-1) + (k-1)$	$SS_A + SS_B$		
Factor A	$l-1$	SS_A	$SS_A/(l-1)$	MS_A/MS_R
Factor B	$k-1$	SS_B	$SS_B/(k-1)$	MS_B/MS_R
Interacción	$(l-1)(k-1) = \text{inter}$	SS_{inte}	$SS_{\text{inte}}/\text{inter}$	MS_{inter}/MS_R
Residual	$t - (k-1) - (l-1) - (l-1)(k-1) = r$	SS_R	SS_R/r	
Total	$n_{jkl} - 1 = t$	SS_T		

EXCEL

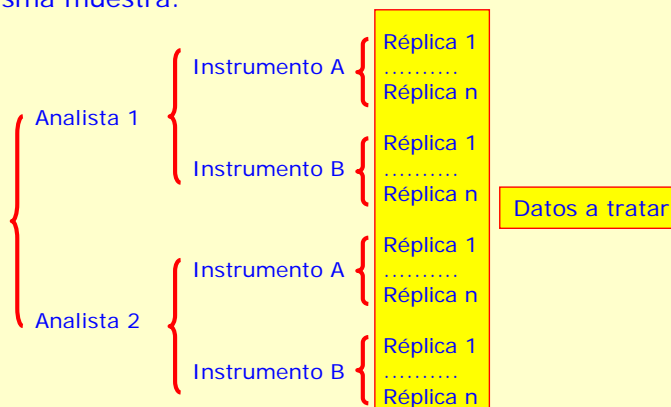
- Ejemplo

Fuente	g.d.l	SS	MS	F	p
Efectos globales	3	1,7501			
Factor A	1	0,5041	0,5041	22,81	$4 \cdot 10^{-5}$
Factor B	2	1,2460	0,6230	28,19	$1 \cdot 10^{-7}$
Interacción	2	0,8962	0,4481	20,27	$3 \cdot 10^{-6}$
Residual	30	0,6639	0,0221		
Total	35	3,3102			

todos los efectos son significativos

ANOVA ENCAJADO (NESTED ANOVA)

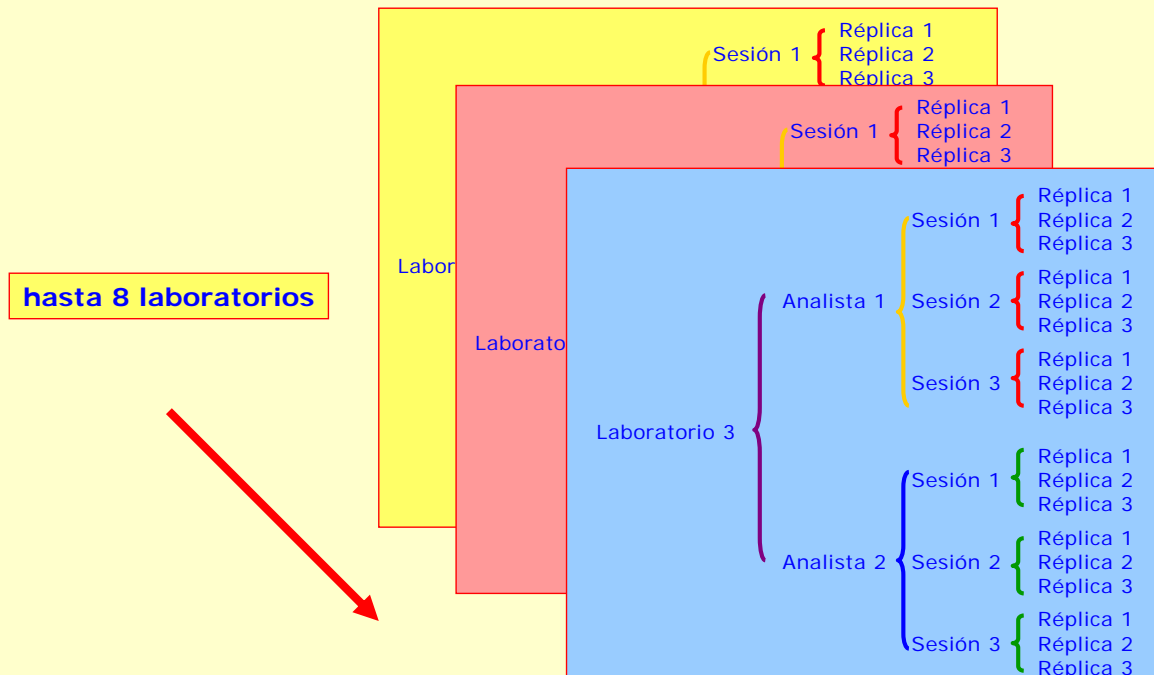
- Tienen gran interés en **ensayos de validación**
- Permiten encontrar las **contribuciones de los diferentes factores** que pueden tener efecto sobre un resultado
- Son imprescindible para determinar la **incertidumbre** de los resultados (que es la raíz cuadrada de la varianza combinada)
- Si se quiere determinar el efecto de **2 diferentes analistas y 2 diferentes instrumentos** sobre la varianza (incertidumbre) de los resultados generados.
- Cada Analista (1 y 2) utilizaría los dos Instrumentos (A y B) para hacer determinaciones Replicadas de una misma muestra:



- Se tiene siempre un **esquema jerárquico en forma de árbol ramificado**.

ANOVA ENCAJADO (NESTED ANOVA)

- Ejemplo. En **ocho** laboratorios ($a=8$), **dos** analistas diferentes ($b=2$) analizan en **tres** sesiones diferentes ($c=3$), **tres** réplicas ($n=3$) de un mismo material.



ANOVA ENCAJADO (NESTED ANOVA)

- Ejemplo. En **ocho** laboratorios ($a=8$), **dos** analistas diferentes ($b=2$) analizan en **tres** sesiones diferentes ($c=3$), **tres** réplicas ($n=3$) de un mismo material.
- El reparto de las SS aditivas de un ANOVA encajado sería:

$$(x_{ijkl} - \bar{x}) = (x_{ijkl} - \bar{x}_{ijk}) + (\bar{x}_{ijk} - \bar{x}_{ij}) + (\bar{x}_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x})$$

- x_{ijkl} , el valor obtenido para la réplica l , el día k , por el analista j , en el laboratorio i
- \bar{x}_{ijk} , la media de todas las réplicas para el día k del analista j en el laboratorio i
- \bar{x}_{ij} , la media de todos los días del analista j en laboratorio i
- \bar{x}_i , la media de todos los analistas del laboratorio i
- \bar{x} , la gran media o centroide

- Si elevamos al cuadrado y hacemos los sumatorios correspondientes (sobre i, j, k y l), los términos cruzados se anularán y, reordenando, resultará finalmente:

$$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c \sum_{l=1}^n (x_{ijkl} - \bar{x})^2 = bcn \sum_{i=1}^a (\bar{x}_i - \bar{x})^2 + cn \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{ij} - \bar{x}_i)^2 + n \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (\bar{x}_{ijk} - \bar{x}_{ij})^2 + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c \sum_{l=1}^n (x_{ijkl} - \bar{x}_{ijk})^2$$

$$SS_T = SS_A + SS_B + SS_C + SS_D$$

cada laboratorio respecto de la gran media
cada analista respecto de la media de su laboratorio
cada sesión del analista respecto de su propia media
cada réplica en la misma sesión del mismo analista

- Grados de libertad:

$$abcn-1 = a-1 + a(b-1) + ab(c-1) + abc(n-1)$$

ANOVA ENCAJADO (NESTED ANOVA)

La tabla del ANOVA

Fuente de variación	g.d.l.	SS	MS	F
Laboratorios (A)	a-1	SS _A	MS _A	MS _A / MS _B
Analistas dentro de cada laboratorio (B)	a(b-1)	SS _B	MS _B	MS _B / MS _C
Días dentro de cada analista (C)	ab(c-1)	SS _C	MS _C	MS _C / MS _D
Réplicas dentro de días (D)	abc(n-1)	SS _D	MS _D	
Total	abcn-1	SS _T		

- Cada F se obtiene **dividiendo su estimación MS por la que que está inmediatamente debajo**
- Los valores de las varianzas correspondientes a cada uno de los factores (o niveles):
 - σ_D^2 es la varianza debida a las réplicas (que suele coincidir con la repetitividad)
 - σ_C^2 la debida a los diferentes días
 - σ_B^2 la debida a los diferentes analistas
 - σ_A^2 la debida a los laboratorios.
- Cada uno de ellas se estima a partir de la correspondiente MS por medio de:
 - MS_D estima σ_D^2
 - MS_C estima $\sigma_D^2 + n\sigma_C^2$
 - MS_B estima $\sigma_D^2 + n\sigma_C^2 + cn\sigma_B^2$
 - MS_A estima $\sigma_D^2 + n\sigma_C^2 + cn\sigma_B^2 + bcn\sigma_A^2$
- Por tanto se pueden estimar **todos los componentes individuales**, ver cual es el más importante y (si son independientes) calcular la varianza combinada y de ahí la **incertidumbre**.

ANOVA ENCAJADO (NESTED ANOVA)

Ensayos inter-laboratorios

- Un **material homogéneo** se reparte entre **a laboratorios**. Cada laboratorio hace **b determinaciones** analíticas, cada una de las cuales consta de **n réplicas**.
- Hay pues tres niveles y el reparto se basa en:
$$e_{ij} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x}) + (\bar{x} - \mu_0)$$
que origina: $SS_T = SS_C + SS_B + SS_A$
- Los componentes de la varianza se pueden estimar como antes: σ_C^2 , σ_B^2 y σ_A^2 :

σ_C **repetitividad** ←
 σ_B **incertidumbre causada por los bias del laboratorio** ←
 σ_A **incertidumbre debida al bias del método** ←

de acuerdo a:

$$\begin{aligned}\sigma_C^2 &= MS_C \\ \sigma_B^2 &= (MS_B - MS_C)/n \\ \sigma_A^2 &= (MS_A - MS_B)/(n*b)\end{aligned}$$

- **La reproducibilidad** se podrá estimar a partir de $\sqrt{\sigma_C^2 + \sigma_B^2}$ ← **interacción aleatorio ↔ bias**
- Es importante notar que solo de esta forma se pueden separar los efectos correspondientes a **todas las etapas** de un método analítico, con objeto de estimar los puntos críticos y las incertidumbres de cada una de las etapas (muestreo, almacenamiento, submuestreo...).