EVALUACIÓN DE FUENTES DE VARIACIÓN DE DATOS. ANOVA

1 INTRODUCCIÓN

Anteriormente se han discutido los tests de hipótesis (pruebas de significación) de comparación de medias, pero en ocasiones es preciso comparar más de dos medias. La tabla siguiente presenta los resultados obtenidos al determinar Cu en una muestra de suelo agrícola aplicando siete procedimientos diferentes de mineralización.

	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

En vez de comparar las medias de las columnas dos a dos, podemos plantear una cuestión más general: El factor que hace que las columnas difieran, ¿tiene algún efecto sobre las medias de esas columnas?. Dicho de otra manera: ¿todos los procedimientos de mineralización originan el mismo resultado?.

Si el factor no tuviera efecto alguno, la varianza total de la tabla dependería exclusivamente de la precisión de los procedimientos, por lo que todos los valores de la tabla pertenecerían a la misma población, y podríamos poner:

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

siendo x_{ij} el número que aparece en la fila i y la columna j, μ el valor verdadero y e_{ij} un error aleatorio de media cero y varianza $\sigma_e^2 = \sigma^2$.

Por el contrario, si los procedimientos de mineralización tienen efecto sobre los resultados experimentales obtenidos, podemos escribir:

$$x_{ij} = \mu + a_j + e_{ij} \quad ; \quad \sum a_{ij} = 0$$

siendo a_j el efecto del pretratamiento j sobre la media global. El término a_j introduce una varianza adicional en los datos de la columna j, de manera que tendrán una varianza mayor de σ_e^2.

La tabla es de una vía, puesto que hay un solo factor (la mineralización previa) a siete niveles. La prueba de significación a aplicar para determinar si ese factor ejerce algún efecto, se denomina Análisis de Varianza de una vía (one-way ANOVA)

Existen otras situaciones que conducen a resultados experimentales similares. P.e. estudios inter-laboratorios en los que una misma muestra se envía a varios laboratorios. Si se s envía una muestra totalmente homogénea y todos emplean el mismo procedimiento, las posibles diferencias serían achacables a los laboratorios. Si emplean diferentes procedimientos, las diferencias serían una mezcla del laboratorio y del método empleado. Ambos estudios son del tipo laboratory-proficiency.

Otras veces, lo que se quiere comprobar es la calidad del material analizado (material-testing) y se envía la muestra a varios laboratorios cuyos resultados estén exentos de bias. Si hay diferencias, deberían ser achacables a la heterogeneidad de la muestra.

En cualquiera de los casos, existe un único factor a varios niveles, y el tratamiento es idéntico.

2 ANÁLISIS DE VARIANZA DE UNA VÍA

2.1 Fuentes de varianza y significación

Pongamos la tabla de forma más general

	MUESTRA					
	1	**2**	**....**	**j**	**....**	**k**
	x_{11}	x_{21}	....	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	$\bar{x}_1$	$\bar{x}_2$	....	$\bar{x}_j$	....	$\bar{x}_k$
Varianza	s_1^2	s_2^2	....	s_j^2	....	s_k^2

Supongamos que el lote es **homogéneo** y que la **única fuente** de variación son las **incertidumbres** de las medidas. En ese caso la varianza podría ser estimada a partir de la primera columna:

$$s_1^2 = \sum_{i=1}^{n1} \left(x_{i1} - \bar{x}_1\right)^2 / (n_1 - 1)$$

o bien a partir de la segunda, o de la tercera, o de cualquiera de las k columnas. Si todos los datos proceden de la misma población de media μ y σ^2, las medias y varianzas de cada columna ($\bar{x}_j, s_j^2$) son estimaciones diferentes de esos parámetros. El valor μ se estimaría mediante $\bar{\bar{x}}$ (media de las k medias columnares $\bar{x}_j$ o gran media) mientras que la varianza puede calcularse como una **varianza promediada (pooled) de las varianzas de las k columnas** (Esto implica que todas las varianzas pertenecen a la misma población: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_j^2 = ... = \sigma_k^2 = \sigma^2$)

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

y s_p^2 será una estimación de σ^2 mejor que cualquiera de las s_j^2 individuales.

Otra posibilidad para estimar σ^2 calcular la **varianza de las medias de las columnas** $s_{\bar{x}}^2$:

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

Evidentemente, $s_{\bar{x}}^2$ estima $\sigma_{\bar{x}}^2$, y como hay n_j datos en cada columna (para facilitar los cálculos supondremos que todas las columnas tienen el mismo número de datos $n_1 = n_2 = \ldots = n_k$.) debido al **Teorema del Límite Central** se cumplirá $\sigma_{\bar{x}}^2 = \sigma^2/n_j$ o bien $\sigma^2 = n_j \sigma_{\bar{x}}^2$, por lo que σ^2 es estimada mediante $n_j s_{\bar{x}}^2$:

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

Las dos estimaciones de σ^2 serán iguales si el material es homogéneo (y todas las columnas son muestras representativas de la población). **Si el material es heterogéneo las dos cantidades estiman cosas diferentes.**

Si el material fuera heterogéneo y las columnas fueran en realidad diferentes, la estimación s_p^2 no resultaría afectada, pues sus componentes (s_j^2) se determinan dentro de cada columna y por tanto dependen única y exclusivamente de la precisión de la determinación analítica: La existencia de un bias en alguna de las columnas, afectaría de igual manera a todos los componentes de esa columna y a su media $\bar{x}_j$ y el correspondiente s_j^2 sería el mismo. Por tanto s_p^2 describe la **varianza dentro de las columnas (within-column variance).** Por tanto s_p^2 aún estima σ^2.

Por su parte, la varianza de las medias columnares $s_{\bar{x}}^2$ o **varianza entre columnas (between-column)** sí que va ser afectada por la heterogeneidad, pues la existencia de un bias va a originar que alguna de las $\bar{x}_j$ utilizadas en el cálculo, va a ser muy diferente de las restantes**.** En ese caso, $s_{\bar{x}}^2$ no estima únicamente σ^2/n_j sino que debe añadirse un **componente adicional** σ_a^2 que estima la varianza adicional debida a la heterogeneidad de las columnas: $\sigma_{\bar{x}}^2 = \sigma^2 / n_j + \sigma_a^2$, por tanto $n_j s_{\bar{x}}^2$ estima $\sigma^2 + n_j \sigma_a^2$.

Todo ello nos permite formular una hipótesis y comprobarla.

1) Si el material es homogéneo s_p^2 y $n_j s_{\bar{x}}^2$ estiman σ^2:

$H_0 : \sigma_p^2 = n_j \sigma_{\bar{x}}^2$ o $H_0 : \sigma_a^2 = 0$

2) Si el material es heterogéneo $n_j s_{\bar{x}}^2$ estima una varianza más grande que s_p^2

$H_1 : \sigma_p^2 < n_j \sigma_{\bar{x}}^2$ o $H_1 : \sigma_a^2 > 0$

Estas varianzas pueden ser comparadas con una prueba F de una cola.

2.2 La tabla del ANOVA

La forma práctica real de llevar a cabo el ANOVA y comprender mejor los cálculos es considerar el ANOVA como la **separación de la varianza total en sus componentes**. La varianza total es:

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

De forma general, una estimación de varianza es un cociente de una suma de cuadrados y unos grados de libertad. Por razones de facilidad de cálculo trabajaremos primero con las sumas de cuadrados SS (sums of squares)

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

Como se cumple que

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

al elevar al cuadrado

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

Al sumar primero sobre las filas (i) y luego sobre las columnas (j) el último término se anula y el resultado es:

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

o lo que es igual

$SS_T = SS_R + SS_A$

SS_R se llama **suma residual de cuadrados** o suma de cuadrados residuales y coincide con la SS calculada **dentro** de las columnas. **SS_A** es la **suma de cuadrados debida al efecto** del factor estudiado (la heterogeneidad entre muestras) y coincide con la SS calculada **entre** columnas. A veces se llama $SS_{TRATAMIENTO}$ debido al origen agronómico del ANOVA.

Los **grados de libertad** (g.d.l.) disponibles para SS_T son (n-1) ya que se gasta uno de los n disponibles en estimar $\overline{x}$. Esos g.d.l. de SS_T se reparten entre los de SS_R y SS_A. Para SS_A se gastan (k-1) y para SS_R quedan (n-1)-(k-1) = (n-k). Con ellos y con las SS se calculan las medias cuadradas o cuadrados medios (mean squares):

$MS_A = SS_A/(k-1)$
$MS_R = SS_R/(n-k)$

Las MS son estimaciones de las varianzas, así MS_R es una estimación de σ^2, mientras que MS_A es una estimación de $\sigma^2 + n_j\sigma_a^2$ (o de $\sigma^2 + \sigma_a^2 \frac{n - (\sum n_j^2 / n)}{k-1}$ cuando los k valores n_j no son iguales)

El **test de hipótesis** o prueba de significación es una prueba F (de Fischer) de comparación de las varianzas MS_A y MS_R (que estiman respectivamente ($\sigma^2 + n_j\sigma_a^2$) y σ^2). La H_0 es que las varianzas son homogéneas (o idénticas) y la H_1 es que $MS_A > MS_R$. Para ello se calcula un valor de $F = \frac{MS_A}{MS_R} = \frac{SS_A/(k-1)}{SS_R/(n-k)}$ que se compara con el F_{crit} tabulado de una cola con k-1 y n-k grados de libertad.

Los resultados se presentan en forma de tabla

Fuente	**g.d.l**	**SS**	**MS**	**F**
Entre columnas (A)	k-1	SS_A	$SS_A/(k-1)$	MS_A/MS_R
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)=.....

Para el ejemplo

Fuente	**g.d.l**	**SS**	**MS**	**F**
Entre columnas (A)	6	2,6834	0,4472	23,1529
Dentro de columnas (residual)	35	0,6761	0,0193	
Total	41	3,3595		

F_{crit} (0,05,6,35)= 2,38

Además del F_{crit} se puede calcular el p 'a posteriori' (Cf. Lección 3), que en este caso vale $7.727.10^{-11}$.

2.3 Suposiciones implícitas

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

Una forma de comprobarla es hacer una inspección visual de los datos antes del ANOVA. Un auxiliar muy potente son los box-plots, pero existen otras pruebas para comprobar que las varianzas son similares.

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

2.4 Modelos de efectos fijos y efectos aleatorios

Hasta ahora hemos supuesto que los efectos son aleatorios. La diferencia no afecta ni a los experimentos reales ni al test F, pero los objetivos del ANOVA son diferentes.

Puesto que $x_{ij} = \mu + a_j + e_{ij}$, cada resultado experimental se descompone en varios componentes, uno de los cuales es el efecto del factor a_j. Hay dos posibilidades:

1) **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. Es el caso del ejemplo inicial en la que se estudia el efecto de la mineralización. En este caso cada resultado consiste por un lado en $\mu + a_j$ (la media + el efecto del pretratamiento) y por otro de los errores aleatorios o residuales e_i:

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

En rigor MS_A estima $\sigma^2 + \frac{\sum_j n_j a_j^2}{k-1}$ y la hipótesis a comprobar es:

$H_0 : a_1 = a_2 = ... = a_k = 0$
$H_1 : a_j \neq 0$ **para al menos** un j

Esto no tiene consecuencias para los cálculos. En este modelo se rechaza H_0, si **al menos una columna tiene un valor medio diferente del resto**, es decir es significativamente diferente de las otras (al menos un método de pretratamiento es diferente de los otros).

En ocasiones se **desea saber qué columna es la diferente y cuál es la cuantía de la diferencia**. El método más simple por su sencillez es el de las diferencia menos significativa (LSD), y es parecido a las pruebas t de comparación de medias. Para ello podría calcularse un valor de t, para cada pareja de columnas:

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{MS_R\left[(1/n_1) + (1/n_2)\right]}} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{MS_R\left[2/n_j\right]}} \text{ si } n_1 = n_2$$ que se compararía con el t crítico.

La prueba se hace en la práctica chequeando cada par $\left|\bar{x}_1 - \bar{x}_2\right|$ frente a LSD, calculado como $LSD = t_{crit}\sqrt{MS_R(2/n_j)}$ donde t_{crit} se calcula al nivel de significación que se desea y con un número de g.d.l. igual al que se utilizó para calcular MS_R.

Ejemplo:

Se tienen los resultados correspondientes a la normalización de una disolución de hidróxido sódico por 5 alumnos (volumen en ml de reactivo gastado), cada uno de los cuales valora con dicho hidróxido sódico y por triplicado una misma disolución patrón preparada con ftalato ácido de potasio tipo primario. A la vista de los resultados indique si alguno de los alumnos origina resultados significativamente diferentes del resto, y en caso afirmativo identifíquele.

	Alumno				
	1	**2**	**3**	**4**	**5**
	10,2	9,8	9,7	9,6	9,5
	10,1	9,7	9,3	9,6	9,5
	10,9	9,7	9,8	9,7	9,5
Media	10,40	9,73	9,60	9,63	9,50

ANÁLISIS DE VARIANZA

Fuente	SS	g.d.l.	MS	F_{cal}	$F_{crí}$	*p*
Entre grupos (A)	1,556	4	0,3890	7,294	3,478	0,0051
Dentro grupos (R)	0,533	10	0,0533			
Total	2,089	14				

Modelo de efectos fijos t 2,228
LSD 0,420

Se observa que el factor (distinto alumno) sí que tiene efecto significativo sobre los resultados. Al comparar las diferencias existentes entre las medias de los alumnos, se encuentra que el alumno número 1 origina diferencias superiores a la LSD en todos los casos, por lo que dicho alumno es el que, en principio, origina las diferencias. Una vez eliminado, se debe realizar de nuevo el ANOVA.

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

En este caso no estamos interesados en un efecto específico debido a una cierta columna, sino en un efecto general sobre todas las columnas y en que dicho efecto esté normalmente distribuido.

Las estimaciones son las indicadas anteriormente, así MS_A estima $\sigma^2 + \sigma_a^2 \frac{n - (\sum n_j^2 / n)}{k - 1}$ o para iguales n_j : $\sigma^2 + n_j \sigma_a^2$.

Puesto que el efecto es aleatorio no tiene sentido conocer qué media columnar es significativamente diferente de las otras, pero si el efecto existe hay que conocer su **cuantía**. Cuando todos los n_j son idénticos, MS_A estimaba $\sigma^2 + n_j \sigma_a^2$, mientras MS_R estima σ^2. Por tanto la varianza debida a la heterogeneidad de la muestra es:

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

La diferencia entre ambos modelos no es siempre evidente: En un ejercicio de inter-comparación podríamos estar interesados en comprobar si todos los laboratorios trabajan bien (proficiency testing) en cuyo caso se trataría de un modelo de efectos fijos. Por otro lado, podemos suponer que todos los laboratorios trabajan bien, por lo que si el material es homogéneo 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, En este caso es un modelo de efectos aleatorios.

La diferencia es solo de tipo filosófico y no afecta a los cálculos

Ejemplo

Se desea comprobar la homogeneidad de las diferentes partidas de vinagre de una determinada marca comercial. Para ello se analizan por quintuplicado y en condiciones homogéneas (mismo analista, mismo método y en corto espacio de tiempo) muestras procedentes de 4 partidas distintas. A la vista de los resultados (grados de acidez) demuestre si el vinagre es o no homogéneo entre partidas.

Botella			
1	2	3	4
5,95	5,98	5,94	6,00
5,98	5,99	5,95	5,99
5,97	6,00	5,93	6,01
5,95	5,97	5,95	6,01
5,96	5,99	5,93	5,99

ANÁLISIS DE VARIANZA

Fuente	SS	g.d.l.	MS	$F_{calculado}$	$F_{crítico}$	*p*
Entre grupos (A)	0,01051	3	0,0035029	28,212	3,239	1,263E-06
Dentro grupos (R)	0,00199	16	0,0001242			
Total	0,01250	19				

Modelo de efectos aleatorios

s_A^2 **0,0006757** s_A **0,0260**

s_R^2 **0,0001242** s_R **0,0111**

El material ha resultado ser heterogéneo y la varianza introducida por ese efecto es bastante mayor (más de cinco veces) que la debida a la determinación experimental.

NOTA: En este ejemplo, si el material fuese en realidad homogéneo, el que se rechace la H0 podría deberse al muestreo o al pretratamiento (pipeteado, enrasado) de la muestra. Para encontrar la verdadera causa, habría que estudiar el problema por medio de ANOVA's encajados (véase más adelante).

Existe una situación en procedimientos de validación ANOVA de efectos aleatorios que es la determinación de la **precisión intermedia** (cf. Lección 2). Para ello se analizan varias réplicas de una misma muestra en diferentes condiciones, p.e. en varias sesiones, y los resultados se estudian por medio de un ANOVA de efectos aleatorios.

Por ejemplo, supongamos que se analizan por duplicado muestras de un M.R.C. con un método exento de bias y en condiciones homogéneas a lo largo de 5 sesiones, obteniéndose los siguientes resultados

Sesión				
1	**2**	**3**	**4**	**5**
5,95	6,01	6,27	5,89	5,91
5,98	6,02	6,29	5,87	5,89

ANÁLISIS DE VARIANZA

Fuente	SS	g.d.l.	MS	F_{cal}	$F_{crít}$	*p*
Entre sesiones (A)	0,2079	4	0,05197	236,20	5,19	$7{,}01.10^{-06}$
Dentro sesiones (R)	0,0011	5	0,00022			
Total	0,2090	9				

Modelo de efectos aleatorios

s_A^2 : 0,0259

s_R^2 : 0,00022

El valor s_R **es la repetitividad**, mientras que la **precisión intermedia** se calcula combinando s_A^2 y s_R^2 de acuerdo a:

$$\text{precisión intermedia} = \sqrt{s_A^2 + s_R^2}$$

3 ANOVA DE DOS VÍAS Y MULTI-VÍA

3.1 Fuentes de varianza y significación

En ocasiones se desean tener en cuenta **dos o más factores**. Por ejemplo, podemos estar estudiando el efecto que tiene sobre una la determinación de Cu en suelos tres procedimientos diferentes de mineralización y, simultáneamente, varias formas de desecar la muestra, por ejemplo al aire o en estufa. Los datos podrían ser:

	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

De manera más general, la tabla anterior se puede poner:

	FACTOR B						
FACTOR A	1	2	....	j	....	k	**Medias factor A**
1	x_{11}	x_{21}	....	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_{ij}	....	x_{lk}	$\bar{x}_{l.}$
Medias Factor B	$\bar{x}_{.1}$	$\bar{x}_{.2}$	....	$\bar{x}_{.j}$	....	$\bar{x}_{.k}$	**Gran media** $\bar{x}$

Las tablas de este estilo 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 doble clasificación. Cada intersección se denomina **celda** y puede contener uno o más datos. Si hay **replicación** cada celda contiene más de un dato (en el ejemplo, cada celda contiene 6 datos).

3.2 La tabla del ANOVA de dos vías

La tabla sigue un modelo lineal:

$$x_{hj} = \mu + a_h + b_j + e_{hj}$$

es decir que cada valor viene afectado por el efecto del factor a, el efecto del factor b y queda un residual que debería ser aleatorio.

Supongamos primero que no hay replicación, es decir que solo hay un resultado en cada celda, como en la tabla general.

La gran media de los datos es:

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

Hay l niveles del factor A y la media de cada uno de estos niveles está dada por $\bar{x}_{1.}, \bar{x}_{2.}, \ldots \bar{x}_{h.}, \ldots \bar{x}_{l.}$, tal que

$$\bar{x}_{h.} = \sum_j x_{hj} / k$$

Similarmente hay k niveles del factor B y la media de cada nivel viene dada por

$$\bar{x}_{.j} = \sum_h x_{hj} / l$$

La SS_T se obtiene de forma similar a como antes y puede dividirse también en diversos componentes debidos a los diferentes factores y a los residuales

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

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

Para el factor A

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

$$\bar{x}_{h.} = \sum_j x_{hj} / k$$

Similarmente hay k niveles del factor B y la media de cada nivel viene dada por

$$\bar{x}_{.j} = \sum_h x_{hj} / l$$

Para el factor B

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

y para los residuales

$SS_R = SS_T - SS_A - SS_B$; g.d.l. = (kl - 1) - (k - 1) - (l - 1) ; $MS_R = SS_R$/g.d.l.

Puede demostrarse que:

$$SS_R = \sum_h \sum_j (x_{hj} - \bar{x}_{h.} - \bar{x}_{.j} + \bar{x})^2 \; ; \; \text{g.d.l.} = (k - 1)(l - 1) \; ; \; MS_R = SS_R/(k - 1)(l - 1)$$

Se puede distinguir también aquí entre modelos de efectos fijos y aleatorios, y existe también un modelo mixto en el cual un factor es de efectos fijo y otro aleatorio.

Cuando se hacen réplicas, en cada celda existe más de un valor. En nuestro ejemplo numérico hay 3x2 celdas y cada una contiene además 6 réplicas. Aunque no es preciso que todas las celdas contengan el mismo número de réplicas, debería evitarse el que haya grandes diferencias. A partir de los valores anteriores se construye una tabla del ANOVA que es similar a la del ANOVA de una vía y que resulta ser:

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	
Total	$n_j$kl-1=t	SS_T		

Se hacen dos pruebas F de 1 cola con los correspondientes g.d.l. de numerador y denominador para comprobar si los efectos de los factores estudiados A y B son estadísticamente significativos

En el caso del ejemplo

Fuente	**g.d.l**	**SS**	**MS**	**F**
Efectos globales	3	1,7501		
Factor A	1	0,5041	0,5041	10,34
Factor B	2	1,2460	0,6230	12,78
Residual	32	1,5601	0,0488	
Total	35	3,3102		

Los F críticos son: Factor A (1 y 32 g.d.l.) 4,15; Factor B (2 y 32 g.d.l.) 3,30. Luego ambos factores tienen efectos significativamente estadísticos sobre los resultados.

3.3 Interacción

En ocasiones el **efecto de uno de los factores depende del nivel del otro factor** (En el ajemplo indicado, uno de los métodos de desecado de la muestra de suelo podría afectar de diferente forma a los procedimientos de determinación de Cu). Este fenómeno se denomina interacción. La forma de considerarlo, consiste en añadir otro término adicional como fuente de varianza, o sea aumentar el modelo lineal subyacente:

$$x_{hj} = \mu + a_h + b_j + (ab)_{hj} + e_{hj}$$

El número de g.d.l. necesarios para estimar el efecto de la interacción será el producto de (k -1)(l -1), y ese es precisamente el número de g.d.l. sobrantes que quedan para calcular los residuales si no se hacen réplicas (ver más arriba). Por tanto si se considera la interacción y no se hacen réplicas no quedarían g.d.l. para los residuales. En resumen, si se desea estudiar la interacción deben hacerse réplicas. Es decir alguna o todas las combinaciones de niveles de los factores debe analizarse más de una vez.

En nuestro ejemplo concreto sí que la había, luego puede calcularse (No entraremos aquí en detalles). 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) = inter	SS_{inte}	SS_{inte}/inter	MS_{inter}/MS_R
Residual	t-(k-1)-(l-1)-(l -1)(k -1)=r	SS_R	SS_R/r	
Total	$n_j$kl-1=t	SS_T		

En nuestro caso concreto:

Fuente	**g.d.l**	**SS**	**MS**	**F**
Efectos globales	3	1,7501		
Factor A	1	0,5041	0,5041	22,81
Factor B	2	1,2460	0,6230	28,19
Interacción	2	0,8962	0,4481	20,27
Residual	30	0,6639	0,0221	
Total	35	3,3102		

Luego los dos efectos, incluida la interacción son significativos.

En el caso de que la interacción se estime y no sea significativa, puede ser incorporada en el residual. Para ello simplemente se suman las SS y los g.d.l. correspondientes.

4 ANOVA encajado (nested ANOVA)

Los ANOVA encajados (nested en inglés) son de gran interés en los ensayos de validación, ya que permiten encontrar las contribuciones de los diferentes factores que pueden tener efecto sobre un resultado. De hecho, se determinan las varianzas de todos los factores, lo cual es imprescindible en ensayos de validación si se quiere determinar la incertidumbre de los resultados (que es la raiz cuadrada de la varianza combinada).

Supongamos que se quiere determinar el efecto de diferentes analistas y diferentes instrumentos sobre la varianza de los resultados generados. Por ejemplo, dos analista y dos instrumentos: Cada analista (1 y 2) utilizaría los dos instrumentos (A y B) para hacer determinaciones replicadas de una misma muestra. De forma jerárquica:

- Analista 1
  - Instrumento A
    - Réplica 1
    - .....
    - Réplica n
  - Instrumento B
    - Réplica 1
    - .....
    - Réplica n
- Analista 2
  - Instrumento A
    - Réplica 1
    - .....
    - Réplica n
  - Instrumento B
    - Réplica 1
    - .....
    - Réplica n

Si además hubiera dos Laboratorios L1 y L2, habría que repetirlo todo en los dos. Se tiene siempre un esquema jerárquico en forma de árbol que se va ramificando: dos laboratorios, y en cada uno dos analistas, y en cada uno con dos instrumentos, hacen varias réplicas de la determinación analítica de un mismo material homogéneo.

Un ejemplo típico (y complicado) podría ser: En ocho laboratorios (a=8), dos analistas diferentes (b=2) analizan en tres sesiones diferentes (c=3), dos réplicas (n=3) de un mismo material. Se puede hacer un ANOVA encajado. El reparto de las SS aditivas sería:

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

x_{ijkl} el valor obtenido en el laboratorio i, por el analista j, el día k para la réplica l

$\bar{x}_{ijk}$, la media del laboratorio i por el analista j para el día k

$\bar{x}_{ij}$, la media del laboratorio i por el analista j

$\bar{x}_{i}$, la media del laboratorio i

$\bar{\bar{x}}$, 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 resultará finalmente:

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

En esa expresión se cumple que:

$$SS_A = bcn\sum_{i=1}^{a}\left(\bar{x}_i - \bar{\bar{x}}\right)^2$$

$$SS_B = cn\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{x}_{ij} - \bar{x}_i\right)^2$$

$$SS_C = n\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(\bar{x}_{ijk} - \bar{x}_{ij}\right)^2$$

Es decir: $SS_T = SS_A + SS_B + SS_C + SS_D$

Los g.d.l se reparten de forma progresiva a partir de A, y avanzando de forma jerárquica hasta llegar a T. Los resultados se muestran en la Tabla del ANOVA correspondiente:

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_B
Réplicas dentro de días (D)	abc(n-1)	SS_D	MS_D	
Total	abcn-1	SS_T		

Cada valor de F se obtiene dividiendo cada estimación MS por la que tiene inmediatamente debajo. Lo más importante es ver los valores de las varianzas correspondientes a cada unoi 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 y 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.

En los estudios interlaboratorio se suele hacer un diseño de este estilo: Se analiza un material **homogéneo**, que se reparte entre los **a** laboratorios. Cada laboratorio hace **b** determinaciones analíticas, de manera que obtenga **n** réplicas. Hay pues tres niveles y el reparto se basa en el desarrollo de:

$$e_{ij} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x}) + (\bar{x} - \mu_0)$$

(Cf. Lección 2). Esto origina, si empleamos la misma nomenclatura del desarrollo inicial:

$$SS_T = SS_A + SS_B + SS_C$$

de donde se pueden estimar σ_C^2 , σ_B^2 y σ_A^2 que serán respectivamente los componentes de la repetitividad (σ_C), la incertidumbre causada por los bias del laboratorio (σ_B), y de la incertidumbre debida al bias del método (σ_A). La reproducibilidad se podrá estimar a partir de $\sqrt{\sigma_C^2 + \sigma_B^2}$.(A modo de ejercicio deduzca la tabla del correspondiente ANOVA encajado).

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...).

5 BIBLIOGRAFÍA

Massart D.L., Vandegintse B.G.M., Buydens L.M.C., De Jong S., Lewi P.J. and Smeyers-Verbeke J., *Handbook of Chemometrics and Qualimetrics*. Elsevier, Amsterdam, 1997

6 EJERCICIOS PROPUESTOS

1.- La Tabla siguiente presenta los valores de la concentración de una disolución reserva de hidróxido sódico originados por 20 alumnos repartidos en 4 mesas, en cada una de las cuales se preparó una disolución diferente de ftalato ácido de potasio tipo primario. Demuestre si existen diferencias significativas entre los resultados obtenidos.

1	2	3	4
0,10043	0,10145	0,10525	0,10297
0,10832	0,10683	0,10383	0,10297
0,09974	0,10145	0,10415	0,10297
0,09847	0,10004	0,10308	0,10223
0,10175	0,10217	0,10383	0,10275

2.- La disolución anteriormente normalizada fue empleada por cada uno de los veinte alumnos para determinar el grado de acidez de una muestra de vinagre comercial. A partir de los resultados generados, demuestre si existen diferencias significativas entre las mesas.

1	2	3	4
5,95	6,01	6,27	5,89
6,37	5,94	6,30	5,95
5,89	5,97	6,31	6,24
5,85	5,90	6,19	5,93
6,25	5,95	6,30	6,01

3.- Durante un ejercicio de inter-comparación se ha determinado el fósforo extraíble en HCl (mg/kg). Demuestre si el material es homogéneo o no.

BO1291	**BO1292**	**BO1293**	**BO1294**
39,3	38,7	38,9	39,3
40,6	39,9	39,8	38,7
39,9	39,7	39,7	39,1

4.- Se desea evaluar la repetitividad y la precisión intermedia de un laboratorio que se está acreditando. Para ello se analizan en cinco sesiones diferentes dos réplicas de un M.R.C., obteniéndose los siguientes resultados en mg/L:

Sesión				
1	**2**	**3**	**4**	**5**
5,95	5,98	6,01	5,99	5,99
5,98	6,00	5,98	5,97	6,00

5.- El contenido en cobre de muestras de suelos puede ser determinado mediante tres procedimientos diferentes: ICP, ETAAS y ASV, que pueden ser aplicados a muestras secas al aire o desecadas en estufa. A partir de los datos de la Tabla demuestre si el método de análisis y/o el método de secado ejercen influencia sobre el resultado final

	Método de análisis		
Método de secado	ICP	ETAAS	ASV
Aire	5,48	5,58	5,47
Estufa	5,34	5,55	5,43

6.- Los mismos métodos de análisis y secado se aplican ahora a 6 réplicas, de manera que se obtienen los siguientes resultados experimentales. Demuestre si el método de análisis y/o el método de secado ejercen influencia sobre el resultado final, así como si ele efecto de alguno de los factores estudiados depende del nivel del otro (es decir si hay interacción)

	Método de análisis		
Método de secado	**ICP**	**ETAAS**	**ASV**
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
Estufa	5,90	5,90	5,81
	5,75	6,01	5,90
	6,07	5,85	5,81
	5,90	5,54	5,81
	6,01	5,81	5,90
	6,06	5,70	5,90

6.- Se desea evaluar las características de un nuevo procedimiento de análisis. Para ello se reparte un material homogéneo entre 3 laboratorios. Cada laboratorio aplica el método a 4 sub-muestras (variando aleatoriamente los analistas, el día, las disoluciones etc.), analizando 3 réplicas de cada una de ellas. Los resultados aparecen en la siguiente tabla. Dibuje el esquema del ANOVA encajado resultante, deduzca las fórmulas, identifique los componentes que pueden determinarse (repetitividad, bias...) y calcule sus estimaciones.

Laboratorio 1			
1	**2**	**3**	**4**
35.90	36.74	36.53	35.56
34.89	35.96	35.87	35.71
36.10	35.84	35.53	35.34

Laboratorio 2			
1	**2**	**3**	**4**
37.60	37.84	37.92	38.12
37.89	37.34	38.04	38.03
37.64	37.88	37.99	38.08

Laboratorio 3			
1	**2**	**3**	**4**
36.56	35.87	35.78	35.56
35.67	35.54	35.83	35.78
35.97	35.75	35.12	35.98