EL ANÁLISIS DE LA VARIANZA (ANOVA)
1. Comparación de múltiples poblaciones

Ricard Boqué, Alicia Maroto
Grupo de Quimiometría y Cualimetría. Universitat Rovira i Virgili.
Pl. Imperial Tàrraco, 1. 43005-Tarragona

El análisis de la varianza (ANOVA) es una potente herramienta estadística, de gran utilidad tanto en la industria, para el control de procesos, como en el laboratorio de análisis, para el control de métodos analíticos. Los ejemplos de aplicación son múltiples, pudiéndose agrupar, según el objetivo que persiguen, en dos principalmente: la comparación de múltiples columnas de datos y la estimación de los componentes de variación de un proceso. Nos ocupamos en este artículo de la primera de ellas.

Comparación de múltiples poblaciones

La comparación de diversos conjuntos de resultados es habitual en los laboratorios analíticos. Así, por ejemplo, puede interesar comparar diversos métodos de análisis con diferentes características, diversos analistas entre sí, o una serie de laboratorios que analizan una misma muestra con el mismo método (ensayos colaborativos). También sería el caso cuando queremos analizar una muestra que ha estado sometida a diferentes tratamientos o ha estado almacenada en diferentes condiciones. En todos estos ejemplos hay dos posibles fuentes de variación: una es el error aleatorio en la medida y la otra es lo que se denomina *factor controlado* (tipo de método, diferentes condiciones, analista o laboratorio,...). Una de las herramientas estadísticas más utilizadas que permite la separación de las diversas fuentes de variación es el análisis de la varianza (ANOVA, del inglés *Analysis of Variance*) [Massart, 1997].

El ANOVA también puede utilizarse en situaciones donde ambas fuentes de variación son aleatorias. Un ejemplo sería el análisis de algún compuesto de un vino almacenado en un depósito. Supongamos que las muestras se toman aleatoriamente de diferentes partes del depósito y se realizan diversos análisis replicados. Aparte de la variación natural en la medida tendremos una variación en la composición del vino de les diferentes partes del depósito.

Cuando tengamos un factor, controlado o aleatorio, aparte del error propio de la medida, hablaremos del ANOVA de un factor. En el caso de que estuviésemos desarrollando un nuevo método colorimétrico y quisiéramos investigar la influencia de diversos factores independientes sobre la absorbancia, tales como la concentración de reactivo A y la temperatura a la que tiene lugar la reacción, entonces hablaríamos de un ANOVA de dos factores. En los casos donde tenemos dos o más factores que influyen, se realizan los experimentos para todas las combinaciones de los factores estudiados, seguido del ANOVA. Se puede deducir entonces si cada uno de los factores o una interacción entre ellos tienen influencia significativa en el resultado.

Para utilizar el ANOVA de forma satisfactoria deben cumplirse tres tipos de hipótesis, aunque se aceptan ligeras desviaciones de las condiciones ideales:

1. Cada conjunto de datos debe ser independiente del resto.
2. Los resultados obtenidos para cada conjunto deben seguir una distribución normal.
3. Las varianzas de cada conjunto de datos no deben diferir de forma significativa.

ANOVA de un factor

Tomemos como ejemplo la comparación de 5 laboratorios que analizan n_k veces con el mismo procedimiento la concentración de Pb en una misma muestra de agua de río. El objetivo del ANOVA aquí es comparar los errores sistemáticos con los aleatorios obtenidos al realizar diversos análisis en cada laboratorio. Hemos comentado antes que son condiciones importantes que cada laboratorio analice sus muestras de manera independiente y con precisiones parecidas a las del resto de laboratorios. En la tabla 1 se muestran los resultados obtenidos (expresados en μg/L).

Tabla 1. Resultados del análisis de plomo en agua de río realizado por 5 laboratorios (*k* indica el nº de laboratorio).

Resultados	Laboratorio A	Laboratorio B	Laboratorio C	Laboratorio D	Laboratorio E
1	2.3	6.5	1.7	2.1	8.5
2	4.1	4.0	2.7	3.8	5.5
3	4.9	4.2	4.1	4.8	6.1
4	2.5	6.3	1.6	2.8	8.2
5	3.1	4.4	4.1	4.8	-
6	3.7	-	2.8	3.7	-
7	-	-	-	4.2	-
Suma	20.6	25.4	17.0	26.2	28.3
Valor medio, $\overline{x}_k$	3.4	5.1	2.8	3.7	7.1
n_k	6	5	6	7	4
Media aritmética de todos los resultados, $\overline{\overline{x}} = 4.2$					
Número total de resultados, $N = 28$					

Observando los valores medios todo parece indicar que existen diferencias entre los laboratorios. Ahora bien, ¿son dichas diferencias significativas? El ANOVA responde a esta cuestión. El objetivo del ANOVA es comparar los diversos valores medios para determinar si alguno de ellos difiere significativamente del resto. Para ello se utiliza una estrategia bien lógica: si los resultados proporcionados por los diversos laboratorios no contienen errores sistemáticos, los valores medios respectivos no diferirán mucho los unos de los otros y su dispersión, debida a los errores aleatorios, será comparable a la dispersión presente individualmente en cada laboratorio.

El secreto está, pues, en descomponer la variabilidad total de los datos en dos fuentes de variación: la debida a los laboratorios y la debida a la precisión dentro de cada laboratorio. Matemáticamente, la suma de cuadrados total, SS_T, puede descomponerse como una suma de dos sumas de cuadrados:

$$SS_T = SS_R + SS_{lab}$$

SS_T es la suma de las diferencias al cuadrado de cada resultado individual respecto a la media de todos los resultados y por tanto, representa la variación total de los datos. SS_R mide las desviaciones entre los resultados individuales (x_{kj}), de cada laboratorio (donde *j* indica el nº de repetición) y la media del laboratorio ($\overline{x}_k$) y, por lo tanto, es una medida de la dispersión dentro de los laboratorios. Cuando se divide SS_R por los correspondientes grados de libertad, ($N - K$), se obtiene el cuadrado medio (o *MS*, del inglés *Mean Square*) "dentro de los laboratorios", MS_R.

Por su lado, SS_{lab} mide las desviaciones entre los resultados medios de los laboratorios y el resultado medio global y, dividido por sus grados de libertad, (k - 1), constituye el cuadrado medio "entre laboratorios", MS_{lab}. La Tabla 2 muestra las diferentes expresiones para calcular las sumas de cuadrados y las correspondientes varianzas.

Tabla 2 Expresiones para el cálculo del ANOVA de un factor (K indica el número de laboratorios y N el número total de resultados).

Fuente	Suma de cuadrados	Grados de libertad	Varianza	F_{cal}
Entre laboratorios	$SS_{lab} = \sum_{k=1}^{K} n_k(\bar{x}_k - \bar{\bar{x}})^2$	K - 1	$MS_{lab} = \frac{SS_{lab}}{K\text{ -}1}$	$F = \frac{MS_{lab}}{MS_R}$
Dentro de los laboratorios	$SS_R = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(x_{kj} - \bar{x}_k)^2$	$N - K$	$MS_R = \frac{SS_R}{N - K}$	
Total	$SS_T = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(x_{kj} - \bar{\bar{x}})^2$	N - 1	$MS_T = \frac{SS_T}{N\text{ -}1}$	

Se calculan, por tanto, MS_{lab} y MS_R como una medida de las dispersiones comentadas y se comparan mediante una prueba de hipótesis F. Si no existe diferencia estadísticamente significativa entre ellas, la presencia de errores aleatorios será la causa predominante de la discrepancia entre los valores medios. Si, por el contrario, existe algún error sistemático, MS_{lab} será mucho mayor que MS_R, con lo cual el valor calculado de F será mayor que el valor tabulado F_{tab} para el nivel de significación α escogido y los grados de libertad mencionados.

A continuación se muestra la típica tabla ANOVA obtenida para los resultados del ejemplo de la Tabla 1:

Tabla 3. Tabla ANOVA para los resultados de la Tabla 1.

Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_{cal}	*Probabilidad*
Entre laboratorios	53.13	4	13.28	10.30	$6.23 \cdot 10^{-5}$
Dentro de los laboratorios	29.64	23	1.29		
Total	82.77	27			
F_{tab} = 2.79 (α = 0.05, 4, 23, 1 cola)					

Como $F_{cal} > F_{tab}$, en este caso se podría concluir que al menos uno de los laboratorios ha producido resultados la media de los cuales difiere de forma estadísticamente significativa del resto de laboratorios. El valor de probabilidad que aparece en la Tabla 3 indica aquel valor de α a partir del cual el ANOVA no detectaría ninguna diferencia significativa. Así pues, a menor valor de probabilidad, mayor seguridad de que existen diferencias significativas.

El ANOVA no indica cuántos laboratorios difieren ni cuáles son. Una inspección visual de los resultados puede proporcionar sin duda alguna pista, pero si se quieren tener criterios más sólidos, hay diversas pruebas estadísticas que permiten saber de qué laboratorios se trata [Massart, 1997].

En el ejemplo que hemos presentado, todos los laboratorios han analizado la muestra siguiendo un procedimiento analítico común. Se hubiese podido plantear que cada laboratorio utilizase dos procedimientos comunes, por ejemplo el método oficial y un método alternativo. En este caso dispondríamos de los resultados del contenido en plomo obtenidos por una serie de laboratorios con dos métodos distintos, y el ANOVA nos proporcionaría información sobre la existencia de discrepancias entre laboratorios y entre métodos. Sería un ejemplo de ANOVA de dos factores.

Conclusiones

En este artículo hemos visto que el ANOVA puede utilizarse para comparar entre sí las medias de los resultados obtenidos por diversos laboratorios, analistas, métodos de análisis, etc. En el siguiente artículo mostraremos cómo utilizar el ANOVA para descomponer la variación total de un proceso en las fuentes de variación parciales. Esto nos puede resultar muy útil para, por ejemplo, determinar cuáles son los factores que afectan más a un determinado procedimiento analítico.

Desde el punto de vista práctico, existen múltiples paquetes estadísticos que permiten ejecutar rápidamente los cálculos del ANOVA. Lo que es interesante, sin embargo, es que el usuario tenga capacidad para extraer conclusiones químicas de los resultados obtenidos.

Referencias bibliográficas

D.L. Massart, B.M.G. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, “Handbook of Chemometrics and Qualimetrics: Part A”, Elsevier (1997), Amsterdam.

Los autores agradecen todos los comentarios relacionados con los contenidos de este artículo. Pueden dirigirse, mediante mensaje electrónico, a la dirección: quimio@quimica.urv.es.

Una versión en soporte electrónico de este artículo e información suplementaria puede encontrarse en:
http://www.quimica.urv.es/quimio