Análisis de la varianza

Preparado por Luis M. Molinero (Alce Ingeniería)
CorreoE: **bioestadistica@alceingenieria.net**

Artículo en formato PDF

www.seh−lelha.org/stat1.htm

Junio 2003

Las técnicas englobadas bajo la denominación de análisis de la varianza o abreviadamente ANOVA (del inglés analysis of variance) han jugado un papel crucial en la metodología estadística moderna, desde que fueran ideadas por R.A. Fisher en 1925, y como sucede en tantas ocasiones, aunque conocidas por la gran mayoría, quizás no son adecuadamente comprendidas por los no especialistas.

Casi siempre se introduce el tema del análisis de la varianza como respuesta a la necesidad de utilizar una técnica de comparación de más de dos grupos, es decir como un método para comparar más de dos tratamientos: si disponemos de medidas cuantitativas continuas, que se puede suponer como procedentes de una distribución de probabilidad normal, y queremos comparar dos grupos −dos tratamientos−, la prueba estadística que se utiliza es un contraste de medias basado en la t de Student, y cuando se dispone de más de dos grupos, la prueba a emplear es el análisis de la varianza. Personalmente, aunque el enfoque es adecuado, me parece que refleja solo una parte del interés de la técnica, ideada no sólo para analizar los datos sino también para planificar los experimentos, y creo más apropiado hablar de que el análisis de la varianza es un procedimiento estadístico que nos permite dividir la variabilidad observada en componentes independientes que pueden atribuirse a diferentes causas de interés.

En el planteamiento más simple de análisis de la varianza tenemos una variable numérica cuantitativa (resultado), y queremos determinar en qué medida se puede atribuir la variabilidad de ésta a otra variable cualitativa nominal que vamos a denominar **factor**. Estamos hablando por tanto de análisis de la varianza para un solo factor, que puede tener 2 o más categorías o niveles.

Este factor, cuyo posible efecto sobre la variable medida queremos analizar, puede tener unos niveles fijos, por ejemplo el nivel educativo alcanzado por los sujetos que intervienen (sin estudios, estudios primarios, secundarios, formación universitaria), y hablamos entonces de **modelo de efectos fijos**; o bien puede tratarse de una muestra procedente de un conjunto de niveles más amplio, como puede ser por ejemplo el caso de un estudio en el que se seleccionan varios hospitales y se analiza las posibles diferencias entre hospitales. Entonces lo denominamos **modelo de efectos aleatorios**. En el análisis de la varianza de 1 factor es mucho más frecuente el modelo de efectos fijos.

Vamos a plantear el problema y comentar los cálculos que se efectúan en un análisis de la varianza para un factor. Estudiamos *K* grupos clasificados de acuerdo a los niveles *1,2 .. K* del factor. En cada nivel tenemos *n1, n2, ... nk* observaciones independientes y obtenidas de forma aleatoria. Si designamos de forma general cada observación como y_{ij}, el subíndice *i* indica el grupo al que pertenece, *j* es el número de la observación dentro de ese grupo, de tal manera que por ejemplo y_{25} corresponderá al valor observado en el quinto sujeto del segundo grupo. Por tanto en el grupo 2 tenemos las observaciones y_{21} hasta y_{2n_2}.

Si juntamos todas las observaciones *N=n1+n2+...+nk*, calculamos la media global que vamos a denominar $\bar{y}$.

También podemos calcular la media dentro de cada uno de los *K* grupos. La media para el grupo *i* la designamos como $\bar{y}_i$.

Es obvio que la diferencia entre cada observación y la media global se puede descomponer de la siguiente forma:

$$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}) \quad [1]$$

Es decir que la diferencia entre el valor observado y la media global es igual a la suma de la diferencia de la observación con la media de su grupo y la diferencia de la media del grupo con la media global.

Se puede comprobar que si cada término de esa expresión se eleva al cuadrado y se suma para todas las observaciones, se mantiene la igualdad, lo que curiosamente no es más que la aplicación del famoso teorema de Pitágoras a este diseño:

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - \bar{y})^2$$

Cada uno de los términos es pues una suma de desviaciones cuadráticas, que denominaremos de forma abreviada como *suma de cuadrados (SC)*. La primera *SC* del lado de la derecha corresponde a las desviaciones de cada observación respecto de la media de su propio grupo, por lo que se la conoce como "*dentro del grupo*" o "*intra grupo*" (en inglés *within*). El segundo sumando de la derecha corresponde a las desviaciones de la media de cada grupo respecto de la media global, por lo que cuantifica las diferencias medias entre los grupos, y se conoce como suma de cuadrados "*entre grupos*" (en inglés *between*):

$$SC_{\boldsymbol{T}otal} = SC_{\boldsymbol{I}ntra\ grupo} + SC_{\boldsymbol{E}ntre\ grupos}$$

El *cuadrado medio intra−grupo*, equivalente a una varianza, lo calculamos dividiendo la suma de cuadrados entre los grados de libertad

$$MS_I = \frac{SC_I}{N - K}$$

y se puede comprobar que es en realidad una media ponderada de las varianzas muestrales de cada grupo, con la siguiente expresión:

$$MS_I = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \ldots + (n_K - 1)S_K^2}{n_1 + n_2 + \ldots + n_K - K}$$

Queda claro que constituye por tanto una estimación de la varianza común σ^2.

De igual manera podemos calcular el *cuadrado medio entre grupos*:

$$MS_E = \frac{SC_E}{K - 1}$$

Si la media de todos los grupos es la misma, MS_E también es una estimación de la varianza común σ^2. Esto se puede entender mejor de una forma intuitiva si consideramos el caso particular en el que todos los grupos tienen el mismo tamaño *n*. Sabemos que la desviación estándar al cuadrado (varianza) de la media obtenida en muestras de tamaño *n* extraídas de una población normal es σ^2/n (es lo que conocemos como error estándar de la media), por lo tanto $\sum (\bar{y}_i - \bar{y})^2 / (K - 1)$ será una estimación de σ^2/n y por tanto $\sum n(\bar{y}_i - \bar{y})^2 / (K - 1)$ es una estimación de σ^2.

Ahora bien, si las medias de los grupos sí son diferentes, MS_E no sólo contiene el valor de la varianza intrínseca σ^2, sino que además estará aumentada según las variaciones entre las medias de los tratamientos, y será tanto mayor cuanto mayor sean estas diferencias. El cociente:

$$F = \frac{MS_E}{MS_I}$$

que compara la variabilidad *entre grupos* y la variabilidad *intra grupos*, será por tanto próximo a 1 si las medias de los grupos son similares y tanto mayor que 1 cuanto mayores sean las diferencias entre los grupos. El valor de F obtenido se contrastará con el valor de la distribución teórica con grados de libertad $K-1, N-K$, y si la probabilidad de obtener un valor tan grande como el observado es baja, rechazaremos la hipótesis de igualdad de medias entre los grupos. La utilización de este parámetro de contraste, que tiene una rigurosa justificación metodológica estadística, también tiene pues una interpretación intuitiva: estamos comparando la variabilidad entre los grupos con la variabilidad intrínseca dentro de los grupos.

Por otro lado hemos visto que la variabilidad total la hemos dividido en dos partes: una variabilidad debida o explicada por pertenecer a cada uno de los grupos o niveles del factor, y una parte de variabilidad individual, que no atribuimos a ninguna causa concreta, y que por ello se suele denominar también *variabilidad residual*. Esto podemos reflejarlo de una forma clara manipulando un poco la fórmula [1] en la que se desglosa la variabilidad de cada observación en dos términos:

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i) = \mu + \tau_i + \varepsilon_{ij} \qquad [2]$$

Es decir que el modelo postulado (término de la derecha) para nuestras observaciones corresponde a tres sumandos: una media global μ, un efecto diferencial debido a la pertenencia al grupo o tratamiento τ_i y un termino residual no explicado ε_{ij}.

Diseño en Bloques aleatorizados

En un artículo anterior se habló de la ventaja que presentan las pruebas pareadas para aumentar la eficiencia, al controlar parte de la variabilidad no atribuible al factor que estamos estudiando. Cuando se analizan más de dos niveles o grupos el concepto de prueba pareada se puede generalizar al análisis de la varianza. Aquí se denomina bloque a cada unidad de observación, y para un factor o tratamiento tenemos el siguiente diseño experimental:

	Tratamiento 1	**Tratamiento 2**	...	**Tratamiento K**
Bloque 1	Y_{11}	Y_{12}	...	Y_{1K}
Bloque 2	Y_{21}	Y_{22}	...	Y_{1K}
...	...	...	...	...
Bloque n	Y_{n1}	Y_{n2}	...	Y_{nK}

En este diseño, de manera análoga a la expresada en la fórmula [2] podemos descomponer la variabilidad individual según el siguiente modelo:

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

donde aparece un nuevo término β_j que corresponde a la variabilidad atribuida al bloque, con lo que el término ε_{ij} correspondiente a la variabilidad no explicada disminuye, obteniéndose por tanto una prueba más eficiente.

Los bloques o unidades de observación pueden ser cada paciente, un hospital, un grupo de pacientes con unas características específicas, etc. A veces también se habla de **análisis estratificado**, donde los conceptos bloque y estrato son equivalentes.

Aunque uno de los motivos fundamentales de la asignación aleatoria de los pacientes a cada grupo de tratamiento es precisamente evitar la presencia de sesgos en las características de los pacientes que puedan afectar a las diferencias de eficacia que se observen, sin embargo cuando se sabe que factores como la edad del paciente, la presencia de diabetes, antecedentes de tabaquismo, etc influyen en el resultado, puede ocurrir que finalmente por azar las proporciones de los diferentes niveles de estos factores no se repartan "equitativamente" entre los grupos de tratamiento, lo que conlleva a que los resultados queden bajo sospecha, incluso aunque después en el análisis se acuda a técnicas multivariantes para "ajustar" los resultados en función de los valores basales en los grupos, atribuyendo parte de la variación observada a esas diferencias, y corrigiendo o disminuyendo la diferencia encontrada atribuible al efecto del tratamiento. La utilización de técnicas de diseños aleatorizados en bloques y diseños factoriales nos permite anticiparnos a esa situación, por lo que han sido ampliamente empleadas no sólo en experimentación agrícola donde se originaron, sino también en farmacología y en la industria, y en mucha menor medida, por lo que se comentará más adelante, en la investigación médica clínica.

En este diseño aletorizado por bloques disponemos de dos valores de *F* para contrastar: uno relativo a la influencia del tratamiento y otro para la influencia del bloque; aunque el contraste en el que seguramente estamos interesados es solo el primero, ya que de entrada se supone que el bloque sí que influye en la variable medida y precisamente por eso se ha acudido a este tipo de diseño.

Diseños factoriales

Los denominados diseños factoriales permiten al investigador planificar un trabajo para evaluar el efecto combinado de dos o más variables de forma simultánea en el resultado medido, obteniéndose también información en cuanto a la posible **interacción** entre los diversos factores.

Así podemos extender el modelo presentado en la fórmula [2] para considerar en cada observación la influencia de dos factores que vamos a denominar *A* y *B*. Expresamos la observación número *k* en el nivel *i* del factor *A*, nivel *j* del factor *B*, como:

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

donde se ha separado en un término correspondiente a la media global, otro debido al efecto diferencial por pertenecer a un nivel determinado del factor *A*, un efecto debido al factor *B*, un efecto de la interacción entre los factores *A* y *B*, y una variabilidad residual no atribuible.

Este modelo es la base del **análisis de la varianza para dos factores**.

El problema de los diseños factoriales clásicos cuando se aplica a la investigación clínica, en la que predominan los diseños observacionales y donde casi siempre es por tanto difícil fijar el número de sujetos en cada uno de los niveles de los diferentes factores, radica en que para que sea aplicable un análisis de la varianza clásico para más de un factor, es necesario que se cumpla también la igualdad de la suma de cuadrados, y esto sólo ocurre cuando el número de sujetos por celda (llamamos celda a cada combinación de

niveles de los distintos factores) es el mismo para todas las celdas. Es decir que la igualdad:

$$SC_{Total}=SC_A+SC_B+SC_{AB}+SC_{Residual}$$

sólo es cierta cuando todas las celdas tienen el mismo número de sujetos. Si ese número no es igual no podemos aplicar el análisis de la varianza.

Afortunadamente existe una relación directa entre el modelo de efectos postulado y la regresión lineal múltiple, en la que intervendrán los factores como variables independientes. Es lo que se conoce como **modelos lineales** y serán objeto de un nuevo artículo.

Obviamente en ese modelo de regresión los factores entrarán adecuadamente codificados como **variables diseño o dummy**, procedimiento que ya fue comentado en el artículo relativo a la regresión logística.

Enlaces de interés

- Statistics Notes: Comparing several groups using analysis of variance
  Douglas G Altman and J Martin Bland
  BMJ 1996; 312: 1472−1473. [Full text]

- Análisis de la varianza
  V. Abraira. Unidad de Bioestadística clínica del Hosp. Ramón y Cajal

- ¿Cómo comparar m medias? ANOVA
  J.F. Recasens i Collado
  JANO EMC. Viernes 05 Diciembre 1997. Volumen 53 − Número 1236 p. 74

-
  - Introduction to ANOVA
  - Factorial Between−Subjects ANOVA
  - Within−Subjects ANOVA

  HyperStat Online Textbook. Texto de estadística de David M. Lane Associate Professor of Psychology, Statistics, and Administration at Rice University.

- Calculadoras on line para diferentes modelos del análisis de la varianza

  - VassarStats Statistical Computation Web Site
  - StatPoint On−Line Statistical Computing Center

- ANOVA
  G. David Garson. North Carolina State University