ANALISIS DE VARIANZA

En estadística, el análisis de la varianza (ANOVA, ANalysis Of VAriance, según terminología inglesa) es una colección de modelos estadísticos y sus procedimientos asociados, en el cual la varianza está particionada en ciertos componentes debidos a diferentes variables explicativas.

Las técnicas iniciales del análisis de varianza fueron desarrolladas por el estadístico y genetista R. A. Fisher en los años 1920 y 1930 y es algunas veces conocido como "Anova de Fisher" o "análisis de varianza de Fisher", debido al uso de la distribución F de Fisher como parte del contraste de hipótesis

El análisis de la varianza parte de los conceptos de regresión lineal.

El primer concepto fundamental es que todo valor observado puede expresarse mediante la siguiente función:

$$Y = B_0 + B_1 * X + e$$

Donde Y sería el valor observado (variable dependiente), y X el valor que toma la variable independiente.

B_0sería una constante que en la recta de regresión equivale a la ordenada en el origen, B_1es otra constante que equivale a la pendiente de la recta, y $e$es una variable aleatoria que añade a la función cierto error que desvía la puntuación observada de la puntuación pronosticada.

Por tanto, a la función de pronóstico la podemos llamar "Y prima":

$$Y' = B_0 + B_1 * X$$

Podemos resumir que las puntuaciones observadas equivalen a las puntuaciones esperadas, más el error aleatorio:

$$Y = Y' + e \quad (1.1)$$

Sabiendo este concepto, podemos operar con esta ecuación de la siguiente forma:

1) Restamos a ambos lados de la ecuación (para mantener la igualdad) la media de la variable dependiente:

$$Y - \overline{Y} = Y' + e - \overline{Y}$$

2) Substituimos el error por la ecuación resultante de despejar la ecuación 1.1:

$$e = Y - Y'$$

Por tanto...

$$Y - \overline{Y} = Y' + (Y - Y') - \overline{Y}$$

Y reorganizando la ecuación:

$$Y - \overline{Y} = (Y' - \overline{Y}) + (Y - Y')$$

Ahora hay que tener en cuenta que la media de las puntuaciones observadas es exactamente igual que la media de las puntuaciones pronosticadas:

$$\overline{Y} = \overline{Y'}$$

Por tanto:

$$Y - \overline{Y} = (Y' - \overline{Y'}) + (Y - Y')$$

Podemos ver que nos han quedado 3 puntuaciones diferenciales. Ahora las elevamos al cuadrado para que posteriormente, al hacer el sumatorio, no se anulen:

$$(Y - \overline{Y})^2 = [(Y' - \overline{Y'}) + (Y - Y')]^2$$

Y desarrollamos el cuadrado:

$$(Y - \overline{Y})^2 = (Y' - \overline{Y'})^2 + (Y - Y')^2 + 2 * (Y' - \overline{Y'})(Y - Y')$$

Podemos ver que tenemos los numeradores de las varianzas, pero al no estar divididas por el número de casos (n), las llamamos Sumas de Cuadrados., excepto en el último término, que es una Suma Cruzada de Cuadrados (el numerador de la covarianza), y la covarianza en este caso es cero (por las propiedades de la regresión lineal, la covarianza entre el error y la variable independiente es cero).

Por tanto:

$$(Y - \overline{Y})^2 = (Y' - \overline{Y}')^2 + (Y - Y')^2$$

O lo mismo que:

$$SS_{total} = SS_{fact} + SS_{error}$$

de un factor, que es el caso más sencillo, la idea básica del análisis de la varianza es comparar la variación total de un conjunto de muestras y descomponerla como:

$$SS_{total} = SS_{fact} + SS_{int}$$

Donde:

> SS_{fact}es un número real relacionado con la varianza, que mide la variación debida al "factor", "tratamiento" o tipo de situación estudiado.
> SS_{int}es un número real relacionado con la varianza, que mide la variación dentro de cada "factor", "tratamiento" o tipo de situación.

En el caso de que la diferencia debida al factor o tratamiento no sean estadísticamente significativa puede probarse que las varianzas muestrales son iguales:

$$\hat{s}_{fact} = \frac{SS_{fact}}{a-1}, \qquad \hat{s}_{int} = \frac{SS_{int}}{a(b-1)}$$

Donde:

> $a$es el número de situaciones diferentes o valores del factor se están comparando.
> $b$es el número de mediciones en cada situación se hacen o número de valores disponibles para cada valor del factor.

Así lo que un simple test a partir de la F de Snedecor puede decidir si el factor o tratamiento es estadísticamente significativo.

Visión general

Existen tres clases conceptuales de estos modelos:

1. El Modelo de efectos fijos asume que los datos provienen de poblaciones normales las cuales podrían diferir únicamente en sus medias. (Modelo 1)
2. El Modelo de efectos aleatorios asume que los datos describen una jerarquía de diferentes poblaciones cuyas diferencias

quedan restringidas por la jerarquía. Ejemplo: El experimentador ha aprendido y ha considerado en el experimento sólo tres de muchos más métodos posibles, el método de enseñanza es un factor aleatorio en el experimento. (Modelo 2)

3. El Modelo de efectos mixtos describen situaciones que éste puede tomar. Ejemplo: Si el método de enseñanza es analizado como un factor que puede influir donde están presentes ambos tipos de factores: fijos y aleatorios. (Modelo 3)

Supuestos previos

El ANOVA parte de algunos supuestos que han de cumplirse:

- La variable dependiente debe medirse al menos a nivel de intervalo.
- Independencia de las observaciones.
- La distribución de los residuales debe ser normal.
- Homocedasticidad: homogeneidad de las varianzas.

La técnica fundamental consiste en la separación de la suma de cuadrados (SS, 'sum of squares') en componentes relativos a los factores contemplados en el modelo. Como ejemplo, mostramos el modelo para un ANOVA simplificado con un tipo de factores en diferentes niveles. (Si los niveles son cuantitativos y los efectos son lineales, puede resultar apropiado un análisis de regresión lineal)

$$SS_{\text{Total}} = SS_{\text{Error}} + SS_{\text{Factores}}$$

El número de grados de libertad (gl) puede separarse de forma similar y corresponde con la forma en que la distribución chi-cuadrado (χ^2 o Ji-cuadrada) describe la suma de cuadrados asociada.

$$gl_{\text{Total}} = gl_{\text{Error}} + gl_{\text{Factores}}$$

Tipos de modelo

Modelo I: Efectos fijos

El modelo de *efectos fijos* de análisis de la varianza se aplica a situaciones en las que el experimentador ha sometido al grupo o material analizado a varios factores, cada uno de los cuales le afecta sólo a la media, permaneciendo la "variable respuesta" con una distribución normal.

Este modelo se supone cuando el investigador se interesa únicamente por los niveles del factor presentes en el experimento, por lo que cualquier variación observada en las puntuaciones se deberá al error

experimental.

Modelo II: Efectos aleatorios (componentes de varianza)

Los modelos de *efectos aleatorios* se usan para describir situaciones en que ocurren diferencias incomparables en el material o grupo experimental. El ejemplo más simple es el de estimar la media desconocida de una población compuesta de individuos diferentes y en el que esas diferencias se mezclan con los errores del instrumento de medición.

Este modelo se supone cuando el investigador está interesado en una población de niveles, teóricamente infinitos, del factor de estudio, de los que únicamente una muestra al azar (*t* niveles) están presentes en el experimento.