

Anova de una vía

Guillermo Ayala Gallego

Anova de una vía

Guillermo Ayala Gallego

2024-03-25

Análisis de la varianza de una vía

- Tenemos una variable de interés (Y) y pretendemos estudiar su posible dependencia de un factor experimental.
- El experimentador tiene un **factor** de interés con distintos niveles y se pretende evaluar su influencia en la variable respuesta.
- Si el factor tiene dos niveles entonces podemos comparar las medias mediante un test de la t (t-test).
- Con más de dos niveles del factor hemos de utilizar otras herramientas.
- Nos ocupamos (de un modo muy simple) de lo que se conoce como **experimentos con un solo factor completamente aleatorizado**.

Comparando grupos

- Tenemos (I) condiciones distintas.
- En cada una de ellas (n_i) muestras.
- $(\sum_{i=1}^I n_i = n)$ es el total de muestras.
- (Y_{ij}) denota la respuesta aleatoria en la (j) -ésima muestra de la (i) -ésima condición.

Modelo

- Suponemos que (Y_{ij}) con $(j=1, \dots, n_i)$ son independientes y con la misma distribución.
- El modelo de análisis de la varianza de una vía es
$$Y_{ij} = \beta_0 + \beta_i + \epsilon_{ij}$$
- Donde se asume que $(\epsilon_{ij} \sim N(0, \sigma^2))$ y son independientes entre si para los distintos grupos y dentro de cada grupo.
- Estamos asumiendo que $(Y_{ij} \sim N(\beta_0 + \beta_i, \sigma^2))$.

Interpretación de los parámetros

- β_0 es la media global de todos los grupos.
- β_i sería la diferencia de la media del grupo (i) respecto de esta media global.
- ϵ_{ij} sería el error aleatorio de la observación $((i,j))$ respecto del valor medio $(\beta_0 + \beta_i)$.

- Se trata de evaluar si hay diferencias entre grupos.
- Bajo el modelo que acabamos de formular se traduce en la siguiente hipótesis nula $H_0: \beta_1 = \beta_2 = \dots = \beta_I = 0$ frente a que algún par de los (β_i) sean distintos.

Contraste

- ¿Cómo podemos contrastar la hipótesis nula anterior?
- Denotamos por y_{ij} la j -ésima muestra observada bajo la condición (i) ($(i=1, \dots, I)$ y $(j=1, \dots, n_i)$).
- Denotamos las medias muestrales para cada grupo como $\bar{y}_i = \sum_{j=1}^{n_i} \frac{y_{ij}}{n_i}$.
- La media de todas las observaciones o media total como $\bar{y} = \sum_{i=1}^I \sum_{j=1}^{n_i} \frac{y_{ij}}{n}$.
- Definimos la **suma de cuadrados intra** o **del error** como $SS(\text{Within}) = \sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$, y la **suma de cuadrados entre** como $SS(\text{Between}) = \sum_{i=1}^I n_i (\bar{y}_i - \bar{y})^2$.
- El estadístico para contrastar esta hipótesis nula es $F = \frac{SS(\text{Between}) / (I-1)}{SS(\text{Within}) / (n-I)}$.

Tabla de análisis de la varianza

Source	SS	df	MS	F	p
Between	SS(B)	I-1	$\frac{SS(B)}{I-1}$	$\frac{SS(B) / (I-1)}{SS(W) / (n-I)}$	
Within	SS(W)	n-I	$\frac{SS(W)}{n-I}$		

Source	SS	df	MS	F	p
Total	SS(B) + SS(W)				

Contraste (\dots)

- **Bajo la hipótesis nula de que todas las medias son la misma** (y puesto que asumimos una misma varianza) tendríamos una distribución común bajo todas las condiciones.
- Asumiendo la hipótesis nula el estadístico (F) se distribuye como un (F) con $(I-1)$ y $(n-I)$ grados de libertad, $[F \sim F_{I-1, n-I}]$
- Bajo la hipótesis alternativa, los valores de (F) tenderán a ser **grandes** o mayores que los esperables bajo la hipótesis nula.
- La **región crítica** (donde rechazamos la hipótesis nula) será un intervalo de la forma $[c, +\infty)$.
- Si tomamos como valor (c) el valor observado tendremos el p-valor.

Representación matricial

- El vector de respuestas aleatorias de modo que las (n_1) primeras posiciones las ocupan $(Y_{11}, \dots, Y_{1 n_1})$, las (n_2) posiciones siguientes $(Y_{21}, \dots, Y_{2 n_2})$ y así sucesivamente.
- Haciendo lo análogo con los errores aleatorios tendríamos el vector $(\mathbf{\epsilon})$.
- $$[\mathbf{Y} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} \\ \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} \\ \mathbf{0}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_{n_I} & \mathbf{1}_{n_I} \\ \mathbf{0}_{n_I} & \cdots & \mathbf{0}_{n_I} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_{n_I} & \mathbf{1}_{n_I} \\ \mathbf{0}_{n_I} & \cdots & \mathbf{0}_{n_I} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_I \end{bmatrix} + \mathbf{\epsilon}]$$

Otra formulación del modelo

- Consideramos las variables que indican las categorías y que consideran como categoría de referencia el primer grupo.
- $[E[Y_{1j}] = \beta_0]$ y $[E[Y_{ij}] = \beta_0 + \beta_i]$ para $(i=2, \dots, I)$. De un modo conjunto: $[E[Y_{ij}] = \beta_0 + \beta_2 v_{2j} + \dots + \beta_I v_{Ij}]$ donde $(v_{ij} = 1)$ si estamos en el grupo (i) y cero en otro caso.
- $$[\mathbf{Y} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} \\ \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} \end{bmatrix}]$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
time2Pi	3	37.52	12.505	12.98	6.22e-05 ***
Residuals	20	19.26	0.963		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1