



Introducción a la Estadística (II)

RAQUEL SOCORRO LEÓN

ANOVA

- ▶ Análisis de varianza (ANOVA) también conocida como análisis factorial estudia el efecto de uno o más factores (o grupos), es decir, una o más variables independientes (X) sobre una variable respuesta o dependiente continua (Y). Este o estos factores pueden tener distintas modalidades o lo que se conoce como niveles.
- ▶ Utiliza las varianzas para determinar si las medias son diferentes.
 - **Hipótesis nula**, H_0 : las medias de los grupos son iguales. $\mu_1 = \mu_2 = \dots = \mu_k$
 - **Hipótesis alternativa**, H_1 : alguna de las medias es distinta. $\mu_i \neq \mu_j$ para algún i y j

Donde μ es la media del grupo y k el número de grupos.

- ▶ Fue desarrollada por Fisher en 1930.

Supuestos de ANOVA

1. La variable dependiente o respuesta debe ser continua
2. La variable independiente o explicativa debe estar formada por dos o más niveles, es decir, k de medias (k mayor o igual a 2)
3. Independencia de las observaciones
4. Normalidad de las observaciones: la distribución de la variable dependiente, Y , sigue una distribución normal $(0,1)$
5. Homogeneidad de la varianza (Homocedasticidad): las varianzas de la respuesta en cada grupo son iguales

Tipos de ANOVA

- ▶ Según nº de factores: se refiere al total de variables independientes que incluimos.
 - **Unifactoriales:** Una sola variable independiente
 - Multifactoriales: Más de una variable independiente

- ▶ Tipo de muestro de niveles: se refiere al nivel que tiene la variable independiente.
 - **Efectos fijos:** Seleccionamos los niveles exclusivos que queremos probar
 - Efectos aleatorios: Seleccionamos de manera aleatoria un nº de niveles.
 - Efectos mixtos: Cuando disponemos de dos o mas factores y unos son fijos y otros aleatorios

- ▶ Tipo de aleatorización: se refiere a cómo asignamos los sujetos a las condiciones
 - **Aleatorios:** asignados aleatoriamente, el nº observaciones coincide con el nº de participantes
 - Medidas Repetidas: un único grupo de sujetos recibe x condiciones
 - Split-Plot: disponemos de factores aleatorizados y medidas repetidas. Por ejemplo: conocer la posible influencia diferencial de una campaña publicitaria entre hombres y mujeres, podríamos medir todos los sujetos antes y después de la campaña. Tendríamos un diseño mixto 2 x 2 con cuatro condiciones experimentales: 2 relativas al factor sexo (H/M) y 2 de medida repetidas (medida antes y después de la campaña)

Modelo ANOVA

- ▶ La diferencia entre medias se detecta a través del estudio de la varianza entre grupos y dentro del grupo.
- ▶ Se descompone la variabilidad total o lo que es lo mismo SST (sum of squares total) de la variable dependiente en dos componentes:
 - Suma Cuadrática Intergrupo o SSB (*sum of squares between*)
 - Suma Cuadrática Intragrupo o SSW (*sum of squares within*)

Modelo ANOVA (II)

Esto significa que:

Variabilidad Total (SST) =

Suma Cuadrática Intergrupo (SSB) + Suma Cuadrática Intragrupo (SSW)

lo que es equivalente:

Variabilidad debida a los diferentes niveles del factor + variabilidad residual

lo que es equivalente:

Variabilidad explicada por el factor + variabilidad no explicada por el factor

lo que es equivalente:

Varianza entre niveles + varianza dentro de los niveles

Modelo ANOVA (III)

- **Suma de Cuadrados Total (SST):** mide la variabilidad total de los datos, se define como la suma de los cuadrados de las diferencias de cada observación respecto a la media general de todas las observaciones.

Los grados de libertad es igual al número total de observaciones menos uno ($N-1$).

- **Suma de Cuadrados Intergrupo (SSB):** mide la variabilidad en los datos asociada al efecto del factor sobre la media (la diferencia de las medias entre los diferentes niveles. Se obtiene como la suma de los cuadrados de las desviaciones de la media de cada proveedor respecto de la media general, ponderando cada diferencia al cuadrado por el número de observaciones de cada grupo.

Los grados de libertad son igual al número niveles del factor menos uno ($J-1$).

- **Suma de Cuadrados Intragrupo/residual/error (SSW):** mide la variabilidad dentro de cada nivel, es decir, la variabilidad que no es debida a variable cualitativa o factor. Se calcula como la suma de los cuadrados de las desviaciones de cada observación respecto a la media del nivel al que pertenece.

Los grados de libertad es igual a la diferencia entre los grados de libertad totales y los grados de libertad del factor, o lo que es lo mismo ($N-J$).

Modelo ANOVA (IV)

- Tabla resumen del Modelo ANOVA

Fuente de variación	Sumas Cuadráticas	Grados de libertad	Medias Cuadráticas	Estadístico de contraste
Inter	$\sum_j n_j \bar{Y}_j - \bar{Y}^2$	J-1	SC_{Inter} / gl_{Inter}	$F = \frac{MC_{Inter}}{MC_{Intra}}$
Intra	$\sum_j \sum_i Y_{ij} - \bar{Y}_j^2$	N-J	SC_{Intra} / gl_{Intra}	
Total	$\sum_j \sum_i Y_{ij} - \bar{Y}^2$	N-1		

Estadístico de contraste F

Una vez descompuesta la estimación de la varianza, obtenemos el estadístico F o también conocido como F ratio. Para ello, dividimos las medias cuadráticas inter e intra. Es decir:

$$F_{ratio} = \frac{MC\ Inter}{MC\ Intra}$$

La F_{ratio} o proporción de varianza intergrupo versus la intragrupo, sigue la distribución F centrada con n_1 y n_2 grados de libertad si H_0 es cierta.