ANOVA

jueves, 28 de abril de 2022 18:44

A partir de ahora, para cada análisis vamos a intentar contestar unas preguntas:

- ¿Para qué voy a utilizar este análisis?
- ¿Cómo voy a intrepretar los resultados?
- ¿En qué casos lo voy a poder utilizar?

En primer lugar recordaremos ANOVA:

1. ¿Para qué voy a utilizar este análisis?

Supongamos que:

Población 1 : $x_{11}, x_{12}, \dots, x_{1n_1}$

Población 2 : $x_{21}, x_{22}, \dots, x_{2n_2}$

⋮

Población g : $x_{g1}, x_{g2}, \dots, x_{gn_2}$

} Si tuvieramos 2 poblaciones utilizaríamos un test de hipótesis básico

El objetivo es descubrir si existen diferencias significativas en la media para las g poblaciones

Por ejemplo:

- Nota de un examen para estudiantes que asistieron a diferentes programas escolares:

Nota del estudiante 1 que asistió al P1.

P1 : $x_{11}, \dots, x_{1n_1}$

P2 ...

P2: $X_{21}, \dots, X_{2n_2}$

⋮

PG: $X_{g1}, \dots, X_{gn_g}$

↓

Nota del estudiante g que asistió al PG.

Por tanto:

Notación: X_{ij} → i = programa, j = estudiante

Es decir, utilizaré este test para comprobar si hay diferencias entre más de dos grupos para una variable.

2 ¿Cómo voy a intrepretar los resultados?

"Nota de los est. del P1"

Sean $X_{\ell 1}, \dots, X_{\ell n_\ell}$ muestras aleatorias de $N(\mu_\ell, \sigma^2)$

con $\ell = 1, \dots, g$

Buscamos evaluar $\mu_1 = \mu_2 = \dots = \mu_g$, es decir, si las medias son iguales

Para ello, realizamos el siguiente test de hipótesis

$H_0: \mu_1 = \mu_2 = \dots = \mu_g$

Si definimos $\mu_\ell = \mu + \tau_\ell$

"efecto asociado a la pob. ℓ"

Podemos definir el test como:

$H_0: \tau_1 = \dots = \tau_g = 0$ (Que <u>NO</u> haya efecto)

$H_1: \exists \tau_\ell \neq 0, \ell = 1, \dots, g$

De este modo podemos definir:

$$x_{\ell j} \sim N(\mu + \tau_\ell, \sigma^2) \Rightarrow x_{\ell j} = \overbrace{\mu + \tau_\ell}^{\mu_\ell} + \varepsilon_{\ell j}$$ error aleatorio

Para verificar este test debemos calcular las sumas de cuadrados siguientes:

$$SS_{total} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} x_{\ell j}^2$$ (suma al cuadrado de todas las obs.)

$$SS_{mean} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} \bar{x}^2 = \sum_{\ell=1}^{g} n_\ell \bar{x}^2$$ (suma cuadrada de las medias)

$$SS_{treatment} = \sum_{\ell=1}^{g} \tau_\ell^2 = \sum_{\ell=1}^{g} (\bar{x}_\ell - \bar{x})^2$$ (suma cuadrada de los "efectos")

$$SS_{res} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} e_{\ell j}^2 = \sum\sum (x_{\ell j} - \bar{x}_\ell)$$ (suma cuadrada de los errores)

Definimos

$$F = \frac{SS_{treatment}/(g-1)}{SS_{res}/(n-g)} \sim \underset{\substack{\uparrow \\ \text{F de tablas}}}{F_{(g-1),(n-1)}}$$

Rechazaremos H_0 si $F > F_{(g-1),(n-1)}$

Es decir, si $F > F_{(g-1),(n-1)}$ podré concluir que existen diferencias para alguna población (NO sabemos cuál). En caso contrario, NO existirán diferencias para la variable en estudio en ninguna población.

(3) ¿En qué casos voy a poder utilizar ANOVA?

Se deben cumplir los siguientes supuestos:

a) $X_{\ell 1}, \dots, X_{\ell n_\ell}$ es una muestra aleatoria de tamaño n_ℓ de una población con media μ_ℓ independiente entre poblaciones.

b) Todas las poblaciones tienen la misma varianza.

c) Todas las poblaciones son normales multivariadas.

Es decir, para poder utilizar ANOVA, se deben cumplir esos supuestos.

Ejemplo:

Example 1

Consider the following independent samples.

Population 1 : $9, 6, 9$
Population 2 : $0, 2$
Population 3 : $3, 1, 2$

Obtain the sum of squares decomposition for univariate ANOVA, and test for treatment effects.

Example 1

Since, for example,

media para la 3a población

$$\bar{x}_3 = (3+1+2)/3 = 2 \text{ and}$$

$$\bar{x} = (9+6+9+0+2+3+1+2)/8 = 4, \rightarrow \text{media total}$$

we find that

$$\left.\begin{aligned} 3 = x_{31} &= \bar{x} + (\bar{x}_3 - \bar{x}) + (x_{31} - \bar{x}_3) \\ &= 4 + (2-4) + (3-2) \\ &= 4 + (-2) + 1 \end{aligned}\right\} \text{Descomposición}$$

Repeating this operation for each observation, we obtain the arrays

$$\begin{pmatrix} 9 & 6 & 9 \\ 0 & 2 & \\ 3 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 4 & 4 \\ 4 & 4 & \\ 4 & 4 & 4 \end{pmatrix} + \begin{pmatrix} 4 & 4 & 4 \\ -3 & -3 & \\ -2 & -2 & -2 \end{pmatrix} + \begin{pmatrix} 1 & -2 & 1 \\ -1 & 1 & \\ 1 & -1 & 0 \end{pmatrix}$$

observation $=$ mean $+$ treatment effect $+$ residual
$(x_{\ell j})$ $\quad (\bar{x})$ $\quad (\bar{x}_\ell - \bar{x})$ $\quad (x_{\ell j} - \bar{x}_\ell)$

Example 1

$$\begin{pmatrix} 9 & 6 & 9 \\ 0 & 2 & \\ 3 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 4 & 4 \\ 4 & 4 & \\ 4 & 4 & 4 \end{pmatrix} + \begin{pmatrix} 4 & 4 & 4 \\ -3 & -3 & \\ -2 & -2 & -2 \end{pmatrix} + \begin{pmatrix} 1 & -2 & 1 \\ -1 & 1 & \\ 1 & -1 & 0 \end{pmatrix}$$

observation $=$ mean $+$ treatment effect $+$ residual

- For the observations, we construct the vector $\mathbf{y}' = [9, 6, 9, 0, 2, 3, 1, 2]$. Its squared length is

$$\mathrm{SS}_{\mathrm{obs}} = 9^2 + 6^2 + 9^2 + 0^2 + 2^2 + 3^2 + 1^2 + 2^2 = 216$$

- $\mathrm{SS}_{\mathsf{mean}} = 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 = 8\left(4^2\right) = 128$
- $\mathrm{SS}_{\mathrm{tr}} = 4^2 + 4^2 + 4^2 + (-3)^2 + (-3)^2 + (-2)^2 + (-2)^2 + (-2)^2$
  $= 3\left(4^2\right) + 2(-3)^2 + 3(-2)^2 = 78$
- $\mathrm{SS}_{\mathrm{res}} = 1^2 + (-2)^2 + 1^2 + (-1)^2 + 1^2 + 1^2 + (-1)^2 + 0^2 = 10$
- $\mathrm{SS}_{\mathrm{obs}} = \mathrm{SS}_{\mathrm{mean}} + \mathrm{SS}_{\mathrm{tr}} + \mathrm{SS}_{\mathrm{res}}$, or $216 = 128 + 78 + 10$.

Example 1

Source of variation	Sum of squares	Degrees of freedom
Treatments	$\mathrm{SS}_{\mathrm{tr}} = 78$	$g - 1 = 3 - 1 = 2$
Residual	$\mathrm{SS}_{\mathrm{res}} = 10$	$\sum_{\ell=1}^{g} n_\ell - g = (3+2+3) - 3 = 5$
Total (corrected)	$\mathrm{SS}_{\mathsf{cor}} = 88$	$\sum_{\ell=1}^{g} n_\ell - 1 = 7$

Therefore,

$$F = \frac{\mathrm{SS}_{\mathrm{tr}}/(g-1)}{\mathrm{SS}_{\mathrm{res}}/\left(\Sigma n_\ell - g\right)} = \frac{78/2}{10/5} = 19.5$$

Since $F = 19.5 > F_{2,5}(.01) = 13.27$, we reject $H_0 : \tau_1 = \tau_2 = \tau_3 = 0$ (no treatment effect) at the 1% level of significance.