Dos Muestras

Mediante el método del pivote podemos obtener tests de hipótesis que involucren dos muestras usando los pivotes con los que trabajamos en el contexto de intervalos de confianza.

Volvamos al ejemplo de las parcelas

Dos Muestras

Mediante el método del pivote podemos obtener tests de hipótesis que involucren dos muestras usando los pivotes con los que trabajamos en el contexto de intervalos de confianza.

Volvamos al ejemplo de las parcelas

Ejemplo: Comparamos dos tratamientos

- **Objetivo**: comparar el rendimiento de dos fertilizantes.
- Se divide un terreno en 30 parcelas homogéneas (por ej. cant. sol, humedad, etc.).
- En las 30 parcelas se cultiva la misma variedad de maíz.
- En 15 de ellas se utiliza el fertilizante A y en las restantes el B.

Ejemplo: Comparamos dos tratamientos

Datos

- Parcelas con Tratamiento A

$$238, 237, 235, 220, 233, 203, 228, 220, 221,$$
$$215, 218, 217, 232, 225, 209$$

- Parcelas con Tratamiento B

$$253, 227, 241, 245, 237, 248, 250, 218, 239,$$
$$243, 257, 208, 215, 240, 229$$

- Queremos comparar el rendimiento medio de esta variedad de maíz cuando el terreno se trata con estos dos fertilizantes.

Modelo 1: Muestras normales independientes con igual varianza

Supongamos que:

- $X_1, \ldots, X_{n_1} \sim \mathcal{N}\left(\mu_1, \sigma^2\right),$
- $Y_1, \ldots, Y_{n_2} \sim \mathcal{N}\left(\mu_2, \sigma^2\right)$

Además, asumimos que:

Las dos muestras aleatorias, independientes entre sí.

Test para $H_0 : \mu_1 - \mu_2 = \Delta$ vs. $H_1 : \mu_1 - \mu_2 \neq \Delta$

En muchas situaciones $\Delta = 0$. Notemos que en nuestro ejemplo podemos reescribir las hipótesis como:

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2$$

Varianzas DESconocidas pero IGUALES $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1) \ , \quad \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{N}(0,1)$$

Necesitamos estimar σ^2. Tenemos dos estimadores:

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \overline{X}_{n_1})^2 \ , \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \overline{Y}_{n_2})^2$$

Test para $H_0 : \mu_1 - \mu_2 = \Delta$ vs. $H_1 : \mu_1 - \mu_2 \neq \Delta$

En muchas situaciones $\Delta = 0$. Notemos que en nuestro ejemplo podemos reescribir las hipótesis como:

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2$$

Varianzas DESconocidas pero IGUALES $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1) \ , \quad \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{N}(0,1)$$

Necesitamos estimar σ^2. Tenemos dos estimadores:

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \overline{X}_{n_1})^2 \ , \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \overline{Y}_{n_2})^2$$

Los combinamos:

$$S_p^2 = \frac{S_1^2 (n_1 - 1) + S_2^2 (n_2 - 1)}{n_1 + n_2 - 2}$$

El pivote

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1+n_2-2} \; , \quad S_p^2 = \frac{S_1^2 \,(n_1 - 1) + S_2^2 \,(n_2 - 1)}{n_1 + n_2 - 2}$$

El pivote

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}\,, \quad S_p^2 = \frac{S_1^2\,(n_1 - 1) + S_2^2\,(n_2 - 1)}{n_1 + n_2 - 2}$$

$$\frac{\dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta}{\sqrt{\sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}}{\sqrt{\dfrac{S_1^2\,(n_1 - 1) + S_2^2\,(n_2 - 1)}{\sigma^2(n_1 + n_2 - 2)}}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\dfrac{\chi^2_{n_1+n_2-2}}{n_1 + n_2 - 2}}} \sim t_{n_1+n_2-2}$$

El pivote

$$\frac{\dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta}{\sqrt{\sigma^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)}}}{\sqrt{\dfrac{S_1^2\,(n_1 - 1) + S_2^2\,(n_2 - 1)}{\sigma^2 (n_1 + n_2 - 2)}}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\dfrac{\chi^2_{n_1+n_2-2}}{n_1 + n_2 - 2}}} \sim t_{n_1+n_2-2}$$

ya que son independientes

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta}{\sqrt{\sigma^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)}} \sim \mathcal{N}(0,1),$$

$$\frac{S_1^2\,(n_1 - 1)}{\sigma^2} \sim \chi^2_{n_1-1}$$

y

$$\frac{S_2^2\,(n_2 - 1)}{\sigma^2} \sim \chi^2_{n_2-1}$$

$H_0 : \mu_1 - \mu_2 = \Delta$ vs. $H_1 : \mu_1 - \mu_2 \neq \Delta$

Test de nivel α

Rechazamos H_0 si

$$\frac{|\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta|}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > t_{n_1+n_2-2,\alpha/2}$$

$H_0 : \mu_1 - \mu_2 = \Delta$ vs. $H_1 : \mu_1 - \mu_2 \neq \Delta$

Test de nivel α

Rechazamos H_0 si

$$\frac{|\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta|}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > t_{n_1+n_2-2,\alpha/2}$$

.....¿y cómo sabemos que las varianza son iguales????

El test más tradicional para chequear esto es el test de F que asume que las dos poblaciones son normales.

Comparación de varianzas

Supongamos que tenemos dos muestras independientes:

- $X_1, \ldots, X_{n_1} \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$,
- $Y_1, \ldots, Y_{n_2} \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$

Queremos testear a nivel α: $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 \neq \sigma_2^2$

Parece razonable considerar el cociente

$$\frac{S_1^2}{S_2^2}$$

Como bajo H_0 esperamos que sea cercano a 1, rechazaremos H_0 si es mucho más grande o mucho más pequeño que 1.

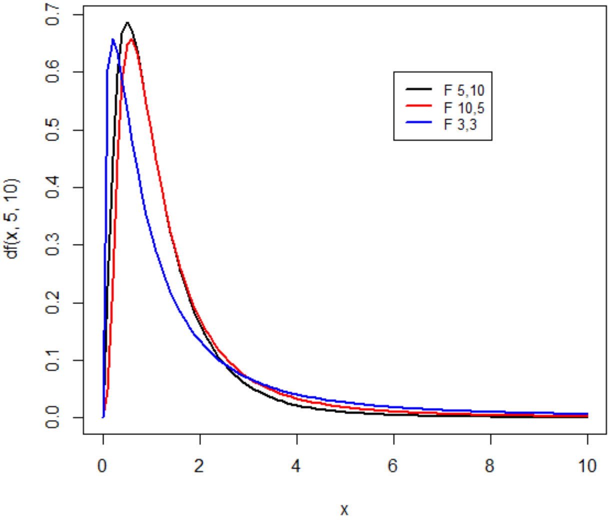
¿Pero conocemos su distribución?

Tengamos en cuenta que si $U \sim \chi_k^2$ y $W \sim \chi_m^2$ independientes, entonces

$$V = \frac{U/k}{W/m} \sim F_{k,m}$$

.

Densidad F



El pivote

Llamemos $F_{k,m,\beta}$ al punto que cumple

$$\beta = \mathbb{P}(V \geq F_{k,m,\beta}) \text{ para } V \sim F_{k,m}$$

.

Notemos que bajo H_0

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}$$

por lo tanto, rechazaremos H_0 si

$$\frac{S_1^2}{S_2^2} \leq F_{n_1-1,n_2-1,1-\alpha/2} \text{ o bien } \frac{S_1^2}{S_2^2} \geq F_{n_1-1,n_2-1,\alpha/2}$$

El pivote

Llamemos $F_{k,m,\beta}$ al punto que cumple

$$\beta = \mathbb{P}(V \geq F_{k,m,\beta}) \text{ para } V \sim F_{k,m}$$

.

Notemos que bajo H_0

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}$$

por lo tanto, rechazaremos H_0 si

$$\frac{S_1^2}{S_2^2} \leq F_{n_1-1,n_2-1,1-\alpha/2} \text{ o bien } \frac{S_1^2}{S_2^2} \geq F_{n_1-1,n_2-1,\alpha/2}$$

¿Cómo comparamos las medias si rechazamos con este test la igualdad de varianzas?

Modelo 2: Muestras normales independientes con distintas varianzas

- $X_1, \ldots, X_{n_1} \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right),$
- $Y_1, \ldots, Y_{n_2} \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$

Las dos muestras aleatorias, independientes entre sí.

$H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$

El problema de comparar las medias en estas condiciones se conoce como el problema de Fisher-Behrens.

Parece natural considerar el estadístico:

$$T^* = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

La distribución de T^* ya no es una t de Student bajo H_0, su distribución es la de Fisher-Behrens y depende de n_1, n_2 y de σ_1^2/σ_2^2.

Modelo 2: Muestras normales independientes con distintas varianzas

$$T^* = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Hay distintas soluciones a este problema, una de las más usadas se basa la solución de Welch (1947) en la que la distribución de T^* se aproxima por una t_ν, en la que los grados de libertad ν están dados por

$$\nu = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(S_1^2/n_1\right)^2/(n_1-1) + \left(S_2^2/n_2\right)^2/(n_2-1)}$$

Vayamos a R para terimar el ejemplo de las parcelas tratadas con dos fertilizantes.

Regiones de confianza

Dado un vector $\mathbf{X}$ con distribución perteneciente a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$, *una región de confianza $\mathcal{S}(\mathbf{X})$ para $\boldsymbol{\theta}$ con nivel de confianza* $1-\alpha$ será una función que a cada $\mathbf{X}$ le hace corresponder un subconjunto $\mathcal{S}(\mathbf{X}) \subset \Theta$ de manera que

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in \mathcal{S}(\mathbf{X})) = 1 - \alpha, \qquad \forall \boldsymbol{\theta} \in \Theta$$

Es decir, $\mathcal{S}(\mathbf{X})$ cubre el valor verdadero del parámetro con probabilidad $1-\alpha$.

Regiones de confianza

Dado un vector $\mathbf{X}$ con distribución perteneciente a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$, *una región de confianza* $\mathcal{S}(\mathbf{X})$ *para* $\boldsymbol{\theta}$ *con nivel de confianza* $1-\alpha$ será una función que a cada $\mathbf{X}$ le hace corresponder un subconjunto $\mathcal{S}(\mathbf{X}) \subset \Theta$ de manera que

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in \mathcal{S}(\mathbf{X})) = 1-\alpha, \qquad \forall \boldsymbol{\theta} \in \Theta$$

Es decir, $\mathcal{S}(\mathbf{X})$ cubre el valor verdadero del parámetro con probabilidad $1-\alpha$.

Caso particular: Si $\boldsymbol{\theta} \in \mathbb{R}$ se dirá que $\mathcal{S}(\mathbf{X})$ es un intervalo de confianza

$$\mathcal{S}(\mathbf{X}) = [a(\mathbf{X}), b(\mathbf{X})]$$

La longitud de $\mathcal{S}(\mathbf{X})$ es

$$L = b(\mathbf{X}) - a(\mathbf{X})$$

Relación entre Tests e Intervalos de confianza

- $\mathbf{X} \sim F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.
  Para cada $\boldsymbol{\theta}_0$ fijo sea $\phi_{\boldsymbol{\theta}_0}$, un test no aleatorizado de nivel α, para
  $$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Relación entre Tests e Intervalos de confianza

- $\mathbf{X} \sim F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.
  Para cada $\boldsymbol{\theta}_0$ fijo sea $\phi_{\boldsymbol{\theta}_0}$, un test no aleatorizado de nivel α, para
  $$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$
- $\mathcal{A}(\boldsymbol{\theta}_0)$: Región de aceptación de $\phi_{\boldsymbol{\theta}_0}$
- Notemos que
  $$P_{\theta_0}(\mathbf{X} \in \mathcal{A}(\boldsymbol{\theta}_0)) = 1 - \alpha$$
- Familias de regiones de aceptación:
  $\mathcal{S}(\mathbf{X}) = \{\boldsymbol{\theta} : \phi_{\boldsymbol{\theta}}(\mathbf{X}) = 0\} = \{\boldsymbol{\theta} : \mathbf{X} \in \mathcal{A}(\boldsymbol{\theta})\}$

  $\Longrightarrow \mathcal{S}(\mathbf{X})$ es una región de confianza de nivel $1 - \alpha$ para $\boldsymbol{\theta}$

Relación entre Test e Intervalos de confianza

- Recíprocamente, si $\mathcal{S}(\mathbf{X})$ es una región de confianza de nivel $1-\alpha$ para $\boldsymbol{\theta}$, el test

$$\phi_{\boldsymbol{\theta}_0}(\mathbf{X}) = \begin{cases} 1 & \text{si } \boldsymbol{\theta}_0 \notin \mathcal{S}(\mathbf{X}) \\ 0 & \text{si } \boldsymbol{\theta}_0 \in \mathcal{S}(\mathbf{X}) \, . \end{cases}$$

es un test de nivel de α para testear

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Relación entre Test e Intervalos de confianza

En general podemos decir que:

- Para obtener un intervalos de confianza $[L(\mathbf{X}), U(\mathbf{X})])$ podemos invertir la región de aceptación de un test de la forma $H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \neq \theta_0$
- Para obtener una cota inferior de confianza $[L(\mathbf{X}), \infty)$ podemos invertir la región de aceptación de un test de la forma $H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta > \theta_0$
- Para obtener una cota superior de confianza $(-\infty, U(\mathbf{X})]$ podemos invertir la región de aceptación de un test de la forma $H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta < \theta_0$