Grupo MAT ESI UCLM

Sesión 9: Inferencia Estadística

1

EUSEBIO ANGULO SÁNCHEZ-HERRERA

CURSO 2022/2023

ESCUELA SUPERIOR DE INFORMÁTICA CIUDAD REAL

Contenidos Principales

- Intervalos de confianza (CI) y Constraste de Hipótesis (HT)
  - dos muestras
  - Reglas de significación
  - p-valor
    - Rechazamos Ho cuando p-valor ≤ α

independientes: primero hacer test de igualdad
pareadas:

- Datos independientes
  - Test de igualdad devarianzas -> Test de medias
- Datos pareados
  - Test de medias para datos pareados

Distribución Normal

n grande -> teorema central del limite para asumir normalidad

- Métodos paramétricos
  - Solo utilizables cuando los datos siguen una distribución normal.
  - De lo contrario, sería necesario utilizer métodos no paramétricos.

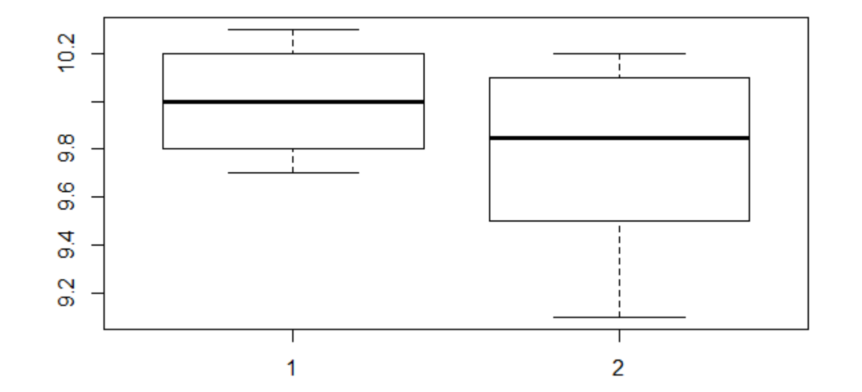
- El TeoremaCentral del Límite
  - Si "n es lo suficientemente grande", podemos asumir la normalidad, entonces temenos un modelo que responde a los parámetros…

Ejemplo Grosor de Aislamiento

```
insulation1<-c(10.3,10.1,9.8,9.9,10.2,10.1,9.7,9.9,9.7,10.2)
insulation2<-c(9.9,10.1,9.5,9.3,10.2,10,9.1,9.5,9.8,10.1)
```

Ejemplo Grosor de Aislamiento

- boxplot(insulation1,insulation2)



  - summary(insulation1)

```
Min. 1st Qu. Median Mean  3rd Qu. Max.
9.700 9.825  10.000 9.990 10.170  10.300
```

  - summary(insulation2)

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
9.10 9.50    9.85   9.75  10.07   10.20
```

Ejemplo Grosor de Aislamiento

- ¿Hay diferencias significativas entre las medias de las muestras?

- Estadísticamente

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{cases}$$

- Estamos asumiendo normalidad pero n=10 es pequeño y nosotros necesitaríamos métodos no paramétricos.

Distintos planteamientos

- Datos independientes:
  - Varianzas iguales   var.equal = true

– Varianzas desconocidas e iguales: $T = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)S_{c_1}^2 + (n_2-1)S_{c_2}^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$.

  - Varianzas distintas   var.equal = false

– Varianzas desconocidas y distintas: $T = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_{c1}^2}{n_1}+\frac{S_{c2}^2}{n_2}}}$

- Datos pareados: paired = true

- Dos medias (muestras relacionadas): $T = \dfrac{\bar{X}_1 - \bar{X}_2}{\frac{S_{cD}}{\sqrt{n}}}$.

Distintos planteamientos

Datos independientes o pareados

1º) Test varianzas desconocidas e iguales
> var.test

2º) Test de media
> t.test with var equal (or unequal)

1º) Test de medias (muestras relacionadas)

> t.test with paired data

¿Independiente o Pareado?

1) El grosor del aislamiento medido en dos edificios diferentes en paredes distintas.

2) El grosor del aislamiento medido en varias paredes de un edificio en verano e invierno.

¿Independiente o Pareado?

1) El grosor del aislamiento medido en dos edificios diferentes en paredes distintas. **INDEPENDIENTE**

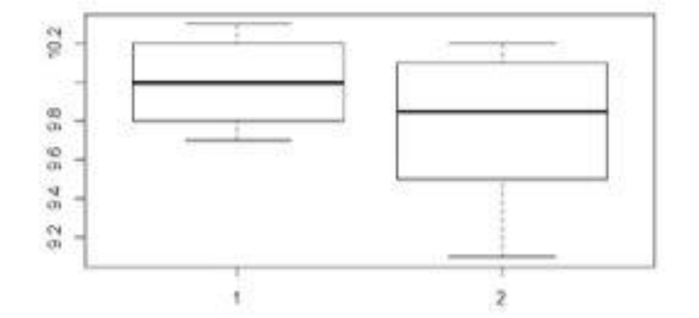
2) El grosor del aislamiento medido en varias paredes de un edificio en verano e invierno. **PAREADO**

2 medias independientes

Primero, nosotros necesitamos un test de varianzas:

test de varianzas

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases} \sim \begin{cases} H_0: & \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1: & \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{cases}$$



var.test()

var.test(insulation1,insulation2)

F test to compare two variances

data: insulation1 and insulation2

se acepta H0 0,1178 < 0,05 las varianzas son iguales

F = 0.33385, num df = 9, denom df = 9, p-value = 0.1178

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.08292415 1.34408677

sample estimates:

ratio of variances

0.3338521

Decision Rule...

'We reject H_0 when p-value $\leq \alpha$'

$\Uparrow$

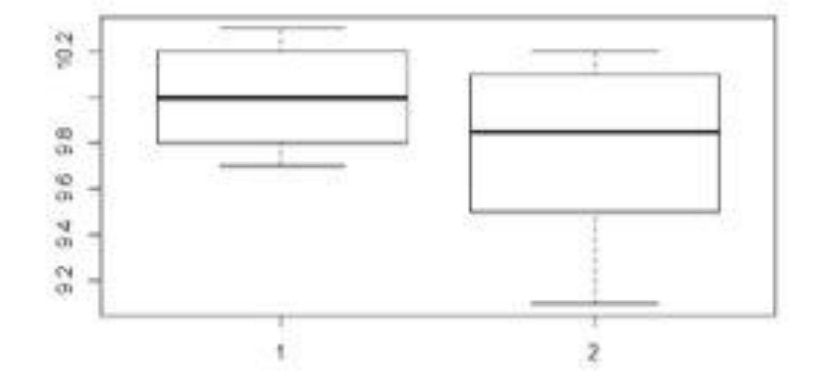
desired significance level (beforehand)

In general $\alpha = 0.05$

2 medias independientes: test de medias

Contraste de Hipótesis bilateral:

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{cases} \quad \sim \quad \begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 - \mu_2 \neq 0 \end{cases}$$



2 medias independientes

$\sigma_1 = \sigma_2$ (como hemos comprobado anteriormente)

t.test(insulation1,insulation2,var.equal = TRUE)

Two Sample t-test

data: insulation1 and insulation2

t = 1.7391, df = 18, p-value = 0.09909 se rechaza H0

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.04993062 0.52993062

sample estimates:

mean of x mean of y

9.99 9.75

Decision Rule...

'We reject H_0 when

- statistic ∈ Critical Region (α)
- p-value ≤ α
- reference value for "parameter" ∉ (1- α)% CI

Tipos de hipótesis en R

$$H_1 : \ldots \neq \ldots \quad \overset{\text{R}}{\Rightarrow} \text{ `}\texttt{two.sided}\text{' (by default)}$$

$$H_1 : \ldots < \ldots \quad \overset{\text{R}}{\Rightarrow} \text{ `}\texttt{less}\text{'}$$

$$H_1 : \ldots > \ldots \quad \overset{\text{R}}{\Rightarrow} \text{ `}\texttt{greater}\text{'}$$

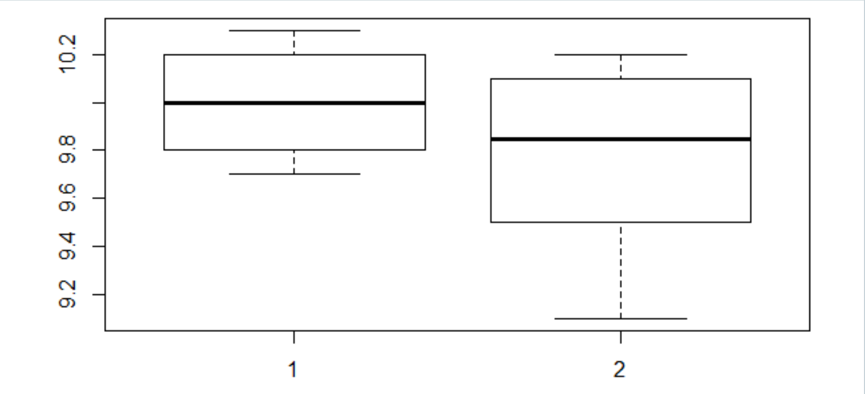
$$H_1 : \mu_x < \mu_y$$

$$H_1 : \mu_y > \mu_x$$

is the same! $\Rightarrow$ **Data order** $\Downarrow$

```
t.test(x,y,...alt="less"...)
t.test(y,x,...alt="greater"...)
```

Ejemplos tests unilaterales



todas las combinaciones que se pueden demostrar

```
> t.test(insulation1, insulation2, alternative = "less" )
> t.test(insulation1, insulation2, alternative = "greater" )
> t.test(insulation2, insulation1, alternative = "less" )
> t.test(insulation2, insulation1, alternative = "greater" )
```

Media con datos pareados

Grosor del aislamiento medido en varias paredes de un edificio en verano e invierno

Test de Hipótesis Bilateral:

$$sample1_i - sample2_i = D_i$$

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{cases} \sim \begin{cases} H_0: & \mu_D = 0 \\ H_1: & \mu_D \neq 0 \end{cases}$$

Media con datos pareados

> t.test(insulation1, insulation2, paired = TRUE)

pareados

Paired t-test

data: insulation1 and insulation2

t = 2.9794, df = 9, p-value = 0.01547

se rechaza H0, medias distintas

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.05777501 0.42222499

sample estimates:

mean of the differenc

0.24

Decision Rule...

'We reject H_0 when

- statistic ∈ Critical Region (α)
- p-value ≤ α
- reference value for "parameter" ∉ (1- α)% CI

Ejercicio 1

- Se han probado dos marcas de automóviles comparando las velocidades máximas que alcanzan. Se han hecho 12 pruebas con cada uno de los vehículos obteniendo los siguientes resultados:

  A: 256, 239, 222, 207, 228, 241, 212, 216, 236, 219, 225, 230

  B: 212, 240, 263, 275, 237, 271, 261, 223, 234, 220, 232, 230

- Suponiendo normalidad en la variable ¿existe evidencia estadística de que el vehículo de la marca B alcanza velocidades máximas en promedio mayores que el de la marca A? (Considerar una significancia del 5%)

Ejercicio 1 b

- Se han probado dos marcas de automóviles comparando las velocidades máximas que alcanzan. Se han hecho 12 pruebas con cada uno de los vehículos obteniendo los siguientes resultados:

  A: 256, 239, 222, 207, 228, 241, 212, 216, 236, 219, 225, 230

  B: 247, 219, 270, 282, 244, 278, 268, 230, 227, 241, 239, 237

- Suponiendo que los datos siguen una distribución normal, ¿existe evidencia estadística de que el vehículo de la marca B es 5km/h mayor que la velocidad máxima media de la marca A con una significancia del 5%?

Ejercicio 2

- El departamento de calidad de una empresa quiere comprobar si se mantiene el nivel de calidad de un producto de una semana a otra, tras un incidente registrado durante el fin de semana. La calidad del artículo se supone normalmente distribuida y se han obtenido los siguientes resultados de una m.a.s de 8 productos:

  Semana1: 93, 86, 90, 90, 94, 91, 92, 96

  Semana2: 93, 87, 97, 90, 88, 87, 84, 93

- Realizar el contraste correspondiente a un nivel de significación del 1% que permita discriminar si se mantiene el nivel de calidad.

Ejercicio 3

Ejercicio 5.2.8 Examen Extraordinario 2018

- El número diario de twits realizados durante los 5 días previos al examen que contienen el hashtag #laestadisticamola han sido: 50, 42, 53, 60, 37. Mientras tanto, el número de twits que contenían el hashtag #aprobareestadistica esos mismos días ha sido de: 40, 51, 62, 55 y 64. Si suponemos que ambas distribuciones de tuits se pueden considerar normales y considerando un nivel de significación del 5%, se pide:

- ¿Podemos concluir que el número de twits medio que contienen el hastagh #laestadisticamola es mayor que 39?
- Comprobar si podemos llegar a la conclusión de que las medias del número de tuits que usan esos hashtags difieren.