Fundamentos de ciencia de datos con R - Módulo 5

Clase 2: Inferencia para medias y diferencia de medias I

CEPAL - Unidad de Estadísticas Sociales

2025-11-07

Introducción

En la clase anterior analizamos la inferencia sobre una media poblacional. Hoy extenderemos ese concepto para comparar dos medias, diferenciando entre:

- Muestras independientes (dos grupos distintos).
- Muestras pareadas (antes y después, o casos emparejados).

Introducción

Nota

Objetivos de esta clase:

- Contrastar hipótesis sobre la diferencia de medias.
- Aplicar pruebas t en R según el tipo de muestra.
- Interpretar los resultados con visualizaciones.

Base de datos y muestra

Usaremos la misma base base_personas_gasto.rds, centrando el análisis en el gasto medio .

```
library(tidyverse)

datos <- readRDS("../Data/base_personas_gasto.rds")
set.seed(123)
muestra <- slice_sample(datos, n = 500)
head(muestra[,2:8], 5)
```

id_pers	upm	estrato	area	fep	pobreza	ingreso_hh
3	14128201182	141282	1	45	3	7236.040
2	14912200300	149122	2	22	3	12036.117
1	4102200159	41022	1	135	3	5414.580
2	3101400134	31014	1	17	3	7923.053
1	5902900064	59029	2	41	3	9714.580

Comparación de medias (muestras independientes)

Supongamos que queremos saber si el gasto promedio de los hogares difiere entre hombres y mujeres.

Formulamos las hipótesis:

$$H_0 : \mu_H = \mu_M$$
$$H_1 : \mu_H \neq \mu_M$$

Prueba t para muestras independientes

Usamos la prueba t de Student:

```
t.test(gasto_hh ~ sexo, data = muestra, var.equal = TRUE)
```

```
    Two Sample t-test

data:  gasto_hh by sexo
t = -0.58955, df = 498, p-value = 0.5558
alternative hypothesis: true difference in means between group Hombre and g
95 percent confidence interval:
 -2518.466  1355.900
sample estimates:
mean in group Hombre  mean in group Mujer
            11005.41              11586.69
```

Prueba t para muestras independientes

Tip

Si el valor-p $< 0.05 \rightarrow$ rechazamos H_0 (hay diferencia significativa).
Si el valor-p $> 0.05 \rightarrow$ no se rechaza H_0 (no hay evidencia de diferencia).

Resumen descriptivo

```
muestra %>%
group_by(sexo) %>%
summarise(
media = mean(gasto_hh, na.rm = TRUE),
sd = sd(gasto_hh, na.rm = TRUE),
n = n()
)
```

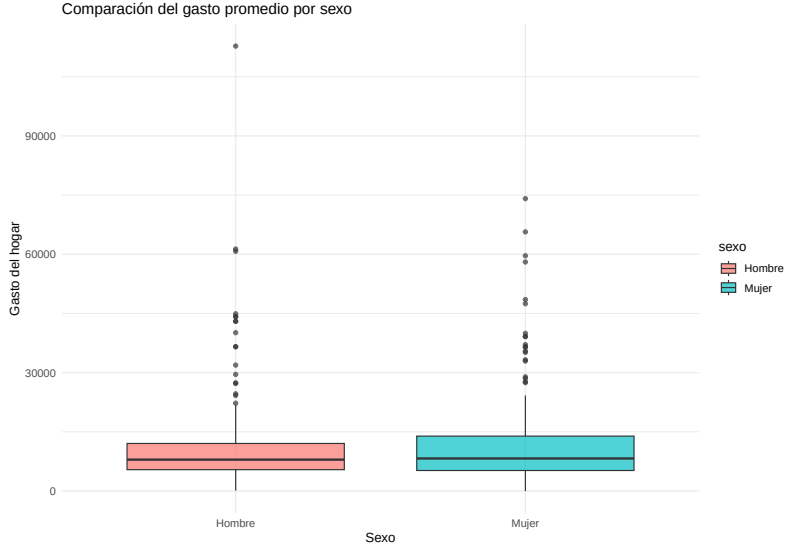
sexo	media	sd	n
Hombre	11005.41	11195.71	248
Mujer	11586.69	10850.72	252

Visualización comparativa

```
grap <- ggplot(muestra, aes(x = sexo, y = gasto_hh, fill = sexo)) +
geom_boxplot(alpha = 0.7) +
labs(title = "Comparación del gasto promedio por sexo",
x = "Sexo", y = "Gasto del hogar") +
theme_minimal(base_size = 12)
```

Visualización comparativa

```
grap
```



Diferencia de medias (muestras pareadas)

En contextos de evaluación o intervención, se dispone de **dos mediciones del mismo grupo**: por ejemplo, el **gasto promedio de los hogares antes y después** de la implementación de una política pública.

En este ejemplo, a partir del gasto observado en la base, se simula un escenario "después" donde los hogares aumentan ligeramente su gasto, reflejando un posible efecto de política o cambio económico.

Diferencia de medias (muestras pareadas)

Formulación de hipótesis (pareadas):

$$H_0 : \mu_{\text{antes}} = \mu_{\text{después}}$$
$$H_1 : \mu_{\text{antes}} \neq \mu_{\text{después}}$$

Donde:

- μ_{antes} = gasto promedio antes de la política
- $\mu_{\text{después}}$ = gasto promedio después de la política

Diferencia de medias (muestras pareadas)

Crear un subconjunto y generar una versión 'después' con pequeña variación

```
df_pareada <- datos %>%
  slice_sample(n = 200) %>%
  transmute(
    id = row_number(),
    gasto_antes = as.numeric(gsub("[^0-9.-]", "", as.character(gasto_hh)))
  ) %>%
  filter(is.finite(gasto_antes)) %>%
  mutate(
    gasto_despues = gasto_antes + rnorm(n(), mean = 50, sd = 100)
  )
```

Diferencia de medias (muestras pareadas)

Crear un subconjunto y generar una versión 'después' con pequeña variación

```
head(df_pareada, 10)
```

id	gasto_antes	gasto_despues
1	6084.147	6121.432
2	3592.920	3448.768
3	4575.000	4743.118
4	9236.520	9472.511
5	10914.580	11071.981
6	9166.073	9213.339
7	3910.773	3957.440
8	15816.190	15714.583
9	6618.364	6747.403
10	6171.300	6200.227

Diferencia de medias (muestras pareadas)

Prueba t pareada

```r
t.test(df_pareada$gasto_antes, df_pareada$gasto_despues, paired = TRUE)
```

```
    Paired t-test

data:  df_pareada$gasto_antes and df_pareada$gasto_despues
t = -6.489, df = 199, p-value = 6.716e-10
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -62.43959 -33.33471
sample estimates:
mean difference
      -47.88715
```

Diferencia de medias (muestras pareadas)

Prueba t pareada

Nota

Interpretación:
Si $p < 0.05$, existe evidencia de cambio significativo entre el gasto antes y después de la intervención.
Si $p > 0.05$, no hay evidencia de cambio.

Conclusiones de la clase

Tipo de prueba	Escenario	Hipótesis nula	Función en R	Interpretación
t de Student (independientes)	Dos grupos distintos con varianzas iguales	$\mu_1 = \mu_2$	`t.test(y ~ grupo, var.equal = TRUE)`	Compara dos medias bajo igualdad de varianzas
t pareada	Dos mediciones del mismo grupo	$\mu_d = 0$	`t.test(x1, x2, paired = TRUE)`	Evalúa cambios antes y después

Conclusiones de la clase

Tip

En resumen: La inferencia para medias nos permite evaluar si los cambios observados son estadísticamente significativos, ya sea entre grupos distintos o dentro del mismo grupo a lo largo del tiempo.