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 definimos los conceptos fundamentales de inferencia estadística. Hoy aplicaremos esas ideas a la estimación e inferencia para una media poblacional, utilizando ejemplos prácticos con R.

Nota

Aprenderemos a:
Calcular intervalos de confianza para una media.
Realizar pruebas de hipótesis sobre una media.
Interpretar resultados de forma estadística y práctica.

Base de datos y muestra

Utilizaremos nuevamente la base base_personas_gasto.rds, que contiene información sobre el ingreso y gasto de los hogares.

```
library(tidyverse)

datos <- readRDS("../Data/base_personas_gasto.rds")

set.seed(123)
muestra <- slice_sample(datos, n = 400)
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

Estimación de la media

```
x <- muestra$ingreso_hh
x <- x[is.finite(x)]
media <- mean(x)
se <- sd(x)/sqrt(length(x))
media; se
```

```
[1] 11897.43
```

```
[1] 672.6117
```

Nota

Interpretación: La media muestral $\bar{X}$ es el mejor estimador puntual de la media poblacional μ. El error estándar $(s/\sqrt{n})$ mide la variabilidad esperada de la media entre muestras.

Intervalo de confianza para la media

```r
ic_95 <- c(media - 1.96 * se, media + 1.96 * se)
ic_95
```

```
[1] 10579.11 13215.75
```

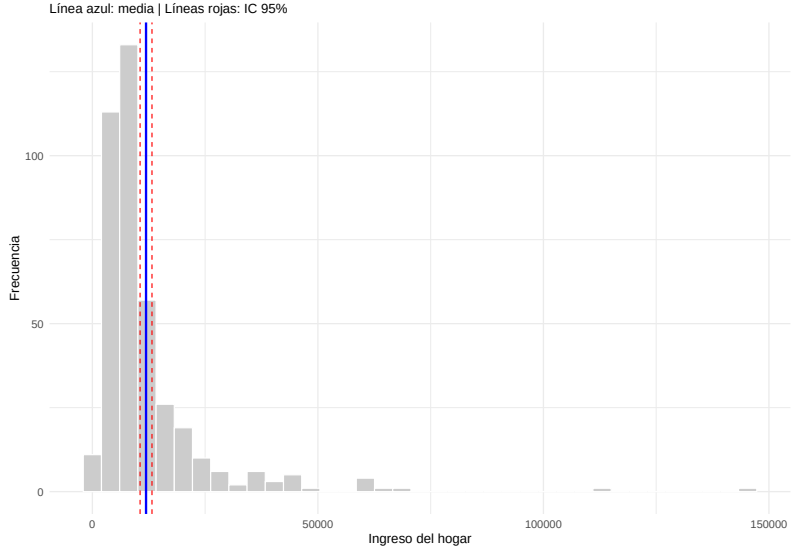
> **Nota**
>
> Interpretación: “Con un 95 % de confianza, la media poblacional se encuentra entre los límites del intervalo calculado.”

Visualización del IC

```
graf <- ggplot(as.data.frame(x), aes(x)) +
geom_histogram(binwidth = 0.3*sd(x), fill = "gray80", color = "white") +
geom_vline(xintercept = media, color = "blue", linewidth = 1) +
geom_vline(xintercept = ic_95, linetype = "dashed", color = "red") +
labs(subtitle = "Línea azul: media | Líneas rojas: IC 95%",
x = "Ingreso del hogar", y = "Frecuencia") +
theme_minimal(base_size = 12)
```

Visualización del IC

```
graf
```



Prueba de hipótesis para la media

Queremos contrastar:

$$H_0 : \mu = 1.500.000 H_1 \quad : \mu \neq 1.500.000$$

Tip

Interpretación:
Si el valor-p $< 0.05 \rightarrow$ se rechaza H_0 (la media difiere de 1.500.000).
Si el valor-p $> 0.05 \rightarrow$ no hay evidencia suficiente para rechazar H_0.

Prueba de hipótesis para la media

Usamos la prueba t de Student:

```r
t.test(muestra$gasto_hh, mu = 1500000)
```

```
    One Sample t-test

data:  muestra$gasto_hh
t = -2526.2, df = 399, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 1500000
95 percent confidence interval:
 10497.23 12813.77
sample estimates:
mean of x
  11655.5
```

Prueba de hipótesis para la media

Con un valor p menor a 2.2×10^{1}, inferior al nivel de significancia del 5 % ($= 0.05$), se rechaza la hipótesis nula que planteaba que el gasto promedio de los hogares es igual a 1.500.000.

El promedio muestral fue de 11.655,5, con un intervalo de confianza al 95 % entre 10.497,23 y 12.813,77, lo que indica que el gasto medio real difiere significativamente del valor hipotético propuesto.

Ejemplo práctico: comparación por sexo

Analizamos si el gasto promedio difiere entre hombres y mujeres.

```
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	11610.56	12176.34	199
Mujer	11700.00	11411.50	201

Ejemplo práctico: comparación por sexo

Prueba de diferencia de medias (independientes)

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

```
    Two Sample t-test

data:  gasto_hh by sexo
t = -0.075804, df = 398, p-value = 0.9396
alternative hypothesis: true difference in means between group Hombre and g
95 percent confidence interval:
 -2408.924  2230.050
sample estimates:
mean in group Hombre  mean in group Mujer
            11610.56              11700.00
```

Ejemplo práctico: comparación por sexo

Prueba de diferencia de medias (independientes)

El resultado fue t=−0.0758, con p-valor = 0.9396, superior al nivel de significancia de 0.05. Por tanto, no se rechaza la hipótesis nula, lo que indica que no existen diferencias estadísticamente significativas en el gasto promedio entre hombres y mujeres. El resultado fue t=−0.0758, con p-valor = 0.9396, superior al nivel de significancia de 0.05. Por tanto, no se rechaza la hipótesis nula, lo que indica que no existen diferencias estadísticamente significativas en el gasto promedio entre hombres y mujeres.

Ejemplo práctico: comparación por sexo

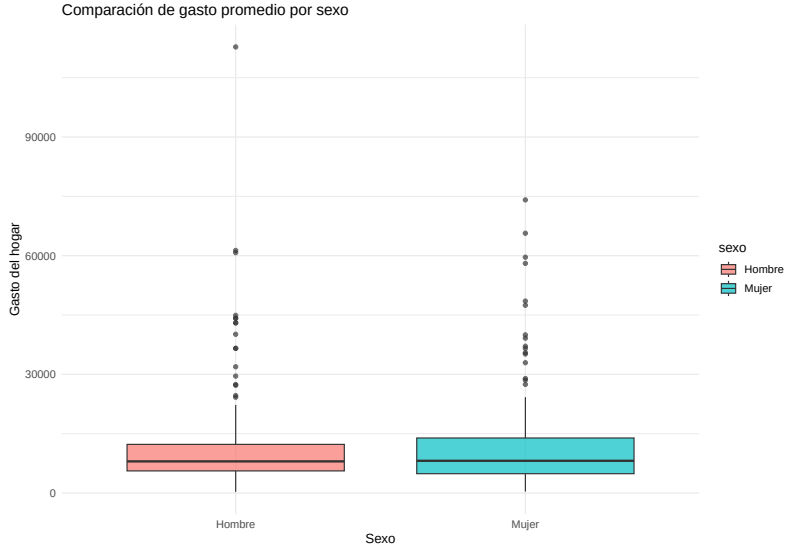
Visualización comparativa

```
graf2 <- ggplot(muestra, aes(x = sexo, y = gasto_hh, fill = sexo)) +
geom_boxplot(alpha = 0.7) +
labs(title = "Comparación de gasto promedio por sexo",
x = "Sexo", y = "Gasto del hogar") +
theme_minimal(base_size = 12)
```

Ejemplo práctico: comparación por sexo

Visualización comparativa

```
graf2
```



Conclusiones de la clase

Concepto	Aplicación en esta clase
Media muestral ($\bar{X}$)	Estimador puntual de μ
Error estándar ($s/\sqrt{n}$)	Medida de precisión del estimador
Intervalo de confianza	Rango plausible para la media poblacional
Prueba t para una media	Contrasta hipótesis sobre la media poblacional
Valor-p	Mide la evidencia contra H_0
Comparación inicial entre grupos	Introducción a la diferencia de medias

i Nota

En resumen: Hoy aplicamos los fundamentos de la inferencia sobre una media. En la Clase 3 profundizaremos en la comparación entre dos medias y sus aplicaciones prácticas.