Fundamentos de ciencia de datos con R - Módulo 5

Clase 1: Definiciones básicas estadísticas

CEPAL - Unidad de Estadísticas Sociales

2025-11-06

Introducción

En esta clase revisaremos los conceptos fundamentales para la inferencia estadística: población, muestra, estimador, intervalo de confianza e hipótesis. Estos conceptos son esenciales para comprender cómo los datos se utilizan para hacer inferencias sobre fenómenos reales.

> **Nota**
>
> "La estadística es el lenguaje de los datos. Nos permite describir, inferir y tomar decisiones basadas en la incertidumbre."

Población y muestra

- Población: conjunto total de elementos sobre los cuales se desea realizar un estudio. Ejemplo: todas las personas mayores de 18 años en un país.
- Muestra: subconjunto representativo de la población, sobre el cual se realizan las mediciones. Ejemplo: 1.200 personas seleccionadas aleatoriamente para una encuesta.

Tip

Una muestra representativa debe reflejar las características esenciales de la población.

Carga base de datos ejemplo

```
library(tidyverse)
 datos <- readRDS("../Data/base_personas_gasto.rds")
 head(datos[,1:8], 5)
```

id_hogar	id_pers	upm	estrato	area	fep	pobreza	ingreso_hh
262	1	1100100006	11001	1	19	3	10783.05
262	2	1100100006	11001	1	19	3	10783.05
265	1	1100100006	11001	1	16	3	21259.72
265	2	1100100006	11001	1	16	3	21259.72
265	3	1100100006	11001	1	16	3	21259.72

Selección de una muestra

El tamaño y la forma en que se elige la muestra son fundamentales para garantizar la **representatividad** y la **validez de las inferencias**.

En este ejemplo, partimos de la base de datos `base_personas_gasto.rds`, que representa nuestra **población**.
A partir de ella, seleccionamos de forma **aleatoria simple** una muestra de 400 registros, que utilizaremos para estimar parámetros como la media y proporciones.

```
set.seed(123) #fijamos semilla
n_total <- nrow(datos)
n_muestra <- 400
muestra <- slice_sample(datos, n = n_muestra)
n_total; n_muestra
```

```
[1] 19427
```

```
[1] 400
```

Estimadores básicos

*Estimador:** es un estadístico calculado a partir de los datos de una muestra, utilizado para aproximar un parámetro desconocido de la población.
Por ejemplo, la media muestral $\bar{X}$ estima la media poblacional μ.

Parámetro poblacional	Estimador muestral	Fórmula
Media (μ)	$\bar{X}$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Varianza (σ^2)	s^2	$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$
Desviación estándar (σ)	s	$s = \sqrt{s^2}$
Proporción (p)	$\hat{p}$	$\hat{p} = \frac{x}{n}$
Correlación (ρ)	r	$r = \frac{s_{XY}}{s_X s_Y}$

Estimadores básicos

Nota

Un buen **estimador** debe cumplir tres propiedades fundamentales:

- **Insesgado:** su valor esperado es igual al parámetro poblacional.
- **Eficiente:** tiene la menor varianza posible entre los estimadores insesgados.
- **Consistente:** se aproxima al valor real del parámetro a medida que aumenta el tamaño de la muestra ($n \to \infty$).

Intervalos de confianza (IC)

El **intervalo de confianza (IC)** proporciona un rango de valores plausibles para el parámetro poblacional, con un cierto nivel de confianza (por ejemplo, 95 %).

$$IC = \bar{X} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Donde:

- $\bar{X} \rightarrow$ media muestral
- $z_{\alpha/2} \rightarrow$ valor crítico de la distribución normal estándar asociado al nivel de confianza deseado
- $s \rightarrow$ desviación estándar muestral
- $n \rightarrow$ tamaño de la muestra

Pruebas de hipótesis

Las **pruebas de hipótesis** son procedimientos estadísticos que permiten **tomar decisiones sobre un parámetro poblacional** a partir de los datos muestrales.

El proceso consiste en contrastar una **hipótesis nula (H_0)** con una **hipótesis alternativa (H_1)**, utilizando la información obtenida de la muestra.

Ejemplo general:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Donde:

- H_0 es la afirmación inicial o de no diferencia (por ejemplo, "la media poblacional es igual a un valor conocido").
- H_1 representa la alternativa (por ejemplo, "la media poblacional es distinta").

Pruebas de hipótesis

El estadístico de prueba para la media, si la varianza poblacional es desconocida, se calcula como:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Este valor se compara con la distribución t de Student para determinar si se **rechaza o no** H_0 al nivel de significancia α.

Valor-p y nivel de significancia ( )

Nivel de significancia (α): umbral de error tipo I que estamos dispuestos a aceptar (típicos: 0.10, 0.05, 0.01).

Se relaciona con el nivel de confianza por: **nivel de confianza =** $1 - \alpha$.

Valor-p: probabilidad, **asumiendo** H_0 **verdadera**, de obtener un estadístico de prueba **tan extremo o más** que el observado.

Valor-p y nivel de significancia ( )

- Prueba **una cola (derecha)**:

$$p\text{-valor} = \Pr(T \geq t_{\text{obs}} \mid H_0)$$

- Prueba **una cola (izquierda)**:

$$p\text{-valor} = \Pr(T \leq t_{\text{obs}} \mid H_0)$$

- Prueba **dos colas**:

$$p\text{-valor} = 2\,\text{mín}\{\Pr(T \leq t_{\text{obs}}),\ \Pr(T \geq t_{\text{obs}})\}$$

Regla de decisión:
Si **valor-p** $\leq \alpha$, **se rechaza** H_0. Si **valor-p** $> \alpha$, **no se rechaza** H_0.

Valor-p y nivel de significancia ( )

Tip

Interpretación correcta: *"Con los datos observados, sería (valor-p) la probabilidad de ver un resultado igual o más extremo* ***si*** H_0 ***fuera cierta****."*
No es "la probabilidad de que H_0 sea verdadera".

Otras pruebas de hipótesis

Además de las pruebas sobre la **media**, existen otras pruebas de hipótesis ampliamente utilizadas para contrastar diferentes tipos de parámetros estadísticos.

A continuación se presentan algunas de las más comunes:

Otras pruebas de hipótesis

Prueba para una proporción

valúa si la proporción poblacional (p) es igual a un valor teórico p_0.

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Prueba para la diferencia de medias

Compara si dos grupos tienen la misma media poblacional.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Otras pruebas de hipótesis

Prueba de independencia (Chi-cuadrado)

Evalúa si dos variables categóricas están asociadas o son independientes.

El estadístico de prueba se calcula como:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Donde:
- O_{ij} = frecuencia observada en la celda (i, j)
- E_{ij} = frecuencia esperada en la celda (i, j) bajo H_0

El estadístico χ^2 sigue una **distribución Chi-cuadrado** con $(r-1)(c-1)$ grados de libertad, donde r es el número de filas y c el número de columnas de la tabla.

Resumen de la clase

A continuación, se presenta una síntesis de los conceptos fundamentales revisados en esta sesión:

Concepto	Descripción	Ejemplo o Fórmula
Población	Conjunto total de elementos sobre los cuales se desea inferir.	Todos los hogares de una ciudad.
Muestra	Subconjunto representativo de la población.	400 hogares seleccionados aleatoriamente.
Estimador	Estadístico calculado con los datos muestrales para aproximar un parámetro poblacional.	$\bar{X}$ estima μ
Propiedades del estimador	Insesgado, eficiente, consistente.	$E(\bar{X}) = \mu$

Resumen de la clase

A continuación, se presenta una síntesis de los conceptos fundamentales revisados en esta sesión:

Concepto	Descripción	Ejemplo o Fórmula
Intervalo de confianza (IC)	Rango de valores plausibles para el parámetro poblacional con cierto nivel de confianza.	$IC = \bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$
Prueba de hipótesis	Procedimiento para evaluar afirmaciones sobre parámetros poblacionales.	$H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$
Valor-p	Probabilidad de obtener un estadístico tan extremo como el observado, si H_0 fuera cierta.	Si $p < \alpha$, se rechaza H_0.
Pruebas comunes	Media, proporción, diferencia de medias, independencia.	t, z, χ^2

Conlusion

En esta clase aprendimos a distinguir entre población y muestra, a definir estimadores, calcular intervalos de confianza y formular pruebas de hipótesis, que serán la base de las aplicaciones prácticas en las siguientes sesiones.