Fundamentos de ciencia de datos con R - Módulo 1

Clase 4: Descripción de los datos

CEPAL - Unidad de Estadísticas Sociales

2025-10-30

Introducción

La descripción de los datos es el primer paso del análisis estadístico. Permite conocer la estructura, distribución y calidad de la información, identificando patrones, valores atípicos y posibles errores.

Como señaló Tukey (1977)

"la exploración de los datos busca descubrir lo que los datos pueden decirnos por sí mismos"

Tipos de datos en R

Los datos en R pueden adoptar diferentes **tipos o clases**, según su naturaleza y propósito analítico.

Tipo de dato	**¿Qué representa?**
`numeric`	Números decimales o enteros
`integer`	Números enteros declarados explícitamente
`character`	Texto: palabras, nombres, frases
`logical`	Valores lógicos: verdadero o falso
`factor`	Categorías o niveles (variables cualitativas)
`Date`	Fechas en formato calendario

Tipos de datos en R

```r
# Numeric (numérico)
x_num <- 12.5
class(x_num)
```

```
[1] "numeric"
```

```r
# Integer (entero)
x_int <- 7
class(x_int)
```

```
[1] "numeric"
```

Tipos de datos en R

```
# Character (texto)
x_char <- "Bogotá"
class(x_char)
```

```
[1] "character"
```

```
# Logical (lógico)
x_log <- TRUE
class(x_log)
```

```
[1] "logical"
```

Tipos de datos en R

```r
# Factor (categorías)
x_fac <- factor(c("Primaria", "Secundaria", "Universitaria"))
class(x_fac)
```

```
[1] "factor"
```

```r
levels(x_fac)
```

```
[1] "Primaria"      "Secundaria"    "Universitaria"
```

```r
# Date (fecha)
x_date <- as.Date("2025-10-21")
class(x_date)
```

```
[1] "Date"
```

Ejemplo práctico

Construimos un data frame que combina distintos tipos de datos (numéricos, categóricos y de texto) para ilustrar cómo R maneja cada uno de ellos.

```r
datos <- data.frame(
  edad = c(23, 30, 27, 45, 50),
  sexo = factor(c("Mujer", "Hombre", "Mujer", "Hombre", "Mujer")),
  ingreso = c(2000, 3500, 2800, 5000, 4200)
)

str(datos)      # Estructura del conjunto de datos
```

```
'data.frame':   5 obs. of  3 variables:
 $ edad   : num  23 30 27 45 50
 $ sexo   : Factor w/ 2 levels "Hombre","Mujer": 2 1 2 1 2
 $ ingreso: num  2000 3500 2800 5000 4200
```

Exploración inicial de los datos

```r
summary(datos) # Resumen estadístico general
```

edad	sexo	ingreso
Min. :23	Hombre:2	Min. :2000
1st Qu.:27	Mujer :3	1st Qu.:2800
Median :30	NA	Median :3500
Mean :35	NA	Mean :3500
3rd Qu.:45	NA	3rd Qu.:4200
Max. :50	NA	Max. :5000

Nota

Interpretación:
Para variables numéricas: muestra mínimo, máximo, media, mediana y cuartiles.
Para factores: muestra la frecuencia de cada categoría.

Estadísticos descriptivos básicos

```r
mean(datos$edad) # Promedio
```

```
[1] 35
```

```r
median(datos$edad) # Mediana
```

```
[1] 30
```

```r
sd(datos$edad) # Desviación estándar
```

```
[1] 11.81101
```

Estadísticos descriptivos básicos

```r
var(datos$edad) # Varianza
```

```
[1] 139.5
```

```r
range(datos$edad) # Rango (mínimo y máximo)
```

```
[1] 23 50
```

Tip

Antes de calcular medidas descriptivas, verifica que las variables sean del tipo correcto. Si una variable numérica aparece como texto, usa as.numeric().

Tablas de frecuencia y proporciones

Las tablas permiten observar la distribución de categorías en variables cualitativas.

```
table(datos$sexo)              # Frecuencia absoluta
```

Hombre	Mujer
2	3

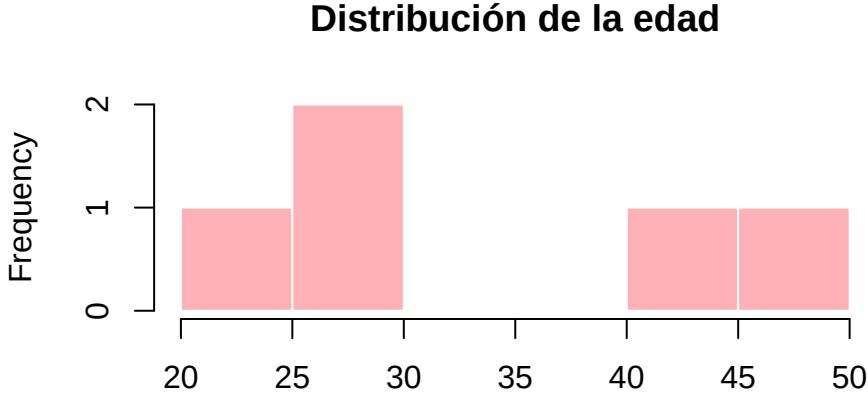
```
prop.table(table(datos$sexo)) # Frecuencia relativa (%)
```

Hombre	Mujer
0.4	0.6

Visualización básica

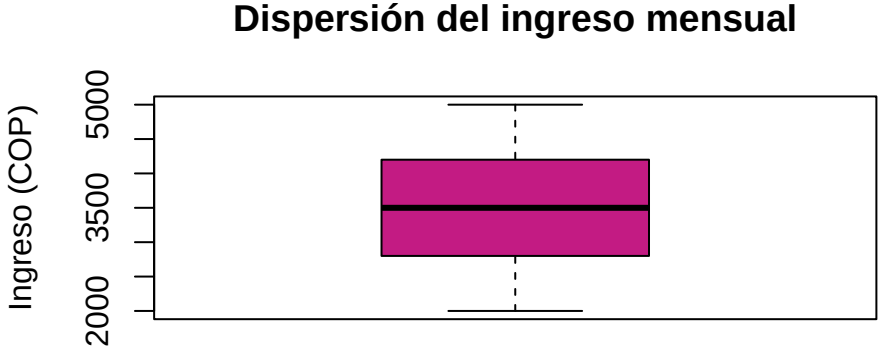
Una forma rápida de comprender los datos es mediante representaciones gráficas simples.

```
hist(datos$edad, main = "Distribución de la edad",
xlab = "Edad", col = "#FEB0B7", border = "white")
```



Visualización básica

```
boxplot(datos$ingreso, col = "#C31B83",
main = "Dispersión del ingreso mensual",
ylab = "Ingreso (COP)")
```



Visualización básica

Nota

Los gráficos ayudan a detectar valores atípicos y patrones de distribución.