

# Lección 3: Tipos de variables

Hans Sigrist

2026-03-19

En esta lección se estudian los distintos tipos de variables estadísticas (numéricas y categóricas) y se introducen los conceptos de asociación, independencia y variables explicativa y respuesta.

## Tabla de contenidos

|          |                                                                           |          |
|----------|---------------------------------------------------------------------------|----------|
| <b>1</b> | <b>Cargar librerías y explorar los datos</b>                              | <b>1</b> |
| <b>2</b> | <b>Tipos de variables: ¿cómo clasificarlas?</b>                           | <b>3</b> |
| <b>3</b> | <b>Efecto de los valores atípicos (outliers) en la media y la mediana</b> | <b>6</b> |
| <b>4</b> | <b>Cuestionario Grupal: Tipos de variables</b>                            | <b>8</b> |

Built with Quarto v1.8.27 · R v4.5.2 · LaTeX v3.141592653-2.6-1.40.27 · openintro v2.0.0

### Advertencia

Los datos utilizados en esta lección (`region30`) corresponden a una versión didáctica simplificada de la data `county` (`openintro`) que simula indicadores regionales en Chile. No representan datos oficiales del INE ni de ningún organismo público, y han sido contruidos únicamente con fines pedagógicos para el estudio de tipos de variables y relaciones estadísticas.

## 1. Cargar librerías y explorar los datos

Cargamos `tidyverse` (colección de paquetes para ciencia de datos) y leemos el archivo `region30.csv`. Luego usamos `glimpse()` para ver la estructura de los datos de forma compacta.



Figura 1: Ilustración utilizada con fines educativos. Fuente: USF ClipArt Collection (University of South Florida).

```
library(tidyverse)
region30 <- read_csv("../datos/region30.csv")
glimpse(region30)
```

```
Rows: 30
Columns: 8
$ region          <chr> "Norte_1", "Norte_2", "Norte_3", "Centro_1", ~
$ tasa_desempleo  <dbl> 8.4, 7.1, 9.3, 6.2, 5.8, 7.4, 10.2, 8.9, 7.6~
$ poblacion_total <dbl> 540000, 620000, 410000, 980000, 1120000, 870~
$ crecimiento_poblacional <dbl> 1.2, -0.5, 0.8, 1.5, 2.1, 0.4, -1.1, -
0.3, 0~
$ ingreso_mediano_hogar <dbl> 720000, 680000, 640000, 880000, 910000, 7600~
$ tasa_propietarios <dbl> 62.1, 59.3, 65.8, 55.4, 52.7, 60.2, 71.5, 68~
$ viviendas_multipiso <dbl> 38.4, 41.2, 32.5, 48.6, 50.1, 36.9, 21.4, 25~
$ nivel_educacional_mediano <chr> "Media completa", "Técnico superior", "Media~
```

Observamos que tenemos 30 filas y 8 columnas. Cada columna es una variable. Con `names()` podemos listar sus nombres:

```
names(region30)
```

```
[1] "region"          "tasa_desempleo"
[3] "poblacion_total" "crecimiento_poblacional"
[5] "ingreso_mediano_hogar" "tasa_propietarios"
[7] "viviendas_multipiso" "nivel_educacional_mediano"
```

## 2. Tipos de variables: ¿cómo clasificarlas?

En estadística, las variables se clasifican en:

**Numéricas continuas** pueden tomar cualquier valor en un intervalo (ej: `tasa_desempleo`).

**Numéricas discretas** toman valores enteros que representan conteos (ej: `poblacion_total`).

**Categorías nominales** categorías sin orden (ej: `region`).

**Categorías ordinales** categorías con un orden natural (ej: `nivel_educativo`).

Pero cuidado: no todo lo que parece número es una variable numérica en sentido estadístico. Por ejemplo, un número de teléfono es un identificador, no tiene sentido calcular su promedio. En cambio, la `tasa_desempleo` sí es una variable numérica con la que podemos hacer operaciones.

Para las variables categóricas, podemos usar `table()` o `count()` para ver frecuencias. Por ejemplo, la variable `region`:

```
table(region30$region)
```

```
Austral_1 Austral_2 Austral_3 Austral_4 Austral_5 Austral_6 Centro_1 Centro_10
      1         1         1         1         1         1         1         1
Centro_2  Centro_3  Centro_4  Centro_5  Centro_6  Centro_7  Centro_8  Centro_9
      1         1         1         1         1         1         1         1
Norte_1   Norte_2   Norte_3   Norte_4   Norte_5   Norte_6   Norte_7   Sur_1
      1         1         1         1         1         1         1         1
      Sur_2      Sur_3      Sur_4      Sur_5      Sur_6      Sur_7
      1         1         1         1         1         1
```

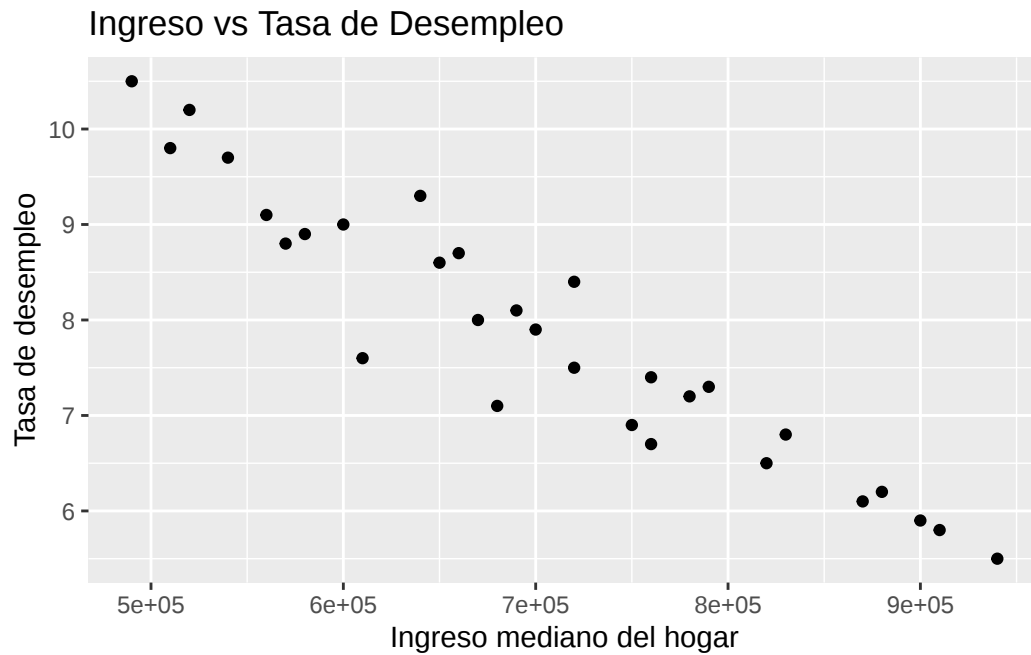
```
region30 |> count(region, sort = TRUE)
```

```
# A tibble: 30 x 2
  region      n
  <chr>    <int>
1 Austral_1    1
2 Austral_2    1
3 Austral_3    1
4 Austral_4    1
5 Austral_5    1
6 Austral_6    1
7 Centro_1     1
8 Centro_10    1
9 Centro_2     1
10 Centro_3     1
# i 20 more rows
```

Relación entre variables: diagramas de dispersión

Cuando queremos explorar la relación entre dos variables numéricas, el gráfico adecuado es el diagrama de dispersión (scatter plot). Veamos si existe relación entre el ingreso mediano del hogar y la tasa de desempleo:

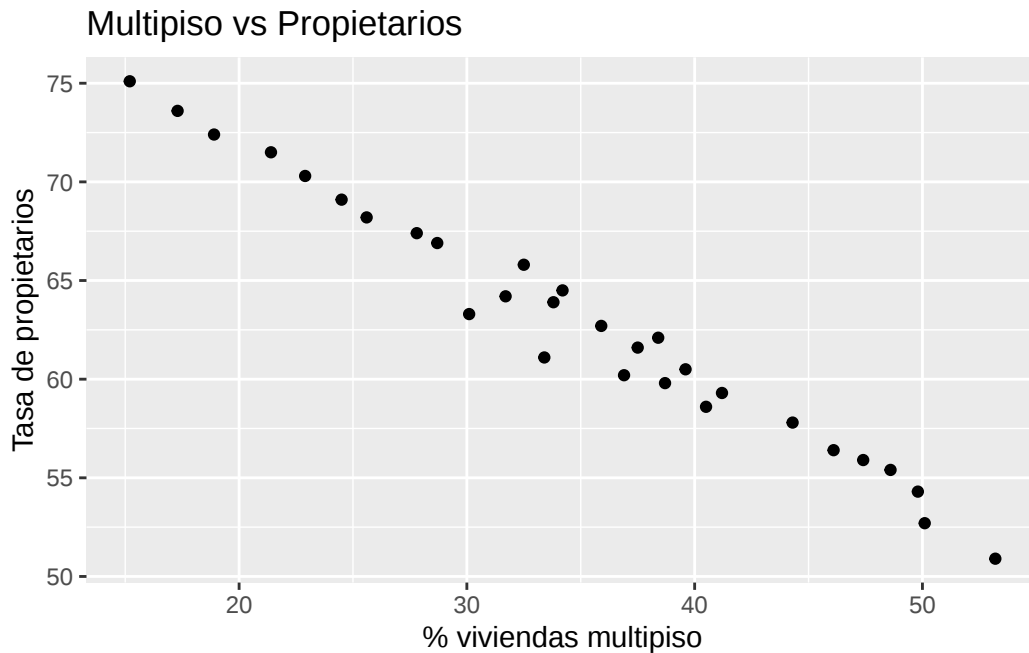
```
ggplot(region30, aes(x = ingreso_mediano_hogar, y = tasa_desempleo)) +
  geom_point() +
  labs(x = "Ingreso mediano del hogar", y = "Tasa de desempleo",
       title = "Ingreso vs Tasa de Desempleo")
```



¿Se observa alguna tendencia? ¿A mayor ingreso, menor desempleo?

Ahora exploremos otra posible relación: porcentaje de viviendas en edificios de varios pisos versus tasa de propietarios.

```
ggplot(region30, aes(x = viviendas_multipiso, y = tasa_propietarios)) +
  geom_point() +
  labs(x = "% viviendas multipiso", y = "Tasa de propietarios",
       title = "Multipiso vs Propietarios")
```



Importante: Que dos variables muestren una asociación en un gráfico no significa que una cause la otra. La correlación no implica causalidad.

### 3. Efecto de los valores atípicos (outliers) en la media y la mediana

Vamos a simular un conjunto de ingresos para entender cómo los valores extremos afectan las medidas de tendencia central.

Primero, generamos 60 ingresos “típicos” usando una distribución log-normal (común para ingresos) y luego añadimos tres valores extremadamente altos.

```
set.seed(123) # para que los resultados sean reproducibles
ingresos_tipicos <- round(rlnorm(60, meanlog = log(650), sdlog = 0.25))
ingresos_extremos <- c(5000, 7000, 12000)
ingresos <- c(ingresos_tipicos, ingresos_extremos)
```

Ahora tenemos 63 valores. Veamos un resumen estadístico:

```
length(ingresos)
```

```
[1] 63
```

```
summary(ingresos)
```

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max.  |
|------|---------|--------|------|---------|-------|
| 398  | 578     | 668    | 1027 | 794     | 12000 |

Calculamos la media y la mediana:

```
media <- mean(ingresos)
mediana <- median(ingresos)
media
```

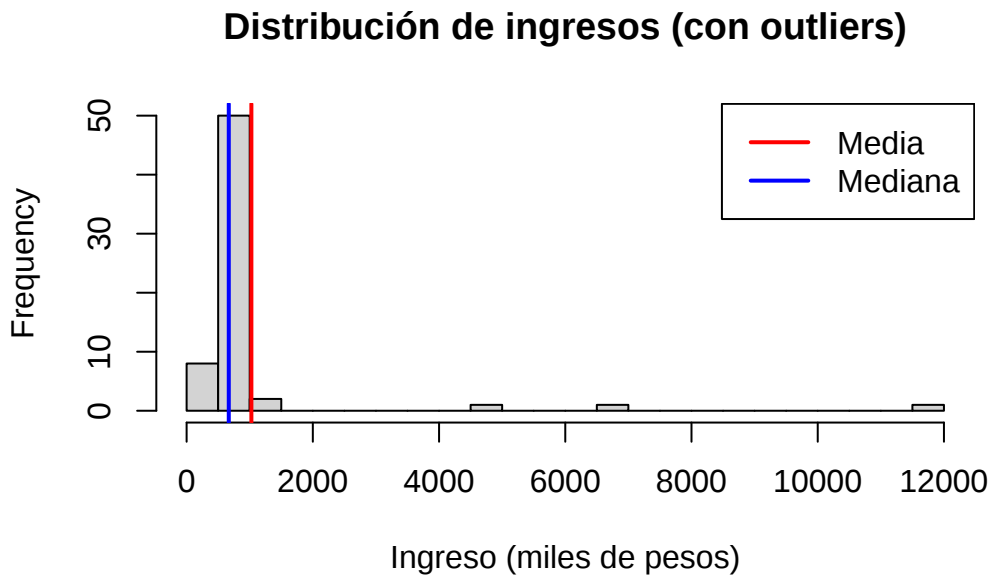
```
[1] 1026.524
```

```
mediana
```

```
[1] 668
```

Observa que la media es mucho mayor que la mediana debido a los valores extremos. Visualicemos la distribución con un histograma:

```
hist(ingresos, breaks = 20, main = "Distribución de ingresos (con outliers)",
     xlab = "Ingreso (miles de pesos)")
abline(v = media, col = "red", lwd = 2)
abline(v = mediana, col = "blue", lwd = 2)
legend("topright", legend = c("Media", "Mediana"), col = c("red", "blue"), lwd = 2)
```



La media es sensible a outliers; la mediana es más robusta. Si eliminamos los valores extremos:

```
ingresos_sin_extremos <- ingresos_tipicos
mean(ingresos_sin_extremos)
```

```
[1] 677.85
```

```
median(ingresos_sin_extremos)
```

[1] 653.5

Ahora media y mediana son mucho más cercanas.

#### 4. Cuestionario Grupal: Tipos de variables

1. **Clasifica las siguientes variables de `region30`.** Para cada una, indica si es: numérica continua, numérica discreta, categórica nominal o categórica ordinal, y justifica brevemente.

| Variable | Tipo | Justificación |
|----------|------|---------------|
|          |      |               |

2. **Explorando la variable region.** Al usar `glimpse(region30)` no se ven todos los valores de la variable `region`. ¿Qué comando usarías para listar todos los valores únicos de esa variable?



 **Tip**

(Ayuda: prueba con `unique()`, `sort()`, `table()` o combinaciones)

```
# Escribe aquí tu código
```

3. **Interpretación de gráficos.** Observa el gráfico de dispersión entre viviendas\_multipiso y tasa\_propietarios. ¿Qué tendencia parece haber? ¿Crees que vivir en un edificio de varios pisos causa una menor tasa de propietarios? Explica.

4. **Media vs mediana.** En el ejemplo de los ingresos simulados, ¿por qué la media es mayor que la mediana? ¿Qué medida es más representativa del “ingreso típico” en presencia de outliers? ¿Por qué?