

Lab 4 - Gráficos, hipótesis

Métodos Cuantitativos

Carlos Eduardo Molina Berumen

El Colegio de México

14 de noviembre de 2025



Recapitulación

Hemos visto: - Comandos básicos de R - Cómo explorar bases de datos -
Cómo sacar estadísticos descriptivos univariados - Tipos de variables

Objetivos de hoy:

- Entender ggplot y cómo hacer buenas gráficas
- Cómo hacer pruebas de hipótesis

ggplot2

```
#install.packages("ggplot2")  
library(ggplot2)
```

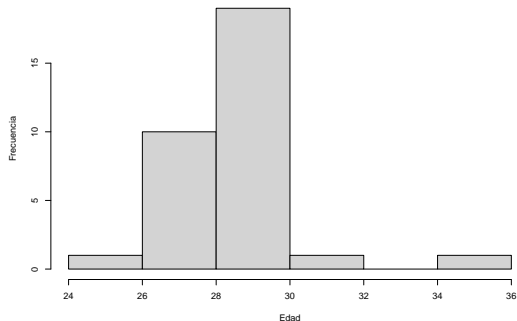
- Con ggplot creamos gráficos empleando una gramática de gráficos que crea objetos de visualización compuestos por capas
- Esta gramática de gráficos se basa en tres componentes: datos, sistema de coordenadas y marcas visuales (geoms)

¿Por qué usar ggplot2 en lugar de r base para nuestros gráficos?

```
library(readr)  
library(dplyr)  
  
data <- read_csv("Estados.csv")  
data2020 <- data %>%  
  filter(Año == "2020")
```

```
hist(data2020$Edad_mediana,  
      xlab = "Edad",  
      ylab = "Frecuencia")
```

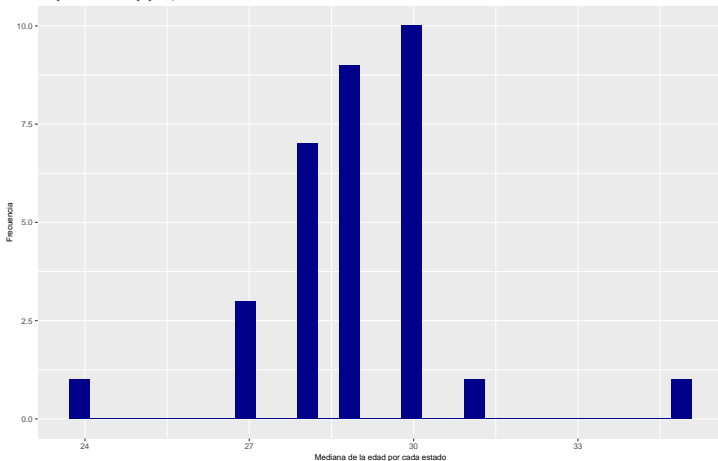
Histogram of data2020\$Edad_mediana



Con ggplot2

Distribución de las edades medianas en México

Histograma con datos desagregados por estado



Fuente: Censo poblacional INEGI 2020

```
ggplot(df, aes(x, y, other aesthetics))
```

- Data frame (df)
- Conjunto de definiciones estéticas (e.g. definición de ejes)
- Agregamos capas al gráfico con +

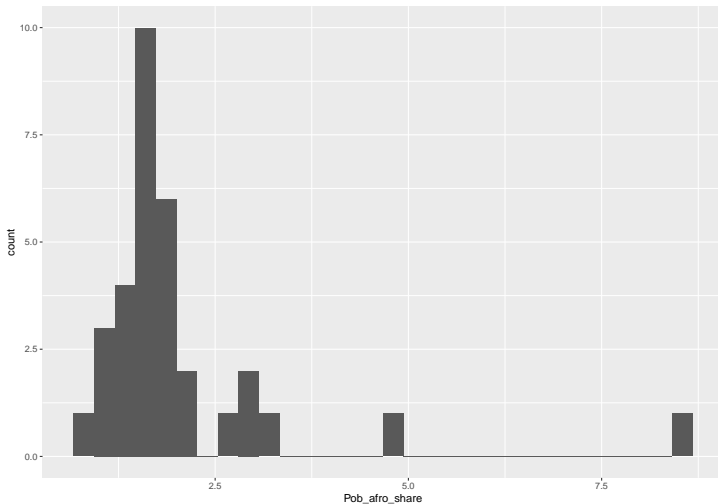
Ejemplo: ¿Cómo crear un histograma en ggplot?

¿Cómo se distribuye el porcentaje de población afromexicana en México?

```
plot <- ggplot(data2020, aes(x = Pob_afro_share)) +  
  geom_histogram()
```



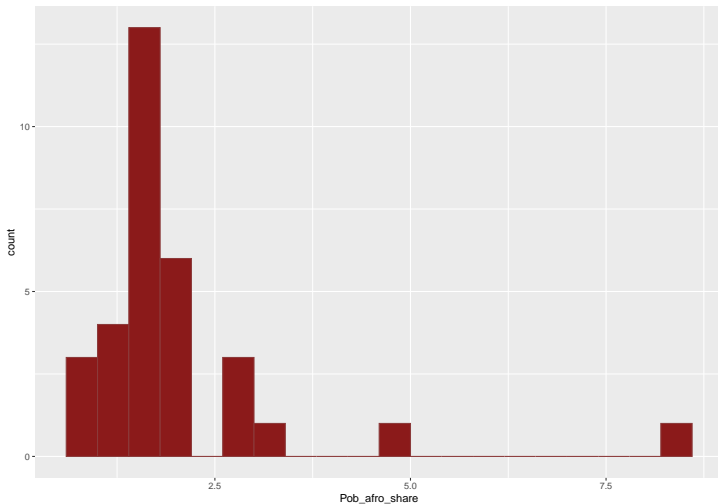
```
print(plot)
```



Si quieren ver todos los colores que tiene R, busquen “R colors cheatsheet”. Cambiemos color, contorno y ancho de nuestras barras.

```
plot_a <- ggplot(data2020, aes(x = Pob_afro_share)) +  
  geom_histogram( fill = "firebrick4", color = "indianred4",  
                 binwidth=0.4)
```

```
print(plot_a)
```



Recordemos: Nuestra gráfica debe comunicarnos toda la información necesaria sin remitirnos a nada más.

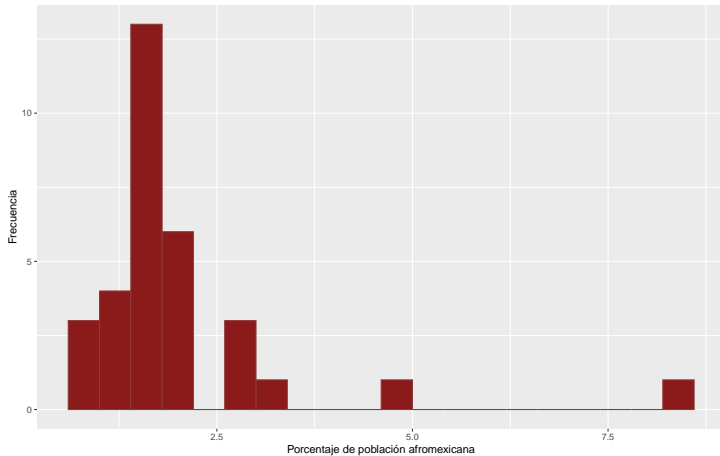
Una buena gráfica lleva:

- Título (y un subtítulo opcional)
- Descripción de los ejes
- La fuente de tus datos

```
plot_b <- ggplot(data2020, aes(x = Pob_afro_share)) +  
  geom_histogram( fill = "firebrick4", color = "indianred4",  
                 binwidth=0.4) +  
  labs(  
    x = "Porcentaje de población afromexicana",  
    y = "Frecuencia",  
    title = "Distribución del porcentaje de población afromexicana",  
    subtitle = "Histograma con datos desagregados por estado",  
    caption = "Fuente: Censo poblacional INEGI 2020"  
  )
```

```
print(plot_b)
```

Distribución del porcentaje de población afromexicana
Histograma con datos desagregados por estado



Fuente: Censo poblacional INEGI 2020

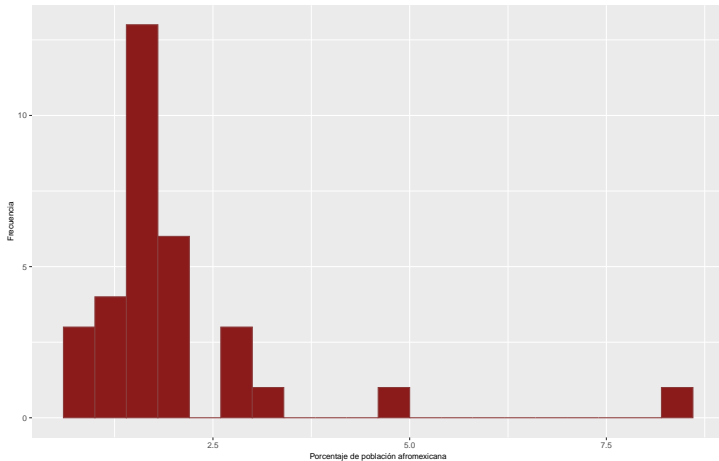
Cambiamos el estilo de nuestras labs

```
plot_c <- ggplot(data2020, aes(x = Pob_afro_share)) +  
  geom_histogram( fill = "firebrick4", color = "indianred4",  
                 binwidth=0.4) +  
  labs(  
    x = "Porcentaje de población afromexicana",  
    y = "Frecuencia",  
    title = "Distribución del porcentaje de población afromexi",  
    subtitle = "Histograma con datos desagregados por estado",  
    caption = "Fuente: Censo poblacional INEGI 2020"  
  ) +  
  theme(  
    plot.title=element_text(size=14, face="bold"),  
    plot.subtitle=element_text(size=8),  
    plot.caption=element_text(size=6),  
    axis.title=element_text(size=8),  
    axis.text=element_text(size=8)  
  )
```

```
print(plot_c)
```

Distribución del porcentaje de población afromexicana

Histograma con datos desagregados por estado



Fuente: Censo poblacional INEGI 2020

Parte 2: Prueba de hipótesis

Objetivo: inferir sobre una población a partir de una muestra.

Conceptos centrales: hipótesis nula (H_0) e hipótesis alternativa (H_1).

1. Planteamiento de H_0 y H_1

El primer paso es definir claramente qué queremos probar.

Hipótesis nula (H_0): no existe diferencia en la población católica entre 2010 y 2020.

Hipótesis alternativa (H_1): sí existe diferencia, es decir, la población católica cambió entre esos años.

2. Selección del nivel de significancia α

Dado que estamos comparando el mismo número de observaciones, ocuparemos un t-test.

Comúnmente se usa

$$\alpha=0.05$$

Representa la probabilidad de rechazar H_0 cuando es verdadera (error tipo I).

3. Cálculo del p-value

R lo calcula automáticamente con `t.test(..., paired = TRUE)`

El p-value indica la probabilidad de observar una diferencia igual o más extrema que la actual si H_0 fuera cierta.

4. Decisión: rechazar o no rechazar H_0

Si $p\text{-value} < \alpha$ (0.05): rechazar $H_0 \rightarrow$ hay evidencia estadística de que la población católica cambió entre 2010 y 2020.

Si $p\text{-value} \geq \alpha$: no se rechaza $H_0 \rightarrow$ no hay evidencia suficiente para afirmar que cambió.

#Población católica entre 2010 y 2020

¿ha cambiado la población católica en México entre 2010 y 2020?

¿Cuáles serían H_0 y H_1 ?

Hipótesis

H_0 : católicos2020 = católicos 2010

H_1 : católicos2020 \neq católicos 2010

Creemos nuestros datos y hagamos la prueba. ¿Qué concluimos?

```
data2010 <- data %>%  
  filter(Año == "2010")  
  
t.test(data2020$Pob_catolica, data2010$Pob_catolica,  
        alternative = "two.sided")
```

Welch Two Sample t-test

```
data: data2020$Pob_catolica and data2010$Pob_catolica  
t = 0.31453, df = 61.855, p-value = 0.7542  
alternative hypothesis: true difference in means is not equal  
95 percent confidence interval:  
 -1005451  1380914  
sample estimates:  
mean of x mean of y  
 2819517  2631786
```