3. Análisis

La idea de este apartado es repasar algunas herramientas de estadística descriptiva para realizar análisis de datos. Se utilizará de insumo la Encuesta Nacional de Hogares (EAH) de la Ciudad de Buenos Aires.

```
library(tidyverse)
```

La EAH es una encuesta anual por muestreo que se propone recabar datos para conocer y analizar la situación socioeconómica y demográfica de la población y de los hogares de la Ciudad. Si bien lo correcto sería tomar y utilizar el factor de expansión, prescindiremos para facilitar el trabajo sobre la base.

```
ruta <- "data/encuentro_3/EAH_2023_ind.csv"

columnas <- c("Número de vivienda","Número de hogar","Número de miembro","Comuna","Monto del
eah <- read_csv(ruta, # seleccionamos archivo
                  col_select=columnas) # elegimos qué columnas cargar

col_nuevas <- c("num_vivienda","num_hogar","num_individuo","comuna","ingreso_per_capita_fami
names(eah) = col_nuevas # renombramos

eah <- eah %>% filter(ingreso_per_capita_familiar > 0) # eliminamos casos sin ingresos

dim(eah)
```

```
[1] 13000     5
```

```
head(eah)
```

```
# A tibble: 6 x 5
  num_vivienda num_hogar num_individuo comuna ingreso_per_capita_familiar
         <dbl>     <dbl>         <dbl>  <dbl>                       <dbl>
1            1         1             1      2                      140040
2            1         2             1      2                      140000
```

3	1	3	1	2	70000
4	1	4	1	2	70000
5	2	1	1	11	215000
6	2	1	2	11	215000

3.1. Repasando conceptos básicos

¿Por qué hablamos de **muestra**? se llama muestra a un subconjunto de una población que es captado para analizar a una **población** específica. Con población nos referimos al conjunto de todos los elementos que forman parte de un universo de interés.

Existen distintas **medidas resumen** para sintetizar y describir las características principales de un conjunto de datos. Permiten obtener una visión general y son fundamentales para el análisis exploratorio.

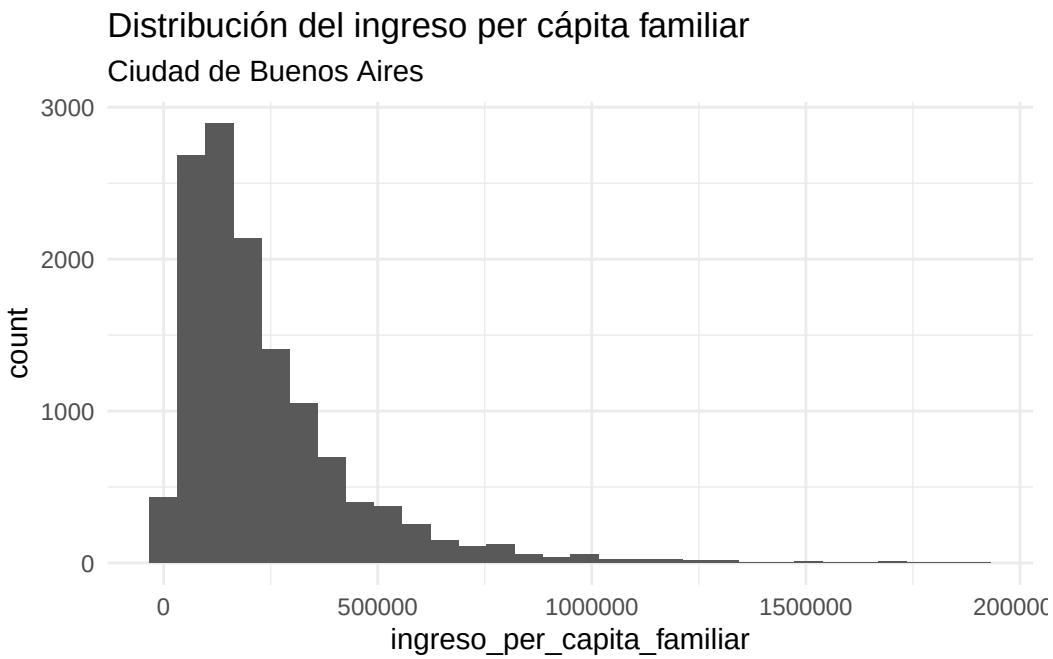
En este apartado vamos a trabajar con la distribución del ingreso familiar per cápita en CABA. Una distribución es simplemente un conjunto de datos determinado; en este caso, una muestra proveniente de la encuesta antedicha.

```
# al ser más de 13 mil casos, es imposible verlos en formato texto
eah$ingreso_per_capita_familiar[1:10]
```

```
 [1] 140040 140000  70000  70000 215000 215000  70000 268333 268333 268333
```

Para visualizar una gran cantidad de datos podemos usar un tipo de gráfico llamado **histograma**. Veremos más sobre esto en el siguiente apartado. El eje X representa los valores presentes en la distribución y el eje Y representa la cantidad de apariciones de cada uno de esos valores.

```
eah %>%
  filter(ingreso_per_capita_familiar < 2000000) %>% # filtramos algunos casos para mejorar la
  ggplot(aes(ingreso_per_capita_familiar))+
  geom_histogram()+
  theme_minimal()+
  labs(title="Distribución del ingreso per cápita familiar", subtitle="Ciudad de Buenos Aires
```



El primer conjunto de medidas resumen que veremos son las **medidas de tendencia central**.

- **Media**: más conocida como promedio. Es la suma de todos los valores dividida la cantidad de valores. Su sensibilidad a los valores atípicos la vuelve poco representativa en ciertas distribuciones.

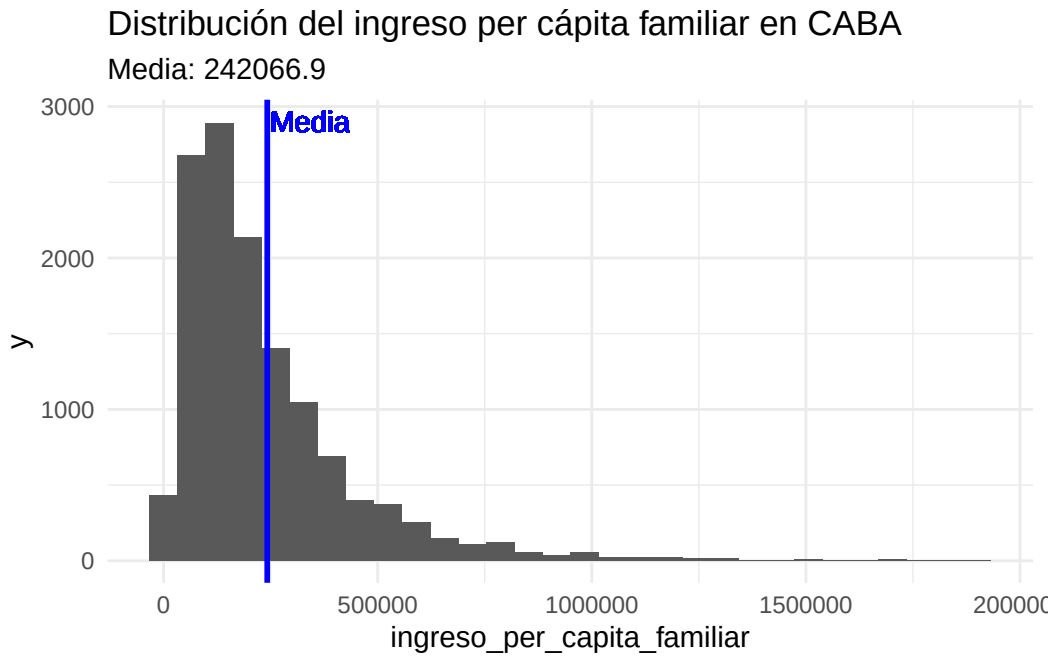
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Figure 1: Fórmula media

- **Mediana**: se obtiene ordenando todos los valores de menor a mayor y tomando el valor que se encuentra justo en la mitad. Evita la sensibilidad a valores atípicos.
- **Moda**: es la observación con mayor frecuencia en la distribución. Puede existir más de una.

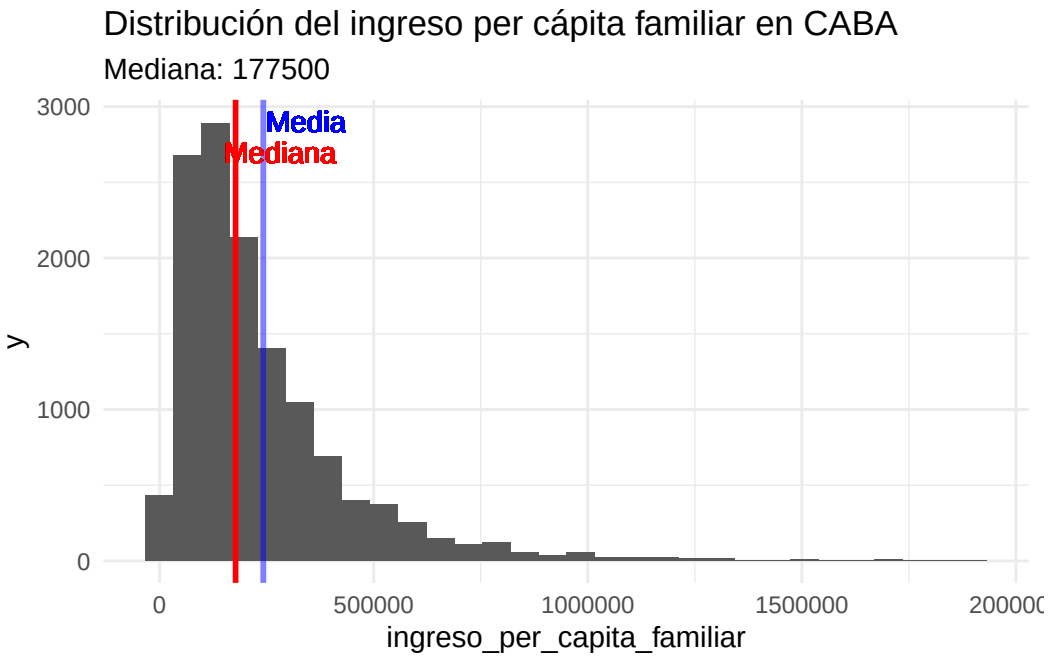
```r
# guardamos los valores
media <- mean(eah$ingreso_per_capita_familiar)
mediana <- median(eah$ingreso_per_capita_familiar)

# veamos la media
eah %>%
  filter(ingreso_per_capita_familiar < 2000000) %>% # filtramos algunos casos para mejorar la
  ggplot()+
  geom_histogram(aes(ingreso_per_capita_familiar))+
  geom_vline(xintercept= media, color="blue", size=1)+
  geom_text(aes(x=media+1e5, y=2900, label="Media"), color="blue")+
  theme_minimal()+
  labs(title="Distribución del ingreso per cápita familiar en CABA", subtitle=paste0("Media:
```




```r
# veamos la mediana
eah %>%
  filter(ingreso_per_capita_familiar < 2000000) %>% # filtramos algunos casos para mejorar la
  ggplot()+
  geom_histogram(aes(ingreso_per_capita_familiar))+
  geom_vline(xintercept= media, color="blue", size=1, alpha=.5)+
  geom_vline(xintercept= mediana, color="red", size=1)+
  geom_text(aes(x=media+1e5, y=2900, label="Media"), color="blue")+
```

```
  geom_text(aes(x=media+4e4, y=2700, label="Mediana"), color="red")+
  theme_minimal()+
  labs(title="Distribución del ingreso per cápita familiar en CABA", subtitle=paste0("Median
```



El segundo conjunto de medidas resumen que veremos son las **medidas de variabilidad**.

- **Varianza**: es la media de las desviaciones cuadráticas respecto de la media. Se elevan al cuadrado para evitar la compensación entre números positivos y negativos.

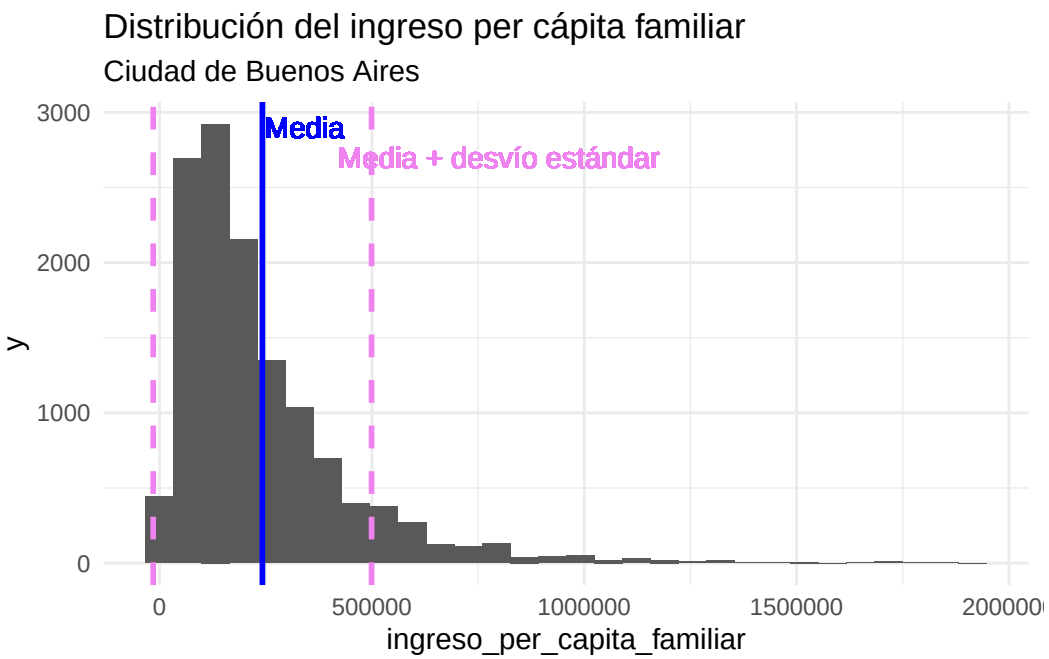
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Figure 2: Fórmula varianza

- **Desvío estándar**: es la raíz cuadrada de la varianza; se utiliza para hacer interpretable la métrica.

```r
# guardamos los valores
sd <- sd(eah$ingreso_per_capita_familiar)

# veamos la media
eah %>%
  filter(ingreso_per_capita_familiar < 2000000) %>% # filtramos algunos casos para mejorar la
  ggplot()+
  geom_histogram(aes(ingreso_per_capita_familiar))+
  geom_vline(xintercept= media, color="blue", size=1)+
  geom_vline(xintercept= media-sd, color="violet", size=1, linetype = "dashed")+
  geom_vline(xintercept= media+sd, color="violet", size=1, linetype = "dashed")+
  geom_text(aes(x=media+1e5, y=2900, label="Media"), color="blue")+
  geom_text(aes(x=media+sd+3e5, y=2700, label="Media + desvío estándar"), color="violet")+
  theme_minimal()+
  labs(title="Distribución del ingreso per cápita familiar", subtitle=paste0("Ciudad de Buen
```



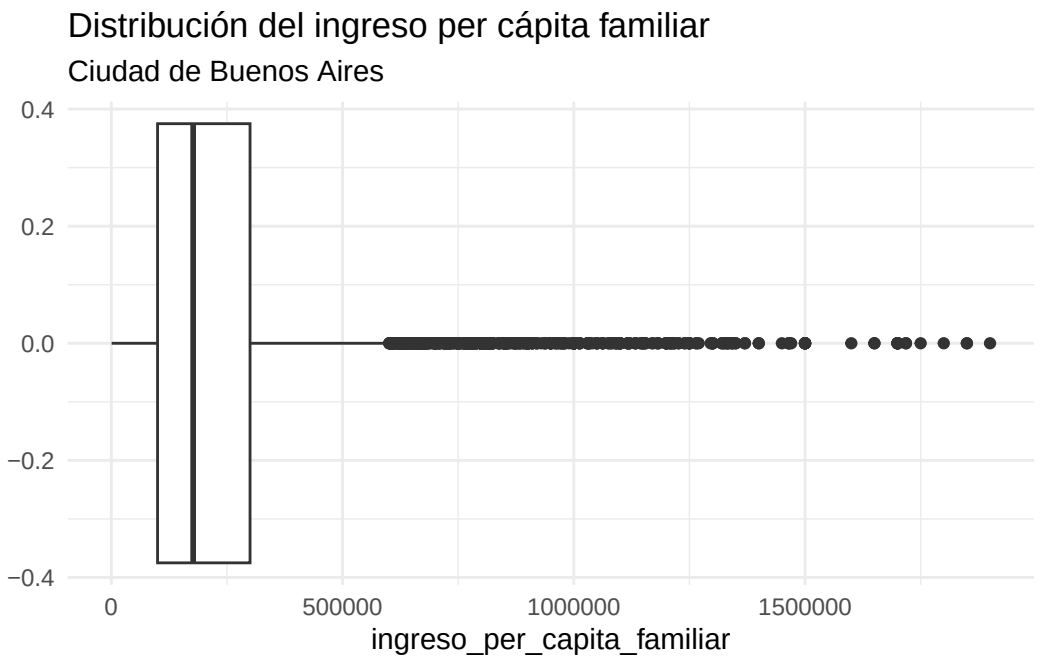
El tercer conjunto de medidas resumen que veremos son las **medidas de posición**. En general trabajamos con **cuantiles**, valores que dividen la distribución en una cantidad arbitraria de partes iguales. Se suelen usar los quintiles, que vimos en el primer encuentro con la función `summary()`.

```r
summary(eah$ingreso_per_capita_familiar)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    643  100000  177500  242067  300000 8134000
```

Una forma típica de ver una distribución según sus quintiles es el gráfico de cajas o bigotes (*boxplot* en inglés). Si un histograma sirve para ver una distribución en particular, los boxplots van a servirnos para comparar distintas distribuciones.

```r
eah %>%
  filter(ingreso_per_capita_familiar < 2000000) %>% # filtramos algunos casos para mejorar la
  ggplot()+
  geom_boxplot(aes(ingreso_per_capita_familiar))+
  theme_minimal()+
  labs(title="Distribución del ingreso per cápita familiar", subtitle=paste0("Ciudad de Buen
```




```r
f <- function(x, pos){
  filter(x,(cargo_nombre == "PRESIDENTE Y VICE"))
}
#data <- read_csv_chunked(ruta, DataFrameCallback$new(f), chunk_size=10000)
#dim(data) # vemos cuántas filas y columnas tiene
#head(data) # vemos las primeras 5 filas
```