Practica 1

Gonzalo Cruz Gómez, Lucía Arnaldo Cuevas

2024-10-20

#Introducción

Práctica 1 de la asignatura de Inferencia Estadística. Se trabajará con los datos de los arrestos en EEUU de la librería 'datasets'

#Pregunta 1

EDA

Cargo los datos de la librería datasets y las librerías que vamos a usar en la práctica

```
library(datasets)
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
datos=datasets::USArrests
```

Estos datos representan los arrestos por cada 100000 habitantes en cada estado de EEUU en 1973, además del porcentaje de la población que vive en las ciudades

```
dim(datos)
```

```
## [1] 50  4
```

```
str(datos)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

Estas variables representan:

- Murder: numeric Murder arrests (per 100,000)
- Assault: numeric Assault arrests (per 100,000)
- UrbanPop: numeric Percent urban population

- Rape: numeric Rape arrests (per 100,000)

Todas las variables son variables continuas, pueden tomar cualquier valor, no tenemos variables discretas, para ello habría que discretizar una variable continua. Hay 50 observaciones y 4 características medidas. No tenemos variables de tipo texto ni variables irrelevantes

#Pregunta 2

Estadísticos resumen de todas las variables (cuartiles, mediana, media, mínimos y máximos), además de las desviaciones típicas de cada variable:

```
summary(datos)
```

```
##      Murder          Assault         UrbanPop          Rape      
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30  
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07  
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10  
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23  
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18  
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

```
print("desviaciones típicas de cada variable: \n")
```

```
## [1] "desviaciones típicas de cada variable: \n"
```

```
sd(datos$Murder)
```

```
## [1] 4.35551
```

```
sd(datos$Assault)
```

```
## [1] 83.33766
```

```
sd(datos$Rape)
```

```
## [1] 9.366385
```

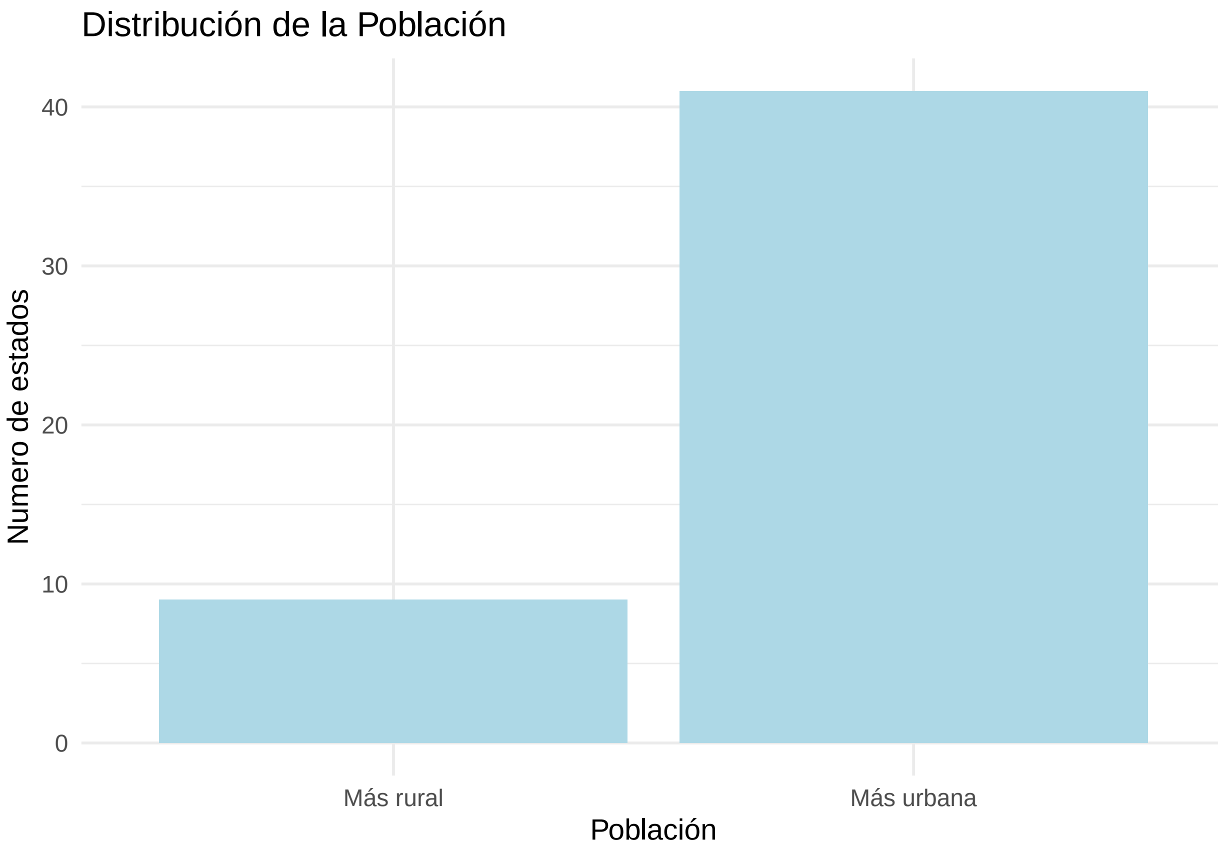
```
sd(datos$UrbanPop)
```

```
## [1] 14.47476
```

#Pregunta 3

Como no tenemos variables discretas, hemos optado por discretizar la variable UrbanPop

```
discretizada <- ifelse(USArrests$UrbanPop > 50, "Más urbana", "Más rural")

# Crear un gráfico de barras para visualizar la discretización
ggplot(USArrests, aes(x = discretizada)) +
  geom_bar(fill = "lightblue") +
  labs(title = "Distribución de la Población",
       x = "Población",
       y = "Numero de estados") +
  theme_minimal()
```

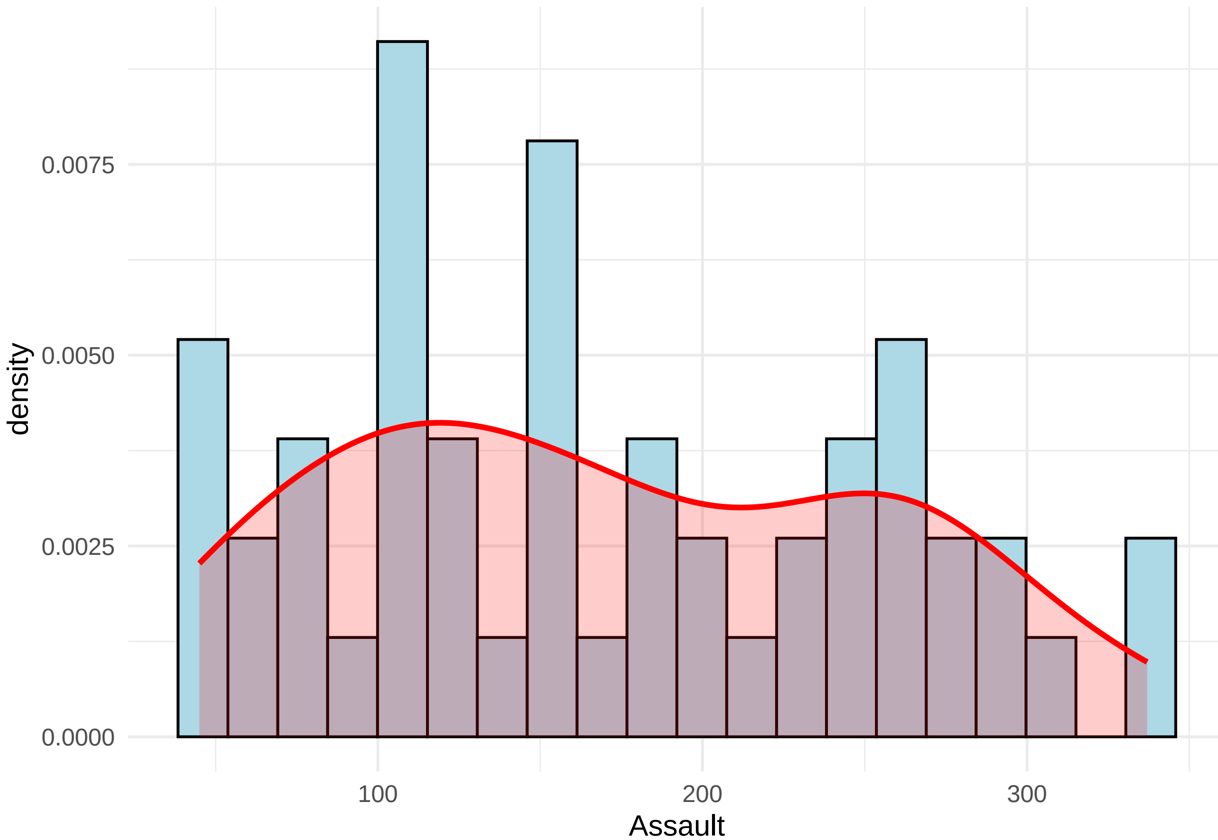


Esta gráfica de la distribucion de la poblacion demuestra que claramente hay más estados en los que predomina la población urbana que estados en los que predomina la población rural

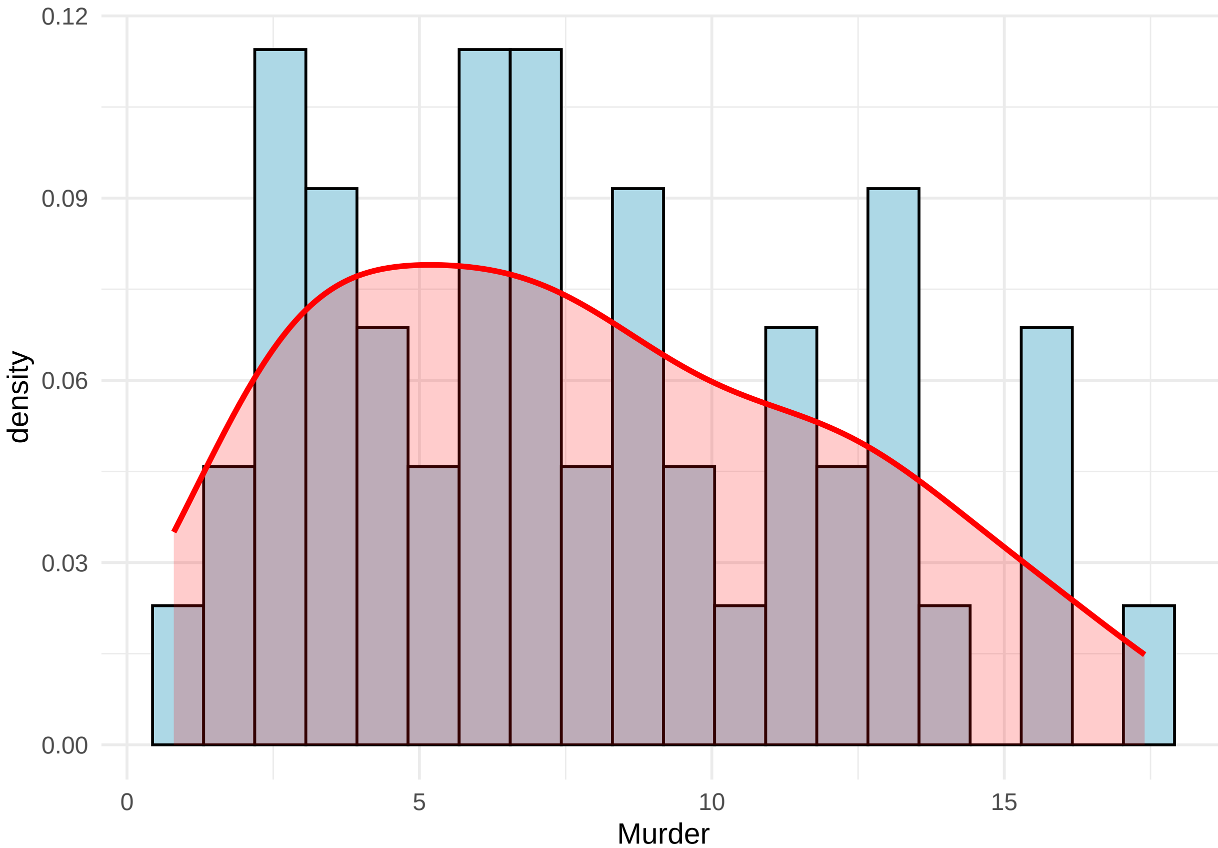
#Pregunta 4

Distribución de los asesinatos, de los asaltos y de las violaciones, que son nuestras variables continuas

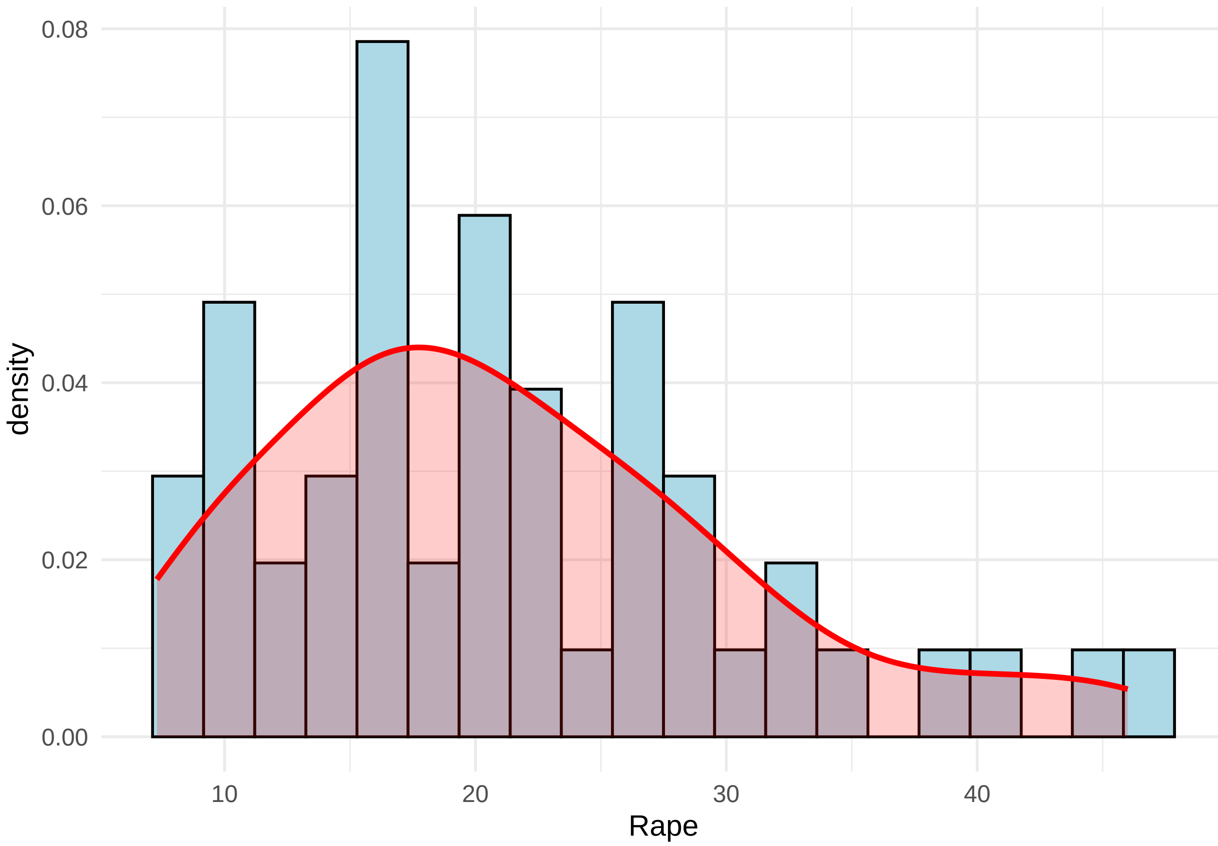
```r
library(ggplot2)

ggplot(datos, aes(x = Assault)) +
  geom_histogram(aes(y = after_stat(density)),
                 colour = 1, fill = "lightblue", bins=20) +
  geom_density(lwd = 1, colour = "red",
               fill = "red", alpha = 0.2) + theme_minimal()
```




```
ggplot(datos, aes(x = Murder)) +
  geom_histogram(aes(y = after_stat(density)),
                 colour = 1, fill = "lightblue", bins=20) +
  geom_density(lwd = 1, colour = "red",
               fill = "red", alpha = 0.2) + theme_minimal()
```




```
ggplot(datos, aes(x = Rape)) +
  geom_histogram(aes(y = after_stat(density)),
                 colour = 1, fill = "lightblue", bins=20) +
  geom_density(lwd = 1, colour = "red",
               fill = "red", alpha = 0.2) + theme_minimal()
```



La variable Assaults se aproxima relativamente bastante a una distribución uniforme, de ello podemos extraer que la cantidad de asaltos es más o menos igual en todo el territorio. En cuanto a las violaciones y los asesinatos, ambas variables se aproximan a una distribución normal aunque con una media poco centrada, por lo cual podemos concluir que es más común que haya un numero menor de violaciones y asesinatos que una gran cantidad de ellos

#Pregunta 5

Vamos a hacerlo para una distribución normal para luego compararlo con la distribución de las violaciones

Se trata de una distribución normal con parámetros $\mathcal{N}(\mu, \sigma^2)$ para una muestra de n datos $x_1, x_2, \ldots, x_n$ con una función de densidad de probabilidad:

$$f(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Su función de verosimilitud es:

$$L(\mu, \sigma^2|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Ahora hacemos la log-verosimilitud, para ello tomamos logaritmos

$$\ell(\mu, \sigma^2|x) = \log L(\mu, \sigma^2|x_1, x_2, \ldots, x_n)$$

$$\ell(\mu, \sigma^2|x) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Para hallar el estimador ML de μ , derivamos con respecto a μ e igualamos a 0

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

$$-\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0$$

$$\sum_{i=1}^{n}(x_i-\mu)=0$$

$$\sum_{i=1}^{n}x_i-n\mu=0$$

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

Hacemos lo mismo para σ^2 partiendo de la log-verosimilitud

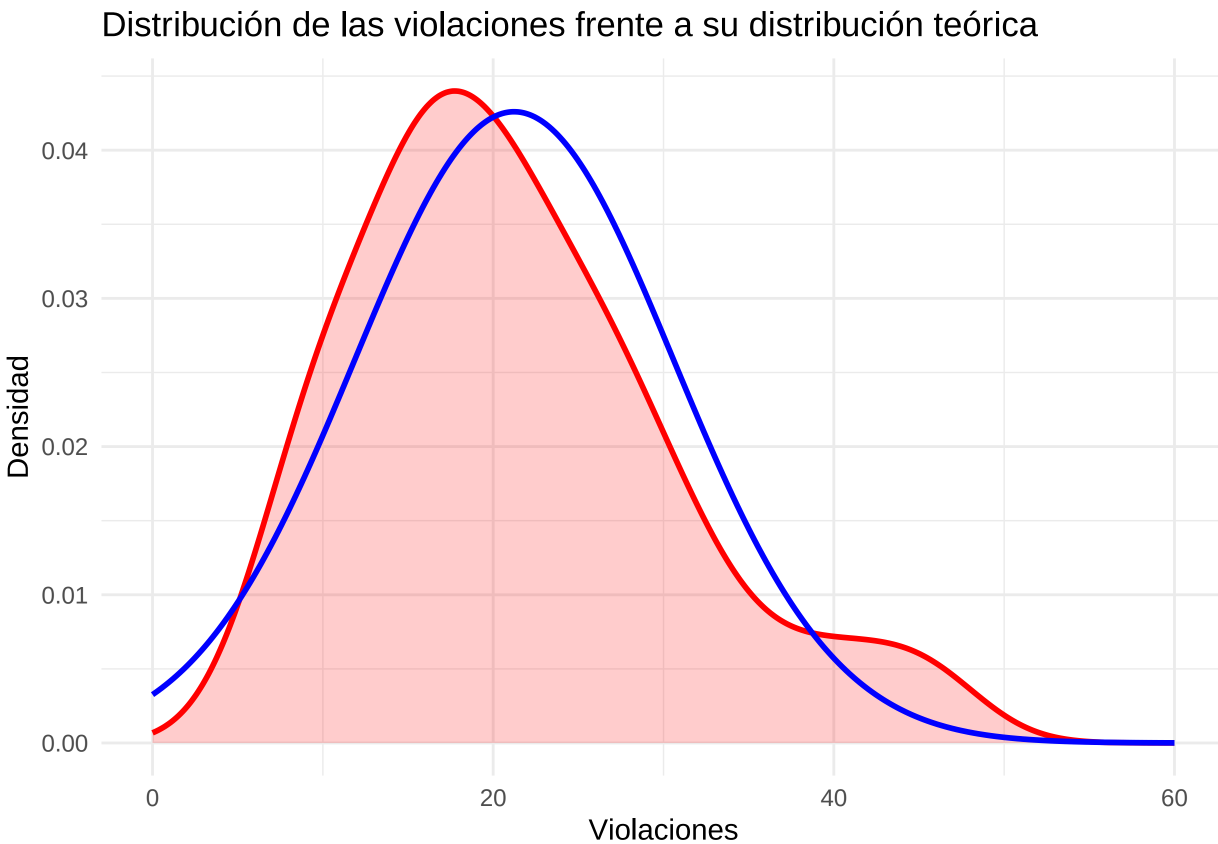
$$\frac{\partial \ell}{\partial \sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Por lo que podemos concluir que los estimadores ML para μ y σ^2 son la media muestral y la varianza muestral

```r
mean_value <- 21.23
sd_value <- 9.366385
x <- seq(0, 60, length.out = 1000)
y <- dnorm(x, mean = mean_value, sd = sd_value)
# Crear un dataframe para ggplot
normal <- data.frame(x = x, y = y)

ggplot(USArrests, aes(x = Rape)) +
  geom_density(lwd = 1, colour = "red", fill = "red", alpha = 0.2) +
  geom_line(data = normal, aes(x = x, y = y), colour = "blue", lwd = 1) +
  theme_minimal() +
  labs(title = "Distribución de las violaciones frente a su distribución teórica",
       x = "Violaciones",
       y = "Densidad")
```



#Pregunta 6

Utilizando el método de los momentos:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Tenemos que igualar los momentos muestrales con los momentos teóricos:

$$\mu_1 = E(X) = \mu$$

$$\mu_2 = E(X^2) = \mathrm{Var}(X) + E(X)^2 = \mu^2 + \sigma^2$$

$$m_1 = \bar{X} = \mu$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu_2 = \mu^2 + \sigma^2$$

Resolvemos el sistema

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\hat{\mu} = \bar{X}$$

Y como resultado tenemos los estimadores de los parámetros Los valores de estos estimadores son:

```
print(paste("media =", mean(USArrests$Rape)))
```

```
## [1] "media = 21.232"
```

```
print(paste("desviación típica = " , sd(datos$Rape)))
```

```
## [1] "desviación típica =  9.36638453105965"
```

!!!!falta calcular el intervalo de confianza!!!!

#Pregunta 7

##Apartado a

Vamos a calcular la probabilidad de que haya más de 10 asesinatos, sabiendo que la media es 7,78 y la desviación típica es 4.355 (pregunta 2)

$$\text{Murder} \sim N(\mu, \sigma^2)$$
$$Z = \frac{X - \mu}{\sigma}$$
$$P(\text{Murder} > x) = P\left(Z > \frac{x - \mu}{\sigma}\right)$$
$$z = \frac{10 - 7.788}{4.355} \approx 0.507$$
$$P(Z > 0.507) = 1 - 0.6943 = 0.3057$$

Por lo tanto la probabilidad de que en un estado haya mas de 10 asesinatos por cada 100000 habitantes es de aproximadamente un 30%

##Apartado b

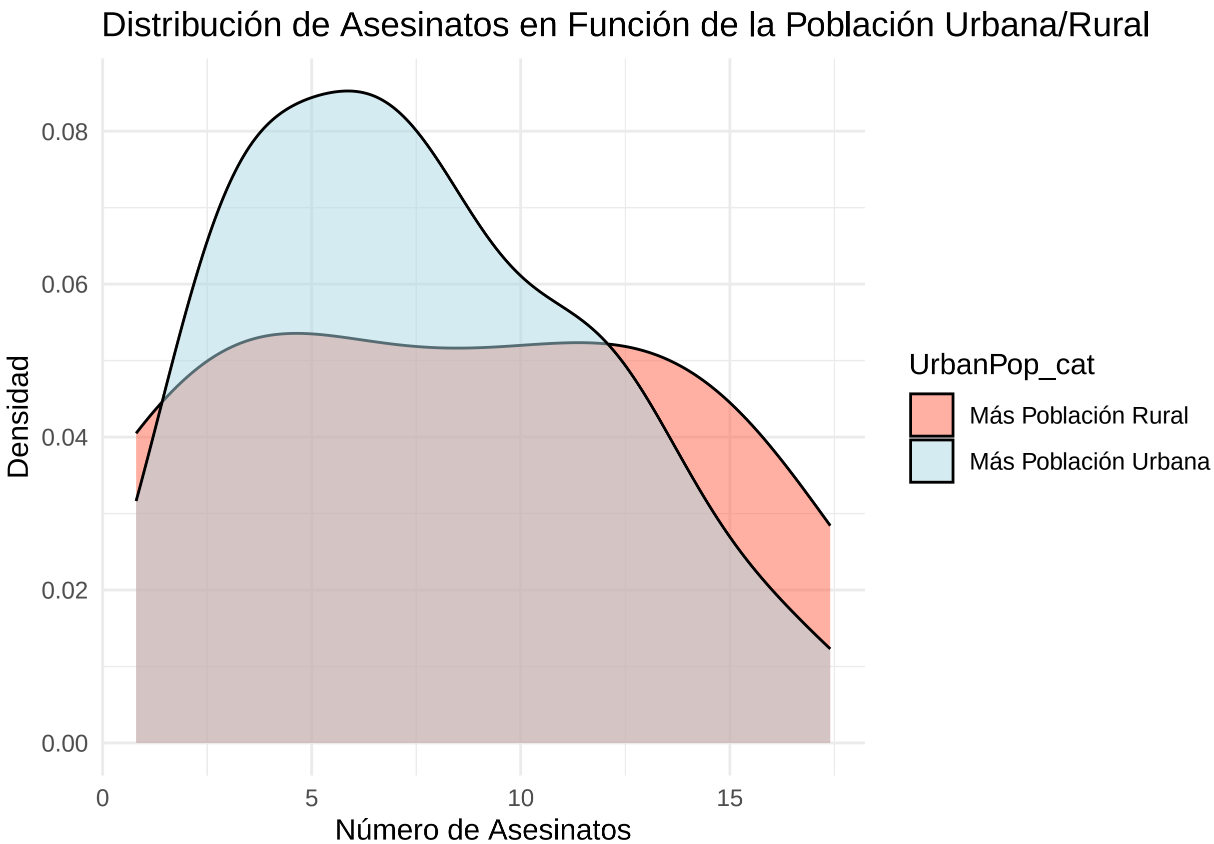
Si ahora en vez de calcular el valor de manera teórica, lo hacemos mediante una simulación en r:

```
x <- 10
n <- 10000
simulacion <- rnorm(n, mean = mean(USArrests$Murder), sd = sd(USArrests$Murder))
prob <- mean(simulacion > x)
print(paste("Probabilidad de que haya más de 10 asesinatos = ", prob))
```

```
## [1] "Probabilidad de que haya más de 10 asesinatos =  0.3141"
```

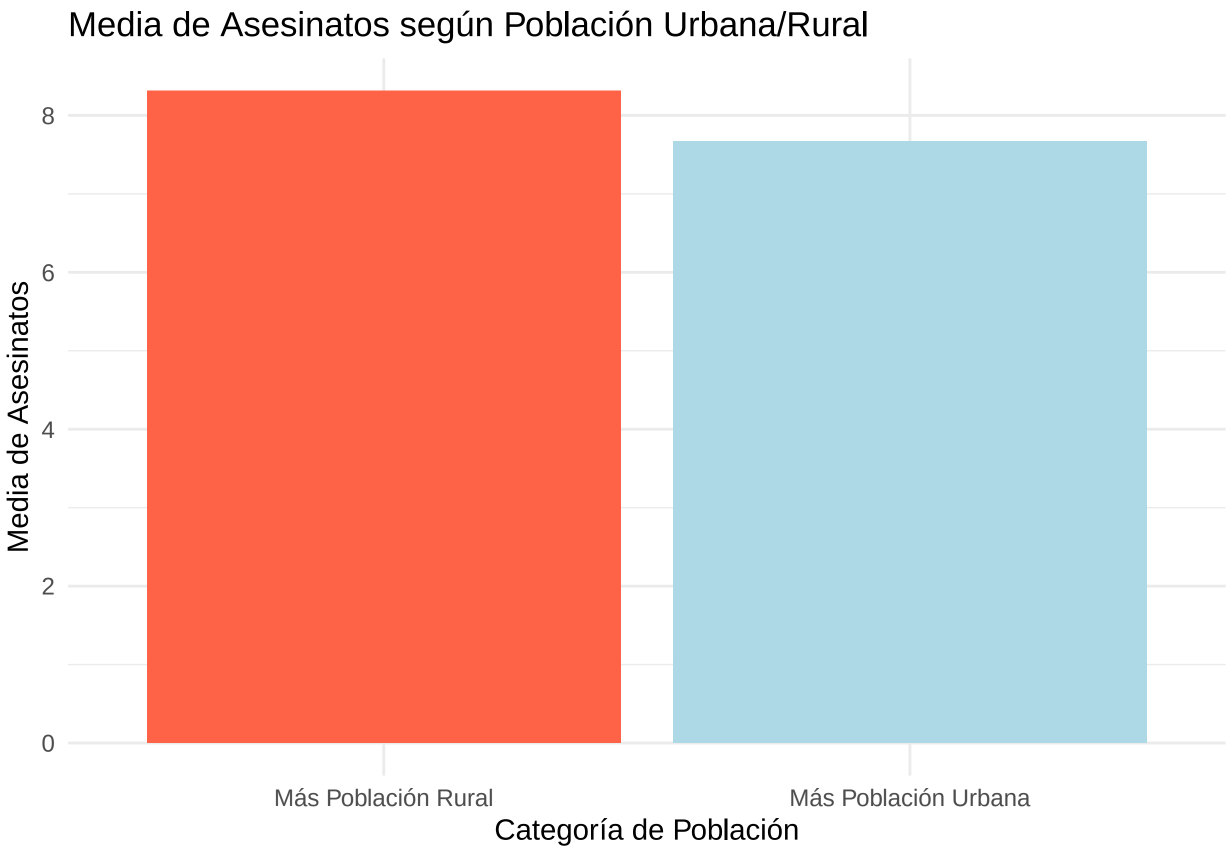
#Pregunta 8

```
# Discretizar la variable UrbanPop
USArrests$UrbanPop_cat <- ifelse(USArrests$UrbanPop > 50, "Más Población Urbana", "Más Población Rural")

# Crear el gráfico de densidad para comparar la variable Murder
ggplot(USArrests, aes(x = Murder, fill = UrbanPop_cat)) +
  geom_density(alpha = 0.5, show.legend = TRUE) +
  labs(title = "Distribución de Asesinatos en Función de la Población Urbana/Rural",
       x = "Número de Asesinatos",
       y = "Densidad") +
  scale_fill_manual(values = c("Más Población Urbana" = "lightblue", "Más Población Rural" = "tomato"))
  theme_minimal()
```



Con esto podemos observar que la distribución de los asesinatos en los estados con mas población rural se aproxima mucho a una distribución uniforme, mientras que en los estados con población más urbana se acerca más a una normal.

```
# Discretizar la variable UrbanPop
USArrests$UrbanPop_cat <- ifelse(USArrests$UrbanPop > 50, "Más Población Urbana", "Más Población Rural")

# Calcular la media de asesinatos por categoría
murder_summary <- USArrests %>%
  group_by(UrbanPop_cat) %>%
  summarize(Media_Asesinatos = mean(Murder))

# Crear el gráfico de barras
ggplot(murder_summary, aes(x = UrbanPop_cat, y = Media_Asesinatos, fill = UrbanPop_cat)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  labs(title = "Media de Asesinatos según Población Urbana/Rural",
       x = "Categoría de Población",
       y = "Media de Asesinatos") +
  scale_fill_manual(values = c("Más Población Urbana" = "lightblue", "Más Población Rural" = "tomato"))
  theme_minimal()
```



Aunque la distribución de los asesinatos pueda indicar lo contrario, la media de asesinatos en poblacion rural es ligeramente superior a la urbana. La anterior gráfica puede llevar a equivocación debido a que hay muchos más estados con más población urbana que rural, por lo cual más asesinatos se cometen en estados urbanos, a pesar de que de media se cometan más en los estados rurales

#Pregunta 9

Teniendo en cuenta los resultados anteriores, podemos plantear el siguiente contraste de hipótesis: Hay el mismo número de asesinatos independientemente de si más población vive en las ciudades o en los pueblos