Tarea 1

Estadística Actuarial II

José Ignacio Rojas Zárate, C16911 Montserrat Beirute Abarca, C10997
Valeria Vásquez Venegas, C18373

17 de enero de 2024

Índice

1. Introducción **1**

2. Parte I **3**

2.1. Análisis descriptivo de las variables Cuotas y Salarios con respecto a la variable Sexo 3

2.2. Gráfico plotbox para el salario para comparar entre las categorías de Sexo 6

2.3. Conclusiones con respecto a los salarios y sexo 6

2.4. Prueba de hipótesis sobre las medias de las categorías de sexo 7

3. Parte II **8**

3.1. Construcción del Histograma de los salarios 8

3.2. Densidad de los salarios por kernel (no paramétrica) 9

1. Introducción

El presente documento prestan el código y resultados a los ejercicios de la Tarea 1 del curso. Para su realización, se utilizó el programa R studio.

Este trabajo requirió de los siguientes paquetes:

```
#---Paquetes-------------------------------------------

library(stringr) # para cambiar todas las commas a puntos en la columna Ultimo Salario
library(ggplot2) # para hacer los gráficos
library(univariateML) # para hacer el análisis AIC
library(rriskDistributions) # para hacer el análisis AIC
library(boot) # para hacer el bootstrap
library(ks) # para Kernel smoothing
library(knitr) # para presentar mejor los datos en formato tabla
options(scipen = 999) # para no utilizar notación científica
library(cowplot) # para hacer graficos más atractivos
```

El presente trabajo utilizó la base de datos “BaseSalarios” brindada por el profesor Esteban Bermúdez Aguilar. A continuación se presenta el código para la lectura de la base de datos, así como también unos ajustes para poder manipular y extraer la información de los datos.

```
#---CSV y formatos----------------------------------------

# Abrir y dar formato a la base de datos

base_salarios <- read.csv("BaseSalarios.csv",  sep = ";")

base_salarios$Fec.Nac <- as.Date(base_salarios$Fec.Nac, format = "%d/%m/%Y")

colnames(base_salarios)[5] ="Ultimo.Salario"

base_salarios <- subset(base_salarios, select = -X)

base_salarios$Ultimo.Salario <- as.numeric(str_replace_all(base_salarios$Ultimo.Salario, ",", "."))
```

La base de datos tiene 5 columnas. A continuación, se presentan los primeros seis registros de la base de datos “BaseSalarios” para poder explicar la información de cada columna.

```
kable(head(base_salarios_temp), caption = "Primeros 6 registros de la base de datos", align = "r")
```

Cuadro 1: Primeros 6 registros de la base de datos

ID	Fec.Nac	Sexo	Coutas	Ultimo.Salario
1	1945-01-30	1	323	2 784 093.18
2	1949-06-17	2	342	532 324.93
3	1951-04-19	2	13	225 517.69
4	1950-02-05	2	323	1 612 733.09
5	1952-02-06	2	185	1 411 502.30
6	1954-01-10	2	284	1 915 751.19

La columna **ID** cuenta el número de registros disponibles, en este caso, se dispone de 106,002 registros. En la columna **Fec.Nac** se encuentra la información de la fecha de nacimiento de cada persona. En la tercera columna, **Sexo**, se observan los números 1 y 2; específicamente, si el valor de Sexo es 1, corresponde a hombres, y si es 2, corresponde a mujeres. La cuarta columna, **Cuotas** indica la cantidad de cuotas que cada persona ha aportado a un fondo de pensiones. Finalmente, la columna **Ultimo.Salario** contiene el último salario reportado, expresado en colones.

Es importante señalar que se tomó la decisión de eliminar dos registros de la base de datos, ya que se consideraron datos anómalos, es decir, outliers. Estos dos registros correspondían a salarios de 13 199 892 y 13 110 539 colones y estaban asociados a individuos del género masculino. Ambos salarios superaban la media salarial de hombres en al menos 11 953 340 colones.

```
base_salarios<- base_salarios[-84807, ] # corresponde al salario 13 199 892
base_salarios<- base_salarios[-96367, ] # corresponde al salario 13 110 539
```

2. Parte I

2.1. Análisis descriptivo de las variables Cuotas y Salarios con respecto a la variable Sexo

De las 106,000 personas que contiene la base de datos utilizada, el 69.1 % son mujeres (73,279 mujeres). Por su parte, el 30.1 % restante corresponde a hombres (32,721 hombres).

Es de interés conocer las diferencias o similitudes en los datos del número de cuotas y los salarios según el sexo. Para ello, se presenta el siguiente código:

```
resumen_cuotas_hombres <-summary(base_salarios$Coutas[base_salarios$Sexo == 1])

varianza_cuotas_hombres <- var(base_salarios$Coutas[base_salarios$Sexo == 1])

resumen_cuotas_hombres <- c(resumen_cuotas_hombres, Varianza = varianza_cuotas_hombres)

resumen_cuotas_mujeres <-summary(base_salarios$Coutas[base_salarios$Sexo == 2])

varianza_cuotas_mujeres <- var(base_salarios$Coutas[base_salarios$Sexo == 2])

resumen_cuotas_mujeres <- c(resumen_cuotas_mujeres, Varianza = varianza_cuotas_mujeres)

cuotas_resumen <- data.frame(
  Estadistico = c("Mínimo", "Primer cuartil (Q1)", "Mediana", "Promedio", "Tercer cuartil (Q3)", "Máxim
  Hombres = as.numeric(round(resumen_cuotas_hombres,2)),
  Mujeres = as.numeric(round(resumen_cuotas_mujeres,2))
)
```

A continuación, se presenta un cuadro con un resumen estadístico de los datos del número de cuotas para hombres y mujeres.

Cuadro 2: Resumen estadístico de número de coutas por sexo

Estadistico	Hombres	Mujeres
Mínimo	1.00	1.00
Primer cuartil (Q1)	61.00	66.00
Mediana	124.00	135.00
Promedio	135.19	142.97
Tercer cuartil (Q3)	198.00	213.00
Máximo	371.00	373.00
Varianza	7658.22	8187.47

Como se puede observar, tanto para mujeres como para hombres, el mínimo de cuotas aportadas es tan solo una.

Por su parte, se puede decir que en general y se va a justificar a continuación, las mujeres de la base de datos aportaron más cuotas que los hombres.

Se puede notar que el 25 % de los hombres aportó 61 cuotas o menos, mientras que en el caso de las mujeres, el 25 % aportó 66 cuotas o menos.

Además, el 50 % de las mujeres aportó más de 135 cuotas, mientras que en el caso de los hombres, fue de 124 cuotas.

El promedio de cuotas es más alto para las mujeres (142.97) que para los hombres (135.19).

También, el 75 % de las mujeres aportó 213 cuotas o menos, mientras que en el caso de los hombres, fue de 198 cuotas.

El valor máximo de cuotas aportadas para las mujeres fue de 373, mientras que para los hombres fue de 371.

Es importante resaltar que para ambos sexos, la varianza indica que los datos están alejados de la media y se presenta mucha variabilidad en los datos de número de cuotas aportadas al fondo de pensiones.

Por otra parte, a continuación, se presenta un cuadro con un resumen estadístico de los datos del último salario reportado para hombres y mujeres.

```
resumen_salarios_hombres <-summary(base_salarios$Ultimo.Salario[base_salarios$Sexo == 1])

varianza_salarios_hombres <- var(base_salarios$Ultimo.Salario[base_salarios$Sexo == 1])

resumen_salarios_hombres <- c(resumen_salarios_hombres, Varianza = varianza_salarios_hombres)

resumen_salarios_mujeres <-summary(base_salarios$Ultimo.Salario[base_salarios$Sexo == 2])

varianza_salarios_mujeres <- var(base_salarios$Ultimo.Salario[base_salarios$Sexo == 2])

resumen_salarios_mujeres <- c(resumen_salarios_mujeres, Varianza = varianza_salarios_mujeres)

salarios_resumen <- data.frame(
  Estadistico = c("Mínimo", "Primer cuartil (Q1)", "Mediana", "Promedio", "Tercer cuartil (Q3)", "Máxim
  Hombres = format(as.numeric(round(resumen_salarios_hombres,2)),big.mark = " "),
  Mujeres = format(as.numeric(round(resumen_salarios_mujeres,2))
,big.mark = " "))
```

Cuadro 3: Resumen de Último Salario reportado por sexo

Estadistico	Hombres	Mujeres
Mínimo	10 880.92	10 223.99
Primer cuartil (Q1)	552 490.82	584 757.36
Mediana	1 104 956.42	1 062 461.32
Promedio	1 156 468.24	1 046 661.46
Tercer cuartil (Q3)	1 611 734.74	1 403 025.79
Máximo	8 154 905.08	7 290 150.00
Varianza	500 311 658 906.86	287 419 606 660.60

En el Cuadro 3, se observa que el salario más bajo reportado le pertenece a una mujer, siendo tan solo 656.93 colones menos que el salario más bajo de los hombres.

En cuanto al Q1, se destaca que el 25 % de los hombres tuvieron un último salario reportado de 552,491 colones o menos, mientras que el 25 % de las mujeres tuvieron un último salario de 584,757 colones o menos. Es decir, en los últimos salarios más bajos, las mujeres experimentaron ingresos superiores a los de los hombres.

Ahora en relación a la mediana, la mitad de los hombres ganó más de 1 104 956 colones, mientras que la mitad de las mujeres ganó menos (1 046 661 colones).

Además, el promedio de los últimos salarios fue mayor para los hombres, aproximadamente 109 807 colones más alto.

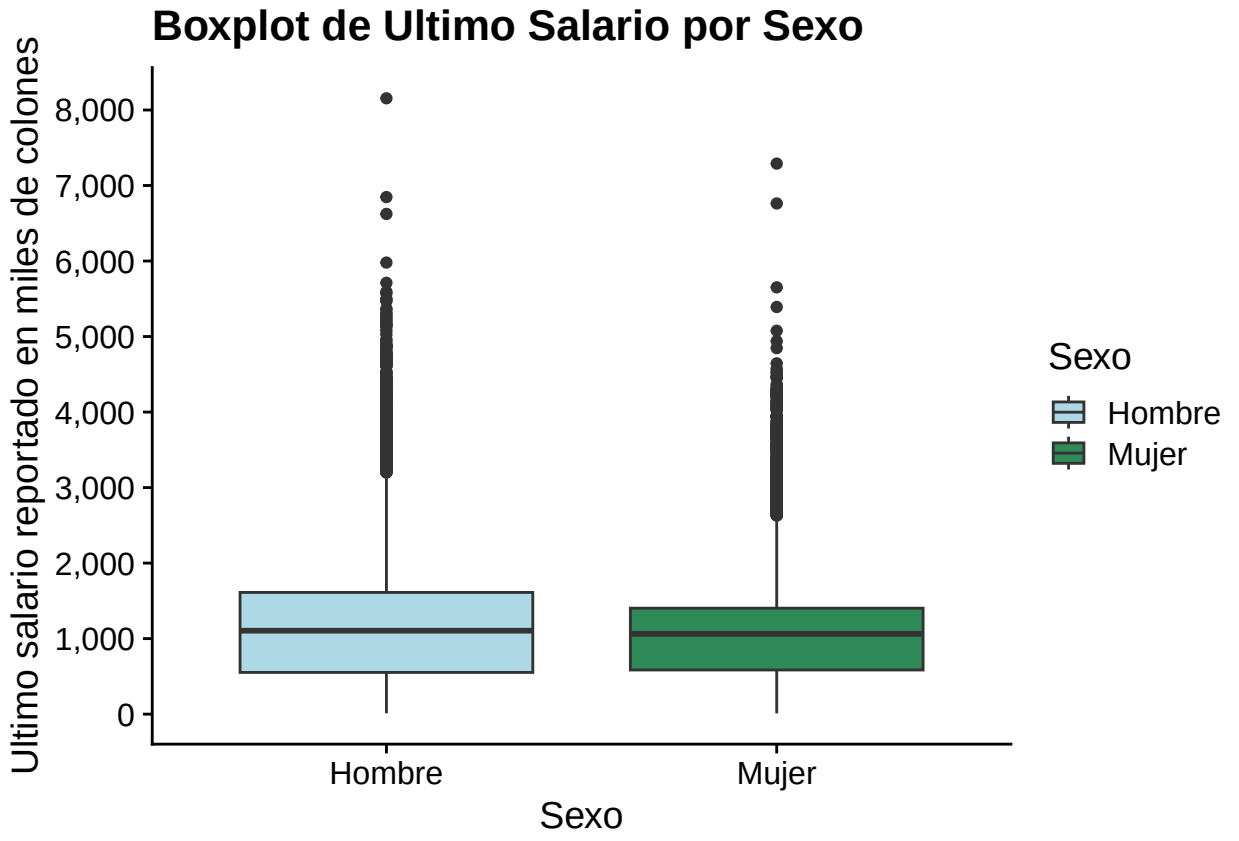
Las diferencias se incrementan a medida que se consideran los últimos salarios más altos. El 75 % de las mujeres ganaron 1 403 025.79 o menos, mientras que el 75 % de los hombres ganaron 1 611 735 colones o menos.

El salario máximo reportado por hombres es 864 755 colones más alto que el de las mujeres.

Es importante resaltar que la varianza para ambos sexos señala que los datos están muy dispersos de la media, es decir hay una gran variabilidad en los datos de último salario reportado.

2.2. Gráfico plotbox para el salario para comparar entre las categorías de Sexo

El siguiente gráfico plotbox permite tener una representación visual de como se distribuyen los datos para cada sexo.



Este gráfico destaca la presencia de salarios atípicos, identificados como aquellos que están fuera de cada una de las cajas. Además, en este gráfico, es de utilidad para visualizar el análisis presentado en la sección anterior.

2.3. Conclusiones con respecto a los salarios y sexo

Después de realizar el análisis descriptivo de los datos y para complementar el gráfico anterior, se concluye que al dividir la población por sexo, el 25 % de los hombres con los salarios más bajos ganan menos que el 25 % de las mujeres con salarios más bajos. Sin embargo, cuando se considera el 50 % de los hombres que ganan más, estos perciben un salario mayor que el 50 % de las mujeres que ganan más. Esta tendencia se evidencia claramente en el gráfico anterior, donde la parte superior de la caja, que está por encima de la línea de la mediana, es más amplia en el caso de los hombres.

Para respaldar este argumento, se observa que el promedio de los últimos salarios fue mayor para los hombres, aproximadamente 109 807 colones más alto que el de las mujeres.

No obstante, es importante señalar que la varianza de los salarios es mayor para los hombres, y los valores atípicos en los salarios masculinos también son más altos que los de las mujeres.

En este sentido, los salarios de los hombres parecen ser relativamente más altos que los de las mujeres, aunque las diferencias no resultan significativas. Ambos tienen personas con salarios más altos que la mayoría.

2.4. Prueba de hipótesis sobre las medias de las categorías de sexo

En relación a las medias, se cree que las medias de los salarios para hombres y mujeres son diferentes. Para verificar lo expresado anteriormente, se llevó a cabo la siguiente prueba de hipótesis.

En este caso, se llevó a cabo un **Welch Two Sample t-test**, la cual es una prueba estadística que permite comparar las medias de dos muestras. En este caso, el tamaño de las muestras y sus varianzas difieren.

Para realizar una prueba de hipótesis se necesitan tres cosas: la hipótesis nula (H0), el estadístico de prueba (t) y la distribución del estadístico de prueba.

En este caso particular:

1. H0: La diferencia entre la media de los últimos salarios de los hombres y la media de los últimos salarios de las mujeres es 0.
2. El estadístico de prueba es t, el cual se puede calcular de la siguiente manera:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

   donde $\bar{X}_1$ y $\bar{X}_2$ son las medias de cada una de las muestras, s_1 y s_2 son las desviaciones estándar de los dos muestras, n_1 y n_2 son el tamaño de cada una de las muestras.
3. Por último, el estadístico t sigue una distribución t con v grados de libertad, la cual se calcula utilizando la equación Welch–Satterthwaite.

A continuación se presenta el código de la prueba y la interpretación de los resultados:

```
t.test(
  x           = base_salarios$Ultimo.Salario[base_salarios$Sexo == 1],
  y           = base_salarios$Ultimo.Salario[base_salarios$Sexo == 2],
  paired      = FALSE,
  alternative = "two.sided",
  conf.level  = 0.95
)
```

```
##
##  Welch Two Sample t-test
##
## data:  base_salarios$Ultimo.Salario[base_salarios$Sexo == 1] and base_salarios$Ultimo.Salario[base_sa
## t = 25.052, df = 50185, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  101215.7 118397.9
## sample estimates:
## mean of x mean of y
##   1156468   1046661
```

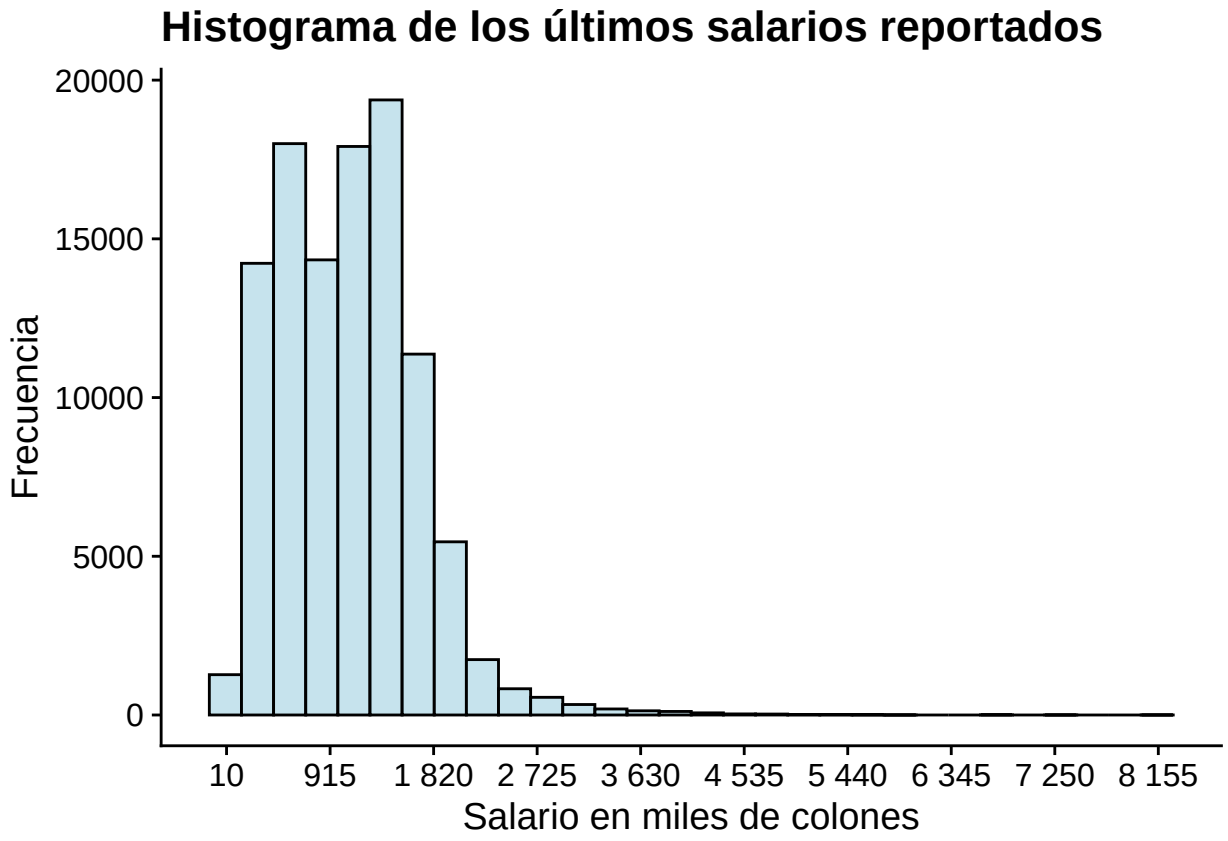
El p-valor lo que dice es si se asume la H0 como verdadera, la probabilidad de que H0 sea verdadera. En este caso, el valor p fue menor a 0,00000000000000022 lo cual es muy cercano a 0, es decir, hay una probabilidad muy pequeña de que la media de los últimos salarios entre hombres y mujeres sea 0, es decir, que sea la misma. Por tanto, existe suficiente evidencia para rechazar la hipotesis nula.

3. Parte II

3.1. Construcción del Histograma de los salarios

A coninaución se presenta el histograma de los salarios de la base de datos. Se decidió dividir los salarios en 30 grupos (30 rectángulos). El histograma evidencia una cantidad pequeña de salarios bajos. Luego, se puede observar que la gran mayoría de las personas están entre el rectángulo dos y el rectángulo seis. En particular, el rectángulo seis es el que más registros contiene y señala que se trata de salarios menor a 1 820 000 colones.

```
ggplot(base_salarios, aes(x = Ultimo.Salario)) +
  geom_histogram(binwidth = (max(base_salarios$Ultimo.Salario) - min(base_salarios$Ultimo.Salario)) / 2
                 fill = "lightblue", color = "black", alpha = 0.7) +
  labs(title = "Histograma de los últimos salarios reportados",
       x = "Salario en miles de colones",
       y = "Frecuencia") +
  theme_cowplot() +
  scale_x_continuous(
    breaks = seq(min(base_salarios$Ultimo.Salario), max(base_salarios$Ultimo.Salario), length.out = 10)
    labels = scales::label_number(big.mark = " ", decimal.mark = ".", scale = 1e-3)
  )
```

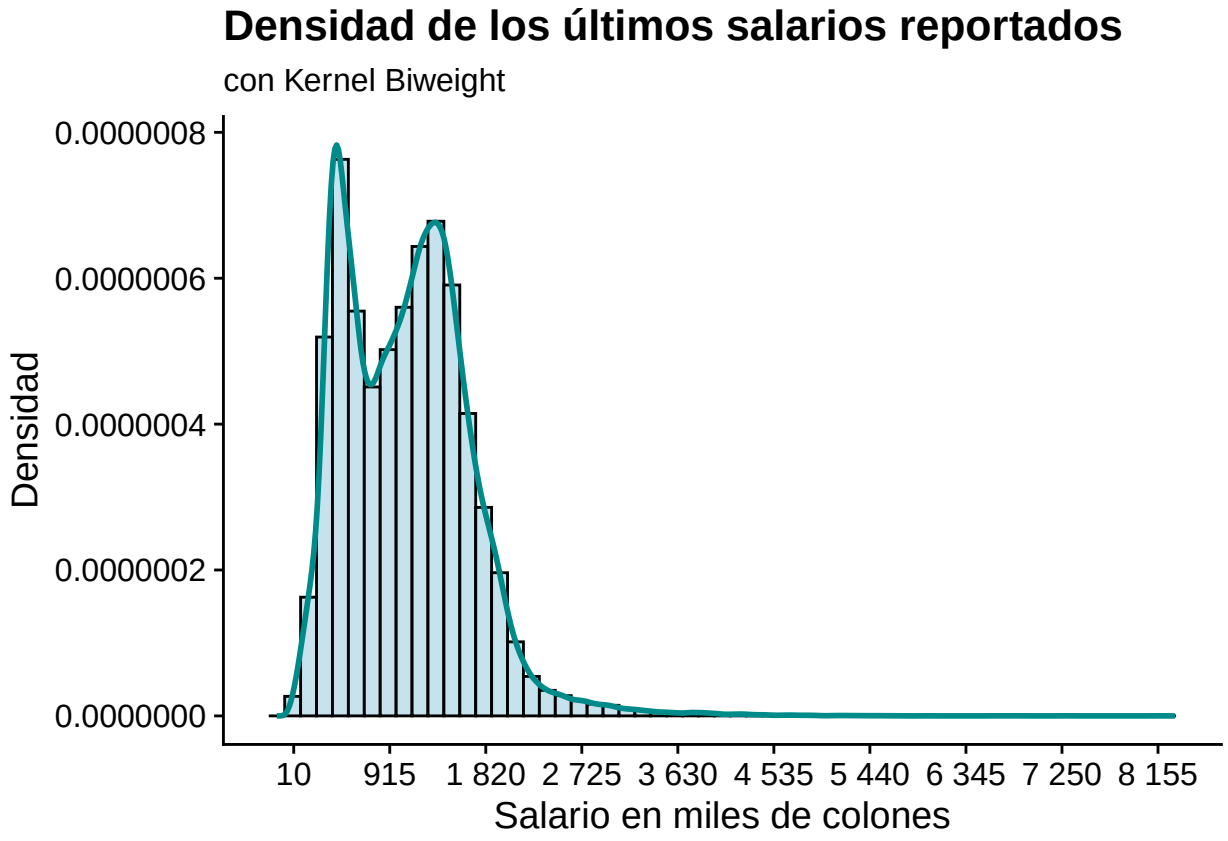


3.2. Densidad de los salarios por kernel (no paramétrica)

Una forma de aproximar la densidad de una muestra es utilizar kernels. Existen diferentes tipos de kernels: kernel normal, kernel epanechnikov, kernel triangular, kernel coseno, kernel rectangular. A continuación se presentan los gráficos de la densidad de los salarios utilizando cada uno de los tipos de kernels utilizando la función Density en el programa R.

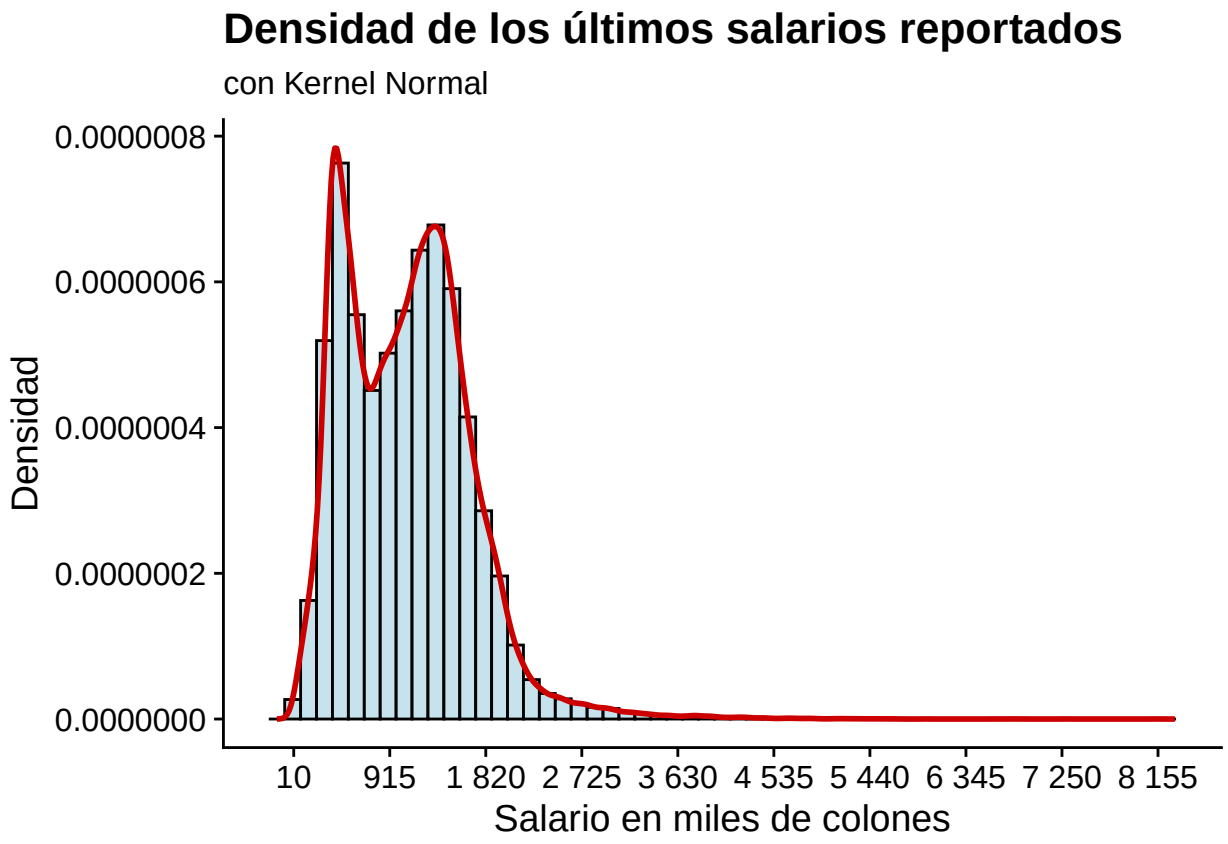
3.2.1. Densidad de los salarios por kernel usando como kernel Biweigth

```
density_biweight <- density(base_salarios$Ultimo.Salario, kernel = "biweight")

ggplot(base_salarios, aes(x = Ultimo.Salario)) +
  geom_histogram(binwidth = 150000, fill = "lightblue", color = "black", alpha = 0.7, aes(y = ..density
  geom_line(data = data.frame(x = density_biweight$x, y = density_biweight$y),
            aes(x, y), color = "cyan4", size = 1) +
  labs(title = "Densidad de los últimos salarios reportados ", subtitle = "con Kernel Biweight", x = "Sa
  theme_cowplot() +
  scale_x_continuous(
    breaks = seq(min(base_salarios$Ultimo.Salario), max(base_salarios$Ultimo.Salario), length.out = 10)
    labels = scales::label_number(big.mark = " ", decimal.mark = ".", scale = 1e-3)
  )
```



3.2.2. Densidad de los salarios por kernel usando como kernel Normal

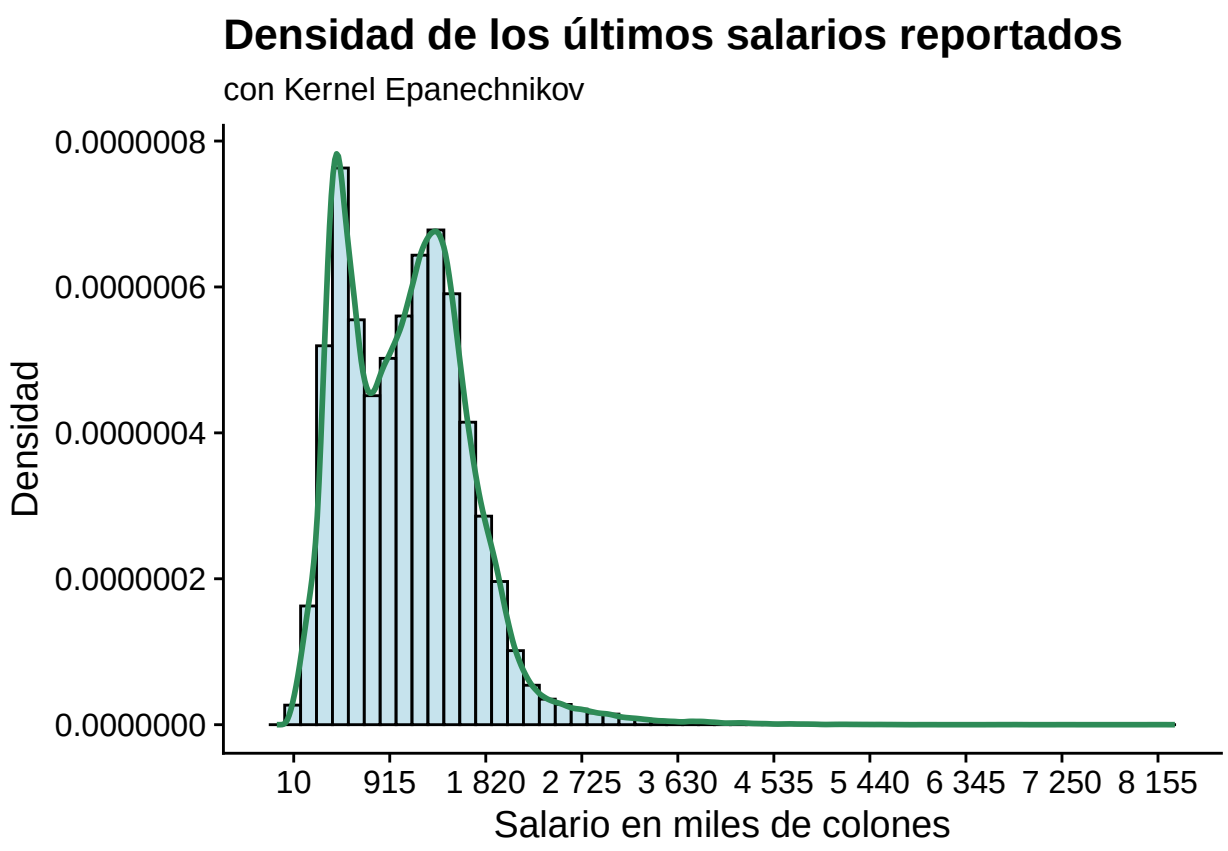
```
density_normal <- density(base_salarios$Ultimo.Salario,kernel = "gaussian")

ggplot(base_salarios, aes(x = Ultimo.Salario)) +
  geom_histogram(binwidth = 150000, fill = "lightblue", color = "black", alpha = 0.7, aes(y = ..density
  geom_line(data = data.frame(x = density_normal$x, y = density_normal$y),
            aes(x, y), color = "red3", size = 1) +
  labs(title = "Densidad de los últimos salarios reportados ", subtitle = "con Kernel Normal", x = "Sala
  theme_cowplot() +
  scale_x_continuous(
    breaks = seq(min(base_salarios$Ultimo.Salario), max(base_salarios$Ultimo.Salario), length.out = 10)
    labels = scales::label_number(big.mark = " ", decimal.mark = ".", scale = 1e-3)
  )
```



3.2.3. Densidad de los salarios por kernel usando como kernel Epanechnikov

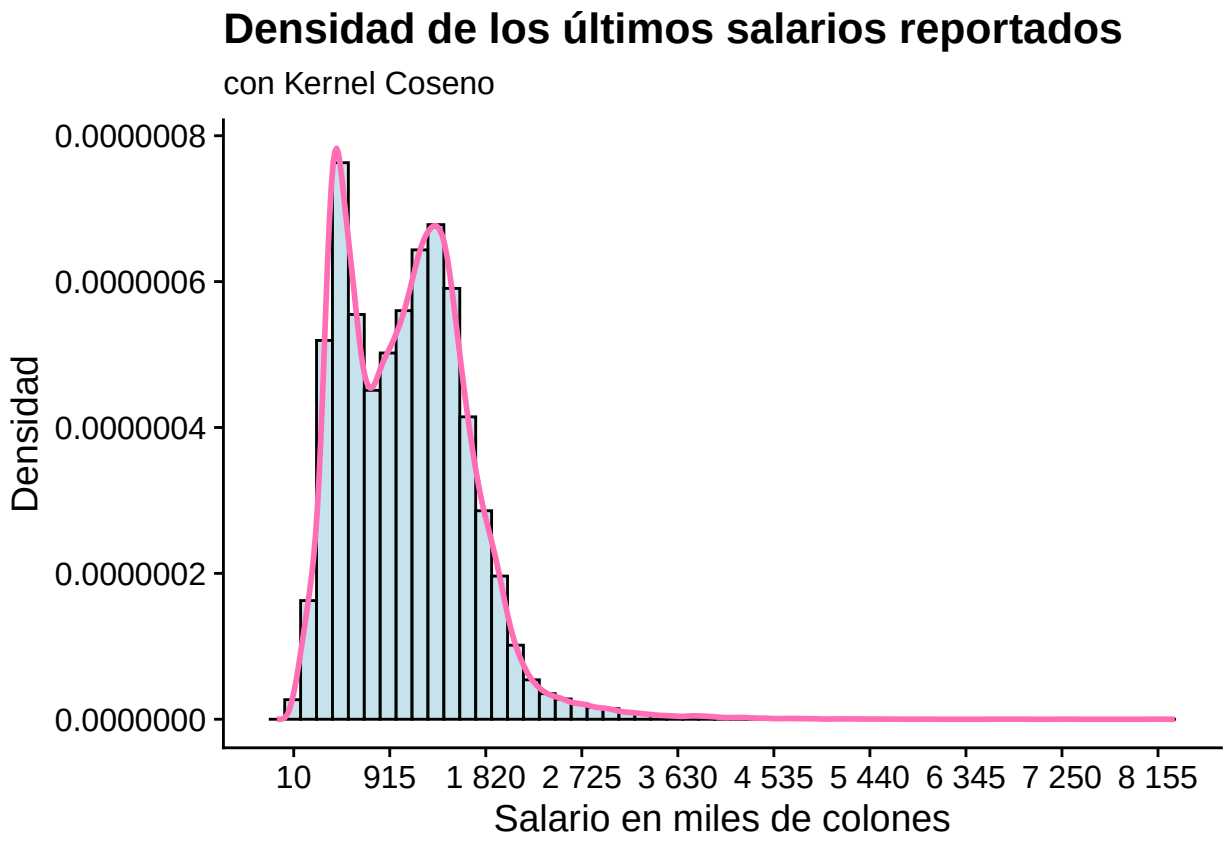
```
density_epanechnikov <- density(base_salarios$Ultimo.Salario,kernel = "epanechnikov")

ggplot(base_salarios, aes(x = Ultimo.Salario)) +
  geom_histogram(binwidth = 150000, fill = "lightblue", color = "black", alpha = 0.7, aes(y = ..density
  geom_line(data = data.frame(x = density_epanechnikov$x, y = density_epanechnikov$y),
            aes(x, y), color = "seagreen", size = 1) +
```

```
  labs(title = "Densidad de los últimos salarios reportados ",subtitle = "con Kernel Epanechnikov", x =
  theme_cowplot() +
  scale_x_continuous(
    breaks = seq(min(base_salarios$Ultimo.Salario), max(base_salarios$Ultimo.Salario), length.out = 10)
    labels = scales::label_number(big.mark = " ", decimal.mark = ".", scale = 1e-3)
  )
```



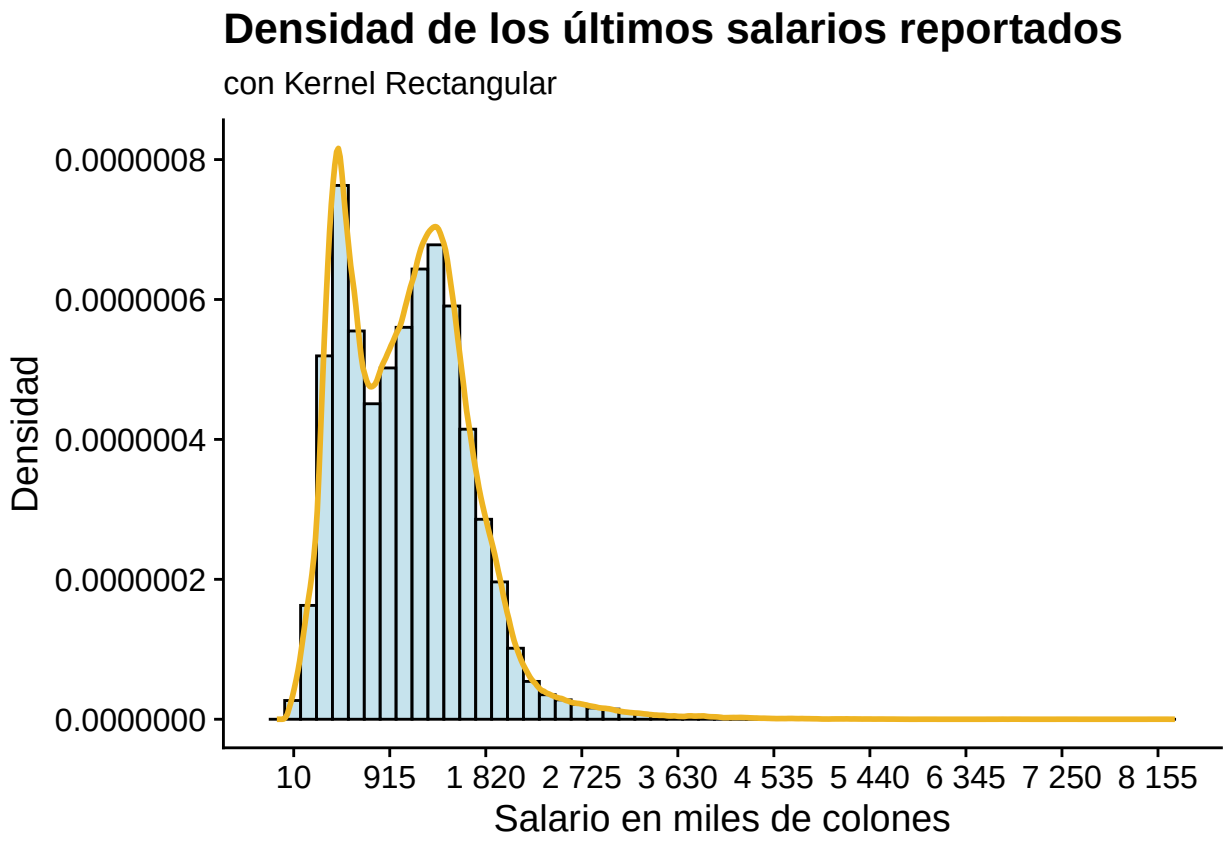
3.2.4. Densidad de los salarios por kernel usando como kernel Coseno

```
density_coseno <- density(base_salarios$Ultimo.Salario,kernel = "cosine")

ggplot(base_salarios, aes(x = Ultimo.Salario)) +
  geom_histogram(binwidth = 150000, fill = "lightblue", color = "black", alpha = 0.7, aes(y = ..density
  geom_line(data = data.frame(x = density_coseno$x, y = density_coseno$y),
            aes(x, y), color = "hotpink1", size = 1) +
  labs(title = "Densidad de los últimos salarios reportados ",subtitle = "con Kernel Coseno", x = "Sala
  theme_cowplot() +
  scale_x_continuous(
    breaks = seq(min(base_salarios$Ultimo.Salario), max(base_salarios$Ultimo.Salario), length.out = 10)
    labels = scales::label_number(big.mark = " ", decimal.mark = ".", scale = 1e-3)
  )
```



3.2.5. Densidad de los salarios por kernel usando como kernel Rectangular

```
density_rectangular <- density(base_salarios$Ultimo.Salario,kernel = "rectangular")

ggplot(base_salarios, aes(x = Ultimo.Salario)) +
  geom_histogram(binwidth = 150000, fill = "lightblue", color = "black", alpha = 0.7, aes(y = ..density
  geom_line(data = data.frame(x = density_rectangular$x, y = density_rectangular$y),
            aes(x, y), color = "goldenrod2", size = 1) +
  labs(title = "Densidad de los últimos salarios reportados ",subtitle = "con Kernel Rectangular", x =
  theme_cowplot() +
  scale_x_continuous(
    breaks = seq(min(base_salarios$Ultimo.Salario), max(base_salarios$Ultimo.Salario), length.out = 10)
    labels = scales::label_number(big.mark = " ", decimal.mark = ".", scale = 1e-3)
  )
```



3.2.6. Densidad de los salarios con todos los kernel

```
ggplot(base_salarios, aes(x = Ultimo.Salario)) +
  geom_histogram(binwidth = 150000, fill = "lightblue", color = "black", alpha = 0.7, aes(y = ..density
  geom_line(data = data.frame(x = density_normal$x, y = density_normal$y),
            aes(x, y, color = "Normal"), size = 2) +
  geom_line(data = data.frame(x = density_biweight$x, y = density_biweight$y),
            aes(x, y, color = "Biweight"), size = 1.7) +
  geom_line(data = data.frame(x = density_coseno$x, y = density_coseno$y),
            aes(x, y, color = "Coseno"), size = 1.4) +
  geom_line(data = data.frame(x = density_rectangular$x, y = density_rectangular$y),
            aes(x, y, color = "Rectangular"), size = 1.1) +
  geom_line(data = data.frame(x = density_epanechnikov$x, y = density_epanechnikov$y),
            aes(x, y, color = "Epanechnikov"), size = 0.8) +
  labs(title = "Densidad de los últimos salarios reportados", x = "Salario en miles de colones", y = "D
  theme_cowplot() +
  scale_x_continuous(
    breaks = seq(min(base_salarios$Ultimo.Salario), max(base_salarios$Ultimo.Salario), length.out = 10)
    labels = scales::label_number(big.mark = " ", decimal.mark = ".", scale = 1e-3)
  ) +
  scale_color_manual(values = c("red3", "cyan4", "hotpink1", "goldenrod2", "seagreen"),
                     name = "Kernel",
                     labels = c("Normal", "Biweight", "Coseno", "Rectangular", "Epanechnikov")) +
  theme(legend.text = element_text(size = 6),
```

```
        legend.title = element_text(size = 9))
```

En el siguiente gráfico, se colocó todos los kernel en un sólo gráfico. En este gráfico lo que se puede observar es que la gráfica de kernel epanechnikov es la que más se distancia del resto de las gráficas.