Trabajo Práctico Nº1

2024-07-11

Integrantes: Jeremías Laria Guaza (1629/21), Jerónimo Barragán (1472/21), Manuel Horn (321/21).

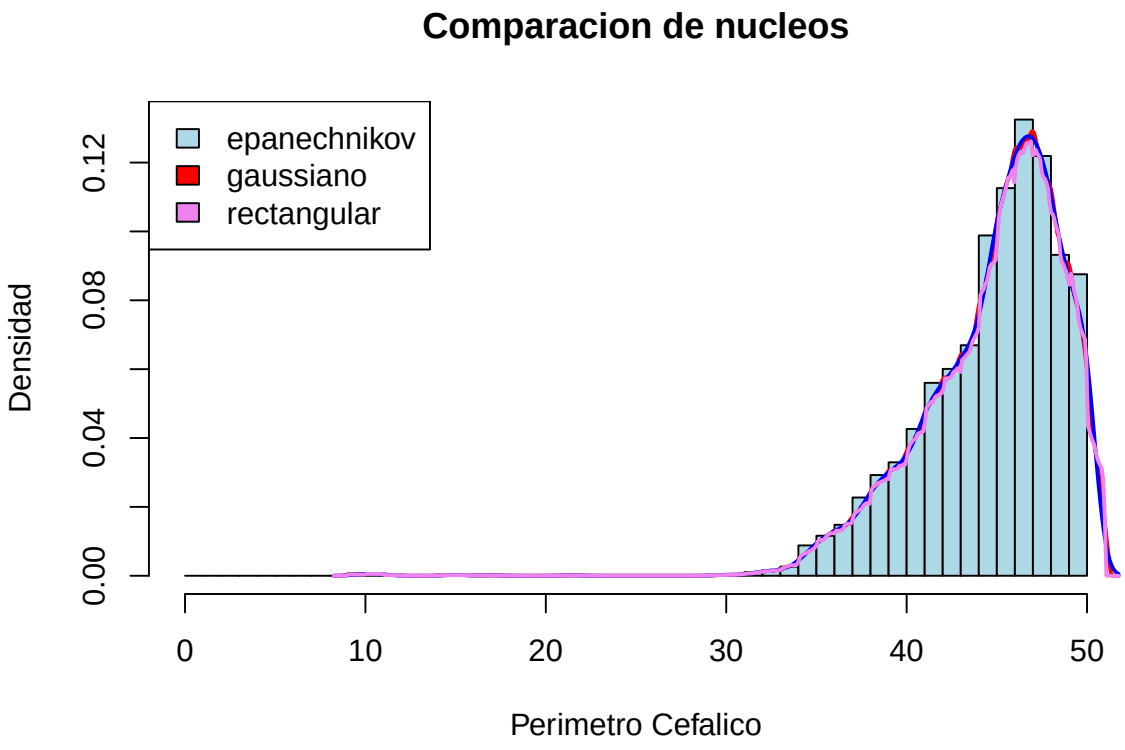
```
data <- read.table("ENNyS_menorA2.txt", header=TRUE)

attach(data)

as.factor(Sexo)
as.factor(Tipo_embarazo)
```

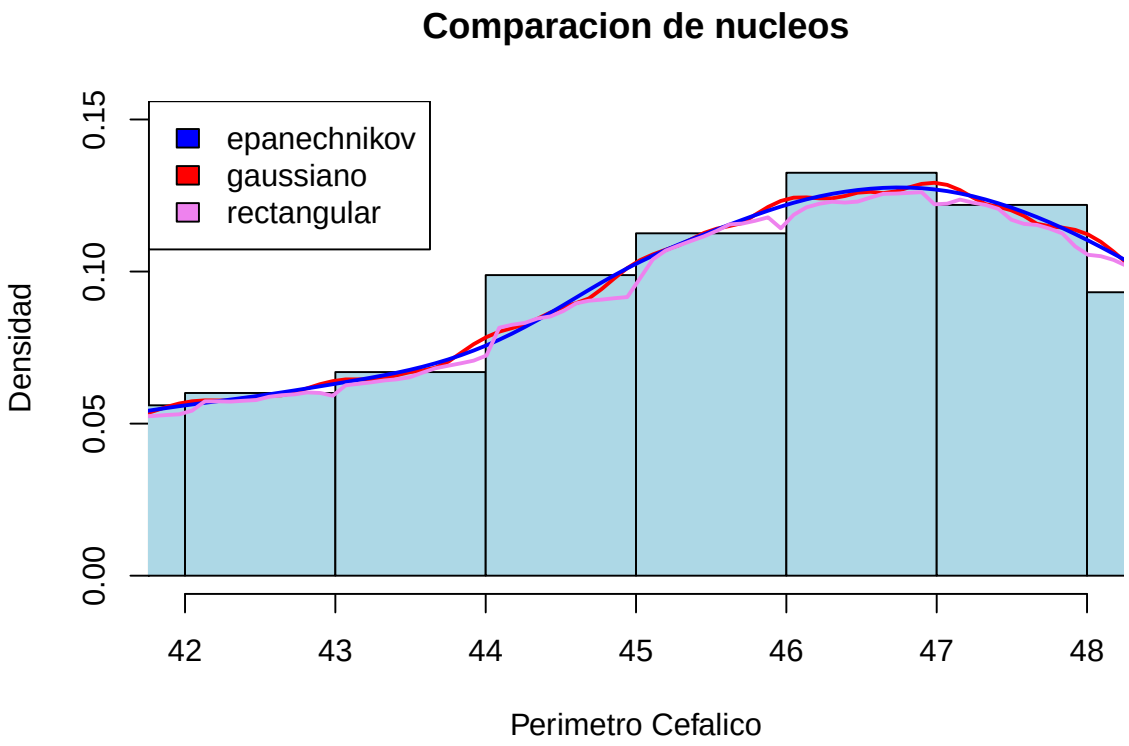
Ejercicio 1

```
hist(Perim_cef, breaks = seq(from = 0, to = max(Perim_cef), by = 1), freq = FALSE,
    col = "lightblue", main = "Comparacion de nucleos", xlab = "Perimetro Cefalico",
    ylab = "Densidad")
lines(density(Perim_cef, kernel = "epanechnikov"), col = "red", lwd = 2)
lines(density(Perim_cef, kernel = "gaussian"), col = "blue", lwd = 2)
lines(density(Perim_cef, kernel = "rectangular"), col = "violet", lwd = 2)
legend("topleft", legend = c("epanechnikov", "gaussiano", "rectangular"), fill = c("lightblue",
    "red", "violet"), bg = "white")
```



En conclusión, podemos ver que las densidades estimadas por los distintos núcleos nos dan como resultado valores muy similares entre si. Podemos analizarlo en una región más acotada para ver si hay diferencias:

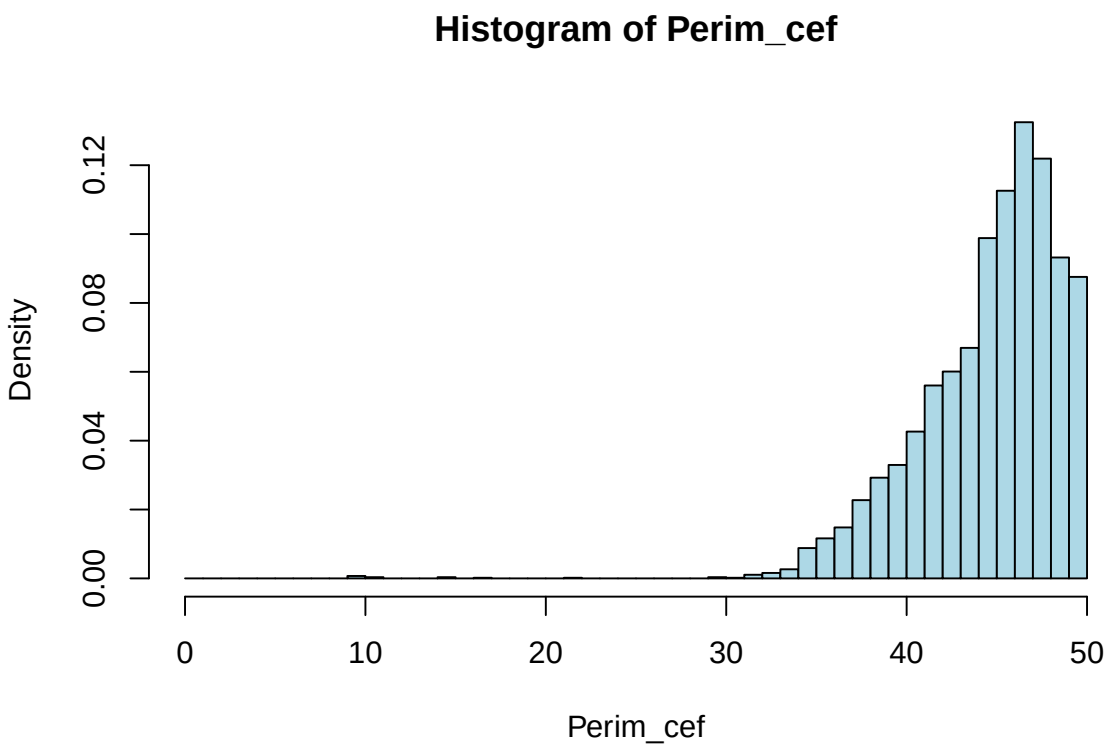
```
hist(Perim_cef, breaks = seq(from = 0, to = max(Perim_cef), by = 1), xlim = c(42,
    48), ylim = c(0, 0.15), freq = FALSE, col = "lightblue", main = "Comparacion de nucleos",
    xlab = "Perimetro Cefalico", ylab = "Densidad")
lines(density(Perim_cef, kernel = "epanechnikov"), col = "red", lwd = 2)
lines(density(Perim_cef, kernel = "gaussian"), col = "blue", lwd = 2)
lines(density(Perim_cef, kernel = "rectangular"), col = "violet", lwd = 2)
legend("topleft", legend = c("epanechnikov", "gaussiano", "rectangular"), fill = c("blue",
    "red", "violet"), bg = "white")
```



Acá podemos ver cómo tanto el núcleo epanechnikov como el núcleo gaussiano nos dan una gráfica mas "suave" que el núcleo rectangular.

Ejercicio 2

```
histograma <- hist(Perim_cef, breaks = seq(from = 0, to = max(Perim_cef), by = 1),
    freq = FALSE, col = "lightblue")
```




```
densidad_bines_filtradas <- histograma$density[42:48]
probabilidad_histograma <- sum(densidad_bines_filtradas)

densidad_epanechnikov <- density(Perim_cef, kernel = "epanechnikov")
probabilidad_densidad <- integrate(function(x) {
    approx(densidad_epanechnikov$x, densidad_epanechnikov$y, xout = x)$y
}, lower = 42, upper = 48)

cat("Probabilidad calculada con el histograma: ", probabilidad_histograma, "\n")
```

```
## Probabilidad calculada con el histograma:  0.6488724
```

```
cat("Probabilidad calculada con la densidad: ")
```

```
## Probabilidad calculada con la densidad:
```

```
probabilidad_densidad
```

```
## 0.5752209 with absolute error < 3.3e-05
```

Para calcular la probabilidad estimada usando el histograma, recordamos que es igual a #{42 <= Perim_cef <= 48} / n donde n es el total de realizaciones. Como el alto de cada bin Cj es #{Cj} / (n . |Cj|) y |Cj| = 1 para todo j en nuestro gráfico, tenemos directamente que la probabilidad es la suma de las alturas de los bins entre 42 y 48.

En cuanto a la densidad, para calcular la probabilidad, integramos la función de densidad estimada con el núcleo de epanechnikov entre 42 y 48.

Ejercicio 3

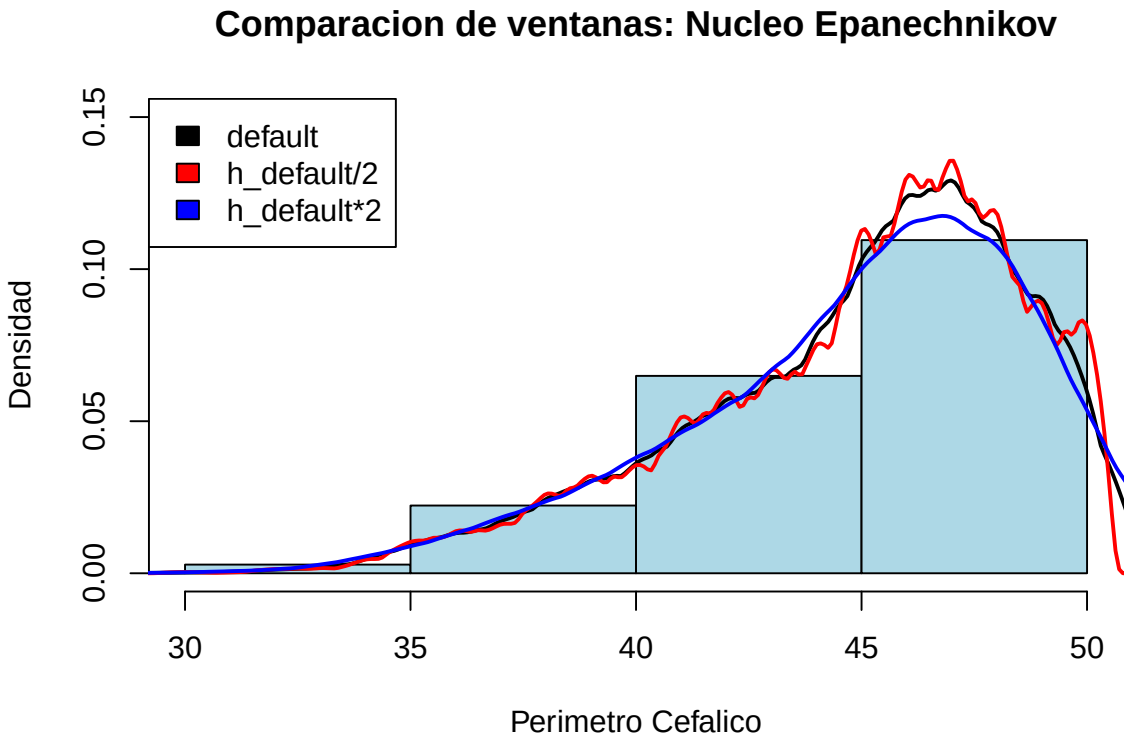
El valor de la ventana R lo ajusta dependiendo del conjunto de datos. Para Perim_cef, el valor utilizado es:

```
h_default <- bw.nrd0(Perim_cef)
h_default
```

```
## [1] 0.5841513
```

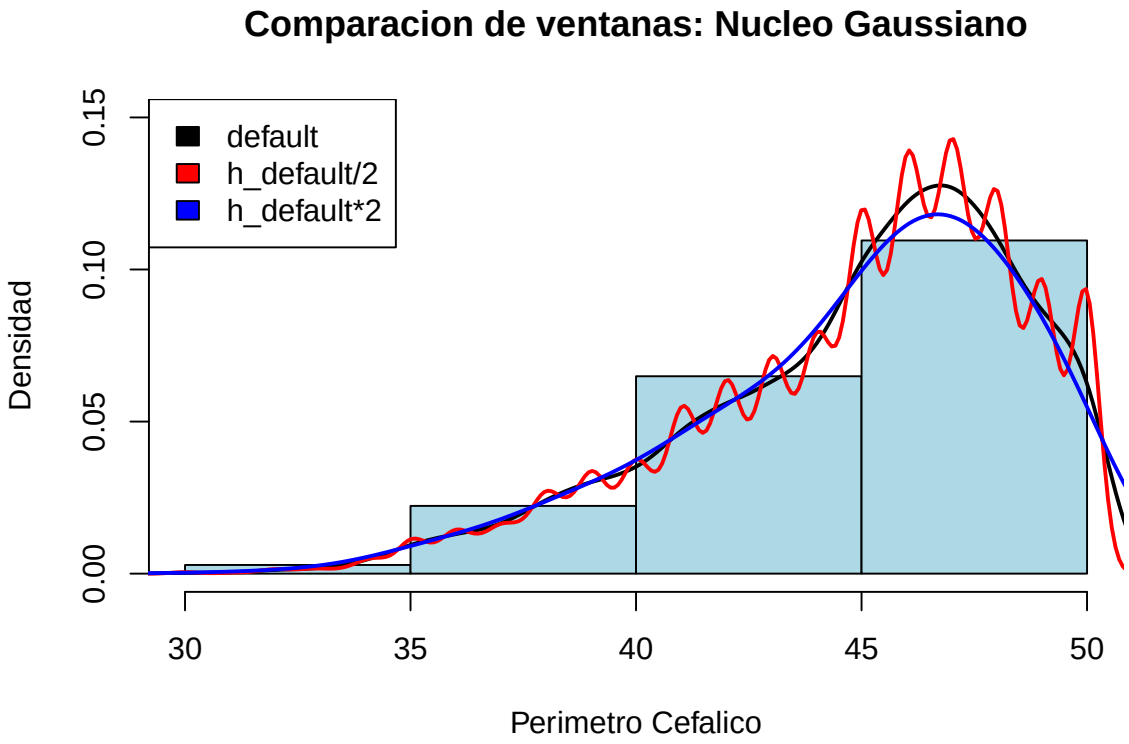
Veamos la comparación de los graficos si utilizamos la mitad de la ventana y el doble de esta.

```
hist(Perim_cef, xlim = c(30, max(Perim_cef)), ylim = c(0, 0.15), freq = FALSE, col = "lightblue",
    main = "Comparacion de ventanas: Nucleo Epanechnikov", xlab = "Perimetro Cefalico",
    ylab = "Densidad")
lines(density(Perim_cef, kernel = "epanechnikov"), col = "black", lwd = 2)
lines(density(Perim_cef, kernel = "epanechnikov", bw = h_default/2), xlim = c(30,
    50), col = "red", lwd = 2)
lines(density(Perim_cef, kernel = "epanechnikov", bw = h_default * 2), xlim = c(30,
    50), col = "blue", lwd = 2)
legend("topleft", legend = c("default", "h_default/2", "h_default*2"), fill = c("black",
    "red", "blue"), bg = "white")
```



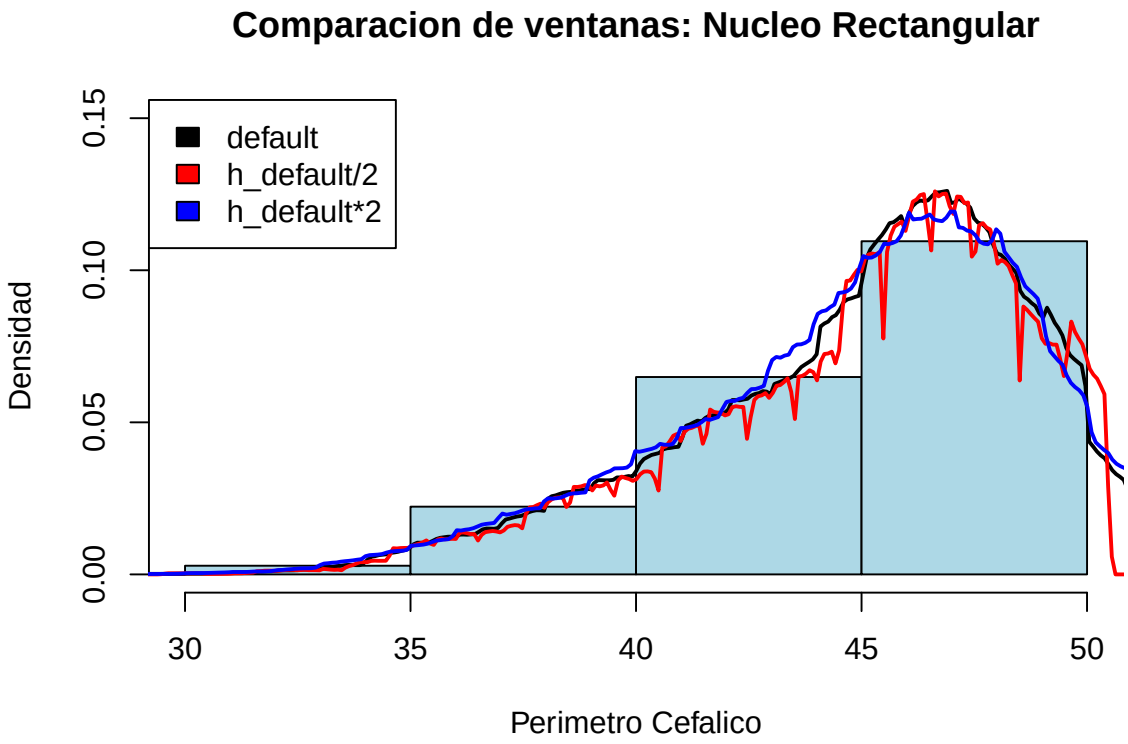
Podemos ver que utilizando la mitad del valor de la ventana que utiliza por defecto R, el gráfico va a tener más picos, pero se encuentra mas cerca de la densidad estimada con la ventana calculada por defecto. Por otro lado, la densidad estimada con el doble del valor calculado por defecto es mas “suave” pero en comparación a la anterior se aleja más de la densidad calculada por defecto.

```r
hist(Perim_cef, xlim = c(30, max(Perim_cef)), ylim=c(0, 0.15), freq = FALSE, col = "lightblue", main
lines(density(Perim_cef, kernel="gaussian"), col = "black", lwd = 2)
lines(density(Perim_cef, kernel="gaussian",  bw = h_default/2),xlim=c(30,50) , col = "red", lwd = 2)
lines(density(Perim_cef, kernel="gaussian",  bw = h_default*2), xlim=c(30,50), col = "blue", lwd = 2
legend("topleft",
       legend = c("default", "h_default/2", "h_default*2"),
       fill = c("black", "red", "blue"),
       bg = "white")
```



Al igual que para el núcleo de Epanechnikov, vemos que utilizando la ventana más chica la curva obtenida varía mucho ("oscila"), obteniendo muchos picos en el gráfico. Por otro lado, para un valor de ventana más grande, la curva de la densidad estimada se hace más suave, y no presenta picos.

```r
hist(Perim_cef, xlim = c(30, max(Perim_cef)), ylim=c(0, 0.15), freq = FALSE, col = "lightblue", main
lines(density(Perim_cef, kernel="rectangular"), col = "black", lwd = 2)
lines(density(Perim_cef, kernel="rectangular",  bw = h_default/2),xlim=c(30,50) , col = "red", lwd =
lines(density(Perim_cef, kernel="rectangular",  bw = h_default*2), xlim=c(30,50), col = "blue", lwd
legend("topleft",
       legend = c("default", "h_default/2", "h_default*2"),
       fill = c("black", "red", "blue"),
       bg = "white")
```



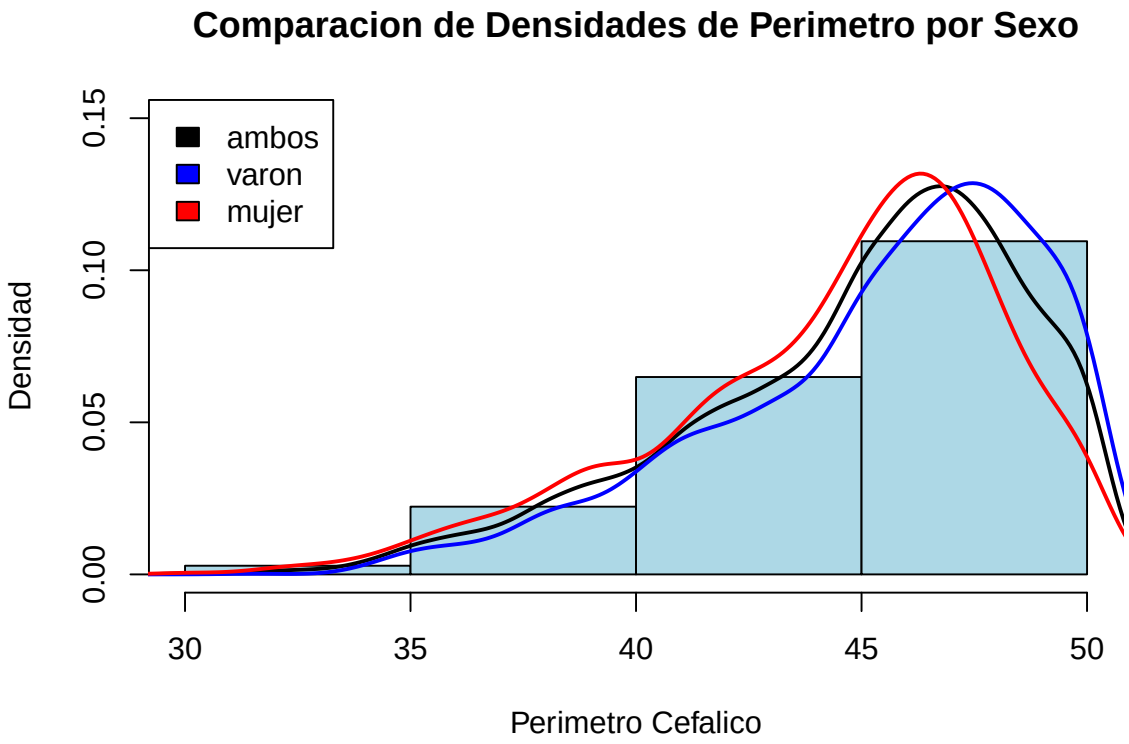
Igual que para los dos núcleos anteriores, podemos ver que para el valor de ventana más chico, el gráfico varía considerablemente ya que tiene muchos picos (en particular vemos que los picos "se caen"). Para el mayor valor de ventana, vemos que los picos se vuelven menos bruscos, o sea que el gráfico varía menos en general, pero no se llega a observar una estimación tan suave como se obtuvo con los núcleos anteriores.

Ejercicio 4

```
require(tidyverse)
```

```
Perim_cef_varon <- data %>%
    filter(Sexo == "Varon") %>%
    pull(Perim_cef)
Perim_cef_mujer <- data %>%
    filter(Sexo == "Mujer") %>%
    pull(Perim_cef)


hist(Perim_cef, xlim = c(30, max(Perim_cef)), ylim = c(0, 0.15), freq = FALSE, col = "lightblue",
    main = "Comparacion de Densidades de Perimetro por Sexo", xlab = "Perimetro Cefalico",
    ylab = "Densidad")
lines(density(Perim_cef, kernel = "gaussian"), col = "black", lwd = 2)
lines(density(Perim_cef_varon, kernel = "gaussian"), col = "blue", lwd = 2)
lines(density(Perim_cef_mujer, kernel = "gaussian"), col = "red", lwd = 2)
legend("topleft", legend = c("ambos", "varon", "mujer"), fill = c("black", "blue",
    "red"), bg = "white")
```

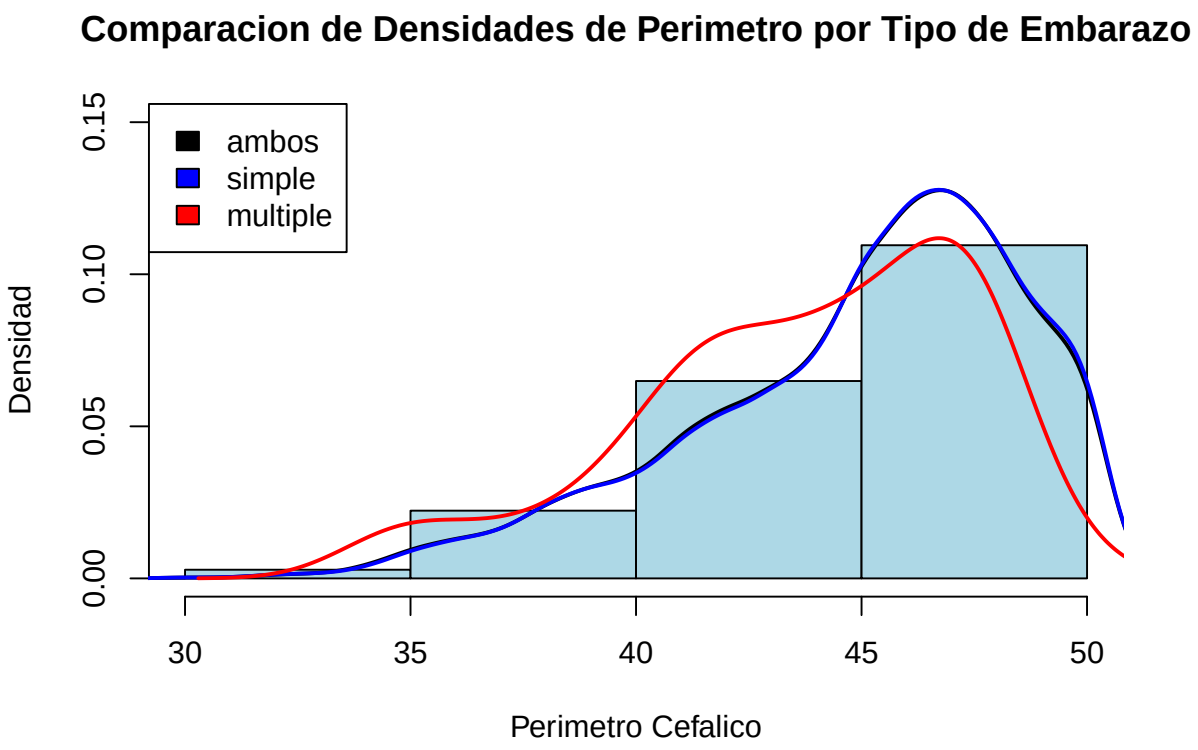


En el gráfico vemos que los varones tienden a tener un perimetro cefálico mas grande respecto a las mujeres. Esto se llega a ver dado que en perímetros cefálicos más bajos hay una mayor densidad de mujeres, y en perimetros cefálicos más grandes hay una mayor densidad de varones. La gráfica de varones está más desplazada hacia la derecha respecto a la de mujeres.

Ejercicio 5

```
Perim_cef_simple <- data %>%
    filter(Tipo_embarazo == "Simple") %>%
    pull(Perim_cef)
Perim_cef_multiple <- data %>%
    filter(Tipo_embarazo == "Multiple") %>%
    pull(Perim_cef)


hist(Perim_cef, xlim = c(30, max(Perim_cef)), ylim = c(0, 0.15), freq = FALSE, col = "lightblue",
    main = "Comparacion de Densidades de Perimetro por Tipo de Embarazo", xlab = "Perimetro Cefalico
    ylab = "Densidad")
lines(density(Perim_cef, kernel = "gaussian"), col = "black", lwd = 2)
lines(density(Perim_cef_simple, kernel = "gaussian"), col = "blue", lwd = 2)
lines(density(Perim_cef_multiple, kernel = "gaussian"), col = "red", lwd = 2)
legend("topleft", legend = c("ambos", "simple", "multiple"), fill = c("black", "blue",
    "red"), bg = "white")
```



Podemos ver que la densidad de los tipos de embarazo simple es casi idéntica a la densidad total, esto se debe a que la cantidad de tipos de embarazo múltiples es muy baja respecto a la cantidad total de datos.

```
table(Tipo_embarazo)
```

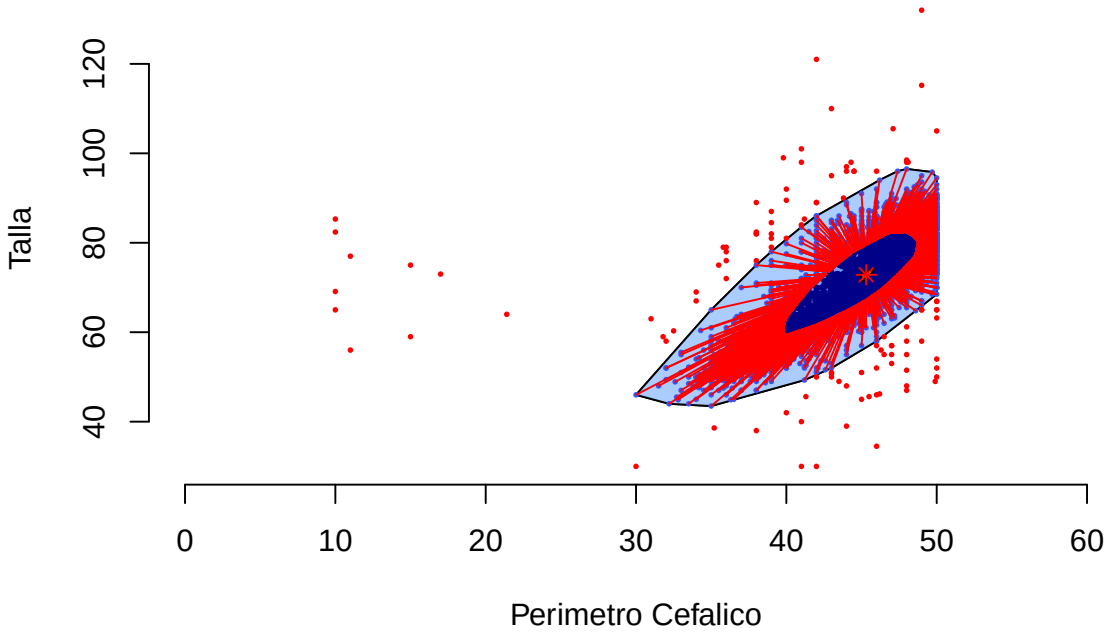
```
## Tipo_embarazo
## Multiple   Simple
##      190     5486
```

Por otro lado, el perimetro cefálico de los bebes que nacieron en un embarazo múltiple suele ser más chico respecto a los bebes que nacen de un tipo de embarazo simple. Esto se debe a que en perimetros cefálicos más bajos hay una mayor cantidad de tipos de embarazo multiple y en perimetros cefálicos mas altos hay una mayor cantidad de tipos de embarazo simples.

Ejercicio 6

```
library(aplpack)
```

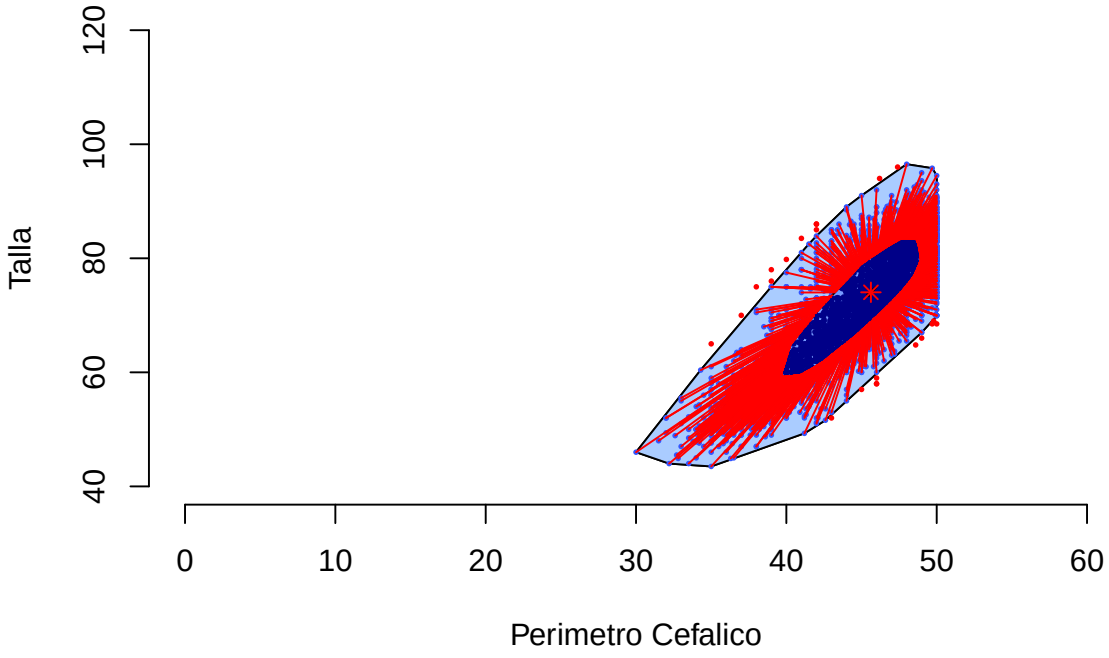
```
bag <- bagplot(Perim_cef, Talla, xlim = c(0, 60), xlab = "Perimetro Cefalico", ylab = "Talla")
```



Se registraron algunos valores atípicos, como por ejemplo casos con perímetro cercano a 10 y talla en aproximadamente 90. De la misma manera, hay casos que tienen un perímetro de aproximadamente 45 y de talla cercana a 40. Aún así, hay una tendencia clara en el boxplot: a medida que el perímetro cefálico aumenta, la talla también, lo cuál tiene sentido considerando las proporciones del cuerpo humano.

Ejercicio 7

```
out <- as.data.frame(bag$pxy.outlier)
names(out) <- c("Perim_cef", "Talla")
filtered_data <- anti_join(data, out, by = c("Perim_cef", "Talla"))
bag <- bagplot(filtered_data$Perim_cef, filtered_data$Talla, xlim = c(0, 60), ylim = c(40,
    120), xlab = "Perimetro Cefalico", ylab = "Talla")
```

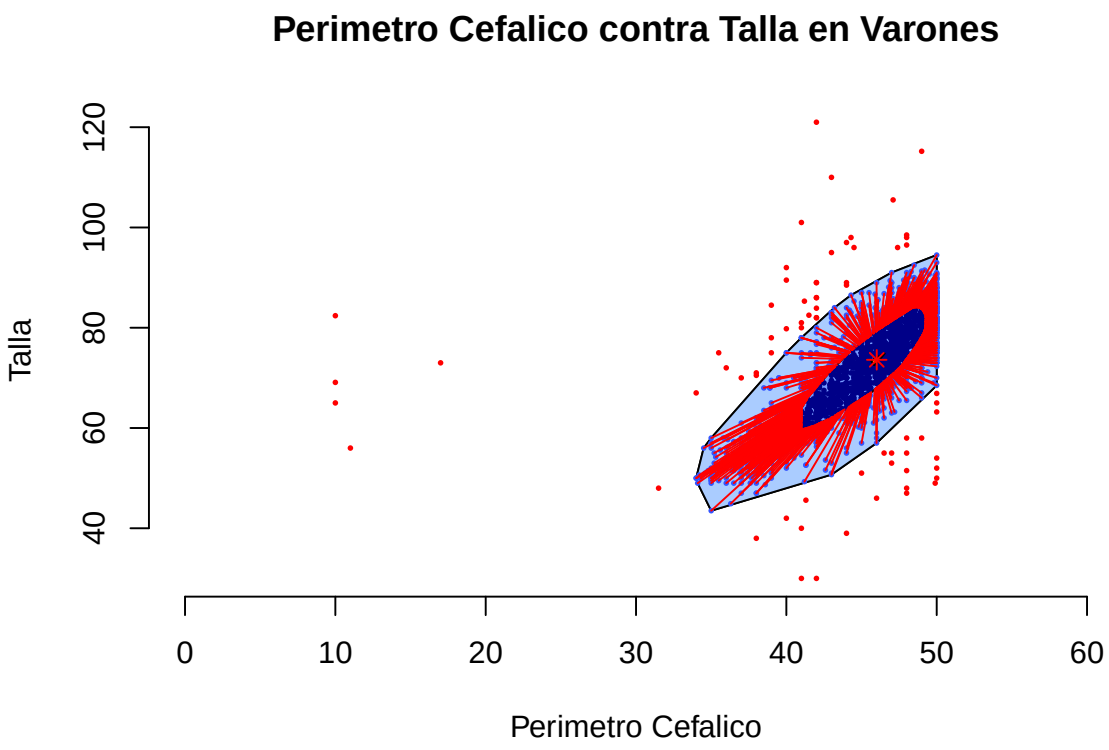


Podemos ver que al filtrar todos los datos atípicos del gráfico anterior, el nuevo bagplot sigue conteniendo datos atípicos, aunque ahora son muy pocos. En general, la estructura del nuevo bagplot no se modifica considerablemente con respecto al gráfico del punto anterior.

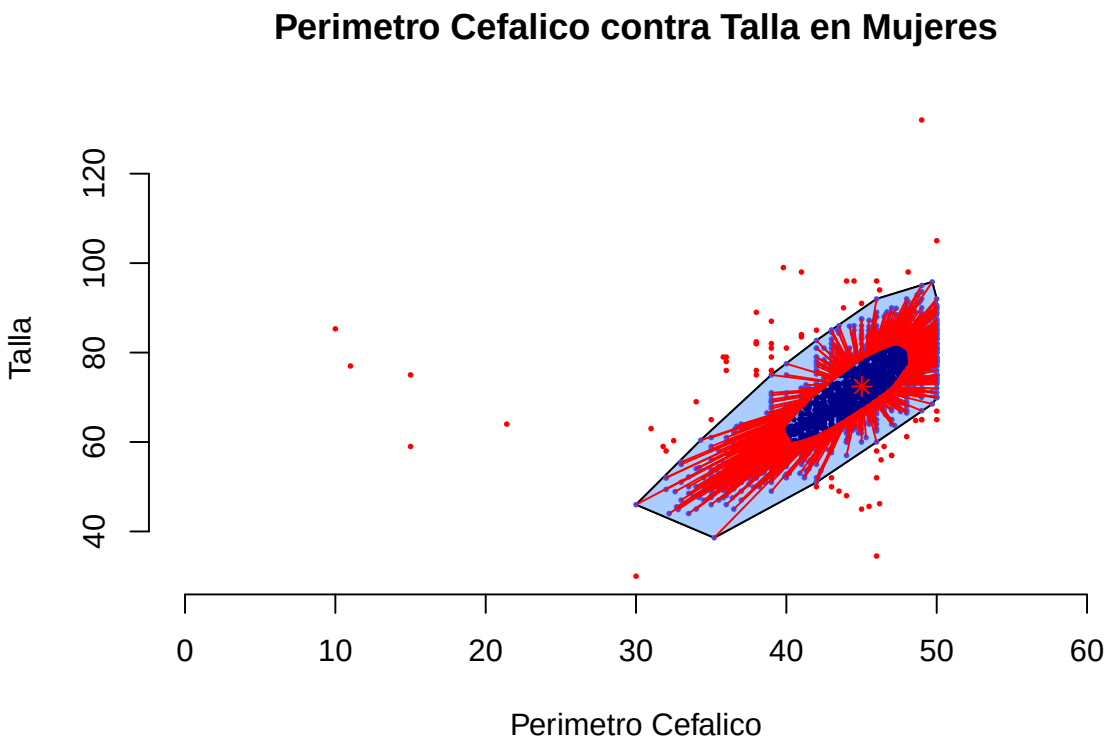
Ejercicio 8

```
Talla_varon <- data %>%
    filter(Sexo == "Varon") %>%
    pull(Talla)
Talla_mujer <- data %>%
    filter(Sexo == "Mujer") %>%
    pull(Talla)

bagplot(Perim_cef_varon, Talla_varon, xlim = c(0, 60), xlab = "Perimetro Cefalico",
    ylab = "Talla")
title(main = "Perimetro Cefalico contra Talla en Varones")
```




```
bagplot(Perim_cef_mujer, Talla_mujer, xlim = c(0, 60), xlab = "Perimetro Cefalico",
    ylab = "Talla")
title(main = "Perimetro Cefalico contra Talla en Mujeres")
```



Se puede ver como el bagplot del perímetro cefálico en varones está levemente desplazado hacia la derecha respecto al de mujeres, lo cual tiene sentido ya que anteriormente graficamos la distribución de estos y pudimos ver que los varones tienden a tener un perimetro cefálico más grande respecto a las mujeres. Los valores atípicos seguimos viéndolos al igual que en el item 6.