Maestría en Computo Estadístico
Inferencia Estadística
Tarea 2
11 de septiembre de 2020
Enrique Santibáñez Cortés
Repositorio de Git: Tarea 2, IE.

4. El siguiente conjuntos de datos contiene mediciones del diámetro de un agave, medido en decímetros, en distintas localizaciones no cercanas.

23.37	21.87	24.41	21.27	23.33	15.20	24.21	27.52	15.48	27.19
25.05	20.40	21.05	28.83	22.90	18.00	17.55	25.92	23.64	28.96
23.02	17.32	30.74	26.73	17.22	22.81	20.78	23.17	21.60	22.37

a) Escriba una función en R que calcule la función de distribución empírica para un conjunto de datos dado D. La función debe tomar como parámetros al valor x donde se evalúa y al conjunto de datos D. Utilizando esta función grafique la función de distribución empírica asociada al conjunto de datos de agave. Ponga atención a los puntos de discontinuidad. ¿Qué observa? **Nota**: Escriba la función mediante el algoritmo descrito en las notas de la clase; para este ejercicio no vale usar la funciones implementadas en R que hacen lo pedido.

RESPUESTA
Sea $X_i, \cdots, X_n \sim F$, la función de densidad empírica $\hat{F}$ es la función de distribución acumulada que asigna masa 1/n en cada punto X_i. Formalmente,

$$\hat{F}(x) = \frac{\sum_{1=i}^{n} 1_{X_i \leq x}}{n},$$

donde

$$1_{X_i \leq x} = \begin{cases} 1 & \text{si } X_i \leq x \\ 0 & \text{si } X_i > x \end{cases}$$

Utilizando la definición anterior construímos la función en R:

```
fde <- function(x, D){
  n <- length(D)
  fde_x <- sum(D<x)/n
}
```

Procedemos a calcular con la función anterior la fde en los puntos de discontinuidad, para ello primero ingresemos los datos:

```
diametro_agave<-c(23.37, 21.87, 24.41, 21.27, 23.33, 15.20, 24.21, 27.52, 15.48, 27.19,
25.05, 20.40, 21.05, 28.83, 22.90, 18.00, 17.55, 25.92, 23.64, 28.96, 23.02, 17.32,
30.74, 26.73, 17.22, 22.81, 20.78, 23.17, 21.60, 22.37)



sort(diametro_agave)
```

```
##  [1] 15.20 15.48 17.22 17.32 17.55 18.00 20.40 20.78 21.05 21.27 21.60 21.87
## [13] 22.37 22.81 22.90 23.02 23.17 23.33 23.37 23.64 24.21 24.41 25.05 25.92
## [25] 26.73 27.19 27.52 28.83 28.96 30.74
```

Los puntos de discontinuidad serían cuando cambia fde, es decir, cuando se evalua en un $x + \epsilon$ tal que $\hat{F}(x) \neq \hat{F}(x + \epsilon)$

b) Usando la desigualdad de Dvoretzky-Kiefer-Wolfowitz, escriba una función en R que calcule y grafique una región de confianza para la función de distribución empírica. La función debe tomar como parámetros al conjunto de datos que se usan para contruir la función de distribución empírica.

c) Escriba una función en R que determine la gráfica Q-Q normal de un conjunto de datos. La función debe tomar como parámetro al conjunto de datos y deberá graficar contra el percentil estandarizado de la normal. Para poder comparar el ajuste más claramente, la función además deberá ajustar en rojo a la recta $sx + \bar{x}$ s(=desviación estándar muestral y x=media muestral). Usando esta función, determine la gráfica Q-Q normal. ¿Qué observa? **Nota:** La misma del inciso a). d) Escriba una función en R que determine el gráfico de probabilidad normal. La función debe tomar como parámetro al conjunto de datos. ¿Qué observa? Nota: La misma del inciso a).

d) ¿Los datos anteriores se distribuyen normalmente? Argumente.

5. En este ejercicio repasará la estimación de densidades.

a) Escriba una función en R que estime una densidad por el método de kerneles. La función deberá recibir al punto x donde se evalúa al estimador, al parámetro de suavidad h, al kernel que se utilizará en la estimación y al conjunto de datos.

RESPUESTA

Para el caso univariado, el KDE está dado por

$$\hat{f} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_1}{h}\right), \qquad x \in \mathbb{R},\ h > 0.$$

K es la función Kernel, y h es el acho de banda que determina la suavidad de la estimación. Procedemos a definir las funciones Kernel's, solo consideraremos las siguientes: gaussian, epanechnikov, rectangular, triangular, consine.

```
# Definimos las funciones kernel:
kernel_gaussian <- function(x,x_i,h){
  dnorm((x-x_i)/h)
}

kernel_rectangular <- function(x,x_i,h){
  dunif((x_i - x)/h, min = -1, max = 1)
}

kernel_triangular <- function(x, x_i, h){
  u = (x_i-x)/h
  (1-abs(u))*(abs(u) <= 1)
}

kernel_epanechnikov <- function(x, x_i, h){
  u = (x_i - x)/h
  3/4*(1-u^2)*(abs(u) <= 1)
}

kernel_cosine <- function(x, x_i, h){
  u = (x_i - x)/h
  (1+cos(pi*u))/2*(abs(u) <= 1)
}
```

Una vez definido lo anterior procedemos a definir la función para estimar la densidad por el método de kerneles (https://github.com/JonasMoss/kdensity/blob/master/R/builtin_kernels.R):

```
kde <- function(x, h, kernel, D){
  f_hat <- 0
  n <- length(D)
  for (i in 1:n){
    if (kernel=="gaussian"){
      f_hat <- f_hat + kernel_gaussian(x, D[i], h)
    }
    else if(kernel=="epanechnikov"){
      f_hat <- f_hat + kernel_epanechnikov(x, D[i], h)
    }
    else if(kernel=="rectangular"){
      f_hat <- f_hat + kernel_rectangular(x, D[i], h)
    }
    else if(kernel=="triangular"){
      f_hat <- f_hat + kernel_triangular(x, D[i], h)
    }
    else if(kernel=="consine"){
      f_hat <- f_hat + kernel_cosine(x, D[i], h)
    }
    else{
      print("Kernel incorrecto")
      break()
    }
  }
  f_hat/(n*h)
}
```

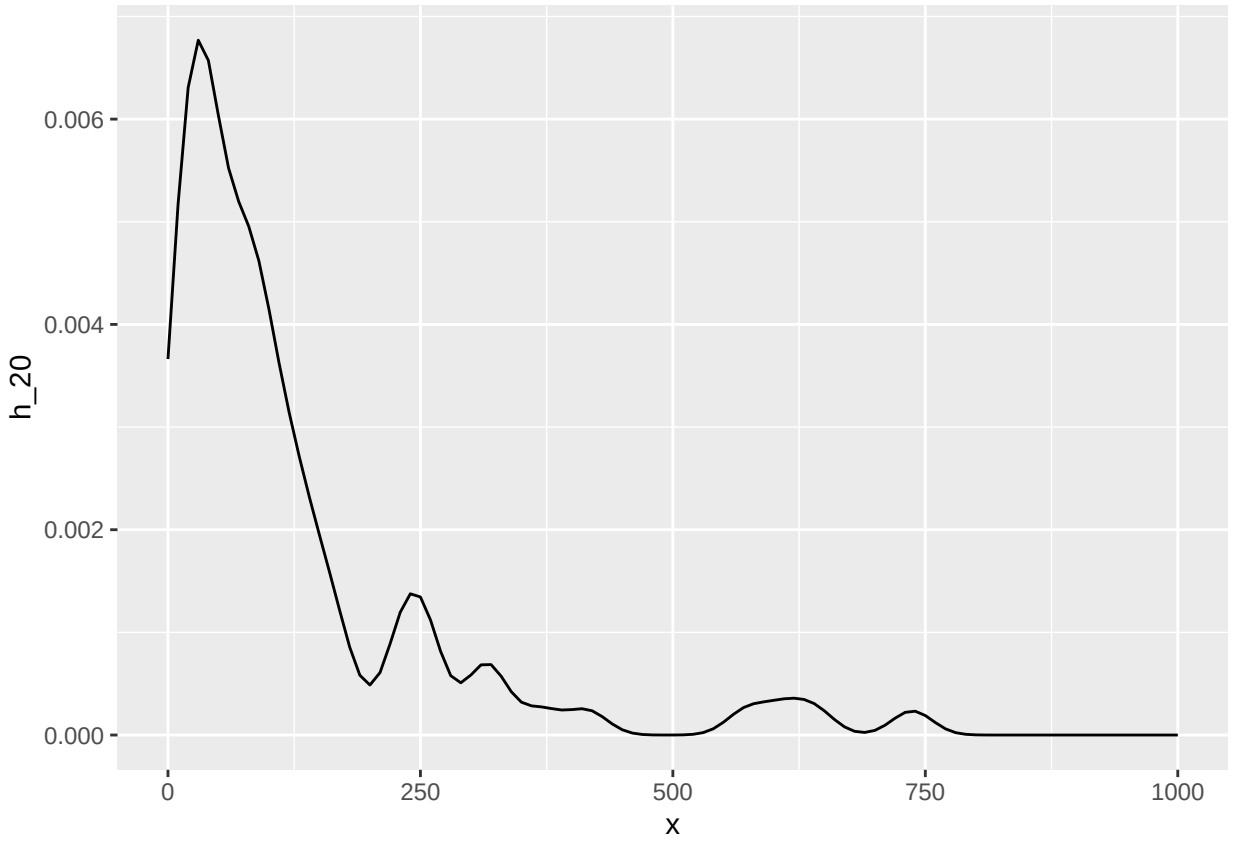
b) Cargue en R al archivo "Tratamiento.csv", el cual contiene la duración de los períodos de tratamiento (en días) de los pacientes de control en un estudio de suicidio. Utilice la función del inciso anterior para estimar la densidad del conjunto de datos para $h = 20, 30, 60$. Grafique las densidades estimadas. ¿Cuál es el mejor valor para h? Argumente.

RESPUESTA

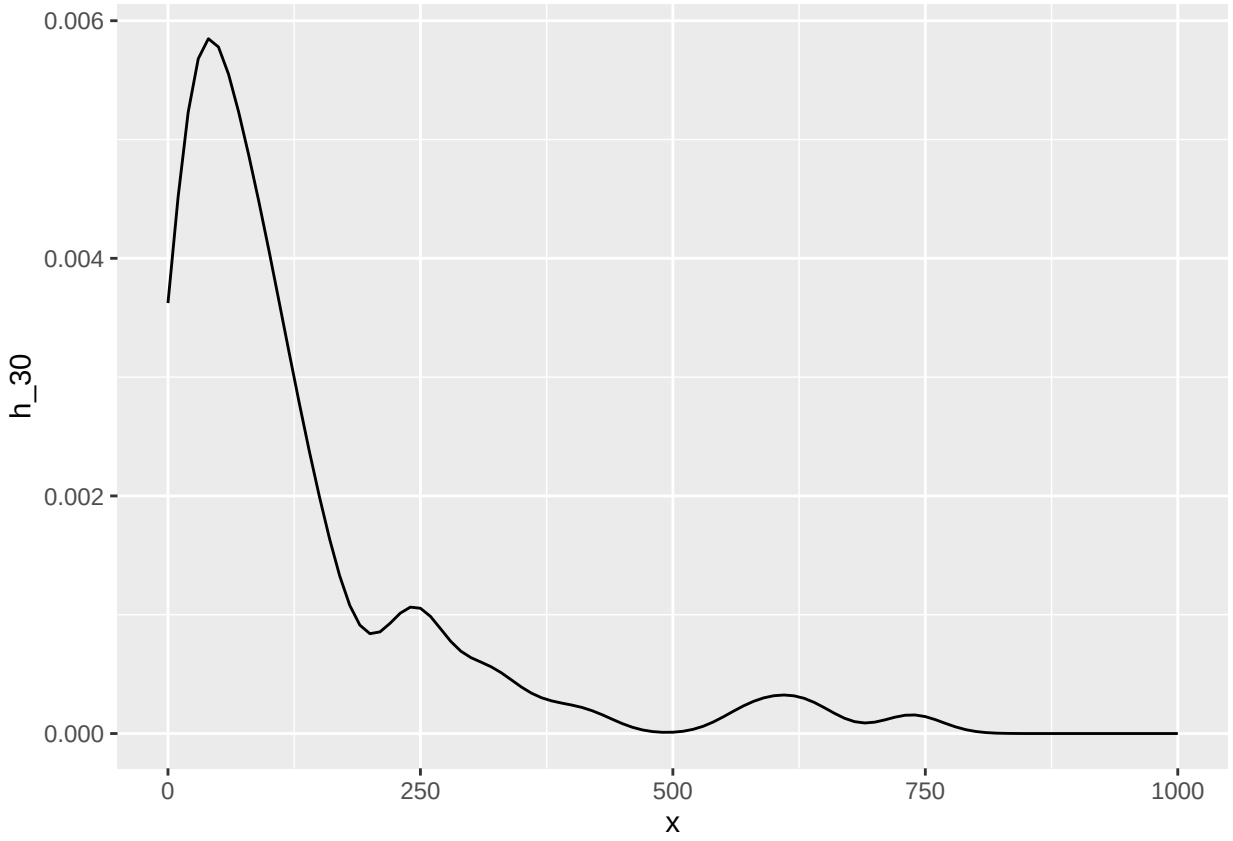
Cargamos los datos:

```
library(tidyverse)
tratamiento <- read.csv("Tratamiento.csv")
```

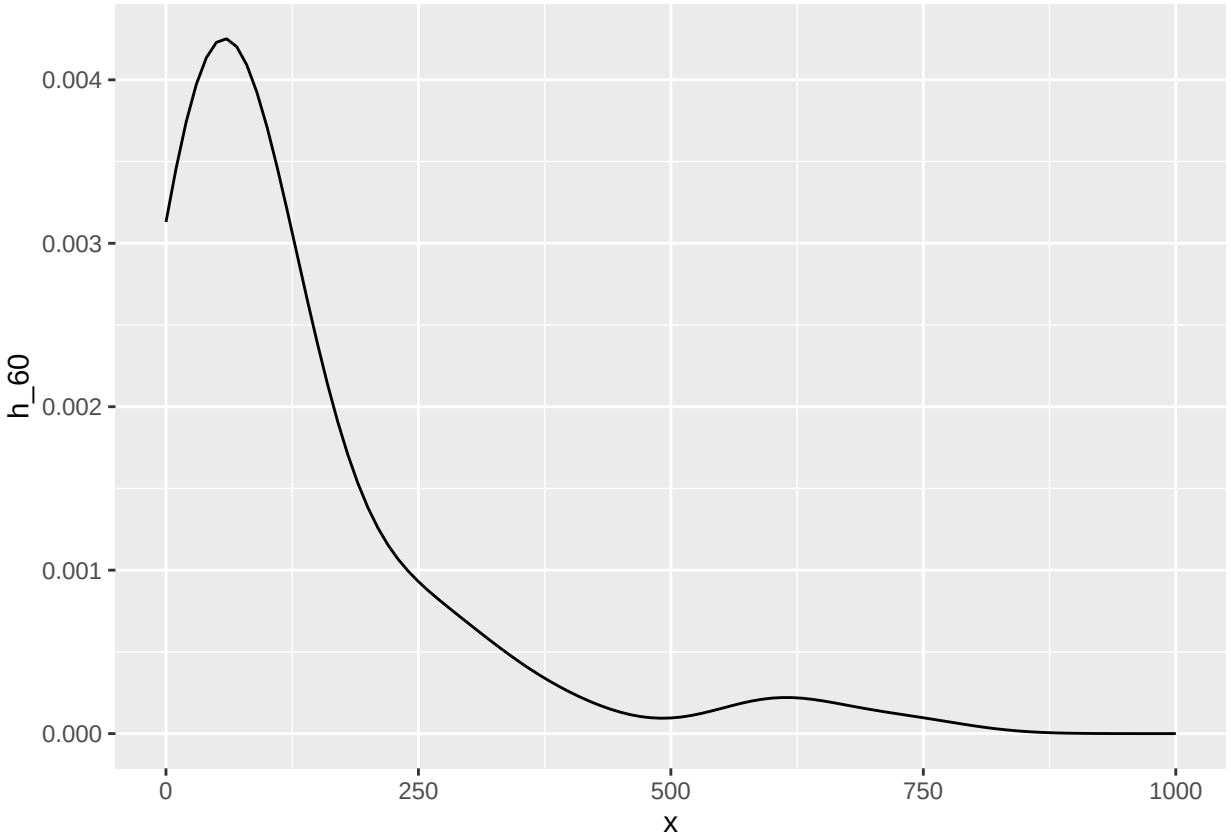
```
fd_tratamiento <-data.frame(x=seq(0,1000, 10))

fd_tratamiento$h_20<- kde(fd_tratamiento$x, 20, "gaussian", tratamiento$X1)

fd_tratamiento$h_30 <- kde(fd_tratamiento$x, 30, "gaussian", tratamiento$X1)

fd_tratamiento$h_60 <- kde(fd_tratamiento$x, 60, "gaussian", tratamiento$X1)

ggplot(fd_tratamiento, aes(x, h_20))+
  geom_line()
```




```
ggplot(fd_tratamiento, aes(x, h_30))+
  geom_line()
```



```
ggplot(fd_tratamiento, aes(x, h_60))+
  geom_line()
```



c) En el contexto de la estimación de densidades, escriba una función en R que determine el ancho de banda que optimiza al ISE. Grafique la densidad con ancho de banda óptimo para el conjunto de datos de "Tratamiento.csv".

6. Cargue en R al conjunto de datos "Maíz.csv", el cual contiene el precio mensual de la tonelada de maíz y el precio de la tonelada de tortillas en USD. En este ejercicio tendrá que estimar los coeficientes de una regresión lineal simple.

a) Calcule de forma explícita la estimación de los coeficientes via mínimos cuadrados y ajuste la regresión correspondiente. Concluya.

b) Calcule de forma explícita la estimación de los coeficientes via regresión no-paramétrica tipo kernel (ver Nadaraya, E. A. (1964). "On Estimating Regression". Theory of Probability and its Applications. 9 (1): 141–2. doi:10.1137/1109020) y ajuste la regresión correspondiente. Concluya.

c) Compare ambos resultados. ¿Qué diferencias observa?

8. En este ejercicio se comprobará que tan buena es la aproximación dada por las reglas empíricas para algunas de las distribuciones estudiadas en la clase. Considerese las distribuciones $Unif(a = -3, b = 3)$, $Normal(0, 1)$, $Exponencial(2)$, $Gamma(\alpha = 2, \beta = 1)$, $Gamma(\alpha = 3, \beta = 1)$, $Beta(\alpha = 2, \beta = 2)$, $Weibull(\alpha = 4, \beta = 1)$ y Lognormal$(\mu = 3, \sigma = 2)$.

a) Para cada una de las distribuciones anteriores, haga una tabla que muestre las probabilidades contenidas en los intervalos $(\mu - k\sigma, \mu + k\sigma)$, para $k = 1, 2, 3$. Utilice las fórmulas de las medias y varianzas contenidas en las notas para determinar μ y σ en cada caso. Puede usar R para determinar las probabilidades pedidas.

RESPUESTA

Determinemos para cada distribución su media y varianza,

- Si $X \sim Unif(a, b)$,
$$E[X] =, \quad Var(X)$$
- Si $X \sim Normal(\mu, \sigma)$,
$$E[X] =, \quad Var(X)$$
- Si $X \sim Exponencial(\theta)$,
$$E[X] =, \quad Var(X)$$
- Si $X \sim Gamma(\alpha, \beta)$,
$$E[X] =, \quad Var(X)$$
- Si $X \sim Beta(\alpha, \beta)$,
$$E[X] =, \quad Var(X)$$
- Si $X \sim Weibull(\alpha, \beta)$,
$$E[X] =, \quad Var(X)$$
- Si $X \sim LogNormal(\mu, \sigma)$,
$$E[X] =, \quad Var(X).$$

b) En R, simule $n = 1000$ muestras de cada una de las distribuciones anteriores y calcule la media muestral $\bar{x}$ y la varianza muestral s^2 como se mencionó en la clase. En cada caso, calcule la proporción de observaciones que quedan en los intervalos $(\bar{x} - k\sigma, \bar{x} + k\sigma))$, para k = 1, 2, 3. Reporte sus hallazgos en una tabla como la del inciso anterior. ¿Qué tanto se parecen la tabla de este inciso y la del anterior?