

Tarea sobre los Temas 3 a 5

Grado en Estadística. Universidad de Sevilla

Marta Venegas Pardo

Contents

1	Pregunta 1	1
1.1	Realizar una estimación no paramétrica de la función de densidad por el método del núcleo. .	2
1.2	Realizar una estimación no paramétrica de la función de densidad por el método de los logsplines.	3
1.3	Estimar $P[\text{precip} > 42]$ y el cuantil 0.90.	5
2	Pregunta 2	5
2.1	Fijar M (número de ofertas, se recomienda al menos 1000).	5
2.2	Definir una matriz Mx4 donde se irán almacenando los valores generados.	5
2.3	Repetir M veces:	5
3	Pregunta 3	6
4	Pregunta 4	6

1 Pregunta 1

En el dataset `datasets::precip` se recoge la cantidad media de precipitaciones de 70 ciudades de Estados Unidos (unidad = inches).

Extraigo los datos

```
precip<-datasets::precip
# View(precip)

#dotchart(precip[order(precip)], main = "precip data")
#title(sub = "Average annual precipitation (in.)")

precip_data<-as.data.frame(precip)
names<-names(precip)
precipdf<-cbind.data.frame(names,precip_data)
colnames(precipdf)<-c("City","Precip")

cbind.data.frame(precipdf[1:35,],precipdf[36:70,])%>%
  kable(booktabs=TRUE,longtable=TRUE) %>%
  kable_styling(latex_options = c("striped","scale_down")) %>%
  footnote(general = "Annual Precipitation in US Cities")

## Warning in styling_latex_scale_down(out, table_info): Longtable cannot be
## resized.
```

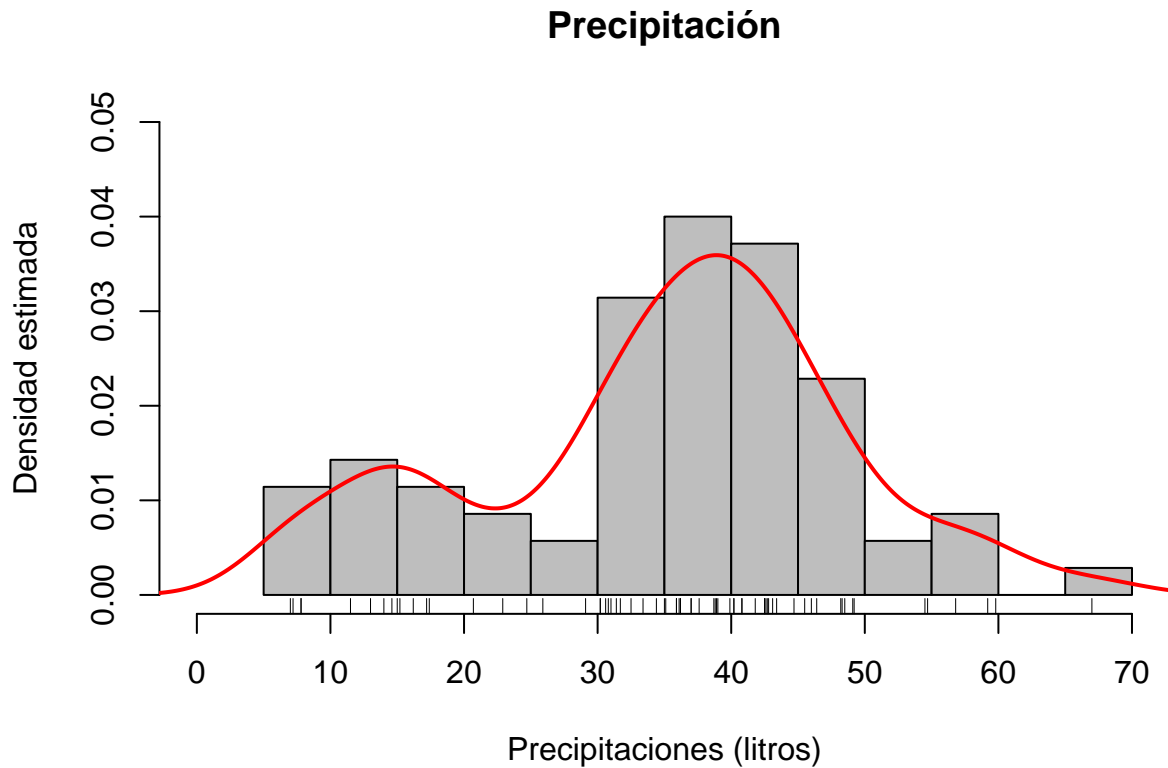
City	Precip	City	Precip
Mobile	67.0	Reno	7.2
Juneau	54.7	Concord	36.2
Phoenix	7.0	Atlantic City	45.5
Little Rock	48.5	Albuquerque	7.8
Los Angeles	14.0	Albany	33.4
Sacramento	17.2	Buffalo	36.1
San Francisco	20.7	New York	40.2
Denver	13.0	Charlotte	42.7
Hartford	43.4	Raleigh	42.5
Wilmington	40.2	Bismark	16.2
Washington	38.9	Cincinnati	39.0
Jacksonville	54.5	Cleveland	35.0
Miami	59.8	Columbus	37.0
Atlanta	48.3	Oklahoma City	31.4
Honolulu	22.9	Portland	37.6
Boise	11.5	Philadelphia	39.9
Chicago	34.4	Pittsburg	36.2
Peoria	35.1	Providence	42.8
Indianapolis	38.7	Columbia	46.4
Des Moines	30.8	Sioux Falls	24.7
Wichita	30.6	Memphis	49.1
Louisville	43.1	Nashville	46.0
New Orleans	56.8	Dallas	35.9
Portland	40.8	El Paso	7.8
Baltimore	41.8	Houston	48.2
Boston	42.5	Salt Lake City	15.2
Detroit	31.0	Burlington	32.5
Sault Ste. Marie	31.7	Norfolk	44.7
Duluth	30.2	Richmond	42.6
Minneapolis/St Paul	25.9	Seattle Tacoma	38.8
Jackson	49.2	Spokane	17.4
Kansas City	37.0	Charleston	40.8
St Louis	35.9	Milwaukee	29.1
Great Falls	15.0	Cheyenne	14.6
Omaha	30.2	San Juan	59.2

Note:

Annual Precipitation in US Cities

1.1 Realizar una estimación no paramétrica de la función de densidad por el método del núcleo.

```
x=precipdf[,2]
hist(precipdf[,2], br=10,prob=TRUE, main="Precipitación",
     col="gray",
     xlab="Precipitaciones (litros)", ylab="Densidad estimada",
     ylim = c(0,0.05),xlim = c(0,70))
lines(density(x, bw="SJ"),lwd=2,col="red")
rug(x)
```



1.2 Realizar una estimación no paramétrica de la función de densidad por el método de los logsplines.

```
library(logspline)
ajuste <- logspline(precip_data)
ajuste # 7nudos, criterio BIC
```

```
## knots A(1)/D(2) loglik AIC minimum penalty maximum penalty
## 4 2 -282.85 578.45 15.02 Inf
## 5 2 -275.34 567.68 0.51 15.02
## 6 2 -275.29 571.82 NA NA
## 7 2 -274.83 575.15 0.41 0.51
## 8 2 -274.79 579.31 NA NA
## 9 2 -274.41 582.82 0.02 0.41
## 10 1 -274.40 587.04 0.00 0.02
## the present optimal number of knots is 5
## penalty(AIC) was the default: BIC=log(samplesize): log( 70 )= 4.25
```

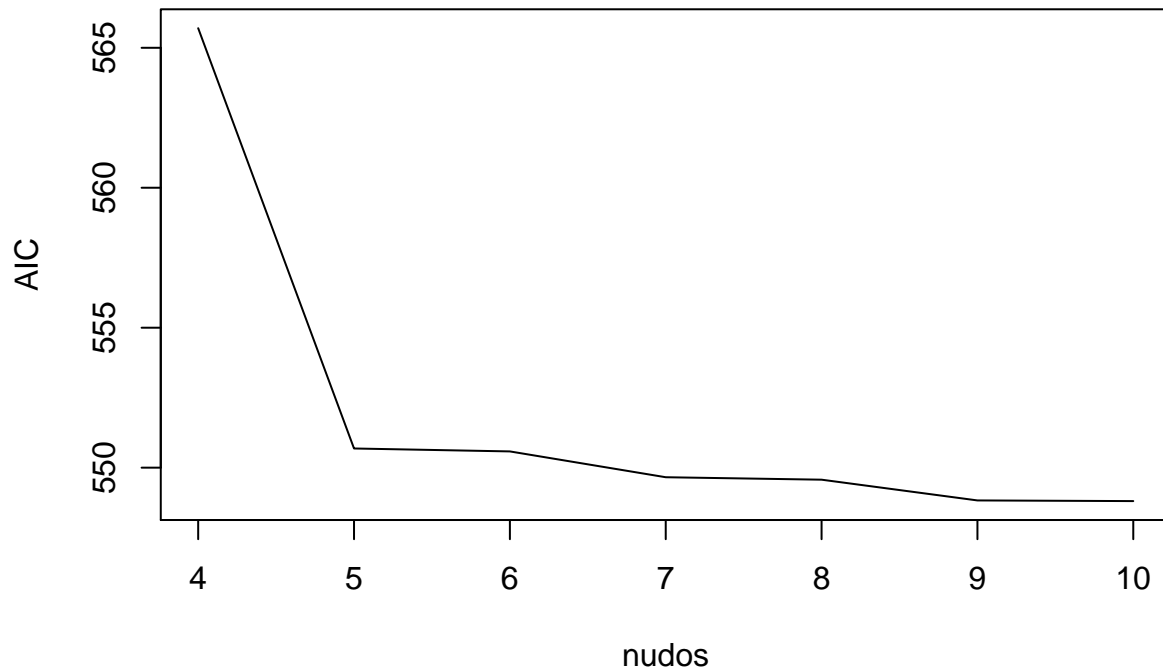
Vamos a comprobar los valores AIC

```
resul<- ajuste$logl
nudos<- resul[,1]
logL<- resul[,3]
AIC<- -2*logL+log(length(precip_data))*(nudos-1)
AIC
```

```
## [1] 565.7001 550.6848 550.5803 549.6576 549.5704 548.8295 548.8057
```

Lo dibujamos

```
plot(nudos,AIC,type="l")
```



Ahora el histograma

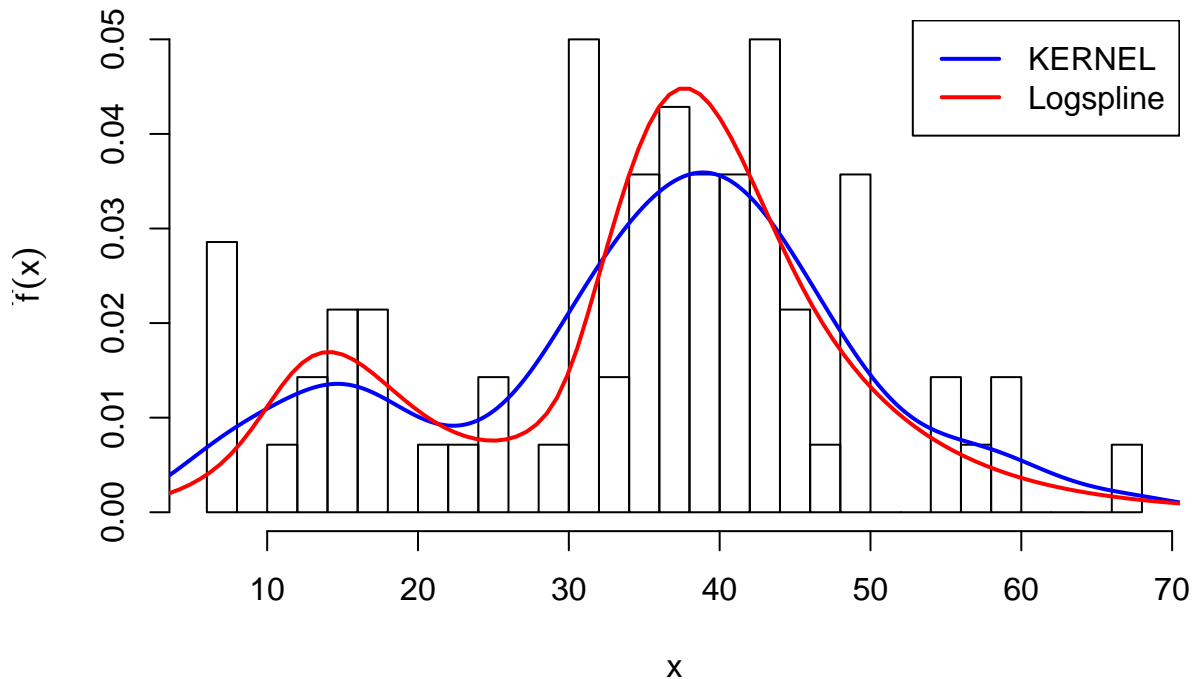
```
hist(x,br=30, prob=TRUE,
     main="Histograma y estimac. de la densidad",
     ylab = expression(hat(f)(x)),
     xlab="x")

lines(density(precipdf[,2],bw="SJ"),col="blue",lwd=2)

plot(ajuste,col="red",lwd=2,add=TRUE)

legend("topright",
      col=c("blue","red"),
      lwd=2,
      legend=c("KERNEL","Logspline"))
```

Histograma y estimac. de la densidad



1.3 Estimar $P[\text{precip} > 42]$ y el cuantil 0.90.

```
1-plogspline(42, ajuste)
```

```
## [1] 0.2865462
```

```
qlogspline(0.9, ajuste)
```

```
## [1] 50.17582
```

2 Pregunta 2

Cuando cierta empresa recibe una invitación para optar a un contrato, la oferta no se puede completar hasta que sea revisada por cuatro departamentos: Ingeniería, Personal, Legal y Contabilidad. Los departamentos empiezan a trabajar al mismo tiempo, pero lo hacen de forma independiente. El tiempo en semanas que emplean en completar la revisión es una variable aleatoria con las siguientes distribuciones. Ingeniería: Exponencial con media 3 semanas; Personal: Normal con media 4 y desviación típica 1; Legal: 2 o 4 semanas, siendo ambos valores equiprobables; Contabilidad: Uniforme continua en el intervalo (1,5). Se trata de simular el tiempo W que tarda la empresa en preparar una oferta. Para ello se pueden implementar los siguientes pasos:

2.1 Fijar M (número de ofertas, se recomienda al menos 1000).

2.2 Definir una matriz $M \times 4$ donde se irán almacenando los valores generados.

2.3 Repetir M veces:

- Generar de forma independiente los cuatro tiempos según las cuatro distribuciones.
- Guardar esos cuatro tiempos en una fila de la matriz.

- Calcular W como el máximo de los cuatro tiempos.

Se pide:

- Estudiar gráficamente la distribución de la variable aleatoria W “tiempo que transcurre hasta completar la oferta”. Estimar su media y su mediana.
- Estimar la probabilidad de que W supere las 6 semanas.
- ¿Cuál es el departamento que suele tardar más en completar la revisión?
- ¿Cuál es la ordenación más frecuente de los cuatro tiempos?

3 Pregunta 3

Generar aleatoriamente un conjunto de datos donde tengo sentido construir un modelo de clasificación o de predicción. Ajustar el modelo y estimar su capacidad de generalización mediante Jackknife y mediante Validación Cruzada (K=10).

4 Pregunta 4

Bootstrap. Implementar una función que calcule el estadístico de Fisher de comparación de coeficientes de correlación lineal:

$$T = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

Donde:

- Z1 y Z2 representan la transformación de Fisher de los respectivos coeficientes de correlación lineal para dos grupos
- n1 y n2 son las frecuencias absolutas respectivas de los grupos.

Sobre un conjunto de datos apropiado (que puede ser generado), utilizando el anterior estadístico, realizar e interpretar un test bootstrap bilateral de comparación de los coeficientes de correlación lineal (B=1999).