Práctica 11

Prueba de normalidad analítica

Luis Eduardo Galindo Amaya (1274895)

Asignatura	Estadística Avanzada
Docente	Olivia Mendoza Duarte
Fecha	23-11-2022

Prueba de normalidad analítica

Luis Eduardo Galindo Amaya (1274895)

23-11-2022

Información del dataset[1]

This is one of the best known datasets in statistics and machine learning. Fisher's paper is a classic in the field and is frequently used for tutorial and teaching purposes. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are not linearly separable from each other.

Predicted attribute: class of iris plant.

1. sepal length
2. sepal width
3. petal length
4. petal width
5. class

Desarollo de la práctica

Distribusion normal (sepal width)

Observacion previa

Se puede notar como el ancho del sépalo se distribuye de manera normal, el tamaño del sépalo tiende a ser de un tamaño especifico y no tanto del tamaño de los pétalos

Resultados

Efectivamente los valores coinciden en una distribucion normal $RQ = 0{,}9925113$.

[1] `https://archive-beta.ics.uci.edu/ml/datasets/iris`

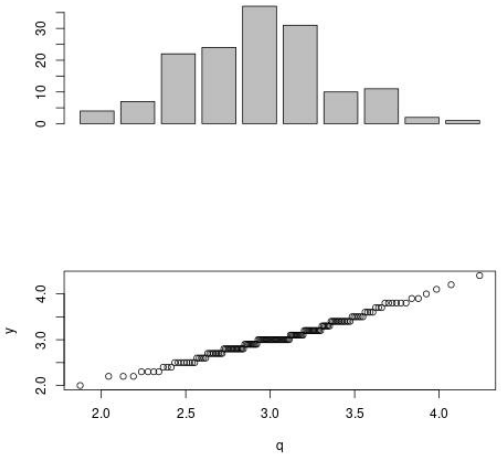

Figura 1: columna "sepal width"

Distribucion no normal (sepal length)

Observacion previa

El largo del sépalo se distribuye de manera serrada sobre los datos, no parece haber un patrón en la gráfica

Resultados

Aunque de a primera vista no es muy claro los datos de esta columna tambien cumplen la prueba de normalidad $rQ = 0{,}9891878$.

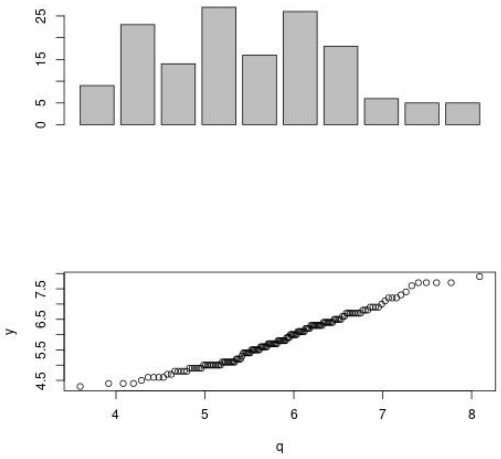

Figura 2: columna "sepal length" del dataset bezdekIris

Distribucion no normal (petal length)

Observacion previa

Por otro lado el largo de los pétalos es muy interesante, en la parte derecha parece haber una distribución normal pero tiene unos sectores que sobresalen a la izquierda.

Resultados

Esta no es una distribución normal, la gráfica sale completamente dividida $rQ = 0{,}9378633$

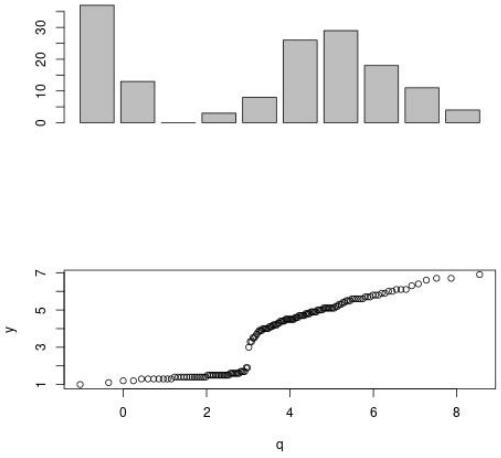

Figura 3: columna "petal length" del dataset bezdekIris

Conclusiones

Usar métodos que solo dependen de la observación es bastante peligroso, ya que no podemos determinar con total precisión si nuestras gráficas coinciden o no y que un simple cambio en el numero de clases puede hacer que la gráfica sea completamente diferente, el método analítico nos permite tener mucha precisión para determinar si nuestros casos coinciden

Código

```
## AUTHOR: Luis Eduardo Galindo Amaya
##   DATE: 24-11-2022
##   DESC: Prueba de normalidad grafica y analitica

## ENTRADAS
n_clases <- 10
nombre_del_archivo <- "./bezdekIris.csv"

columna <- 3
filas <- 1:150

archivo_salida <- "./practica 11/img/i3.jpeg"

```

```
## OPERACIONES GRAFICA ####
archivo <- read.csv(nombre_del_archivo)
data <- archivo[filas, columna]

valor_minimo <- min(data)
valor_maximo <- max(data)
amplitud_de_clase <- (valor_maximo - valor_minimo) / n_clases

## numero de elementos dentro de cada clase
frecuencias <- array(0, dim = (n_clases))

for (i in 1:n_clases) {
  rango_min <- valor_minimo + amplitud_de_clase * (i - 1)
  rango_max <- valor_minimo + amplitud_de_clase * i
  frecuencias[i] <- sum(data >= rango_min & data < rango_max)
}


## OPERACIONES ANALÍTICA ####
n <- length(data) # tamaño del arreglo

## Se calcula las fracciones las fracciones correspondientes a
## los acuartiles teoricos

fracciones <- c()
xdata <- c()

for (i in 1:n) {
  fraccion <- (i - 0.5) / n
  fracciones <- c(fracciones, fraccion)
  xdata <- c(xdata, data[i])
}

# se ordena el vector x y se asiga ordenado a la variable y
y <- sort(xdata)

# se calcula la media y la desviación estándar de y
media_y <- mean(y)
std_y <- sd(y)

## se calculan los cuartiles teóricos usando la distribució
## normal inversa qnorm

q <- c()
for (i in 1:n) {
  qi <- media_y + std_y * qnorm(fracciones[i])
  q <- c(q, qi)
}

datosQQ=data.frame(q,y)
rQ=cor(q,y)




## grafica
jpeg(file = archivo_salida)
par(mfrow = c(2, 1))
barplot(frecuencias)
plot(datosQQ)
dev.off()

## SALIDAS GRAFICAS ####
frecuencias
n_clases
valor_minimo
valor_maximo
amplitud_de_clase
length(data)
sum(frecuencias)
## SALIDAS ANALÍTICAS ####
rQ
```