Práctica 10

Prueba de normalidad

Luis Eduardo Galindo Amaya (1274895)

Asignatura	Estadística Avanzada
Docente	Olivia Mendoza Duarte
Fecha	16-11-2022

Prueba de normalidad

Luis Eduardo Galindo Amaya (1274895)

16-11-2022

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.

Desarollo de la práctica

1. Probar el código proporcionado en recursos para la prueba de normalidad, cambiando el número de clases a 5, 6, 8 y 10

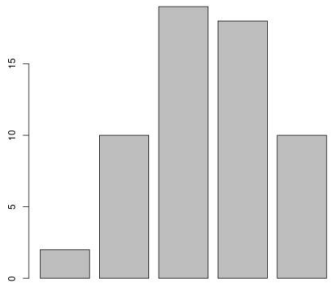

Figura 1: 5 clases

2. Probar este código adaptándolo al menos hasta encontrar alguna que muestre evidencia de distribución normal

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

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

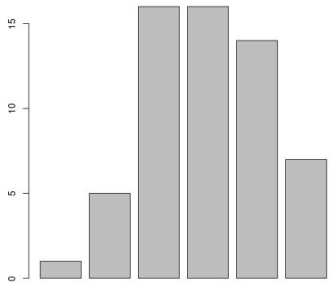

Figura 2: 6 clases

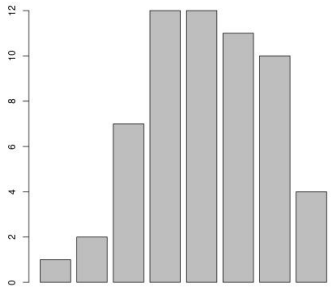

Figura 3: 8 clases

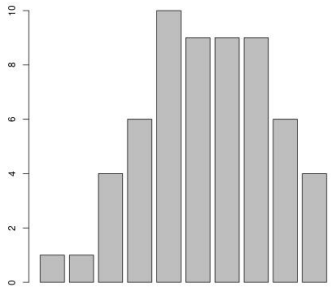

Figura 4: 10 clases

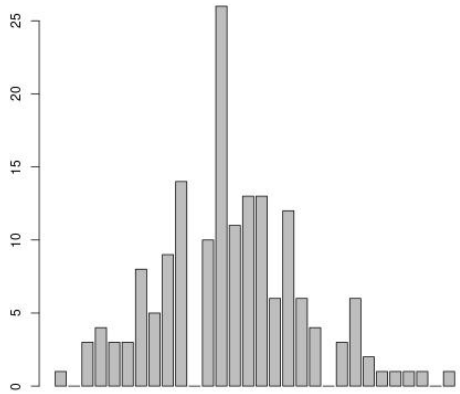

Figura 5: columna "sepal width" del dataset bezdekIris

3. Reportar en un documento todos los resultados obtenidos aún que el comportamiento de los datos no de evidencia de distribución normal

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

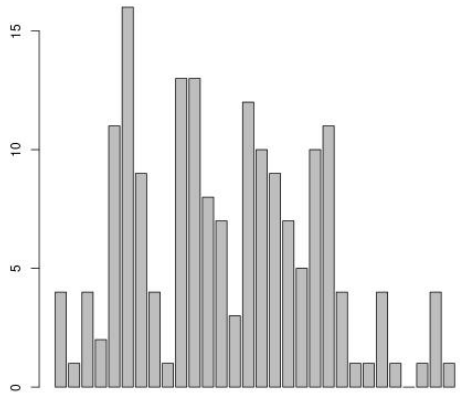

Figura 6: columna "sepal length" del dataset bezdekIris

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.

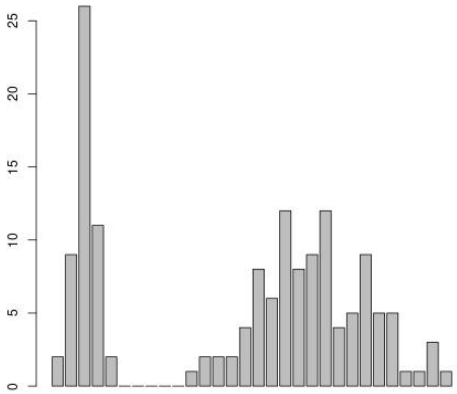

Figura 7: columna "petal length" del dataset bezdekIris

Código

```
## AUTHOR: Luis Eduardo Galindo Amaya
## DATE:   21-10-2022

## ENTRADAS
n_clases <- 30
nombre_del_archivo <- "./bezdekIris.csv"
## nombre_del_archivo <- "./practica 10/test.csv"
## salida <- "./practica 10/img/iris1.jpeg"
columna <- 1

## OPERACIONES
archivo <- read.csv(nombre_del_archivo)
data <- archivo[, 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 + 0.00001
  frecuencias[i] <- sum(data >= rango_min & data < rango_max)
}

## SALIDAS
n_clases
valor_minimo
valor_maximo
amplitud_de_clase
length(data)
sum(frecuencias)
frecuencias

## grafica
## jpeg(file = salida)
barplot(frecuencias)
## dev.off()
```