Análisis de la distribución de la distancia

Code

Usando la colección de datos *iris* vamos a proceder análizar como se comporta la distancia Euclediana entre dos puntos aleatorios del espacio vectorial.

Hide

```
View(iris)
```

Para ello vamos a definir primero la función distancia aplicado a dos vectores multidimensionales.

Hide

```
ED <- function(X, Y){
  return(sqrt(sum((X - Y)^2)))
}
```

También se requiere una función para calcular la distancia entre N pares de puntos y almacenarlos en un array.

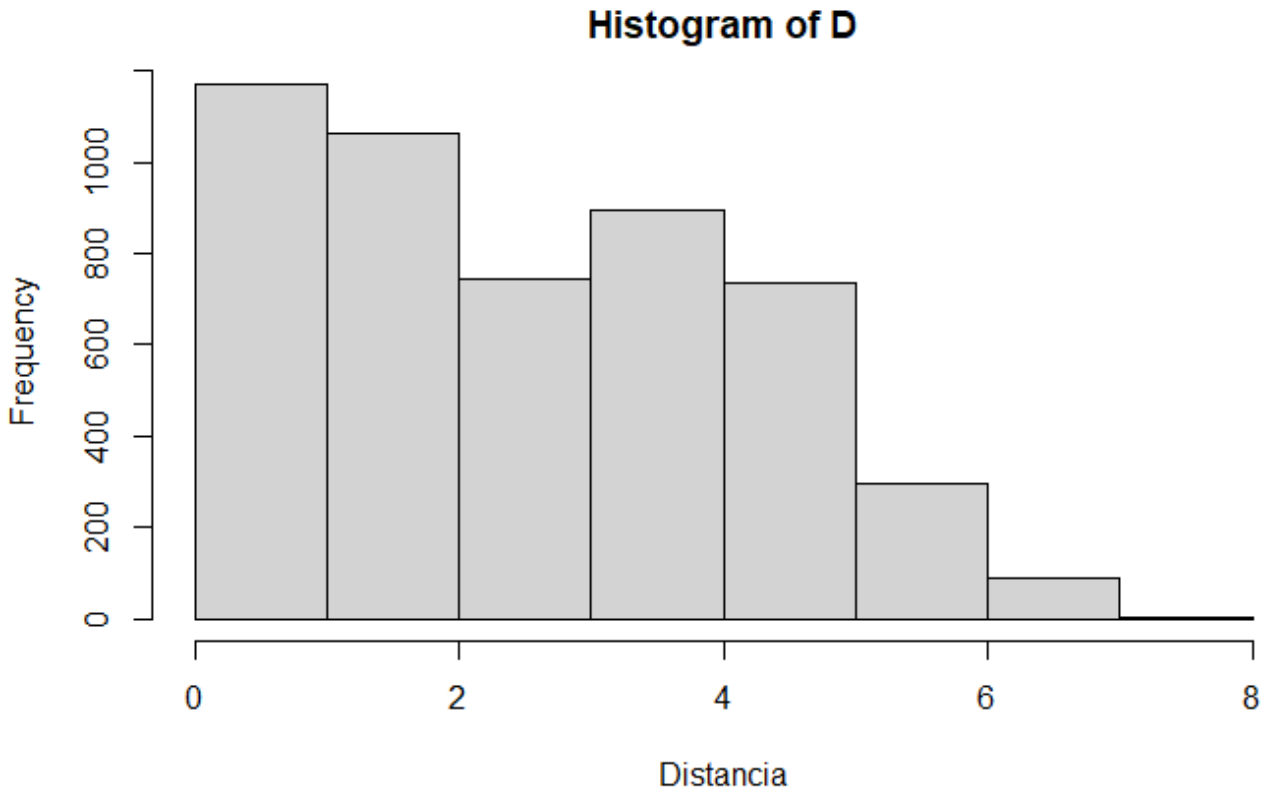
Hide

```
genDistancias<-function(data, N){
  v<-rep(0, N)
  for (i in 1:N) {
    ind<-sample(1:nrow(data), size=2)
    P<-data[ind[1], ]
    Q<-data[ind[2], ]
    v[i]<-ED(P, Q)
  }
  return(v)
}
```

Generamos 5000 distancias aleatorias y procedemos a visualizar los resultados en un histograma,

Hide

```
D<-genDistancias(iris[, 1:4], 5000)
#generamos el histograma
H<-hist(D, xlab = "Distancia", breaks = 10)
```



Para poder saber que radio usar en las búsquedas por rango, podemos guiarnos del porcentaje de elementos que cubre la distancia conforme va creciendo.

Hide

```
#porcentaje de cobertura
for (i in 1:(length(H$counts)-1)) {
  print(paste("Radio <=" , H$breaks[i+1],": ", round(100*sum(H$counts[1:i])/sum(H$counts)),
"%"))
}
```

```
[1] "Radio <= 1 :  23 %"
[1] "Radio <= 2 :  45 %"
[1] "Radio <= 3 :  60 %"
[1] "Radio <= 4 :  78 %"
[1] "Radio <= 5 :  92 %"
[1] "Radio <= 6 :  98 %"
[1] "Radio <= 7 :  100 %"
```