Tarea # 1
Teoría de la probabilidad y estadística
Fecha de entrega: 13 de diciembre de 2021, hasta 11 pm.

El trabajo puede hacerse en grupos de 2 personas. Es permitida solo la discusión de los problemas de la tarea entre los integrantes del grupo; sin embargo, cada estudiante debe montar una copia de la respuesta al classroom y especificar en la solución el compañero con el que trabajó.

Problema 1. Hace cerca de dos años a nivel mundial se desató la pandemia generada por el virus SARS-CoV-2. En Colombia se han llevado registros de los casos que han resultado positivos como portadores del virus, lo cual ha permitido a los científicos desarrollar estrategias que permitan disminuir los efectos negativos de la pandemia. En el archivo 2021-11-20.zip podrá encontrar la base de datos donde se han registrado los casos positivos, reportados por el instituto nacional de salud. Con el fin de que usted ponga en práctica los nuevos conocimientos adquiridos en el curso, se proponen las siguientes actividades:

a. Cree una función en R que permita construir (Dibujar) un histograma con un conjunto cualquiera de datos. Esta función debe permitir al usuario seleccionar que tipo de histograma desea (Frecuencias, Frecuencias relativas o densidades). También debe permitir al usuario indicar cuantas clases desea en el histograma, pero si el usuario no lo indica, la función debe seleccionar la cantidad de clases con el criterio visto en clase.

Para realizar la tarea se importaron los datos del archivo de Excel original, en el que estaban todos los datos, este fue el primer archivo subido por el profesor y decidí hacerlo con este ya que creo que tenía un poco más de complejidad y era más interesante. Se añaden las filas del documento a una lista llamada *datos*.

```
1  datos<-read.csv("C:\\Users\\santi\\Desktop\\Tarea_1\\Salida_Datos_Abiertos.csv")
2  for (i in length(datos)){
3    datos[i]<-c(datos[i])
4  }
```

Debido a que todos los datos del documento no eran numéricos y R solo permite realizar histogramas con valores numéricos, se crea una función llamada ***convertir*** que se encarga de convertir los valores de las columnas a valores numéricos para así tener mayor facilidad a la hora de hacer los histogramas. El argumento de entrada para esta función es la lista con los datos a los que se les desea realizar el cambio por valores numéricos.

La función retorna una lista de listas, la primera lista es la que tiene los valores numéricos y la segunda lista es la que contiene los valores en letras o fechas según sea el caso.

```
convertir <- function(columna) {
  j=1
  casos_unicos<-c()
  for (i in 1:length(datos[[columna]])){
    if ((datos[[columna]][[i]] %in% casos_unicos)==FALSE){
      casos_unicos[[j]]=datos[[columna]][[i]]
      j=j+1
    }
  }
  lista<-c()
  for (k in casos_unicos){
    pos=which(datos[[columna]]==k)
    for (l in pos){
      lista[l]=pos[1]
    }
  }
  resultados<-c()
  resultados[[1]]=lista
  resultados[[2]]=casos_unicos
  return(resultados)
}
```

Luego se utiliza la función ***convertir*** para cada una de las columnas necesarias, se separan por columnas con datos no numéricos y columnas con datos numéricos.

```
#No numericos
fecha_hoy_casos=convertir(1)[[1]]
Fecha_Not=convertir(3)[[1]]
Departamento_nom=convertir(5)[[1]]
Ciudad_municipio_nom=convertir(7)[[1]]
Sexo=convertir(10)[[1]]
Fuente_tipo_contagio=convertir(11)[[1]]
Ubicacion=convertir(12)[[1]]
Estado=convertir(13)[[1]]
Pais_viajo_1_cod=convertir(14)[[1]]
Pais_viajo_1_nom=convertir(15)[[1]]
Recuperado=convertir(16)[[1]]
Fecha_inicio_sintomas=convertir(17)[[1]]
Fecha_muerte=convertir(18)[[1]]
Fecha_diagnostico=convertir(19)[[1]]
Fecha_recuperado=convertir(20)[[1]]
Tipo_recuperacion=convertir(21)[[1]]
nom_grupo=convertir(23)[[1]]

#Numericos:
Caso=datos[[2]]
Departamento=datos[[4]]
Ciudad_municipio=datos[[6]]
Edad=datos[[8]]
unidad_medida=datos[[9]]
per_etn_=datos[[22]]
```

Ahora sí se crea la función para graficar cualquiera de las columnas que hay en el documento, la función se llama ***histograma*** y tiene como argumentos de entrada el conjunto de datos que se desea graficar, cualquiera de las columnas del documento, tipo de histograma que se desea realizar y el número de clases, para este caso si es un argumento vacío (“”), la función toma como el numero de clases el valor de 2246 que es el valor aproximado de la raíz cuadrada del total de datos que hay en cada una de las columnas.

```
histograma <- function(conjunto,tipo,clases){
  if (clases==""){
    clases=2246
  }
  else{
  }
  largo=length(conjunto)
  conjunto=sort(conjunto)
  ancho=ceiling((conjunto[[largo]]-conjunto[[1]])/clases)
  minimo=conjunto[[1]]-1
  maximo=conjunto[[largo]]+1
  limites<-c()
  a=1
  while (maximo>minimo){
    limites[[a]]=minimo+ancho
    a=a+1
    minimo=limites[[a-1]]
  }
  frecuencia<-c()
  b=1
  conteo=0
  for (i in 1:length(conjunto)){
    lim_sup=limites[[b]]
    if (lim_sup>=conjunto[[i]]){
      conteo=conteo+1
    }
    else{
      frecuencia<- c(frecuencia,conteo)
      b=b+1
      conteo=0
```

```r
histograma <- function(conjunto,tipo,clases){
  if (clases==""){
    clases=2246
  }
  else{
  }
  conjunto=sort(conjunto)
  ancho=ceiling((conjunto[[5045412]]-conjunto[[1]])/clases)
  minimo=conjunto[[1]]-1
  maximo=conjunto[[5045412]]+1
  limites<-c()
  a=1
  while (maximo>minimo){
    limites[[a]]=minimo+ancho
    a=a+1
    minimo=limites[[a-1]]
  }
  frecuencia<-c()
  b=1
  conteo=0
  for (i in 1:length(conjunto)){
    lim_sup=limites[[b]]
    if (lim_sup>=conjunto[[i]]){
      conteo=conteo+1
    }
    else{
      frecuencia<- c(frecuencia,conteo)
      b=b+1
      conteo=0
    }
  }
  frecuencia<- c(frecuencia,conteo)
  relativa<-c()
  densidad<-c()
  for (i in 1:length(frecuencia)){
    f=frecuencia[[i]]
    relativa<- c(relativa,f/length(conjunto))
    densidad<- c(densidad,f/(length(conjunto)*ancho))
  }
  if (tipo="Frecuencia"){
    barplot(frecuencia)
  }
  else if (tipo="Relativa"){
    barplot(relativa)
  }
  else if (tipo="Densidad"){
    barplot(densidad)
  }
  else{
    print("Seleccione entre Frecuencia, Relativa o Densidad")
  }
}
```

b. Utilice la función construida por usted mismo, para realizar un histograma de densidades con las edades de las personas contagiadas en Medellín. Interprete los resultados, mostrando los factores relevantes de su resultado.

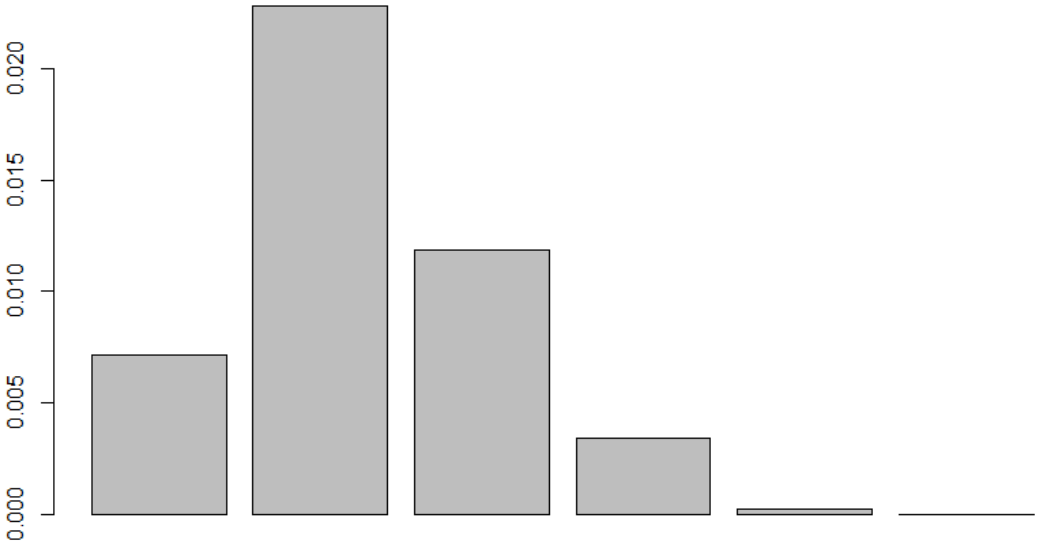
Para este punto se realizó otra función en la cual se filtran los resultados para solo los habitantes de la ciudad de Medellín quienes aparecen en el documento como *MEDELLIN*.

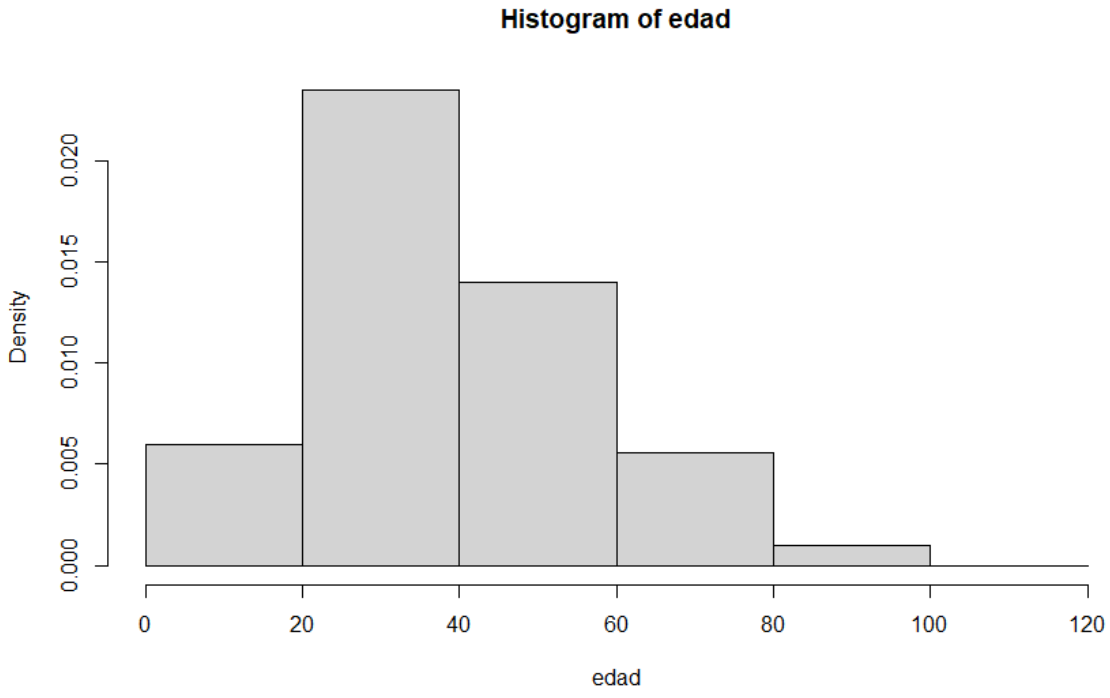
Se creó la función *puntob*, la cual tiene como argumentos de entrada las columnas de ciudad y edad en ese orden, la función nos retorna una lista con las edades de las personas de la ciudad de Medellín.

```
109 puntob <- function(ciudades,edades){
110   edad<-c()
111   for (i in 1:length(ciudades)){
112     if (ciudades[[i]]=="MEDELLIN"){
113       edad<- c(edad,edades[[i]])
114     }
115   }
116   return(edad)
117 }
```

Se grafican ambos histogramas, primero el histograma con la función realizada por el estudiante y el segundo histograma es utilizando la función *hist* de R.

```
119 edad=puntob(datos[[7]],datos[[8]])
120 clases=5
121 histograma(edad,"Densidad",clases)
122 hist(edad,freq=FALSE,breaks=clases)
```





Según los histogramas, (gráficamente no es posible identificar una diferencia notable entre ellos), la mayor cantidad de personas tiene o tenía entre los 20 y los 40 años, siendo muy evidente en la gráfica como sobresale frente a los demás, era de esperarse ya que la mayoría

de los trabajadores se encuentran en esta edad, entre los 100 y los 120 años no se tiene mucha cantidad puesto que llegar a esta edad es difícil.

Podría decirse que ambos gráficos son casi iguales dados sus valores en el eje y correspondientes a la densidad y a la altura de las barras del gráfico.

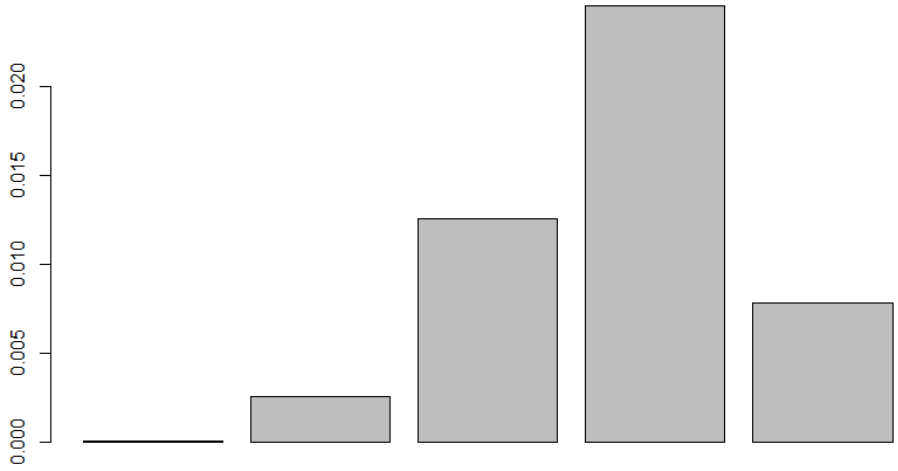
c. Utilice la función construida por usted mismo, para realizar un histograma de densidades con las edades de las personas fallecidas en Medellín. Interprete los resultados y compare el nuevo histograma con el del numeral b.

Se crea una función de manera similar a la anterior, en donde se filtran por la ciudad de Medellín y ahora los fallecidos, obteniendo así la edad de los fallecidos de la ciudad de Medellín.

```
126 puntoc <- function(ciudades,edades,estado){
127   fallecidos<-c()
128   for (i in 1:length(ciudades)){
129     if (ciudades[[i]]=="MEDELLIN"&(estado[[i]]=="Fallecido")){
130       fallecidos<- c(fallecidos,edades[[i]])
131     }
132   }
133   return(fallecidos)
134 }
```

Se reutiliza la función histograma con el fin de obtener el histograma de densidades, el numero de clases es el mismo que se utilizó en el punto anterior, *clases=5*.

```
135 fallecido=puntoc(datos[[7]],datos[[8]],datos[[13]])
136 histograma(fallecido,"Densidad",clases)
```

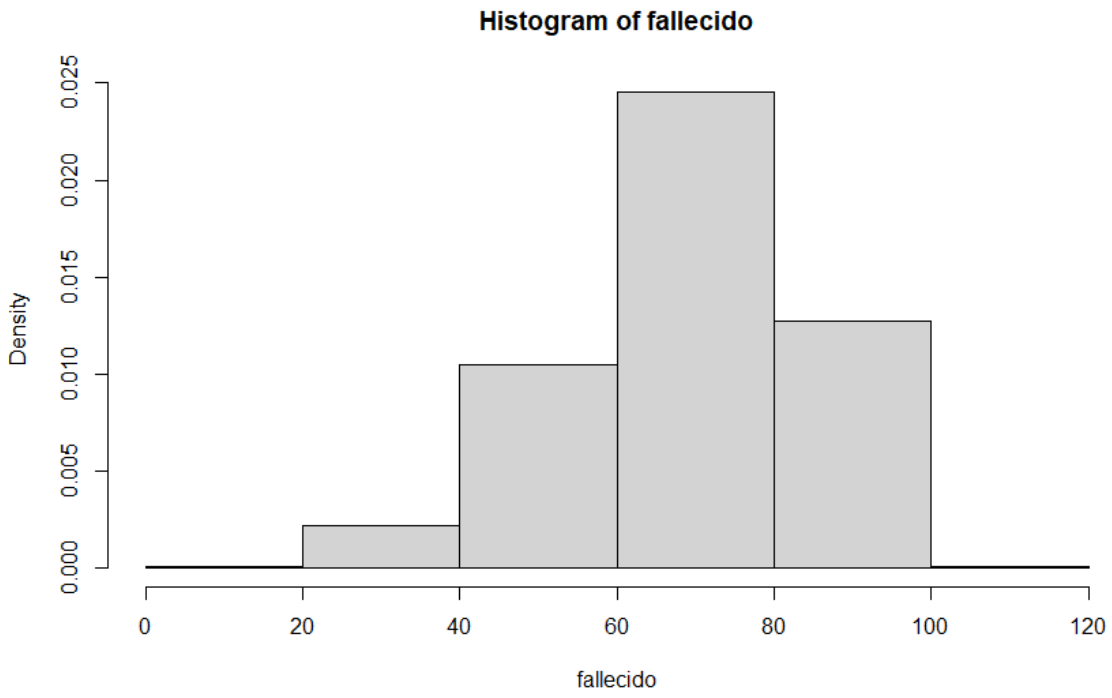


En este histograma podemos observar que aunque los más jóvenes fueron los más afectados en cuanto a contagios, no fue en igual proporción para las muertes, la peor parte la llevaron las personas entre los 60 y 80 años, puesto que como se mencionó en el punto anterior, es muy difícil tener alrededor de los 100 años y por eso no se ven tantas muertes en ese sentido, es claro que los menos afectados en cuanto a muertes fueron las personas entre 0 y 20 años, las personas entre 40 y 60 también sufrieron grandes pérdidas.

d. Realice el histograma de densidades con las edades de las personas fallecidas en Medellín, usando la función hist de R. Compare sus resultados con los de dicha función y explique las diferencias (si las hay).

Por último se tiene el histograma de densidades para las edades de los fallecidos en la ciudad de Medellín, debido a que ya se tenían filtrados los datos se pudo realizar este punto con una sola línea de código.

```
137  hist(fallecido,freq=FALSE,breaks=clases)
```



La diferencia que se tiene respecto al histograma anterior es que el anterior no muestra las muertes entre las personas de 100 y 120 años, esto es debido a que es un numero insignificante para el programa, habría que hacer un análisis mas profundo de los datos para obtener el mismo resultado. Sin embargo la gráfica a simple vista luce igual y los valores están de ambos ejes y están similares.