Estadística Descriptiva

true

9 de julio de 2020

Contents

Para regresar a Exploratorio R 

Nota: Puede descargar este ejercición haciendo clic aquí

Descripción

El presente documento se encuentra dividido en dos partes. La primera parte es un caso de ejemplo que muestra los diferentes procedimientos para realizar un análisis estadístico descriptivo. La segunda parte es el ejercicio que deberá desarrollar como parte de la tarea.

Caso de ejemplo

Para el caso de ejemplo usaremos datos del Ministerio de Salud Pública sobre los casos de amebiasis en la provincia de Loja. Cargaremos estos datos a R y realizaremos un estudio descriptivo. Con el fin del ejercicio vamos a realizar una descripción general de todos los datos y los evaluaremos separadamente entre hombres y mujeres.

Cargar los datos desde la web

Vamos a cargar los datos desde la web, para esto usaremos la función `read.csv()`esta función nos permitirá leer los datos directamente desde el repositorio web (Github). Otra forma es descargar los datos aquí. Haga clic en *View raw* y descargue los datos. Una vez descargados guárdelo en la carpeta y puede usar la función `read_excel()` del paquete **`readxl`**.

```
#Desde la web

amebW <- read.csv("https://github.com/Ciespinosa/datos_practicas/blob/master/AMEBIASIS_LOJA.csv")

##desde su computador
# install.packages("readxl") #Este código solo debe correrlo la primera vez si no
                             #ha sido corrido antes. Para ejecutar borre el # del principio

library(readxl)
ameb <- read_excel("AMEBIASIS_LOJA.xlsx")
```

Para poder asegurar que nuestras matrices están bien vamos a revisar los datos. Usaremos la función `str()`, esta función nos permite ver cuáles son las características de las variables. Por ejemplo, *Canton* es un caracter "chr" y *Edad* es numérico "num"

```
str(ameb)
```

```
## tibble [3,019 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Cantón      : chr [1:3019] "LOJA" "LOJA" "LOJA" "LOJA" ...
##  $ Distrito    : chr [1:3019] "11D01" "11D01" "11D01" "11D01" ...
##  $ Sexo        : chr [1:3019] "Hombre" "Hombre" "Hombre" "Hombre" ...
##  $ Edad en años: num [1:3019] 1 13 14 2 2 22 3 30 36 4 ...
##  $ Consultas   : num [1:3019] 1 2 1 1 1 1 2 1 1 1 ...
##  $ Parroquia   : chr [1:3019] "CHUQUIRIBAMBA" "CHUQUIRIBAMBA" "CHUQUIRIBAMBA" "CHUQUIRIBAMBA" ...
```

Como podemos ver nuestra matriz tiene variables cualitativas nominales (canton, parroquia, sexo) y cuantitativas discretas (número de consultas, Edad). Vamos a generar una par de variables que nos permitan tener un mayor tipo de variables.

Vamos a generar una variable cualitativa nominal con los datos de edad, generaremos una escala de edad de joven hasta adulto. Usaremos la función `cut()`, esta función nos permite cortar un vector numérico y convertirlo en caracteres ordinales. Primero veremos los valores máximo y mínimo de edad para realizar los cortes, usaremos las funciones `min() y max()` que devuelven el valor máximo y mínimo.

> Nota: es mejor que la tabla de datos no tenga caracteres latinos como la tilde o ñ. R y muchos otros softwares suele tener problemas con estos caracteres. En la generación de las variables evitaremos poner tildes.

```
min(ameb$`Edad en años`); max(ameb$`Edad en años`) # el punto y coma ";" nos permite poner en
```

```
## [1] 0
```

```
## [1] 117
```

```
                                                    # la misma línea de código dos órdenes
ameb$C.edad <- cut(ameb$`Edad en años`, breaks=c(-0.1, 16, 25, 40,60, 117),
                   labels = c("infantes", "jovenes", "adultos", "maduros", "adultos mayores"))
```

Para terminar, puesto que no tenemos variables continuas generaremos una variable de temperatura y la incluiremos en nuestra matriz. Usaremos la función `rnorm()`para generar una variable de temperatura de media 38 y desviación de 2.5.

```
ameb$temp <- rnorm(nrow(ameb), 38, 2.5)
```

Muy bien ahora tenemos los datos listos para poder realizar un análisis descriptivo de nuestros datos.

Análisis descriptivo

Variables nominales

Las variables nominales normalmente sirven para separar grupos, estas variables al no tener orden no son factibles de realizar un análisis numérico. Lo que podemos hacer con estas variables es graficar con el fin de comprender por ejemplo la frecuencia de cada carácter o la proporción de cada carácter.

Una variable

Usaremos los datos de Amebiasis con el fin de evaluar la frecuencia de los casos por cantón. Usaremos la función `table()` que nos permite transformar los caracteres en una tabla de frecuencias y la función `prop.table()` para transformar los datos de la tabla a proporciones.

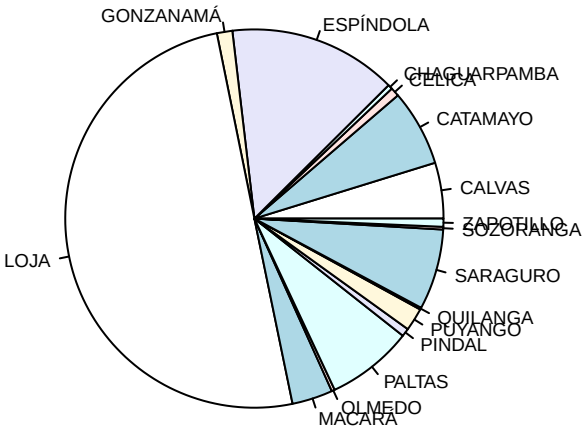
```
canT <- table(ameb$Cantón)
canPT <- prop.table(canT)

canT; canPT
```

```
##
##       CALVAS     CATAMAYO       CELICA CHAGUARPAMBA    ESPÍNDOLA    GONZANAMÁ
##          144          197           23           12          435           40
##         LOJA       MACARÁ       OLMEDO       PALTAS       PINDAL      PUYANGO
##         1511          105            8          223           21           59
##     QUILANGA     SARAGURO    SOZORANGA    ZAPOTILLO
##            5          208            7           21
```

```
##
##       CALVAS     CATAMAYO       CELICA CHAGUARPAMBA    ESPÍNDOLA    GONZANAMÁ
##  0.047697913  0.065253395  0.007618417  0.003974826  0.144087446  0.013249420
##         LOJA       MACARÁ       OLMEDO       PALTAS       PINDAL      PUYANGO
##  0.500496853  0.034779728  0.002649884  0.073865518  0.006955946  0.019542895
##     QUILANGA     SARAGURO    SOZORANGA    ZAPOTILLO
##  0.001656178  0.068896986  0.002318649  0.006955946
```

Aunque podemos ver como se distribuyen los datos entre los cantones no son claras las diferencias entre los cantones. Un gráfico que es muy utilizado es el gráfico de pastel.

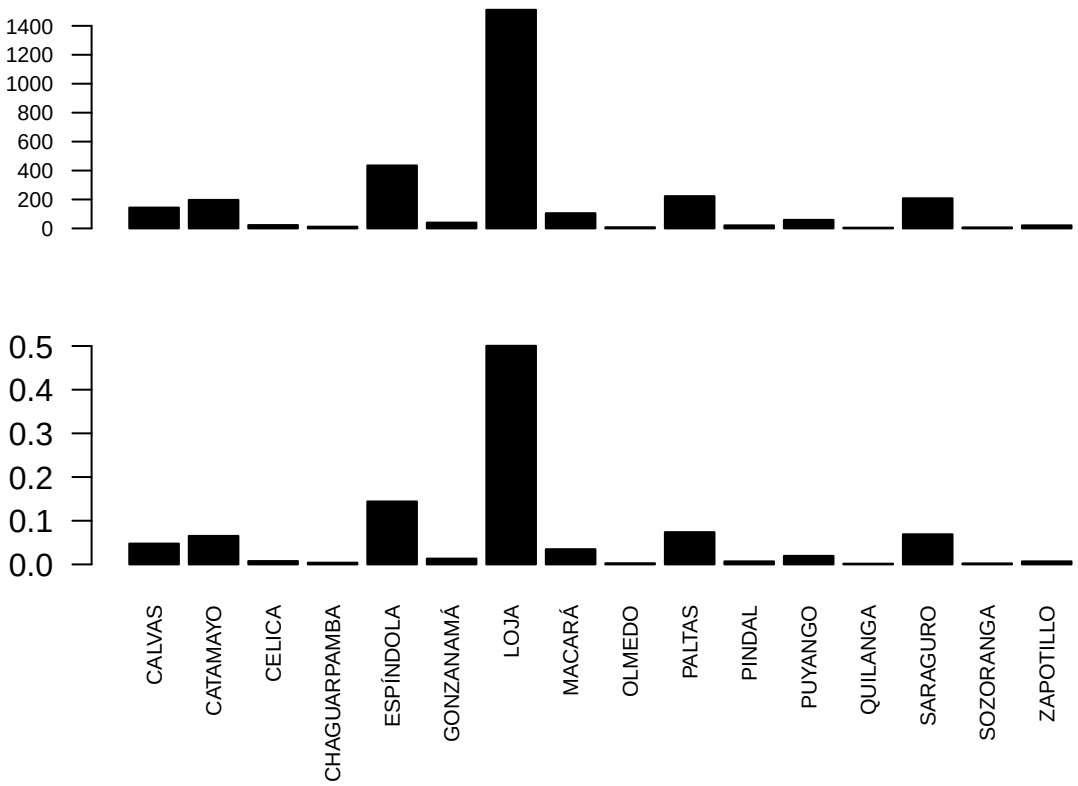
```
pie(canT, cex = 0.6)
```



Aunque inicialmente podemos ver diferencias entre los cantones con mayor número, los seres humanos somos malos para estimar los tamaños relativos de los ángulos presentados de esta forma, por lo que los gráficos de pastel no son una buena forma de presentar datos.

Una mejor estrategia es hacer una gráfica de barras, esta nos permita destacar las diferencias entre los cantones y nos es más fácil comprender visualmente. Si te interesa conocer profundizar algo más en las gráficas en R puedes ir al enlace.

```
par(mfcol=c(2,1), mar=c(2,3,1,1), oma=c(4,1,1,1))
barplot(canT, col="black", names=FALSE, las=2, cex.axis = 0.7)
barplot(canPT, col="black", cex.names=0.7, las =2)
```



Vemos que en el cantón Loja hay más casos de Amebiasis y representa alrededor del 50% de todos los casos de amebiasis. Sin embargo, hay que tomar estos datos con cuidado, ya que un mayor número de casos no necesariamente implica una mayor incidencia de la enfermedad, puesto que Loja también es el cantón más poblado. En todo caso, con la gráfica podemos ver como se reparten nuestros casos en la Provincia de Loja.

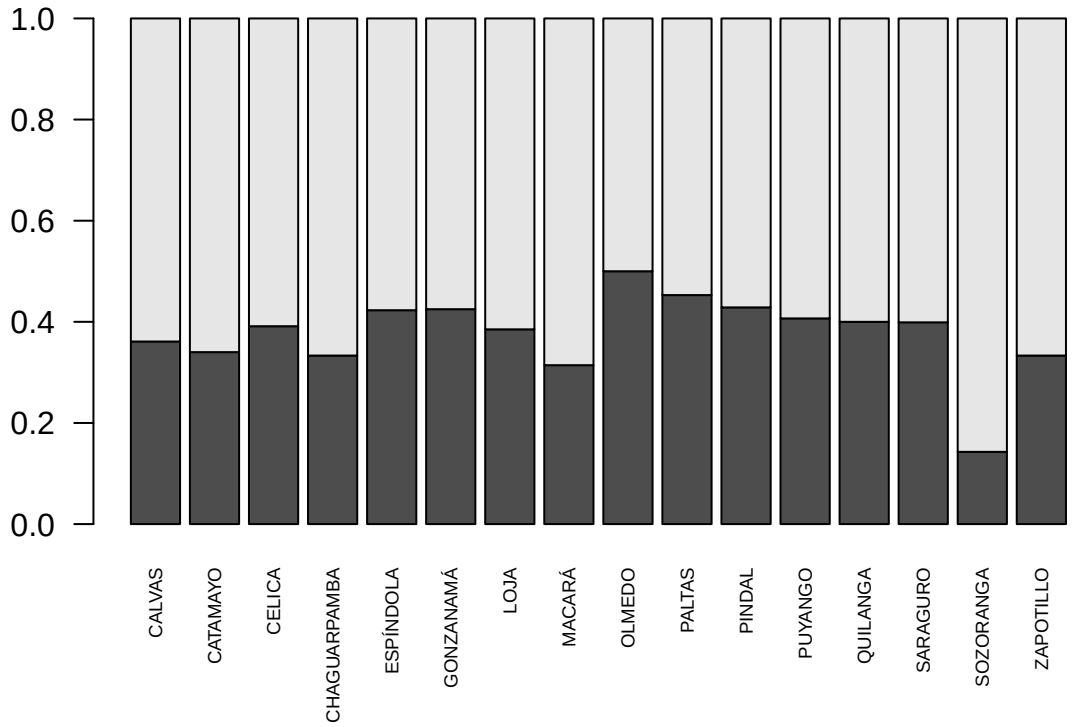
Actividad:
Revisa los datos de el tamaño poblacional de cada uno de los cantones y calcula la incidencia de amebiasis en Loja. Haz un gráfico con los datos de incidencia y analiza si el resultado cambia. ¿Cuál es ahora el cantón con mayor y menor incidencia de amebiasis?

Dos variables

Cuando desarrollamos una descripción de las frecuencias de una variable ordinal nos permite comprender como esa característica aparece en mi muestra, sin embargo, muchas veces nos interesa entender si esa frecuencia se mantiene o cambia en función de otra variable ordinal.

En nuestro ejemplo, vimos como la frecuencia de casos de Amebiasis cambia entre cantones, pero, ¿esta frecuencia es similar en hombres y mujeres?. Bueno, vamos a responder esta pregunta usando la misma función **table**, pero esta vez usando las dos variables.

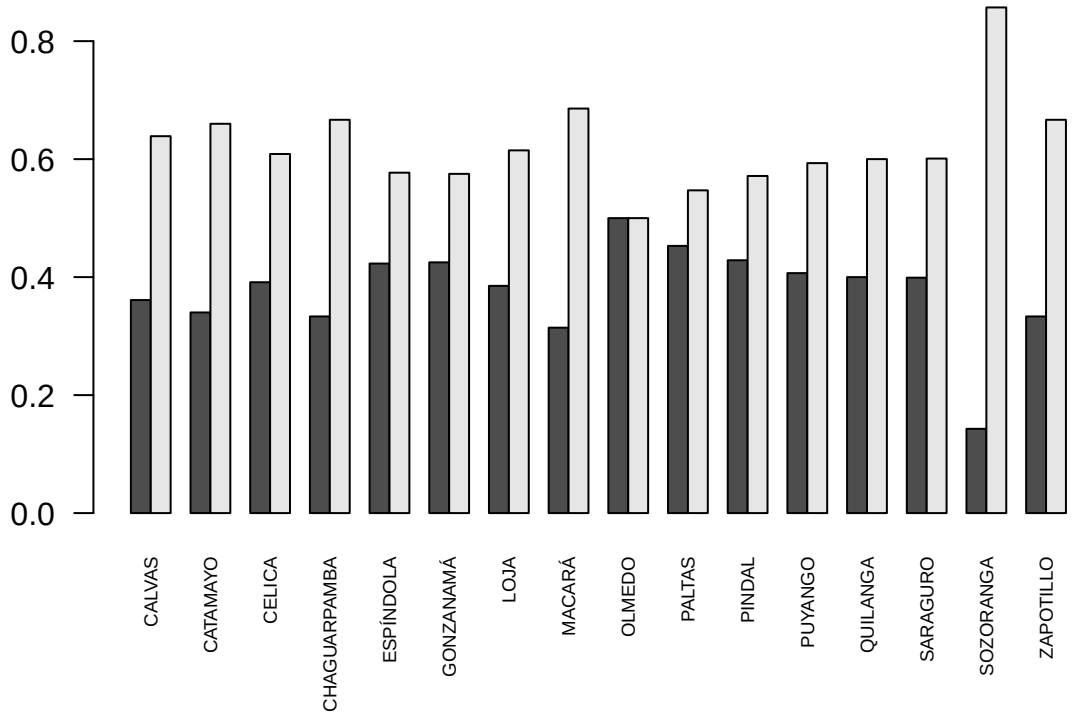
```
tGC <- table(ameb$Sexo, ameb$Cantón)

ptGC <- prop.table(tGC, 2) # El número 2 le dice que calcule la proporción por columnas

barplot(ptGC, cex.names = 0.55, las=2) # las cambia la dirección del texto
```




```
                                    # cex.names cambia el tamaño de letra
```

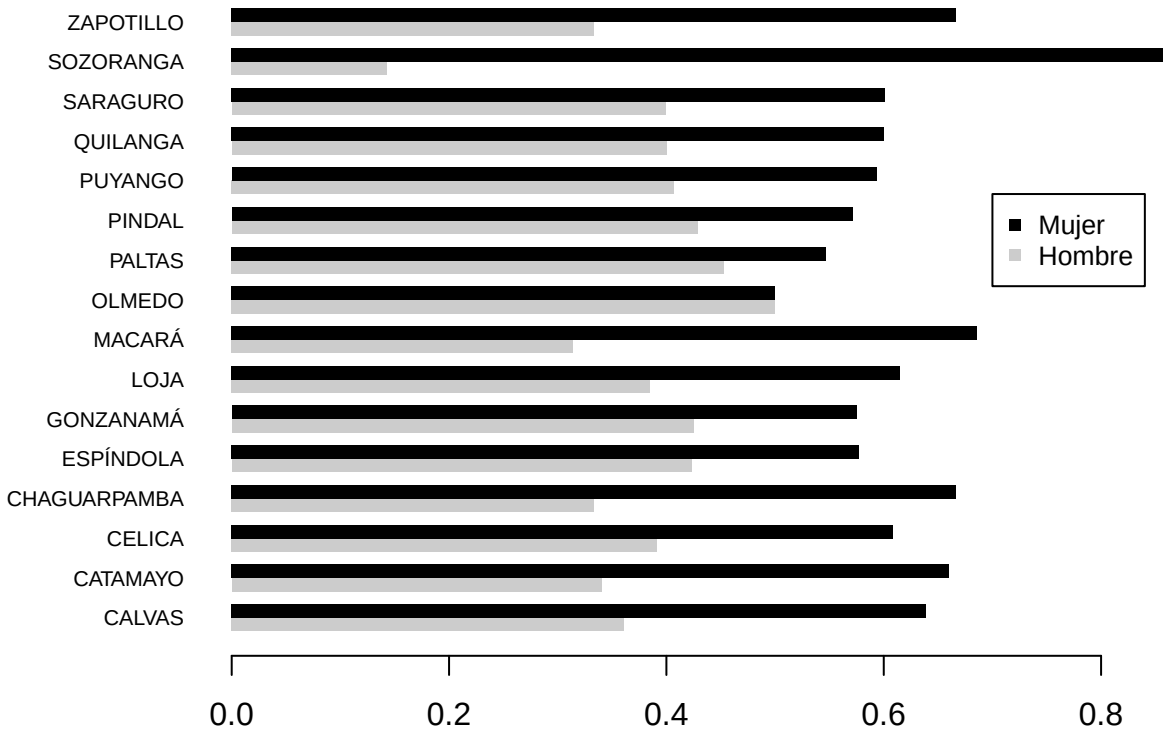
Aunque hemos logrado tener nuestra gráfica, me parece que no es del todo claro, la superposición de las columnas no nos permite apreciar adecuadamente el patron que queremos rescatar. Podemos cambiar esto con un argumento sencillo en nuestra función **barplot()**.

```
barplot(ptGC, beside = T, cex.names = 0.55, las = 2)
```



Algunas veces podríamos estar interesados en mostrar este gráfico de forma horizontal, nuevamente un pequeño argumento nos ayuda a cambiar el gráfico.

```
par(mar = c(3,6,2,2))
barplot(ptGC, beside = T, cex.names = 0.7, las = 1, horiz = T,
        col = c("grey80", "black"), border = NA )
legend(x = 0.7, y = 34, legend = c("Mujer", "Hombre"),
       pch = 15, col = c("black", "grey80"), cex = 0.8)
```



Actividad:
Modifica esta gráfico pero esta vez puedes realizarlo los datos de incidencia, analiza si el resultado cambia. ¿Cuál es ahora el cantón con mayor y menor incidencia de amebiasis?. Con los datos de incidencia obten las proporciones no por columnas como lo hicimos, sino por filas. Puedes graficar esta nueva tabla. ¿Qué información te brinda el primer gráfico y este segundo gráfico?

Variables ordinales

Las variables ordinales son caracteres que presentan orden, así en nuestro ejemplo las personas *adultas* tienen mayor edad que los *jovenes* aunque no sepamos en que magnitud. Con el fin de describir estas variables usaremos medidas de posición, como son los cuantiles en sus distintas versiones, y como medida de dispersión al recorrido intercuartílico.

Para poder usar análisis numéricos de las variables ordinales, debemos estar seguros que estas son categorías y que están ordenadas de forma correcta. Podemos preguntar a R el tipo de vector que tenemos usando la función `class()`, el orden podemos verlo usando la función `levels`. Si nuestros datos no son factores podemos convertirlos en factores con la función `factors`. Veamos nuestros datos.

```
class(ameb$C.edad) #En nuestro caso es un factor
```

```
## [1] "factor"
```

```r
levels(ameb$C.edad) #En nuestro caso el orden es correcto.
```

```
## [1] "infantes"        "jovenes"         "adultos"         "maduros"
## [5] "adultos mayores"
```

Ahora, con el fin de que ustedes puedan realizar el la conversión de los datos, si nos son factores, voy a pedirle a R que transforme a caracteres y luego transformar en factor.

```r
ameb$C.edad <- as.character(ameb$C.edad)#convertimos en carácter
class(ameb$C.edad) #ahora ya no es un factor
```

```
## [1] "character"
```

```r
ameb$C.edad <- factor(ameb$C.edad)#lo reconvertimos en factor
levels(ameb$C.edad) #Comprobamos el orden. En este caso no está en el orden que queremos
```

```
## [1] "adultos"         "adultos mayores" "infantes"        "jovenes"
## [5] "maduros"
```

```r
##cambiamos el orden
ameb$C.edad <- factor(ameb$C.edad,
                      levels = c("infantes", "jovenes", "adultos", "maduros", "adultos mayores") )
levels(ameb$C.edad) #ahora el orden es correcto
```

```
## [1] "infantes"        "jovenes"         "adultos"         "maduros"
## [5] "adultos mayores"
```

Una vez que tenemos listos los datos podemos calcular los cuantiles de nuestra variable y el recorrido intercuartílico. Para calcular los cuantiles usaremos la función `quantile()`, esta función necesita valores numéricos, como tenemos factores podemos convertir los factores en números con la función `as.numeric()`.

```r
quantile(as.numeric(ameb$C.edad))
```

```
##   0%  25%  50%  75% 100%
##    1    1    2    4    5
```

Usamos los cuantiles como una medida de posición, así los resultados nos muestran que la mediana de nuestra variable ordinal se encuentra en la categoría de edad 2 (*jovenes*). Ahora nos interesa calcular el recorrido intercuatílico y recorrido intercuatílico relativo como medida de dispersión, definido como el cociente entre la diferencia de los cuartiles tercero y primero, y dividido para la mediana el Q2 en el caso del relativo.

$$RI = Q3 - Q1$$

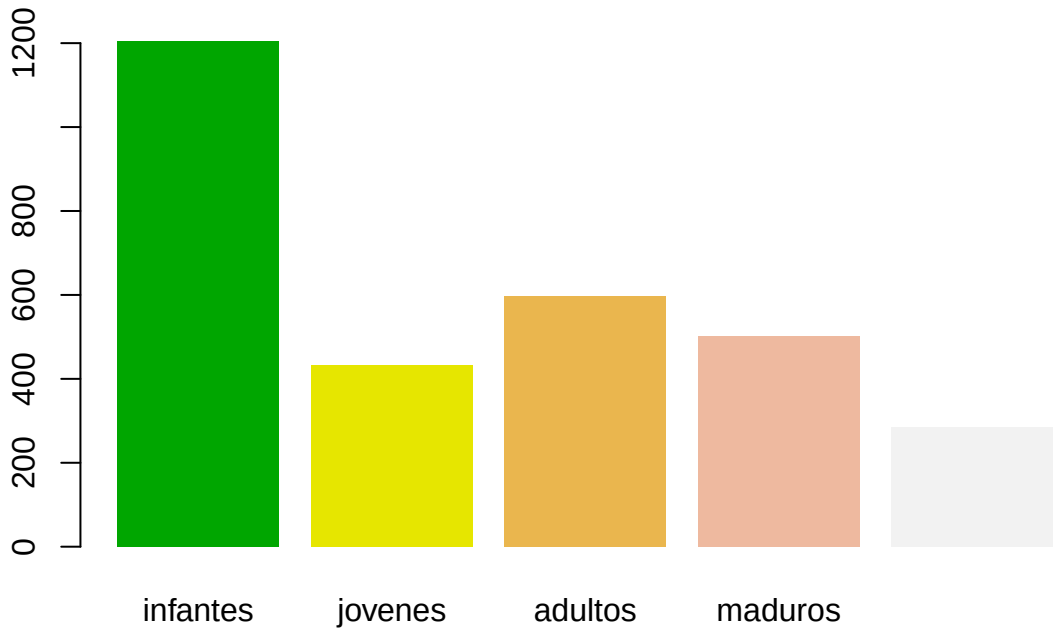
$$RIR = \frac{Q3 - Q1}{Mediana}$$

```r
RI <- quantile(as.numeric(ameb$C.edad), 0.75) - quantile(as.numeric(ameb$C.edad), 0.25)

RIR <- (quantile(as.numeric(ameb$C.edad), 0.75)- quantile(as.numeric(ameb$C.edad), 0.25))/quantile(as.nu
```

Estas medidas nos permiten conocer que tan dispersos están nuestros datos. Valores más altos de RI o de RIR muestran una mayor dispersión.

Podemos graficar esta variable usando la misma función de antes `barplot()`.

```
barplot(table(ameb$C.edad), col= terrain.colors(5), border=FALSE)
```



Actividad:
Vamos a trabajar un poco con los datos. Imaginen que nos interesa conocer si los rangos de edad son diferentes entre dos cantones, usaremos las medidas de posición y dispersión para comparar lo que sucede en cada cantón. Para el ejercicio usaremos los cantones de Loja y Esíndola. Una vez que calcule las medidas grafique los datos de los dos cantones. Puede usar la función subset para extraer los datos de cada cantón de la siguiente forma: loja <- subset(ameb, Cantón==“Loja”) ¿Qué conclusión puede sacar en torno a la incidencia por edades en estos dos cantones?

Variables numéricas

En el caso de las variables numéricas sean estas continuas o discretas usaremos las medidas de tendencia central más usadas como la media y moda, y medidas de dispersión como la desviación estándar y la varianza. Usaremos las funciones `mean()` y `sd()` que miden la media y la desviación estándar de una variable.

```
mediaE <- mean(ameb$`Edad en años`)
modaE <- as.numeric(names(which.max(table(ameb$`Edad en años`))))
```

```r
mediaT <- mean(ameb$temp)


mediaE; mediaT; modaE
```

```
## [1] 27.58496
```

```
## [1] 38.01183
```

```
## [1] 5
```

La media de edad es de 27.6 años y la de temperatura es de 38.5 °C.

```r
sdE <- sd(ameb$`Edad en años`)
sdT <- sd(ameb$temp)

sdE; sdT
```

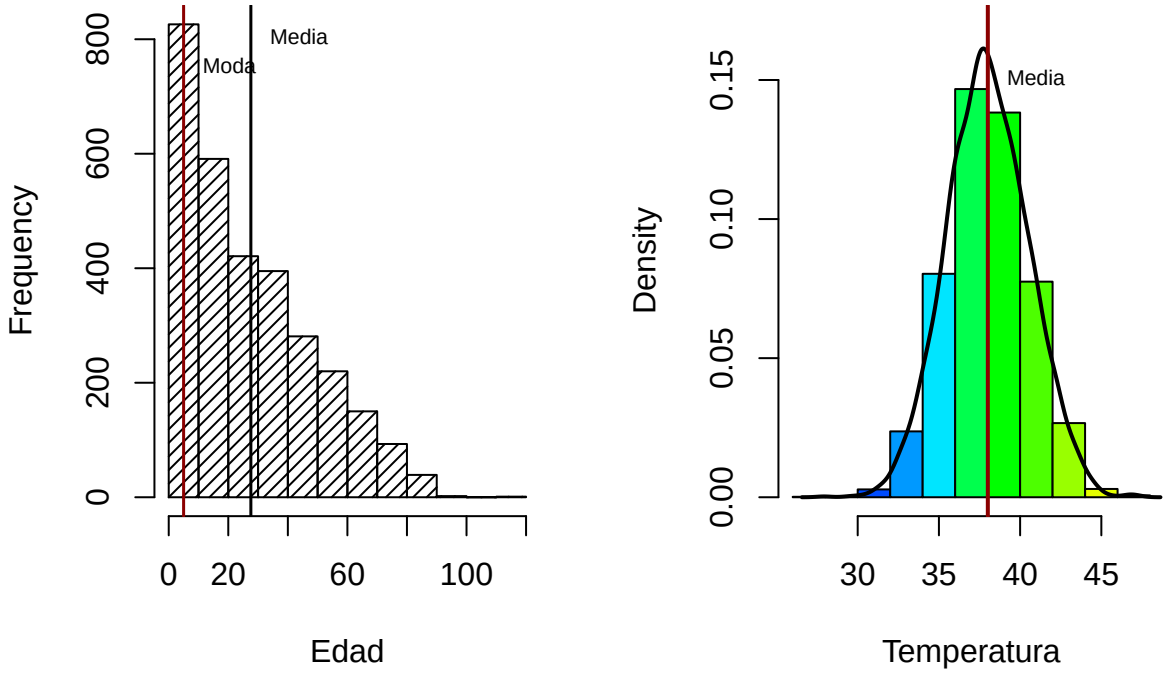
```
## [1] 21.27373
```

```
## [1] 2.508999
```

La desviación estándar es de 21.3 años y el de temperatura es de 2.5 °C.

Ahora podemos graficar las variables. Podemos ver el tipo de distribución de los datos usando la función `hist()` y además la función `density()` para graficar una curva.

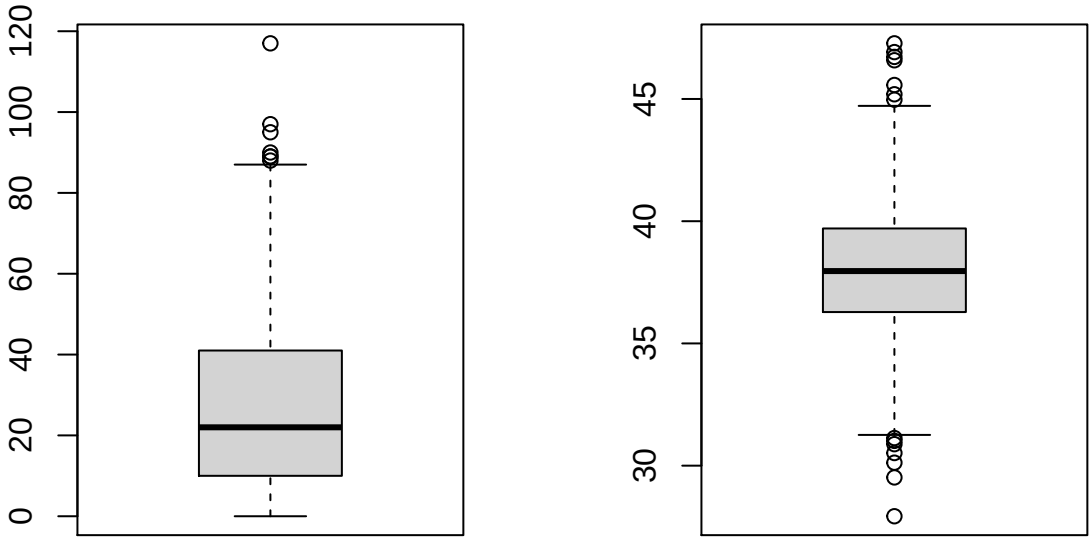
```r
par(mfcol=c(1,2))
hist(ameb$`Edad en años`, col="black", density = 20, xlab="Edad", main="")
abline(v=modaE, col="darkred", lwd=1.5)
abline(v=mediaE, lwd=1.5)
text(mediaE, 800, "Media", pos = 4, cex = 0.7)
text(modaE, 750, "Moda", pos = 4, cex = 0.7)

hist(ameb$temp, prob = TRUE, col = topo.colors(15), main="", xlab="Temperatura", ylim=c(0,0.17))
lines(density(ameb$temp), lwd=2)
abline(v=mediaT, lwd=2, col="darkred")
text(mediaT, 0.15, "Media", pos=4, cex=0.7)
```



Como podemos ver muchas veces en las variables discretas la media no es el mejor valor de centralidad, el problema que tenemos es que las variables discretas normalmente no responden a una distribución de campana (*distribución normal*) por lo que es mejor usar una medida como la moda.

Finalmente, podemos graficar la variable como un grafico de cajas y bigotes (boxplot) para ello usaremos la función `boxplot()`, esta gráfica nos ayuda a ver la distribución de los datos, nos marca los cuantiles y los puntos que están como outlayers (datos extremos dentro de mi variable)

```
par(mfcol=c(1,2))
boxplot(ameb$`Edad en años`)
boxplot(ameb$temp)
```



Como podemos ver en estas gráficas, nuestra variable discreta no es uniforme, tenemos más datos en edades menores de 20 años. En el caso de la temperatura vemos que es una variable muy simétrica, el tamaño de las cajas y los bigotes es bastante parecido por arriba y abajo de la mediana (línea negra en el centro de la caja).

Comparando las medidas descriptivas.

En este punto queremos comparar los parámetros descriptivos que calculamos pero esta vez dividiremos los datos entre hombres y mujeres. Usaremos la función `subset()` para dividir los datos entre hombres y mujeres.

```
amebH <- subset(ameb, ameb$Sexo=="Hombre")
amebM <- subset(ameb, ameb$Sexo=="Mujer")
```

Medidas de tendencia central

Ahora calculemos las medidas para las variables numéricas. La moda para los años y la media para la temperatura.

```
as.numeric(names(which.max(table(amebH$`Edad en años`)))); as.numeric(names(which.max(table(amebM$`Edad
```

```
## [1] 3
```

```
## [1] 5
```

```
mean(amebH$temp); mean(amebM$temp)
```

```
## [1] 37.93743
```

```
## [1] 38.05951
```

Como podemos ver los datos de es diferente entre los dos grupos, la moda en las mujeres es 5 años, mientras que en los hombres es 3 años. La temperatura no difiere entre los dos grupos.

Finalmente, usaremos el diagrama de cajas para poder observar las diferencias en la dispersión entre la edad de los dos grupos de personas.

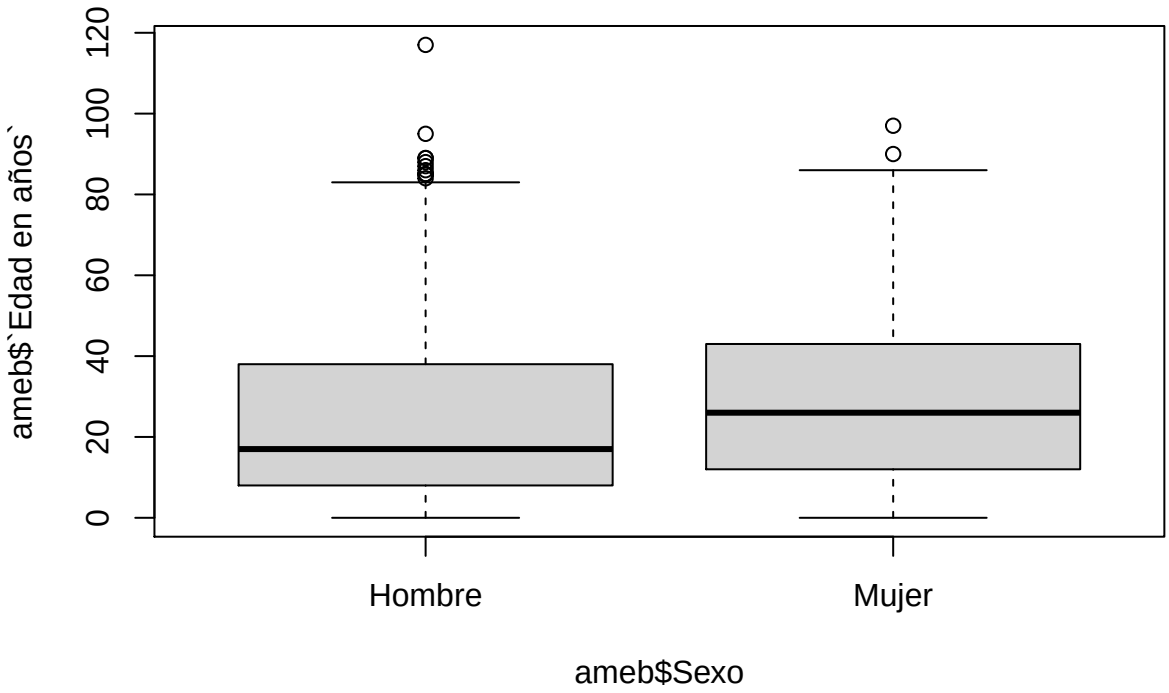
```
sd(amebH$`Edad en años`)
```

```
## [1] 21.88161
```

```
sd(amebM$`Edad en años`)
```

```
## [1] 20.70701
```

```
boxplot(ameb$`Edad en años`~ameb$Sexo)
```



Como vemos la desviación estandar es 1 año mayor en hombres que en las mujeres, eso significa que este grupo muestra una mayor dispersión. Además vemos que en el caso de las mujeres la distribución es un tanto más simétrica.

Ejercicio

A continuación espero que realicen una evaluación de las características de las variables de datos de un estudio sobre cirrosis hepática (Counting Processes and Survival Analysis by T. Fleming & D. Harrington, (1991). Estudio sobre cirrocis biliar primaria (PBC), published by John Wiley & Sons.). Pueden descargar el archivo de Excel haciendo clic aquí

Descargue la tabla, revise la hoja dos donde se encuentran los metadatos de las variables, una explicación de que se refieren las variables que se encuentran en la tabla.

Una vez este claras cada una de las variables vamos a realizar un análisis de algunas de ellas y las vamos a comparar entre tratamientos. Las variables que analizaremos son; fase, colesterol, albumina y triglicéridos.

Desarrollar las siguientes partes:

1. Las variables cualitativas ordinales obtener la mediana y el Rango Intercuartílico para el tratamiento y placebo. ¿Hay diferencia en la mediana y el rango entre los dos tratamientos, que significa eso?
2. Realizar un gráfico de barras para cada tratamiento con los datos de la variable ordinal.
3. Para las variables continuas. Calcular la media y la desviación estándar para cada tratamiento. ¿Existen diferencias entre tratamientos? ¿Cómo puede interpretar esta información?.
4. Realice un histograma y una línea con la densidad que permita ver la distribución de los datos para cada tratamiento. Incluya la media en cada gráfica.
5. Realice un gráfico de cajas para analizar las diferencias entre cada tratamiento.