Ejercicios tema 2

Juan Andrés Peraira Pérez

Ejercicio 1: Utiliza los datos "iris" que corresponden a mediciones (en centímetros) de 4 variables: largo y ancho de los pétalos y sépalos; para 50 flores de 3 especies distintas de plantas Iris setosa, versicolor, y virginica.

Queremos responder a las siguientes preguntas:

¿Cuántos datos (o casos) tenemos para cada especie? y ¿qué porcentaje representan del total de casos? Realice los gráficos pertinentes para cada tipo de variable (cualitativa vs. cuantitativa).

```
datos<-iris #-- cargamos los datos
head(datos)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
attach(datos) #-- activamos las variables
table(Species)
```

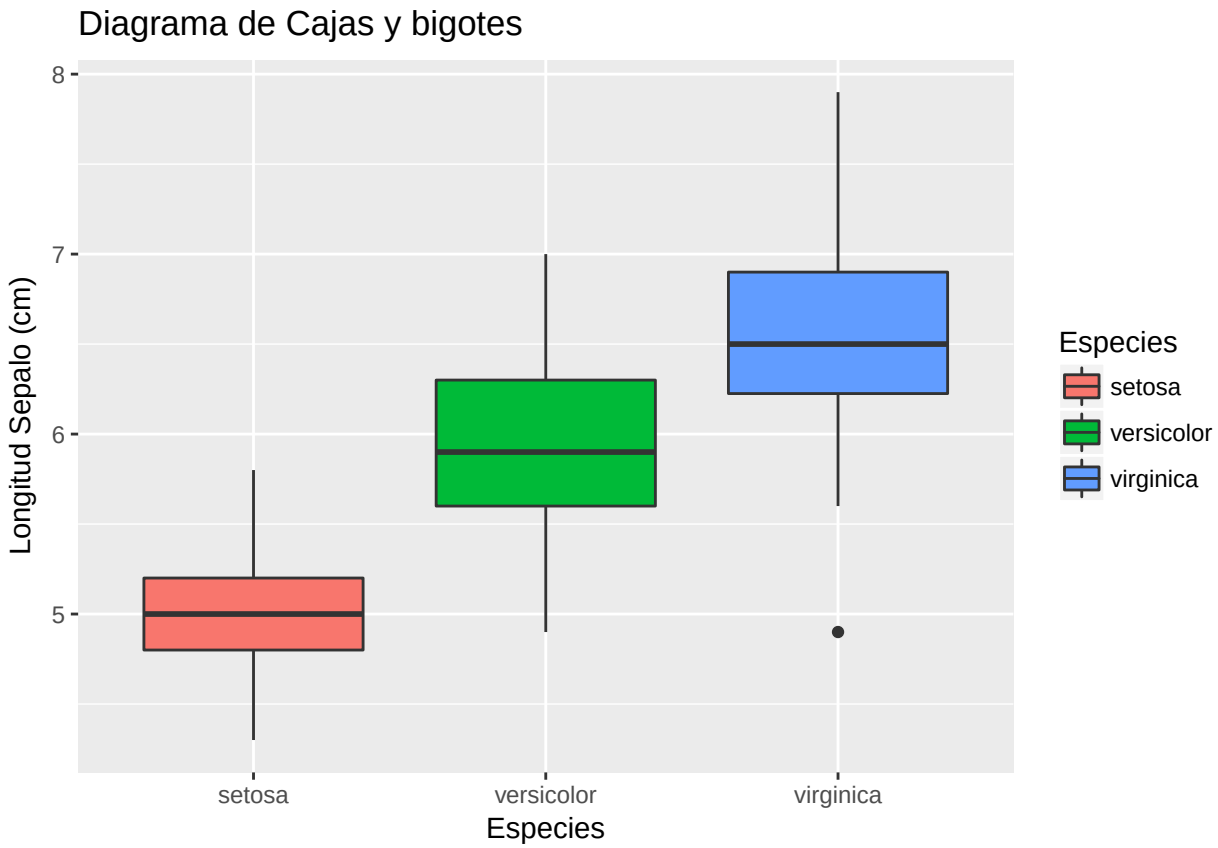
```
## Species
##     setosa versicolor  virginica
##         50         50         50
```

```
#-- Podemos observar que tenemos 50 casos para cada especie
table(Species)/length(Species)
```

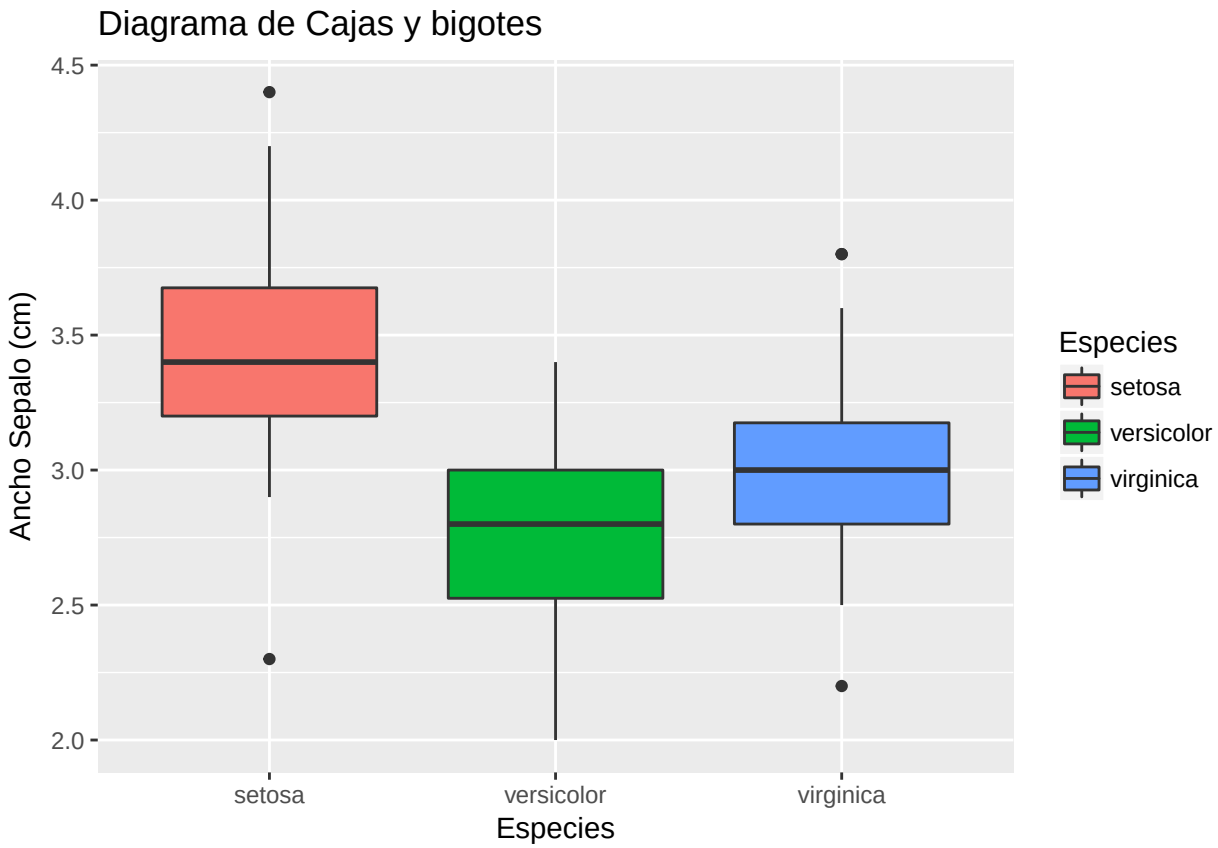
```
## Species
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
```

El porcentaje es de un 33,33% por especie con respecto al total de datos. Comenzaremos con los Gráficos para las variables cuantitativas, en primer lugar se realizaran los diagramas de cajas y bigotes

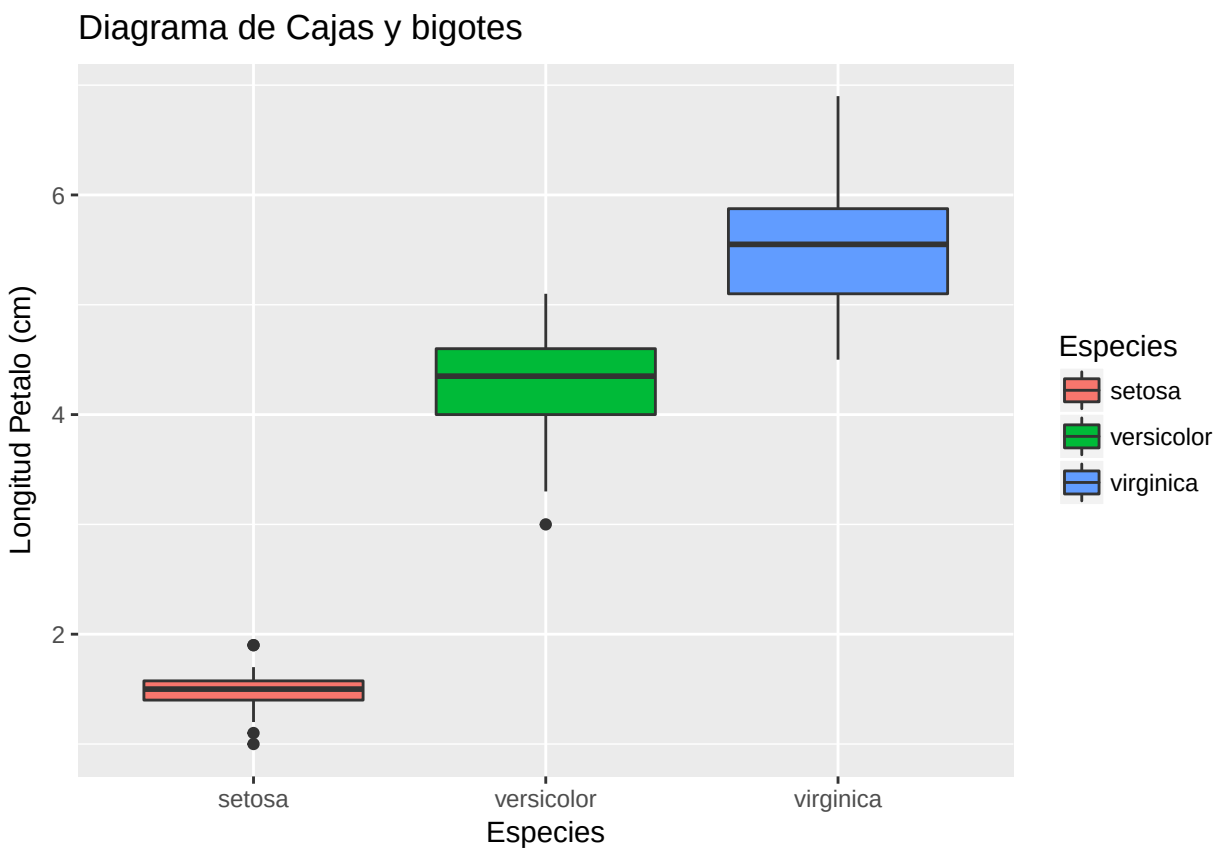
```
library(ggplot2)
p <- ggplot(datos,aes(Species,Sepal.Length))
p +
  labs(x = "Especies",y = "Longitud Sepalo (cm)",
       title="Diagrama de Cajas y bigotes")+
  scale_fill_discrete(guide_legend(title = "Especies"))+
  geom_boxplot(aes(fill=Species))
```



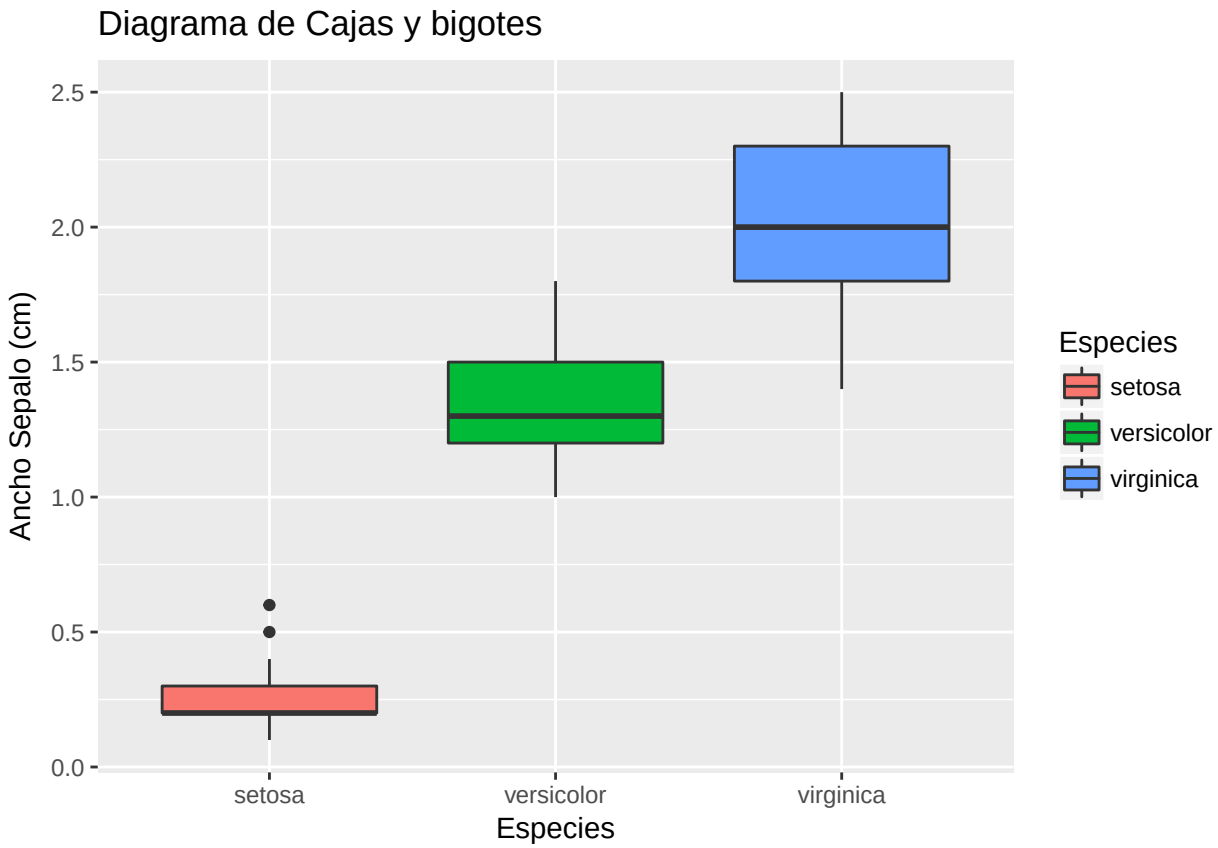

```
p <- ggplot(datos,aes(Species,Sepal.Width))
p +
  labs(x = "Especies",y = "Ancho Sepalo (cm)",
       title="Diagrama de Cajas y bigotes")+
  scale_fill_discrete(guide_legend(title = "Especies"))+
  geom_boxplot(aes(fill=Species))
```




```
p <- ggplot(datos,aes(Species,Petal.Length))
p +
  labs(x = "Especies",y = "Longitud Petalo (cm)",
       title="Diagrama de Cajas y bigotes")+
  scale_fill_discrete(guide_legend(title = "Especies"))+
  geom_boxplot(aes(fill=Species))
```

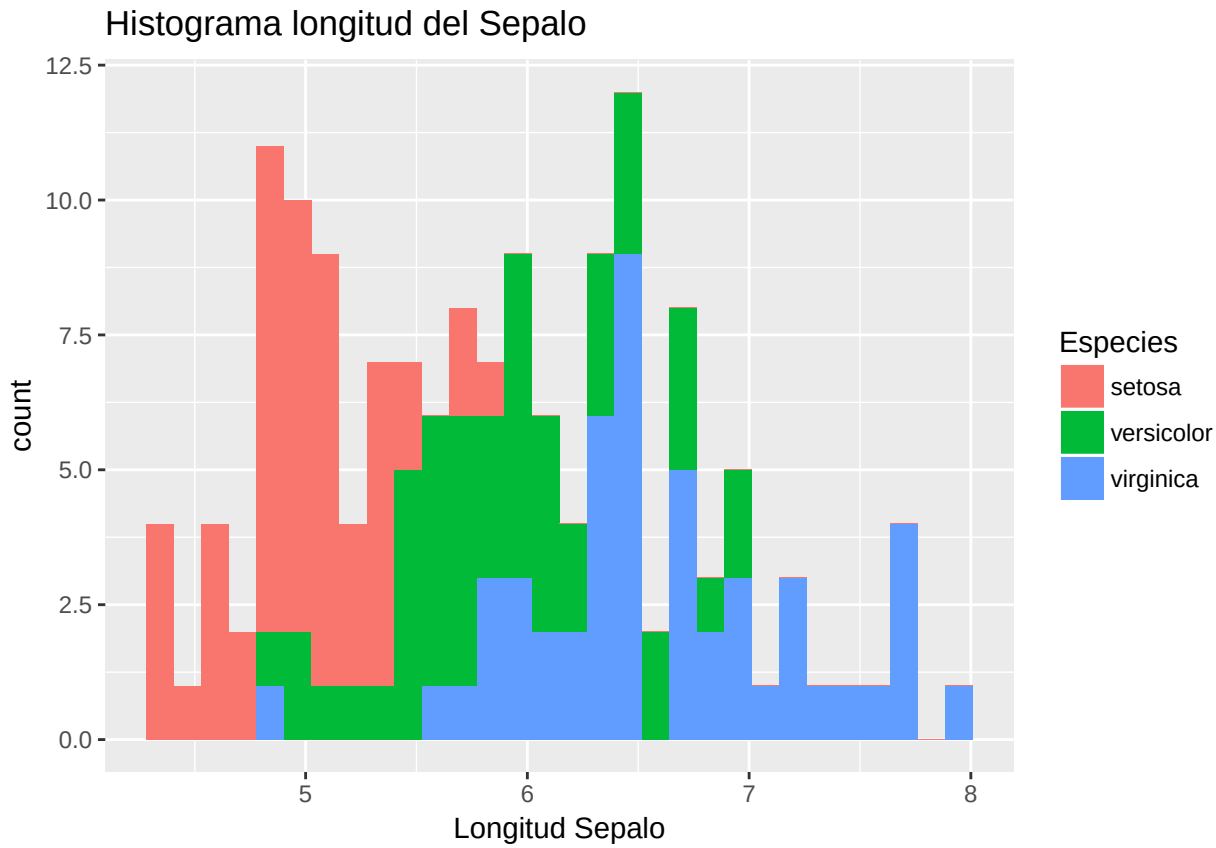



```
p <- ggplot(datos,aes(Species,Petal.Width))
p +
  labs(x = "Especies",y = "Ancho Sepalo (cm)",
       title="Diagrama de Cajas y bigotes")+
  scale_fill_discrete(guide_legend(title = "Especies"))+
  geom_boxplot(aes(fill=Species))
```

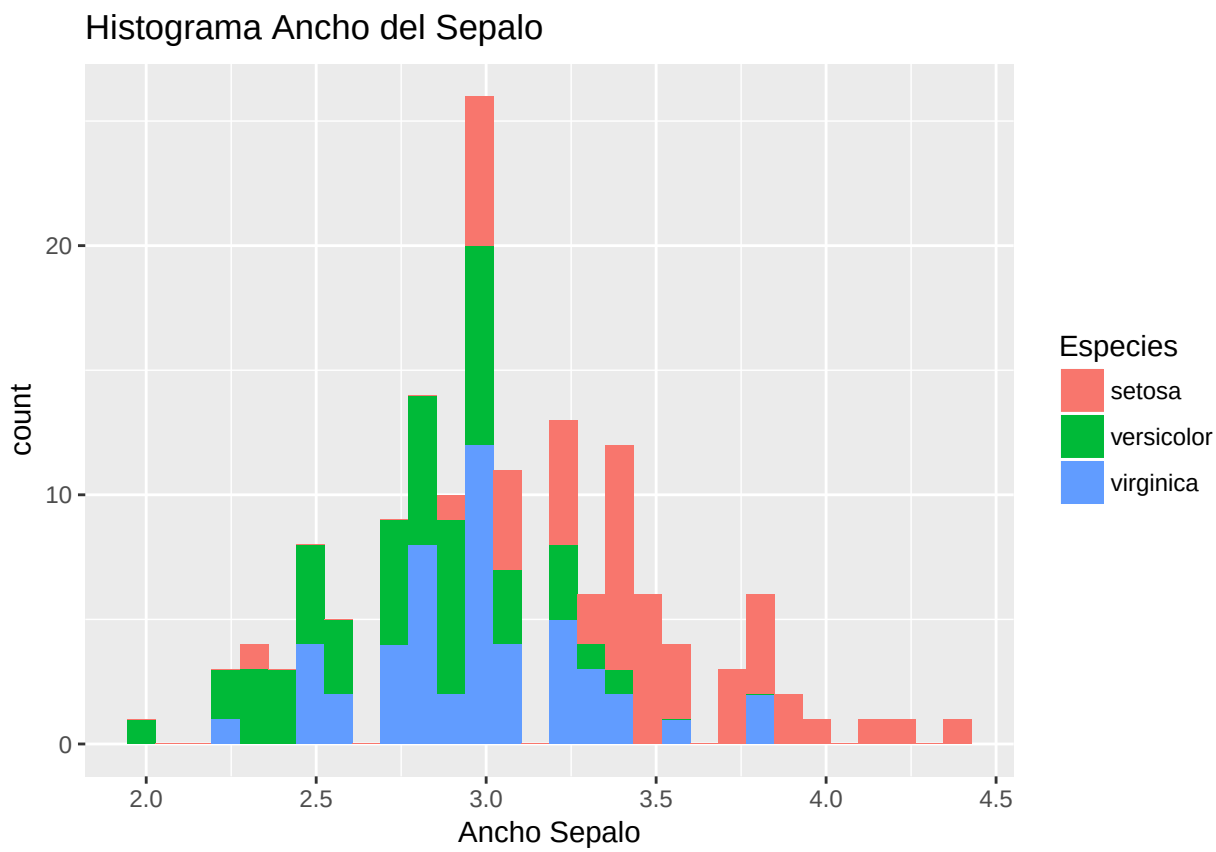


A continuación se presentan los histrogramas para cada variable por especie.

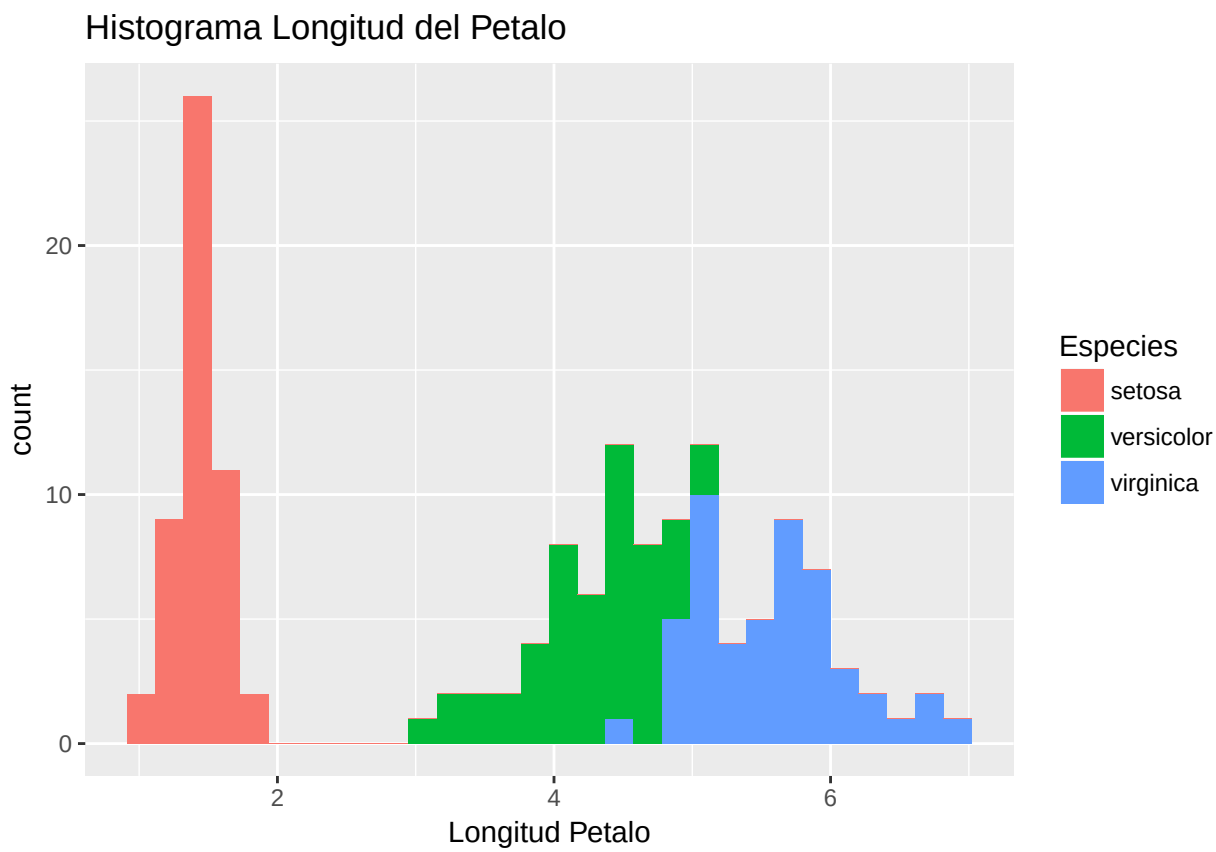
```
ggplot(datos, aes(x = Sepal.Length, fill=Species)) +
  labs(x = "Longitud Sepalo",
       title="Histograma longitud del Sepalo")+
  scale_fill_discrete(guide_legend(title = "Especies"))+
geom_histogram()
```



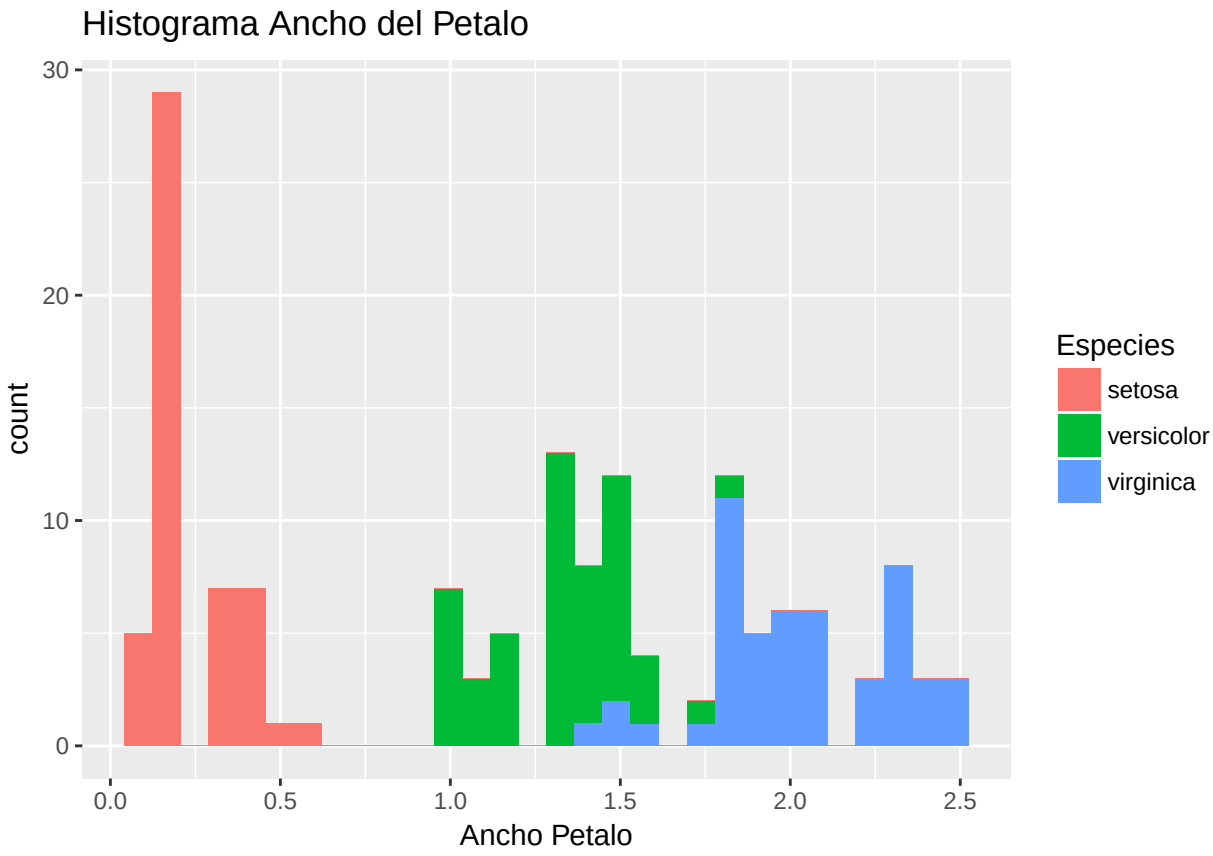

```
ggplot(datos, aes(x = Sepal.Width, fill=Species)) +
  labs(x = "Ancho Sepalo",
       title="Histograma Ancho del Sepalo")+
  scale_fill_discrete(guide_legend(title = "Especies"))+
geom_histogram()
```




```
ggplot(datos, aes(x = Petal.Length, fill=Species)) +
  labs(x = "Longitud Petalo",
       title="Histograma Longitud del Petalo")+
  scale_fill_discrete(guide_legend(title = "Especies"))+
geom_histogram()
```




```
ggplot(datos, (aes(x = Petal.Width, fill=Species))) +
  labs(x = "Ancho Petalo",
       title="Histograma Ancho del Petalo")+
  scale_fill_discrete(guide_legend(title = "Especies"))+
geom_histogram()
```



¿Cuál es la media clásica y robusta del ancho del sépalo para cada especie?. Realice diagrama de cajas.

```
#-- Media clásica del ancho del sépalo para cada especie.
media_ancho_sepalo<-aggregate(Sepal.Width~Species, datos, mean)
media_ancho_sepalo
```

```
##      Species Sepal.Width
## 1     setosa       3.428
## 2 versicolor       2.770
## 3  virginica       2.974
```

```
#-- Media robusta del ancho del sépalo para cad especie.
library(WRS2)
media_robusta_ancho_sepalo<-aggregate(Sepal.Width~Species, datos, mest)
media_robusta_ancho_sepalo
```

```
##      Species Sepal.Width
## 1     setosa    3.418994
## 2 versicolor    2.782828
## 3  virginica    2.962500
```

¿Qué correlaciones existen entre las distintas medidas tomadas a cada planta? Realice un gráfico bidimensional para observarlo.

```
datos$Species = NULL
correlacion <-cor(datos)
correlacion
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
```

```
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Podemos observar que existe una alta correlación entre la longitud del sepalo con la longitud y ancho del petalo y la longitud del petalo con el ancho del petalo.

Ejercicio 2: Utiliza los datos "Davis" (paquete "car") para calcular el IMC como se indicó en el tema 2 (IMC=Peso/Estatura^2). Realia:

Gráfico de barras y de sectores para las categorías del IMC por sexo.

```
#-- cargamos los datos
library(car)
datos<-Davis
head(datos)
```

```
##   sex weight height repwt repht
## 1   M     77    182    77   180
## 2   F     58    161    51   159
## 3   F     53    161    54   158
## 4   M     68    177    70   175
## 5   F     59    157    59   155
## 6   M     76    170    76   165
```

```
attach(datos) #-- Activamos las variables

#-- Creamos la función
imc=function(w,h){w/(h/100)^2}
#-- Calculamos el imc para los datos
datos_imc<-imc(datos$weight,datos$height)
# creamos las categorías de IMC
imcc_datos=cut(datos_imc, breaks=c(0, 15, 18.5, 25, 30))
#-- Creamos la tabla
imcfrec=table(imcc_datos)
cbind(imcfrec)
```

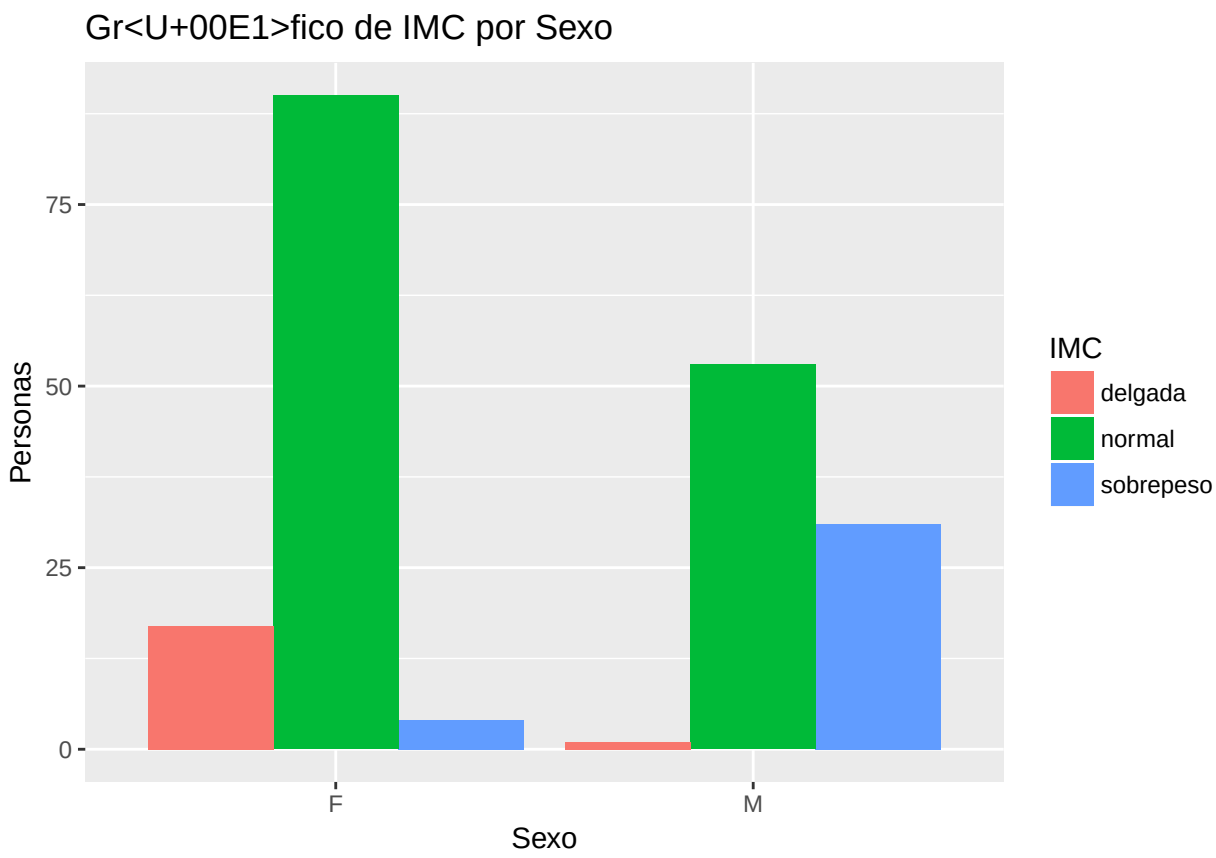
```
##           imcfrec
## (0,15]          0
## (15,18.5]      18
## (18.5,25]     143
## (25,30]        35
```

```
#-- agregamos las etiquetas
levels(imcc_datos)=c("infrapeso","delgada","normal","sobrepeso","obesidad")
#-- Creamos el data frame
datos_davis<-data.frame(datos, IMC=datos_imc, IMCc=imcc_datos)
# eliminamos los errores
datos_davis<-datos_davis[!is.na(datos_davis$IMCc),]
head(datos_davis)
```

```
##   sex weight height repwt repht      IMC      IMCc
## 1   M     77    182    77   180 23.24598    normal
## 2   F     58    161    51   159 22.37568    normal
## 3   F     53    161    54   158 20.44674    normal
## 4   M     68    177    70   175 21.70513    normal
```

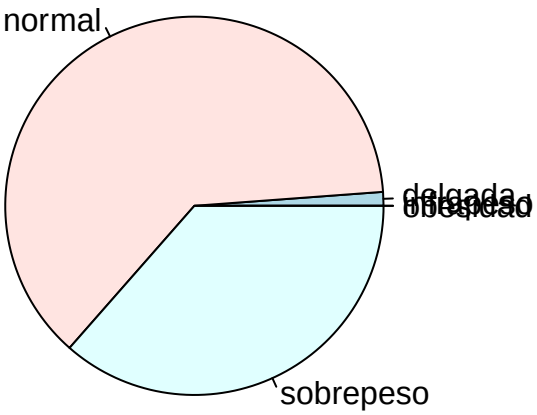
```
## 5   F     59    157    59   155 23.93606    normal
## 6   M     76    170    76   165 26.29758 sobrepeso
```

```r
library(ggplot2)
library(gridExtra)
ggplot(datos_davis,aes(x=factor(sex),fill=factor(IMCc))) +
  geom_bar(stat = "count", position="dodge")+
  labs(title = "Gráfico de IMC por Sexo") +
  labs(fill = "IMC") +
  labs(aes(x="Sexo",y="Personas"))
```

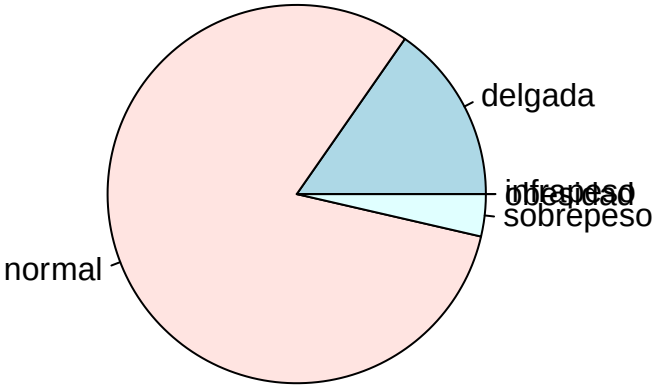



```r
#-- Gráfico de sectores
#-- Realizaremos dos gráficos separando los hombres ylas mujeres

#-- Gráfrico para los hombres
datos_hombres<-subset(datos_davis,datos_davis$sex=="M")
library(plotrix)
grafico_tarta<-pie(table(datos_hombres$IMCc),
                   main="Gráfico Hombres")
```

Gr<U+00E1>fico Hombres

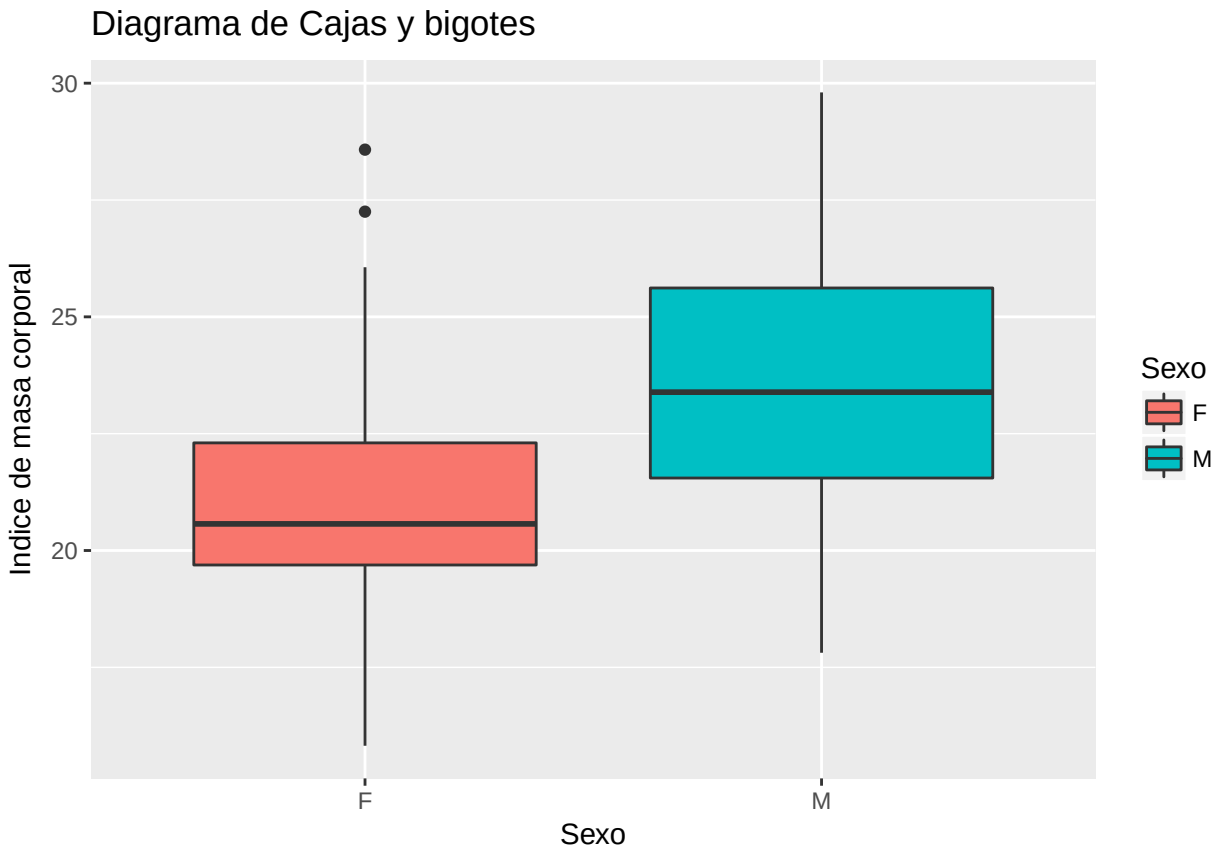



```
#-- Gráfico para las mujeres

datos_mujeres<-subset(datos_davis,datos_davis$sex=="F")
grafico_tarta<-pie(table(datos_mujeres$IMCc))
```



Gráficos de cajas e histogramas para la variable IMC numérica, también por sexo. ¿Existe algún outlier?, ¿cuáles?.

```
#-- Realizamos el gráfico de cajas y bigotes para el IMC por sexo
p <- ggplot(datos_davis,aes(sex,IMC))
p +
  labs(x = "Sexo",y = "Indice de masa corporal",
       title="Diagrama de Cajas y bigotes")+
  scale_fill_discrete(guide_legend(title = "Sexo"))+
  geom_boxplot(aes(fill=sex))
```




```
#-- Realizamos el hitográma del IMC
p <- ggplot(data=datos_davis, aes(x=IMC)) +
  geom_histogram(fill="steelblue") +
  ggtitle("Histograma Indice de masa corporal") +
  labs(aes(x="IMC")) +
  theme_minimal()

#-- Realizamos el histográma para el IMC por sexo
p <- ggplot(data=datos_davis, aes(x=IMC)) +
  geom_histogram(fill="steelblue") +
  ggtitle("Histograma IMC por Sexo") +
  facet_wrap(~sex)+
  theme_minimal()
```

Podemos observar en el gráfico de cajas y bigotes que exiten dos outliers para el sexo Femenino, estos outliers superan el IMC = 25 (mujeres con sobrepeso).

Ejercicio 3: Utiliza los datos "Arthritis" (paquete "vcd") sobre un ensayo clínico de doble ciego que investiga un nuevo tratamiento para la artritis reumatoide. Tenemos información de 84 observaciones de 5 variables: la identificación del paciente (ID), el tratamiento (Treatment: Placebo, Treated), el sexo (Sex: Female, Male), la edad (Age) y la mejoría (Improved: None, Some, Marked). Obtener las tablas de frecuencias y medidas de asociación entre estas variables. Interpreta los resultad

```
#-- Cargamos el paquete
library(vcd)
datos<-Arthritis
attach(datos) #-- Activamos las variables
head(datos)
```

```
##   ID Treatment  Sex Age Improved
## 1 57   Treated Male  27     Some
## 2 46   Treated Male  29     None
## 3 77   Treated Male  30     None
## 4 17   Treated Male  32   Marked
## 5 36   Treated Male  46   Marked
## 6 23   Treated Male  58   Marked
```

```
#-- Las tablas de frecuencia se realizarán por variable.
#-- Tabla de frecuencias de la variable Tratamiento
table(Treatment)
```

```
## Treatment
## Placebo Treated
##      43      41
```

```
#-- Tabla de frecuencias de la variable Sexo
table(Sex)
```

```
## Sex
## Female   Male
##     59     25
```

```
#-- Tabla de frecuencias de la variable Edad
table(Age)
```

```
## Age
## 23 27 29 30 31 32 33 37 41 44 45 46 48 49 50 51 52 53 54 55 56 57 58 59 60
##  2  1  1  3  1  3  1  3  2  2  1  2  3  1  1  2  1  2  3  3  1  5  3  8  1
## 61 62 63 64 65 66 67 68 69 70 74
##  2  4  4  3  1  4  1  3  3  2  1
```

```
#-- Tabla de frecuencias de la variable Improved
table(Improved)
```

```
## Improved
##   None   Some Marked
##     42     14     28
```

Una vez que hemos obtenido las tablas de frecuencia, el siguiente paso es realizar las medidas de asociación para las variables cualitativas y cuantitativas. Vamos a analizar en primer lugar el sexo con el tratamiento:

```r
summary(assocstats(table(datos$Sex,datos$Treatment)))
```

```
##
## Number of cases in table: 84
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0.7365, df = 1, p-value = 0.3908
##                      X^2 df P(> X^2)
## Likelihood Ratio 0.73748  1  0.39047
## Pearson          0.73653  1  0.39078
##
## Phi-Coefficient   : 0.094
## Contingency Coeff.: 0.093
## Cramer's V        : 0.094
```

Observamos que no existe una relación significativa (p>0.05) entre el sexo y el tratamiento y dado el coeficiente de Cramer tienen una asociación bastante baja.

Ahora vamos a analizar el sexo con la mejoria:

```r
summary(assocstats(table(datos$Sex,datos$Improved)))
```

```
##
## Number of cases in table: 84
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 4.841, df = 2, p-value = 0.08889
##  Chi-squared approximation may be incorrect
##                     X^2 df P(> X^2)
## Likelihood Ratio 5.0131  2 0.081550
## Pearson          4.8407  2 0.088891
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.233
## Cramer's V        : 0.24
```

Observamos que no existe una relación significativa (p>0.05) entre el sexo y la mejoria del paciente y dado el coeficiente de Cramer tienen una asociación baja.

Analizamos el tratamiento con la mejoria:

```r
summary(assocstats(table(datos$Treatment,datos$Improved)))
```

```
##
## Number of cases in table: 84
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 13.055, df = 2, p-value = 0.001463
##                     X^2 df  P(> X^2)
## Likelihood Ratio 13.530  2 0.0011536
## Pearson          13.055  2 0.0014626
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.367
## Cramer's V        : 0.394
```

En este caso podemos ver que existe una relación significativa (p<0.05) entre el tratamiento y la mejoría, el coeficiente de Cramer es bajo.

Ahora vamos a realizar el estudio con la edad y la mejoria:

```
summary(assocstats(table(datos$Age,datos$Improved)))
```

```
##
## Number of cases in table: 84
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 65.42, df = 70, p-value = 0.6329
##  Chi-squared approximation may be incorrect
##                     X^2 df P(> X^2)
## Likelihood Ratio 75.455 70  0.30656
## Pearson          65.417 70  0.63287
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.662
## Cramer's V        : 0.624
```

Según el coeficiente de Cramer exite una alta asociación entre estas variables, pero el $p<0.05$ y por lo tanto no existe una relación significativa.