Visualización de datos y estadísticos

Visualización de datos y estadísticos

Sitio: eGela 2017-18 UPV/EHU
Curso: Exploración y análisis de datos
Libro: Visualización de datos y estadísticos
Imprimido por: JESUS MARIA YURRAMENDI MENDIZABAL
Día: lunes, 25 de septiembre de 2017, 10:49

Tabla de contenidos

1 Ejemplo de aplicación. Descripción de las variables una a una

```
#####################################################################
#
# Introduccion de los datos: 'iris'
#
# UCI: Iris Data Set
#
# Wikipedia: Iris Data Set
#
?iris
#
iris
#
dim(iris)
#
head(iris)
tail(iris)
#
sapply(iris,class)
#
names(iris)
#[1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"
rownames(iris)
#
levels(iris$Species)
#
summary(iris)
#
#######
#
iris[,] # notacion matricial
#
iris[,2]
iris[,1:4]
iris[,-5]
#
iris[1,]
iris[1:10,]
#
iris[1,3]
iris[1:10,3:4]
#
#####################################################################
#
# Analisis descriptivo de una sola variable: CUANTITATIVA
#
#######
#
```

```
# Tabla de la distribucion de frecuencias
#
?table
table(iris$Sepal.Length) # Horrible!
#
# La tabla es estadisticamente valida
# cuando el numero de modalidades es pequeño
#
#######
#
# Graficas
#
# 1
#
?dotchart
dotchart(iris$Sepal.Length, pch=19, col="red",
         xlab="Sepal Length", main="Iris Data Set")
axis(2, las=1) # Horrible!
#
# Esta grafica es valida
# cuando el numero de objetos es pequeño
#
dotchart(iris$Sepal.Length[1:20], pch=19, col="red",
         xlab="Sepal Length", main="Iris Data Set")
axis(2, at=1:20, labels=rownames(iris[1:20,]), las=1)
#
# Mejor aún si se ordenan los elementos
# en función de su valor
#
orden <- order(iris$Sepal.Length[1:20])
dotchart(iris$Sepal.Length[orden], pch=19, col="red",
         xlab="Sepal Length", main="Iris Data Set")
axis(2, at=1:20, labels=rownames(iris[orden,]), las=1)
# 2
#
?barplot
marcas <- barplot(iris$Sepal.Length[orden], col="red", space=c(5,1),
              horiz=TRUE, xlab="Sepal Length", main="Iris Data Set")
axis(2, at=marcas, labels=rownames(iris[orden, ]), las=1)
#
marcas <- barplot(iris$Sepal.Length[orden], col="red", space=c(5,1),
                  axes=FALSE, horiz=FALSE,
                  ylab="Sepal Length", main="Iris Data Set")
axis(1, at=marcas, labels=rownames(iris[orden, ]))
axis(2, las=1)
#
# Si el conjunto de objetos fuera mas grande,
# la grafica no seria valida
#
#
# 3
#
```

```
?plot
plot(iris$Sepal.Length, type="l", col="red",
     xlab="Conjunto totalmente ordenado", ylab="X",
     las=1, frame.plot=FALSE)
#
#
# 4
#
?stripchart
stripchart(iris$Sepal.Length, method="stack", pch=19, col="red",
xlab="Sepal Length", main="Iris Data Set")
#
# Si el conjunto de objetos fuera mas grande,
# la grafica no seria valida
#
#
# 5
#
# http://en.wikipedia.org/wiki/Stem-and-leaf_display
#
?stem
stem(iris$Sepal.Length, scale=0.5)
#
#
# 6
#
?hist
hist(iris$Sepal.Length, col="red",
     xlab="Sepal Length", main="Iris Data Set")
#
# Muy valido cuando el conjunto de objetos es grande
#
#
# 7
#
# https://en.wikipedia.org/wiki/Box_plot
#
?boxplot
boxplot(iris$Sepal.Length, col="red", las=1, horizontal=TRUE,
        xlab="Sepal Length", main="Iris Data Set")
#
# Muy valido cuando el conjunto de objetos es grande
#
#######
#
# Estadisticos (indices)
#
summary(iris$Sepal.Length)
#
var(iris$Sepal.Length) # varianza
sd(iris$Sepal.Length) # desviacion tipica = sqrt(varianza)
#
```

```
#
# Analisis descriptivo de una sola variable: CUALITATIVA
#
#######
#
# Tabla de la distribucion de frecuencias
#
table(iris$Species) # frecuencias absolutas
#
prop.table(table(iris$Species)) # frecuencias relativas (porcentajes)
#
cbind(freq=table(iris$Species) ,
      porcentaje=prop.table(table(iris$Species)))
#
#######
#
# Graficas
#
# http://en.wikipedia.org/wiki/Bar_chart
#
?barplot
barplot(table(iris$Species) , col="red",
        xlab="Species", main="Iris Data Set")
#
# En el diagrama de barras las frecuencias se comparan mejor que en
# el diagrama de sectores https://en.wikipedia.org/wiki/Pie_chart
#
#######
#
summary(iris$Species)
#
#####################################################################
#
# Es preciso hacer notar que se ha usado la función  barplot()
# en dos contextos aparentemente diferentes, a saber:
#  barplot(iris$Sepal.Length, ...) y
#  barplot(table(iris$Species), ...)
#
# Sin embargo, responden a la misma situación:
# Una variable cualitativa y una variable cuantitativa
#
# En el primer caso la variable cualitativa es el identificador,
# y la cuantitativa 'Sepal.Length'.
# En el segundo caso la variable cualitativa es 'Species',
# y la cuantitativa la frecuencia asociada a cada modalidad.
#
# Por tanto, realizar una tabla de frecuencias consiste en
# dar un 'salto'
# por el que las modalidades de la variable cualitativa
# pasan a ser 'identificadores' en un siguiente nivel,
# en el que la variable cuantitativa la constituyen las frecuencias
```

```
#
# En este 'salto' se pierde la informacion aportada (si la hay) por
# el 'identificador' original.
#
#####################################################################
#
# Analisis descriptivo de todas las variables a la vez
#
#######
#
# Graficas
#
attach(iris)
par(mfrow=c(3,2))
#
hist(Sepal.Length, col="red", xlab="Sepal Length", main="")
hist(Sepal.Width,  col="red", xlab="Sepal Width",  main="")
hist(Petal.Length, col="red", xlab="Petal Length", main="")
hist(Petal.Width,  col="red", xlab="Petal Width",  main="")
#
barplot( table(iris$Species) , col="red", xlab="Species", main="")
#
par(mfrow=c(1,1))
detach(iris)
#
#######
#
# Estadisticos (indices)
#
summary(iris)
#
apply(iris, 2, var) # varianzas ; atencion 'cualitativa'!
apply(iris[,-5], 2, var) # varianzas
#
#####################################################################
#####################################################################
#
# Transformacion: categorizacion de una variable CUANTITATIVA
#
#######
#
(Sepal.Length.categ <- cut(iris$Sepal.Length , breaks=4))
?cut
summary(Sepal.Length.categ)
#
Sepal.Length.categ.2 <- cut(iris$Sepal.Length,
               breaks=quantile(iris$Sepal.Length, seq(0, 1, 0.25)),
               include.lowest=TRUE)
summary(Sepal.Length.categ.2)
#
#####################################################################
```

```
#######################################################################
#
# Transformacion: binarizacion de una variable CUALITATIVA
#
#######
#
iris.bin <- iris
for(k in 1:length(levels(iris$Species))){
  iris.bin <- cbind(iris.bin,
            as.integer(iris.bin$Species == levels(iris$Species)[k]))
  names(iris.bin)[ncol(iris)+k] <- levels(iris$Species)[k]
}
#
head(iris.bin)
#
#######################################################################
```

2 Ejemplo de aplicación. Descripción de las asociaciones entre cada par de variables

```
#####################################################################
#
# Introduccion de los datos: 'iris'
#
iris
#
dim(iris)
#
attach(iris)
#
#####################################################################
#
# Analisis descriptivo de la asociacion
# de un par de variables CUANTITATIVAS
#
#######
#
# Tabla de la distribucion de frecuencias
#
table(Sepal.Length, Sepal.Width) # Horrible!
#
# La tabla es estadisticamente valida
# cuando el numero de modalidades es pequeño
#
#######
#
# Graficas
#
?plot
plot(Sepal.Length, Sepal.Width, las=1, type="p", pch=19, col="red",
     xlab="Sepal Length", ylab="Sepal Width", main="Iris Data Set")
#
#
#######
#
# Estadisticos (indices)
#
# Correlacion lineal
#
?cor
cor(Sepal.Length, Sepal.Width)
#
# Coeficientes de correlación : http://en.wikipedia.org/wiki/Correlation
#
# Coeficiente de correlación lineal: http://en.wikipedia.org
/wiki/Pearson_correlation
#
```

```
cor(Sepal.Length, Sepal.Width) ==
   cov(Sepal.Length, Sepal.Width)/(sd(Sepal.Length)*sd(Sepal.Width)) #
'cov()' covarianza
cov(Sepal.Length, Sepal.Length) == var(Sepal.Length)
#
# -1 <= cor(x1,x2) <= +1
# Si cor(x1,x2)=+1, entonces relación lineal directa (creciente)
# Si cor(x1,x2)=-1, entonces relación lineal inversa (decreciente)
# Si cor(x1,x2)= 0, entonces incorrelados linealmente (no hay relación
lineal)
# cor(x1,x2)= 0 y (x1,x2) distribucion binormal si y sólo si independencia
(no hay relación)
#
# Coeficientes para datos ordenados
# - Spearman
# - Kendall
# - Goodman & Kruskal
#
#
# http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient
#
cor(Sepal.Length, Sepal.Width, method="spearman")
#
# http://en.wikipedia.org/wiki/Kendall_tau_rank_correlation_coefficient
# http://en.wikipedia.org/wiki/Kendall_tau
#
cor(Sepal.Length, Sepal.Width, method="kendall")
#
#######################################################################
#
# Analisis descriptivo de la asociacion entre
# una variable CUANTITATIVA y una variable CUALITATIVA
#
#######
#
# Tabla de la distribucion de frecuencias
#
table(Sepal.Length, Species) # Horrible!
#
# La tabla es estadisticamente valida
# cuando el numero de modalidades de la variable cuantitativa
# es pequeño
#
#######
#
# Graficas
#
#
# 1
#
stripchart(Sepal.Length ~ Species, pch=19 , method="stack",
           col=c("red","green3","blue"),
```

```
     xlab="Sepal Length", ylab="Species", main="Iris Data Set")
#
#
# 2
#
par(mfrow=c(3,1))
#
hist(Sepal.Length[Species=="virginica"],
     xlim=c(min(Sepal.Length),max(Sepal.Length)), col="blue",
     xlab="Sepal Length", main="virginica")
hist(Sepal.Length[Species=="versicolor"],
     xlim=c(min(Sepal.Length),max(Sepal.Length)), col="green3",
     xlab="Sepal Length", main="versicolor")
hist(Sepal.Length[Species=="setosa"],
     xlim=c(min(Sepal.Length),max(Sepal.Length)), col="red",
     xlab="Sepal Length", main="setosa")
#
par(mfrow=c(1,1))
#
#
# 3
#
boxplot(Sepal.Length ~ Species, pch=19, horizontal=TRUE,
        col=c("red","green3","blue"),
        xlab="Sepal Length", ylab="Species", main="Iris Data Set")
#
#######
#
# Estadisticos (indices)
#
#
# Razon de correlacion (indice):
# http://fr.wikipedia.org/wiki/Rapport_de_corr%C3%A9lation
#
# Se puede definir una funcion que calcule
# la razon de correlacion, eta2
#
eta2 <- function(x, factor){
   niv <- levels(factor)
   numniv <- length(niv)
   SSB <- 0
   for(i in 1:numniv){
     xx <- x[factor==niv[i]]
     nxx <- length(xx)
     SSB <- SSB + nxx*(mean(xx)-mean(x))^2
   }
   SST <- (length(x)-1)*var(x)
#
   eta2value <- SSB/SST
#
   return(eta2value)
}
```

```
#
eta2(Sepal.Length, Species)
#
# 0 <= eta2 <= 1
# Si eta2=0, entonces no hay asociacion entre 'x' e 'y',
#            las medias parciales son todas iguales
# Si eta2=1, entonces hay una dependencia funcional entre 'x' e 'y'
#            no hay variabilidad dentro de las categorias
#
#########################################################################
#
# Analisis descriptivo de la asociacion
# de un par de variables CUALITATIVAS
#
########################
#
# Preparacion de los datos
#
?HairEyeColor
HairEyeColor
HairEyeColor[,,"Female"]
HairEyeColor[,,"Male"]
#
# Suma sobre 'Sex'
#
x <- apply(HairEyeColor, c(1, 2), sum)
x
#
coloreshair <- c("black", "brown", "red", "yellow")
coloreseye <- c("brown", "turquoise1", "#8E7618", "#BCEE68")
#
########################
#
# Frecuencias absolutas
#
#
# Diagrama de barras
#
barplot(x, las=1, col=coloreshair, beside=TRUE,
        xlab="Eye", main="Relation between hair and eye color")
par("usr")
legend(15, 110, legend=rownames(x), pch=15,
       col=coloreshair, box.col="black",
       title="Hair")
#
barplot(t(x), las=1, col=coloreseye, beside=TRUE,
        xlab="Hair", main="Relation between hair and eye color")
par("usr")
legend(10, 110, legend=colnames(x), pch=15,
       col=coloreseye, box.col="black",
       title="Eye")
#
```

```
#
# Diagrama de barras acumuladas
#
barplot(x, las=1, col=coloreshair, beside=FALSE,
        xlab="Eye", main="Relation between hair and eye color")
par("usr")
legend(3.75, 220, legend=rownames(x), pch=15,
       col=coloreshair, box.col="black",
       title="Hair")
#
barplot(t(x), las=1, col=coloreseye, beside=FALSE,
        xlab="Hair", main="Relation between hair and eye color")
par("usr")
legend(3.755, 320, legend=colnames(x), pch=15,
       col=coloreseye, box.col="black",
       title="Eye")
#
########################
#
# Porcentajes, Frecuencias relativas
#
#
# Diagrama de barras
#
barplot(prop.table(x, margin=2)*100, las=1,
        col=coloreshair, beside=TRUE,
        xlab="Eye", main="Relation between hair and eye color")
#
barplot(prop.table(t(x), margin=2)*100, las=1,
        col=coloreseye, beside=TRUE,
        xlab="Hair", main="Relation between hair and eye color")
#
# Diagrama de barras acumuladas
#
barplot(prop.table(x, margin=2)*100, las=1,
        col=coloreshair, beside=FALSE,
        xlab="Eye", main="Relation between hair and eye color")
#
barplot(prop.table(t(x), margin=2)*100, las=1,
        col=coloreseye, beside=FALSE,
        xlab="Hair", main="Relation between hair and eye color")
#
# Diagrama de mosaicos (barras acumuladas proporcionales)
#?mosaicplot
mosaicplot(x, main = "Relation between hair and eye color", las=1,
           color=coloreseye)
mosaicplot(t(x), main = "Relation between hair and eye color", las=1,
           color=coloreshair)
#
########################
#
# Gráficos relacionados con la relación de independencia
```

```
#
?assocplot
# desviaciones respecto a la independencia entre ambas variables
assocplot(x, col=c("red", "blue"),
          main="Relation between hair and eye color")
assocplot(t(x), col=c("red", "blue"),
          main="Relation between hair and eye color")
#
#
# Otro ejemplo
#
?occupationalStatus
occupationalStatus
class(occupationalStatus)
attributes(occupationalStatus)
#
assocplot(occupationalStatus, col=c("red", "blue"),
          main="Occupational Status of\nFathers and their Sons")
mosaicplot(occupationalStatus, las=1)
#
#
#####################################################################
#####################################################################
#
# Analisis descriptivo de todos los pares de variables CUANTITATIVAS
# de una vez
#
#######
#
# Graficas
#
pairs(iris[,-5], pch=19, col="red")
#
#######
#
# Estadisticos (indices)
#
#
# Correlacion lineal
#
cor(iris[,-5])
#
#####################################################################
detach(iris)
#
#####################################################################
```

3 Ejemplo de aplicación. Descripción de las asociaciones entre tres variables

```
#######################################################################
#
# Analisis descriptivo de tres variables CUANTITATIVAS
#
#######
#
# Graficas
#
library(scatterplot3d)
scatterplot3d(iris$Petal.Length, iris$Petal.Width, iris$Sepal.Length)
#
#######################################################################
#
# Analisis descriptivo de dos variables CUANTITATIVAS y una CUALITATIVA
#
#######
#
# Graficas
#
plot(iris$Sepal.Length, iris$Sepal.Width,
     col=c("red", "green3", "blue")[iris$Species],
     bg= c("red", "green3", "blue")[iris$Species],
     pch=c(22, 24, 25)[iris$Species]               )
#
#######
#
# Estadisticos
#
cor(iris$Sepal.Length[iris$Species=="setosa"],
    iris$Sepal.Width[iris$Species=="setosa"])
cor(iris$Sepal.Length[iris$Species=="versicolor"],
    iris$Sepal.Width[iris$Species=="versicolor"])
cor(iris$Sepal.Length[iris$Species=="virginica"],
    iris$Sepal.Width[iris$Species=="virginica"])
#
#######################################################################
#
# Analisis descriptivo de todos los pares de variables CUANTITATIVAS
#                       y una variable CUALITATIVA de una vez
#
#######
#
# Graficas
#
pairs(iris[,-5], main="Edgar Anderson's Iris Data",
      pch=21, bg = c("red", "green3", "blue")[iris$Species])
#
```

4 Ejemplo de aplicación. Descripción de las asociaciones entre multiples variables

```
#################################################################
#
# Cuatro variables
#
# CUANTITATIVA, CUANTITATIVA, CUANTITATIVA, CUALITATIVA
#
#######
#
# Graficas
#
library(scatterplot3d)
?scatterplot3d
#
scatterplot3d(iris$Petal.Length, iris$Petal.Width, iris$Sepal.Length,
              pch=19, color=c("red", "green3", "blue")[iris$Species])
#
scatterplot3d(iris$Petal.Length, iris$Petal.Width, iris$Sepal.Length,
              pch=19, color=c("red", "green3", "blue")[iris$Species],
angle=60)
#
scatterplot3d(iris$Petal.Length, iris$Petal.Width, iris$Sepal.Length,
              pch=19, color=c("red", "green3", "blue")[iris$Species],
angle=75)
#
#
#################################################################
#################################################################
#
# Multiple variables
#
#
#######
#
# Graficas
#
#
# Coordenadas paralelas
#
# https://en.wikipedia.org/wiki/Parallel_coordinates
#
library(MASS)
#
parcoord(iris[,1:4],
col=c("red", "green3", "blue")[unclass(iris$Species)])
#
parcoord(iris[,sample(4)],
col=c("red", "green3", "blue")[unclass(iris$Species)])
```

```
#
cor(iris[, 1:4])
#
parcoord(iris[,c(3,4,1,2)],
col=c("red", "green3", "blue")[unclass(iris$Species)])
#
#######
#
# Diagrama de estrella o radar
#
# https://en.wikipedia.org/wiki/Radar_chart
#
?stars
#
stars(iris[,1:4], ncol=10,
      col.stars=c("red", "green3", "blue")[iris$Species])
#
#######
#
# Caras de Chernoff
#
# https://en.wikipedia.org/wiki/Chernoff_face
#
#####################################################################
```

5 Ejercicio. Vertebral column

El fichero "column0.RData", que es la base de datos a analizar, se ha tomado desde Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Vertebral+Column)

En este conjunto de datos cada paciente está representado por seis atributos biomecánicos derivados de la forma y la orientación de la pelvis y de la columna lumbar (PI (Pelvic incidence), PT (Pelvic tilt), LLA (Lumbar lordosis angle), SS (Sacral slope), PR (Pelvic radius) y GS (grade of spondylolisthesis)), y un diagnóstico (DIAG, en el que las modalidades son DH (Disk Hernia), SL (Spondylolisthesis), y NO (Normal)).

1. **Análisis univariado**. Trazar un **diagrama de cajas** para cada variable cuantitativa. Dar un resumen de los **estadísticos** que describen las variables. Se van a suprimir 4 pacientes, ya que sus perfiles son muy *raros* respecto a los demás ('*outlier*'), y se va a constituir el fichero "column.RData".
2. **Análisis bivariado**. Trazar un **diagrama de cajas en relación a las clases** ou modalidades de DIAG para cada variable cuantitativa. Calcular los valores de la razón de correlación (η^2 o eta2) de cada variable cuantitativa y la variable DIAG. Comentar cuáles son las variables más asociadas a DIAG, et su capacidad de separar las modalidades de DIAG.

6 Ejercicio. User Knowledge Modeling

El fichero "DUMtraining.RData", que es la base de datos a analizar, se ha tomado desde Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling)

En este conjunto de datos cada usuario está representado por cinco atributos STG (nivel de tiempo de estudio para los materiales del objetivo), SCG (nivel del número de repeticiones del usuario para lograr los materiales del objetivo), STR (nivel del tiempo de estudio del usuario para objetos relacionados con el objetivo), LPR (resultado del examen del usuario para objetos relacionados con el objetivo), PEG (resultado del examen del usuario para el objetivo), y un diagnóstico UNS (nivel de conocimiento del usuario) en el que las modalidades son Very Low, Low, Middle y High.

1. **Análisis univariado**. Trazar un **diagrama de cajas** para cada variable (ordinal) cuantitativa.
2. **Análisis bivariado**. Trazar un **diagrama de cajas en relación a las clases** o modalidades de UNS para cada variable cuantitativa. Calcular los valores de la razón de correlación (η^2 o eta2) de cada variable cuantitativa y la variable UNS. Comentar cuáles son las variables más asociadas a UNS, y su capacidad de separar las modalidades de UNS.