Actividad 2 - Análisis estadístico

Francisco Javier Melchor González

9/11/2020

Contents

1. Estadística descriptiva

En primer lugar, realizamos la lectura del fichero **ChildCarSeats_clean**, aplicando para ello la función *read.csv*.

En este caso, indicaremos como parámetros que el dataset sí tiene header *(header=TRUE)*, que el separador de columnas es la ',' *(sep=",")*, que los strings a interpretar como NA son tanto los campos vacíos, los que tienen un espacio en blanco y en los que aparece la cadena "NA" *(na.strings=c(""," ","NA"))* y por último, que las columnas de tipo String, sean consideradas como factores, ya que todas las columnas que son de tipo String, en este caso son factores.

```
childCarSeats_clean_filename <- "../Data/ChildCarSeats_clean.csv"
childCarSeats_clean <- read.csv(file=childCarSeats_clean_filename, header=TRUE, sep=",",
                                na.strings=c("","  ","NA"), stringsAsFactors=TRUE)
head(childCarSeats_clean)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

```
str(childCarSeats_clean)
```

```
## 'data.frame':    400 obs. of  11 variables:
##  $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice  : int  138 111 113 117 141 124 115 136 132 132 ...
##  $ Income     : int  73 48 35 100 64 113 105 81 110 113 ...
##  $ Advertising: int  11 16 10 4 3 13 0 15 0 0 ...
##  $ Population : int  276 260 269 466 340 501 45 425 108 131 ...
##  $ Price      : int  120 83 80 97 128 72 108 120 124 124 ...
##  $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
##  $ Age        : int  42 65 59 55 38 78 71 67 76 76 ...
##  $ Education  : int  17 10 12 14 13 16 15 10 10 17 ...
##  $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
##  $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

Como se puede observar, todos los tipos de las columnas han sido asignagos correctamente.

A continuación, procederemos a realizar una visualización de las diferentes columnas o variables que forman el dataset, para ver como se distribuyen las mismas.

1.1 Representación gráfica de variables categóricas o cualitativas

```
unlist(lapply(childCarSeats_clean, is.factor))
```

```
##       Sales   CompPrice      Income Advertising  Population       Price
##       FALSE       FALSE       FALSE       FALSE       FALSE       FALSE
##   ShelveLoc         Age   Education       Urban          US
##        TRUE       FALSE       FALSE        TRUE        TRUE
```

Como se puede observar, las únicas variables categóricas son:

- **ShelveLoc**, que indica la calidad de la ubicación de las sillas en la tienda (tres posibles valores: Bad, Good y Medium)
- **Urban**, que indica si la tienda se encuentra en una ubicación urbana o rural (dos posibles valores: Yes y No)
- **US**, que indica si la tienda se encuentra en EUA o no (dos posibles valores: Yes y No)

Para representar gráficamente las variables anteriores, realizaremos un **diagrama de barras** en el caso de la variable **ShelveLoc** y un **diagrama de sectores** para las variables **Urban** y **US**.

La razón por la cual he considerado más oportuno utilizar para las variables Urban y US un diagrama de sectores y no un diagrama de barras, es porque estas solo pueden tomar dos posibles valores, por lo que considero que un diagrama de sectores permitirá captar mejor a simple vista la distribución de las categorías, que si se representa mediante un diagrama de barras.

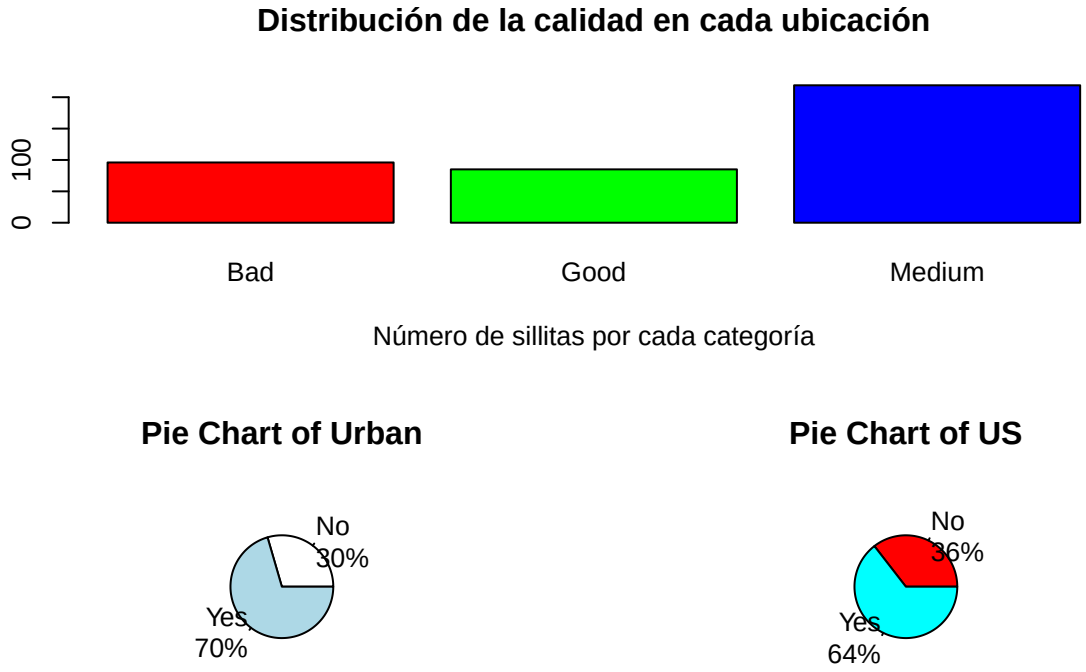
```
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))


counts <- table(childCarSeats_clean$ShelveLoc)
barplot(counts, main="Distribución de la calidad en cada ubicación",
```

```
        xlab="Número de sillitas por cada categoría", col = rainbow
        (length(levels(childCarSeats_clean$ShelveLoc))))

mytableUrban <- table(childCarSeats_clean$Urban)
pctUrban <- round(mytableUrban/sum(mytableUrban)*100)
lblsUrban <- paste(names(mytableUrban), "\n", pctUrban, sep="")
lblsUrban <- paste (lblsUrban, '%', sep="")
pie(mytableUrban, labels = lblsUrban,
    main="Pie Chart of Urban\n")


mytableUS <- table(childCarSeats_clean$US)
pctUS <- round(mytableUS/sum(mytableUS)*100)
lblsUS <- paste(names(mytableUS), "\n", pctUS, sep="")
lblsUS <- paste (lblsUS, '%', sep="")
pie(mytableUS, labels = lblsUS, col=rainbow(length(lblsUS)),
    main="Pie Chart of US\n")
```



- La primera gráfica, nos indica que **la mayoría de las sillas tienen un nivel de calidad de ubicación medio**, y que el grupo que presenta la minoría es el que se corresponde con la calidad de ubicación buena, lo que quiere decir que **sólo una pequeña parte del total de tiendas tienen ubicadas correctamente las sillas**.
- La segunda gráfica empezando por la izquierda, nos indica que **la mayoría de las tiendas analizadas se encuentran en una población urbana**, pues la clase "Yes" representa el 70% de los casos.

- Por último, la última gráfica indica que **la mayoría de las tiendas que se están analizando se encuentran dentro de USA**, pues la clase "Yes", representa un 64%.

1.1 Representación gráfica de variables numéricas

```
unlist(lapply(childCarSeats_clean, is.numeric))
```

```
##       Sales   CompPrice      Income Advertising  Population       Price
##        TRUE        TRUE        TRUE        TRUE        TRUE        TRUE
##   ShelveLoc         Age   Education       Urban          US
##       FALSE        TRUE        TRUE       FALSE       FALSE
```

Como se puede observar, las variables numéricas representan la gran mayoría de las variables del dataset, y estas son:

- **Sales**, que indica el número de ventas unitarias, en miles, en cada ubicación
- **ComPrice**, que indica el precio que cobra la competencia en cada ubicación.
- **Income**, que indica el nivel de ingresos comunitarios, en miles de dólares
- **Adversiting**, que indica el presupuesto de publicidad local de la empresa en cada ubicación, en miles de dólares.
- **Population**, que indica el tamaño de la población en la región, en miles.
- **Price**, que indica el precio de las sillitas de coche en cada ubicación
- **Age**, que indica la edad media de la población local.

Todas ellas son de tipo numérico, pero solo la primera es de tipo decimal, las demás son de tipo entero. Por ello, para representarlas gráficamente, en el caso de la variable **Sales**, la representaremos mediante un diagrama de puntos y en el caso de las demás variables enteras, se representarán mediante un histograma de frecuencias relativas.

La razón por la que se ha decidido representar la variable **Sales** mediante un diagrama de punto es porque al poder tomar valores decimales, esta representación permitirá ver mejor como se distribuyen los diferentes valores, mientras que el histograma nos permitirá ver mejor las demás variables que son de tipo entero.

```
par(mfrow=c(2,4))

dotchart(childCarSeats_clean$Sales,labels=,cex=0.7,
         main="Ventas por ubicación",
         xlab="Ventas por mil", cex.main=0.8, cex.lab=0.8)

colorForHistograms = rainbow(table (unlist(lapply(childCarSeats_clean, is.numeric)))["TRUE"] - 1)

hist(childCarSeats_clean$CompPrice, breaks=sqrt(dim(childCarSeats_clean)[1]), col=colorForHistograms[1]
     xlab="Precio en euros",cex.main=0.8, cex.lab=0.8)

hist(childCarSeats_clean$Income, breaks=sqrt(dim(childCarSeats_clean)[1]), col=colorForHistograms[2],ma
     xlab="Nivel de ingresos en miles de dólares",cex.main=0.8, cex.lab=0.8)

hist(childCarSeats_clean$Advertising, breaks=sqrt(dim(childCarSeats_clean)[1]), col=colorForHistograms[
     xlab="Presupuesto en miles de dólares",cex.main=0.8, cex.lab=0.8)

hist(childCarSeats_clean$Population, breaks=sqrt(dim(childCarSeats_clean)[1]), col=colorForHistograms[4
```

```
     xlab="Tamaño de la población en miles",cex.main=0.8, cex.lab=0.8)

hist(childCarSeats_clean$Price, breaks=sqrt(dim(childCarSeats_clean)[1]), col=colorForHistograms[5],mai
     xlab="Precio en euros",cex.main=0.8, cex.lab=0.8)

hist(childCarSeats_clean$Age, breaks=sqrt(dim(childCarSeats_clean)[1]), col=colorForHistograms[6],main=
     xlab="Edad media en años",cex.main=0.8, cex.lab=0.8)

hist(childCarSeats_clean$Age, breaks=sqrt(dim(childCarSeats_clean)[1]), col=colorForHistograms[7],main=
     de la población",
     xlab="Años de educación",cex.main=0.8, cex.lab=0.8)
```