Guió de les classes de laboratori sobre anàlisi descriptiva

1. Començar a treballar amb R

1. Comencem fent una ullada a l'script d'R que es troba a `http://www-eio.upc.es/teaching/pe/read-data` i l'executem:

   ```
   > source(url("http://www-eio.upc.es/teaching/pe/read-data"))
   ```

   Fixem-nos que, entre d'altres, s'ha creat un conjunt de dades, un *data frame*, que es diu `DAT`. Abans de començar es recomanable mirar-nos aquestes dades, per exemple executant els següents comandes:

   ```
   > DAT
   > View(DAT)
   > head(DAT)
   > str(DAT)
   ```

2. El que ens interessa és fer una descripció d'aquestes dades. Com ho podem fer en cas de les variables numèriques?
   ⇝ Transparències 4 i 7 a 10 de `ED.ppt`

   Amb R:

   ```
   > mean(DAT$mysql)
   > median(DAT$mysql)
   > summary(DAT$mysql)
   > var(DAT$mysql)
   > sd(DAT$mysql)
   ```

   Com s'interpreten aquests valors?

3. Apart de calcular els indicadors numèrics de la tendència central i de la dispersió, es recomanable fer una representació gràfica de la distribució d'aquesta variable.
   ⇝ Transparències 12 a 16 de `ED.ppt`

   ```
   > hist(DAT$mysql)
   > boxplot(DAT$mysql)
   ```

   Què s'hi observa? Quin dels dos gràfics es preferible en aquest cas?

4. En canvi, en cas de la variable categòrica `op`, què ens interessa saber? Com la podem descriure?
   ⇝ Transparències 12, 17 i 19 de `ED.ppt`

   ```
   > table(DAT$op)
   > prop.table(table(DAT$op))
   > barplot(table(DAT$op))
   ```

5. Seria desitjable que la taula de freqüència contingués tant les freqüències absolutes com les relatives. Existeixen varies funcions en diferents paquets d'R que es poden instal·lar. Per exemple podem utilitzar la funció `freq` del paquet `descr`:

```
> # install.packages("descr")  # Instal·lació del paquet (només si no està instal·lat)
> library(descr)               # Es carrega el paquet
> freq(DAT$op, plot = FALSE)
```

Important: Un cop instal·lat un paquet d'R en un ordinador, ja no cal fer-ho a les pròximes sesions d'R. En canvi, se'l ha de carregar en cada nova sessió d'R.

6. Si volem exportar les dades podem utilitzar la funció `write.table` i per guardar el contingut d'una sessió d'R podem fer servir la funció `save.image`:

```
> write.table(DAT, file = "DadesDAT.txt", quote = FALSE, row.names = FALSE)
> save.image("DadesDAT.RData")
```

2. Exercicis

`Nota:` Copieu tant les preguntes com les instruccions d'R següents i pegueu-les a un document WORD. A continuació completeu el document WORD amb les instruccions completes, els resultats i vostres comentaris.

1. Feu una descripció de la segona variable numèrica `post`. Com aquesta variable te alguna dada mancant (*missing*), que es denota amb `NA` en R, les funcions `mean`, `var`, etc. tanmateix tornen un `NA`. Per resoldre aquest problema podeu mirar l'ajuda de la funció `mean` (executant `?mean` en R) o mirar l'apartat 3.1.1 del tutorial d'R a `http://www-eio.upc.es/teaching/pe/R/`:

```
> ?mean
> mean(...)
> sd(...)
```

   Un cop resolt el petit problema i fet els càlculs, interpreteu els resultats obtinguts.

2. Feu també una representació gràfica de la variable `post`:

```
> windows(width = 10)
> par(mfrow = c(1, 2))
> hist(...)
> boxplot(..)
```

   Com es pot descriure aquesta distribució?

3. Per saber si el comportament d'aquesta variable varia d'un `op` a un altre, s'han de fer els càlculs dels indicadors numèrics per cada grup. En R ho podem fer amb la funció `tapply`.
   ⇝ Transparència 11 de `ED.ppt` o pàgina 39 del tutorial

```
> with(DAT, tapply(...))
```

4. Dibuixeu un diagrama de caixa de la variable `post` en funció de `op`:
   ⇝ Transparència 14 de `ED.ppt` o pàgines 52 i 53 del tutorial

```
> boxplot(post ~ ...)
```

   Comenteu les diferències que hi podeu observar.

3. Anàlisi descriptiva bivariant

1. A continuació utilitzarem un dels conjunts de dades de la llibreria `datasets` d'R. Es tracta de dades dels 50 estats dels Estats Units:

```
> ?state
> View(state.x77)
> str(state.x77)
> str(as.data.frame(state.x77))
> state.region
> state77 <- data.frame(state.x77, state.region)
> head(state77)
> summary(state77)
> names(state77)[c(4, 6, 9)] <- c("LifeExp", "HSGrade", "Region")
> names(state77)
```

2. Ens interessa ara fer una anàlisi descriptiva bivariant, tant de dues variables numèriques com d'un parell de variables categòriques. Per al primer cas és molt recomanable fer un diagrama de dispersió (*scatterplot*), que ens dona una idea de la relació entre ambdues variables.
   ⇝ Transparències 12, 20, 21 i 24 de `ED.ppt`

```
> windows(height = 10, width = 10)
> par(las = 1)
> pairs(state77[, 1:8], pch = 16)
```

   Entre quines variables sembla haver-hi més relació? Mirem amb més detall les relacions entre algunes de les variables:

```
> windows(height = 5, width = 15)
> par(mfrow = c(1, 3), las = 1, font.lab = 2, font.axis = 3)
> with(state77, plot(Illiteracy, LifeExp, pch = 17, col = 2))
> with(state77, plot(Illiteracy ~ Income, pch = 18, col = 3, cex = 1.3))
> plot(LifeExp ~ Frost, data = state77, pch = 19, col = 4,
+      ylab = "Life expectancy", cex = 1.5)
```

   Què hi podem observar? Com podem descriure les relacions?

3. En cas de que podem suposar una relació lineal entre dues variables numèriques, es pot calcular el coefficient de correlació (lineal), que quantifica el grau de relació lineal:
   ⇝ Transparències 22 i 23 de `ED.ppt`

```
> cor(state77)
> round(cor(state77[, 1:8]), 3)
> with(state77, round(cor(Area, Illiteracy), 3))
> with(state77, round(cor(LifeExp, Illiteracy), 3))
```

   Interpreteu aquests valors.

4. Per categoritzar una variable numèrica per tal de crear una variable ordinal podem usar la funció `cut`:
   ⇝ Pàgina 54 del tutorial

```
> cut(state77$Income, c(0, 4000, 4500, 5000, 10000))
> state77$Income.cat <- cut(state77$Income, c(0, 4000, 4500, 5000, 10000),
+                           labels = c("<= 4000", "4001--4500", "4501--5000",
+                                      ">5000"))
> head(state77, 10)
```

5. La relació entre dues variables categòriques es pot presentar mitjançant taules de contingència. Aquestes poden mostrar la distribució conjunta o la distribució condicional d'una de les dues variables en funció de l'altra:

```
> with(state77, table(Region, Income.cat))
> # install.packages("descr")  # només si no està instal·lat el paquet
> library(descr)
> with(state77, CrossTable(Region, Income.cat, prop.c = FALSE, prop.t = FALSE,
+                          prop.chisq = FALSE, format = "SPSS"))
```

   Sembla existir una relació entre les dues variables?

6. A més a més es poden fer diferents diagrames –un diagrama de barres o un diagrama de mosaic– per visualitzar aquesta relació:
   ⇝ Transparències 18, 19 i 25 i 28 de `ED.ppt`

```
> windows(height = 8, width = 16)
> par(mfrow = c(1, 2), las = 1, font.lab = 2, font.axis = 3)
> with(state77, barplot(table(Income.cat, Region), col = 1:4, legend = TRUE,
+                       xlab = "Region"))
> mosaicplot(Region ~ Income.cat, data = state77, col = 1:4, ylab = "Income",
+            main = "Income per region")
```

   Interpreteu els dos diagrames.

4. Exercicis

`Nota:` Copieu tant les preguntes com les instruccions d'R següents i pegueu-les a un document WORD. A continuació completeu el document WORD amb les instruccions completes, els resultats i vostres comentaris.

1. Tornem a treballar amb les dades dels dos gestors de bases de dades:

```
> source(url("http://www-eio.upc.es/teaching/pe/read-data"))
> head(DAT)
> summary(DAT)
```

2. Estudieu la relació entre les dues variables numèriques fent un diagrama de dispersió i feu un altre per a les variables transformades amb logaritme.

```
> plot(post ~ ...)
> plot(log(post) ~ ...)
```

Què hi podeu observar?

3. Calculeu la correlació de les dues parelles de variables i n'interpreteu els seus valors:

```
> with(DAT, cor(...))
```

4. Creeu una variable ordinal amb els temps de `post` utilitzant com punts de tall els tres quartils. A continuació feu una taula de contingència amb la nova variable i la variable `op`.

```
> DAT$post.cat <- cut(DAT$post, ...)
> with(DAT, CrossTable(...))
```

Sembla haver-hi diferències entre les operacions pel que fa als temps que triga `post`?

5. Feu una representació gràfica de la taula de contingència anterior amb un diagrama de barres i també amb un diagrama de mosaic.
⇝ Veure també les pàgines 56 a 57 del tutorial

```
> with(DAT, barplot(...))
> mosaicplot(...)
```

6. Executeu l'script `letsmakeadeal.R` i comenteu el resultat. És el que heu esperat?

```
> source("Letsmakeadeal.R")
```