Grau interuniversitari (UB-UPC) d'Estadística
Software Estadístic: Solució de la pràctica final amb R

Exercici 1 (5.5 punts)
Heu d'aconseguir una base de dades (ASCII, EXCEL, SPSS o SAS) en què es mesurin una sèrie de variables sobre **un mínim de 50 individus**. En concret, cada observació constarà d'una variable identificadora `id` (pot ser un nom de persona, una marca, un codi, etc), i estarà acompanyada de set variables més: Una variable categòrica binària, una variable categòrica politòmica (amb més de 2 categories) i cinc variables de tipus numèric. Es valorarà la dificultat i l'originalitat de les dades triades, i en cap cas s'admetran els següents conjunts de dades:

- Les utilitzades a pràctiques d'R d'altres anys o en altres assignatures (els alumnes repetidors de l'assignatura que van fer el treball sí podran fer servir les seves dades de l'any passat en el cas que realitzin la pràctica de forma individual).
- Les ja utilitzades per realitzar qualsevol tipus d'informe estadístic que es pugui trobar.
- Les que estiguin incorporades dins d'alguna llibreria d'R o bé incloses en qualsevol altre programa estadístic.

La utilització de dades no permeses suposarà la invalidació de l'exercici.

La memòria en PDF per a aquest exercici contindrà la resolució dels següents apartats:

a) Redacteu un apartat d'introducció que contingui els següents punts:
- El tipus de dades que utilitzeu en el treball.
- La font d'obtenció de les dades (incloent l'enllaç a internet en cas d'haver-ne).
- El significat de les diverses variables segons la seva nomenclatura dins la base de dades.
- L'objectiu de l'anàlisi que es realitza.

b) Importeu les dades a R com un *data frame* anomenat `df` i substituïu el nom de cada registre de `df` pel corresponent nom de la variable `id` (la qual desapareix al fer la substitució). Independentment de que `df` contingui valors perduts, afegiu-hi aleatòriament 1 *missing* a la 1a variable numèrica i 2 *missings* a la 2a variable numèrica.

Per tal d'esquematitzar la resolució d'alguns dels apartats de la pràctica, suposarem que s'ha importat el següent data frame:

```
> set.seed(75)
> num1 <- round(rnorm(50, 50, 10), 2)
> num2 <- round(runif(50, 75, 150), 2) + num1
> num3 <- round(rgamma(50, 2) - num1 * runif(50, 0.7, 0.8))
> num4 <- round(runif(50, 0.5, 1) * num1)
> num5 <- round(runif(50, - 1, - 0.75) * (num1) ^ (7 / 8))
> df <- data.frame(id = paste0("id.", 1:50),
+         catbin = sample(LETTERS[1:2], 50, replace = TRUE),
+         catpol = sample(paste0("Group.", 1:4), 50, replace = TRUE),
+         num1, num2, num3, num4, num5)
```

S'anomena cada fila amb el nom de la variable identificadora `id`:

```
> df <- transform(df, row.names = id, id = NULL)
```

Es construeix una funció per fer l'assignació de missings especificada a l'enunciat:

```
> addna <- function(seed, data) {
+   if(!is.numeric(seed))
+     stop("La llavor ha de ser un valor numèric")
+   if(!is.data.frame(data))
+     stop("Les dades introduïdes han de tenir format de data frame")
+   nums <- which(sapply(data, is.numeric))
+   set.seed(seed)
+   for (i in 1:2) {
+     data[, names(nums[i])][sample(nrow(data), i)] <- NA
+   }
+   return(data)
+ }
```

S'obté el següent data frame modificat:

```
> df <- addna(seed = 75, data = df)
```

c) Realitzeu una anàlisi descriptiva univariant de totes les vostres variables (tant categòriques com numèriques), posant especial atenció sobre aquelles variables que continguin algun valor anòmal (*outlier*). Comenteu tots els resultats numèrics i gràfics.

Les anàlisis descriptives realitzades dependran de les dades escollides pels estudiants.

d) Estimeu la matriu de correlacions `matR` amb tota la informació disponible de les vostres variables numèriques, obtenint el nivell de significació d'aquelles correlacions que considereu més rellevants. A continuació realitzeu una anàlisi descriptiva multivariant entre aquelles variables (del tipus que siguin) que més poden contribuir a comprendre la informació. Comenteu tots els resultats numèrics i gràfics.

Càlcul de la matriu de correlacions de Pearson amb tota la informació disponible:

```
> matR <- round(cor(df[, nums], use = "pairwise.complete.obs"), 3)
> matR
```

```
       num1   num2   num3   num4   num5
num1  1.000  0.492 -0.948  0.624 -0.902
num2  0.492  1.000 -0.386  0.290 -0.439
num3 -0.948 -0.386  1.000 -0.550  0.871
num4  0.624  0.290 -0.550  1.000 -0.569
num5 -0.902 -0.439  0.871 -0.569  1.000
```

A continuació s'identifiquen els parells de variables amb les correlacions lineals més fortes (independentment del seu signe), explicant el significat de la correlació estimada i obtenint el seu nivell de significació.

La resta d'anàlisis descriptives multivariants dependran de les dades escollides.

e) Redacteu un apartat de conclusions que inclogui els principals resultats obtinguts a la vostra anàlisi descriptiva. Realitzeu un judici crític d'aquests resultats, tot esmentant les causes que poden explicar en cada cas el comportament de les variables.

Redacció d'un apartat de conclusions on es comentin aquells resultats que vosaltres trobeu més rellevants.

Exercici 2 (4.5 punts)

Programeu una funció que faci el següent donat un *data frame* (`dades`) i el nom d'una de les seves variables categòriques (`cvar`) com a arguments principals:

1. Comprovar que `dades` sigui un *data frame*. En cas contrari, la funció ha d'avortar la seva execució tornant un missatge d'error en català o castellà.
2. Comprovar que `dades` contingui una variable categòrica amb nom `cvar`. En cas contrari, la funció ha d'avortar la seva execució tornant un missatge d'error en català o castellà.
3. Comprovar que `dades` contingui almenys una variable numèrica. En cas contrari, la funció ha d'avortar la seva execució tornant un missatge d'error en català o castellà.
4. Tornar la següent informació sobre el *data frame* `dades`:
   (a) Nombre de files i columnes,
   (b) Taula de freqüències dels tipus de variables,
   (c) Nombre de *missings* per variable,
   (d) Fila amb més *missings*.
5. Calcular diferents indicadors numèrics (mitjana, mediana, desviació estàndard, etc.) de totes les variables numèriques per a les diferents categories de `cvar`.
6. De forma opcional, dibuixar gràfics de mosaics per a la resta de variables categòriques per representar les distribucions condicionals d'aquestes variables en funció de `cvar`.
7. De forma opcional, guardar tots aquests gràfics en un sol document pdf.

Apliqueu la funció a vostres dades de l'Exercici 1 i comenteu la sortida de la funció.

Solució

Una possible solució és la funció a continuació, els arguments de la qual són:

`dades:` Un *data frame*.
`cvar:` Nom de la variable categòrica. Ha de ser un caràcter.
`mosaic:` Dibuixar gràfics de mosaics? Valor per defecte: FALSE
`pdf:` Guardar els gràfics en format pdf? Valor per defecte: FALSE. En cas afirmatiu, el nom del fitxer pdf serà mosaics.pdf i no es mostren els gràfics en R.

```
> myFunc <- function(dades, cvar, mosaic = F, pdf = F) {
+   # Apartat 1
+   if (!is.data.frame(dades)) {
+     stop("Prengui nota: dades ha de ser un data frame!")
+   }
+
+   # Apartat 2
+   kats <- which(sapply(dades, function(x) is.factor(x) | is.character(x)))
+   if (!cvar %in% names(kats)) {
+     stop("Recordi: cvar ha de ser una variable categòrica del data frame!")
+   }
+
+   # Apartat 3
+   if (!any(sapply(dades, is.numeric))) {
+     stop("Error: dades ha de contenir almenys una variable numèrica!")
+   }
+
+   # Funció auxiliar per subratllar
+   underline <- function(txt) {
+     cat("\n", txt, "\n", sep = "")
+     cat(rep("=", nchar(txt)), "\n", sep = "")
+   }
+
+   # Carrega de paquets
+   require(descr)
+   require(doBy)
+
+   # Apartat 4a: Nombre de files i columnes
+   dims <- dim(dades)
+   underline("Dimensió del data frame")
+   cat(dims[1], " files, ", dims[2], " columnes.\n", sep = "")
+
+   # Apartat 4b: Tipus de variables
+   Tipus <- sapply(dades, class)
+   underline("Tipus de variables")
+   print(freq(Tipus, plot = F))
+
+   # Apartat 4c: Nombre de dades perdudes per variable
+   nas <- colSums(is.na(dades))
+   nas1 <- sort(nas[nas > 0], decreasing = T) # Almenys un missing
+   nas0 <- nas[nas == 0]                      # Cap missing
+   underline("Dades perdudes per variable")
+   if (max(nas) > 0) {
+     cat(paste0(names(nas1), ": ", nas1, collapse = ", "), ".\n\n",
+         sep = "")
+   }
```

```
+   cat("Sense cap missing: "); cat(cat(names(nas0), sep = ", "), ".\n", sep = "")
+
+   # Apartat 4d: Fila amb més dades perdudes
+   rmiss <- rowSums(is.na(dades))
+   underline("Files amb dades perdudes")
+   if (sum(rmiss) == 0) {
+     cat("No hi ha cap fila amb dades perdudes.\n")
+   } else {
+     fls <- which(rmiss == max(rmiss))
+     cat("Fila (Files) amb més dades perdudes: ", paste(fls, collapse = ", "),
+         ".\n", sep = "")
+   }
+   cat("\n")
+
+   # Apartat 5
+   underline("Anàlisis descriptives de les variables numèriques")
+   # Funció auxiliar per calcular tots els indicadors numèrics
+   sumfun <- function(x, ...) {
+     c(m = mean(x, ...), Med = median(x, ...), sd = sd(x, ...),
+       Min = min(x, ...), Max = max(x, ...))
+   }
+   # Variables numèriques
+   nums <- which(sapply(dades, is.numeric))
+
+   dades$cvar <- dades[, cvar]
+   for (i in nums) {
+     dades$vari <- dades[, i]
+     txt <- paste0("Variable: ", names(dades)[i])
+     cat("\n", txt, "\n", sep = "")
+     cat(rep("-", nchar(txt)), "\n", sep = "")
+     sumby <- summaryBy(vari~cvar, dades, FUN = sumfun, na.rm = T)
+     names(sumby) <- c(cvar, "Mean", "Median", "SD", "Min.", "Max.")
+     print(sumby, digits = 3)
+     dades$vari <- NULL
+   }
+   dades$cvar <- NULL
+
+   # Apartats 6 i 7: Els gràfics de mosaics (opcionals)
+   if (mosaic) {
+     cat("\n")
+     if (length(kats) == 1) {
+       stop("No hi ha altres variables categòriques per fer un gràfic de mosaics!")
+     }
+
+     # Noms de les variables categòriques diferents a cvar
+     kats <- kats[-which(names(kats) == cvar)]
```

```
+
+     ncols <- length(unique(dades[, cvar]))
+     if (pdf) {
+       # Els gràfics de mosaics en format pdf
+       pdf("mosaics.pdf", width = 8)
+       par(las = 1, font.lab = 2, font.axis = 2)
+       for (k in kats) {
+         mosaicplot(dades[, cvar]~dades[, k], col = 1:ncols, main = NULL,
+                    xlab = cvar, ylab = "", cex.axis = 1)
+         title(paste("Variable", names(dades)[k], "en funció de", cvar))
+       }
+       dev.off()
+     } else{
+       # Els gràfics de mosaics s'obren en R
+       for (k in kats) {
+         windows(width = 8)
+         par(las = 1, font.lab = 2, font.axis = 2)
+         mosaicplot(dades[, cvar]~dades[, k], col = 1:ncols, main = NULL,
+                    xlab = cvar, ylab = "", cex.axis = 1)
+         title(paste("Variable", names(dades)[k], "en funció de", cvar))
+       }
+     }
+   }
+ }
```

A continuació es mostren alguns exemples. El gràfic de mosaics de l'exemple 4 es mostra a la Figura 1 (Pàgina 11).

```
> # Les dades per als exemples 1 a 4
> library(Hmisc)
> states <- data.frame(state.x77, Region = state.region)
> states["Alabama", 1:2] <- states["Wyoming", 1] <- NA
> states$Income2 <- cut2(states$Income, c(4000, 4500, 5000))
> levels(states$Income2)[c(1, 4)] <- c("< 4000", ">= 5000")
> head(states, 5)

           Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
Alabama            NA     NA        2.1    69.05   15.1    41.3    20  50708
Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
California      21198   5114        1.1    71.71   10.3    62.6    20 156361
           Region     Income2
Alabama     South        <NA>
Alaska       West     >= 5000
Arizona      West [4500,5000)
Arkansas    South      < 4000
California   West     >= 5000
```

```
> summary(states)

   Population        Income       Illiteracy       Life.Exp
 Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96
 1st Qu.: 1122   1st Qu.:4022   1st Qu.:0.625   1st Qu.:70.12
 Median : 2838   Median :4530   Median :0.950   Median :70.67
 Mean   : 4340   Mean   :4452   Mean   :1.170   Mean   :70.88
 3rd Qu.: 5064   3rd Qu.:4815   3rd Qu.:1.575   3rd Qu.:71.89
 Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60
 NA's   :2       NA's   :1
     Murder          HS.Grad          Frost             Area
 Min.   : 1.400   Min.   :37.80   Min.   :  0.00   Min.   :  1049
 1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985
 Median : 6.850   Median :53.25   Median :114.50   Median : 54277
 Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean   : 70736
 3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81163
 Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432

           Region          Income2
 Northeast    : 9   < 4000     :12
 South        :16   [4000,4500):11
 North Central:12   [4500,5000):18
 West         :13   >= 5000    : 8
                    NA's       : 1

> # Exemple 1
> myFunc(myFunc)

Error in myFunc(myFunc) : Prengui nota: dades ha de ser un data frame!

> # Exemple 2
> myFunc(states, "Regio")

Error in myFunc(states, "Regio") :
  Recordi: cvar ha de ser una variable categòrica del data frame!

> # Exemple 3
> myFunc(states[, c("Region", "Income2")], "Region")

Error in myFunc(states[, c("Region", "Income2")], "Region") :
  Error: dades ha de contenir almenys una variable numèrica!

> # Exemple 4
> myFunc(states, "Region", mosaic = T, pdf = T)

Dimensió del data frame
=======================
50 files, 10 columnes.

Tipus de variables
==================
```

```
Tipus
        Frequency Percent
factor          2      20
numeric         8      80
Total          10     100

Dades perdudes per variable
===========================
Population: 2, Income: 1, Income2: 1.

Sense cap missing: Illiteracy, Life.Exp, Murder, HS.Grad, Frost, Area, Region.

Files amb dades perdudes
========================
Fila (Files) amb més dades perdudes: 1.


Anàlisis descriptives de les variables numèriques
=================================================

Variable: Population
--------------------
         Region Mean Median   SD Min.  Max.
1     Northeast 5495   3100 6080  472 18076
2         South 4248   3806 2872  579 12237
3 North Central 4803   4255 3703  637 11197
4          West 3127   1174 5772  365 21198

Variable: Income
----------------
         Region Mean Median  SD Min. Max.
1     Northeast 4570   4558 559 3694 5348
2         South 4038   3875 617 3098 5299
3 North Central 4611   4594 283 4167 5107
4          West 4703   4660 664 3601 6315

Variable: Illiteracy
--------------------
         Region Mean Median    SD Min. Max.
1     Northeast 1.00   1.10 0.278  0.6  1.4
2         South 1.74   1.75 0.552  0.9  2.8
3 North Central 0.70   0.70 0.141  0.5  0.9
4          West 1.02   0.60 0.608  0.5  2.2

Variable: Life.Exp
------------------
         Region Mean Median    SD Min. Max.
1     Northeast 71.3   71.2 0.744 70.4 72.5
2         South 69.7   70.1 1.022 68.0 71.4
3 North Central 71.8   72.3 1.037 70.1 73.0
4          West 71.2   71.7 1.352 69.0 73.6
```

```
Variable: Murder
----------------
         Region  Mean Median   SD Min. Max.
1     Northeast  4.72   3.30 2.67  2.4 10.9
2         South 10.58  10.85 2.63  6.2 15.1
3 North Central  5.28   3.75 3.57  1.4 11.1
4          West  7.22   6.80 2.68  4.2 11.5

Variable: HS.Grad
-----------------
         Region Mean Median   SD Min. Max.
1     Northeast 54.0   54.7 3.93 46.4 58.5
2         South 44.3   41.7 5.74 37.8 54.6
3 North Central 54.5   53.2 3.62 48.8 59.9
4          West 62.0   62.6 3.50 55.2 67.3

Variable: Frost
---------------
         Region  Mean Median   SD Min. Max.
1     Northeast 132.8  127.0 30.9   82  174
2         South  64.6   67.5 31.3   11  103
3 North Central 138.8  133.0 23.9  108  186
4          West 102.2  126.0 68.9    0  188

Variable: Area
--------------
         Region   Mean Median     SD  Min.   Max.
1     Northeast  18141   9027  18076  1049  47831
2         South  54605  46113  57965  1982 262134
3 North Central  62652  62906  14967 36097  81787
4          West 134463 103766 134982  6425 566432


> # Les dades per a l'exemple 5
> head(iris)

  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

> # Exemple 5
> myFunc(iris, "Species", mosaic = T)


Dimensió del data frame
=======================
150 files, 5 columnes.
```

```
Tipus de variables
==================
Tipus
        Frequency Percent
factor          1      20
numeric         4      80
Total           5     100

Dades perdudes per variable
===========================
Sense cap missing: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, Species.

Files amb dades perdudes
========================
No hi ha cap fila amb dades perdudes.


Anàlisis descriptives de les variables numèriques
=================================================

Variable: Sepal.Length
----------------------
     Species Mean Median    SD Min. Max.
1     setosa 5.01    5.0 0.352  4.3  5.8
2 versicolor 5.94    5.9 0.516  4.9  7.0
3  virginica 6.59    6.5 0.636  4.9  7.9

Variable: Sepal.Width
---------------------
     Species Mean Median    SD Min. Max.
1     setosa 3.43    3.4 0.379  2.3  4.4
2 versicolor 2.77    2.8 0.314  2.0  3.4
3  virginica 2.97    3.0 0.322  2.2  3.8

Variable: Petal.Length
----------------------
     Species Mean Median    SD Min. Max.
1     setosa 1.46   1.50 0.174  1.0  1.9
2 versicolor 4.26   4.35 0.470  3.0  5.1
3  virginica 5.55   5.55 0.552  4.5  6.9

Variable: Petal.Width
---------------------
     Species  Mean Median    SD Min. Max.
1     setosa 0.246    0.2 0.105  0.1  0.6
2 versicolor 1.326    1.3 0.198  1.0  1.8
3  virginica 2.026    2.0 0.275  1.4  2.5

Error in myFunc(iris, "Species", mosaic = T) :
  No hi ha altres variables categòriques per fer un gràfic de mosaics!
```

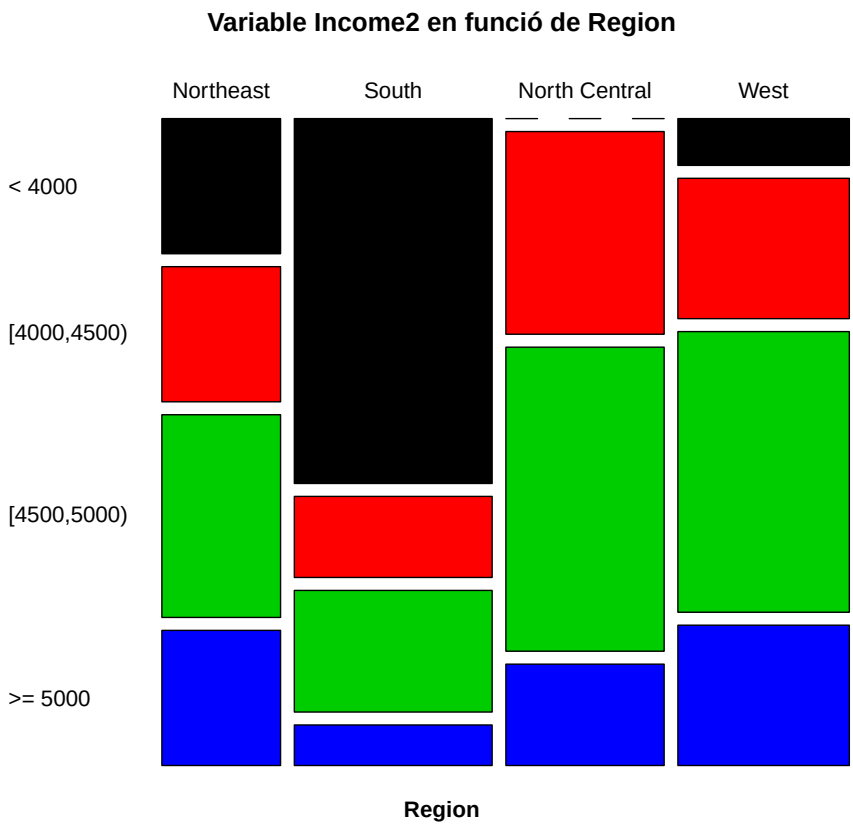

Figura 1: Gràfic de mosaics de la variable Income2 produït amb la funció `myFunc`.