Grau interuniversitari (UB-UPC) d'Estadística

Software Estadístic: Solució del Test 2 amb R

Exercici 1 ($0.8 + 0.5 + 0.5 + 0.6 + 0.8 + 0.6 + 0.7 + 0.5 = 5$ punts)
Utilitzem un cop més les dades dels 50 estats dels Estats Units dels anys 70, que es troben al fitxer tipus text `USstates77.txt`. Les variables son les següents:

`state:` Estat
`pop:` Població (en milers)
`area:` Superfície (en milles quadrades)
`income:` Salari mitjà (en dòlars)
`illit:` Taxa d'analfabetisme (en % de la població)
`lifexp:` Esperança de vida al néixer

a) Importeu les dades a un *data frame* amb nom `states77`.

```
> states77 <- read.table('USstates77.txt', header=T, na.strings=c(NA, '*', '**'))
> head(states77)
```

```
       state   pop   area income illit lifexp
1    Alabama  3615  50708   3624   2.1  69.05
2     Alaska    NA 566432   6315   1.5  69.31
3    Arizona  2212 113417   4530    NA  70.55
4   Arkansas  2110  51945   3378    NA  70.66
5 California 21198 156361   5114   1.1  71.71
6   Colorado  2541 103766   4884   0.7  72.06
```

```
> summary(states77)
```

```
        state          pop             area            income
 Alabama   : 1   Min.   :  376   Min.   :  1049   Min.   :3098
 Alaska    : 1   1st Qu.: 1144   1st Qu.: 36985   1st Qu.:3993
 Arizona   : 1   Median : 2861   Median : 54277   Median :4519
 Arkansas  : 1   Mean   : 4326   Mean   : 70736   Mean   :4436
 California: 1   3rd Qu.: 4981   3rd Qu.: 81163   3rd Qu.:4814
 Colorado  : 1   Max.   :21198   Max.   :566432   Max.   :6315
 (Other)   :44   NA's   :1
     illit          lifexp
 Min.   :0.500   Min.   :67.96
 1st Qu.:0.600   1st Qu.:70.12
 Median :0.900   Median :70.69
 Mean   :1.142   Mean   :70.89
 3rd Qu.:1.425   3rd Qu.:71.89
 Max.   :2.800   Max.   :73.60
 NA's   :2       NA's   :3
```

b) Utilitzeu el nom de l'estat com a identificador de fila (*rowname*) i esborreu la variable `state`.

```
> states77 <- transform(states77, row.names=state, state = NULL)
> head(states77)
```

```
             pop   area income illit lifexp
Alabama     3615  50708   3624   2.1  69.05
Alaska        NA 566432   6315   1.5  69.31
Arizona     2212 113417   4530    NA  70.55
Arkansas    2110  51945   3378    NA  70.66
California 21198 156361   5114   1.1  71.71
Colorado    2541 103766   4884   0.7  72.06
```

c) El vector `state.region` del paquet `datasets` és un factor que conté la regió corresponent a cadascun dels estats. Afegiu-lo com a nova variable a `states77` amb nom `reg`.

```
> states77 <- data.frame(states77, reg=state.region)
> head(states77)
```

```
             pop   area income illit lifexp   reg
Alabama     3615  50708   3624   2.1  69.05 South
Alaska        NA 566432   6315   1.5  69.31  West
Arizona     2212 113417   4530    NA  70.55  West
Arkansas    2110  51945   3378    NA  70.66 South
California 21198 156361   5114   1.1  71.71  West
Colorado    2541 103766   4884   0.7  72.06  West
```

d) Afegiu com a nova variable a `states77` la densitat de població (en habitants per milla quadrada). La nova variable ha de tenir una decimal.

```
> states77$dens <- with(states77, round(pop*1000/area, 1))
> head(states77)
```

```
             pop   area income illit lifexp   reg  dens
Alabama     3615  50708   3624   2.1  69.05 South  71.3
Alaska        NA 566432   6315   1.5  69.31  West    NA
Arizona     2212 113417   4530    NA  70.55  West  19.5
Arkansas    2110  51945   3378    NA  70.66 South  40.6
California 21198 156361   5114   1.1  71.71  West 135.6
Colorado    2541 103766   4884   0.7  72.06  West  24.5
```

```
> summary(states77$dens)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    3.9    27.9    74.7   152.3   147.2   975.0       1
```

e) Creeu la nova variable ordinal `incoCat` que ha de classificar la variable `income` en tres categories amb nombres d'observacions semblants a cada categoria. Poseu-los etiquetes adients a les tres categories.

```
> quantile(states77$income, 1:3/3)

33.33333% 66.66667%      100%
 4174.000  4692.333  6315.000

> states77$incoCat <- cut(states77$income, c(0, 4200, 4700, 6500),
+   labels = c("<= 4200", "4200--4700", "> 4700"))
> table(states77$incoCat)

   <= 4200 4200--4700     > 4700
        18          15          17
```

f) Feu un gràfic de mosaics que mostri la distribució condicionada de la variable `incoCat` per regió. El gràfic ha de tenir un títol i etiquetes d'eixos adients. Guardeu-lo en format jpg.

```
> # Figura 1
> windows(width=8)
> par(las=1, font.lab=2, font.axis=2, cex.lab=1.2)
> mosaicplot(reg~incoCat, states77, xlab='Regio', ylab='Salari mitjà', col=2:4,
+   main = 'Salari mitjà per regió', cex.axis=1.1)

> savePlot('Mosaics', 'jpg')
```

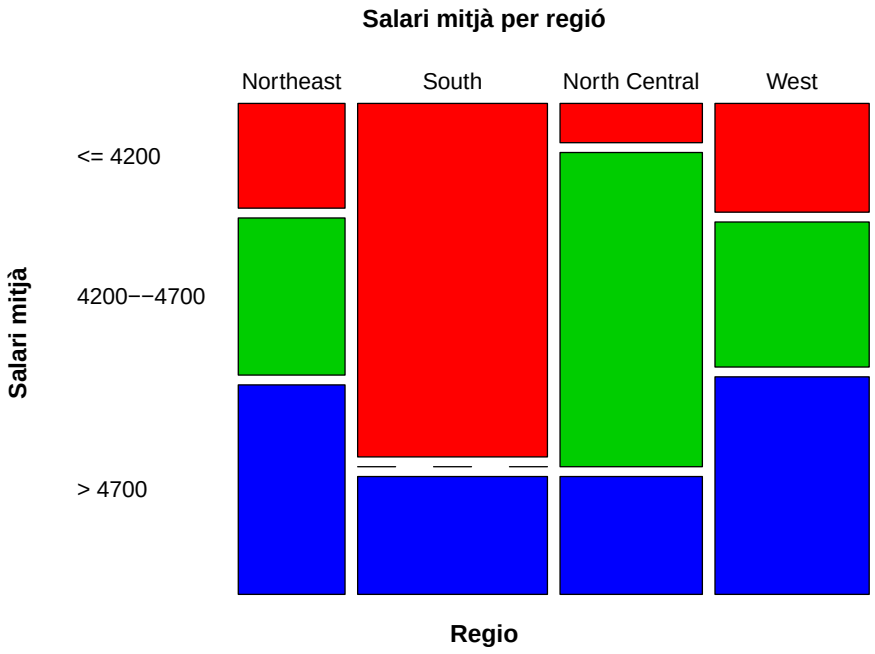

Figura 1: Gràfic de mosaics corresponent a l'apartat **f**

g) Quina regió té la mediana de l'esperança de vida més alta i quin és aquest valor?

```
> (mediana <- with(states77, tapply(lifexp, reg, median, na.rm=T)))
```

```
    Northeast         South North Central          West
       71.230        70.080        72.080        71.715
```

```
> mediana[which.max(mediana)]
```

```
North Central
        72.08
```

h) Guardeu states77 en un àrea de treball d'R.

```
> save(states77, file="states77.RData")
```

Exercici 2 (2,5 punts)

Programeu un bucle amb `for` per dibuixar dos *boxplots* per cada variable numèrica: el primer en funció de la variable `reg` i el segon en funció de la variable `incoCat`. Els dos gràfics de cada variable s'han de dibuixar en una finestra gràfica que sigui el doble d'ample que d'alt i utilitzeu diferents colors per a cada categoria i poseu un títol global als dos gràfics. A més a més, tots els gràfics s'han de guardar en fitxers pdf amb els noms de les variables numèriques.

```
> nums <- which(sapply(states77, is.numeric))
> for (i in nums){
+   pdf(paste0(names(states77)[i], '.pdf'), width=12, height=6)
+   par(las=1, mfrow=c(1, 2), font.lab=2, font.axis=2, oma=c(0, 0, 1, 0),
+     mar=c(5, 4, 2, 2), cex.axis=1.3)
+   boxplot(states77[, i]~reg, states77, col=2:5, pch=16)
+   boxplot(states77[, i]~incoCat, states77, col=2:4, pch=16)
+   title(paste('Variable', names(states77)[i]), outer=T, cex.main=1.5)
+   dev.off()
+ }
```

Exercici 3 (2,5 punts)

Programeu una funció que, donats un vector numèric $x = (x_1, \ldots, x_n)'$ i un altre de pesos $w = (w_1, \ldots, w_n)'$, calculi la mitjana ponderada segons la fórmula següent:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i \cdot w_i}{\sum_{i=1}^{n} w_i}.$$

L'heu de programar de tal manera

- que la funció torni un missatge d'error, si el vector `x` o el vector `w` no són numèrics,
- que torni el valor $\bar{x}_w$ amb d decimals prenent $d = 2$ com a valor per defecte.

Apliqueu la vostra funció a dos vectors de longitud 10 qualsevol.
Nota: No cal considerar el cas que x o w tinguin valors perduts, se suposa que no en tenen.

Solució:

```
> wmeans <- function(x, w, d=2){
+    stopifnot(is.numeric(x) & is.numeric(w))
+    xw <- round(sum(x*w)/sum(w), d)
+    return(xw)
+ }
```

Alguns exemples:

```
> set.seed(2910)
> (x <- rpois(10, 20))
 [1] 24 23 20 22 20 24 20 25 25 19
> (w <- sample(1:10, 10, replace = T))
 [1] 10 10  9  8  3  8  4  1  8  9
> mean(x)
[1] 22.2
> wmeans(x, w)
[1] 22.2
> wmeans(x, w, d=1)
[1] 22.2
> wmeans(x, 'A')
Error: is.numeric(x) & is.numeric(w) is not TRUE
```