Instituto Superior de Agronomia
Modelos Matemáticos e Aplicações – 2017/2018

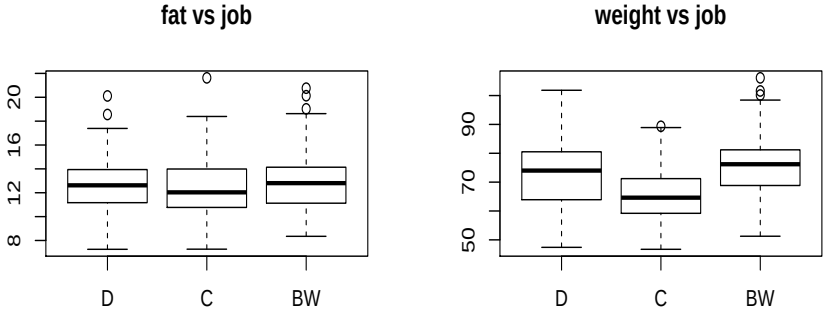
Teste do Módulo I– duração: 2h30 14 March, 2018

1. Consider the data `diet` from the package `Epi` in R, referring to a study that aimed to analyze the existence of some association between diet and the incidence of coronary heart disease (CHD). The `diet` data frame has 337 rows and 15 columns. Let's consider only the following variables:

`fat`: fat intake (g/day), a numeric vector
`weight`: (kg), a numeric vector
`chd`: CHD event, a numeric vector (1=CHD event, 0=no event)
`job`: occupation, a factor with levels Driver Conductor Bank worker

A few commands have been made in R, the results appear in the output that is in the Appendix and that you should use to answer the following questions:

a) Complete the values referenced by **A**, **B**, **C** and **D**, missing in the output.
b) Sketch the Histogram, which is identified in the output, but was not drawn.
c) Sketch a boxplot for the variable `fat`. Clearly identify the limits of the boxplot.
d) Comment, comparing, the 2 groups of *boxplots* shown below.
e) Obtain a 95% confidence interval for the proportion of individuals who **are not referred as having** coronary disease. Comment.
f) Using the results presented in the output compare the mean value of the variable `fat` between the two groups of occurrence or not of coronary disease. Justify.



2. Agronomists are developing an improved variety of green peppers. Supermarket managers have indicated customers are not likely to purchase green peppers weighing less then 45 g. The current variety of green plants produces green peppers that weigh 48 grms on average, but 13% weigh less than 45 grams. Assume the weight of the current variety of green peppers follows a normal distribution.

a) What is the standard deviation of the weights of the current variety of green peppers?
b) The agronomists want to reduce the frequency of green pepper weighing less than 45 grams to no more than 5%. One way to reduce the frequency of underweight green peppers is to increase the weight of the green peppers. If the standard deviation remains the same what mean value should the agronomists target as a goal? (if you have not resolved a) consider $\sigma = 2.6$)
c) The agronomists produced the new variety of green peppers, in order to guarantee a mean weight of 50 grams with a standard deviation of 2.5 grams. A supermarket customer buys 25. What is the probability that the average weight of these peppers is less than 49 grams?

3. Suppose that the number of attempts a basketball player makes to hit the basket is a random variable, X, with the following probability mass function:

$$P[X = x|\theta) = \frac{1}{\theta}\left(1 - \frac{1}{\theta}\right)^{x-1}, \quad \text{with} \quad x = 1, 2, 3, ..., \quad (\theta > 1)$$

Remark: It is known that $E[X] = \theta$ and $Var[X] = \theta(\theta - 1)$.

Consider that you have a random sample of size n, $(X_1, X_2, .., X_n)$, associated with X.

a) Obtain the etimator for θ by the method of moments.

b) Verify if the moment estimator obtained in (a) is an unbiased estimator of θ and calculate its Mean Square Error.

c) Get the maximum likelihood estimator for θ.

d) Consider now the observed sample, with size 20, extracted from that population, with which the following calculations were performed:

```
> sample
 [1] 2 7 1 1 1 3 5 6 4 5 3 2 4 6 6 6 1 7 5 1
> sum(sample)                  > sum(sample*sample)
[1] 76                          [1] 380
```

i) Obtain an estimate for θ.

ii) Obtain an estimate for $P[X = 2]$, i.e., the probability that the player hits the second attempt.
Remark: the maximum likelihood estimator has a property, called the invariance, which establishes - textit The maximum likelihood estimator of $g(\theta)$ is $g(\hat{\Theta})$, if $g(.)$is a biunivocal function of θ.

4. Indicate, justifying, whether the following statements are **True** or **False**. **Correct the false ones**.

a) If $X \frown Poisson(10)$ we can calculate $P[X > 8]$, in the R doing `dpois(8,10)`.

b) Suppose that $Y \frown Binomial(n, p)$. Then $P[0 \leq Y \leq n] = 1$

c) Let be $X \frown N(0, 2)$ and $Y \frown N(1, 1)$, with X and Y independent. Then $2X - Y \frown N(1, 5)$

d) Let $(X_1, ..., X_n)$ be a random sample of size n , coming from a population with mean value μ and variance $\sigma^2 < +\infty$. For a large value of n we have $P[S_n \leq n\mu] \approx 1/2$, with $S_n = \sum_{i=1}^{n} X_i$.

5. The sulfur dioxide (SO_2) concentration X (in parts per million, or ppm) in a certain city is postulated to have a gamma density function. As you know the general form of the gamma density function was presented with two parameters α e β, ($\alpha > 0$, $\beta > 0$. Let us consider, to simplify, that $\alpha = 1/2$, remaining β unknown. So the density function is

$$f(x|\beta) = \begin{cases} \frac{1}{\sqrt{\beta\pi x}}\, e^{-x/\beta} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Let $(X_1, ..., X_n)$ be a random sample of size n, coming from that density, where $(x_1, ..., x_n)$ is the corresponding observed sample.

a) Using the Central Limit Theorem obtain an approximate distribution for $\overline{X}$.

b) Recalling that if a r.v. V is such that $V \sim Normal(0, 1)$, one can write

$$P[-1.96 < V < 1.96] \approx 0.95,$$

obtain an asymptotic interval at 95% confidence for the β parameter.

APPENDIX I

```
########  Question 1   ########

> library(Epi)
> data(diet)
> dim(diet)
[1] 337  15

> names(diet)
 [1] "id"         "doe"        "dox"        "dob"        "y"
 [6] "fail"       "job"        "month"      "energy"     "height"
[11] "weight"     "fat"        "fibre"      "energy.grp" "chd"


> hist(diet$fat,breaks=c(5,10,12,15,18,22),plot=F) #Histogram
$breaks
[1]  5 10 12 15 18 22

$counts
[1]  34 106 146  41  10

$density
[1] 0.020178042 0.157270030 0.144411474 0.040553907 0.007418398

$mids
[1]  A  11.0  13.5  16.5  20.0

>
> sort(diet$fat)
[1]  7.260000  7.272000  7.578000  8.302000  8.348001  8.536000  8.577000
. . .
[323] 17.388000 17.409000 17.445000 17.639000 17.769000 18.389000 18.391000
[330] 18.558000 18.588000 18.628000 19.032000 20.112000 20.132000 20.763000
[337] 21.629000


> summary(diet$fat)
   Min.  1st Qu.   Median    Mean   3rd Qu.    Max.
   B     11.12     12.59     12.75   14.01     21.63


> par(mfrow=c(1,2))
> boxplot(fat~job,data=diet,main="fat vs job",names=c("D","C","BW"))
> boxplot(weight~job,data=diet,main="weight vs job",names=c("D","C","BW"))


> diet$fat[diet$chd==0]    # C - write briefly what results from this command
> length(diet$fat[diet$chd==0])
[1] 291
> length(diet$fat[diet$chd==1])
[1] D
```

```
> shapiro.test(diet$fat[diet$chd==0])

        Shapiro-Wilk normality test

data:  diet$fat[diet$chd == 0]
W = 0.9757, p-value = 7.604e-05


> shapiro.test(diet$fat[diet$chd==1])

        Shapiro-Wilk normality test

data:  diet$fat[diet$chd == 1]
W = 0.9743, p-value = 0.3947


> t.test(fat~chd,data=diet)

        Welch Two Sample t-test

data:  fat by chd
t = 2.9536, df = 62.484, p-value = 0.004423
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.337497    1.750317
sample estimates:
mean in group 0 mean in group 1
       12.88791        11.84400


> t.test(fat~chd,data=diet,alternative="greater")

        Welch Two Sample t-test

data:  fat by chd
t = 2.9536, df = 62.484, p-value = 0.002211
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.4537988       Inf
sample estimates:
mean in group 0 mean in group 1
       12.88791        11.84400



> t.test(fat~chd,data=diet,alternative="less")

        Welch Two Sample t-test

data:  fat by chd
t = 2.9536, df = 62.484, p-value = 0.9978
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 1.634016
sample estimates:
mean in group 0 mean in group 1
       12.88791        11.84400

>
```