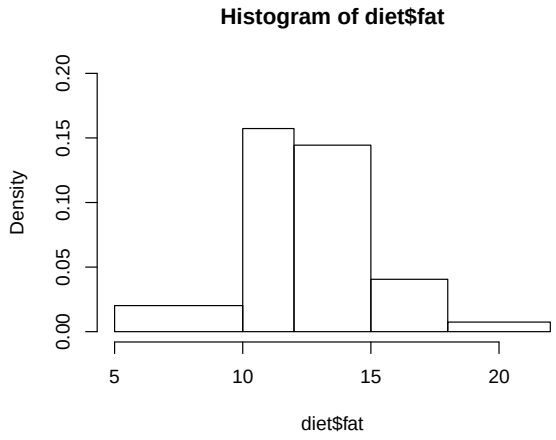
Instituto Superior de Agronomia
Modelos Matemáticos – 2017/2018

Topics of a resolution of the Test – Module I 14 de Março de 2018

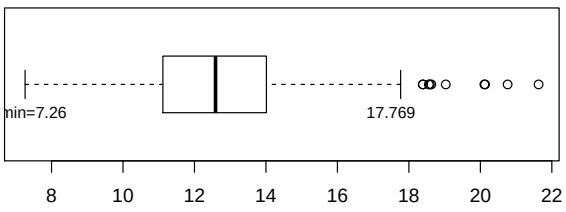
1. a) Let's then indicate the missing values in the *output* (it's worthwhile to enter and try to execute the commands, to see and understand the results)

```
A  (middle point of class 1) 7.5
B  7.26
C  Here we obtain a sub-vector of the variable fat, when chd is set to 0;
 it has 291 observations
D  337-291=46
```

b) See that this histogram has classes of different amplitudes, so the height of each classes is given in *density*=relative frequency/amplitude.



c) To construct the *boxplot* the upper and lower barriers must be calculated to see if there exist *outliers*.
$BI = Q_1 - 1.5(Q_3 - Q_1)$ $BS = Q_3 + 1.5(Q_3 - Q_1)$, como $Q_1 = 11.12$, $Q_3 = 14.01$ tem-se $BI = 6.785$ e $BS = 18.345$, so there are no observed values lower than the lower barrier (so there are no left-tail outliers), but values above 18.345 are all *outliers*. See below the *boxplot*.



d) We see that the boxplots of *fat* against *job* all present *outliers*, but still some homogeneity. In the boxplots of *weight* against *job* it is verified that for the Driver the weights are lower, for Driver there is greater dispersion and there is no occurrence of *outliers*. Is at *Bankworker* that more *outliers* appeared.

e) An estimate of the proportion of individuals in which coronary disease **will not occur** is given by $p^\star = 291/337 = 0.8635$

As $n = 337$ is large, we can consider the asymptotic confidence interval at 95% for the proportion, p, of individuals in whom there will be no occurrence of coronary disease

$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \iff 0.8635 - 1.96\sqrt{\frac{0.8635 \times 0.1365}{337}} < p < 0.8635 + 1.96\sqrt{\frac{0.8635 \times 0.1365}{337}}$

donde o IC a 95% para p é]0.8268462 , 0.9001568[

f) As it was observed that the mean of *fat* is 12.88791 for 'group 0' and for 'group 1' is11.844; it makes sense to investigate the following hypotheses:

$H_0 : \mu_0 = \mu_1 \quad vs \quad H_1 : \mu_0 > \mu_1,$ where μ_0 denotes the mean value of *fat* in the group for which $chd = 0$ e μ_1, the mean value of *fat* in the group for which $chd = 1$. Let us consider the level of significance $\alpha = 0.05$

Since the samples have a large dimension, $n_0 = 291$ and $n_1 = 46$, we can consider the approximation to normal to be good. Samples are independent.

Nota: Os testes

```
> 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
```

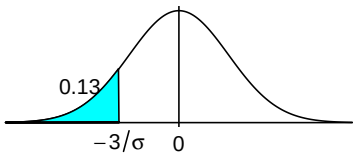
would lead to rejection of the normality hypothesis of *fat*, when $chd = 0$, but since n_0 is too large, we can use the test based on normal approximation

```
> 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
```

As $p - value = 0.00221$, is less than α, we are led to reject H_0, so with that *p-value* we can say that mean value of *fat* is larger in the group that did not have coronary diseases.

2. Let X be a r.v. which designates the weight of each pod of green pepper; $X \frown Normal(48, \sigma)$. σ is unknown. We know that $P[X < 45] = 0.13$

a) Ora $P[X < 45] = 0.13 \iff P\left[\frac{X-48}{\sigma} < \frac{45-48}{\sigma}\right] = 0.13 \iff \Phi\left(\frac{45-48}{\sigma}\right) = 0.13 \iff \Phi\left(\frac{-3}{\sigma}\right) = 0.13$, so $\frac{-3}{\sigma}$ is the quantile of probability 0.13, in normal standard, see figure



It canj be calculated in R

$\frac{-3}{\sigma} = qnorm(0.13) = -1.126391 \iff \sigma = 2.6634$ gramas

b) We want that $P[X < 45] = 0.05$, with $\sigma = 2.6$ and changing the mean value, so μ unknown, i.e., $P\left[\frac{X-\mu}{2.6} < \frac{45-\mu}{2.6}\right] = 0.05 \iff \Phi\left(\frac{45-\mu}{2.6}\right) = 0.05 \iff \frac{45-\mu}{2.6} = qnorm(0.05) \iff \frac{45-\mu}{2.6} = -1.645 \iff \mu = 49.277$ gramas.

c) We then have a sample of $n = 25$ green peppers of a variety whose weight $X \frown Normal(50, 2.5)$. One has for the average weight of 25 peppers $\overline{X} \frown Normal(50, 2.5/\sqrt{25})$, i.e. $\overline{X} \frown Normal(50, 0.5)$
$P[\overline{X} < 49] = pnorm(49, 50, 0.5) = 0.02275$

3. a) Temos $\theta > 1$ e $E[X] = \theta$ e $Var[X] = \theta(\theta - 1)$.
The method of moments establishes that the estimator is the solution of

$$E[X] = \frac{\sum X_i}{n} \iff \theta = \overline{X}$$

The estimator is then $\Theta^* = \overline{X}$

b) Um estimator T de θ is unbiased if and only if $E[T] = \theta$.
We know that $E[\Theta^*] = E[\overline{X}] = \mu = \theta$, so Θ^* is an unbiased estimator of θ.
If an estimator is unbiased its Mean Square Error, $EQM[\Theta^*] = Var[\Theta^*]$, because de bias is zero
Ora se $\Theta^* = \overline{X}$ então $Var[\Theta^*] = Var[\overline{X}] = Var[X]/n = \theta(\theta - 1)/n$

c) To obtain the Maximum Likelihood Estimator, it is necessary to obtain the likelihood function:

$$L(\theta|x_1, \ldots, x_n) = \Pi_{i=1}^n \left[\frac{1}{\theta}\left(1 - \frac{1}{\theta}\right)^{x_i - 1}\right] = \left(\frac{1}{\theta}\right)^n \left(1 - \frac{1}{\theta}\right)^{\sum x_i - n}$$

The log of the likelihood function is:

$$\log L(\theta|x_1, \ldots, x_n) = n \log(1/\theta) + \left(\sum_{i=1}^n x_i - n\right) \log\left(1 - \frac{1}{\theta}\right)$$

Calculando a derivada:

$$\frac{d \log L}{d\theta} = -\frac{n}{\theta} + \left(\sum_{i=1}^n x_i - n\right) \frac{1/\theta^2}{1 - 1/\theta} = -\frac{n}{\theta} + \left(\sum_{i=1}^n x_i - n\right) \frac{1}{\theta(\theta - 1)}$$

e agora igualando a zero:

$$-\frac{n}{\theta} + \left(\sum_{i=1}^n x_i - n\right) \frac{1}{\theta(\theta - 1)} = 0 \iff -n(\theta - 1) + \sum x_i - n = 0 \iff -\theta + 1 + \overline{x} - 1 = 0 \iff \theta = \overline{x}$$

Logo o estimador de máxima verosiminhança é $\widehat{\Theta} = \overline{X}$

d) i) Since the two estimators are equal an estimate of θ based on the observed sample is $\theta^* = \overline{x} = 3.8$

ii) As we have the formula of calculation of probability has $P[X=2] = \frac{1}{\theta}\left(1-\frac{1}{\theta}\right)$.

In view of the property given, an estimate of maximum likelihood of $P[X=2]$ é $\frac{1}{\widehat{\theta}}\left(1-\frac{1}{\widehat{\theta}}\right)$,

i.e. $\widehat{P}[X=2] = 0.1939$

4. a) It is **False**.

Se $X \frown Poisson(10)$ para calcularmos $P[X>8]$=`ppois(8,10,lower.tail=FALSE)` $\equiv$ `1-ppois(8,10)`.

O comando dado, `dpois(8,10)`, calcula $P[X=8]$.

b) Suponha que $Y \frown Binomial(n,p)$. Então $P[0 \leq Y \leq n] = 1$. É **True**, because it means the sum of the probability of all possible values for X, then equal to 1.

c) Seja $X \frown N(1,2)$ e $Y \frown N(1,1)$, com X e Y independentes.

$2X-Y$ será de facto normal; o **valor médio** é: $2\times1-1=1$ e a **variância** é $4Var[X]+Var[Y] = 4\times4+1=17$, O parâmetro que aparece na lei de $2X-Y$ é o desvio padrão que seria $\sqrt{17} \neq 3$. Logo é **False**.

d) Seja $(X_1,...,X_n)$ é uma amostra aleatória de tamanho n, proveniente de uma população com valor médio μ e variância $\sigma^2 < +\infty$ e n suficientemente elevado, pelo Teorema Limite Central $S_n \sim N(n\mu, \sigma\sqrt{n})$, portanto $P[S_n \leq n\mu] \approx 1/2$ é **True**.

5. X concentration of sulfur dioxide (SO_2) com uma distribuição gama, com $\alpha = 1/2$ e $\beta > 0$ desconhecido, cuja função densidade é então:

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

$(X_1,...,X_n)$ uma amostra aleatória de tamanho n,

a) We know that for $X \frown Gama(\alpha,\beta)$ one has $\mu = E[X] = \alpha\beta$ e $\sigma^2 = Var[X] = \alpha\beta^2$. Então neste caso $\mu = E[X] = \beta/2$ e $\sigma^2 = Var[X] = \beta^2/2$.

By the Central Limit Theorem $\overline{X} \sim Normal(\mu, \sigma/\sqrt{n}) \iff \overline{X} \sim Normal(\beta/2, \beta/\sqrt{2n}) \iff \dfrac{\overline{X}-\beta/2}{\beta/\sqrt{2n}} \sim Normal(0,1)$.

b) Aplicando a sugestão de $V \sim Normal(0,1)$ então $P[-1.96 < V < 1.96] \approx 0.95$ a variável definida atrás temos:

$$P\left[-1.96 < \frac{\overline{X}-\beta/2}{\beta/\sqrt{2n}} < 1.96\right] \approx 0.95 \iff P\left[-1.96 < \sqrt{2n}\,\frac{\overline{X}-\beta/2}{\beta} < 1.96\right] \approx 0.95$$

$$P\left[\frac{-1.96}{\sqrt{2n}} < \frac{\overline{X}}{\beta} - 1/2 < \frac{1.96}{\sqrt{2n}}\right] \approx 0.95 \iff P\left[\frac{-1.96}{\sqrt{2n}} + 1/2 < \frac{\overline{X}}{\beta} < \frac{1.96}{\sqrt{2n}} + 1/2\right] \approx 0.95$$

$$P\left[\frac{\overline{X}}{\frac{1.96}{\sqrt{2n}} + 1/2} < \beta < \frac{\overline{X}}{\frac{-1.96}{\sqrt{2n}} + 1/2}\right] \approx 0.95$$

therefore an asymptotic interval at 95 % confidence for the parameter β is:

$$\frac{\overline{x}}{\frac{1.96}{\sqrt{2n}} + 1/2} < \beta < \frac{\overline{x}}{\frac{-1.96}{\sqrt{2n}} + 1/2}$$