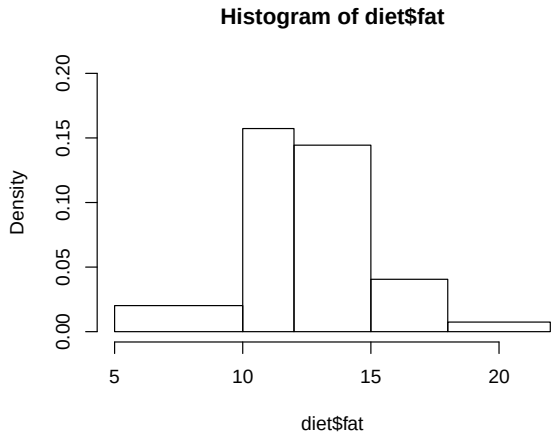
Instituto Superior de Agronomia
Modelos Matemáticos – 2017/2018

Tópicos de uma resolução do Teste do Módulo I 14 de Março de 2018

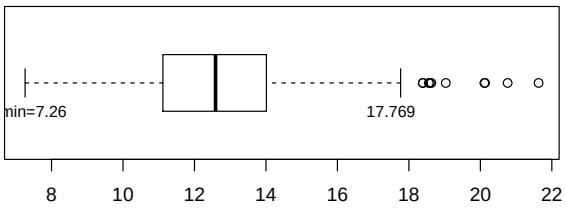
1. a) Vamos então indicar os valores em falta no *output* (vale a pena introduzir no R e tentar executar os comandos, para ver e compreender os resultados)

```
A  (ponto médio da classe 1) 7.5
B  7.26
C  Aqui obtém-se uma sub-vector da variável fat, quando chd toma o valor 0;
     tem  291 observações
D  337-291=46
```

b) Note-se que este histograma tem classes de amplitude diferente, pelo que a altura das classes é dada em *density*=frequência relativa/amplitude.



c) Para construir o *boxplot* devem calcular-se as barreiras superiores e inferiores, para se fazer uma pesquisa de candidatos a *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 $BS = 18.345$, portanto não há valores observados inferiores à barreira inferior (logo não há outliers na cauda esquerda), mas os valores supriores a 18.345 são todos candidatos a outliers. Ver abaixo o *boxplot*.



d) Verifica-se que os boxplots de *fat* como função do *job* todos apresentam *outliers*, mas ainda assim alguma homogeneiddade. Nos boxplots de *weight* como função do *job* verifica-se que para o Conductor os pesos são mais baixos, para Driver há maior dispersão e não há ocorrência de *outliers*. É em *Bankworker* que surgem mais *outliers*.

e) Uma estimativa da proporção de indivíduos em que **não terá havido ocorrência** de doença coronária é dada por $p^\star = 291/337 = 0.8635$

Como $n = 337$ é grande, podemos considerar o intervalo de confiança assintótico a 95% para a proporção, p, de indivíduos em que não terá havido ocorrência de doença coronária,

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

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

f) Como se observou que a média de *fat* é 12.88791 para o 'group 0' e para 'group 1' é 11.844; faz sentido averiguar as seguintes hipóteses:

$H_0 : \mu_0 = \mu_1$ *vs* $H_1 : \mu_0 > \mu_1$, onde μ_0 designa o valor médio de *fat* no grupo em que $chd = 0$ e μ_1, o valor médio de *fat* no grupo em que $chd = 1$. Vamos considerar o nível de significância $\alpha = 0.05$

Como as amostras têm dimensão 'grande', $n_0 = 291$ e $n_1 = 46$ podemos considerar boa a aproximação pela normal. As amostras são independentes.

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
```

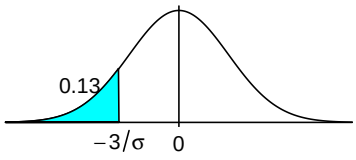
levariam à rejeição da hipótese da normalidade de *fat*, quando *chd=0*, mas como n_0 é muito grande, podemos usar o teste baseado na apoximação pela normal (a não verificação da normalidade é de esperar para amostras desta dimensão). Aqui podemos considerar o comando e respectivo *output*

```
> 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
```

Como $p - value = 0.00221$, é inferior aos nível α, somos levados a rejeitar H_0, portanto com aquele *p-value* podemos dizer que se verifica em média maior valor de *fat* no grupo que não teve doenças coronárias.

2. Seja X a v.a que designa o peso de cada vagem de pimenta verde; $X \frown Normal(48, \sigma)$. Não é dado σ. Sabe-se que $P[X < 45] = 0.13$

a) Ora $P[X < 45] = 0.13 \iff P\left[\frac{X-48}{\sigma} < \frac{45-48}{\sigma}\right] = 0.13 \iff \Phi\left(\frac{45-48}{\sigma}\right) = 0.13 \iff \Phi\left(\frac{-3}{\sigma}\right) = 0.13$, portanto $\frac{-3}{\sigma}$ é o quantil de probabilidade 0.13, na normal standard, ver figura



Uma maneira seria fazer no R
$\frac{-3}{\sigma} = qnorm(0.13) = -1.126391 \iff \sigma = 2.6634$ gramas

b) Pretende-se agora que $P[X < 45] = 0.05$, supondo $\sigma = 2.6$ e agora alterando o valor médio, portanto μ desconhecido, 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) Temos então uma amostra de $n = 25$ pimentas verdes de uma variedade cujo peso $X \frown Normal(50, 2.5)$. Tem-se para a média do peso das 25 pimentas $\overline{X} \frown Normal(50, 2.5/\sqrt{25})$, i.e. $\overline{X} \frown Normal(50, 0.5)$
Pede-se $P[\overline{X} < 49] = pnorm(49, 50, 0.5) = 0.02275$

3. a) Temos $\theta > 1$ e $E[X] = \theta$ e $Var[X] = \theta(\theta - 1)$.
O método dos momentos estabelece que um estimador é a solução de

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

Então o estimador é $\Theta^* = \overline{X}$

b) Um estimador T de θ é centrado se e só se $E[T] = \theta$.
Sabemos que $E[\Theta^*] = E[\overline{X}] = \mu = \theta$, portanto Θ^* é estimador centrado de θ.
Se um estimador é centrado o seu Erro Quadrático Médio, $EQM[\Theta^*] = Var[\Theta^*]$, pois no viés é nulo.
Ora se $\Theta^* = \overline{X}$ então $Var[\Theta^*] = Var[\overline{X}] = Var[X]/n = \theta(\theta - 1)/n$

c) Para obter o Estimador de Máxima Verosimilhança é preciso escrever a função de verosimilhança:

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

O logaritmo da verosimilhança é

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

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:

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

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

d) i) Como os dois estimadores são iguais uma estimativa de θ com base na amostra observada é $\theta^* = \overline{x} = 3.8$

ii) Como temos a fórmula de cálculo da probabilidade tem-se $P[X=2] = \frac{1}{\theta}\left(1-\frac{1}{\theta}\right)$.
Atendendo à propriedade referida, uma estimativa de máxima verosimilhança de $P[X=2]$ é $\frac{1}{\widehat{\theta}}\left(1-\frac{1}{\widehat{\theta}}\right)$, i.e. $\widehat{P}[X=2] = 0.1939$

4. a) A afirmação é **Falsa**.
Se $X \frown Poisson(10)$ para calcularmos $P[X>8]$=`ppois(8,10,lower.tail=FALSE)` ≡ `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 **Verdadeira**, pois significa a soma da probabilidade de todos os valores possíveis para X, logo igual a 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 **Falsa**.

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 **Verdadeira**.

5. X concentração de dióxido de enxofre (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) Sabemos que se $X \frown Gama(\alpha,\beta)$ se tem $\mu = E[X] = \alpha\beta$ e $\sigma^2 = Var[X] = \alpha\beta^2$. Então neste caso $\mu = E[X] = \beta/2$ e $\sigma^2 = Var[X] = \beta^2/2$.

Pelo Teorema Limite Central $\overline{X} \sim Normal(\mu, \sigma/\sqrt{n}) \Longleftrightarrow \overline{X} \sim Normal(\beta/2, \beta/\sqrt{2n}) \Longleftrightarrow \dfrac{\overline{X}-\beta/2}{\beta/\sqrt{2n}} \sim Normal(0,1)$.

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:

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

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

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

portanto um intervalo assintótico a 95% de confiança para o parâmetro β é:

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