Instituto Superior de Agronomia
Bioinformática – 2015/2016

Tópicos de uma resolução das perguntas do 1º e 2º teste de "Testes, Máxima Verosimilhança, Bootstrap"

1º Teste

Pergunta:

No estudo de um dado tipo de ave em quatro habitas diferentes, recolheram-se 40 aves em cada um desses quatro habitats. Registou-se depois o sexo de cada ave, tendo-se obtido os seguintes resultados:

Habitat	H1	H2	H3	H4	Total
Machos	18	19	20	25	82
Fêmeas	22	21	20	15	78
Total	40	40	40	40	120

Consulte o Anexo abaixo com os resultados obtidos no R.

a) Complete os valores que foram substituídos pelas letras A, B, C e D
b) Poder-se-á dizer que a proporção de cada sexo difere consoante os habitats? Justifique convenientemente.

Anexo

```
> habitat_sexo<-matrix(c(18,19,20,25,22,21,20,15),nc=4,byrow=T,
+ dimnames=list(c("Machos", "Fêmeas"),c("H1", "H2","H3", "H4")))

> margin.table(habitat_sexo,1)
Machos Fêmeas
    A     B

> chisq.test(habitat_sexo)

        Pearson's Chi-squared test

data:  habitat_sexo
X-squared = 2.9018, df = 3, p-value = 0.407


> chisq.test(habitat_sexo)$expected
         H1   H2   H3   H4
Machos 20.5 20.5 20.5 20.5
Fêmeas 19.5 19.5 19.5 19.5

> chisq.test(habitat_sexo)$residuals^2
           H1        H2         H3        H4
Machos      C 0.1097561 0.01219512 0.9878049
Fêmeas 0.3205128 0.1153846 0.01282051 1.0384615

> pchisq(2.9018,3)
[1]    D
```

Resposta:

a) $A = 82, \quad B = 78\ , \quad C = \dfrac{(18 - 20.5)^2}{20.5} = 0.304878$
$D = P[\chi^2_{(3)} < 2.9018] = 1 - P.value = 1 - 0.407 = 0.593$

b) Estamos perante uma tabela de contingência com **uma margem fixa - o total em cada habitat**, portanto vamos realizar um teste de homogeneidade, i.e, pretende-se testar

H_0 : os diferentes habitas apresentam a mesma proporção de aves relativamente ao sexo
H_1 : em pelo menos dois habitats a distribuição das aves de cada sexo não é a mesma
A estatítica de teste é $X^2 = \sum_{i=1}^{2} \sum_{j=1}^{4} \dfrac{(O_{(ij)} - e_{(ij)})^2}{e_{(ij)}}$
Sob a validade da hipótese nula, tem-se, $X^2 \sim \chi^2_{(1)\times(3)} = \chi^2_{(3)}$

Como $p.value = 0.407 > 0.05$ (nível de significância habitualmente considerado), **não se rejeita** H_0, portanto podemos dizer que podemos considerar que a proporção de machos/fêmesa não difere de habitat para habitat.

2º Teste

Pergunta:

O modelo de Rayleigh tem sido usado para caracterizar situações no domínio das ciências biológicas em que poderão ocorrer valores "severos". Vamos considerar X uma variável aleatória seguindo um modelo de Rayleigh numa forma muito simples, apenas com um parâmetro desconhecido, $\theta > 0$, com função densidade definida como:

$$f(x|\theta) = \frac{x}{\theta} \exp\left[-\frac{x^2}{2\theta}\right], \quad \text{se } x > 0, \quad \text{e nula nos restantes valores de } x$$

Sabe-se que $E[X] = \sqrt{\frac{\theta\pi}{2}}$ e $Var[X] = \frac{4-\pi}{2}\theta$.

Considere que se tem uma amostra aleatória de dimensão n, $(X_1, X_2, .., X_n)$, associada a X.

a) Determine o estimador de θ pelo método dos momentos.
b) Determine o estimador de máxima verosimilhança para θ.
c) Tendo observado a seguinte amostra extraída daquela população,

2.6	1.7	3.9	0.9	4.1	0.4	1.9	3.0	4.7	3.2
3.0	2.8	0.6	2.1	1.7	3.9	2.2	1.6	2.4	1.6

determine estimativas para θ (consulte o Anexo).
d) Suponha que é sugerido o seguinte estimador para θ, $T = 2S^2$, onde S^2 designa a variância amostral.
i) Explique sucintamente o que se pretende com os comandos do Anexo identificados por **A** e **B**.
ii) Com recurso ao *bootstrap*, indique uma estimativa do parâmetro θ e um intervalo de confiança a 95% para θ.

Anexo

```
> amostra
[1] 2.6  1.7  3.9  0.9  4.1  0.4  1.9  3.0  4.7  3.2  3.0  2.8  0.6  2.1  1.7  3.9  2.2  1.6  2.4

> sum(amostra)                    > sum(amostra^2)
[1] 48.3                                [1] 143.01

> # Metodologia bootstrap

A  > boot <-numeric(1000)
B  > for (i in 1:1000)   boot[i] <- 2*var(sample(amostra,replace=T))

> mean(boot)                          >var(boot)
[1] 2.617516                           [1] 0.4925298

> quantile(boot,prob=c(0.025,0.975))
    2.5%         97.5%
1.335007      4.047878
```

Resposta:

a) Esta função só tem um parâmetro desconhecido, então para aplicar o método dos momentos bastanos uma só equação, i.e,
$E[X] = \overline{X} \Leftrightarrow \sqrt{\frac{\theta\pi}{2}} = \overline{X}$, portanto o estimador pelo método dos momentos para θ é $\Theta^* = \frac{2\overline{X}^2}{\pi}$.

b)Comecemos por pensar na amostra observada $(x_1, x_2, ..., x_n)$, na população X contínua.
Define-se **verosimilhança** como

$$\mathcal{L}(\theta|x_1, ..., x_n) = f(x_1|\theta) \times ... \times f(x_n|\theta) = \Pi_{i=1}^n f(x_i|\theta) = \frac{x_1}{\theta} e^{-\frac{x_1^2}{2\theta}} \times ... \times \frac{x_n}{\theta} e^{-\frac{x_n^2}{2\theta}} = \frac{\Pi x_i}{\theta^n} e^{-\frac{\sum x_i^2}{2\theta}}$$

Para determinarmos o máximo desta função é mais fácil trabalhar com $\log \mathcal{L}()$ (representação simplificada)

$$\log \mathcal{L}() = \log(\Pi_{i=1}^n x_i) - n\log(\theta) - \sum_{i=1}^n \frac{x_i^2}{2\theta}$$

Então agora basta derivar em ordem a θ e depois igualar a zero, para obtermos o ponto crítico (aqui não é necessário ir calcular as segundas derivadas, porque propriedades da máxima verosimilhança garantem que a solução encontrada é um maximizante)

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

portanto vamos agora igualar a zero

$$-\frac{n}{\theta} + \frac{\sum_{i=1}^n x_i^2}{2\theta^2} = 0 \Leftrightarrow -n + \frac{\sum_{i=1}^n x_i^2}{2\theta} = 0 \Rightarrow n = \frac{\sum_{i=1}^n x_i^2}{2\theta}$$

Temos então a estimativa e o estimador de máxima verosimilhança para θ, respectivamente,

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i^2}{2n} \qquad \hat{\Theta} = \frac{\sum_{i=1}^n X_i^2}{2n}$$

c) Face a esta amostra de 20 valores da variável X e usando a valor de $\sum_{i=1}^n x_i = 48.3$ e também $\sum_{i=1}^n x_i^2 = 143.01$ temos duas estimativas para θ, dadas pelo método dos momentos e pelo método da máxima verosimilhança, respectivamente:
$\theta^* = 2 * (48.1/20)^2/\pi = 3.6822 \qquad \hat{\theta} = 143.01/40 = 3.575$

d) i) **A** – cria um vector designado *boot*, com 1000 componentes iguais a 0
B trata-se de um ciclo com B=1000 runs. Em cada réplica é gerada uma amostra aleatória simples com reposição a partir da amostra inicial, com a mesma dimensão dessa amostra.
Sobre cada amostra é calculada uma estimativa de T, sendo cada estimativa guardada numa componente de *boot*.

ii) A estimativa *bootstrap* de θ é $\theta^*_B = 2.617516$.
Um intervalo de confiança *bootstrap* a 95%, obtido pelo método dos percentis é]1.335007 , 4.047878[