Instituto Superior de Agronomia
Bioinformática – 2019/2020

Tópicos de uma resolução das perguntas 1 a 4

1. Vamos colocar o valor das margens na tabela

Sexo do 1º filho	Idade do pai-idade da mãe			Total
	-9 a -1	0 a 5	5 a 15	
Masculino	14	117	37	**168**
Feminino	29	84	20	**133**
	43	**201**	**57**	**301**

Resposta:

a) 301

b) A=TRUE, B=estudo.sexo, $C = 2$, $D = \dfrac{43 * 133}{301} = 19$

$E = \dfrac{20 - 25.18605}{\sqrt{25.18605}} = -1.03337$ $F = P[\chi^2_{(2)} < 11.8106] = 1 - P.value = 0.9973$

c) Estamos perante uma tabela de contingência com **margens livres**, portanto vamos realizar um teste de independência, i.e, pretende-se testar se há independência entre a diferença de idades dos pais e o sexo do primeiro filho ou pelo contrário existirá alguma relação.

H_0 : a diferença de idades dos pais e o sexo do primeiro filho são independentes
H_1 : existe alguma associação.
i.e. $H_0 : p_{ij} = p_{i\bullet}p_{\bullet j}$ $\forall(i,j)$ H_1 : pelo menos 2 daquelas igualdades não se verificam
A estatística de teste é $X^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \dfrac{(O_{(ij)} - e_{(ij)})^2}{e_{(ij)}}$
Sob a validade da hipótese nula tem-se, $X^2 \sim \chi^2_{(1)\times(2)} = \chi^2_{(2)}$

Como $p.value = 0.002725 < 0.05$ (nível de significância habitualmente considerado), **rejeita-se** H_0, portanto podemos dizer que há alguma relação entre a diferença de idades dos pais e o sexo do primeiro filho.
Daquela tabela teste$expected vemos que parece haver influência no sexo da criança quando o pai é mais novo que a mãe, sendo o sexo feminino mais frequente do que seria expectável se houvesse independência.
Note-se que as condições de validade do teste são verificadas, todas as frequências esperadas são mesmo superiores a 5.

d) Da análise dos resultados apresentados em teste$residuals vemos que as parcelas "responsáveis" pelo valor elevado da estatística de teste X^2 são as referentes à diferença de idades entre pai e mãe estar entre -9 e -1. Pelo sinal dos resíduos também se confirma que há mais tendência para haver mais crianças do sexo feminino do que seria expectável se houvesse independência.

2. As respostas são dadas à frente de cada comando

Resposta:

```
> dna<-c("A","C","G","T")   ## é criado um vector de alfanuméricos
                            ##  A, C, G, T  com o nome dna
   >  seq2<-sample(dna,1000,replace=T,
  +  prob=c(0.20,0.30,0.18,0.32))

 ## do vector dna é retirada uma amostra ao acaso, de dimensão 1000, com reposição,
 ## em que os valores A, C, G e T são amostrados com probabilidades definidas no vector
 ## prob. Essa amostra é guardada no vector seq2

> table(seq2)     # cria uma tabela das frequências absolutas de cada  valor observado
 seq2
   A    C    G    T
  192  289  183  336

> pbinom(192,1000,0.20)
[1] 0.2783474
# Para a v.a. X com dist. Binomial(n=1000,p=0.20) calcula P[X<=192]

> 1-pbinom(207,1000,0.20)
[1] 0.2749125
# Para a v.a. X com dist. Binomial(n=1000,p=0.20) calcula 1- P[X<=207]=P[X>207]=P[X>=208]
```

O que se pede é que se responda ao teste $H_0 : p = 0.20$ *vs* $H_1 : p \neq 0.20$ sendo X a v.a. que conta o número de vezes que se observa "A" na sequência; $X \frown Binomial(n = 1000, p)$

Como formulei um teste bilateral, $p - value = P[X <= 192] + P[X >= 208] = 0.5533$, portanto superior a qq valor de α habitual.

Não se rejeita H_0, portanto o nucleótido "A" pode ocorrer na proporção definida no estudo.

Nota: Podia usar-se o teste

```
prop.test(192, 1000, p = 0.2,alternative = "two.sided")
```

3. Amostra aleatória de dimensão n, $(X_1, X_2, .., X_n)$, retirada de uma população X,

$$f(x;\beta) = \begin{cases} \dfrac{\beta+1}{\mathrm{e}^{\beta+1}} x^{\beta} & \text{se } 0 \leq x \leq \mathrm{e} \\ 0 & \text{outros valores de } x \end{cases}$$

Nota: Sabe-se que $E[X] = \dfrac{(\beta+1)\mathrm{e}}{\beta+2}$.

Resposta:

a) Esta função densidade só tem um parâmetro desconhecido, então para aplicar o método dos momentos basta-nos uma só equação, i.e, o estimador de β é a solução da equação

$$E[X] = \overline{X} \Longleftrightarrow \frac{(\beta+1)\mathrm{e}}{\beta+2} = \overline{X} \Longleftrightarrow (\beta+1)\,\mathrm{e} = \overline{X}(\beta+2) \Longleftrightarrow \beta\mathrm{e} + \mathrm{e} = \beta\overline{X} + 2\overline{X}$$

Logo o estimador de β pelo método dos momentos é $\beta^* = \dfrac{2\overline{X} - \mathrm{e}}{\mathrm{e} - \overline{X}}$

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

$$\mathcal{L}(\beta|x_1, ..., x_n) = f(x_1|\beta) \times ... \times f(x_n|\beta) = \frac{\beta+1}{\mathrm{e}^{\beta+1}} x_1^{\beta} \times ... \times \frac{\beta+1}{\mathrm{e}^{\beta+1}} x_n^{\beta} = \left(\frac{\beta+1}{\mathrm{e}^{\beta+1}}\right)^n \Pi_{i=1}^n x_i^{\beta}$$

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

$$\log \mathcal{L}() = \log \left(\frac{\beta+1}{\mathrm{e}^{\beta+1}} \right)^n + \log(\Pi_{i=1}^n x_i^\beta) = n \log(\beta+1) - n \log(\mathrm{e}^{\beta+1}) + \left(\sum_{i=1}^n \log x_i^\beta \right)$$

$$\log \mathcal{L}() = n \log(\beta+1) - n(\beta+1) + \beta \left(\sum_{i=1}^n \log x_i \right)$$

Então agora basta derivar em ordem a β e depois igualar a zero, para obtermos o ponto crítico, que é um maximizante

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

portanto vamos agora igualar a zero

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

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

$$\hat{\beta} = \frac{\sum_{i=1}^n \log x_i}{n - \sum_{i=1}^n \log x_i} \qquad \hat{\Theta} = \frac{\sum_{i=1}^n \log X_i}{n - \sum_{i=1}^n \log X_i}$$

c) Temos uma amostra observada de dimensão 30, extraída daquela população, com a qual se realizaram os cálculos apresentados:

```
> sum(dados)          > sum(log(dados))
[1] 59.77              [1] 19.39522
```

Face a esta amostra de 30 valores da variável X e usando os valores de $\sum_{i=1}^n x_i = 59.77$ e também $\sum_{i=1}^n \log x_i = 19.39522$ temos duas estimativas para β, dadas pelo método dos momentos e pelo método da máxima verosimilhança, respectivamente:

$\beta^* = \dfrac{2 \times 59.77/30 - \mathrm{e}}{\mathrm{e} - 59.77/30} = 1.7445 \qquad \hat{\beta} = \dfrac{19.39522}{30 - 19.39522} = 1.8289$

4. Estimador do parâmetro β

$$\beta^* = \frac{2(\overline{X} - 1)}{2 - \overline{X}}.$$

a) A=30; B=TRUE; C=2.003918

b) Uma estimativa *bootstrap* de β é $\beta_B^* = 2.003918$

c) Um intervalo *booststrap* a 90% de confiança para β, é dado pelos percentis $Q_{0.05}^* = 1.465054$ e $Q_{0.95}^* = 2.588998$, portanto o I.C *booststrap* a 90% é
$]1.465054, 2.588998[$