Instituto Superior de Agronomia
Bioinformática – 2019/2020

Topics of Questions 1 a 4 resolution

1. Let's write the value of the margins in the table

Sex of 1st child	Father's age - mother's age			Total
	-9 a -1	0 a 5	5 a 15	
Masculino	14	117	37	**168**
Feminino	29	84	20	**133**
	43	**201**	**57**	**301**

Answer:

a) 301

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

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

c) We are facing a contingency table with **free margins**, so we are going to carry out a Test of Independence, i.e., it is intended to test whether there is independence between the age difference of the parents and the sex of the first child or on the contrary there will be some relationship.

H_0 : the age difference of the parents and the sex of the first child are independent
H_1 : there is some relationship.
i.e. $H_0 : p_{ij} = p_{i\bullet}p_{\bullet j}$ $\forall(i,j)$ H_1 : at least 2 of those equalities are not verified
The test statistic is $X^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{(ij)} - e_{(ij)})^2}{e_{(ij)}}$
Under the validity of the null hypothesis, we have, $X^2 \sim \chi^2_{(1)\times(2)} = \chi^2_{(2)}$

As $p.value = 0.002725 < 0.05$ (level of significance usually considered), textbf is rejected H_0, so we can say that there is some relationship between the age difference of the parents and the sex of the first child.
From that test$expected table we see that there seems to be an influence on the sex of the child when the father is younger than the mother, with the female sex being more frequent than would be expected if there were independence.
Note that the test's validity conditions are verified, all expected frequencies are even higher than 5.

d) From the analysis of the results presented in test$residuals we see that the components "responsible" for the high value of the test statistic X^2 are those referring to the age difference between father and mother being between -9 and -1. The sign of the residuals also confirms that there is more tendency to have more female children than would be expected if there were independence.

2. Answers are given in front of each command

```
> dna<-c("A","C","G","T")   ## an alphanumeric vector (A, C, G, T) is created with
             ##  the name dna
   >  seq2<-sample(dna,1000,replace=T,
  +  prob=c(0.20,0.30,0.18,0.32))

 ## of the vector dna a random sample is taken, with size  1000, with replacement,
 ## where values A, C, G and T are sampled with probabilities defined in the vector prob.
 ## This sample is saved in the seq2 vector

> table(seq2)     # creates a table of the absolute frequencies of each observed value
 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
# For the random variable X,  Binomial(n=1000,p=0.20)
# we calculated 1- P[X<=207]=P[X>207]=P[X>=208]
```

What is asked is to answer the test $H_0 : p = 0.20$ *vs* $H_1 : p \neq 0.20$ where X is a r.v. that counts the number of times we observe "A" in the sequence ; $X \frown Binomial(n = 1000, p)$

As I formulated a bilateral test, $p - value = P[X <= 192] + P[X >= 208] = 0.5533$, therefore higher than any value of α habitual.

H_0 is not rejected, therefore the nucleotide "A" can occur in the proportion defined in the study.

Nota: The following test could be used

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

3. $(X_1, X_2, .., X_n)$, random sample withdrawal of a population 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: We know that $E[X] = \dfrac{(\beta+1)\mathrm{e}}{\beta+2}$.

Answer:

a) This density function only has an unknown parameter, so to apply the method of moments we need only one equation, i.e., the estimator of β is the solution of the equation

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

The estimator of β by the method of the moments is $\beta^* = \dfrac{2\overline{X} - \mathrm{e}}{\mathrm{e} - \overline{X}}$

b) Let?s start by thinking about the observed sample $(x_1, x_2, ..., x_n)$, in the continuous X population.

The **Likelihood** is defined as

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

To determine the maximum of this function, it is easier to work with $\log \mathcal{L}()$ (simplified representation)

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

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

So now we just need to derive in order to β and then put equal to zero, to get the critical point, which is a maximizer

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

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

Then we have the estimate and estimator of maximum likelihood for β, respectively,

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

c) We have an observed sample of dimension 30, extracted from that population, with which the calculations presented were performed:

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

Given this sample of 30 values of the variable X and using the values of $\sum_{i=1}^n x_i = 59.77$ and $\sum_{i=1}^n \log x_i = 19.39522$ we have two estimates for β, given by the method of moments and the method of maximum likelihood, respectively:

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

4. Estimator of β

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

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

b) A *bootstrap* estimate of β is $\beta_B^* = 2.003918$

c) A *boostrap* confidence interval at 90% for β, is given by the percentiles $Q_{0.05}^* = 1.465054$ and $Q_{0.95}^* = 2.588998$, therefore the CI textit booststrap at 90% is
$]1.465054, 2.588998[$