Instituto Superior de Agronomia
Modelos Matemáticos – 2015/2016

Test Module I - topics of a short resolution 6.04.16

1. It is agood idea you try to enter the data into **R**.

a) $n = 525$ Let's complete the missing values in the output:

$A = 58/525 = 0.110$; $B = 522/525 = 0.994$

$C =$? let's see – the cumulative relative frequency — $F[1] = 0.448 < 0.5$ and $F[2] = 0.745$ (the first value of x for which $F(x) \geq 0.5$) $C = median = 1$

$D =$? this is the lower limit of the confidence interval(CI). Remember that the CI is symmetric relatively to $\overline{x}$, (x the number of affected leaves in each plant)
i.e. CI=$]\overline{x} - A, \overline{x} + A[$

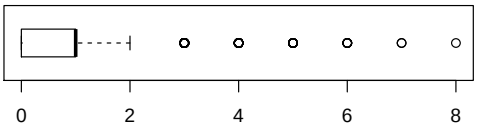
$\overline{x} + A = 1.17987 \Rightarrow A = 1.179870 - 1.060952 = 0.118918 \Rightarrow \overline{x} - A = 0.942034$=D

$E = \sqrt{(Var)} = 1.386996$

b) moda – 0 affected leaves /plant;
mediana – 1 affected leaf/plant;
mean – 1.060952 affected leaves/plant.

The mean $\simeq$ median what could indicate some symmetry, however the mode is **zero** and 75% of the data are 0 and 1, what indicates a strong concentration on the left; the distribution is asymmetric. It is positive skew (the skewness is 1.7886).
Note: The right tail is longer; the mass of the distribution is concentrated on the left, the distribution is said to be right-skewed or skewed to the right.
Try to draw the histogram!!!!

c) For the boxplot we need the following indicators:
min =0; max=8
$Q_1 = 0$; $Q_2 = 1$; $Q_3 = 1$
and now the upper barrier and lower barrier; UB and LB are given by
$UB = Q_3 + 1.5(Q_3 - Q_1) = 2.5$ $LB = Q_1 - 1.5(Q_3 - Q_1) = -1.5$ so there are only possible "upper outliers": 3, 4, 5, 6, 7 and 8.



d) The proportion estimated is $\hat{p} = 1 - 235/525 = 0.5524$

e) As $n = 525$ is large we can consider the 95% CI obtained using

$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Longleftrightarrow]0.5098464; 0.5949155[$

easily obtained in R with

```
>propest<-(525-235)/525;propest
>propest+c(qnorm(0.025)*sqrt(propest*(1-propest)/525),
    qnorm(0.975)*sqrt(propest*(1-propest)/525))
```

The 95% CI for the proportion of plants with affected leaves is :
]0.5098464; 0.5949155[, so with a confidence of 95% the true proportion of plants with affected leaves will be between 0.51 and 0.59.

f)
```
> mean(folhas)+c(qnorm(0.025)*sd(folhas)/sqrt(525),
+ qnorm(0.975)*sd(folhas)/sqrt(525))
[1] 0.9423089 1.1795958
```

g) We have $\overline{x} = 1.060952$ (x the number of affected leaves in each plant)

We suspect that the value is not significantly larger than 1, but we should perform a test.
$H_0 : \mu = 1 \quad vs \quad H_1 : \mu > 1, \quad$ with the level of significance $\alpha = 0.05$

As $n = 525$ is large and we do not know the value of σ^2 we have to use the Test Statisti c $Z = \dfrac{\overline{X} - 1}{s/\sqrt{n}} \simeq \mathcal{N}(0,1)$.

The Rejection Region is $RC : Z > z_\alpha \Leftrightarrow Z > 1.65$

$Z_{calc} = 1.0069$, so as $Z_{calc} \notin RC$ we can not reject H_0, what means that we can not say that, on average, there is more than 1 leaf affected by plant, with a significance level of 5%.
Comment: You could perform a t.test (because the number of degrees of freedom is very large), that will lead to the same result:

```
> t.test(folhas,mu=1,alternative="greater")

        One Sample t-test

data:  folhas
t = 1.0069, df = 524, p-value = 0.1572
alternative hypothesis: true mean is greater than 1
95 percent confidence interval:
 0.9612073       Inf
sample estimates:
mean of x
 1.060952
```

2. a) We have $\theta > 0$ and the Method of Moments establishes that the estimator is the solution of

$$E[X] = \frac{\sum x_i}{n} \Longleftrightarrow \theta^2 = \overline{x} \Longleftrightarrow \theta = \sqrt{\overline{x}}$$

So the estimator is $\Theta^* = \sqrt{\overline{X}}$

b) For obtaining the Maximum Lilelihood Estimator we need to obtain the likelihood funtion

$$L(\theta|x_1,\ldots,x_n) = \frac{\Pi_{i=1}^n x_i^{-1/2}}{(\sqrt{2\pi\theta^2})^n} exp\left(-\frac{\sum_{i=1}^n x_i}{2\theta^2}\right)$$

Now the logarithm is

$$logL(\theta|x_1,\ldots,x_n) = log\left(\frac{\Pi_{i=1}^n x_i^{-1/2}}{(\sqrt{2\pi\theta^2})^n}\right) + \left(-\frac{\sum_{i=1}^n x_i}{2\theta^2}\right)$$

$$logL(\theta|x_1,\ldots,x_n) = log\left(\Pi_{i=1}^n x_i^{-1/2}\right) - \frac{n}{2}log(2\pi\theta^2) - \frac{\sum_{i=1}^n x_i}{2\theta^2}$$

Calculating the derivative:

$$\frac{dlogL}{d\theta} = 0 - \frac{n}{2} \times \frac{4\pi\theta}{2\pi\theta^2} + \frac{\sum_{i=1}^n x_i}{2} \times \frac{2\theta}{\theta^4}$$

and now doing it equal to zero:

$$-\frac{n}{2}\times\frac{4\pi\theta}{2\pi\theta^2}+\frac{\sum_{i=1}^{n}x_i}{2}\times\frac{2\theta}{\theta^4}=0\Longleftrightarrow-\frac{n}{\theta}+\frac{\sum x_i}{\theta^3}\Longleftrightarrow\theta^2=\frac{\sum x_i}{n}\Longleftrightarrow\theta=\sqrt{\overline{x}}$$

The ML estimator is also $\widehat{\Theta}=\sqrt{\overline{X}}$

We know that $E[\overline{X}]=\mu=\theta^2$ however $E[\sqrt{\overline{X}}]\neq\sqrt{E[\overline{X}]}=\sqrt{\theta^2}=\theta$

so $\widehat{\Theta}=\sqrt{\overline{X}}$ is not an unbiased estimator of θ.

c) The two estimators are equal so the estimate of θ based on the observed sample is

$\hat{\theta}=\sqrt{\overline{x}}=1.924058$

d) $CV=\dfrac{\sigma}{\mu}\times 100\%$. For estimating CV we can estimate directly μ and σ.

An estimate of μ is $\overline{x}=1.924058$ and an estimate of σ is $s=\sqrt{s^2}=\sqrt{\dfrac{\sum(x_i)^2-n\overline{x}^2}{n-1}}$, $n=25$, $s=5.594194$

An estimate of CV is 151.1128%, showing a very high dispersion.

3. Please see the script

a)
```
> z<-seq (-4,4,0.01)
> plot(z,dnorm(z),type="l")
> z2<-seq(0,10,0.01)
> plot(z2,dchisq(z2,1,type="l")
```

b) $X\frown\mathcal{N}(0,\sigma)$, with $\sigma^2=0.4$

Then $\dfrac{(X-0)^2}{0.4}\frown\chi^2_{(1)}$.

$P[a<X^2<b]\Longleftrightarrow P\left[\dfrac{a}{0.4}<\dfrac{X^2}{0.4}<\dfrac{b}{0.4}\right]$;

```
> pchisq(b/0.4,1)-pchisq(a/0.4,1)
```

4. a) Here we are considering **only the distances C** and it is intended to test whether or not the observed sample is compatible with the hyphotesis of equal distribution fot the three levels.

It is a goodness-of-fit test and the hyphoteses are:

$H_0: p_1=p_2=p_3=1/3$ *vs* $H_1:$ at least two p_i are different

It is a Chi-Square test – see the script

b) p-value=F=$P[\chi_{(4)}>3.7294]$

`pchisq(3.7294,4,lower.tail=F)=0.4439`

Here it is a Test of Independence, given that a sample of size $n=353$ was classified according two criteria of classification: distance and level

$H_0: p_{ij}=p_i\times p_j\quad\forall(i,j)$ *vs* $H_1: p_{ij}\neq p_i\times p_j$ at least for two pairs

Please see the script