R Short Course | MBBC Students 18/19

Statistics 1

Parametric Inference

Lisete Sousa
lmsousa@fc.ul.pt
Room: 6.4.25

Departamento de Estatística e Investigação Operacional
Centro de Estatística e Aplicações

Ciências ULisboa

CEAUL
Centro de Estatística e Aplicações
Universidade de Lisboa

2019

Summary

1. The Nature of Statistical Inference

2. Confidence Interval Estimation
   - Introduction
   - Some of the Most Used CI
   - Examples in R

3. Hypothesis Testing
   - Parametric Tests

1. The Nature of Statistical Inference

- Statistics deals with data arising from any experiment which result is subject to some random mechanism.
- This means that any time the experiment is performed the result can be different.
- It is not known for certainty what the result will be, but it is known the set of its possible values.

- Experiments are performed in order to draw conclusions.
- However the scientist may want to generalize from that particular experiment to the class of all similar experiments.
- This is the field of inductive inference.
- In inductive inference uncertainty is always present.
- However uncertain inferences can be made, and the degree of uncertainty can be measured if the experiment is performed according with certain principles.

- The theory of Statistics provides techniques for making inductive inferences and for measuring the degree of uncertainty of such inferences.
- Uncertainty is measured in terms of probability.
- For that the result of the experiment is considered to be an observed value of some random variable (or random vector) with a known sample space (the set of possible values to be observed).
- Adequate probabilistic models which may govern the chance mechanism inherent to the observed data are built and relevant inferences are then drawn.

2. Confidence Interval Estimation

Introduction

- Instead of giving a point estimator for a parameter we may instead give an interval estimator which contains the true value of the parameter with a certain probability.
- A **confidence interval** for a parameter is an interval of numbers within which we expect the true value of the population parameter to be contained. The endpoints of the interval are computed based on sample information.
- If confidence intervals are constructed across many separate data analyses of repeated (and possibly different) experiments, the proportion of such intervals that contain the true value of the parameter will match the confidence level.

How to construct a confidence interval (CI)?

Suppose $X_1, \ldots, X_n$ random variables independent identically distributed.

1. Identify the parameter of interest;
2. Determine the confidence level $(1-\alpha)100\%$;
   Note: if not specified, set the confidence to 95%
3. Check the assumptions;
4. Identify the required formula for the CI;
5. Identify the descriptive statistics needed, from the sample $x_1, \ldots, x_n$;
6. Find the required critical value (probability quantile);
7. Compute de CI based on formula in step 4.

Some of the Most Used CI

$(1-\alpha)100\%$ **CI for the mean** μ

1. Parameter: μ (expected value).
2. Confidence level: $(1-\alpha)100\%$.
3. a) Normal population, σ known;
   b) Normal population, σ unknown;
   c) Population not normal, but $n \geq 30$.
4. Formula for the CI:
   a) $\bar{x} \pm z_{critical} \frac{\sigma}{\sqrt{n}}$
   b) $\bar{x} \pm t_{critical} \frac{s}{\sqrt{n}}$
   c) $\bar{x} \pm z_{critical} \frac{s}{\sqrt{n}}$ (approximate)

5 Descriptive statistics needed:
sample mean $\bar{x}$;
standard deviation s;
sample size n.

6 Critical value:
a) $\alpha = 0.10$ −> $z_{0.95} = 1.645$
$\alpha = 0.05$ −> $z_{0.975} = 1.960$
$\alpha = 0.01$ −> $z_{0.995} = 2.576$
b) $\alpha = 0.10$ −> $t_{n-1;0.95}$
$\alpha = 0.05$ −> $t_{n-1;0.975}$
$\alpha = 0.01$ −> $t_{n-1;0.995}$
c) Same as in a).

$(1-\alpha)100\%$ CI for the proportion p

1. Parameter: p (proportion).
2. Confidence level: $(1-\alpha)100\%$.
3. Assumptions: $np \geq 10$ and $n(1-p) \geq 10$.
4. Formula for the CI:
   $\hat{p} \pm z_{critical}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (approximate)
5. Descriptive statistics needed:
   sample proportion $\hat{p}$;
   sample size n.
6. Critical value:
   $\alpha = 0.10$ –> $z_{0.95} = 1.645$
   $\alpha = 0.05$ –> $z_{0.975} = 1.960$
   $\alpha = 0.01$ –> $z_{0.995} = 2.576$.

$(1-\alpha)100\%$ CI for the variance σ^2

1. Parameter: σ^2.
2. Confidence level: $(1-\alpha)100\%$.
3. Assumptions: normal population.
4. Formula for the CI:
   $\left(\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}};\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}\right)$
5. Descriptive statistics needed:
   sample standard deviation s;
   sample size n.
6. Critical value:
   $\alpha = 0.10$ –> $\chi^2_{n-1;0.05}$ and $\chi^2_{n-1;0.95}$
   $\alpha = 0.05$ –> $\chi^2_{n-1;0.025}$ and $\chi^2_{n-1;0.975}$
   $\alpha = 0.01$ –> $\chi^2_{n-1;0.005}$ and $\chi^2_{n-1;0.995}$

Consider now two random variables, X_A and X_B from normal populations A and B, with parameters (μ_A, σ_A) and (μ_B, σ_B), respectively; and two random samples from each population $X_{A1}, \ldots, X_{An_A}$ and $X_{B1}, \ldots, X_{Bn_B}$.

$(1-\alpha)100\%$ **CI for the difference between means:**

σ_A and σ_B known, $\bar{x}_A - \bar{x}_B \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$

$\sigma_A = \sigma_B = \sigma$ unknown, $\bar{x}_A - \bar{x}_B \pm t_{n_A+n_B-2;1-\alpha/2} s_p \sqrt{\frac{n_A+n_B}{n_A n_B}}$

and $s_p^2 = \frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2}$, with s_A^2 and s_B^2 the variances of samples A and B, respectively.

Examples in R

Example 1: CI for the expected value (μ)
(Normal population)

Consider a sample of 20 observations

$\underset{\sim}{x} = (32.81, 37.04, 37.21, 31.15, 26.97, 26.58, 31.85, 30.09, 28.63, 25.12, 31.67, 28.26, 28.57, 37.39, 30.55, 32.98, 24.52, 28.28, 27.37, 26.35)$.

Suppose we want to find the 99% confidence interval for μ. Since the variance σ^2 is unknown, the CI is given by

$$\overline{x} \pm t_{n-1;1-\alpha/2} \, \frac{s}{\sqrt{n}} \, .$$

For the data consider in the Example, the 99% CI for μ is,

$$(27.692 \; , \; 32.647)$$

This interval can be calculated easily in R by using the function `t.test`:

```
> x <- c(32.81,37.04,37.21,31.15,26.97,26.58,31.85,
+ 30.09,28.63,25.12,31.67,28.26,28.57,37.39,30.55,
+ 32.98,24.52,28.28,27.37,26.35)
> t.test(x,alternative="two.sided",conf.level=0.99)$conf.int
```

```
[1] 27.69154 32.64746
attr(,"conf.level")

[1] 0.99
```

Example 2: CI for the ratio of two variances (σ_x^2/σ_y^2)

(Two Normal and independent populations)

Consider another sample of 20 observations $\underset{\sim}{y} = (38.14, 39.07, 37.29, 41.20, 40.31, 39.07, 34.99, 36.82, 35.23, 37.97, 36.21, 45.13, 35.98, 36.55, 37.45, 40.23, 38.45, 45.01, 36.94, 42.09)$. Now, we want to find the 95% confidence interval for σ_x^2/σ_y^2, which is given by

$$\left(\frac{s_x^2 F_{n_y-1, n_x-1; \alpha/2}}{s_y^2}, \frac{s_x^2 F_{n_y-1, n_x-1; 1-\alpha/2}}{s_y^2} \right).$$

For the data in this Example and Example 1, the 95% CI for σ_x^2/σ_y^2 is,

$$(0.709\ ,\ 4.523)$$

This interval can be calculated easily in R by using the function `var.test`:

```
> y <- c(38.14,39.07,37.29,41.20,40.31,39.07,34.99,
+ 36.82,35.23,37.97,36.21,45.13,35.98,36.55,37.45,
+ 40.23, 38.45, 45.01, 36.94, 42.09))
> var.test(x,y)$conf.int
```

```
[1] 0.7086474 4.5232640
attr(,"conf.level")
```

```
[1] 0.95
```

Example 3: CI for the difference of expected values ($\mu_x - \mu_y$)

(Two Normal and independent populations)

From Example 2, we can consider the populations' variances to be equal (we will see further, why) at a significance level of 0.05. Since that variance is unknown, the 95% confidence interval for $\mu_x - \mu_y$ is,

$$(-10.726\ ,\ -6.348)$$

This interval can be calculated easily in R by using the function `t.test`:

```
> y <- c(38.14,39.07,37.29,41.20,40.31,39.07,34.99,
+ 36.82,35.23,37.97,36.21,45.13,35.98,36.55,37.45,
+ 40.23, 38.45, 45.01, 36.94, 42.09))
>
t.test(x,y,alternative="two.sided",var.equal=T,paired=F)$conf.int

[1] -10.725979 -6.348021
attr(,"conf.level")

[1] 0.95
```

3. Hypothesis Testing

Running a Hypothesis Test

Steps in hypothesis testing

- Determine the null and alternative hypothesis, using mathematical expressions if applicable.
- Select a significance level (α).
- Take a random sample from the population of interest.
- Calculate a test statistic from the sample that provides information about the null hypothesis.
- Decision (by classical definition or with p-value).
- Conclusion.

Decision

The decision as to whether H_0 is rejected or not rejected is made on the basis of data using the result of a *test statistic*, say $T(X_1, \ldots, X_n)$, or for short, T.

A good test statistic should be such that the probability of committing a type I error is as small as possible.

How to proceed once a test statistic T is chosen?

Rejection regions

The set of possible values the test statistic T can take is divided into two regions

- The acceptance region $\mathcal{A}$; observed values of the test statistic T, falling in this region lead to non-rejection of the null hypothesis.
- The rejection region $\mathcal{R}$; observed values of the test statistic T falling in this region lead to the rejection of the null hypothesis.
- The alternative hypothesis tells the tale (1-tailed vs 2-tailed tests)

This is accomplished either with the knowledge of the exact sampling distribution of the test statistic, under the null hypothesis, or with the help of asymptotic theory.

The rejection regions relatively to a significance level α, are usually of one of the types,

$$\mathcal{R}_\alpha = \{t : t > t_{1-\alpha}\}, \qquad \mathcal{R}_\alpha = \{t : t < t_\alpha\},$$

$$\mathcal{R}_\alpha = \{t : t < t_1 \quad \text{or} \quad t > t_2\},$$

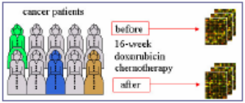
P-value

If for some specific data we observe t_{obs} as the value for the test statistic, then the p-value is the probability of observing a value for the test statistic as "extreme" as t_{obs}. For each of the type of the rejection regions considered above we have:

$$\begin{aligned} p\text{-value} &= P(T > t_{obs} | H_0 \text{true}), \\ p\text{-value} &= P(T < t_{obs} | H_0 \text{true}), \\ p\text{-value} &= 2\ P(T > \text{abs}(t_{obs}) | H_0 \text{true}) \end{aligned}$$

where abs(t_{obs}) means the absolute value of t_{obs}.

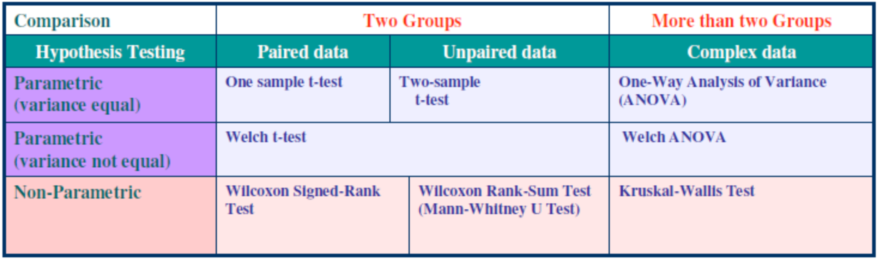
Types of hypothesis tests



Dependent samples

control
treated

Independent samples



Comparison	Two Groups		More than two Groups
Hypothesis Testing	Paired data	Unpaired data	Complex data
Parametric (variance equal)	One sample t-test	Two-sample t-test	One-Way Analysis of Variance (ANOVA)
Parametric (variance not equal)	Welch t-test		Welch ANOVA
Non-Parametric	Wilcoxon Signed-Rank Test	Wilcoxon Rank-Sum Test (Mann-Whitney U Test)	Kruskal-Wallis Test

Parametric Tests

One-sample t-test

- The one-sample t-test compares the mean score of a sample to a known value. Usually, the known value is a population mean.
- Assumption: the variable is normally distributed.
- $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
- Test statistic: under H_0, $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \frown N(0,1)$ or $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \frown t_{n-1}$, if σ is known or unknown, respectively.
- Reject H_0 if $|z_{obs}| > z_{1-\alpha/2}$ or if $|t_{obs}| > t_{n-1;1-\alpha/2}$, respectively
- p-value$=2 \times P(|Z| > z_{obs})$ or p-value$=2 \times P(|T| > t_{obs})$, respectively

NOTES:

1 - The distribution of the data being tested is normal.

- For **paired** t-test, it is the distribution of the subtracted data that must be normal. In R use the argument paired=TRUE.
- For **unpaired** t-test, the distribution of both data sets must be normal. In R use the (default) argument paired=FALSE.
- Plots: Histogram, Density Plot, QQ Plot.
- Test for Normality: Kolmogorov-Smirnov test, Shapiro-Wilk test.

NOTES:

2 - Homoscedasticity: the variances of both populations are equal.

- If the two populations have **equal** variances, then the two-sample t-test may be used. Variance ($\sigma^2 = \sigma_A^2 = \sigma_B^2$) is estimated by s_p^2 (see frame 38). In R use the argument var.equal=TRUE.
- If the two populations have **unequal** variances, then use the two-sample unequal variances t-test (Welch's t-test). In this case, σ_A^2 and σ_B^2 are estimated by s_A^2 and s_B^2, respectively, and the degrees of freedom are given according to Welch's modification. In R use the (default) argument var.equal=FALSE.
- Test for equality of the two variances: variance ratio F-test.

Comparing more than two groups: one-way ANOVA

Used to compare the means of more than two independent groups. Instead of a t-statistic, ANOVA uses a F statistic and its $p - value$ to evaluate the null hypothesis that all of several population means are equal.

In **one-way ANOVA** we classify the populations of interest according to a single categorical explanatory variable that we call a factor.

Assumptions:

1. The distribution of the means by group are normal with equal variances.
   - Bartlett's test (1937)
   - Levene's test (Levene 1960)
   - O'Brien (1979)

2. Sample sizes between groups do not have to be equal, but large differences in sample sizes by group may effect the outcome of the multiple comparisons tests.

(1) The hypotheses for the comparison of independent groups are:

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ (means of the all groups are equal)

H_1 : not all of μ_i are equal

(2) The ANOVA table:

To assess whether several populations all have the same mean, we compare the variation **among** the means of several groups with the variation **within** groups. Because we are comparing variation, the method is called analysis of variance.

Variation is expressed by **sums of squares**. Each sum of squares is the sum of the squares of a set of deviations that expresses a source of variation.

Source of variation	SS	d.f.	Mean squared	F-ratio
Among groups	SSA	$k-1$	$MSA=\frac{SSA}{k-1}$	$F=\frac{MSA}{MSE}$
Within groups	SSE	$N-k$	$MSE=\frac{SSE}{N-k}$	
Total	$SST =$ $= SSA + SSE$	$N-1$	$MST=\frac{SST}{N-1}$	

Where:

1. k is the number of groups;

2. n_i $(i = 1, ..., k)$ is the size of group i;

3. N is the total sample size: $N = n_1 + ... + n_k$;

4. $SSA = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$;
5. $SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$;

When H_0 is true, the statistic F has the $F_{(k-1,N-k)}$ distribution.

(3) Decision:

When H_0 is true, the statistic tends to be small. We reject H_0 in favour of H_1 if the observed value of the statistic F, i.e. F_0, is sufficiently large.

Thus, we reject H_0 for a significance level α, if

$$p - value = P(F_{(k-1,N-k)} > F_0) < \alpha.$$

Example 4:

The following samples refer to the weights (kg) of cows, according to a certain treatment:

A	30.28	27.58	27.91	29.33				
B	34.26	32.55	21.78	25.59	35.08	26.86		
C	39.47	30.15	33.40	27.38	30.39	25.85	29.11	26.22
D	33.54	30.40	29.60	28.82	30.70	30.83	33.84	

Consider $\alpha = 0.01$.

We start by testing the equality of the variances using Bartlett's test:

$$H_0 : \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$$

The p-value obtained is 0.02994 and so, we do not reject the null hypothesis at the significance level of 0.01.

```
> a<-c(30.28,27.58,27.91,29.33)
> b<-c(34.26,32.55,21.78,25.59,35.08,26.86)
> c<-c(39.47,30.15,33.40,27.38,30.39,25.85,29.11,26.22)
> d<-c(33.54,30.40,29.60,28.82,30.70,30.83,33.84)
> observ<-c(a,b,c,d)
> treatm<-factor(rep(c("a","b","c","d"),c(4,6,8,7)))
> bartlett.test(observ˜treatm)

        Bartlett test of homogeneity of variances
data:  observ and treatm
Bartlett's K-squared = 8.952, df = 3, p-value = 0.02994
```

Thus, we may proceed with the ANOVA:

Hypotheses: $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ (treatment effects are equal)

Decision: Since $p - value = P(F_{(3,21)} > 0.41) = 0.75 > 0.01$, for the usual levels of significance, do not reject H_0. This sample does not give evidence for differences between treatment effects.

```
> aov(observ~treatm)

Call:
   aov(formula = observ ~ treatm)
Terms:
                   treatm Residuals
Sum of Squares   17.49895 311.30299
Deg. of Freedom         3        21
Residual standard error: 3.850189
Estimated effects may be unbalanced

> summary(aov(observ treatm))

            Df Sum Sq Mean Sq F value Pr(>F)
treatm       3   17.5   5.833   0.393  0.759
Residuals   21  311.3  14.824
```

Comparing Two Proportions

Suppose that we have two DNA sequences and we want to test the hypothesis that the nucleotide A appears in both sequences with the same frequency.

If we call p_1 the probability of occurrence of nucleotide A in the first sequence and p_2 the probability of the occurrence in the second sequence, then we want to test

$$H_0 : p_1 = p_2 \quad versus \quad H_1 : p_1 \neq p_2.$$

Assume that the sequences were independently generated and let X_1 and X_2 be the number of A nucleotides in subsequence of size n_1 and n_2 from the first and second sequences, respectively.

Again if n_1 and n_2 are large, $\frac{X_1}{n_1}$ and $\frac{X_2}{n_2}$ are approximately normal distributed.

Under the null hypothesis $p_1 = p_2$ they both have expected value p, (the common value of p_1, p_2) and variances $\frac{p(1-p)}{n_1}$ and $\frac{p(1-p)}{n_2}$, respectively.

Since they are independent then $\bar{X}_1 - \bar{X}_2$ is also approximately normally distributed (under H_0) with expected value 0 and variance $\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}$.

To perform the test we can do as follows:

1. Compute $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$, where x_1 and x_2 are observed values of X_1 and X_2 respectively.
2. Compute $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$ as estimate for the common p under the null hypothesis.
3. Compute $\hat{\sigma}_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$
4. Compute the observed value of the test statistic Z as

$$z_{obs} = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\hat{\sigma}_{\hat{p}_1-\hat{p}_2}}$$

5. If this value is outside the interval $[z_{\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]$ then reject the hypothesis $p_1 = p_2$ at the significance level α. Otherwise do not reject the equality of proportions.

Example 5:

As an example, suppose that for $n_1 = n_2 = 100$ we obtained $x_1 = 25, x_2 = 27$.

Then $\hat{p}_1 = 0.25$, $\hat{p}_2 = 0.27$, $\hat{p} = 0.26$, $\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = 0.06203$ and $z_{obs} = -0.322$ which is inside $[-1.96, 1.96]$.

Hence we would not reject the null hypothesis $H_0 : p_1 = p_2$ at the 5% significance level.

Approximate calculations using R functions

As n_1 and n_2 are large an approximate value of z_{obs} can be obtained by using functions `t.test` and `z.test`, however, we have to be careful about how to introduce the data:

t.test

```
> x<-c(rep(1,25),rep(0,75)); y<-c(rep(1,27),rep(0,73))
> t.test(x,y)

Welch Two Sample t-test

data:  x and y

t = -0.3209, df = 197.877, p-value = 0.7486

alternative hypothesis:true difference in means is not equal to 0

95 percent confidence interval:

-0.1429135 0.1029135

sample estimates:

mean of x mean of y

    0.25    0.27
```

Note that, in this case, $\hat{\sigma}_{\frac{X_1}{n_1}-\frac{X_2}{n_2}}$ is equal to $\sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}}$, which is similar to $\hat{\sigma}_{\hat{p}_1-\hat{p}_2}=0.06203$:

```
> sd<-sqrt((var(x1)+var(x2))/100); sd
[1] 0.06232855
```

z.test

```
> library(BSDA)
> z.test(x,y,sigma.x=sqrt(var(x)),sigma.y=sqrt(var(y)))
        Two-sample z-Test
data:  x and y
z = -0.3209, p-value = 0.7483
alternative hypothesis:true difference in means is not equal to 0
95 percent confidence interval:
 -0.1421617  0.1021617
sample estimates:
mean of x mean of y
     0.25      0.27
```

Performing a Chi-Square test for homogeneity

It is also possible two compare two proportions by using R function `prop.test`, but here the considered Statistic is Chi-square distributed.

```
> suc<-c(sum(x1),sum(x2)) # vector - total successes per group
> trials<-c(100,100)      # vector - total trials per group
> prop.test(suc,trials)
> # or
> fail<-100-suc              # vector - total failures per group
> x<-matrix(c(suc,fail),2,2) # contingency table
> prop.test(x)

   2-sample test for equality of proportions with continuity correction
data:  x
X-squared = 0.026, df = 1, p-value = 0.8719
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1515494  0.1115494
sample estimates:
prop 1 prop 2
  0.25   0.27
```

R Functions - Summary

In case of populations Normally distributed, with mean μ and variance σ^2, the following R functions may be used for inferences on the parameters:

H_0	R function	R package	Assumptions
$\mu = \mu_0$ or $\mu_1 - \mu_2 = \mu_0$	`t.test` `z.test`	stats BSDA	unknown variance known variance
$\mu_1 = \ldots = \mu_k$	`aov` or `anova`	stats	equal, unknown var
$\sigma^2 = \sigma_0^2$ $\sigma_1^2/\sigma_2^2 = \sigma_0^2$	`var.test`	stats	

NOTE: In order to apply these tests, it is important to check normality. The most appropriate R function to test normality is `shapiro.test` (Shapiro-Wilk test).

Acknowledgements:

Carina Silva-Fortes (ESTeSL – IPL), for allowing the use of some material produced by both of us in previous courses.

Bibliography:

Keen, K. J. (2010). *Graphics for Statistics and Data Analysis with R*. Chapman & Hall.

Lumley, T. (2010). *Complex Surveys: a Guide to Analysis Using R*. Wiley.

Maindonald, J. (2010). *Data Analysis and Graphics Using R: an Example-Based Approach*. Cambridge University Press.