Lesson 5

Confidence Intervals and Hypothesis Testing.

Lesson 5 Plan

1. Confidence Intervals
   - Confidence intervals for μ, σ^2 and p
   - Confidence intervals - two population means
   - Confidence intervals - two population variances
2. Hypothesis Testing
   - Introductory Example
   - First notions
   - Type I and Type II errors
   - Steps in performing a hypothesis testing
   - The $p - value$
3. Exercícios
4. Goodness-of-fit tests
5. Non-parametric tests (just a taste!)

Confidence Intervals

We have discussed up to now the point estimation and methods of determining parameter estimators of unknown θ.

We will now address the issue of interval estimation.
The intervals are preferred when, instead of proposing an isolated estimate, $\hat{\theta}$, we can associate an error measurement, $\hat{\theta} \pm \epsilon$, to mean that probably the true value of the parameter will be within
$\boldsymbol{\hat{\theta} - \epsilon \, , \hat{\theta} + \epsilon}$.

Confidence Intervals

Definition

Consider a random sample $(X_1, X_2, \cdots X_n)$ from a population with a distribution function $F(x|\theta)$. Let $\Theta_1^\star(X_1, \cdots X_n)$ e $\Theta_2^\star(X_1, \cdots X_n)$ two statistics, such that

$$P(\Theta_1^\star < \theta < \Theta_2^\star) = 1 - \alpha, \qquad 0 < \alpha < 1,$$

where α is a constant, not dependent on the parameter θ.

We say that $\boldsymbol{(\Theta_1^\star, \Theta_2^\star)}$ is an random interval, which contains θ with probability $1 - \alpha$.

With the use of a confidence interval to estimate a parameter we are gaining?

Confidence Intervals

Indeed, let us think, for example, in the $\overline{X}$.
We have $P[\overline{X} = \mu] = 0$, but we already have a positive probability if we consider

$$P\{\mu \in]\overline{X} - a, \overline{X} + a[\ \} \quad \text{com a} > 0$$

that is, there is a positive probability that the random interval contains the unknown parameter.

Definition

Any interval $(\theta_1^\star, \theta_2^\star)$, with $\theta_1^\star < \theta_2^\star$, real numbers, which result from a random interval realization is called confidence interval at $(1-\alpha)100\%$ for θ.

Confidence Intervals

The determination of confidence intervals for the parameters need the knowledge of the distribution of the estimators involved, called sample distributions, that is, they are distributions of functions of the random sample $(X_1, X_2, \cdots, X_n)$, that we will use to get **Confidence Intervals**

Let's remember

- If $X \frown N(\mu, \sigma)$ and σ known $\longrightarrow \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \frown N(0, 1)$

Confidence Intervals

To obtain the Confidence Interval for σ^2

Random variable	Assumptions	Distribution
$\frac{(n-1)S^2}{\sigma^2}$ $S^2 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$	$X_i \frown N(\mu,\sigma)$ $i = 1, 2, \cdots, n$	$\chi^2_{(n-1)}$

Distribution χ^2 definition

If $Z_1, Z_2, \cdots, Z_n$ are independent random variables $N(0,1)$

$$\Downarrow$$

$X = Z_1^2 + \cdots + Z_n^2$ verifies $\quad X \frown \chi^2_{(n)}$

Tem-se $E[X] = n$; $Var[X] = 2n$

Confidence Intervals

To construct the CI for μ with σ unknown

Random variable	Conditions	Distribution
$\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ $S^2 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$	$X_i \frown N(\mu, \sigma)$ $i = 1, 2, \cdots, n$	$t_{(n-1)}$

t — *Student* distribution

If $Z \frown N(0,1)$ and $X \frown \chi^2_{(n)}$ are r.v. independent

$$\Downarrow$$

$$T = \frac{Z}{\sqrt{X/n}} \frown t_{(n)}$$

Confidence intervals for μ

Confidence interval at $(1-\alpha)\times 100\%$ for μ when **$X \frown N(\mu,\sigma)$**

- If σ known

$$\bar{x} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}$$

($z_{\alpha/2} \rightarrow$ value of the r,v. Z such that $P(Z > z_{\alpha/2}) = \alpha/2$)

- If σ unknown

$$\bar{x} - t_{\alpha/2,(n-1)}\,\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,(n-1)}\,\frac{s}{\sqrt{n}}$$

Remarks: The semi-amplitude of the confidence interval and **confidence** or **confidence level** at $(1-\alpha)\times 100\%$
The larger the interval, the greater the degree of confidence, but the lower the precision of the estimate is.

Confidence Intervals (example)

Example of building a CI in R, for the mean value of a normal with known variance (academic example!)

Example Given the sample referring to 10 heights, assume that the measurement errors are normal with mean 0 and standard deviation 1.5.

```
> x<-c(175,176, 173, 175, 174, 173, 173, 176, 173, 179)
> int.conf.z<-function(x,sigma,conf.level=0.95)
  n <-length(x);xbar<-mean(x)
  alpha <- 1 - conf.level
  zstar <- qnorm(1-alpha/2)
  SE <- sigma/sqrt(n)
  xbar + c(-zstar*SE,zstar*SE)   ## we define a function
> int.conf.z(x,1.5) # just do this
```

Obteve-se o I.C a 95% para μ]173.7703; 175.6297[

Confidence Intervals

Confidence Interval at $(1-\alpha)\times 100\%$ for μ

If X has **any distribution, not normal**

It is necessary to have a large sample-size, i.e., $\boldsymbol{n}$ large $\longrightarrow$ application of the Central Limit Theorem

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \qquad \text{if } \underline{\sigma \text{ known}}$$

Or, the usual situation,

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim \mathcal{N}(0,1) \qquad \text{if } \underline{\sigma \text{ unknown}}$$

Interval at $(1-\alpha)\times 100\%$ confidence for$\boldsymbol{\mu}$

$$\overline{x} - z_{\alpha/2}\,\frac{s}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2}\,\frac{s}{\sqrt{n}}$$

Confidence Intervals

Interval at $(1-\alpha)\times 100\%$ confidence for σ^2 in a normal population

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,(n-1)}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,(n-1)}}$$

($\chi^2_\alpha \rightarrow$ is the value of the r.v. χ^2 such that $P[\chi^2 > \chi^2_\alpha] = \alpha$)

Interval at $(1-\alpha)\times 100\%$ confidence for p

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

se $X \frown \mathcal{B}(n,p)$ and n “large”

Confidence intervals - two population means

Confidence intervals at $(1-\alpha)\times 100\%$ for $\mu_1-\mu_2$ with $X_1 \frown \mathcal{N}(\mu_1,\sigma_1)$ and $X_2 \frown \mathcal{N}(\mu_2,\sigma_2)$ and independent samples

- if known variances

$$(\overline{x_1}-\overline{x_2}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} < \mu_1-\mu_2 < (\overline{x_1}-\overline{x_2}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

- if unknown variances but one can admit equal variances.

$$(\overline{x_1}-\overline{x_2}) - t_{\alpha/2}\ s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < \mu_1-\mu_2 < (\overline{x_1}-\overline{x_2}) + t_{\alpha/2}\ s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$t_{\alpha/2} \equiv t_{\alpha/2,(n_1+n_2-2)} \qquad \text{e} \qquad s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

Confidence intervals - two population means

Confidence interval at $(1-\alpha)\times 100\%$ for $\mu_1-\mu_2$ in any two (non-normal) populations, samples independent, of high dimensions n_1 and n_2

$$(\overline{x_1}-\overline{x_2})-z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}<\mu_1-\mu_2<(\overline{x_1}-\overline{x_2})+z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Confidence interval at $(1-\alpha)\times 100\%$ for $\mu_1-\mu_2$, when the variances σ_1^2 and σ_2^2 are unknown and unequal

This situation was handled by Welch-Satterthwaite who considered an approximation of t to v.a. used in the construction of the interval above, the nº of degrees of freedom being calculated approximately— we will come back to this when we deal with Tests of Hypothesis.

Confidence intervals - two population variances

Confidence interval at $(1-\alpha)\times 100\%$ **for** $\boldsymbol{\sigma_1^2/\sigma_2^2}$, with $X_1 \frown \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \frown \mathcal{N}(\mu_2, \sigma_2)$ and independent samples

Here, a new distribution appears, whose characterization was made by Fisher and Snedecor, based on the quotient of two independent r.v., with chi-squared distribution.

The F distribution of Snedecor

Let $U \frown \chi^2_{(m)}$ and $V \frown \chi^2_{(n)}$ be independent random variables, then $X = \frac{U/m}{V/n}$ is said to have distribution F with (m, n) degrees of freedom and is represented by $X = \frac{U/m}{V/n} \frown F_{(m,n)}$.

Confidence intervals - two population variances

Consider now two random samples of dimensions n_1 and n_2, independently drawn from $X_1 \frown \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \frown \mathcal{N}(\mu_2, \sigma_2)$.

You get $\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \frown F_{(n_1-1, n_2-1)}$, variable that allows estimating σ_1^2/σ_2^2.

Under these conditions there is

The confidence interval at $(1-\alpha) \times 100\%$ for $\dfrac{\sigma_1^2}{\sigma_2^2}$ é

$$\frac{s_1^2}{s_2^2 \; f_{\alpha/2;(n_1-1,n_2-1)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2 \; f_{\alpha/2;(n_2-1,n_1-1)}}{s_2^2}$$

Confidence Intervals (Paired Samples)

Confidence intervals for $\mu_1 - \mu_2$ (paired samples)

If in a given experiment the observations are related, i.e. paired by the individual - , the concept of **block** appears here.

Let's consider the paired sample (X_i, Y_i) $(i = 1, ..., n)$

Seja

$D_1 = X_1 - Y_1;\qquad D_2 = X_2 - Y_2;\qquad \ldots\qquad D_n = X_n - Y_n$, i.e.,

let $(D_1, D_2, ..., D_n)$ be the random sample of the differences

Confidence intervals (paired samples)

If $D_1, D_2, \ldots, D_n$ are random variables coming from a normal with mean value $\mu_D = \mu_X - \mu_Y$ and variance σ_D^2, unknown one has

$$\frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \frown t_{(n-1)}$$

Confidence interval at $(1-\alpha) \times 100\%$ **for** μ_D

$$\overline{d} - t_{\alpha/2,(n-1)} \frac{s_D}{\sqrt{n}} < \mu_D < \overline{d} + t_{\alpha/2,(n-1)} \frac{s_D}{\sqrt{n}}$$

If you cannot admit normal D_i, but you have n **'large' the confidence interval** $(1-\alpha) \times 100\%$ **for** μ_D **is**

$$\overline{d} - z_{\alpha/2} \frac{s_D}{\sqrt{n}} < \mu_D < \overline{d} + z_{\alpha/2} \frac{s_D}{\sqrt{n}}$$

Confidence Intervals for $p_1 - p_2$

Let X_1 and X_2 be random variables such that
$X_1 \frown \mathcal{B}(n_1, p_1) \quad \mathrm{e} \quad X_2 \frown \mathcal{B}(n_2, p_2)$;
n_1 and n_2 the dimensions of independent random samples

Confidence interval at $(1-\alpha)\times 100\%$ **for** $\boldsymbol{p_1 - p_2}$ when sample sizes are large

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\ \hat{q}_1}{n_1} + \frac{\hat{p}_2\ \hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\ \hat{q}_1}{n_1} + \frac{\hat{p}_2\ \hat{q}_2}{n_2}}$$

Remark: All these confidence intervals are calculated in R with a function that simultaneously performs a hypothesis test.
For that reason we will do examples later.

Hypothesis Testing

In the statistical inference studied up to this point, we dealt with estimators and construction of Confidence Intervals.

The **Hypothesis Testing** are inference procedures that in view of a statistical assumptions made, assess their plausibility using rules based on an observed sample.

A statistical hypothesis is a statement about an unknown feature of the population.
A hypothesis testing is a statistical procedure that tests whether the data support a statistical hypothesis

Hypothesis Testing Theory “puts our assumptions to be tested".

Hypothesis Testing

Let $\boldsymbol{X}$ be a population with a distribution function known, dependent on a parameter θ, $F(x|\theta)$, or an array of parameters $\underline{\theta} = (\theta_1, ..., \theta_k)$ unknown.

The general idea based on a hypothesis test (parametric test) is the following:
– assume the parameter θ has a certain value θ_0 – this constitutes the hypothesis to be tested - **statistical-hypothesis**, which is usually represented by $\boldsymbol{H_0}$, which can be true or false.
$\boldsymbol{H_0}$ is rejected if there is factual evidence against it, supporting the alternative hypothesis.

To test a statistical hypothesis is to apply a set of rules to decide whether or not the conjecture should be with minimal probability of errors.

Introductory Example

In an olive oil bottling line the quantity of each bottle is a random variable that is supposed to have normal distribution. The filling process is considered regulated if $\mu = 1$ *liter*, not admitting great deviations. To control the filling process, 20 bottles were randomly selected of daily production. Suppose we obtained an average of 0.965 liters with a standard deviation of 0.08 liter.
Can it be said that the process is not regulated? Justify conveniently the answer.

Resolution: To find out if the process is not regulated, we can formulate a test of the hypotheses:

$$\mathbf{H}_0 : \mu = 1 \qquad \textit{v.s.} \qquad \mathbf{H}_1 : \mu \neq 1$$

Hypothesis Testing–First notions

The Rule of a test consists of:

- Divide the sample space into two complementary regions:

RA - acceptance region – the sample values for which the decision is " to accept" H_0

RR or **RC** - rejection region or critical region – the sample values for which H_0 is rejected.

Since Statistical Inference is a "path" that goes from the particular to the general, it can have associated errors:

Possible decisions that can be taken are :

H_0 true and "accept" H_0 → correct decision
H_0 true and reject H_0 → incorrect decision – error
H_0 false and "accept"H_0 → incorrect decision – error
H_0 false and reject H_0 → correct decision

Hypothesis Testing–Type I and Type II Errors

Decision errors of reject or not reject H_0 are designated respectively by **error of 1^a species** or **error of type I** and **error of 2^a species** or **type II error,** with the probabilities associated with each of the errors commonly called

$\alpha =$ P (error of type I) = P (reject $H_0|H_0$ true)
$\beta =$ P (error of type II) = P(not reject $H_0|H_0$ false).

The α is usually called test significance level and
$1-\beta$ = P (reject $H_0|H_0$ false) power of the test

Hypothesis Testing

The following table summarizes what we have just said:

	not rej. H_0	rej. H_0
H_0 true	correct decision	type I error
H_0 false	type II error	correct decision

Hypothesis Testing

Therefore, a hypothesis testing requires two hypotheses:

1. one which is proposed by the experimenter and
2. the other one that is the opposite.

The first one, usually represented by H_1, is call to alternative hypothesis (hypothesis to be searched), the other represented by H_0, is called the null hypothesis.

The hypothesis to be tested is the null hypothesis, which as a general rule we hope to reject.

It is usually written $H_0 \quad vs \quad H_1$.

Hypothesis Testing

The ideal situation would therefore be to have α and β as small as possible. However, this is not possible.

In fact, for fixed sample size, when an error decreases the other increases.

What you usually do – Neyman-Pearson theory – is to set the value of α and try to minimize the β. When there is a test under these conditions it is usually called more potent test.

The Neyman-Pearson Theory established that performing a hypothesis testing is based on the definition of **Test Statistics** and **Critical Regions** associated with the formulated hypotheses.

Let's go back to the example

It is then assumed that the observed sample is the realization of the random sample $(X_1, X_2, \cdots, X_n)$ from a normal population and σ is not known. The hypotheses to be tested are, as we have seen,

$$\boldsymbol{H_0 : \mu = 1 \quad vs \quad H_1 : \mu \neq 1};$$

the **Test Statistics** is $\dfrac{\overline{\boldsymbol{X}} - \boldsymbol{\mu_0}}{\boldsymbol{S/\sqrt{n}}}$, which under the null hypothesis has a known distribution, which we know to be $\boldsymbol{t_{(n-1)}}$. We have then

$$T = \frac{\overline{X} - 1}{S/\sqrt{n}} \frown t_{(19)},$$

The critical region, corresponding to the significance level α is given by $\quad T < -t_{\alpha/2,(19)} \cup T > t_{\alpha/2,(19)}$

Hypothesis Testing

The answer to a hypothesis test is given in the form

- **Reject $\mathbf{H_0}$** - means that the observed data are strongly against $\mathbf{H_0}$ - so, in this case it will be accepted the hypothesis $\mathbf{H}_1$
- **Do not reject $\mathbf{H_0}$** - means that there is not enough evidence to reject $\mathbf{H}_0$.

Steps in performing a hypothesis testing

Steps in performing a **hypothesis testing**:

1. Identify the parameter(s); to specify $\mathbf{H}_0$ **e** $\mathbf{H}_1$ and the level of significance α.
2. Choosing a random variable – statistics test, which under $\mathbf{H}_0$ known distribution (at least approximately).
3. Define the **region of rejection** or **critical region** – **RC** (set of statistical values that are less “plausible” when $\mathbf{H}_0$ is true, so it causes to reject $\mathbf{H}_0$).
4. Calculate **the value** of the test statistic for the observed sample.
5. If the calculated value $\in$ RC $\longrightarrow$ H_0 is rejected
   If the calculated value $\notin$ RC $\longrightarrow$ H_0 is not rejected

The $p-value$

The indication of the observed value of the test statistic, for example z_{cal}, and the indication of a critical value z_α for deciding, for example,

To reject $\mathbf{H}_0$ if $z_{cal} > z_\alpha$ (in a one-sided right test)

has recently been “replaced‘” by the calculation of a probability

– the probability of observing an equal or more extreme value than observed, if the null hypothesis is true – what is denoted as ***p-value***

The $p - value$

Note: this value is a quantity that nowadays any *software* is prepared to calculate when a test is ordered.
We can interpret the ***p-value*** as the

measure of the degree of agreement between the data and H_0

So:
- The smaller the *p-value,* the smaller the consistency between the data and the null hypothesis– for a small p.value the null hypothesis is rejected

It is usually adopted as decision rule:

reject H_0 if *p-value* $\leq \alpha$

Exercício

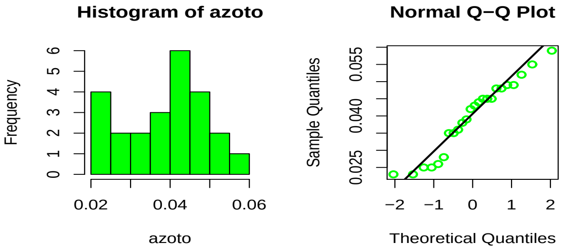
The following data refer to the total nitrogen concentration (ppm) in the water of a lake that is used as a source of urban supply.

0.042	0.023	0.049	0.036	0.045	0.025
0.048	0.035	0.048	0.043	0.044	0.055
0.045	0.052	0.049	0.028	0.025	0.039
0.023	0.045	0.038	0.035	0.026	0.059

1. Determine a confidence interval for μ (at a 99% confidence level).
2. In order to be acceptable as a source of drinking water, the should be less than 0.07 ppm. Do you think the data is compatible with that criterion?

Exercício–resolução

```
> azoto<-c(0.042,0.048,0.045,0.023,0.023,0.035,0.052,
+  0.045,0.049,0.048,0.049,0.038,0.036,0.043, 0.045,
+  0.025, 0.044, 0.055, 0.028, 0.025, 0.039,
+  0.035, 0.026, 0.059)
> qqnorm(azoto)# este é um gráfico para uma
+           #primeira pesquisa da  normalidade
> qqline(azoto)
```



Exercício–resolução

```
>  t.test(azoto,mu=0.0,conf.level=0.99)
#alinea a) pode omitir-se mu=0.0

        One Sample t-test
data:  azoto
t = 18.5066, df = 23, p-value = 2.606e-15
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 0.03382623 0.04592377
sample estimates:
mean of x
 0.039875
```

Exercise–resolution

```
> t.test(azoto, alternative='less', mu=0.07) #alinea b)

        One Sample t-test
data:  azoto
t = -13.9815, df = 23, p-value = 4.944e-13
alternative hypothesis: true mean is less than 0.07
95 percent confidence interval:
       -Inf 0.04356776
sample estimates:
mean of x
 0.039875
```

Hypothesis Testing and CI in R

Confidence interval and hypothesis testing to compare the mean values ??of two populations

```
>t.test(x, y ,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

Performs a test and CI for independent samples, using the Welch-Satterthwaite *t*-test to get an approximation to the nþdegrees of freedom. By default considers `paired = FALSE, var.equal = FALSE`.

```
>data(sleep);>sleep  ## Uma alternativa a um teste a duas médias
  ## forma simples de realizar o teste para comparar 2 grupos
>t.test(extra ~ group, data = sleep, paired = TRUE)
```

Hypothesis Testing and CI in R

Confidence interval and hypothesis testing to compare 2 variances of populations that are assumed to be normal

```
>var.test(x, y, ratio = 1,
         alternative = c("two.sided", "less",
         "greater"),conf.level = 0.95, ...)
```

Confidence interval and hypothesis testing to verify if the proportion of “successes” (in *n* trials in which *x* successes were observed, can be admitted less than 0.4

```
>prop.test(x, n, p = 0.4, alternative = "less",
    conf.level = 0.99, correct = FALSE)
```

Trata-se de um teste a $p = 0.4$ sem correcção de continuidade. É obtido o intervalo a 99% de confiança.

Exercise (TPC)

A study aims to compare an improved type of seed with the type of traditional seed. The improved seed will be used if, on average, the plant growth after 20 days is higher than that of obtained from traditional seeds. 15 different situations are created laboratory, varying temperature and humidity. In each situation one seed of each type is sown and the following results are obtained for the growth (in cm) of the plants after 20 days.

Situation	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Improved seed	3.46	3.48	2.74	2.83	4.00	4.95	2.24	6.92	6.57	6.18	8.30	3.44	4.47	7.59	3.87
Traditional seed	3.18	3.67	2.92	3.10	4.10	4.86	2.21	6.91	6.83	6.19	8.05	3.46	4.18	7.43	3.85

Should improved seeds be used? Respond by justifying and explaining any additional hypotheses that need to be imposed.

Exercise (TPC)

It is intended to test whether the proportion of elms affected by grafiosis is identical in two zones A and B. In zone A, a random sample of 30 elms was taken and 20 were found to be affected by grafiosis. In zone B, a sample of 35 elms was collected and 27 were found to be affected by graphiosis.

What conclusion can you take to the significance level of 0.05?

Goodness of fit tests

The CIs and Hypothesis Testing procedures dealt with so far are usually called parametric methods.

Whenever the procedures used are concerned with the behavior of the population (and not with its parameters) or with the parameters but without requiring knowledge of distributional hypotheses, we say, in general, that we are dealing with nonparametric methods.
These procedures include the ('goodness-of-fit tests') and the tests also called nonparametric tests ('nonparametric tests' or 'distribution -free tests')

Goodness-of-fit tests

We will start by considering the following (‘*goodness-of-fit tests*’) of which we will refer:

- o teste de Shapiro Wilk;
- o teste de Kolmogoroff-Smirnov e
- testes do **Qui-quadrado**

Let’s start with a very important test in our applications - **a normality test** - i.e. allows determine whether a given set of observations can be considered coming from a population with normal distribution – so is a **test of normality**,

The Shapiro Wilk Test

what do you have proved to be one of the most potent. Let us see in summary how process.

The Shapiro-Wilk normality test

Let **X** be the characteristic under study in the population.

The following hypotheses are formulated:
$\mathbf{H}_0$ **:** X has a normal distribution
$\mathbf{H}_1$ **:** X has no normal distribution

The value of the test statistic is calculated $$W_{cal} = \frac{b^2}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}$$

with constant *b* to be determined from the data and using a table (which will not be given here).

We will not use the Shapiro-Wilk Test using tables, we will only run it on R.

The Shapiro-Wilk normality test with R

```
> shapiro.test(azoto)

        Shapiro-Wilk normality test
data:  azoto
W = 0.944, p-value = 0.2001
```

Another fitting test, which, in addition to normality, allows for the adjustment to other distributions taht is very important is the **Kolmogorov-Smirnov test.**
This test, introduced by Kolmogorov and Smirnov in 1933, is based on the difference (distance) between the distribution function postulated in the null hypothesis and the empirical distribution function, built from the sample.

The Kolmogorov Smirnov Test

The hypotheses to be formulated are;
$\mathbf{H}_0$ **:** X has distribution F_0
$\mathbf{H}_1$ **:** X has distribution different from F_0

The test statistic is $D = sup_x|F_0(x) - F_n^*(x)|$, where F_0 designates the considered distribution function and F_n^* the empirical distribution function.

The behavior of that variable D was completely specified by those authors in case X has a normal law with known parameters. Lilliefors studied the same statistics when X was exponential. Currently the softwares are prepared to carry out the test for other distributions.

The Kolmogorov Smirnov with R

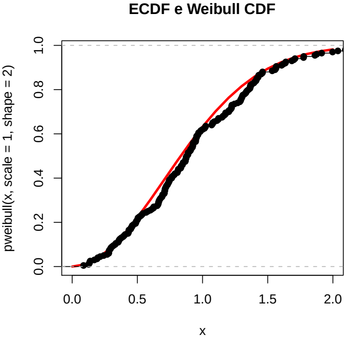
Example: Let’s consider $n = 200$ values ??generated from a Weibull distribution, with the parameters specified below:

```
>x.wei<-rweibull(n=200,shape=2.1,scale=1.1)
+          # gera valores de uma v.a. com
+          # distribuição de Weibull, com
+          # parâmetros de forma=2.1 e escala=1.1
>ks.test(x.wei,"pweibull",shape=2,scale=1)

       One-sample Kolmogorov-Smirnov test
data:  x.wei
D = 0.0765, p-value = 0.1918
alternative hypothesis: two-sided
```

The Kolmogorov Smirnov with R

```
> x<-seq(0,2,0.1)
> plot(x,pweibull(x,scale=1,shape=2),type="l",col="red",
+   lwd=3, main="ECDF e Weibull CDF")
>plot(ecdf(x.wei),add=TRUE)
```



Non-Parametric Tests (just a taste!)

Let us now consider that we intend to perform tests on parameters but the assumptions of normality or approximation to the normal are not verified. As we have already said, such tests are usually called "*distribution-free*" but many authors call them non-parametric tests.

Let's just make a brief note here about two very common tests that can be applied when the $t_Student$ test is not valid.
These tests are based on the orders of the observations, i.e. the position of each observation in the ordered sample.

While parametric tests require the variables in question to be quantitative tests, the nonparametric tests we are going to use can be applied also qualitative variables, provided they are ordinal.

Wilcoxon Tests

Wilcoxon test - non-parametric test for the study of the median of a population or to compare the medians in two paired samples

Wilcoxon-Mann-Whitney test - test no parametric suitable for comparison of two independent samples

`wilcox.test(A,B)` perform the test we called above Wilcoxon-Mann-Whitney for the two independent samples *A* e *B*.

`wilcox.test(A,B,paired=T)` performs the test that we called Wilcoxon for the two samples *A* and *B*, but now considered paired.

Os testes de Wilcoxon no R

The following table gives the percentage of zinc concentration, determined by two different methods, in 9 food samples

Amostra	EDTA tritation	Espectrometria atómica
1	7.2	7.6
2	6.1	6.8
3	4.9	4.8
4	5.9	5.7
5	9.0	9.7
6	8.5	9.1
7	6.6	7.0
8	4	4.7
9	5.2	4.9

Poder-se-á afirmar que existe uma diferença significativa entre os resultados dos dois métodos?