

Experimental Design and Statistical Methods



HYTPOTHESIS TESTING. CONTRAST OF NORMALITY

Jesús Piedrafita Arilla

jesus.piedrafita@uab.cat

Departament de Ciència Animal i dels Aliments

UAB

Items

- Test of hypothesis
 - Null hypothesis and alternative hypothesis
 - Type I and type II errors
- Analyzing normality
 - Numerical tests
 - Quantiles
 - Extreme observations
 - Stem and leaf plots
 - Boxplot
 - Interquartile range
- Basic commands
 - t.test
 - hist, lines, stem
 - boxplot
 - qqnorm
 - skewness, kurtosis
 - shapiro.test, ks.test
- Library
 - moments

Test of Hypothesis

Hypothesis testing is the use of Statistics to determine the probability that a given hypothesis is true. The usual process of hypothesis testing consists of **four steps**.

1. Formulate the **null hypothesis H_0** (commonly, that the **observations are the result of pure chance**) and the **alternative hypothesis H_1** (commonly, that the observations show a **real effect combined** with a component of **chance variation**).
2. Identify a **test statistic** that can be used to assess the truth of the null hypothesis.
3. Compute the **p -value**, which is the probability that a test statistic at least as significant (bigger) as the one observed would be obtained assuming that the null hypothesis were true. The smaller the p -value, the stronger the evidence against the null hypothesis.
4. Construct a **decision rule**: Compare the p -value to an acceptable **significance level α** . If $p \leq \alpha$, that the observed effect is statistically significant, the null hypothesis is rejected, and the alternative hypothesis is accepted.

Test of Hypothesis (cont.)

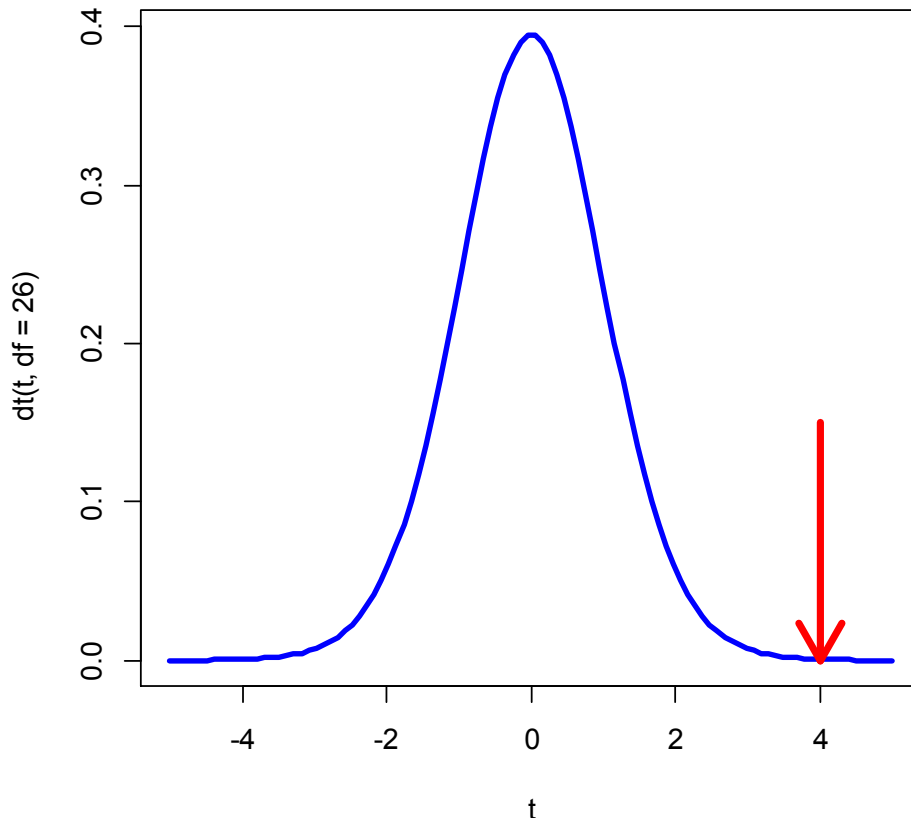
Null hypothesis	Accepted	Rejected
True	Correct decision	Type I error (α) or significance level (False positives)
False	Type II error (β) (False negatives)	Correct decision ($1-\beta$ = Power of test)

Note that:

1. The **asymmetry of the hypothesis**: H_0 is favoured against H_1 . H_0 is rejected when the observations in the sample are inconsistent with the null hypothesis.
2. The decision rule is based upon Type I error.

p-value in a t distribution

```
> t <- seq(-5,5,length=100)
> plot(t,dt(t,df=26),col="blue",lwd=3,type="l")
> arrows(4,0.15,4,0,col="red",lwd=4)
```



The ***p*-value** is the probability of finding a test statistic with a value greater than the observed one. It is the area below the curve starting in the observed value, in our case 4.

p-values and the decision rule

One-sided (right)

1-pt(t, df)	t-test
1-pf(F, df1, df2)	ANOVA
1-pchisq(χ^2 , df)	Contingency tables

Two-sided (left)

2*pt(t, df)	t-test
2*pf(F, df1, df2)	Comparison of variances

Two-sided (right)

2*(1-pt(t, df))	t-test
2*(1-pf(F, df1, df2))	Comparison of variances

Decision rule:

- If *p*-value > 0.05 → accept H_0
- If *p*-value < 0.05 → reject H_0

p-value cartoon

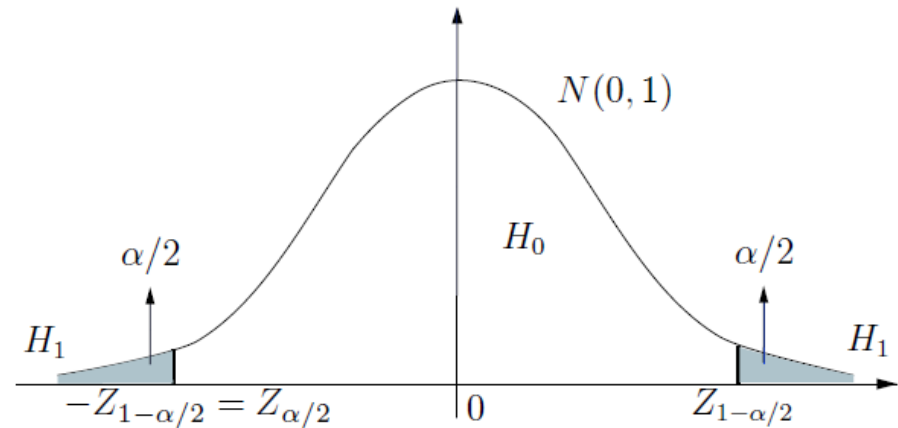
[HTTP://XKCD.COM/1478/](http://xkcd.com/1478/)

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1	

Test for the mean of a normal distribution, known variance (Z-test)

Test statistic:

$$Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$



(Delgado, 2004, p. 117)

Alternative hypothesis

$$H_1 : \mu \neq \mu_0 \quad \rightarrow$$

$$H_1 : \mu > \mu_0 \quad \rightarrow$$

$$H_1 : \mu < \mu_0 \quad \rightarrow$$

Accepted if:

$$z > Z_{1-\frac{\alpha}{2}} \quad \text{or} \quad z < Z_{\frac{\alpha}{2}} \quad \left. \vphantom{z} \right\} \text{Two-sided test}$$

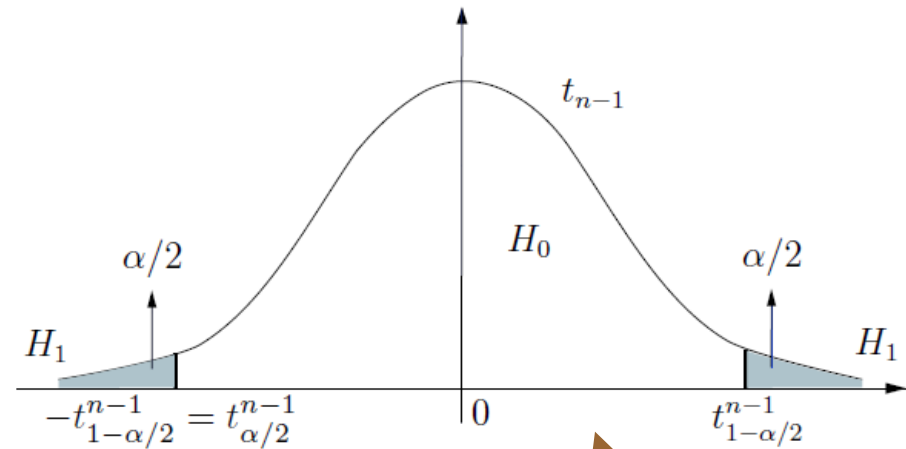
$$z > Z_{1-\alpha} \quad \left. \vphantom{z} \right\} \text{One-sided test}$$

$$z < Z_{\alpha} \quad \left. \vphantom{z} \right\} \text{One-sided test}$$

Test for the mean of a normal distribution, unknown variance (*T*-test)

Test statistic:

$$T = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$



(Delgado, 2004, p. 121)

Alternative hypothesis

$$H_1 : \mu \neq \mu_0 \quad \rightarrow$$

$$H_1 : \mu > \mu_0 \quad \rightarrow$$

$$H_1 : \mu < \mu_0 \quad \rightarrow$$

Accepted if:

$$t > t_{1-\frac{\alpha}{2}}^{n-1} \quad \text{or} \quad t < t_{\frac{\alpha}{2}}^{n-1} \quad \left. \vphantom{t} \right\} \text{Two-sided test}$$

$$t > t_{1-\alpha}^{n-1} \quad \left. \vphantom{t} \right\} \text{One-sided test}$$

$$t < t_{\alpha}^{n-1} \quad \left. \vphantom{t} \right\} \text{One-sided test}$$

Is the mean different from 0?

Student's t

The Student t -test is used to test the null hypothesis that the population **mean** equals μ_0 . The t -statistic is defined to be the difference between the mean and the hypotheses mean (in this case 0) divided by the standard error of the mean. The p -value is the two-tailed probability computed using a t distribution. If the p -value associated with the t -test is small (usually set at $p < 0.05$), there is evidence to reject the null hypothesis in favour of the alternative. In other words, the mean is statistically significantly different than the hypothesized value. If the p -value associated with the t -test is not small ($p > 0.05$), the null hypothesis is not rejected. In our example, our t -value is 48.995 and the corresponding p -value is less than 0.0001. We conclude that there is a statistically significant difference between the mean of the variable **ADG** and 0.

An example with ADG data

Contrasting that the mean is different from 0. In R we can write:

```
> t.test(ADG, mu=0)
```

One Sample t-test

```
data: ADG
```

```
t = 48.9955, df = 37, p-value < 2.2e-16
```

```
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
```

```
1.516678 1.647533
```

```
sample estimates:
```

```
mean of x 1.582105
```

The null hypothesis (H_0) is rejected, as the p -value < 0.05



Note that the value under the null hypothesis (0) is not included into the confidence interval

Computing by hand:

$$t = \frac{\bar{y} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{1.5821 - 0}{\sqrt{\frac{0.0396}{38}}} = \frac{1.5821}{0.0323} = 48.99$$

`pt` gives the cdf of the t distribution.

```
> 2*(1-pt(48.99, 37))  
[1] 0
```

An example with ADG data (cont.)

Contrasting that the mean is different from 1.6. In R we can write:

```
> t.test(ADG, mu=1.6)
```

```
One Sample t-test
```

```
data: ADG
```

```
t = -0.5542, df = 37, p-value = 0.5828
```

```
alternative hypothesis: true mean is not equal to 1.6
```

```
95 percent confidence interval:
```


```
1.516678 1.647533
```

```
sample estimates:
```

```
mean of x
```

```
1.582105
```

The null hypothesis (H_0) is not rejected, as the p -value > 0.05



Note that the value under the null hypothesis (1.60) is within the confidence interval

```
> 2*pt(-.5542,37)  
[1] 0.5827764
```

Contrast of normality

So far we have seen how normality is assumed both for variables and the distribution of parameter estimates. The last one is determined by statistical reasoning, but the adjustment of real data to a certain distribution must be tested. In the case of the **normal distribution**:

1. Graphic tests:

- i. Histogram
- ii. Stem and leaf plot
- iii. Box-plot
- iv. QQ- plot

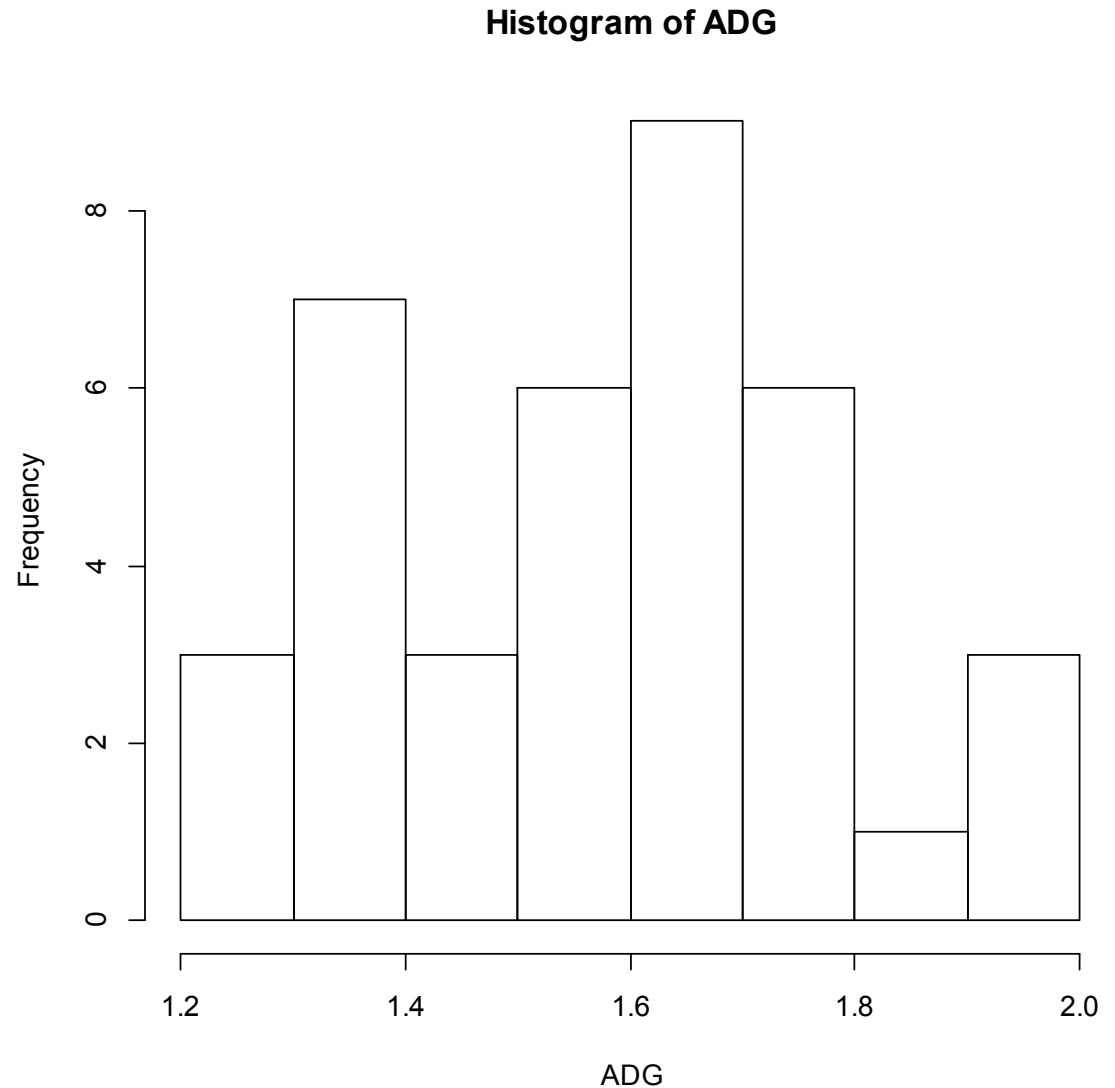
2. Numerical tests (in addition to skewness and kurtosis):

- i. Shapiro-Wilk (for sample sizes 7 – 2000)
- ii. Kolmogorov-Smirnov
- iii. Other tests (Cramer-Von Mises, Anderson-Darling)

Contrasting normality. Histogram (1)

```
> hist(ADG)
```

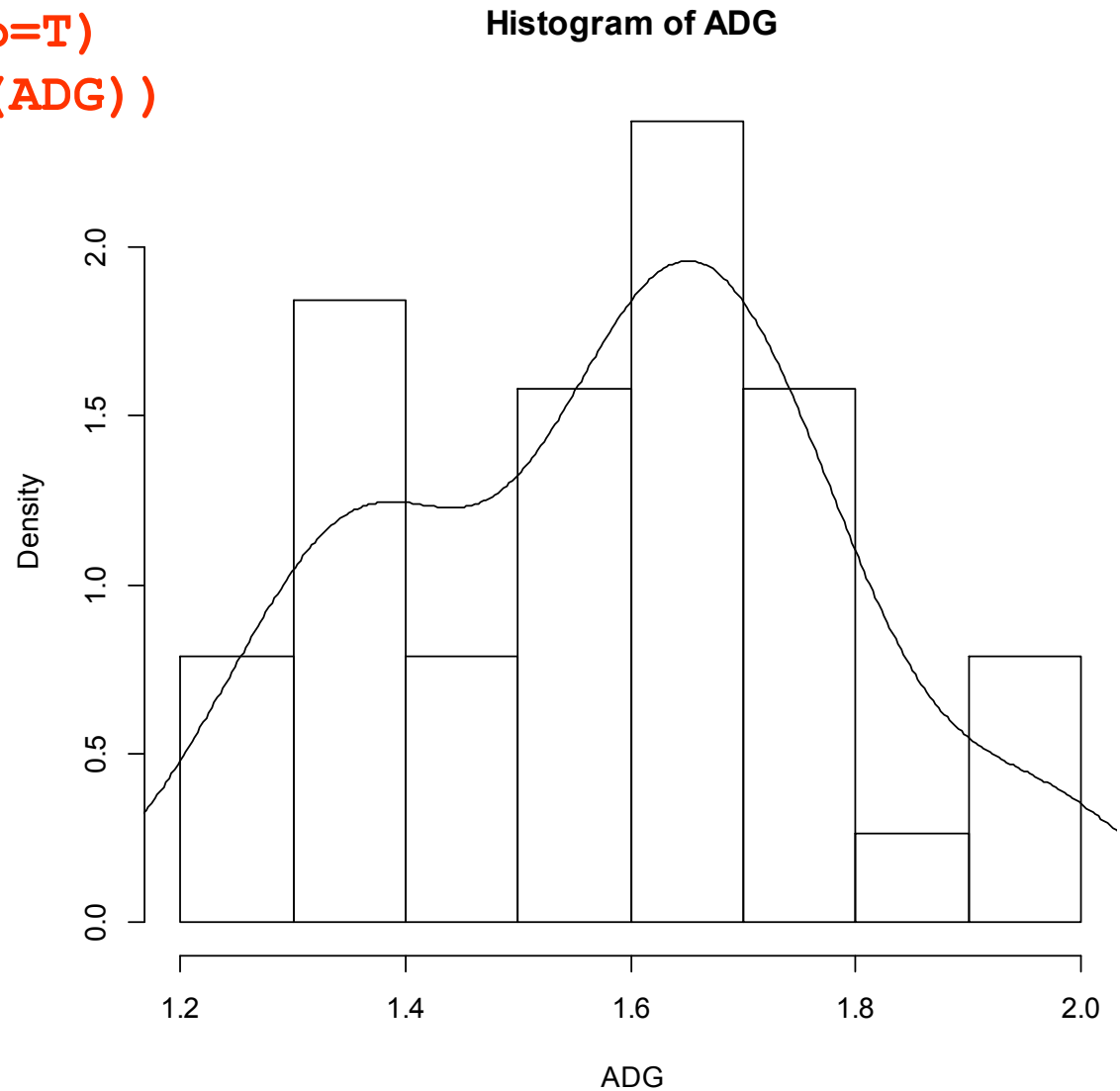
This chart gives a first look at the shape of the distribution



Contrasting normality. Histogram (2)

```
> hist(ADG, prob=T)  
> lines(density(ADG))
```

This chart differs from the previous one in that the y-axis contains the density and there is also a smooth line describing the distribution



Contrasting normality. Stem and leaf plot

> stem(ADG)

The decimal point is 1 digit(s) to the left of the |

```
12 | 14
13 | 02256789
14 | 35
15 | 017888
16 | 0113478889
17 | 124688
18 | 1
19 | 559
```

Stem Leaf

This representation reminds an histogram and allows us to check the data, helping to detect incorrect observations.

This line tell us that we have two observations with a value of 1.95 and one observation that is 1.99

Some more theory on the normal distribution

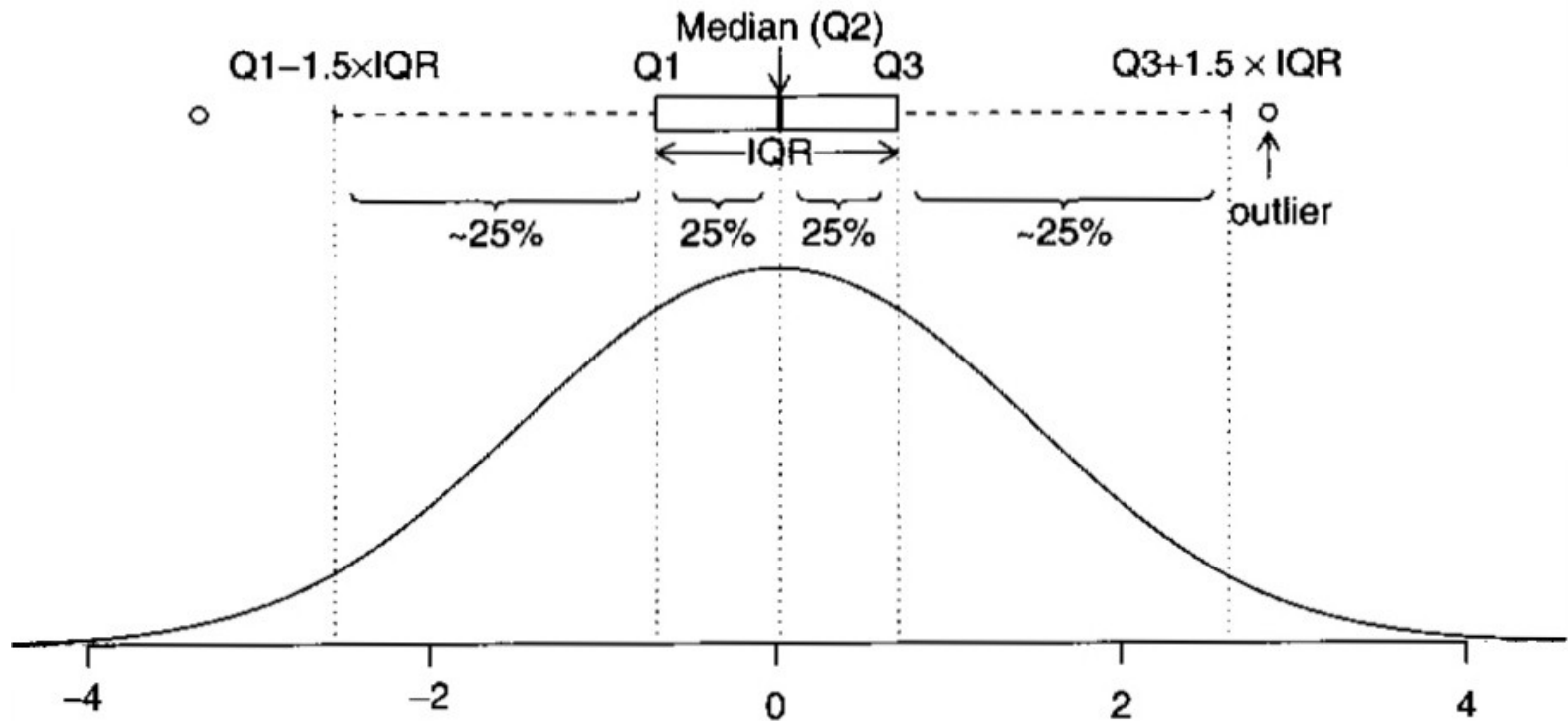
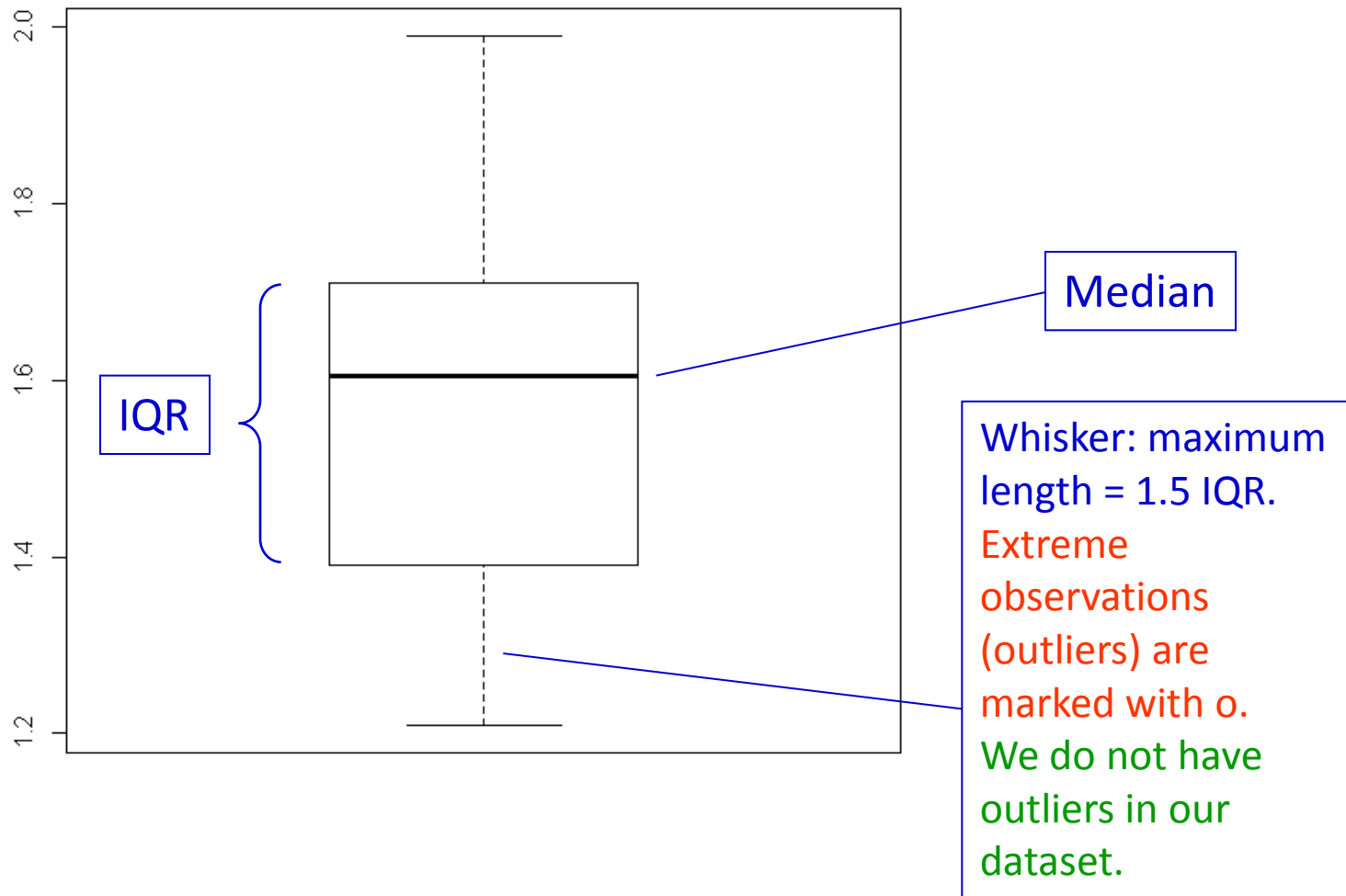


Fig 5.5 Boxplot of a standard normal distribution (mean=0, sd=1).

(Logan, 2010)

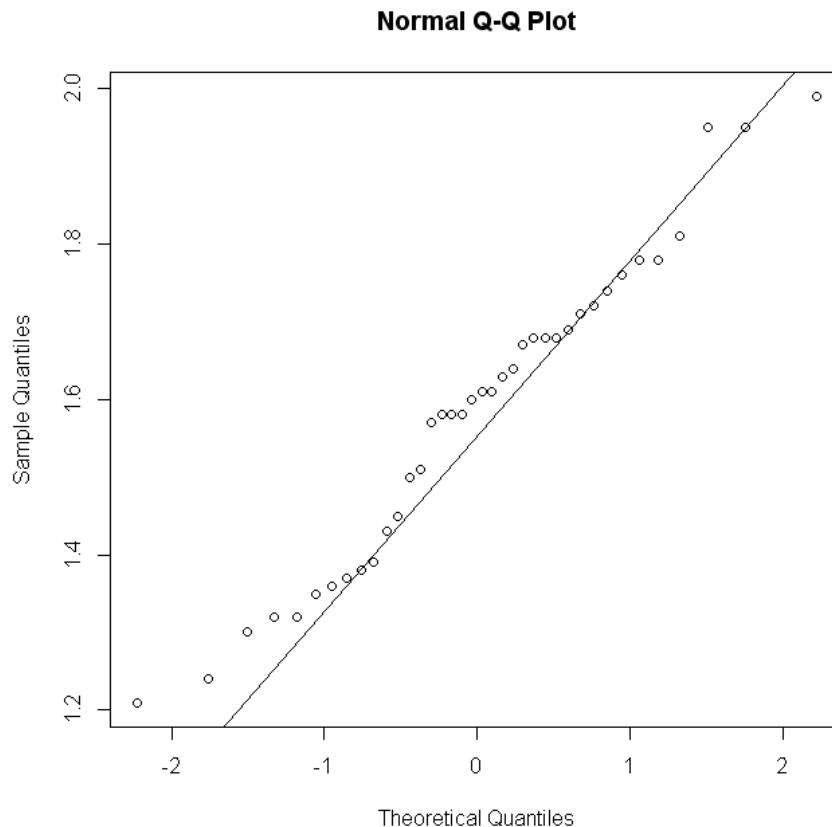
Contrasting normality. Box (and whiskers)-plot

> `boxplot(ADG)`



Contrasting normality. Q-Q plots

```
> qqnorm(ADG) ; qqline(ADG)
```

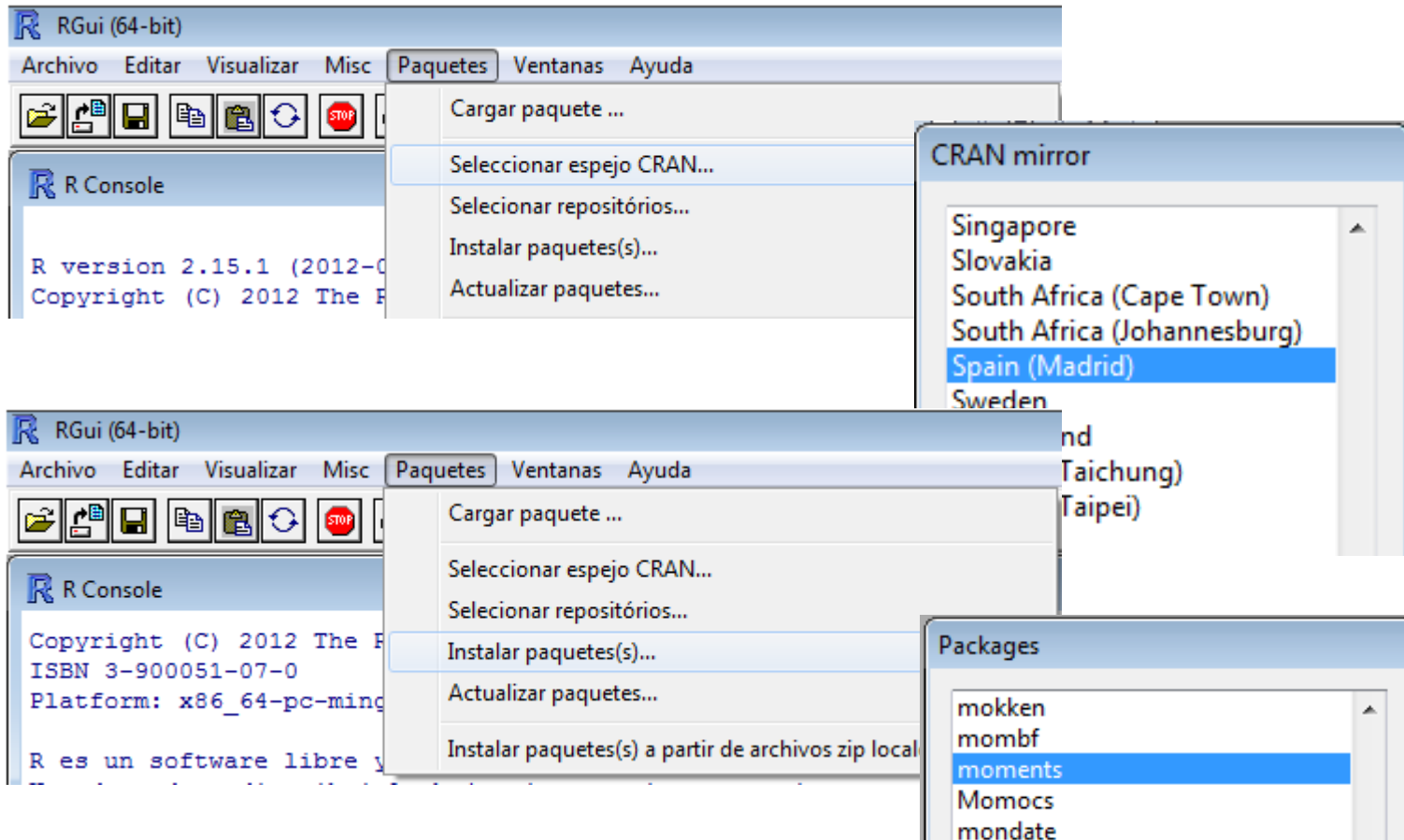


Observe how ; separates two commands in the same line

Q-Q stands for quantile vs quantile plot

A straight line is expected if data come from a normal distribution with *any* mean and standard deviation.

Installing a new package (library)



Skewness and kurtosis

```
> library(moments)
```

```
> skewness(ADG)
```

```
[1] 0.03140257
```

```
> kurtosis(ADG) - 3
```

```
[1] -0.6525667
```

To calculate skewness and kurtosis, the library “moments” must be activated .

To charge a library that is not in the basic package of R, you must go to the console, look for “Packages” and follow the instructions.

Both skewness and kurtosis have values that do not deviated much from 0, i.e., the distribution is fairly symmetric and the tails are not really heavy.

Numerical tests

```
> shapiro.test(ADG)
```

```
Shapiro-Wilk normality test
```

```
data: ADG  
W = 0.9709, p-value = 0.4159
```

The null hypothesis (H_0) is that the distribution is normal

```
> ks.test(ADG, "pnorm", mean = mean(ADG), sd = sqrt(var(ADG)))
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: ADG  
D = 0.1073, p-value = 0.7738  
alternative hypothesis: two-sided
```

```
Mensajes de aviso perdidos
```

```
In ks.test(ADG, "pnorm", mean = mean(ADG), sd = sqrt(var(ADG))) :  
ties should not be present for the Kolmogorov-Smirnov test
```

Both tests tell us that ADG data are normally distributed
→ In this case H_0 cannot be rejected.

An script for the univariate analysis

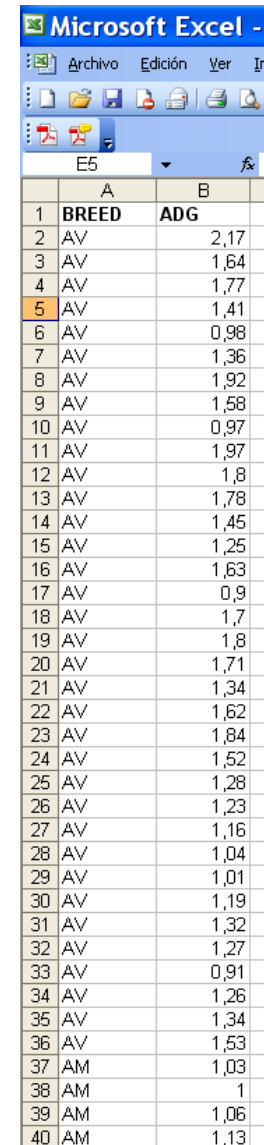
```
R D:\Documents and Settings\jpiedrafita\Escritorio\R\suniva...
#### DESCRIPTIVE ANALYSIS OF A SINGLE DATASET
#Data
ADG <- c(1.99, 1.72, 1.95, 1.67, 1.51, 1.32, 1.39, 1.64, 1.78, 1.50, 1.43,
1.37, 1.60, 1.58, 1.76, 1.57, 1.81, 1.21, 1.45, 1.58, 1.58, 1.68, 1.61,
1.61, 1.78, 1.95, 1.63, 1.68, 1.71, 1.74, 1.69, 1.68, 1.36, 1.30, 1.35,
1.24, 1.38, 1.32)
#Descriptive statistics
length(ADG)
min(ADG); max(ADG);
mean(ADG); median(ADG)
var(ADG); sd(ADG); cv<-sd(ADG)/mean(ADG)*100; cv
quantile(ADG); IQR(ADG)
#Inference about the mean
SEM<-sd(ADG)/sqrt(length(ADG)); SEM
XBAR<-mean(ADG);
LCL<-XBAR-SEM*qt(0.975,length(ADG)-1); LCL
UCL<-XBAR+SEM*qt(0.975,length(ADG)-1); UCL
#Numerical tests for normality
library(moments); skewness(ADG); kurtosis(ADG)-3
shapiro.test(ADG)
ks.test(ADG, "pnorm", mean = mean(ADG), sd = sqrt(var(ADG)))
#Graphic tests for normality
hist(ADG); dev.new()
stem(ADG)
boxplot(ADG); dev.new()
qqnorm(ADG); qqline(ADG)
```

Note that `>` is not needed in a script

`dev.new()` opens a new window for a new graphic

Contrasting normality of several groups. 1. The data.

Suppose we have a sample of ADG measurements in bulls belonging to 6 Spanish beef breeds in an Excel file:



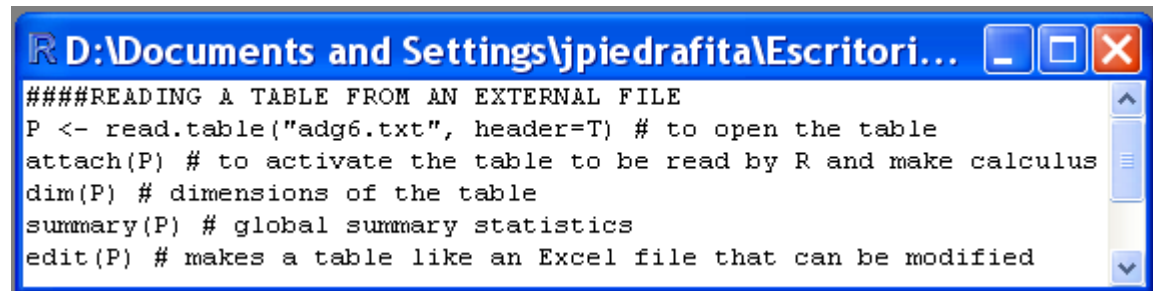
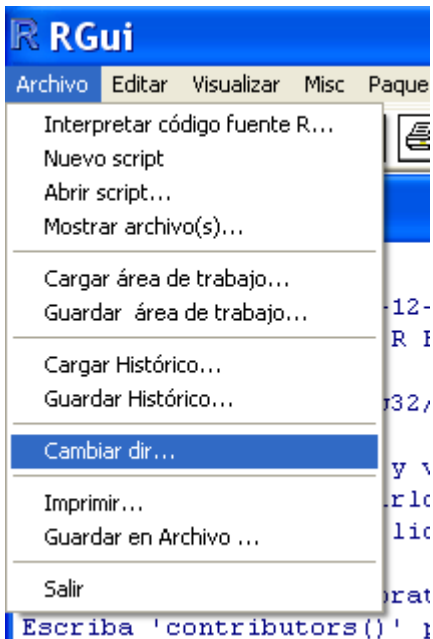
	A	B
1	BREED	ADG
2	AV	2,17
3	AV	1,64
4	AV	1,77
5	AV	1,41
6	AV	0,98
7	AV	1,36
8	AV	1,92
9	AV	1,58
10	AV	0,97
11	AV	1,97
12	AV	1,8
13	AV	1,78
14	AV	1,45
15	AV	1,25
16	AV	1,63
17	AV	0,9
18	AV	1,7
19	AV	1,8
20	AV	1,71
21	AV	1,34
22	AV	1,62
23	AV	1,84
24	AV	1,52
25	AV	1,28
26	AV	1,23
27	AV	1,16
28	AV	1,04
29	AV	1,01
30	AV	1,19
31	AV	1,32
32	AV	1,27
33	AV	0,91
34	AV	1,26
35	AV	1,34
36	AV	1,53
37	AM	1,03
38	AM	1
39	AM	1,06
40	AM	1,13

Observe how each ADG value has a label (left column) corresponding to the breed.

(continues)

Steps to create a data frame in R

1. Put the data in columns in an Excel file, one for each variable (in our case BREED and ADG), with the corresponding header (see previous slide).
2. If there are missing observations, fulfil them with NA.
3. Save the file as a .txt, i.e., delimited by tabulations.
4. Edit the file with Notepad and check that the decimals are separated by points (.). In case the decimals are separated with a comma (,), then substitute all commas by points. Save the file in a directory with name R (for example).
5. Change to the appropriate directory in R before running the following script:

A screenshot of an R script editor window. The title bar shows the path 'R D:\Documents and Settings\jpiedrafita\Escritori...'. The script content is as follows:

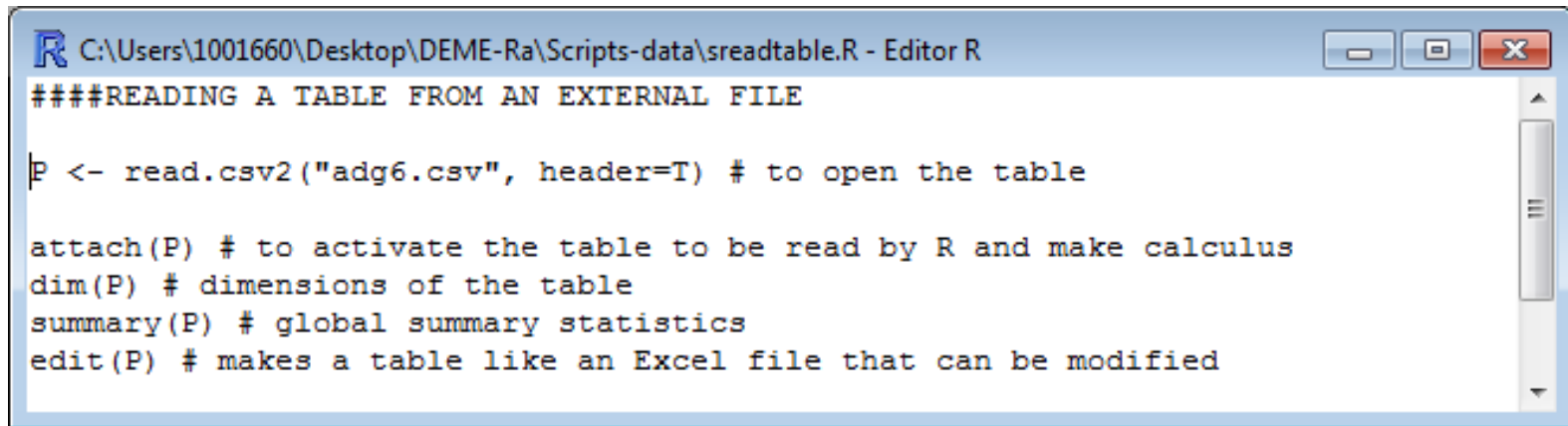
```
#####READING A TABLE FROM AN EXTERNAL FILE
P <- read.table("adg6.txt", header=T) # to open the table
attach(P) # to activate the table to be read by R and make calculus
dim(P) # dimensions of the table
summary(P) # global summary statistics
edit(P) # makes a table like an Excel file that can be modified
```

Note that the file “**adg6.txt**” must be in the same directory that `sreadtable.R`.

It is recommended to create a directory in which we save both data and script files for a certain analysis.

Steps to create a data frame in R, easier!

1. Put the data in columns in an Excel file, one for each variable (in our case BREED and ADG), with the corresponding header (see previous slide).
2. If there are missing observations, fulfil them with NA.
3. Save the file as a csv, i.e., “comma separated values”.
4. Change to the appropriate directory in R before running the following script:



```
C:\Users\1001660\Desktop\DEME-Ra\Scripts-data\sreadtable.R - Editor R
####READING A TABLE FROM AN EXTERNAL FILE

P <- read.csv2("adg6.csv", header=T) # to open the table

attach(P) # to activate the table to be read by R and make calculus
dim(P) # dimensions of the table
summary(P) # global summary statistics
edit(P) # makes a table like an Excel file that can be modified
```

read.csv2 allows to read csv files from “Spanish” Excel, where decimals are separated by a comma (,) and the data columns are separated by a semicolon (;). Using the conventional Excel, you must use **read.csv**.

Note that the file “**adg6.csv**” must be in the same directory that **sreadtable.R**.

It is recommended to create a file in which we save both data and script files for a certain analysis.

Basic statistics and boxplots for several groups (1)

```
R C:\Users\1001660\Desktop\DEME-Ra\Scripts-data\stable-summary.R - Editor R
####TABLE OF DESCRIPTIVE STATISTICS
####of a continuous variable (ADG) according to several
####levels of a clasificatory variable (BREED)

ADG.TABLE<-read.csv2("adg6.csv", header=T)
attach(ADG.TABLE)

#Calculating the elements of a summary table
M<-tapply(ADG, BREED, length)
N<-tapply(ADG, BREED, min)
O<-tapply(ADG, BREED, max)
P<-tapply(ADG, BREED, mean)
Q<-tapply(ADG, BREED, median)
R<-tapply(ADG, BREED, sd)
S<-tapply(ADG, BREED, shapiro.test)
SS<-array(0,dim=c(length(levels(BREED)))) #Creates a vector of 6 zeroes
for(i in 1:length(levels(BREED))) {SS[i] <- S[[i]]$p.value}

#Combining the results into a table
cbind(n=M, min=N, max=O, mean=P, median=Q, std.dev=R, p.shapiro=SS)

#Boxplots per groups
boxplot(ADG~BREED)
```

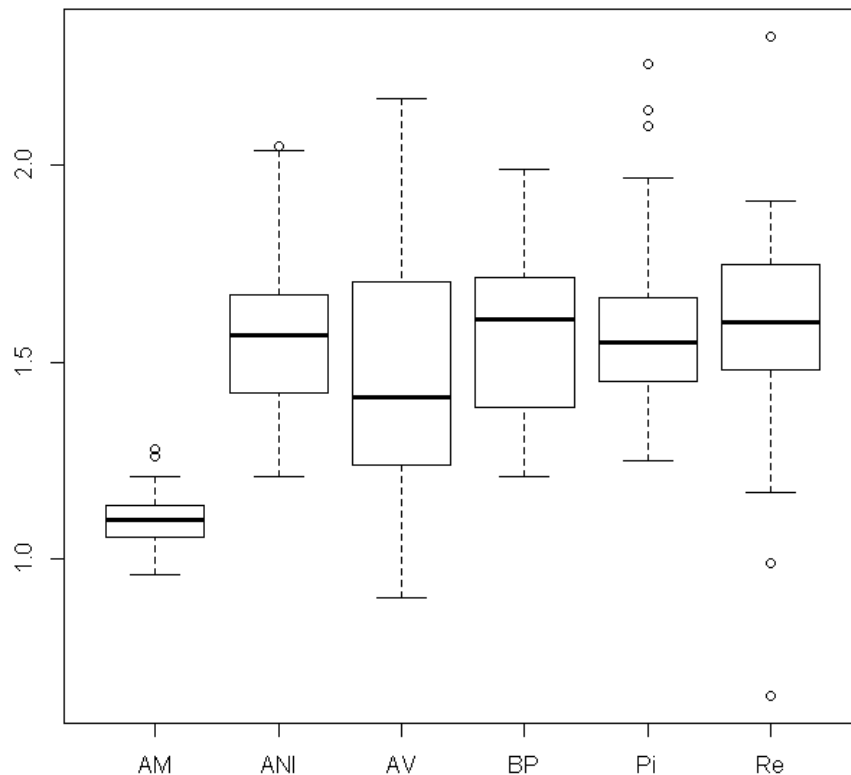
tapply applies a function to each cell of an array

cbind combines columns

~ (ALT+126 in numerical pad) corresponds to “**described by**”

Basic statistics and boxplots for several groups (2)

	n	min	max	mean	median	std.dev	p.shapiro
AM	35	0.96	1.28	1.101429	1.10	0.07904365	0.713089611
ANI	35	1.21	2.05	1.568571	1.57	0.20396284	0.221362022
AV	35	0.90	2.17	1.447143	1.41	0.32849542	0.545417650
BP	36	1.21	1.99	1.585833	1.61	0.20340494	0.312702869
Pi	27	1.25	2.26	1.607407	1.55	0.26440925	0.007537896
Re	35	0.65	2.33	1.586000	1.60	0.29345007	0.032705124



To note:

1. The differences among means and amplitude of distribution.
2. The presence of **outliers** (circles), that not always gives deviation to normality in the Shapiro test.

Selecting a sub-set in a *data-frame*

We can be interested in some particular observations of a data-frame, for example those corresponding to the AM breed. In R we can do that as follows:

```
> ADG.AM <- ADG.TABLE[BREED=="AM" , ]  
> ADG.AM
```

	BREED	ADG
36	AM	1.03
37	AM	1.00
38	AM	1.06
39	AM	1.13
40	AM	1.21
41	AM	1.11
42	AM	1.11
43	AM	1.05

Observe that a new subset of data is generated (**ADG.AM**) in which we can analyze *ADG* or some other variables contained in the sub-set, but only for this breed.

Note that R maintains the original number of the observation in the global data-frame.

Note also that the output presented in this slide is not complete and only includes the first 8 observations.

The student is invited to look for the multiple options that R offers to handle data (internet helps a lot).

Save the information generated in R

1. Open a Word document.
2. After executing the script or some part of it, copy the program (in red) and the results (in black) that you have in the console.
3. Paste them in the Word document. Probably you will get a square with Times New Roman type (depending upon the PC).
4. Change type to **Courier New**, size 10. This action will reorganize the document and will make it similar to what you see in the console. **Courier new** is a type of fixed space, different from other types.
5. If you see some line that continues in the next line, try to reduce margins, for example to 2 cm in each side, or even less, according to what you need to put the information in the same line.
6. To copy a graphic we must put the cursor on it and click the right mouse button. We will obtain a menu from which we will select any one of “Copy as a metafile” or “Copy as a bitmap”, and then paste it in the proper location of the Word document.
7. Save the Word document. A mnemonic name is recommended.

References

- Delgado R. 2004. *Iniciación a la Probabilidad y la Estadística*.
Servei de Publicacions UAB, Materials 153, Barcelona.
- Logan M. 2010. *Biostatistical Design and Analysis Using R*.
Wiley-Blackwell, Chichester.