R commander an Introduction

Natasha A. Karp
nk3@sanger.ac.uk
May 2010

Preface

This material is intended as an introductory guide to data analysis with R commander. It was produced as part of an applied statistics course, given at the Wellcome Trust Sanger Institute in the summer of 2010. The principle aim is to provide a step-by-step guide on the use of R commander to carry out exploratory data analysis and the subsequent application of statistical analysis to answer questions widely asked in the life sciences.

These notes (version 1.1) were written with R commander version 1.4-10 under a Window's operating system. This document is available for download from the Comprehensive R Archive Network (http://cran.r-project.org/) and is provided free-of-charge with no warrantee for its use. It is not to be modified from this form without explicit authorization from the author.

Natasha A. Karp
Biostatistician
Mouse Genetics Group
Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SA
nk3@sanger.ac.uk

Content

1. Starting R commander and importing data

1.1 What is R Commander?

R commander is free statistical software. R commander was developed as an easy to use graphical user interface (GUI) for R (freeware statistical programming language) and was developed by Prof. John Fox to allow the teaching of statistics courses and removing the hindrance of software complexity from the process of learning statistics. This means it has drop down menus that can drive the statistical analysis of data. It is considered the most viable R-alternative to commercial statistical packages like SPSS (Wikipedia). The package is highly useful to R novices, since for each analysis run it displays the underlying R code.

Home page: http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/

It also has a series of plug-ins which extend the range of application

- RcmdrPlugin.Export — Graphically export objects to LaTeX or HTML
- RcmdrPlugin.FactoMineR — Graphical User Interface for FactoMineR
- RcmdrPlugin.HH — Rcmdr support for the HH package
- RcmdrPlugin.IPSUR — Introduction to Probability and Statistics Using R
- RcmdrPlugin.SurvivalT — Rcmdr Survival Plug-In
- RcmdrPlugin.TeachingDemos — Rcmdr Teaching Demos Plug-In
- RcmdrPlugin.epack — Rcmdr plugin for time series
- RcmdrPlugin.orloca — orloca Rcmdr Plug-in

1.2 References and additional reading material

• "The R Commander: A Basic-Statistics Graphical User Interface to R" John Fox
Journal of Statistical Software 2005, Volume 14, Issue 9.
• http://sociology.osu.edu/computing/helpDocs/rcmdr.pdf
• http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf
• http://courses.statistics.com/software/RCommander/RC00.htm
• http://www.eau.ee/~ktanel/DK_0007/DK_prax4_2009.pdf

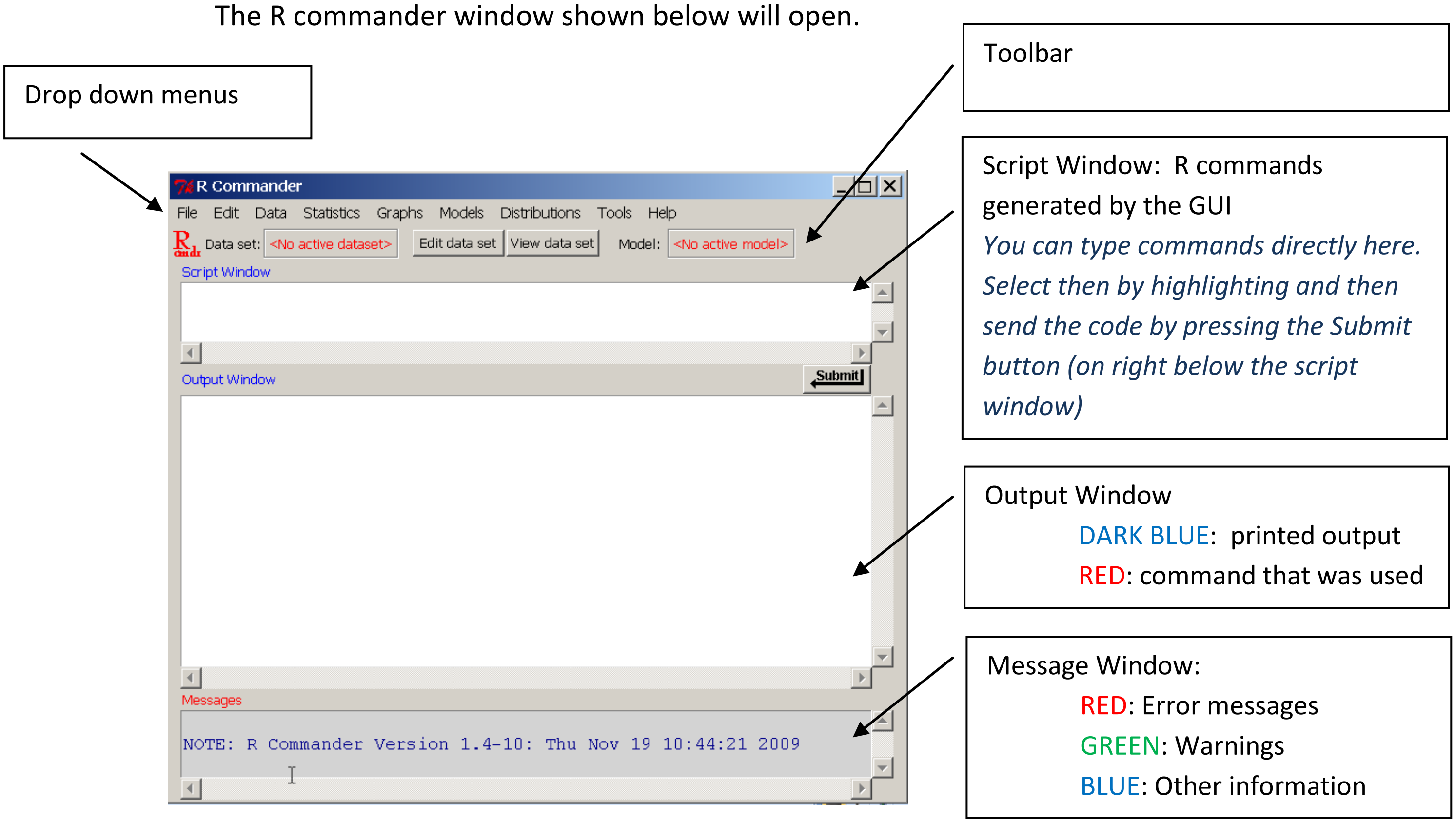
1.3 Installing R commander

You need to first install R and then R commander.
The following link provides good instructions for installation of R:
http://jekyll.math.byuh.edu/other/howto/R/R.shtml
The following link provides good instructions for installation of R commander:
http://jekyll.math.byuh.edu/other/howto/R/Rcmdr.shtml

1.4 Starting the R Commander

i. Open R program
   e.g. double click on R icon or start/all programs/R
ii. To open the R commander program type at the prompt library("Rcmdr") and press return.

The R commander window shown below will open.



Note: Graphs will appear in a separate Graphics Device Window. Only the most recent graph will appear. You can use page up and page down keys to recall previous graphs.

Drop down Menu item	
File	Menu items for loading and saving script files; for saving output and the R workspace; and for exiting.
Edit	Menu items (Cut, Copy, Paste, etc.) for editing the contents of the script and output windows. Right clicking in the script or output window also brings up an edit “context” menu
Data	Submenus containing menu items for reading and manipulating data.
Statistics	Submenus containing menu items for a variety of basic statistical analyses.
Graphs	Menu items for creating simple statistical graphs.
Models	Menu items and submenus for obtaining numerical summaries, confidence intervals, hypothesis tests, diagnostics, and graphs for a statistical model, and for adding diagnostic quantities, such as residuals, to the data set. Distributions Probabilities, quantiles, and graphs of standard statistical distributions (to be used, for example, as a substitute for statistical tables).
Tools	Menu items for loading R packages unrelated to the Rcmdr package (e.g., to access data saved in another package), and for setting some options.
Help	Menu items to obtain information about the R Commander (including an introductory manual derived from this paper). As well, each R Commander dialog box has a Help button.

Toolbar buttons	
Data set	Shows the name of the active dataset Button: allows you choose among dataset currently in memory which to be active
Edit data set	Allows you to open the active dataset
View data set	Allows you to view the active dataset
Model	Shows the name of the active statistical model e.g. linear model Button: allows you to choose among current models in memory

Menu items are inactive (ie, greyed out) if not applicable to the current context.

1.5 Data input

1.5.1 Manual entry

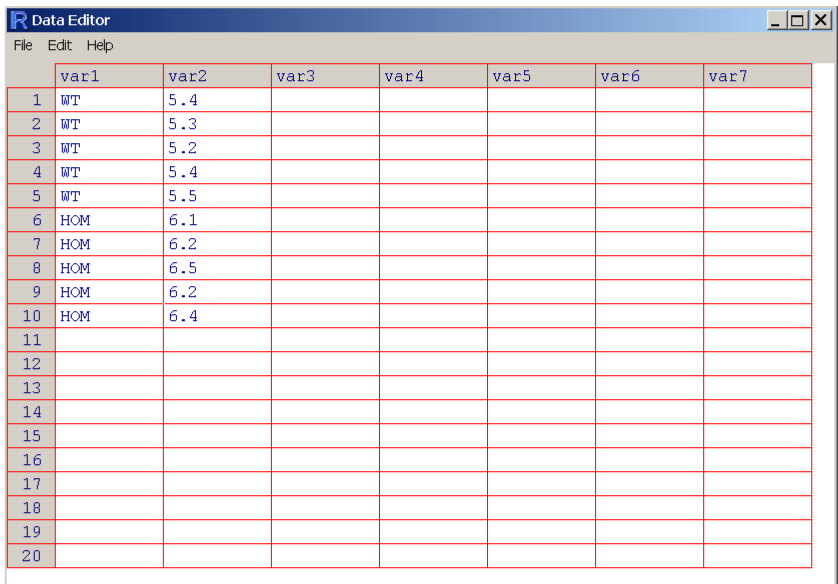
i. Start a new data set through Data -> New data set

ii. Enter a new name for the dataset -> OK

Note: the name cannot have spaces in it

Note: R is case-sensitive hence mydata ≠MyData

iii. A data editor window where you can type in your data using a typical spreadsheet format. Each row corresponds to an independent object e.g. a subject on which a measurement was made.



	var1	var2	var3	var4	var5	var6	var7
1	WT	5.4					
2	WT	5.3					
3	WT	5.2					
4	WT	5.4					
5	WT	5.5					
6	HOM	6.1					
7	HOM	6.2					
8	HOM	6.5					
9	HOM	6.2					
10	HOM	6.4					
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							

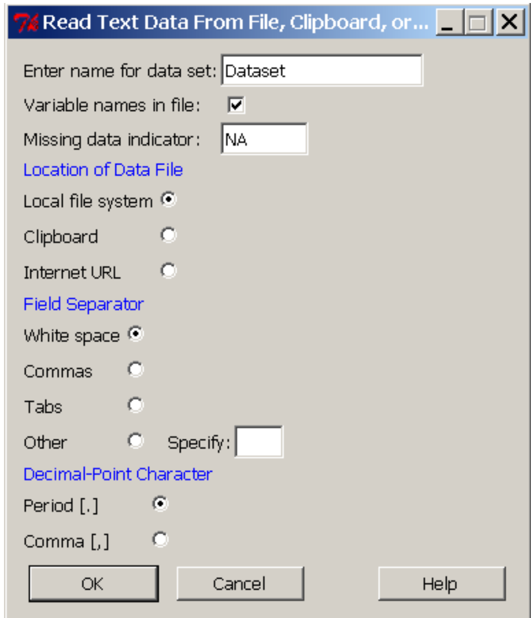
iv. Define the variables (column) by clicking on the column label and then in the resulting dialog box enter the name and type. Where type can be numeric (quantitative) or character (qualitative). Click on the x in the right hand corner to close this dialog box.

v. This data frame is then the active dataset for R commander.

1.5.2 Import from text file

Note: the data file will need to be organized as a classic data frame. Each column represents a single variable e.g. glucose level. Each row represents an individual. The header information needs to be contained in a single row.

i. Data -> Import data -> from text file



ii. Chose a name for the new dataset (note you cannot have spaces)

iii. Specify the characteristics of the data files (e.g. commas for csv files) -> OK

iv. Browse and select the file/Open

Once data is imported you should double-check the file was read-in correctly:

v. Message window: are there any errors?

vi. Do the number of rows and columns look as expected?

vii. View the data via View data set button

1.5.3 Import from Excel

Data files can be read in from Excel, however they often have issues. It is recommended that instead the file is converted to a text file and then import as detailed in 1.5.2.

How?

1. Within Excel: Office -> Save As and select the comma-delimited (.csv) file format.

2. Using R Commander to obtain descriptives

Role of descriptives?

1. Checking for errors

    Looking for values that fall outside the possible values for a variable

    Looking for excess number of missing values

2. As descriptives

    To describe the sample in your report

    To address specific research questions

2.1 Checking categorical variables

i. Statistics -> Summaries -> Frequency Distribution -> Select the variables->OK

ii. Output: For each variable you selected it will tell you the frequency for each level.

The red text following prompt:

R code used to generate output

Red text following #:

Explanation of what the code is doing

The output of analysis is shown in blue

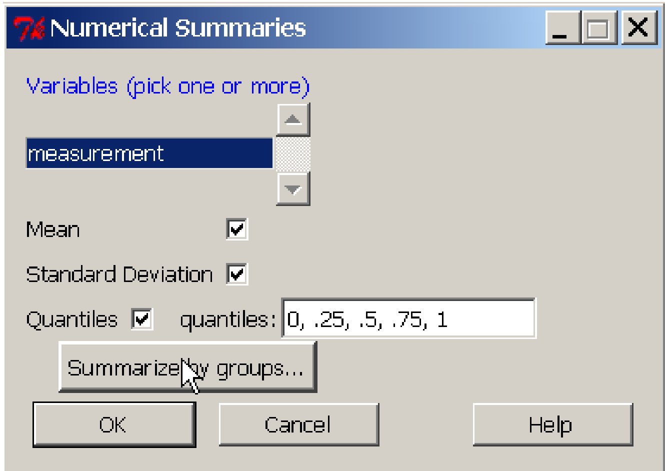
```
> .Table  # counts for Headmorphology

abnormal   normal  missing
       2       18        1

> 100*.Table/sum(.Table)  # percentages for Headmorphology

 abnormal    normal   missing
 9.523810 85.714286  4.761905
```

v. Check for unexpected levels e.g. norm rather than normal.

vi. Check the number of missing values does it seem appropriate?

2.2 Checking continuous variables

i. Statistics -> Summaries -> Numerical summary



ii. If you have multiple groups (e.g. control versus treatment) click on summarize by groups and select the appropriate variable -> OK

Output:

```
        mean        sd   0%  25%  50%  75% 100%  n NA
HOM 4.097778 0.5423969 3.26 3.83 3.99 4.38 5.12  9  1
WT  5.089091 0.6669401 4.11 4.78 4.88 5.37 6.51 11  0
```

Understanding the output:

parameter	What is it?
mean	Measure of central tendency
sd	Standard deviation - a measure of variability in the data
N	Number of readings
NA	Number of missing values
0%	Minimum value
25%	The value below which 25 percent of the observations may be found.
50%	The value below which 50 percent of the observations may be found.
75%	The value below which 75 percent of the observations may be found.
100%	Maximum value

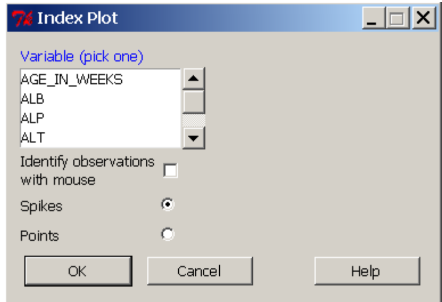
iii. Check your minimum and maximum values – do they make sense?

iv. Check the number of missing values – if there are a lot of missing values you need to ask why?

v. Do the mean score(s) make sense? Is it what you expect from previous experience?

vi. Identifying the outlier

Graphs -> Index Plot



vii. Select the variable of concern

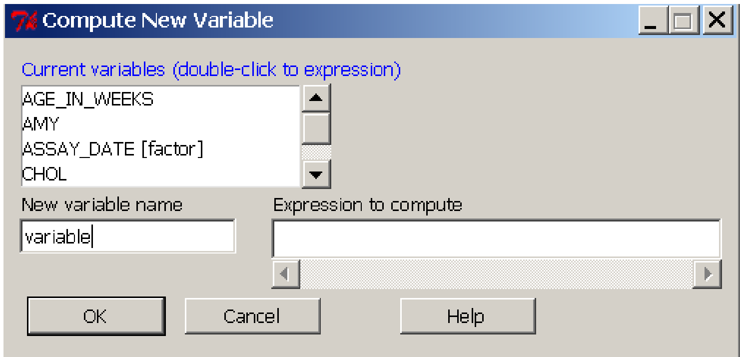
viii. Tick identify observations with mouse

ix. Look at the graphical output and click the mouse on the observation that is the outlier for it index number.

3. Modifying the dataset

3.1 Compute a new variable

i. Data -> Manage variables in active dataset -> compute new variables



ii. Enter new variable name

iii. An expression (equation) is written to reflect the calculation required. The table below indicates the operators available and examples of how it could be used. Note: Double clicking on a variable in the current variables box will send the variable to the expression.

Operators	Function	Example 1	Example 2
x + y	Addition	Variable 1 + Variable 2	Variable 1 + 25
x - y	Subtraction	Variable 1 – Variable 2	35 - Variable 1
x * y	Multiple	Variable 1*Variable 2	100*Variable 1
x / y	Division	Variable 1/Variable 2	Variable 1 / 63
x ^ y	X to the power of Y	Variable 1 ^ Variable2	Variable1^10
log10(x)	Log10 transformation	Log10(Variable 1)	
log(x, base)	Log transformation to a specified base	Log(Variable 1, 2)	

3.2 Converting numeric variables to categorical variables

Categorical variables are measures on a nominal scale i.e. where you use labels. For example, rocks can be generally categorized as igneous, sedimentary and metamorphic. The values that can be taken are called levels. Categorical variables have no numerical meaning but are often coded for easy of data entry and processing in spreadsheets. For example gender is often coded where male =1 and female = 2. Data can thus be entered as characters (e.g. 'normal') or numeric (e.g. 0, 1, 2). It is important to ensure the program distinguishes between categorical variables entered numerically and those variables whose values have a direct numerical meaning.

Assessing whether a variable is entered as categorical:

i. Statistics -> Summaries -> Frequency Distribution

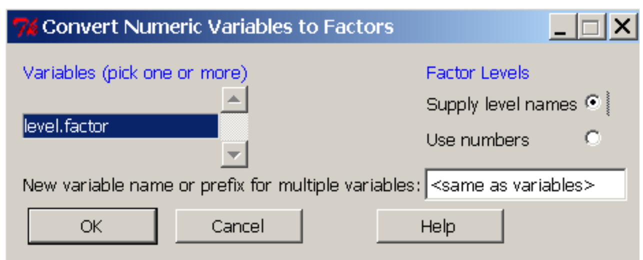
Only categorical variables will be listed

OR

ii. Edit Data Set -> click on each row header and it will tell you it is numeric/categorical

Converting numeric variables to factors:

i. Data -> Manage variables in active data set -> Convert numeric variables to factors…

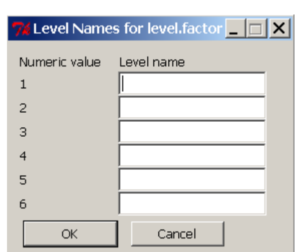


ii. Select the variables

iii. You can generate a new variable by entering a name in box "new variable name…." or over-write the original name.

1. The levels can be formatted as Levels by selecting 'use numbers'
2. Recoded to a name by selecting 'supply level names'

   If this is selected another dialog box will appear to enter the name for each numeric value.

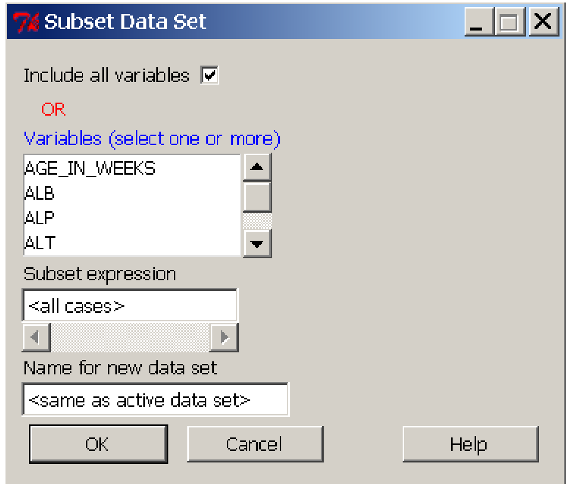


iv. OK

3.3 Sub-dividing data

3.3.1 by columns (variables)

i. data -> active dataset -> subset active dataset

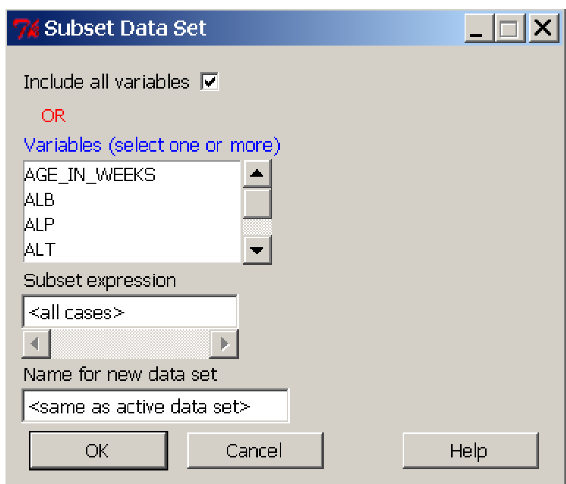


ii. Hold the CTRL key to select the variables you wish to keep

iii. Give the new dataset a name -> OK

3.3.2 by rows (and variables if you wish)

i. Data -> active dataset -> subset active dataset



ii. Select the variables you wish to include in the new dataset

iii. Write a ‘subset expression’ which is a rule to drive the selection of rows

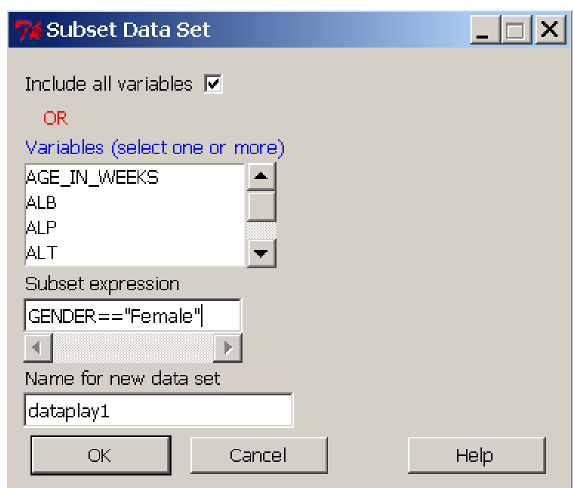
Symbol/code	Name	Use
==	equality	used to indicate the variable should equal
!=	Inequality	used to indicate the variable should not equal
&	And	used to combine multiple expressions
|	Or	used to combine multiple expressions
is.na(varname)		Include the missing values of a variable
!is.na(varname)		Exclude the missing values of a variable
>	Greater than	
<	Less than	
>=		More than or equal to
<=		Less than or equal to

Note 1: If you use a name in an expression you need to surround the name with double quotes e.g. "name".

Note 2: the variable name is case-sensitive (i.e. it has to match exactly the name used as a column header).

Example: GENDER == "Female"

Example 2: GENDER == "Female" & AGE <= 25



iv. Give the dataset a new name -> OK.

4. Using R Commander to explore data

4.1 Graphically

The R commander is able to generate a variety of basic statistical graphs. The graphic output in R commander is limited by the choice offered in the menu. There are too many options to be incorporated sensible. Whilst in R, using the command line, the options are endless. If this becomes an issue I would recommend speaking to an R user, or using books, and web resources to learn more.

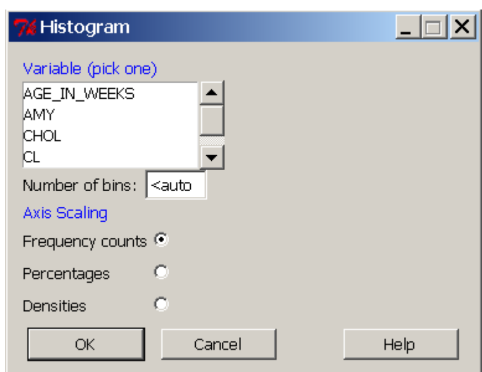
Some references for producing graphs in R

R Graphics (Computer Science and Data Analysis) by [Paul Murrell]
http://www.harding.edu/fmccown/R/
http://www.statmethods.net/graphs/index.html
http://freshmeat.net/articles/creating-charts-and-graphs-with-gnu-r
http://www.ats.ucla.edu/stat/R/library/lecture_graphing_r.htm

4.1.1 Histograms

In statistics, a histogram is a graphical display of tabulated frequencies, shown as bars. It shows what proportion of cases fall into each of several categories.

i. Graph -> Histogram



ii. Select the variable of interest
iii. Select the axis scaling
iv. OK

4.1.2 Norm Q-Q plots

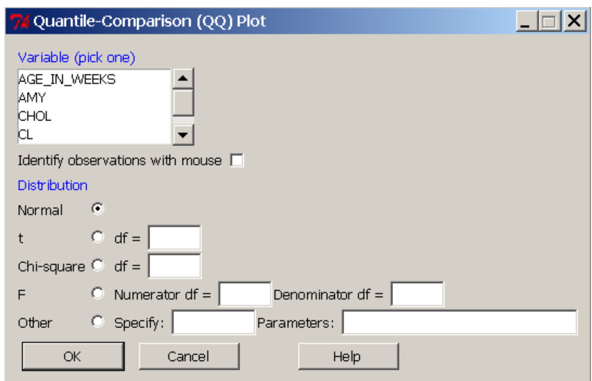
In statistics, a Q-Q plot ("Q" stands for *quantile*) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line *y* = *x*. A norm Q-Q plot compares the sample distribution against a normal distribution.

Additional information:

http://www.cms.murdoch.edu.au/areas/maths/statsnotes/samplestats/qqplot.html

http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Normal_QQ_plot_and_general_QQ_plot

i. Graph -> Quantile-comparison plot



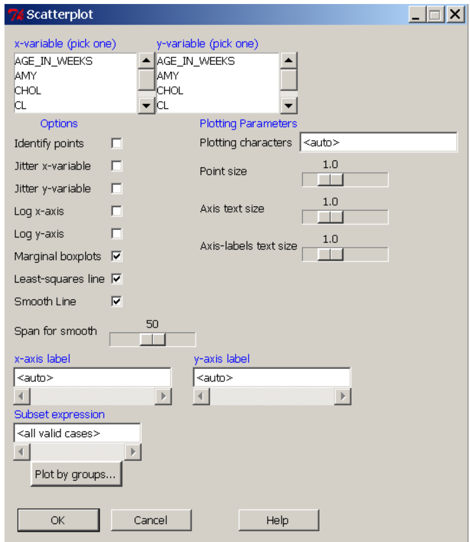
ii. Select variable of interest

iii. Select distribution as normal

iv. OK

4.1.3 Scatterplots

a. Graph -> Scatterplot



b. Select the variables for x-axis and y-axis

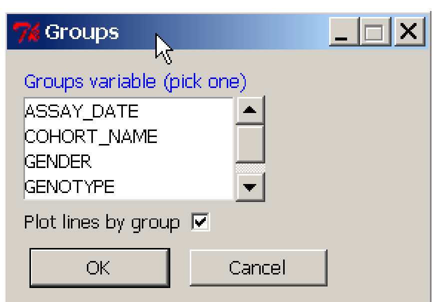
c. Enter the name for the x axis label and the y axis label

d. If you wish the x or y axis can be logged.

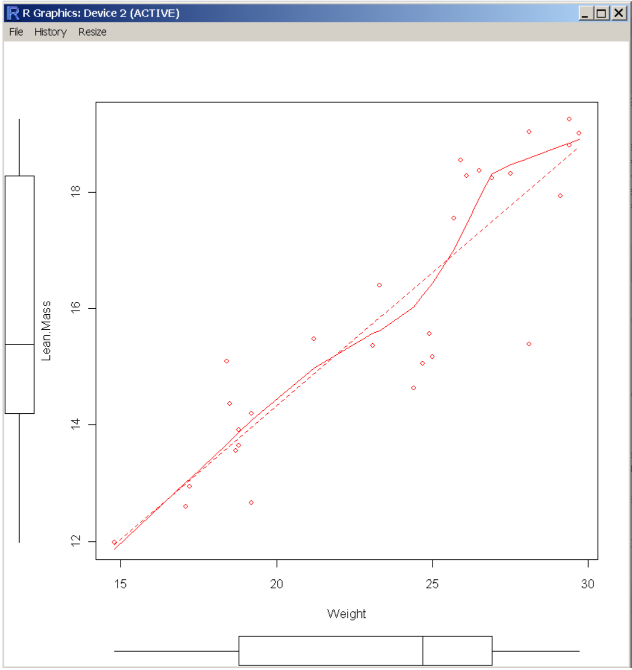
   e. Jitter: this is useful when there are many data points to see if they are overlaying, as a function is used to randomly perturb the points but this does not influence line fitting.

   f. Least-square line can be selected to fit a best fit linear regression line.

   g. Plot by groups will allow a selection of a categorical variable such the scatter plot will use colour to distinguish groups by the categorical variable and fit regression lines independently for each group.



   h. Interpretation of the output?



The dotted line: is the best fit linear regression

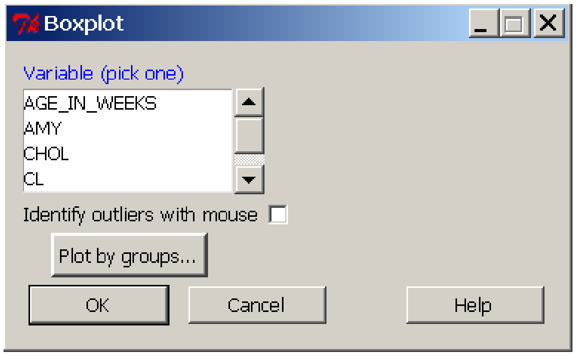
The solid line: is loess line. A loess line is a locally weighted line and is used to assess whether the assumption of linearity is appropriate. Visually you are looking to see whether the loess line suggestions a significant deviation from the linear.

The box plots give an indication to the spread of each variable independently.

4.1.4 Box plots

A boxplot or box and whisker diagram, provides a simple graphical summary of a set of data. It is a convenient way of graphically visualising data through their five-number summaries: the smallest observation (minimum), lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation (maximum). A quartile is any of the three values which divide the sorted dataset into four equal parts, so that each part represents one fourth of the sampled population. Outliers, points which are more than 1.5 the interquartile range (Q3-Q1) away from the interquartile boundaries are marked individually.

a. Select the variable of interest



b. Plot by groups: allows you to have boxplots side by side by splitting the variable by a categorical variable.
c. Identify outliers with mouse: this option allows you to hover over a outlier data point and determine its position in the dataset.
d. OK

4.2 Shapiro-Wilk test for normality

This is a hypothesis tests with the null hypothesis that the data comes from a normal distribution. Hence if the *p*-value is below the significance threshold (typically 0.05), then the null hypothesis is rejected and the alternative hypothesis is accepted. Here the alternative hypothesis is that the data does not come from a normal distribution.

a. Summaries -> Shaprio-Wilk test of normality
b. Select the parameter of interest
c. OK
d. Interpretation: If the *p*-value is below the significance threshold, then there the alternative hypothesis is accepted that the data does not come from a normal distribution.

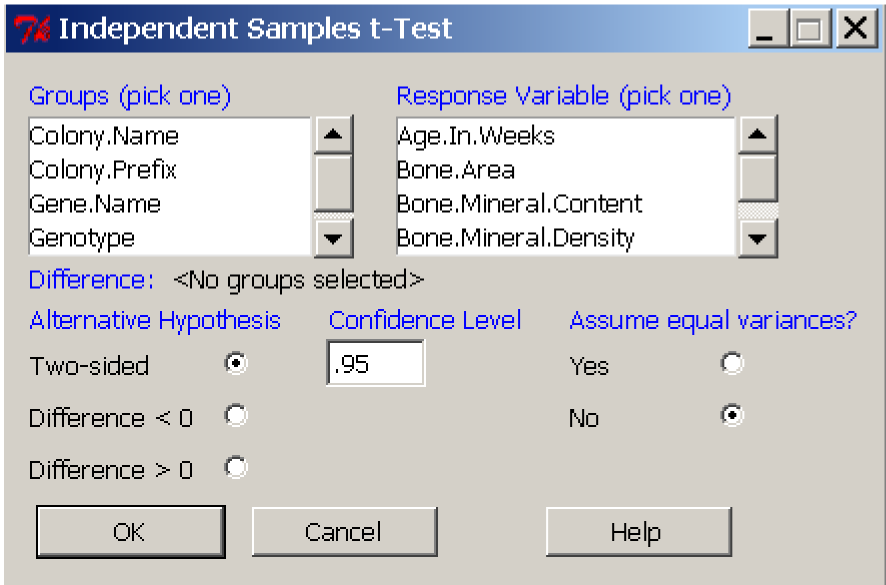
5. Using R commander to apply statistical tests

5.1 Comparing means

5.1.1 Student's *t*-Test

The two-sample Student's *t*-Test is used to determine if two population means are equal.

a. Statistics -> Means -> Independent Samples *t*-Test.



b. Select the grouping variable e.g. genotype

c. Select the response variable (the parameter you are interested in).

d. Typically you select a two-sided hypothesis; this means the change in mean can be either an increase or a decrease.

e. Typically the confidence level of 0.95 is used.

f. If you do not assume equal variance this test is equivalent to the Welch *t*-Test and is considered more robust. Small departures from equal variance significantly affect the robustness of results. The Levene's test (5.3.2) can be used to test whether the variance is equal.

g. OK.

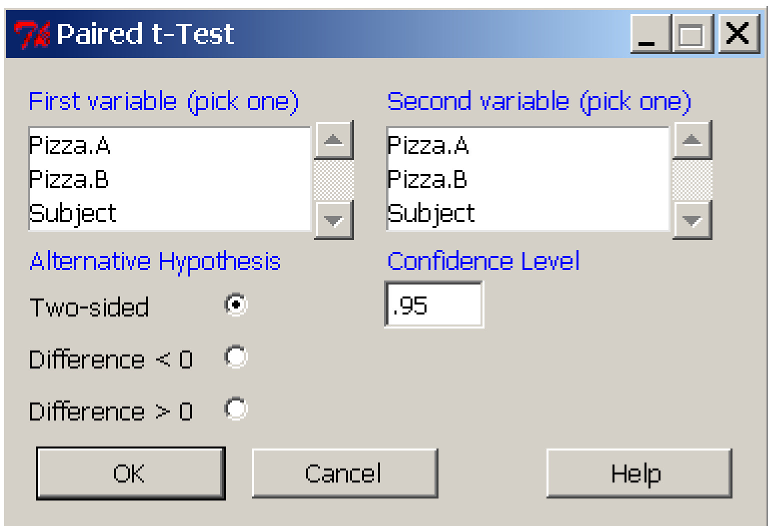
h. Interpretation? If the *p*-value is below the significance threshold, then there is a significant difference in the mean scores for each of the two groups.

5.1.2 Paired student's *t*-Test

The paired test is used to compare means on the same or related subject over time or in differing circumstances. In a paired experiment, there is a one-to-one correspondence between the values in the two samples (e.g. before and after treatment, paired subjects e.g. twins). A paired approach is considered more sensitive as it is looking for a treatment difference excluding initial biological differences.

Note: **Data File Format**
Need two columns; one that contains the first number in each data set pair (e.g., "before" data) and another column that contains the second number in each data set pair. Pairs of numbers must be in the same row.

a. Statistics -> Means -> Paired *t*-Test

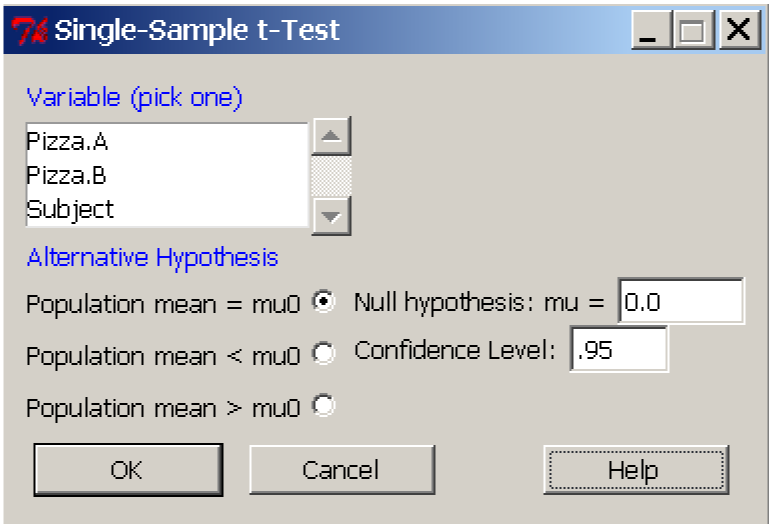


b. Select the first variable
c. Select the second variable
d. Typically you select a two-sided hypothesis; this means the change in mean can be either an increase or a decrease.
e. Typically the confidence level of 0.95 is used.
f. OK.
g. Interpretation?
   - If the *p*-value is below the significance threshold, then the difference in means is not equal to 0
   - The mean of the difference indicates the average difference (variable 1-variable 2)
   - The 95% confidence interval is the confidence interval around the mean difference.

5.1.3 Single sample *t*-Test

The single sample *t*-Test tests a null hypothesis that the population mean is equal to a specified value. If this value is zero (or not entered) then the confidence interval for the sample mean is given.

a. Statistics -> Means -> Single-Sample *t*-Test

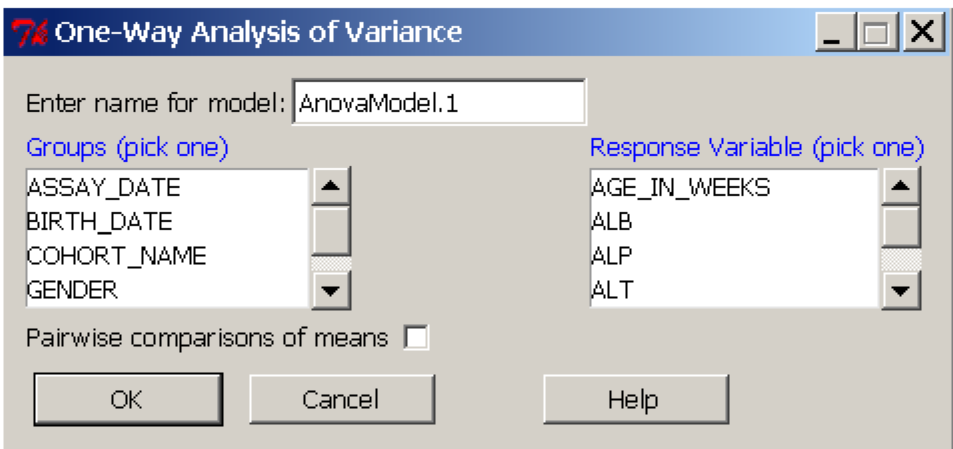


b. Select the variable of interest
c. Enter the proposed mean (Null hypothesis: mu=)
d. Typically the confidence level of 0.95 is used.
e. Three alternative hypothesis are possible:
    a. The mean does not equal the specified value
    b. The mean is less than the specified value
    c. The mean is more than the specified value
f. OK.
g. Interpretation? If the *p*-value is below the significance threshold, then the difference in means is not equal to 0.

5.1.4 One-Way ANOVA

This test is used when you wish to compare the mean scores of more than two groups. Analysis of variance is so called because it compares the variance (variability in scores) between the different groups (believed to be due to the grouping variable) with the variability within each of the groups (believed to be due to chance). The ratio of the variance is converted to a p-value which assesses the chance that this difference in variance arises from sampling affects. A significant *p*-value indicates that we can reject the null hypothesis which states that the populations means are equal. It does not however tell us which of the groups are different. If a significant score is obtained in the one-way ANOVA then post-hoc testing is used to tell where the difference arose. The software uses Tukey post-hoc comparison procedure which is essential like a Student's *t*-Test however the test takes into account the risk of accumulating false positives as multiple tests are being conducted.

a. Statistics -> Means -> One-Way Analysis of Variance



b. Enter a name for the model
c. Select a response variable
d. Select the grouping variable
e. OK
f. Interpretation?

```
> summary(AnovaModel.5)
            Df Sum Sq Mean Sq F value  Pr(>F)
Index        3 2.6545  0.8848  3.8415 0.01723 *
Residuals   37 8.5224  0.2303
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

***p*-value**

```
> numSummary(ANOVA$CHOL , groups=ANOVA$Index, statistics=c("mean", "sd"))
          mean        sd  n
Fem_HET  3.946 0.5468333 10
Fem_WT   4.038 0.4559922 10
Male_HET 4.450 0.4580611 11
Male_WT  4.540 0.4548015 10
```

Group summaries

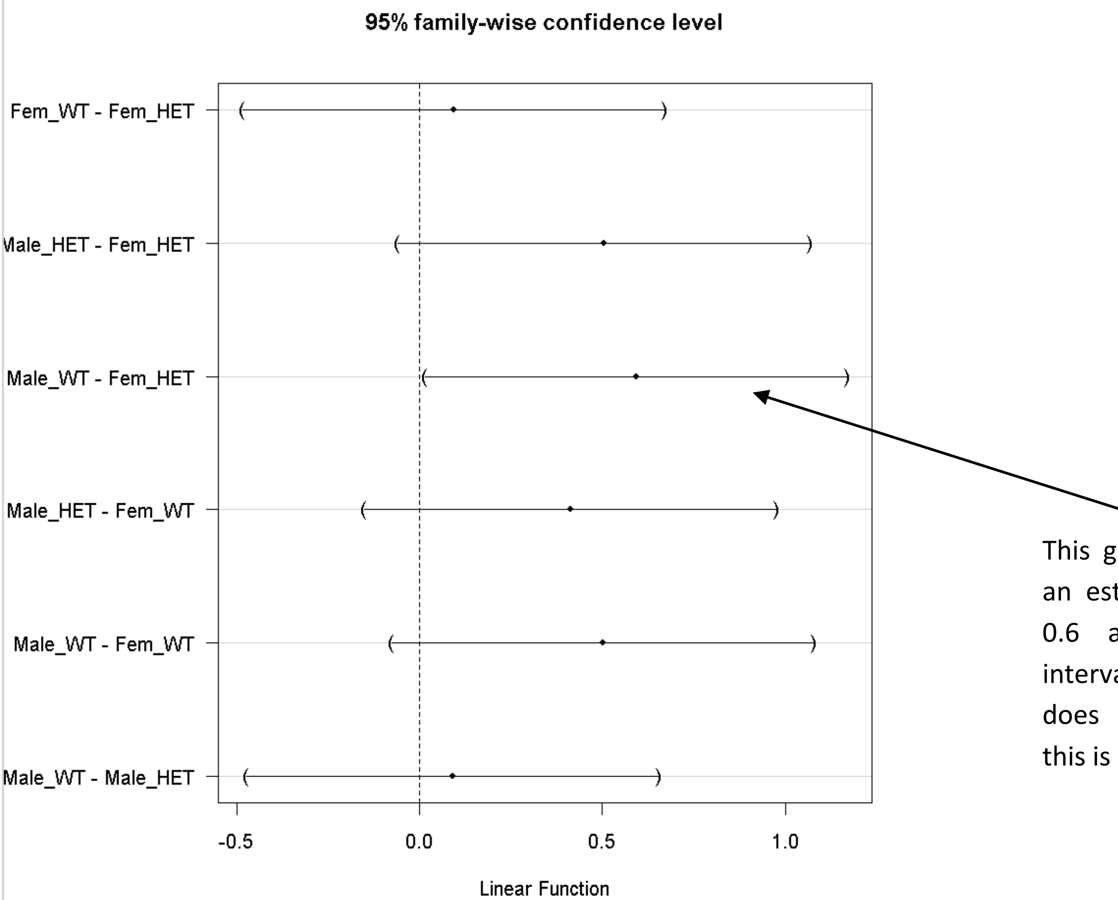
If the *p*-value is below the significance threshold, then the somewhere there is a statistically significant difference in the means of two or more groups.

g. If the *p*-value is significant, repeat the analysis with the pairwise comparisons of means button ticked. This repeats the analysis with the groups being compared to each other group using Tukey contrasts

h. Interpretation?

The output is the mean difference and a 95% confidence interval of this mean difference for each possible comparison. This output is shown mathematically and graphically. You are looking for comparisons where the mean difference confidence interval does not span zero indicating a statistically significant difference in these groups.

```
Linear Hypotheses:
                           Estimate lwr      upr
Fem_WT - Fem_HET == 0       0.09200 -0.48529  0.66929
Male_HET - Fem_HET == 0     0.50400 -0.06002  1.06802
Male_WT - Fem_HET == 0      0.59400  0.01671  1.17129
Male_HET - Fem_WT == 0      0.41200 -0.15202  0.97602
Male_WT - Fem_WT == 0       0.50200 -0.07529  1.07929
Male_WT - Male_HET == 0     0.09000 -0.47402  0.65402
```



This group comparison has an estimated difference of 0.6 and the confidence interval on this estimate does not span zero. Thus this is statistically

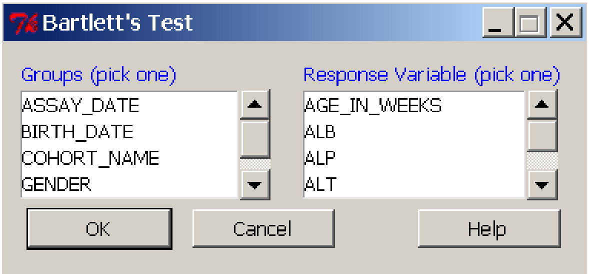
5.2 Comparing the variance

These tests, test if different samples have equal variance (homogeneity of variance). The null hypothesis is that the variance is equal across all groups. When the calculated *p*-value falls below a significance threshold (typically 0.05) then the null hypothesis is rejected and the alternative hypothesis is accepted that the variance is not equal across groups.

5.2.1 Bartlett's test

Bartlett's test is sensitive to departures from normality. That is, if your samples come from non-normal distributions, then Bartlett's test may simply be testing for non-normality. The Levene test (5.3.2) is an alternative to the Bartlett test that is less sensitive to departures from normality.

a. Statistics -> variance -> Bartlett's test

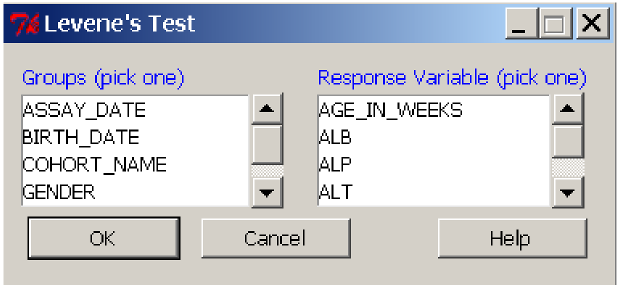


b. Select the grouping variable
c. Select the response variable
d. OK
e. Interpretation: If the *p*-value is below the significance threshold, then the variance in the groups is not equal.

5.2.2 Levene's test

The Levene's test is less sensitive than the Bartlett test (5.3.1) to departures from normality. If you have strong evidence that your data do in fact come from a normal, or nearly normal, distribution, then Bartlett's test has better performance.

a. Statistics -> variance -> Levene's test

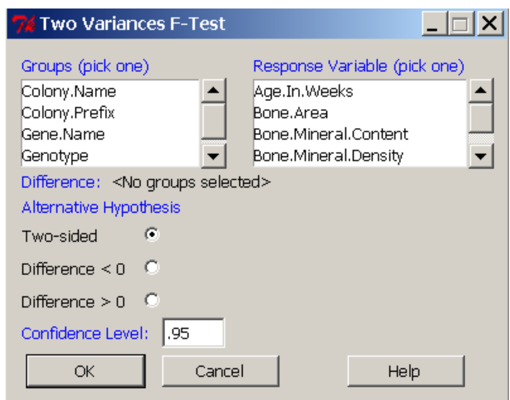


b. Select the grouping variable
c. Select the response variable
d. OK
e. Interpretation: If the *p*-value is below the significance threshold, then the variance in the groups is not equal.

5.2.3 Two variances F-test

An F-Test is used to test if the standard deviations of two populations are equal. This test can be a two-tailed test or a one-tailed test. The two-tailed version tests against the alternative that the standard deviations are not equal. The one-tailed version only tests in one direction that is the standard deviation from the first population is either greater than or less than (but not both) the second population standard deviation. The choice is determined by the problem. For example, if we are testing a new process, we may only be interested in knowing if the new process is less variable than the old process.

a. Statistics -> variance -> Two variances F-test



b. Select the grouping variable
c. Select the response variable
d. Select whether one or two tailed
e. OK
f. Interpretation: When the *p*-value falls below the significance threshold the null hypothesis is rejected and the alternative hypothesis is accepted.

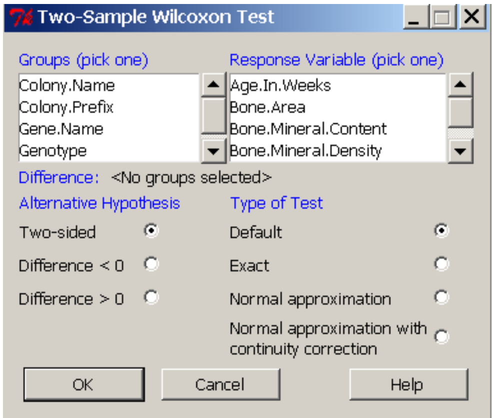
5.3 Non parametric tests

These are statistical tests which are distribution free methods as they do not rely on assumptions that the data are drawn from a given probability distribution.

5.3.1 Two-sample Wilcoxon Test

Non-parametric equivalent to the Student's *t*-Test. Can also be called two-sample Mann-Whitney U test. This test assesses whether the values in two samples differ in size.

a. Statistics -> Non-parametric tests -> Two sample Wilcoxon test



b. Select the grouping variable
c. Select the response variable (variable of interest)
d. If n is low (<50) then exact should be select as the type of test.
e. If the treatment difference can occur in either direction (i.e. increase or a decrease) then select a two-sided test.
f. OK
g. Interpretation: When the *p*-value falls below the significance threshold the null hypothesis is rejected and the alternative hypothesis is accepted.

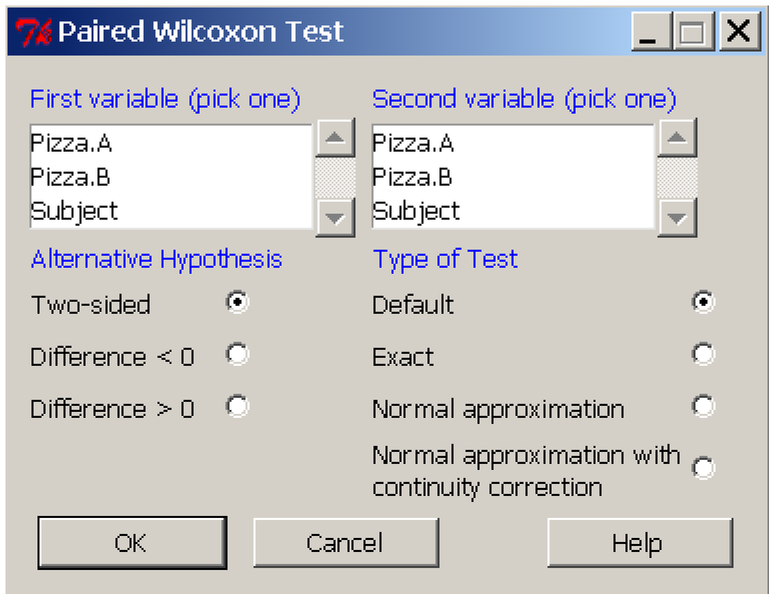
5.3.2 Paired-sample Wilcoxon Test

The Wilcoxon test for paired samples is the non-parametric equivalent of the paired samples *t*-test.

Note: **Data Format**

Need two columns; one that contains the first number in each data set pair (e.g., “before” data) and another column that contains the second number in each data set pair. Pairs of numbers must be in the same row.

a. Statistics -> Non-parametric tests -> Paired- sample Wilcoxon test

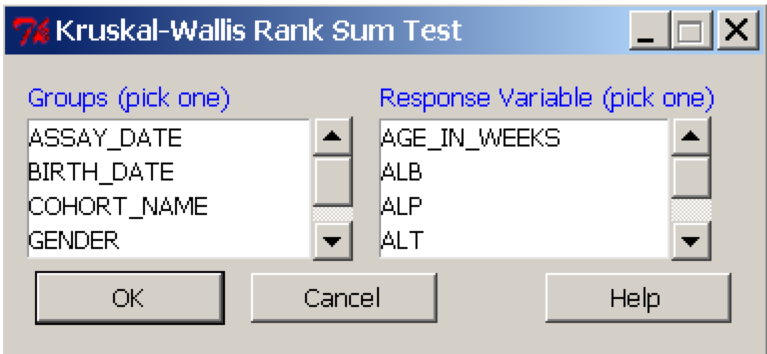


b. Select the first variable
c. Select the second variable
d. If the change can be either an increase or a decrease then select a two-sided test.
e. OK
f. Interpretation: When the *p*-value falls below the significance threshold the null hypothesis is rejected and the alternative hypothesis is accepted.

5.3.3 Kruskal-Wallis Test

This test is a non-parametric method for testing equality of population medians among groups. It is identical to an ANOVA (5.1.4) with the data replaced by their ranks. It is an extension of the Two-sample Wilcoxon test to 3 or more groups.

a. Statistics -> Non-parametric tests -> Kruskal-Wallis test



b. Select the grouping variable
c. Select the response variable (variable of interest)
d. OK

6. Amending the graphical output

One of the main reasons data analysts turn to **R** is for its strong graphic capabilities. However, with R commander, the options on graphs are limited and they don't look too pretty and aren't ideal for reports or presentations. Here I go through some examples of what you can do and then it should give you grounding for proceeding further if you require. The overall strategy is to call the code for the basic graph and then amend the code manually by altering the graphics parameters or by calling a second function to do a particular job (e.g. adding a label).

For future advice and support on R and graphs I recommend:

1. R Graphics by Paul Murrell
2. Data Analysis and Graphics Using R: An Example-based Approach by John Maindonald and John Braun.

Amending code - things to notes

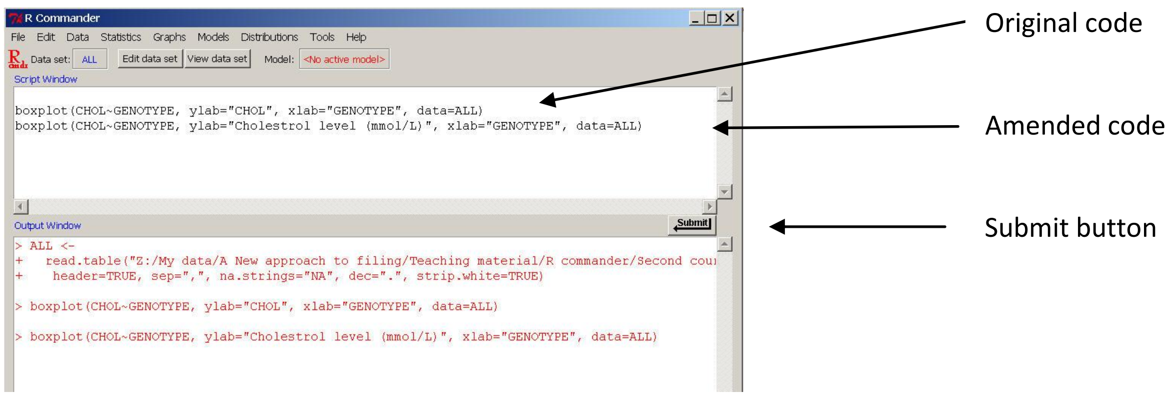
1. If you add another parameter (instruction) to a function it needs to form part of the list so it is placed within the bracket of information passed to that function and a comma is placed between each instruction.
2. If you are using words to describe the colour you want or to add a label then it needs to be surrounded by quote marks (i.e. “”) marks so the software knows that it is looking at string (i.e. text) information.
3. Script is particularly to form so capitals etc. matter.

6.1 Amending the axis labels

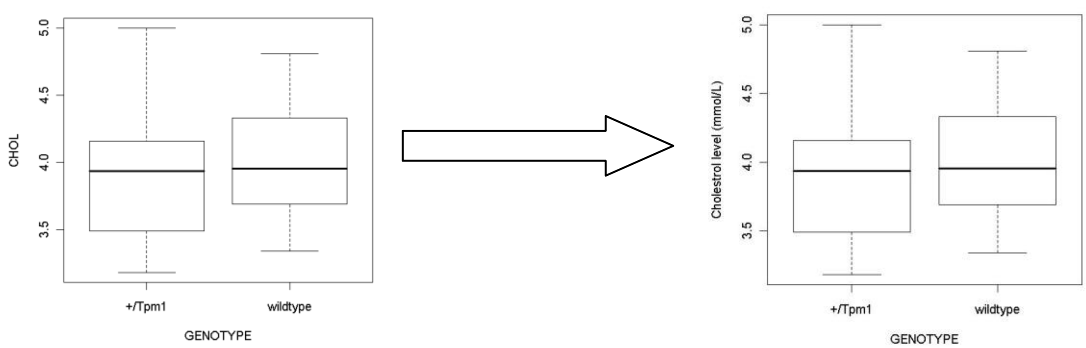
a. Use the drop down menus to request a graph e.g. the box plot (9.1.4).

b. Now you can amend the code. To change the label on the x-axis you either change the text within the quotes for xlab= “XX” and similarly for the ylab or add the text you wish to include.

Example: Changing the label from CHOL to Cholesterol level (mmol/L)



c. Highlight the code and press the submit button to activate the script.



6.2 Adding a main title

a. Use the drop down menus to request a graph.

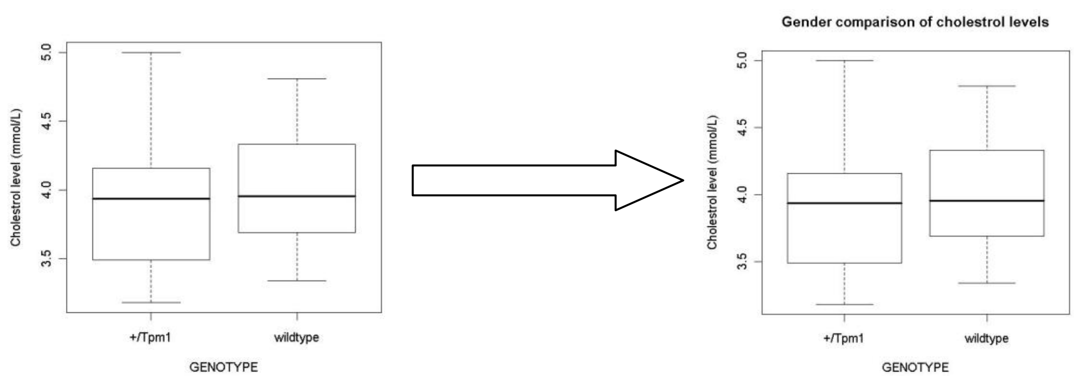
b. The parameter that controls the header is main. You can either change the text if it exists or add the parameter to the instructions for the graph function.

c. Example:

Original code: boxplot(CHOL~GENOTYPE, ylab="Cholestrol level (mmol/L)", xlab="GENOTYPE", data=ALL)

Amended code:
boxplot(CHOL~GENOTYPE, ylab="Cholestrol level (mmol/L)", main="Gender comparison of cholestrol levels", xlab="GENOTYPE", data=ALL)Add to the code

d. Highlight the code and press the send to button to activate the script.



6.3 Adding a line

a. Use the drop down menus to request a scatter graph.

b. Here a second function (abline) is needed to add the line. The parameters within the brackets are used to pass the information to the function. These are used to control the line placement within the graph. If you do not specify the parameter then the parameter will be set to the default settings.

Abline structure: abline(a = NULL, b = NULL, h = NULL, v = NULL, , ...)

parameter		Default
a	intercept	NULL
b	slope	NULL
h	the y-value(s) for horizontal line(s).	NULL
v	the x-value(s) for vertical line(s).	NULL
...	graphical parameters such as `col`, `lty` and `lwd` and the line characteristics `lend`, `ljoin` and `lmitre`.	

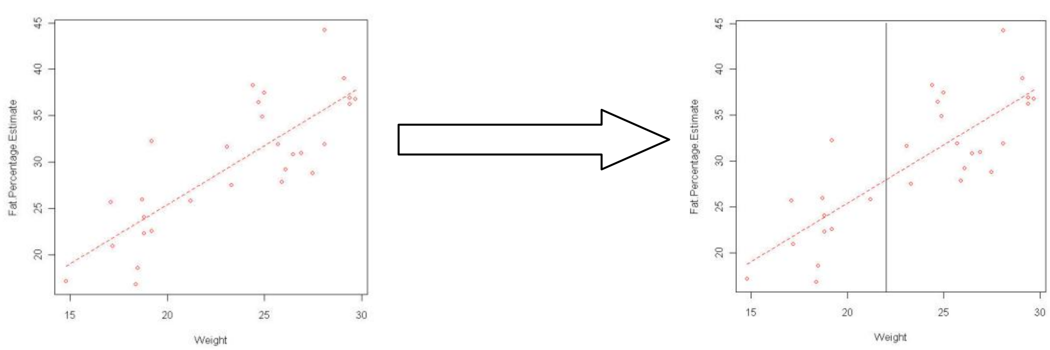
c. Adding a vertical line at point x

   i. Type code abline(v=x) into the script window

   ii. Highlight the code and submit.

Example:

```
scatterplot(Fat.Percentage.Estimate~Weight, reg.line=lm, smooth=FALSE,
  labels=FALSE, boxplots=FALSE, span=0.5, data=DEXA)
abline(v=22.5)
```



d. Adding a horizontal line at point x

    i. Type code abline(h=x) into the script window

    ii. Highlight the code and submit.#

e. Adding a line of a known equation

    i. Type code abline(a=x, b=y) into the script window

    ii. Highlight the code and submit.

f. Adding an equivalence line

    i. Type code abline(b=1) into the script window

    ii. Highlight the code and submit.

6.4 Amending the line parameters

A number of parameters can be added to the abline function to amend the output

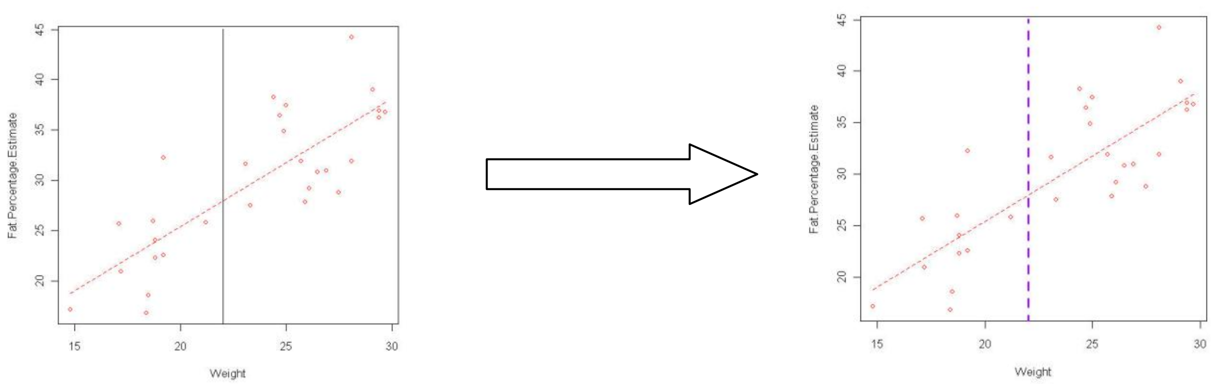
col	The easiest way to specify a colour is to use the name eg "red". R understands 657 different colour names. Type colours() to see a full list of known names.
lty	The line type. Line types can either be specified as an integer (0=blank, 1=solid (default), 2=dashed, 3=dotted, 4=dotdash, 5=longdash, 6=twodash) or as one of the character strings `"blank"`, `"solid"`, `"dashed"`, `"dotted"`, `"dotdash"`, `"longdash"`, or `"twodash"`, where `"blank"` uses 'invisible lines' (i.e., does not draw them).
lwd	The line width, a *positive* number, defaulting to `1`.

Example:

scatterplot(Fat.Percentage.Estimate~Weight, reg.line=lm, smooth=FALSE,

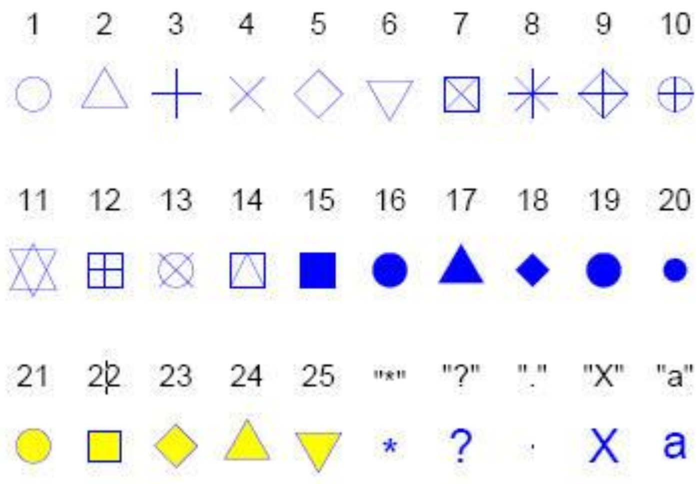
labels=FALSE, boxplots=FALSE, span=0.5, data=DEXA)

abline(v=22.5, col="purple", lty="dashed", lwd=3)



6.5 Amending the plot symbol

R provides a fixed set of 26 data symbols for plotting and the symbol is controlled by the **pch** setting. Pch 21 to 25 allow a fill colour separate from the border colour, with the **bg** setting controlling the fill colour in these cases.



Example:

scatterplot(Fat.Percentage.Estimate~Weight, reg.line=lm, span=0.5, data=DEXA)

scatterplot(Fat.Percentage.Estimate~Weight, reg.line=lm, pch= 2, col= “red”, span=0.5, data=DEXA)

6.6 Adding a text label

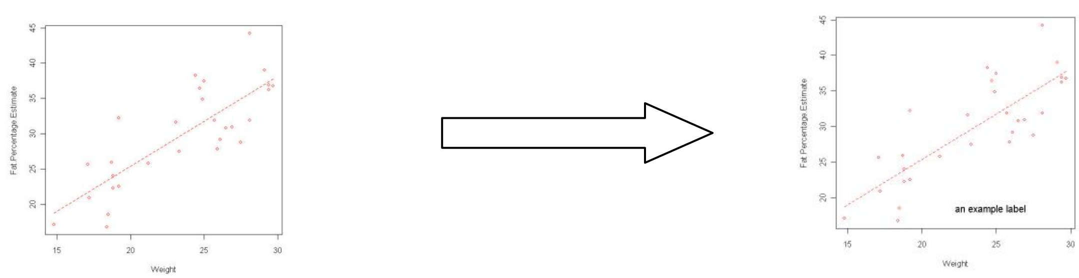
Here a second function (text) is used to add the text. The parameters within the brackets are used to pass the information to the function to drive what text and where the text is placed. If you do not specify the parameter then the parameter will be set to the default settings.

Text function: text (x, y, label, col)

parameter		Default
X, y	Coordinates where the text “labels” should be written	
label	This specifies the text to be written	
col	Colour of the text.	Black

Example 1

```
scatterplot(Fat.Percentage.Estimate~Weight, reg.line=lm, smooth=FALSE,
 labels=FALSE, boxplots=FALSE, span=0.5, data=DEXA)
text(x=25, y=20, label ="an example label")
```



6.7 Amending the plot colours

6.7.1 For a box plot

a. Use the drop down menus to request a boxplot graph.

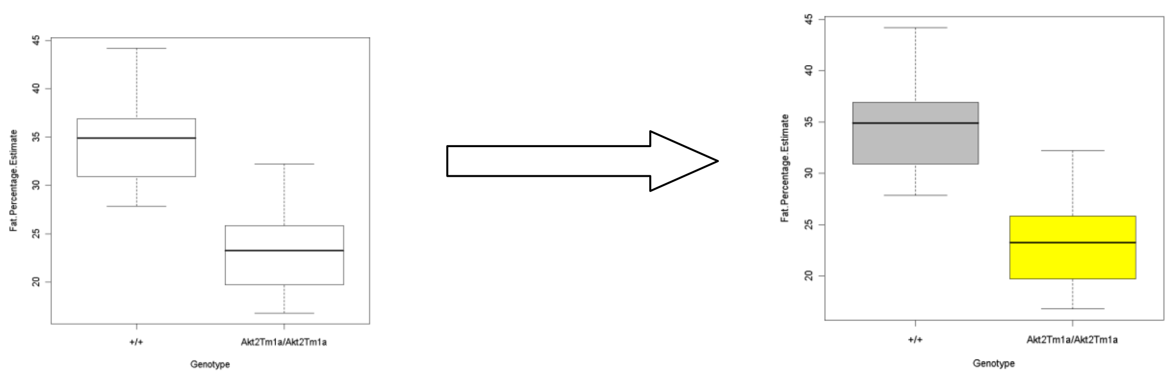
b. Amend the script by adding a col parameter.

   i. To add a single colour to all boxplots add col=("COLOUR OF YOUR CHOICE") to the code.

   ii. To alter each boxplot individually you need to add a list of colours with length matching the number of boxplots to the code.
   Eg. col=c("red", "black", "green")

   iii. Highlight the amended code and submit.

   iv. Example: boxplot(Fat.Percentage.Estimate~Genotype, lab="Fat.Percentage.Estimate", xlab="Genotype", col=c("grey", "yellow"), data=DEXA)



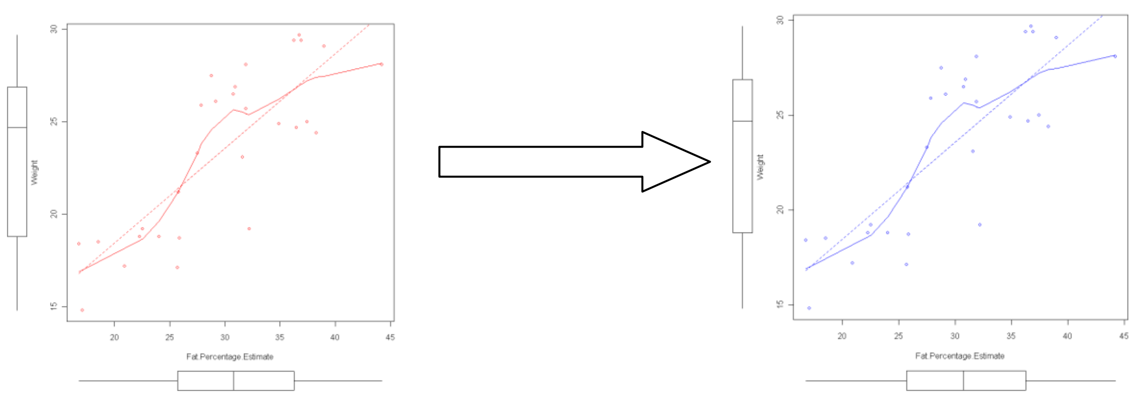
6.7.2 For a scatter plot

a. Using the drop down menus to request a scatter graph.

b. You can change the colour of the scatter graphs by using the col parameter.

   a. For a graph with one group you enter col=c(“black”, “COLOUR OF YOUR CHOICE”) into the list.

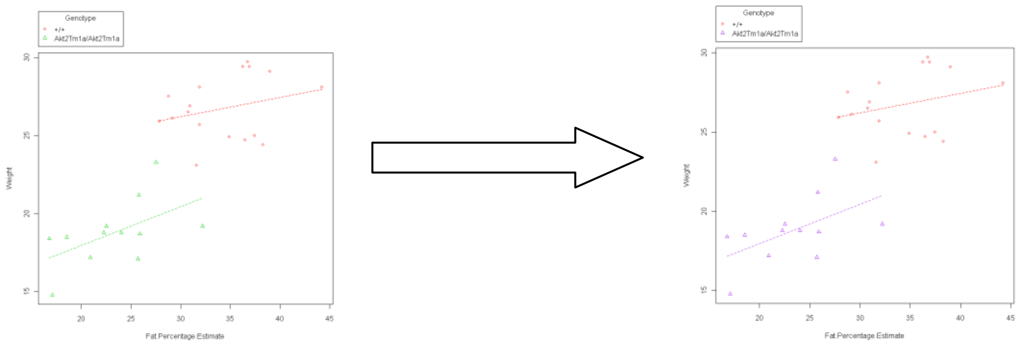
      Example: scatterplot(Weight~Fat.Percentage.Estimate, reg.line=lm, smooth=TRUE, labels=FALSE, boxplots='xy', span=0.5, col=c("black", "blue"), data=DEXA)



   b. For a graph with multiple groups:

      You add the colours as a list (E.g. col=c(“black”, “green”, “pink”, “yellow”). The first colour in the list is for something I cannot work out and I set to black just in case. The subsequent colours are for your groups. The order will match the group levels if they are outputted in alphabetically order.

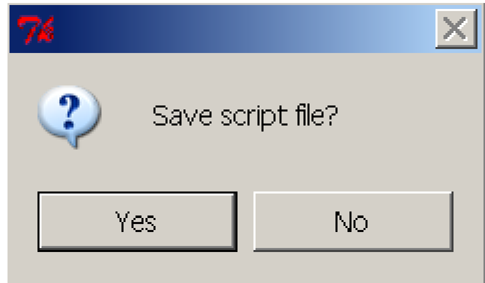
      Example: scatterplot(Weight~Fat.Percentage.Estimate | Genotype, reg.line=lm, smooth=FALSE, labels=FALSE, boxplots=FALSE, span=0.5, by.groups=TRUE, data=DEXA, col=c("black", "red", "purple"))


Genotype
+/+
Akt2Tm1a/Akt2Tm1a
Weight
Fat Percentage Estimate
Genotype
+/+
Akt2Tm1a/Akt2Tm1a
Weight
Fat Percentage Estimate

7. Odds and Ends

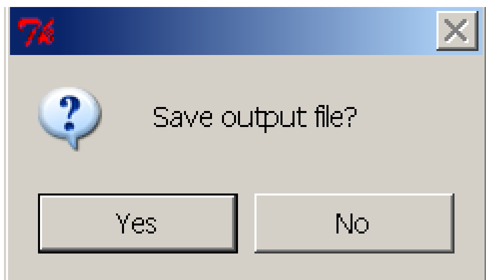
7.1 Exiting and saving script

i. File -> Exit -> From R Commander and R -> OK



ii. There are two advantages to saving the script

a. Provides a record of the analysis completed.
b. During the next session the user can 'get back to where you left off' by opening a saved script and submitting the syntax.



iii. When exiting from R, will ask whether save the workspace image, it is best to answer **NO** because the R program can get confused by objects (datasets/parameters) carried over. Better route is to save script (coding).

7.2 Saving and printing Output

It is recommended that you collect the text output and graphs you want to keep in a word-processor document. In this manner, you can intersperse R output with typed notes and explanations.

7.2.1 Copying text

Highlight the text with the mouse -> ctrl-c and paste ctrl-v as you would for any window application.

7.2.2 Copying graphs

Right-click on the graph, select 'Copy as meta-file' and past directly into Word or PowerPoint.

Alternatively can also save the graph as an independent file:

Graphs -> Save graph to file -> as bitmap/EPS/PDF .....

7.3 Entering commands directly into the script window

Commands generated by the R Commander appear in the script window, and you can type and edit commands in this window. To send this script you have to highlight the relevant text and press the "Submit" button.

Notes:

1. All lines of a multi-line command must be submitted simultaneously for execution.
2. Commands that extend over more than one line should have the second and subsequent lines indented by one or more spaces or tabs.

7.4 Current menu "tree" of the R Commander (version 1.4-10)

- File
  - Change working directory…
  - Open script file…
  - Save script…
  - Save script as…
  - Save output…
  - Save output as…
  - Save R workspace…
  - Save R workspace as…
  - Exit
    - From Commander
    - From Commander and R
- Edit
  - Cut
  - Copy
  - Paste
  - Delete
  - Find…
  - Select all
  - Undo
  - Redo
  - Clear Window
- Data
- New data set…
- Load data set…
- Import data
  - from text file, clipboard, or URL…
  - from SPSS data set…
  - from Minitab data set…

    - from STATA data set…
    - from Excel, Access, or dbase data set…
- Data in packages
    - List data sets in packages
    - Read data set from an attached package…
- Active data set
    - Select active data set…
    - Refresh active data set
    - Help on active data set (if applicable)
    - Variables in active data set
    - Set case names…
    - Subset active data set
    - Remove row(s) from active data set…
    - Stack variables in active data set…
    - Remove cases w/ missing data…
    - Save active data set…
    - Export active data set…
- Manage variables in active data set
    - Recode variables…
    - Compute new variable…
    - Add observation numbers to data set
    - Standardize variables…
    - Convert numeric variables to factors…
    - Bin numeric variable…
    - Reorder factor levels…
    - Define contrasts for a factor…
    - Rename variables…
    - Delete variables from data set…
- Statistics
    - Summaries
        - Active data set
        - Numerical summaries…
        - Frequency distributions…
        - Count missing observations
        - Table of statistics
        - Correlation matrix…
        - Correlation test…
        - Shapiro-Wilk test of normality…
    - Contingency tables
        - Two-way table…
        - Multi-way table…
        - Enter and analyze two-way table…

- Means
  - Single-sample t-test…
  - Independent samples t-test…
  - Paired t-test…
  - One-way ANOVA…
  - Multi-way ANOVA…
- Proportions
  - Single-sample proportion test…
  - Two-sample proportions test…
- Variances
  - Two-variances F-test…
  - Bartlett's test…
  - Levene's test…
- Nonparametric tests
  - Two-sample Wilcoxon test…
  - Paired-samples Wilcoxon test…
  - Kruskal-Wallis test…
  - Friedman rank-sum test…
- Dimensional analysis
  - Scale reliability…
  - Principal-components analysis…
  - Factor analysis…
  - Cluster analysis
    - k-means cluster analysis…
    - Hierarchical cluster analysis…
    - Summarize hierarchical clustering…
    - Add hierarchical clustering to data set…
- Fit models
- Linear regression…
- Linear model…
- Generalized linear model…
- Multinomial logit model…
- Ordinal regression model…

Graphs
- Color palette…
- Index plot…
- Histogram…
- Stem-and-leaf display…
- Boxplot…
- Quantile-comparison plot…

- Scatterplot…
- Scatterplot matrix…
- Line graph…
- XY conditioning plot…
- Plot of means…
- Bar graph…
- Pie chart…
- 3D graph
  - 3D scatterplot…
  - Identify observations with mouse
  - Save graph to file
- Save graph to file
  - as bitmap…
  - as PDF/Postscript/EPS…
  - 3D RGL graph…

Models

- Select active model
- Summarize model
- Add observation statistics to data
- Confidence intervals
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Hypothesis tests
  - ANOVA table…
  - Compare two models…
  - Linear hypothesis…
- Numerical diagnostics
  - Variance-inflation factors
  - Breusch-Pagan test for heteroscedasticity
  - Durbin-Watson test for autocorrelation
  - RESET test for nonlinearity
  - Bonferroni outlier test
- Graphs
  - Basic diagnostic plots
  - Residual quantile-comparison plot
  - Component+residual plots
  - Added-variable plots
  - Influence plot
  - Effect plots

Distributions

Continuous distributions
- Normal distribution
  - Normal quantiles…
  - Normal probabilities…
  - Plot Normal distribution…
  - Sample from Normal distribution…
- t distribution
  - t quantiles…
  - t probabilities…
  - Plot t distribution…
  - Sample from t distribution…
- Chi-squared distribution
  - Chi-squared quantiles…
  - Chi-squared probabilities…
  - Plot Chi-squared distribution…
  - Sample from Chi-squared distribution…
- F distribution
  - F quantiles…
  - F probabilities…
  - Plot F distribution…
  - Sample from F distribution…
- Exponential distribution
  - Exponential quantiles…
  - Exponential probabilities…
  - Plot Exponential distribution…
  - Sample from Exponential distribution…
- Uniform distribution
  - Uniform quantiles…
  - Uniform probabilities…
  - Plot Uniform distribution…
  - Sample from Uniform distribution…
- Beta distribution
  - Beta quantiles…
  - Beta probabilities…
  - Plot Beta distribution…
  - Sample from Beta distribution…
- Cauchy distribution
  - Cauchy quantiles…
  - Cauchy probabilities…
  - Plot Cauchy distribution…
  - Sample from Cauchy distribution…
- Logistic distribution
  - Logistic quantiles…

- Logistic probabilities…
- Plot Logistic distribution…
- Sample from Logistic distribution…

Lognormal distribution

- Lognormal quantiles…
- Lognormal probabilities…
- Plot Lognormal distribution…
- Sample from Lognormal distribution…

Gamma distribution

- Gamma quantiles…
- Gamma probabilities…
- Plot Gamma distribution…
- Sample from Gamma distribution…

Weibull distribution

- Weibull quantiles…
- Weibull probabilities…
- Plot Weibull distribution…
- Sample from Weibull distribution…

Gumbel distribution

- Gumbel quantiles…
- Gumbel probabilities…
- Plot Gumbel distribution…
- Sample from Gumbel distribution…

Discrete distributions

- Binomial distribution
- Binomial quantiles…
- Binomial tail probabilities…
- Binomial probabilities…
- Plot Binomial distribution…
- Sample from Binomial distribution…

Poisson distribution

- Poisson quantiles…
- Poisson tail probabilities…
- Poisson probabilities…
- Plot Poisson distribution…
- Sample from Poisson distribution…

Geometric distribution

- Geometric quantiles…
- Geometric tail probabilities…
- Geometric probabilities…
- Plot Geometric distribution…
- Sample from Geometric distribution…

Hypergeometric distribution

        Hypergeometric quantiles…
        Hypergeometric tail probabilities…
        Hypergeometric probabilities…
        Plot Hypergeometric distribution…
        Sample from Hypergeometric distribution…
    Negative binomial distribution
        Negative binomial quantiles…
        Negative binomial tail probabilities…
        Negative binomial probabilities…
        Plot Negative binomial distribution…
        Sample from Negative binomial distribution…

Tools
- Load package(s)…
- Load Rcmdr plug-in(s)…
- Options…

Help
- Commander help
- Introduction to the R Commander
- Help on active data set (if applicable)
- About Rcmdr