Progress on the British Academy funded Statistical eBook grant

Professor William Browne,
Chris Charlton and Liz Washbrook
Centre for Multilevel Modelling,University of Bristol

What will we cover ?

- Background to CMM and StatJR
- Interoperability and e-Books
- British Academy grant
- Topics to be covered
- Work packages 1 – 5 progress

Background to CMM

- Cross-faculty statistical research group primarily based in Education where we are a Research Centre.
- Produce statistical software packages, MLwiN and StatJR with over 15,000 users.
- Also LEMMA online training materials with nearly 20,000 users.
- Historically research funded by the ESRC via several programme nodes to a total of more than £5M in the past 10 years
- See http://www.bristol.ac.uk/cmm/

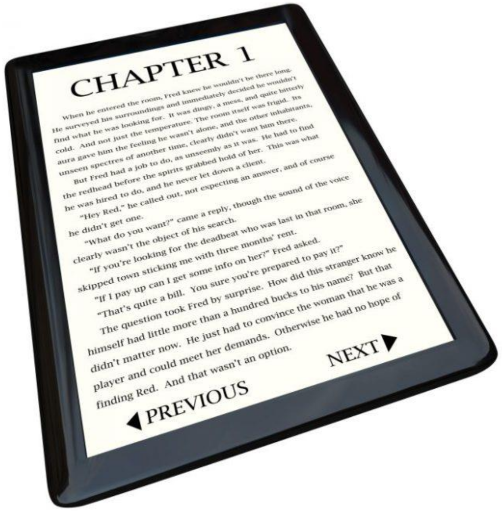
Stat-JR

- A statistical package developed by the team at the Centre for Multilevel Modelling with colleagues at Southampton.
- Contains it's own (MCMC-based) estimation engine.
- System based on the idea of a suite of templates where each template performs a specific operation.
- Also allows interoperability with other software packages, so for example might have a regression template that fits regressions using various software packages.
- The initial TREE interface runs in a web browser.
- There are also newer eBook and workflow interfaces.
- Several ESRC grants have enabled Stat-JR to be written.

eBooks

+

=



CHAPTER 1

When he entered the room, Fred knew he wouldn't be there long. He surveyed his surroundings and immediately decided he wouldn't find what he was looking for. It was dingy, a mess, and quite bitterly cold. And not just the temperature. The room itself was frigid. Its aura gave him the feeling he wasn't alone, and the other inhabitants, unseen spectres of another time, clearly didn't want him there.

But Fred had a job to do, as unseemly as it was. He had to find the redhead before the spirits grabbed hold of her. This was what he was hired to do, and he never let down a client.

"Hey Red," he called out, not expecting an answer, and of course he didn't get one.

"What do you want?" came a reply, though the sound of the voice clearly wasn't the object of his search.

"If you're looking for the deadbeat who was last in that room, she skipped town sticking me with three months' rent.

"If I pay up can I get some info on her?" Fred asked.

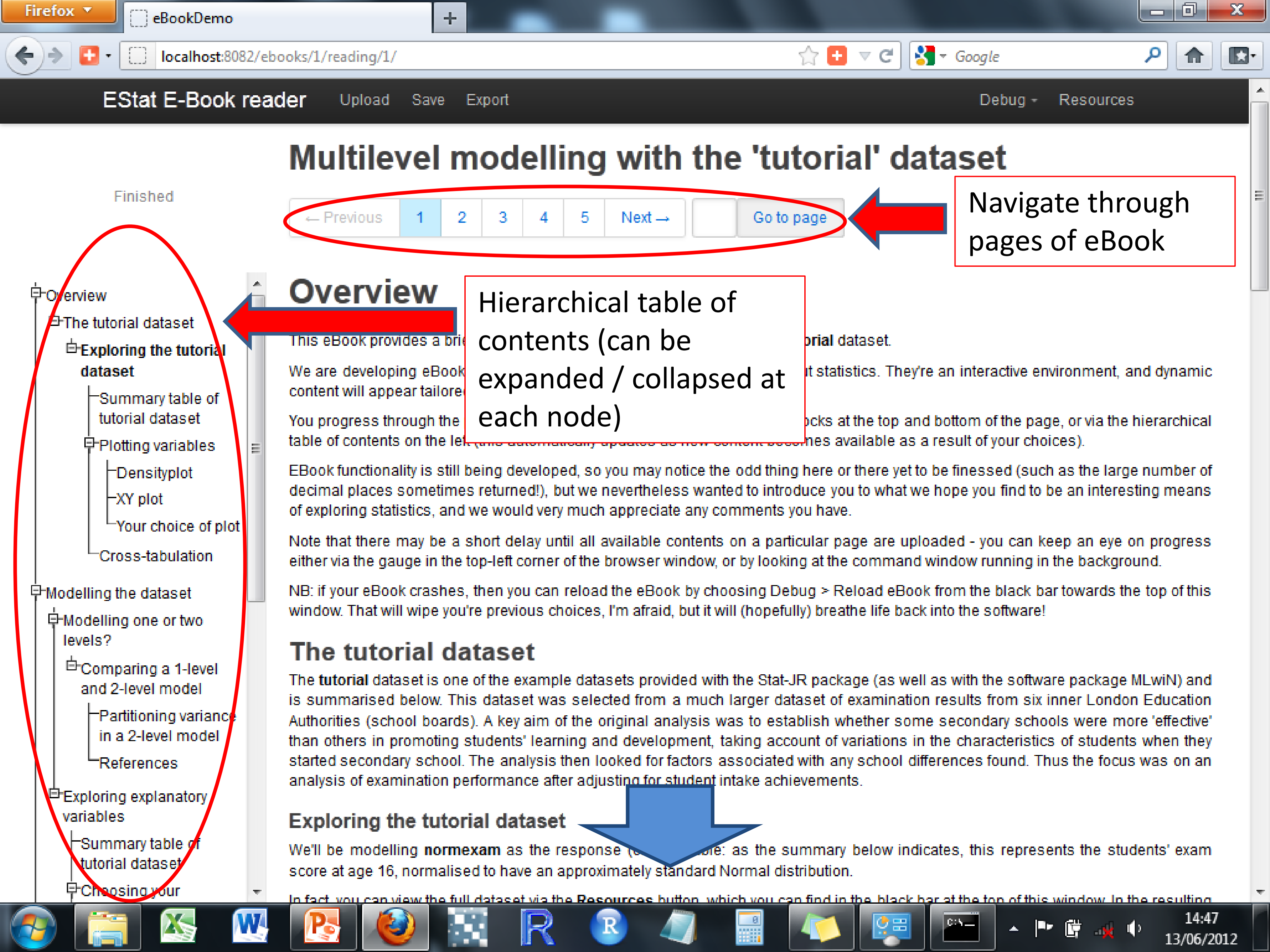
"That's quite a bill. You sure you're prepared to pay it?"

The question took Fred by surprise. How did this stranger know he himself had little more than a hundred bucks to his name? But that didn't matter now. He just had to convince the woman that he was a player and could meet her demands. Otherwise he had no hope of finding Red. And that wasn't an option.

An electronic book is a book-publication in digital form.
In the US more books are published online than distributed in hard copy in book shops.

Statistical (and Mathematical) eBooks

- The idea is can we incorporate statistical content into an eBook? Of course a statistical textbook is no different on paper to any other document when it comes to creating a pdf file (aside from maybe more equations!)
- The difference is in what ‘enhancements’ we can add and so the idea here is combining the text book with the statistics package i.e. interactive examples, allowing the user to include their own dataset etc.



Multilevel modelling with the 'tutorial' dataset

Navigate through pages of eBook

Overview

Hierarchical table of contents (can be expanded / collapsed at each node)

This eBook provides a bri orial dataset.

We are developing eBook t statistics. They're an interactive environment, and dynamic content will appear tailore

You progress through the ocks at the top and bottom of the page, or via the hierarchical table of contents on the le mes available as a result of your choices).

EBook functionality is still being developed, so you may notice the odd thing here or there yet to be finessed (such as the large number of decimal places sometimes returned!), but we nevertheless wanted to introduce you to what we hope you find to be an interesting means of exploring statistics, and we would very much appreciate any comments you have.

Note that there may be a short delay until all available contents on a particular page are uploaded - you can keep an eye on progress either via the gauge in the top-left corner of the browser window, or by looking at the command window running in the background.

NB: if your eBook crashes, then you can reload the eBook by choosing Debug > Reload eBook from the black bar towards the top of this window. That will wipe you're previous choices, I'm afraid, but it will (hopefully) breathe life back into the software!

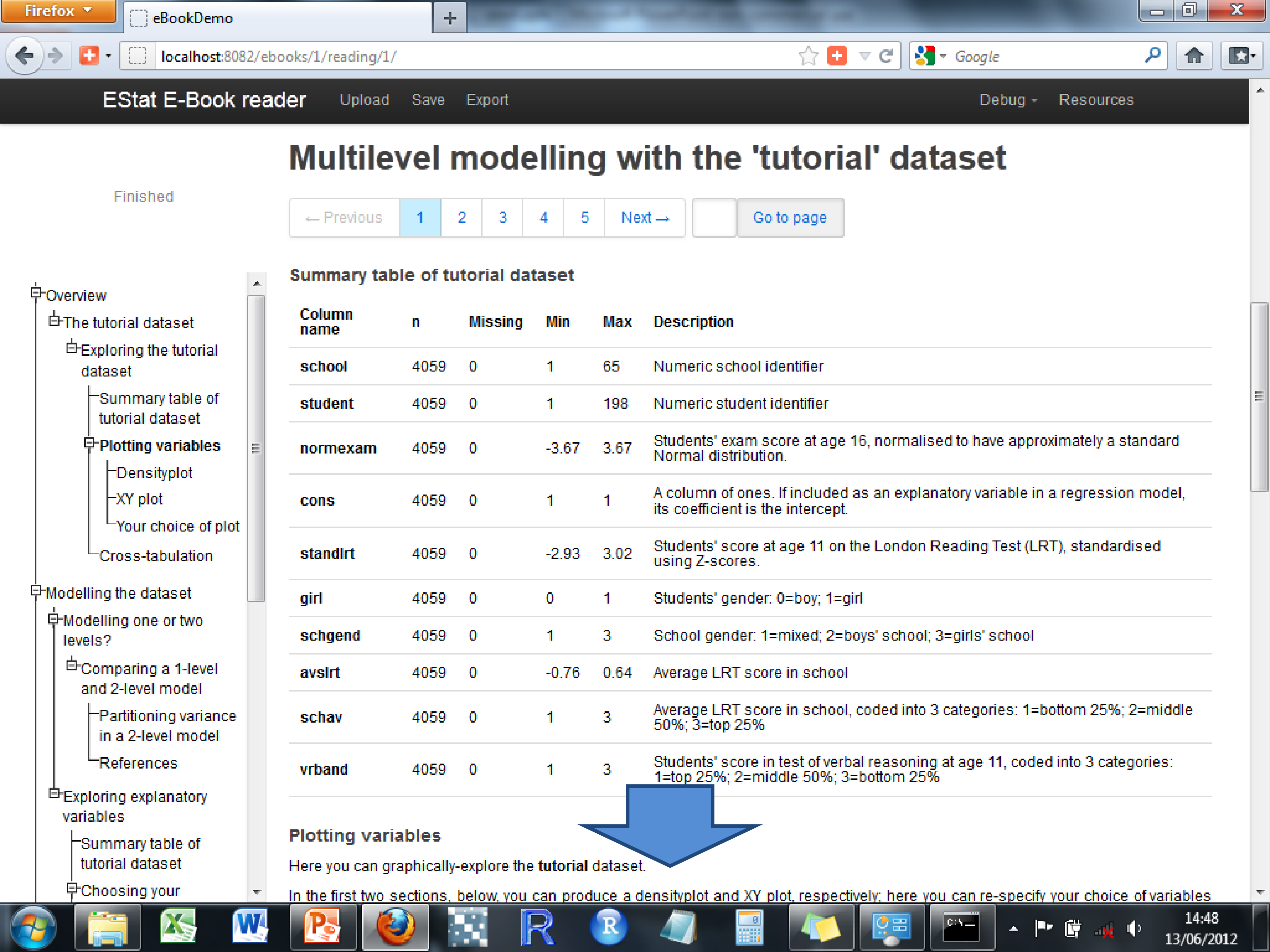
The tutorial dataset

The **tutorial** dataset is one of the example datasets provided with the Stat-JR package (as well as with the software package MLwiN) and is summarised below. This dataset was selected from a much larger dataset of examination results from six inner London Education Authorities (school boards). A key aim of the original analysis was to establish whether some secondary schools were more 'effective' than others in promoting students' learning and development, taking account of variations in the characteristics of students when they started secondary school. The analysis then looked for factors associated with any school differences found. Thus the focus was on an analysis of examination performance after adjusting for student intake achievements.

Exploring the tutorial dataset

We'll be modelling **normexam** as the response (ble: as the summary below indicates, this represents the students' exam score at age 16, normalised to have an approximately standard Normal distribution.

In fact, you can view the full dataset via the **Resources** button, which you can find in the black bar at the top of this window. In the resulting



Multilevel modelling with the 'tutorial' dataset

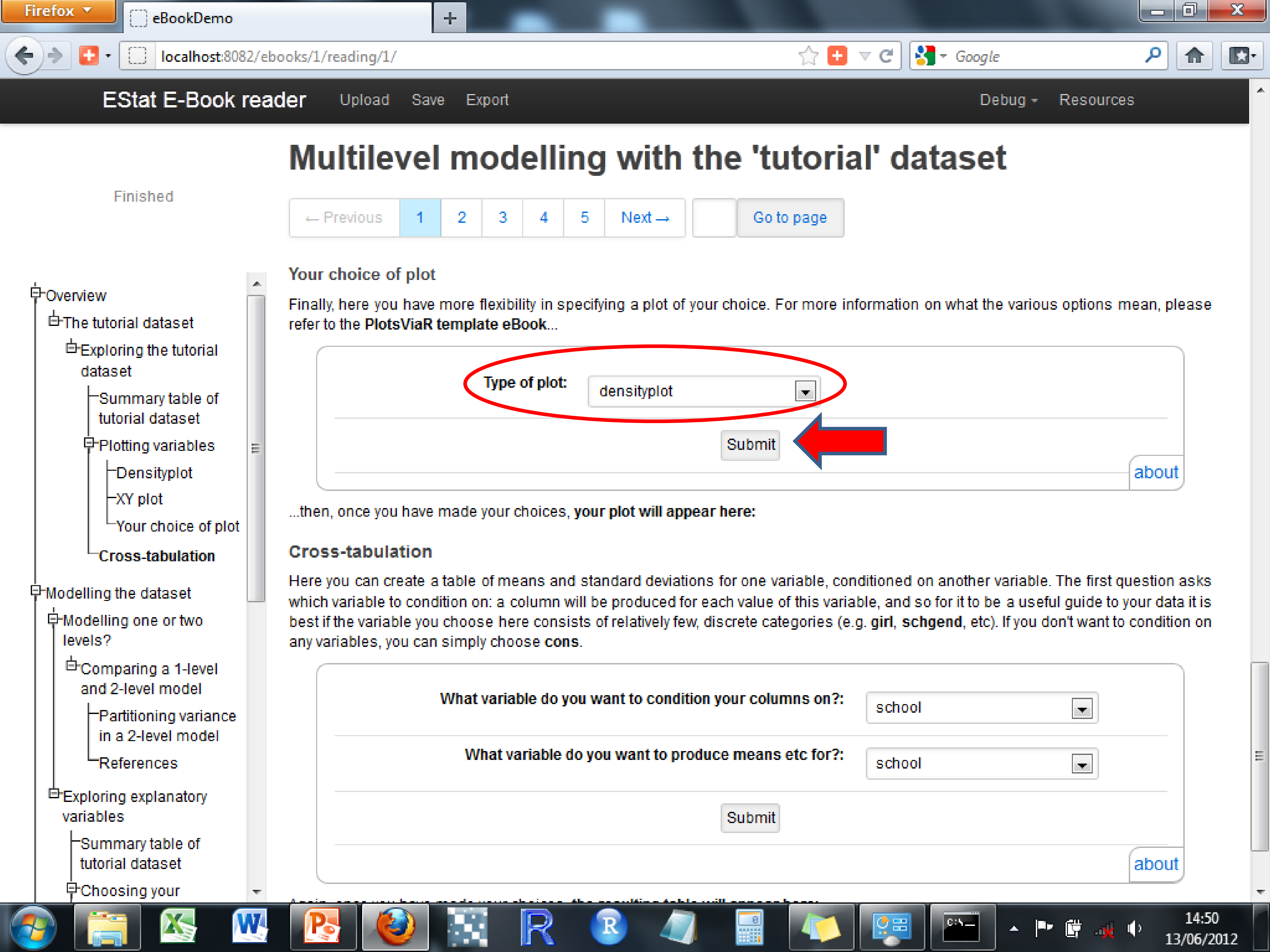
Summary table of tutorial dataset

Column name	n	Missing	Min	Max	Description
school	4059	0	1	65	Numeric school identifier
student	4059	0	1	198	Numeric student identifier
normexam	4059	0	-3.67	3.67	Students' exam score at age 16, normalised to have approximately a standard Normal distribution.
cons	4059	0	1	1	A column of ones. If included as an explanatory variable in a regression model, its coefficient is the intercept.
standlrt	4059	0	-2.93	3.02	Students' score at age 11 on the London Reading Test (LRT), standardised using Z-scores.
girl	4059	0	0	1	Students' gender: 0=boy; 1=girl
schgend	4059	0	1	3	School gender: 1=mixed; 2=boys' school; 3=girls' school
avslrt	4059	0	-0.76	0.64	Average LRT score in school
schav	4059	0	1	3	Average LRT score in school, coded into 3 categories: 1=bottom 25%; 2=middle 50%; 3=top 25%
vrband	4059	0	1	3	Students' score in test of verbal reasoning at age 11, coded into 3 categories: 1=top 25%; 2=middle 50%; 3=bottom 25%

Plotting variables

Here you can graphically-explore the **tutorial** dataset.

In the first two sections, below, you can produce a densityplot and XY plot, respectively; here you can re-specify your choice of variables



Multilevel modelling with the 'tutorial' dataset

Your choice of plot

Finally, here you have more flexibility in specifying a plot of your choice. For more information on what the various options mean, please refer to the **PlotsViaR template eBook**...

Type of plot: densityplot

Submit

...then, once you have made your choices, **your plot will appear here:**

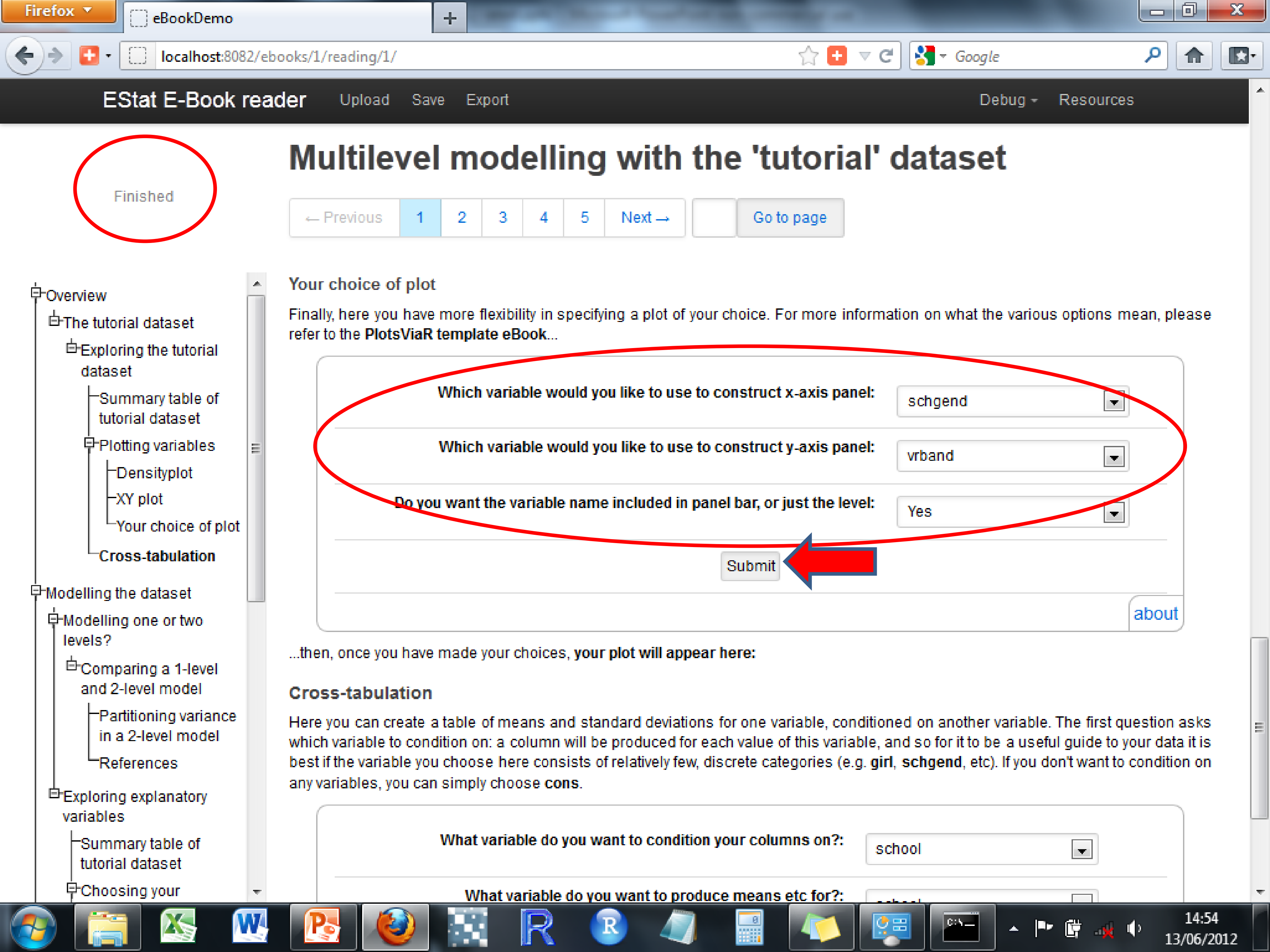
Cross-tabulation

Here you can create a table of means and standard deviations for one variable, conditioned on another variable. The first question asks which variable to condition on: a column will be produced for each value of this variable, and so for it to be a useful guide to your data it is best if the variable you choose here consists of relatively few, discrete categories (e.g. **girl**, **schgend**, etc). If you don't want to condition on any variables, you can simply choose **cons**.

What variable do you want to condition your columns on?: school

What variable do you want to produce means etc for?: school

Submit



- Overview
  - The tutorial dataset
    - Exploring the tutorial dataset
      - Summary table of tutorial dataset
      - Plotting variables
        - Densityplot
        - XY plot
        - Your choice of plot
      - Cross-tabulation
- Modelling the dataset
  - Modelling one or two levels?
    - Comparing a 1-level and 2-level model
      - Partitioning variance in a 2-level model
      - References
  - Exploring explanatory variables
    - Summary table of tutorial dataset
    - Choosing your

Multilevel modelling with the 'tutorial' dataset

Your choice of plot

Finally, here you have more flexibility in specifying a plot of your choice. For more information on what the various options mean, please refer to the **PlotsViaR template eBook**...

Which variable would you like to use to construct x-axis panel: schgend

Which variable would you like to use to construct y-axis panel: vrband

Do you want the variable name included in panel bar, or just the level: Yes

Submit

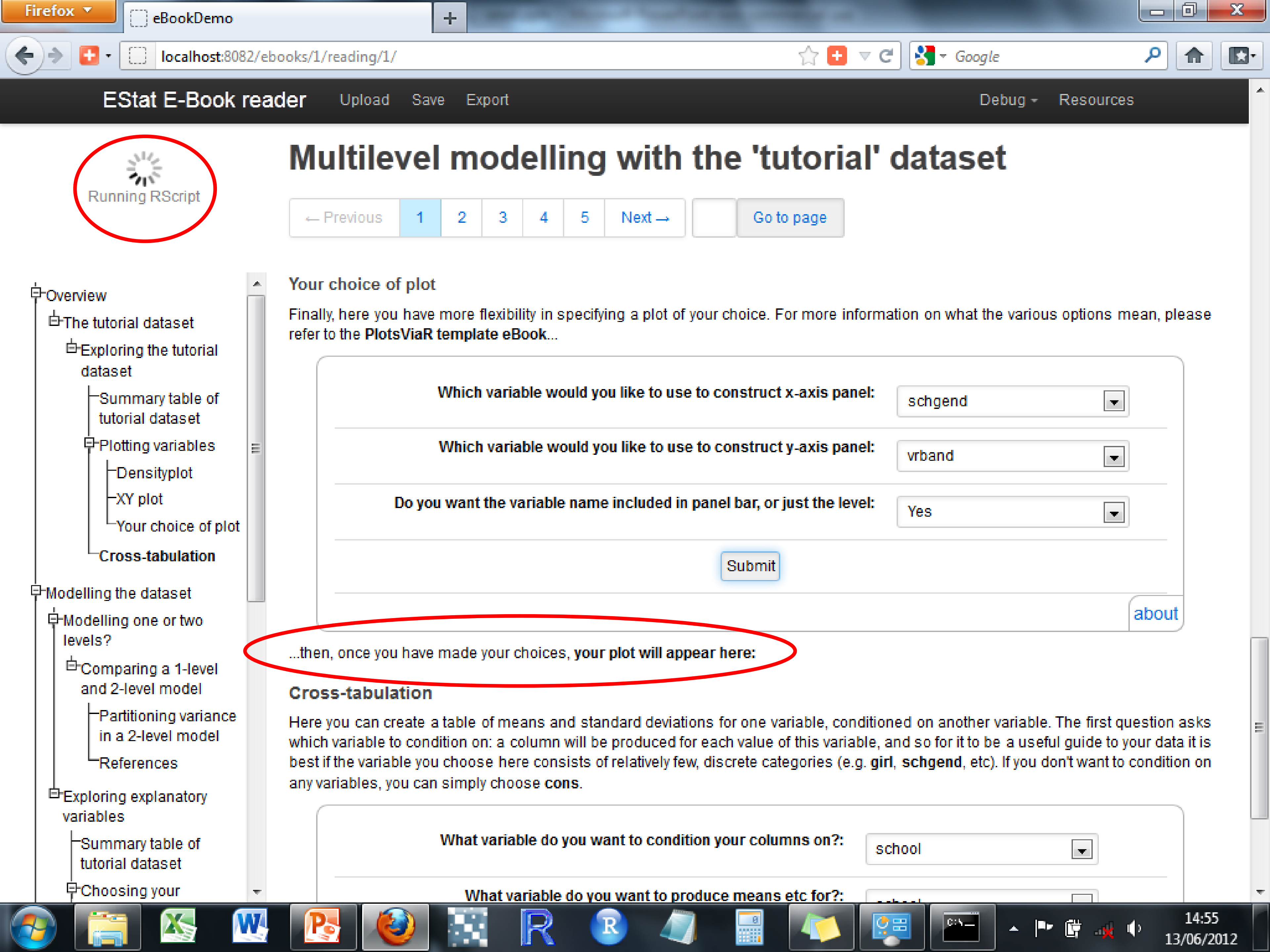
...then, once you have made your choices, **your plot will appear here:**

Cross-tabulation

Here you can create a table of means and standard deviations for one variable, conditioned on another variable. The first question asks which variable to condition on: a column will be produced for each value of this variable, and so for it to be a useful guide to your data it is best if the variable you choose here consists of relatively few, discrete categories (e.g. **girl**, **schgend**, etc). If you don't want to condition on any variables, you can simply choose **cons**.

What variable do you want to condition your columns on?: school

What variable do you want to produce means etc for?:



Multilevel modelling with the 'tutorial' dataset

Your choice of plot

Finally, here you have more flexibility in specifying a plot of your choice. For more information on what the various options mean, please refer to the **PlotsViaR template eBook**...

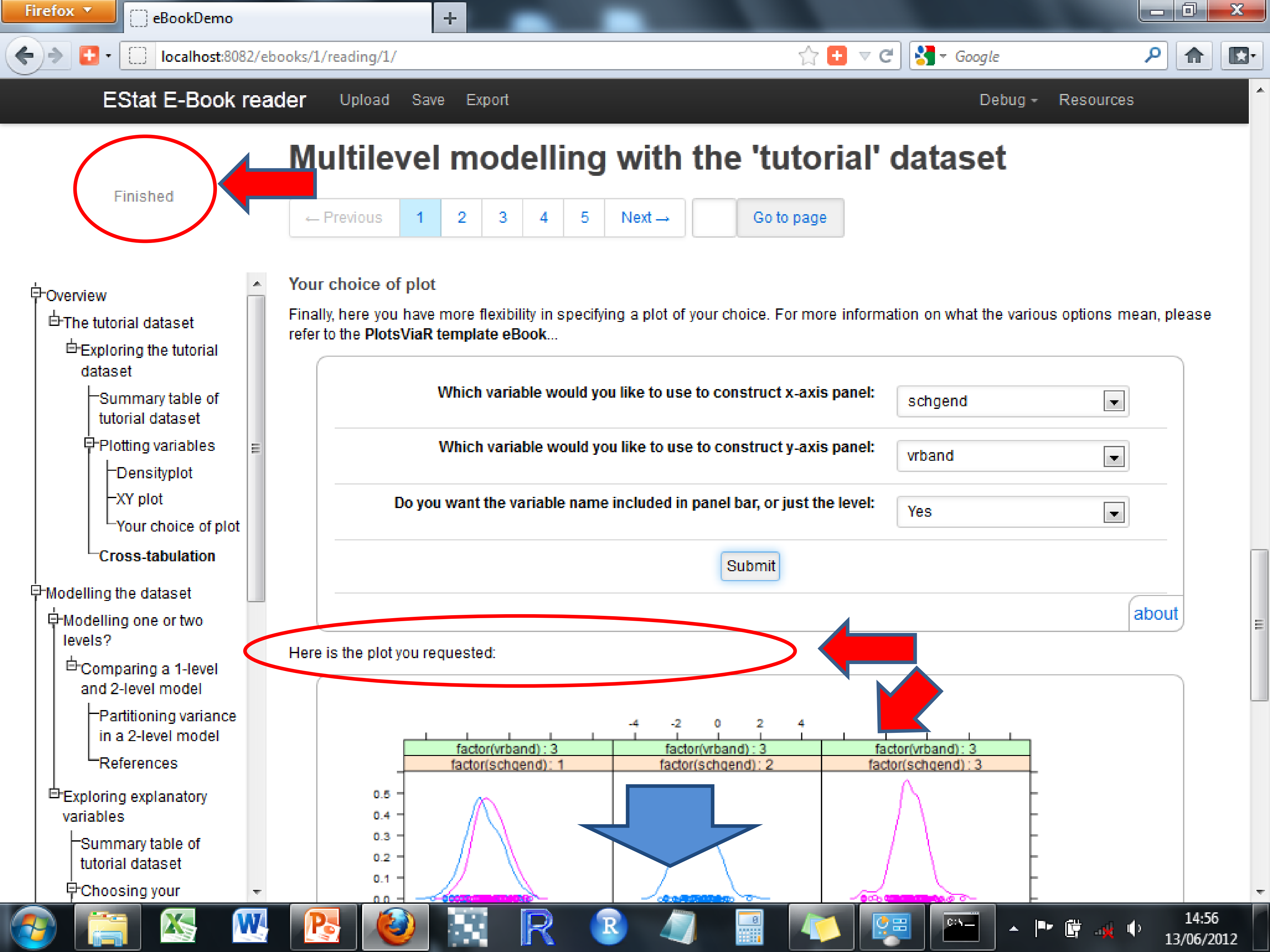
Which variable would you like to use to construct x-axis panel:	schgend
Which variable would you like to use to construct y-axis panel:	vrband
Do you want the variable name included in panel bar, or just the level:	Yes

...then, once you have made your choices, **your plot will appear here:**

Cross-tabulation

Here you can create a table of means and standard deviations for one variable, conditioned on another variable. The first question asks which variable to condition on: a column will be produced for each value of this variable, and so for it to be a useful guide to your data it is best if the variable you choose here consists of relatively few, discrete categories (e.g. **girl**, **schgend**, etc). If you don't want to condition on any variables, you can simply choose **cons**.

What variable do you want to condition your columns on?:	school
What variable do you want to produce means etc for?:	school



Finished

Multilevel modelling with the 'tutorial' dataset

Your choice of plot

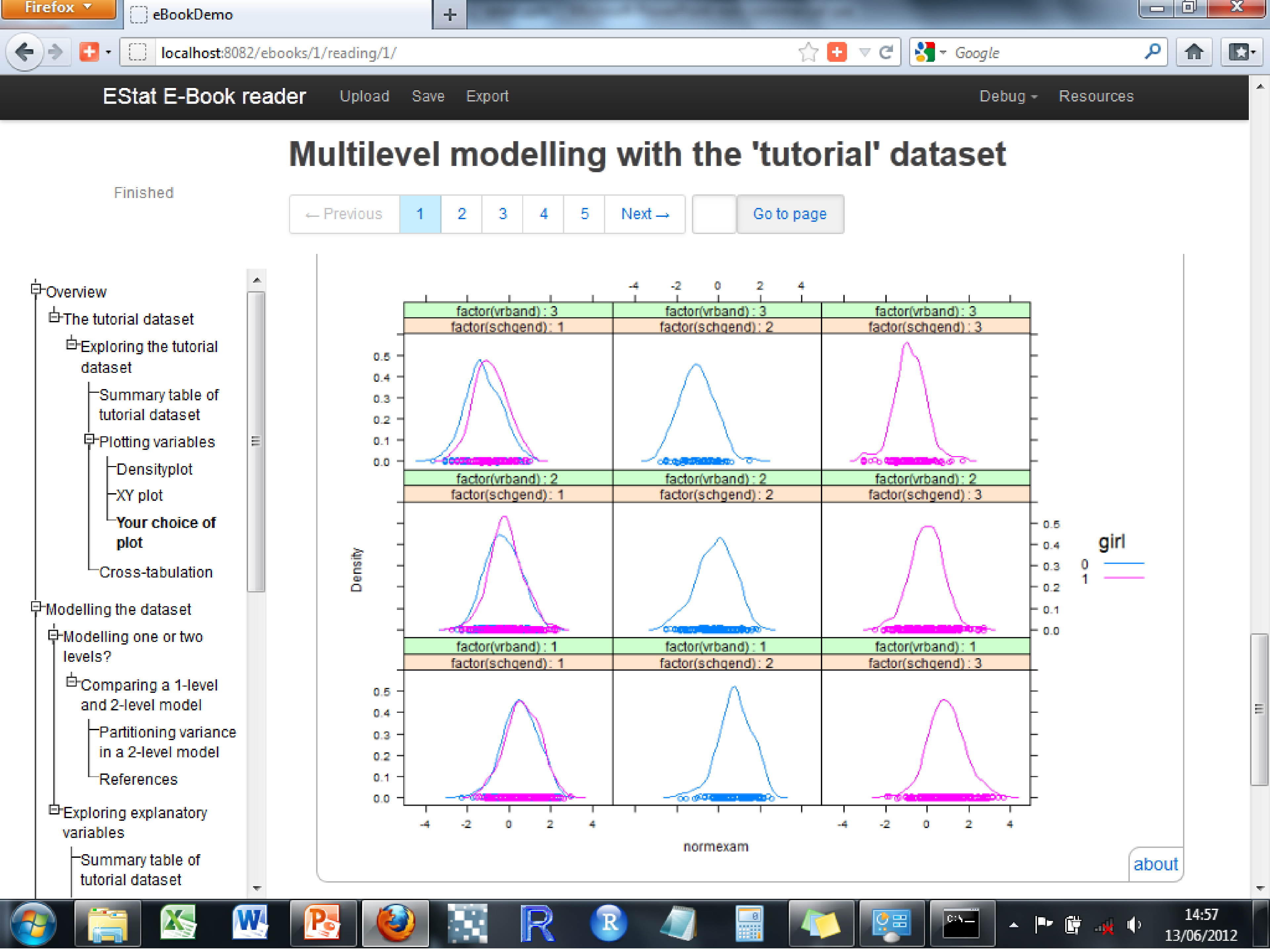
Finally, here you have more flexibility in specifying a plot of your choice. For more information on what the various options mean, please refer to the **PlotsViaR template eBook**...

Which variable would you like to use to construct x-axis panel: schgend

Which variable would you like to use to construct y-axis panel: vrband

Do you want the variable name included in panel bar, or just the level: Yes

Here is the plot you requested:



Multilevel modelling with the 'tutorial' dataset

Motivation for British Academy grant

- We ran a workshop demonstrating some of the new features in StatJR attended by John MacInnes and Rich Harris.
- In current ESRC grant we have been developing Statistical Analysis Assistants (SAAs) which are interactive eBooks that assist you with your analysis.
- As a start we considered automating simple operations.
- John and Rich thought an excellent addition would be using this for teaching and automated teaching material generation.
- The initial proposal was to do everything directly in StatJR but this got switched to creating the materials to use SPSS taking advantage of interoperability.

The end product – what the *student* gets

12 sets of practical exercises (pdfs) with 3 components

1. Takes student through a particular statistical concept in detail, and how to implement it in SPSS, using a specific data example (*learning component)*
2. A worksheet that asks the student to try out their knowledge by applying the techniques to a second dataset or set of variables *(practice component)*
3. Solutions to the worksheet *(self-evaluation component)*

What the *tutor* gets

- The set of static practicals using our choice of data example (PISA data as no restrictions on access)
- Instructions to how to use the Stat-JR software to tailor the practicals to their own choice of datasets/variables
- Makes it quick and easy to
  - Create a suite of discipline-specific materials for teaching and learning
  - Produce multiple versions of worksheets (with solutions) on different substantive topics or using different data sources

Work packages

The grant has 5 work packages:

1. Work package 1 consists of choosing topics and creating a single set of static practicals with solutions
2. Work package 2 consists of extending this to allow the materials to become dynamic and work with other datasets
3. Work package 3 consists of modifying StatJR to give QM teachers tools to customise the materials
4. Work package 4 consists of complementing the practicals / solutions with concept materials (learning component)
5. Work package 5 is demonstrating the materials to the community via a workshop

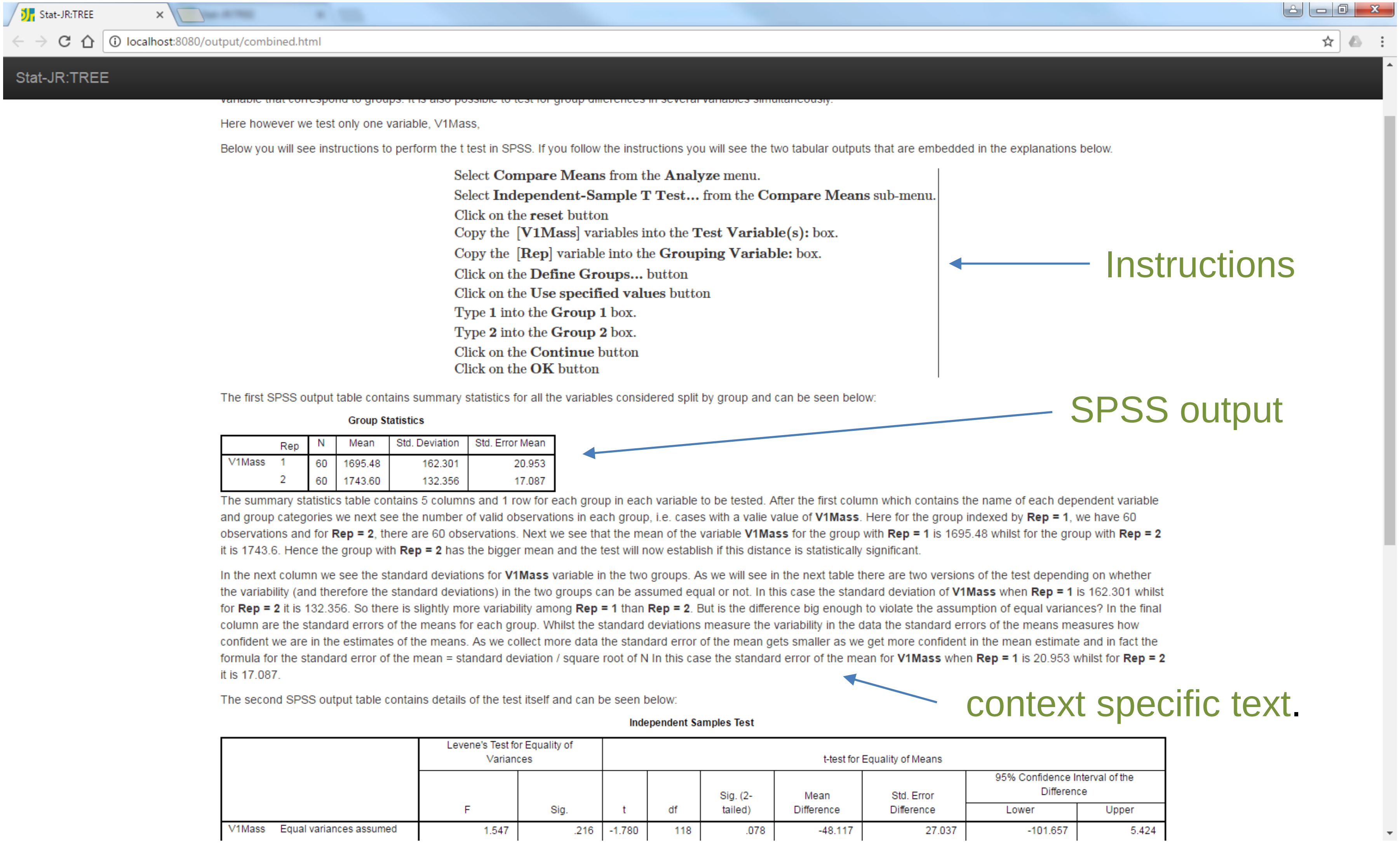
Work package 1

The list of topics is finalised as:

1. Describing categorical variables (summary stats and graphs)
2. Describing continuous variables (summary stats and graphs)
3. Tabulating data
4. Checking for normality
5. Two sample t tests
6. Paired t tests
7. Non parametric tests
8. Chi-squared tests
9. Correlation
10. Linear Regression
11. ANOVA
12. Multiple Regression

Work package 2 (and 1)

- In practice we have constructed the dynamic materials first and from them used test datasets to construct static files
- At this stage we have drafts of the first 10 practicals with 11 and 12 in process.
- In the next couple of slides we show a couple of screen shots to give an idea.
- Basically the practicals contain contextual text in terms of interpretation of the output but not the data context.
- When the materials are complete we intend to then construct a set of static materials using the PISA data and show how to add more data context.



variable that correspond to groups. It is also possible to test for group differences in several variables simultaneously.

Here however we test only one variable, V1Mass,

Below you will see instructions to perform the t test in SPSS. If you follow the instructions you will see the two tabular outputs that are embedded in the explanations below.

Select **Compare Means** from the **Analyze** menu.
Select **Independent-Sample T Test...** from the **Compare Means** sub-menu.
Click on the **reset** button
Copy the [**V1Mass**] variables into the **Test Variable(s):** box.
Copy the [**Rep**] variable into the **Grouping Variable:** box.
Click on the **Define Groups...** button
Click on the **Use specified values** button
Type **1** into the **Group 1** box.
Type **2** into the **Group 2** box.
Click on the **Continue** button
Click on the **OK** button

Instructions

The first SPSS output table contains summary statistics for all the variables considered split by group and can be seen below:

Group Statistics

	Rep	N	Mean	Std. Deviation	Std. Error Mean
V1Mass	1	60	1695.48	162.301	20.953
	2	60	1743.60	132.356	17.087

SPSS output

The summary statistics table contains 5 columns and 1 row for each group in each variable to be tested. After the first column which contains the name of each dependent variable and group categories we next see the number of valid observations in each group, i.e. cases with a valie value of **V1Mass**. Here for the group indexed by **Rep = 1**, we have 60 observations and for **Rep = 2**, there are 60 observations. Next we see that the mean of the variable **V1Mass** for the group with **Rep = 1** is 1695.48 whilst for the group with **Rep = 2** it is 1743.6. Hence the group with **Rep = 2** has the bigger mean and the test will now establish if this distance is statistically significant.

In the next column we see the standard deviations for **V1Mass** variable in the two groups. As we will see in the next table there are two versions of the test depending on whether the variability (and therefore the standard deviations) in the two groups can be assumed equal or not. In this case the standard deviation of **V1Mass** when **Rep = 1** is 162.301 whilst for **Rep = 2** it is 132.356. So there is slightly more variability among **Rep = 1** than **Rep = 2**. But is the difference big enough to violate the assumption of equal variances? In the final column are the standard errors of the means for each group. Whilst the standard deviations measure the variability in the data the standard errors of the means measures how confident we are in the estimates of the means. As we collect more data the standard error of the mean gets smaller as we get more confident in the mean estimate and in fact the formula for the standard error of the mean = standard deviation / square root of N In this case the standard error of the mean for **V1Mass** when **Rep = 1** is 20.953 whilst for **Rep = 2** it is 17.087.

context specific text.

The second SPSS output table contains details of the test itself and can be seen below:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
V1Mass	Equal variances assumed	1.547	.216	-1.780	118	.078	-48.117	27.037	-101.657	5.424

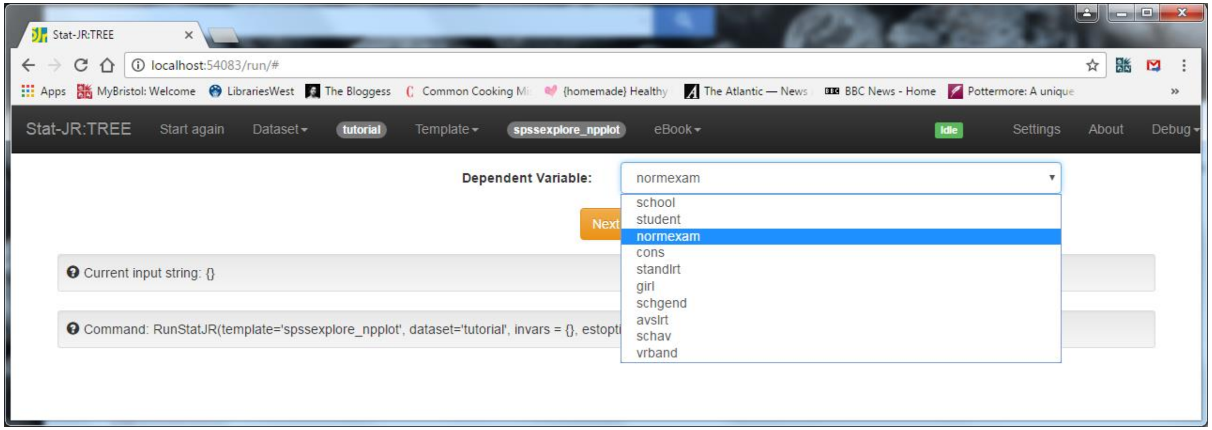
Work package 3

- The first two work packages are largely concerned with content construction whilst work package 3 involves improvements to StatJR specific to this grant. There are three main areas covered:

1) *Better Interfacing with SPSS* – initially it took 30 seconds per SPSS call to use system. Now once started most practicals can be constructed in under 10 seconds.
2) *Improving the eBook writer interface* – We will talk about in the next slide
3) *Improving Exporting of eBooks to PDF for printing* – initially the eBook interface was great for screen display but poor for printing. This has improved to keep SPSS outputs on single pages and to ensure they are rendered appropriately

eBook writer

The first stage is to choose the appropriate template – which aligns either with a full practical or a part of a practical and to choose a dataset



Dependent Variable: normexam

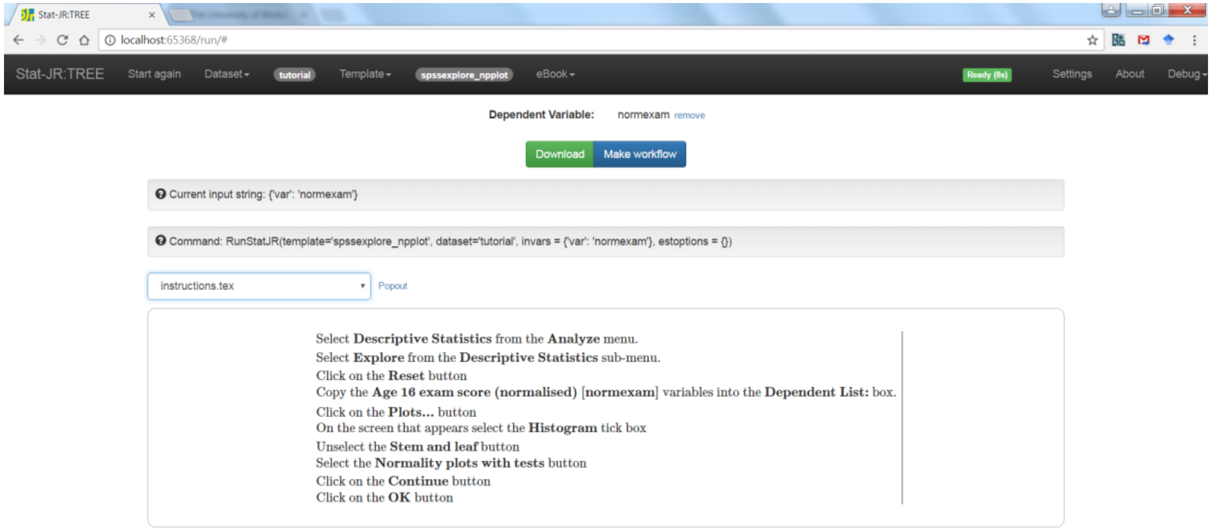
Current input string: {}

Command: RunStatJR(template='spssexplore_npplot', dataset='tutorial', invars = {}, estopti

The QM teacher then chooses the particular inputs that correspond to the variables to be used in the practical.

eBook writer

StatJR then creates lots of objects including SPSS outputs, contextual text describing the outputs and blocks of instructions for using SPSS as illustrated below.



Dependent Variable: normexam remove

Current input string: {'var': 'normexam'}

```
Command: RunStatJR(template='spssexplore_npplot', dataset='tutorial', invars = {'var': 'normexam'}, estoptions = {})
```

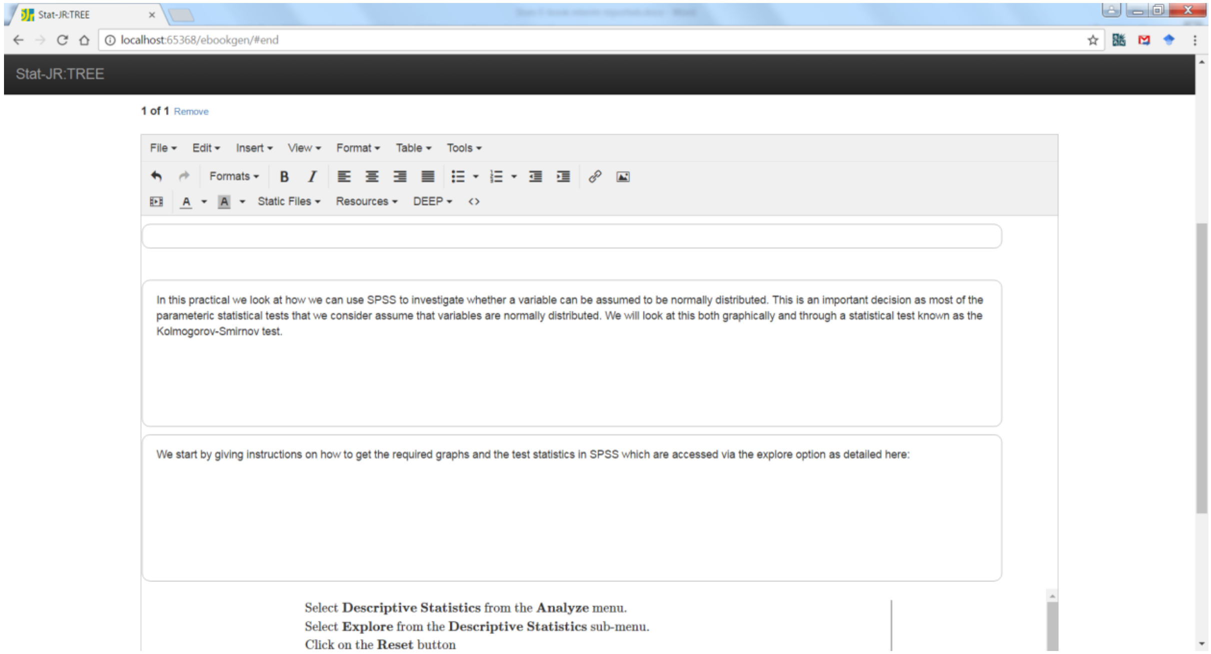
instructions.tex

Select **Descriptive Statistics** from the **Analyze** menu.
Select **Explore** from the **Descriptive Statistics** sub-menu.
Click on the **Reset** button
Copy the **Age 16 exam score (normalised)** [**normexam**] variables into the **Dependent List:** box.
Click on the **Plots...** button
On the screen that appears select the **Histogram** tick box
Unselect the **Stem and leaf** button
Select the **Normality plots with tests** button
Click on the **Continue** button
Click on the **OK** button

There is also a single combined output that puts these objects together

eBook writer

The QM teacher can then piece together the objects in turn as shown below:



In this practical we look at how we can use SPSS to investigate whether a variable can be assumed to be normally distributed. This is an important decision as most of the parameteric statistical tests that we consider assume that variables are normally distributed. We will look at this both graphically and through a statistical test known as the Kolmogorov-Smirnov test.

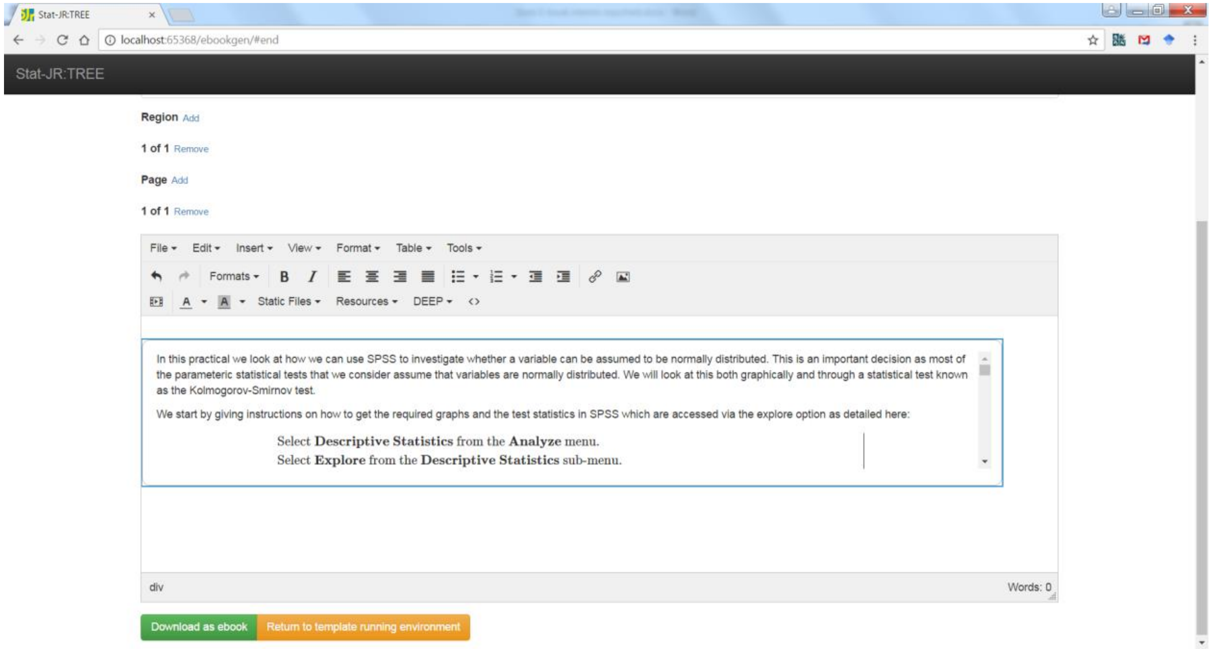
We start by giving instructions on how to get the required graphs and the test statistics in SPSS which are accessed via the explore option as detailed here:

Select **Descriptive Statistics** from the **Analyze** menu.
Select **Explore** from the **Descriptive Statistics** sub-menu.
Click on the **Reset** button

This allows them to add additional dataset specific contextual information and to construct practicals without solutions by omitting specific objects.

eBook writer

The other option is the instantly combined object that does the combining work for the QM teacher but is less customisable:



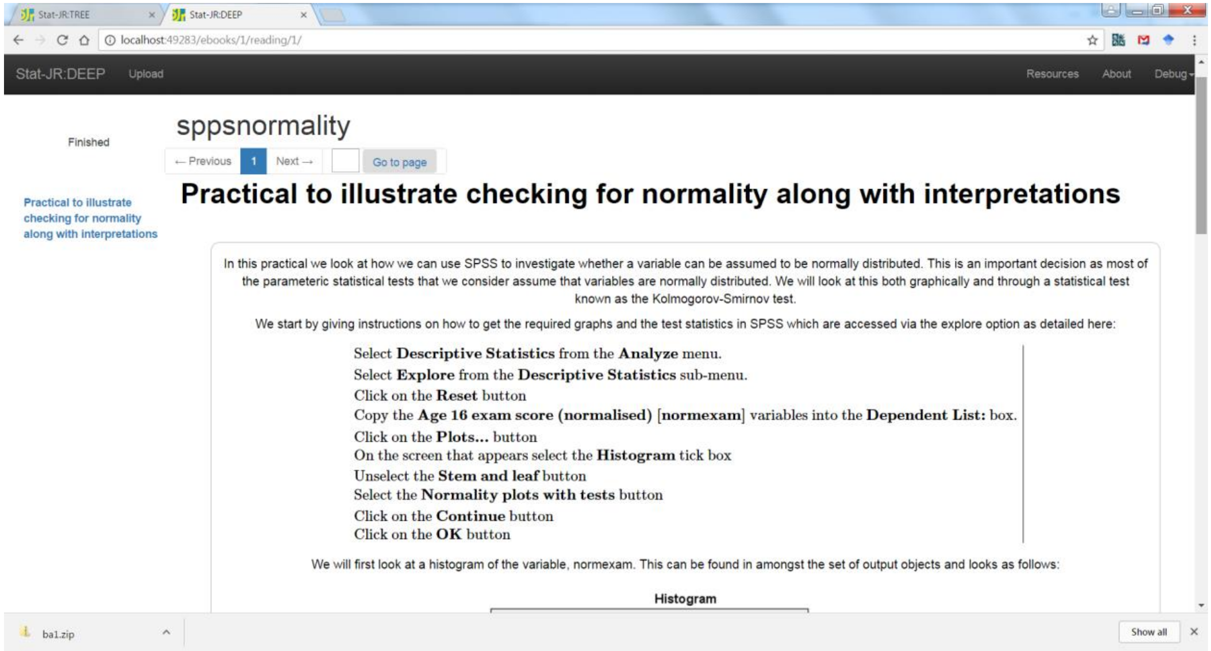
In this practical we look at how we can use SPSS to investigate whether a variable can be assumed to be normally distributed. This is an important decision as most of the parameteric statistical tests that we consider assume that variables are normally distributed. We will look at this both graphically and through a statistical test known as the Kolmogorov-Smirnov test.

We start by giving instructions on how to get the required graphs and the test statistics in SPSS which are accessed via the explore option as detailed here:

Select **Descriptive Statistics** from the **Analyze** menu.
Select **Explore** from the **Descriptive Statistics** sub-menu.

eBook writer

Finally in the eBook (DEEP) system we can see the final product and print to PDF file.



Stat-JR:TREE
Stat-JR:DEEP
localhost:49283/ebooks/1/reading/1/
Stat-JR:DEEP
Upload
Resources
About
Debug
Finished

sppsnormality

← Previous
1
Next →
Go to page
Practical to illustrate checking for normality along with interpretations

Practical to illustrate checking for normality along with interpretations

In this practical we look at how we can use SPSS to investigate whether a variable can be assumed to be normally distributed. This is an important decision as most of the parameteric statistical tests that we consider assume that variables are normally distributed. We will look at this both graphically and through a statistical test known as the Kolmogorov-Smirnov test.

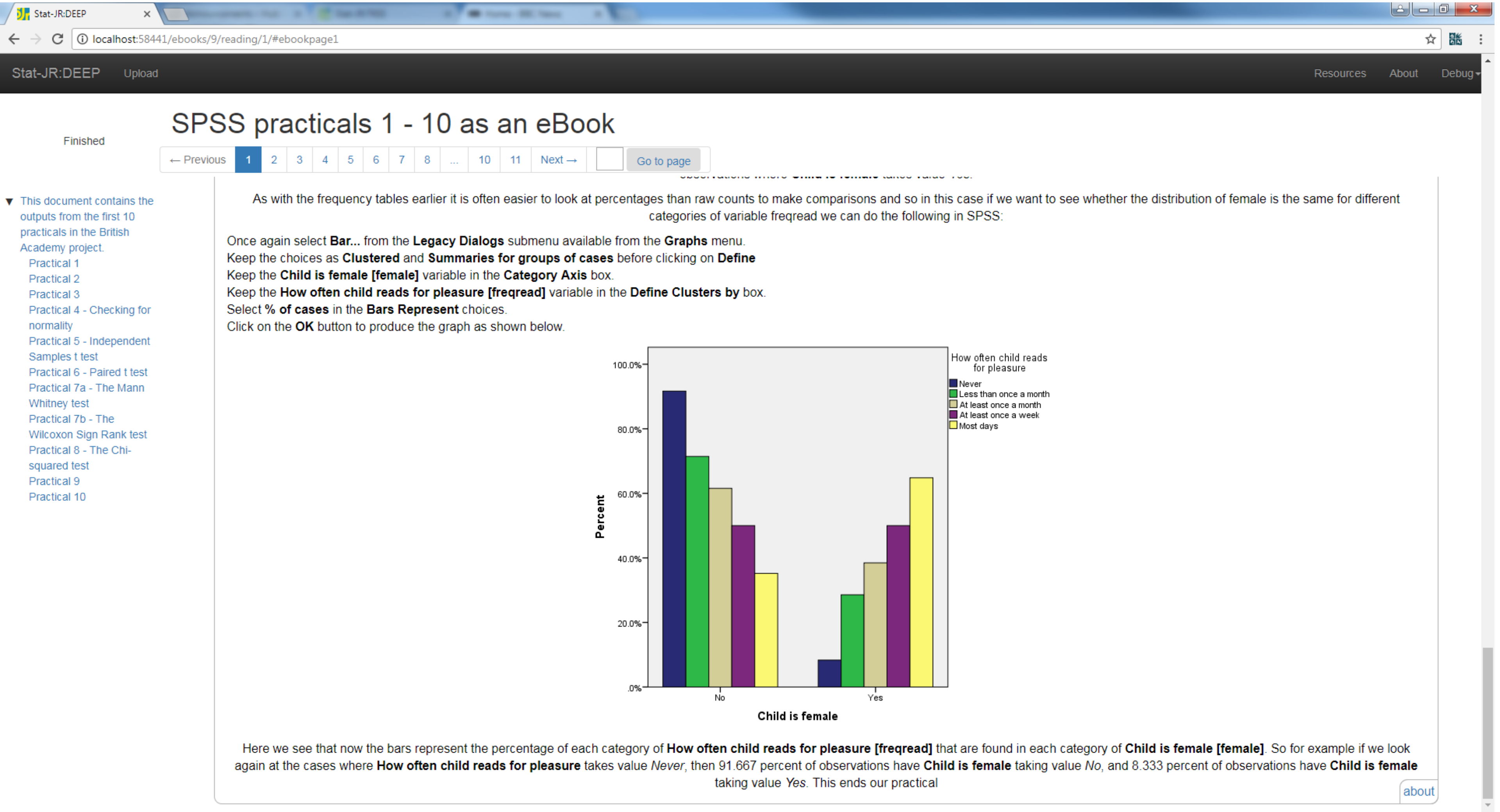
We start by giving instructions on how to get the required graphs and the test statistics in SPSS which are accessed via the explore option as detailed here:

Select **Descriptive Statistics** from the **Analyze** menu.
Select **Explore** from the **Descriptive Statistics** sub-menu.
Click on the **Reset** button
Copy the **Age 16 exam score (normalised)** [**normexam**] variables into the **Dependent List:** box.
Click on the **Plots...** button
On the screen that appears select the **Histogram** tick box
Unselect the **Stem and leaf** button
Select the **Normality plots with tests** button
Click on the **Continue** button
Click on the **OK** button

We will first look at a histogram of the variable, normexam. This can be found in amongst the set of output objects and looks as follows:

Histogram

More outputs



SPSS practicals 1 - 10 as an eBook

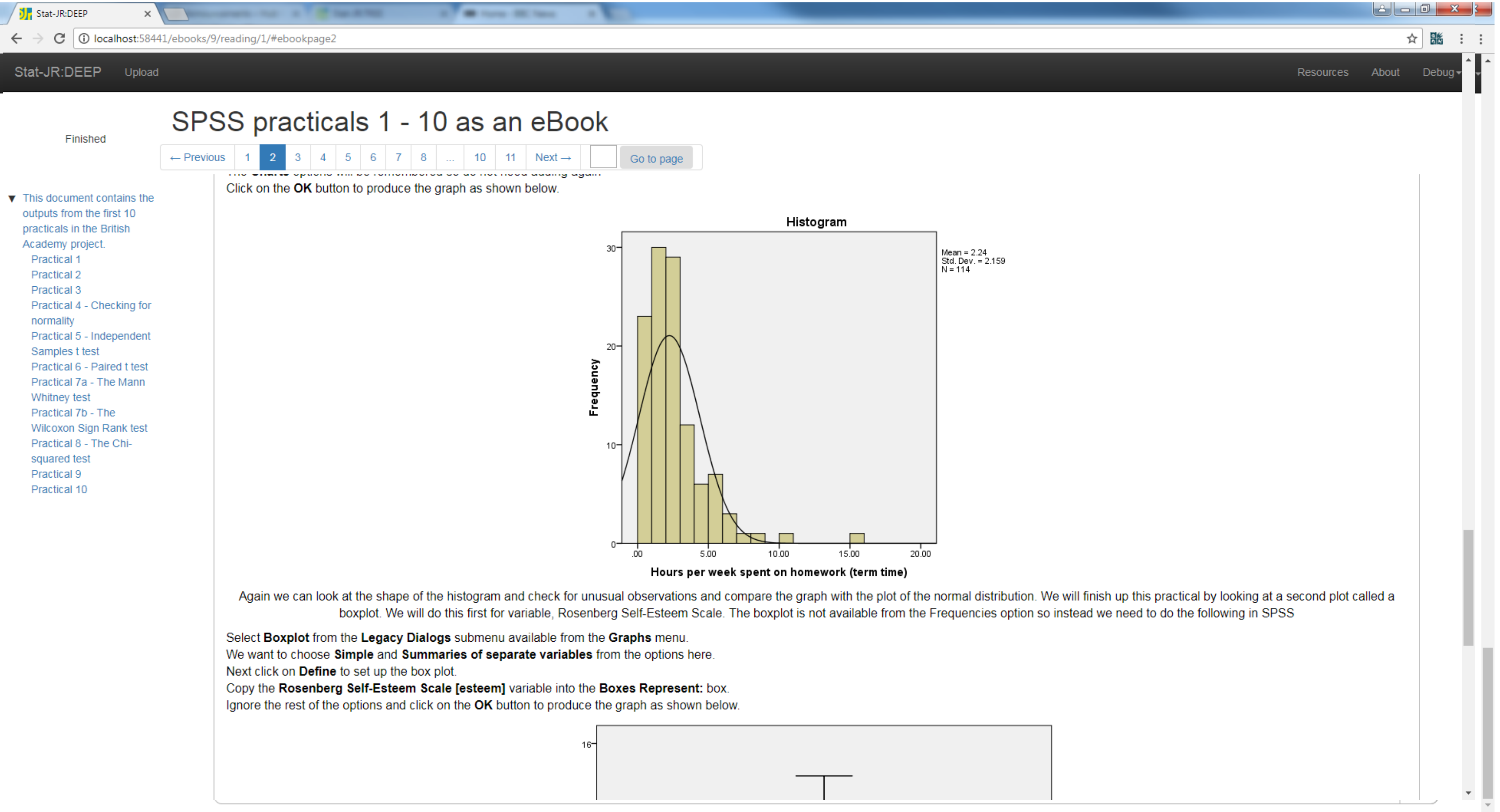
- This document contains the outputs from the first 10 practicals in the British Academy project.
  - Practical 1
  - Practical 2
  - Practical 3
  - Practical 4 - Checking for normality
  - Practical 5 - Independent Samples t test
  - Practical 6 - Paired t test
  - Practical 7a - The Mann Whitney test
  - Practical 7b - The Wilcoxon Sign Rank test
  - Practical 8 - The Chi-squared test
  - Practical 9
  - Practical 10

As with the frequency tables earlier it is often easier to look at percentages than raw counts to make comparisons and so in this case if we want to see whether the distribution of female is the same for different categories of variable freqread we can do the following in SPSS:

Once again select **Bar...** from the **Legacy Dialogs** submenu available from the **Graphs** menu.
Keep the choices as **Clustered** and **Summaries for groups of cases** before clicking on **Define**
Keep the **Child is female [female]** variable in the **Category Axis** box.
Keep the **How often child reads for pleasure [freqread]** variable in the **Define Clusters by** box.
Select **% of cases** in the **Bars Represent** choices.
Click on the **OK** button to produce the graph as shown below.

Here we see that now the bars represent the percentage of each category of **How often child reads for pleasure [freqread]** that are found in each category of **Child is female [female]**. So for example if we look again at the cases where **How often child reads for pleasure** takes value *Never*, then 91.667 percent of observations have **Child is female** taking value *No*, and 8.333 percent of observations have **Child is female** taking value *Yes*. This ends our practical

More outputs



SPSS practicals 1 - 10 as an eBook

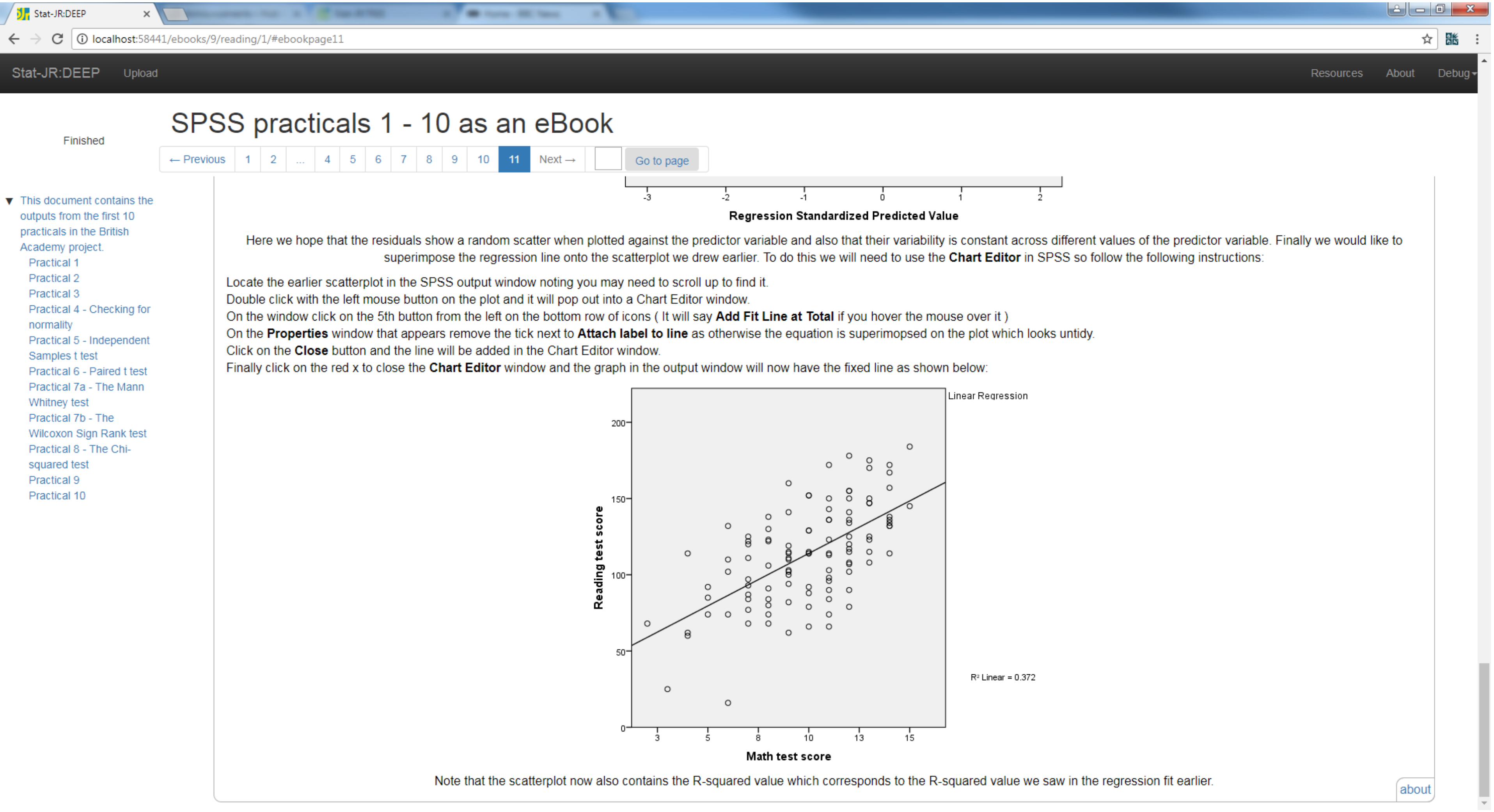
- This document contains the outputs from the first 10 practicals in the British Academy project.
  - Practical 1
  - Practical 2
  - Practical 3
  - Practical 4 - Checking for normality
  - Practical 5 - Independent Samples t test
  - Practical 6 - Paired t test
  - Practical 7a - The Mann Whitney test
  - Practical 7b - The Wilcoxon Sign Rank test
  - Practical 8 - The Chi-squared test
  - Practical 9
  - Practical 10

The **Charts** options will be remembered so do not need adding again.
Click on the **OK** button to produce the graph as shown below.

Again we can look at the shape of the histogram and check for unusual observations and compare the graph with the plot of the normal distribution. We will finish up this practical by looking at a second plot called a boxplot. We will do this first for variable, Rosenberg Self-Esteem Scale. The boxplot is not available from the Frequencies option so instead we need to do the following in SPSS

Select **Boxplot** from the **Legacy Dialogs** submenu available from the **Graphs** menu.
We want to choose **Simple** and **Summaries of separate variables** from the options here.
Next click on **Define** to set up the box plot.
Copy the **Rosenberg Self-Esteem Scale [esteem]** variable into the **Boxes Represent:** box.
Ignore the rest of the options and click on the **OK** button to produce the graph as shown below.

More outputs



SPSS practicals 1 - 10 as an eBook

- This document contains the outputs from the first 10 practicals in the British Academy project.
  - Practical 1
  - Practical 2
  - Practical 3
  - Practical 4 - Checking for normality
  - Practical 5 - Independent Samples t test
  - Practical 6 - Paired t test
  - Practical 7a - The Mann Whitney test
  - Practical 7b - The Wilcoxon Sign Rank test
  - Practical 8 - The Chi-squared test
  - Practical 9
  - Practical 10

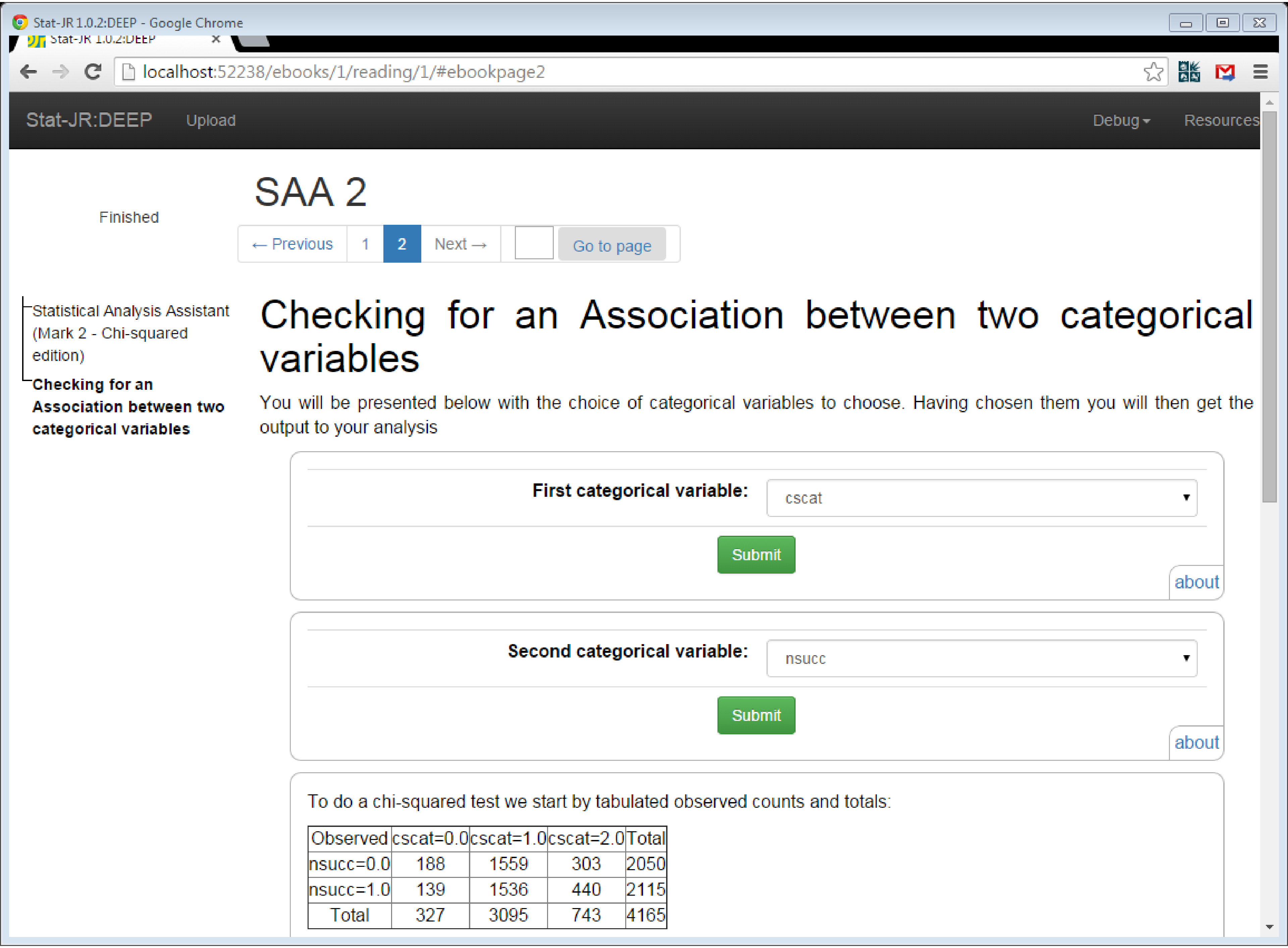
Here we hope that the residuals show a random scatter when plotted against the predictor variable and also that their variability is constant across different values of the predictor variable. Finally we would like to superimpose the regression line onto the scatterplot we drew earlier. To do this we will need to use the **Chart Editor** in SPSS so follow the following instructions:

Locate the earlier scatterplot in the SPSS output window noting you may need to scroll up to find it.
Double click with the left mouse button on the plot and it will pop out into a Chart Editor window.
On the window click on the 5th button from the left on the bottom row of icons (It will say **Add Fit Line at Total** if you hover the mouse over it)
On the **Properties** window that appears remove the tick next to **Attach label to line** as otherwise the equation is superimposed on the plot which looks untidy.
Click on the **Close** button and the line will be added in the Chart Editor window.
Finally click on the red x to close the **Chart Editor** window and the graph in the output window will now have the fixed line as shown below:

Note that the scatterplot now also contains the R-squared value which corresponds to the R-squared value we saw in the regression fit earlier.

Work package 4

- The original plan in the grant was to construct concept materials using StatJR to supplement the students learning.
- An example of such a concept eBook is shown overleaf and we have others for summary statistics and other statistical tests.
- Given the switched focus to SPSS we propose to integrate the conceptual material within the learning component of each practical (so that conceptual understanding and software skills are developed side-by-side)


Stat-JR 1.0.2:DEEP - Google Chrome
Stat-JR 1.0.2:DEEP
localhost:52238/ebooks/1/reading/1/#ebookpage2
Stat-JR:DEEP
Upload
Debug
Resources
Finished
← Previous
1
2
Next →
Go to page
Statistical Analysis Assistant (Mark 2 - Chi-squared edition)
Checking for an Association between two categorical variables
Submit
about
Submit
about

SAA 2

Checking for an Association between two categorical variables

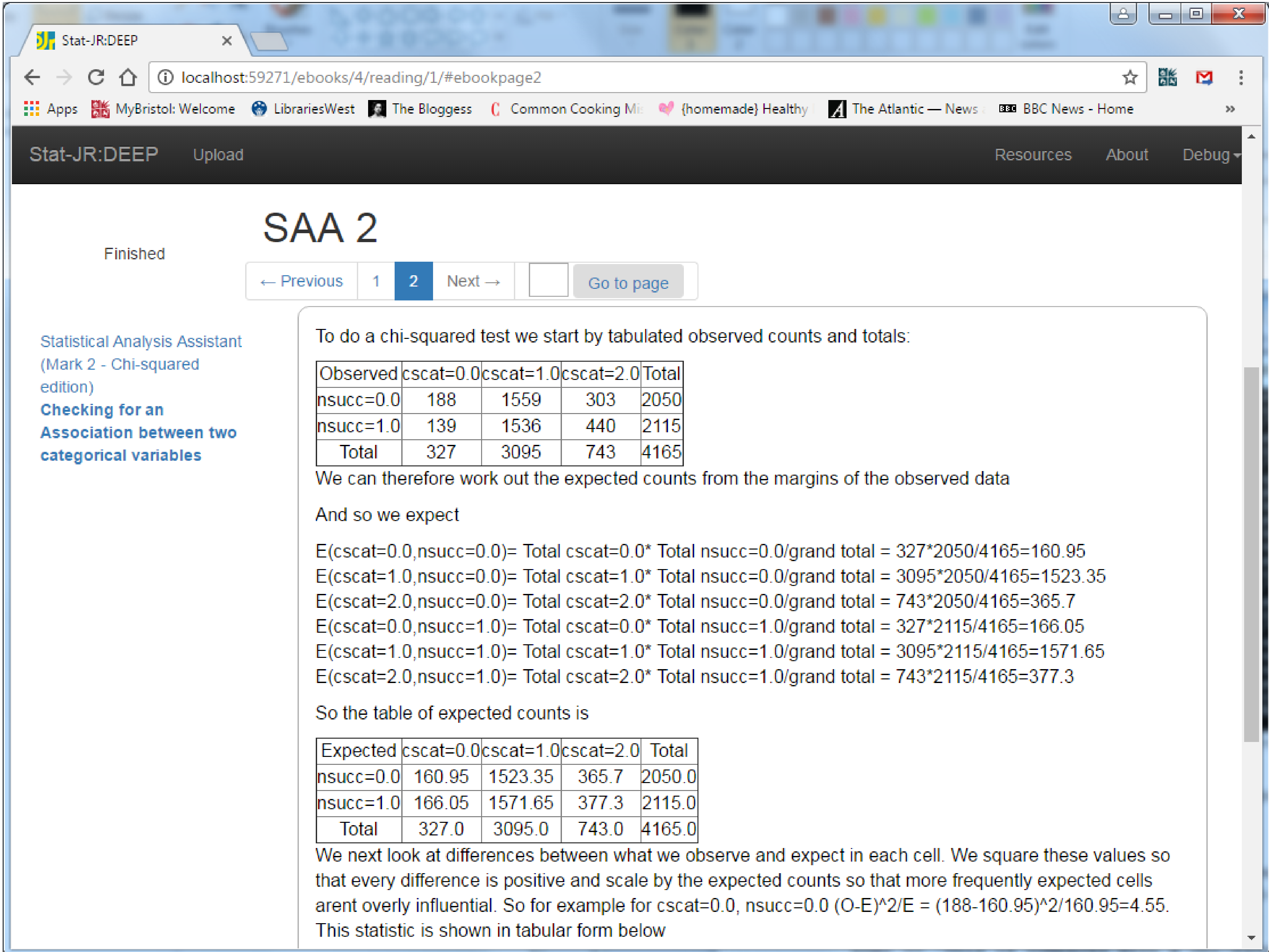
You will be presented below with the choice of categorical variables to choose. Having chosen them you will then get the output to your analysis

First categorical variable: cscat

Second categorical variable: nsucc

To do a chi-squared test we start by tabulated observed counts and totals:

Observed	cscat=0.0	cscat=1.0	cscat=2.0	Total
nsucc=0.0	188	1559	303	2050
nsucc=1.0	139	1536	440	2115
Total	327	3095	743	4165



SAA 2

To do a chi-squared test we start by tabulated observed counts and totals:

Observed	cscat=0.0	cscat=1.0	cscat=2.0	Total
nsucc=0.0	188	1559	303	2050
nsucc=1.0	139	1536	440	2115
Total	327	3095	743	4165

We can therefore work out the expected counts from the margins of the observed data

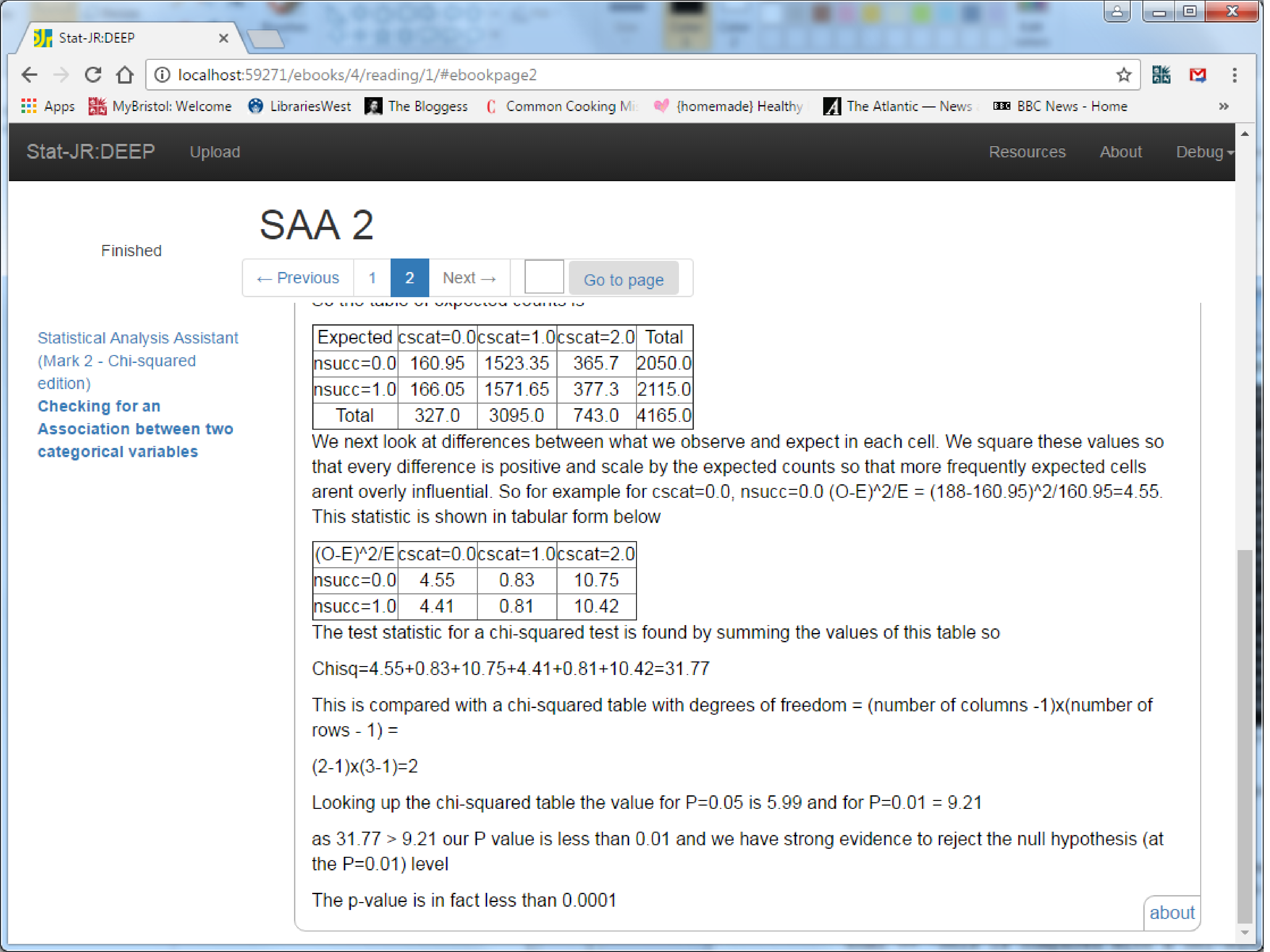
And so we expect

E(cscat=0.0,nsucc=0.0)= Total cscat=0.0* Total nsucc=0.0/grand total = 327*2050/4165=160.95
E(cscat=1.0,nsucc=0.0)= Total cscat=1.0* Total nsucc=0.0/grand total = 3095*2050/4165=1523.35
E(cscat=2.0,nsucc=0.0)= Total cscat=2.0* Total nsucc=0.0/grand total = 743*2050/4165=365.7
E(cscat=0.0,nsucc=1.0)= Total cscat=0.0* Total nsucc=1.0/grand total = 327*2115/4165=166.05
E(cscat=1.0,nsucc=1.0)= Total cscat=1.0* Total nsucc=1.0/grand total = 3095*2115/4165=1571.65
E(cscat=2.0,nsucc=1.0)= Total cscat=2.0* Total nsucc=1.0/grand total = 743*2115/4165=377.3

So the table of expected counts is

Expected	cscat=0.0	cscat=1.0	cscat=2.0	Total
nsucc=0.0	160.95	1523.35	365.7	2050.0
nsucc=1.0	166.05	1571.65	377.3	2115.0
Total	327.0	3095.0	743.0	4165.0

We next look at differences between what we observe and expect in each cell. We square these values so that every difference is positive and scale by the expected counts so that more frequently expected cells arent overly influential. So for example for cscat=0.0, nsucc=0.0 (O-E)^2/E = (188-160.95)^2/160.95=4.55. This statistic is shown in tabular form below



SAA 2

So the table of expected counts is

Expected	cscat=0.0	cscat=1.0	cscat=2.0	Total
nsucc=0.0	160.95	1523.35	365.7	2050.0
nsucc=1.0	166.05	1571.65	377.3	2115.0
Total	327.0	3095.0	743.0	4165.0

We next look at differences between what we observe and expect in each cell. We square these values so that every difference is positive and scale by the expected counts so that more frequently expected cells arent overly influential. So for example for cscat=0.0, nsucc=0.0 (O-E)^2/E = (188-160.95)^2/160.95=4.55. This statistic is shown in tabular form below

(O-E)^2/E	cscat=0.0	cscat=1.0	cscat=2.0
nsucc=0.0	4.55	0.83	10.75
nsucc=1.0	4.41	0.81	10.42

The test statistic for a chi-squared test is found by summing the values of this table so

Chisq=4.55+0.83+10.75+4.41+0.81+10.42=31.77

This is compared with a chi-squared table with degrees of freedom = (number of columns -1)x(number of rows - 1) =

(2-1)x(3-1)=2

Looking up the chi-squared table the value for P=0.05 is 5.99 and for P=0.01 = 9.21

as 31.77 > 9.21 our P value is less than 0.01 and we have strong evidence to reject the null hypothesis (at the P=0.01) level

The p-value is in fact less than 0.0001

Work package 5

- For this work package we intend to run a workshop to demonstrate the system and get feedback.
- The original timetable for this is month 21 or roughly Xmas time and so we will liaise with John MacInnes and Q-step leads to find when precisely works.
- We have demonstrated aspects of the software to John’s group in Edinburgh who were enthusiastic and discussed the software and topics with colleagues at Bristol, Exeter and Cardiff.

Questions

????????????