Stat-JR eBook interface and Statistical Analysis Assistants

Hi. I'm Professor Bill Browne from the Centre for multilevel modelling in Bristol, and this is the second of my StatJR talks. In this talk I will introduce StatJR's eBook interface and talk more about the concept of statistical analysis assistants i.e. computer programs that help you do your statistical analysis.

So, let's begin by asking what is an e-book? Well as this slide suggests if we take a book on the left, a computer on the right, smash them together, out pops an eBook reader as shown. But what we have to ask is the advantage here? In fact, we are interested in statistical eBooks, and clearly we can have statistical textbooks in electronic format. But the question is what enhancements we can add when we're in the electronic world? In some sense what we are trying to do is combine a statistics package with a book. So here behind me now is a screenshot of an example of an ebook which we've made using the StatJR's DEEP eBook interface. This eBook illustrates multilevel modelling with an example dataset. And we see to the left a hierarchical table of contents and at the top a page list to navigate through the eBook. Of course, in the digital world pages do not need to be of standard length, so for this, so far, this looks like a PDF reader, but if we scroll down, here we see after the data set has been described there is a table giving some summary statistics for the data set. Scrolling down further we finally see something that distinguishes this eBook from just any old PDF reader as embedded in the page is a little widget asking the user or the reader to choose a type of plot. So, this is new. If we look at the plot and then select the density plot from the pulldown list, we can then click on the submit button and the eBook will change and ask further questions which will appear underneath the original question. Here we see further questions asking for inputs to fully specify the desired plot. After we press submit more inputs will appear, and after we press submit for the final time having answered all the questions we see that the progress gauge in the top left corner of the screen changes from showing the word finished to indicate that the StatJR eBook is running an execution. This is happening behind the scenes and in this case our eBook is interoperating with the R statistical software package. Furthermore, text has appeared on the screen indicating to the reader where the resulting plot will appear. Once execution finishes we will see the word Finished in the top left corner and the text that we saw earlier is now replaced by the plot that we see here as soon as it is rendered available. As we can see if we scroll down further we can see the

whole plot. This plot is not a particularly interesting plot but has primarily been chosen as it is colourful. This is an example eBook, but we could repeat this exercise if we wish, we could choose a different plot and we would get a different version of the same eBook. That way this eBook is able to have many different versions. So that gives you a bit of background about what a statistical eBook is in the StatJR package.

Now we're going to move on to the topic of statistical analysis assistants. Here what we're going to do is we're going to adapt our eBook system to allow what we call workflows to be constructed which will describe how the steps and the statistical analysis fit together. There may be many statistical analysis assistants adapted to how different researchers approach quantitative research. In this respect opinion is kind of divided on how far one can take the idea from nowhere at all to complete automation i.e. we could pour the data set in at the top and let the computer sort it out. In practice the probable end point will be somewhere in between. So, to start with it's probably easiest to think about how we might automate single statistical operations. Okay, so now behind me you can see our first statistical analysis assistant which as it says simply tells the researcher that runs the eBook to go and ask a statistician to help them with their questions. Perhaps this isn't the most helpful of eBooks but at least it's a starting point which we can then immediately build on. So, if we move on we will now show you our very simple statistical analysis assistant mark 2. Here is a second statistical analysis assistant. This one assumes that we only know, or only have knowledge of one statistical technique, namely the chi-squared test, and what the analysis assistant does is it asks questions to establish if a chi-squared test is what is required. If not, the user gets the usual question of going to ask a statistician, but if they're very lucky and they do need a chi-squared test then they can turn to page 2 of the eBook. Here they will see inputs for the two categorical variables that will be needed to run a chi-squared test. This is an example, and this example comes from ecology, so what we have here for nesting birds is two variables - the birds clutch size, and a binary variable the nest success, okay. So once the inputs are added then the statistical analysis assistant will perform in a completely transparent way a chi-squared test. We first see, as one would with a chi-squared test, the observed data being tabulated. And then we see step by step how we can construct the expected counts for every cell of that table that correspond to those observed data and are required for a chi-square test. From here as we scroll down the eBook the test statistic itself is calculated and displayed and then we have a statement that is given as to whether the test is

significant or not. And this will vary depending on the user inputs. And then we can establish whether or not the hypothesis that we're interested in can be rejected.

So, to summarize, we see that we can expose the four steps that are required for the chi-squared test. So, we see that the assistant can do all of the stats that are required although what it can't do is find contextual meaning in the outcomes. That's beyond it, it's just a computer at the end of the day. We find that statistical tests and tables are easy for a computer to generate, though interpreting figures, as we'll see later, is somewhat harder, so where do we go next? Well when I was a boy there were a series of books, the first of which was called the Warlock of Firetop Mountain. They were a genre of what we would call interactive books published way back in 1982 and lapped up by 10-year olds like myself at the time. It combined a book and a flowchart and worked something like this. You would read the text and the text would say, the goblin advances towards you shouting words that you can't understand; do you try to make a conversation, if so turn to one page, do you run past the Goblin, turn to another page, or do you draw your sword and fight, turn to a third page. Okay so you wonder why I'm talking about fantasy books. Well in reality what we have here is a flow chart in a book which has been disguised by the random page movements, and there were a variety of endings in this case 99% of them would involve you dying. And if you were using one of these books, you'd probably put your thumb in between the pages so that you wouldn't end up with one of those endings.

So flowcharts exist a lot in statistics books often at the back of a statistics book, and they would have questions asking things like what are the number of variables in what you're interested in, what are their types, are they normally distributed, and then the end point would be some sort of test. Thinking about statistical analysis assistants, the idea might be to include branches where we haven't yet written material. In other words, the equivalent of ending up dead in the Warlock of Firetop Mountain scenario, would be the default go and ask the statistician end point, possibly taking your answers to the flowchart with you to help. So, the flowchart idea is appealing as it may to some degree mimic a statistical consultation. In other words, the statistical analysis assistant is acting a bit like the stats consultant. If the system was flexible enough then each statistician could tune their own SAA, and they could tune it to their own approach to doing analysis and how much they feel can be comfortably automated, and we could even include uncertainty and where things are not so clear, options

could be incorporated. eBooks can even contain hyperlinks so that further background on the proposed statistical methods or other examples could be easily found.

So, let's move on to say how we would actually implement this in practice. What we were interested in now is the concept of workflows, and workflows allow the sequencing of a series of operations to perform an analysis. We have another version of StatJR called StatJR leaf, which is based around a new front-end written using the Blockly system, as we'll see in a minute. If any of you have children in school Blockly is a little bit like scratch, and it's just a visual programming language. The user can link up templates themselves in a user-friendly visual way, and the workflows can then be included within an electronic eBook. We will use this system in our SAAs. Here for example is a log- file style workflow. Basically, this workflow allows us to select a dataset, give some inputs, and then fit a histogram to a variable that we've selected, and display several objects. It's a Hist/Skew template that we're using so we're going to look at the histogram and see whether it is symmetric or not. Here is a first run-through of the workflow and we have chosen a fairly symmetrical variable. What we can see at the top is the histogram itself, and then some accompanying text that the SAA has generated which has actually been calculated via a test statistic called a skewness statistic and identifies that in this case the variable is fairly symmetric. Just to show that the system works, here is another go. We have a different variable, and this variable is not symmetric it's very skew, and you see that different messages appear for such variables. Our next step therefore, would be to move on to more complicated models, and here we have, we're going to talk, about a linear regression model. This is a much fuller analysis as we're going to do some pre and post operations. Some summary statistics before and after fitting the regression itself. I will show you the SAA in a minute or two, but first here is a possible analysis for a regression type model. That's quite a scary diagram you see there. This Spaghetti Junction like diagram shows an actual statistical analysis. And what we did when we were running the grant that wrote the StatJR software, is we asked that our statistics reading group each come along with the steps they would use to perform an analysis which had one response variable and one predictor variable. This was one example of the steps involved and hopefully here you can see how there are loads and loads of arrows, lots of pre-processing and post-processing steps, and there's lots of loops in the analysis. There are lots of steps actually in, aside from the main fit in a regression step. And this perhaps will show that there may be some limitations in to how far we can go about generalizing all of

statistics.

So let's move on and let's look instead at the linear regression eBook that we have constructed. As you will see at the top it contains six pages, and as we can see here the first page is visible and inputs are required. Basically, for a linear regression the user will need to choose a dataset and then their choice of a response variable and one predictor variable. Once this is done the eBook will be run and the other pages will be populated with lots of outputs. If we move on to page two, as you can see here page two contain some basic summary statistics and a histogram for the response variable. In this case the data set we're using are the weights of new-born rats, and these are the response variable Y36 is the weights after 36 days. Similarly, page three contains similar information for the predictor variable, in this case the predictor variable chosen was the weights after eight days. Moving on to page four, we now have some information looking at the two variables, the weights at 36 and eight days together, and here we see at the top correlations, both the Pearson correlation and the Spearman correlation and the eBook tells us whether or not they're significant. Then there is a graph that plots not only a linear regression but also a constant, a quadratic, and cubic relationship. With the SAA then informing the user that in this case the linear regression is the best of these options. If we move on to page 5, this contains the estimates for the regression in a tabular form with stars to say where things are significant, along with again a plot of the regression with the data points shown. Finally, on page 6 we have some residual plots because when fitting a regression, we should always look at how well the model fits. So here firstly we see a simple plot of the residuals to look to see if they are normally distributed which is kind of an assumption we're making in this model, and whether there are any outliers. Here we see that in fact there were no outliers and the SAA finds the residuals to be reasonably symmetric. Further down the page our additional residual plots on page 6 here is a quantile plot of the residuals, and a plot of the residuals against the fitted values. In both cases the SAA then gives instructions on what to look for, but this shows that figures are a bit harder for it to interpret so it doesn't interpret the plots itself. A linear regression is a fairly straightforward model, but what about extensions where we have more predictors or predictors of different types? We might look at each predictore individually before trying to find a best model. This would be a much larger SAA and finding a best model may involve some sort of step wise procedure to select between competing models. We have created such an SAA and some things we've found are easy to extend - residual plots are pretty similar if you've got one

predictor or more than one, whereas prediction plots are somewhat less straightforward. You have to think what to do about the other predictors when you construct your prediction plot. So here we show a page from this more general SAA. Here we're illustrating what we would call the univariate modelling of each predictor on its own to assess its relationship with the response. And the table you see here illustrates this. Here is another page which shows a sequence of models fitted in a stepwise fashion, so you can see I think probably about three models behind me, and we're moving from one model to the next, and the significance of the individual predictors has been used to decide on the order we put the predictors in to the model, and where we go for the stepwise procedure. We can see the explanation for what the SAA is doing is concluded for each step.

We have produced a far wider selection of SAAs than we have covered in these slides. Extensions include are the response types, for example binary responses, or counts. And also, multilevel models and models with lots of different sources of variation, and we've also included MCMC and Bayesian methods as well as the standard likelihood methods you see here.

This finishes my presentation and hopefully you have learned something about both eBooks in Stat-JR, and statistical analysis assistants. If you're interested in more details, then please follow the web link that you see on this slide. And thank you once again for listening.