BlogPost #23: CODING CATEGORICAL DATA & USING STUDENT'S T-TEST IN SPSS

Coding categorical data into SPSS

SPSS is a powerful application and is a well-respected analytical tool.

Although the datasets generated by the responses to my eQNR have been captured into Excel – not the least because the output from the PHP form-based QNR has arrived in e-mail format with the data attached as a .csv file – the master spreadsheet that now holds the complete datapool from the 166 good datasets received has been transferred into SPSS for a more comprehensive statistical analysis. Each of the 166 datasets holds 65 data items which means that altogether I have over 10,000 bits of data to explore, understand and try to gain meaning from.

The initial summary analysis has been conducted in Excel as the tools available have been sufficient for this purpose. The results are available on the *Data Summary* page in the ‘ENQUIRY‘ section of the project’s main webpages. However I now need to scrutinize the data in more detail, particularly to explore the inter-relatedness of not only the 8 scales that my QNR is measuring, but also the interactions between individual scale items and more ‘core’ variables of Academic Behavioural Confidence (ABC) and Dyslexia Index (Dx)

At face value, the summary analysis is pointing towards a meaningful difference between the ABC of students with known dyslexia (n = 43) and those in the comparable but albeit small research subgroup of non-dyslexic students who are presenting a dyslexic profile (n = 17). However this early conclusion relies on a very coarse analysis looking at differences between variables’ mean values (using the independent samples t-test), and effect size (using Cohen’s ‘d’).

By transferring the datapool into SPSS I have a means to look more closely at influential factors to try to determine which ones are the most significant, which ones may be irrelevant and indeed which ones conflate or confuse the results and outputs.

SPSS needs to be carefully set up with particular attention being paid to the definition of the variables, particularly to those that are categorical rather than scale variables. In the case of my datasets, this includes variables such as gender, student status and student residency but especially in the definitions of my research groups. Setting up these category names carefully is essential for my research process to progress.

BlogPost #23: CODING CATEGORICAL DATA & USING STUDENT’S T-TEST IN SPSS

Essentially the principle differentiation that my eQNR established at the outset is whether the student respondent has a dyslexic learning difference or not. The former are categorized as Research Group: DI (n = 68), the latter as Research Group: ND (n = 98). From this, I have been able to further sub-divide each of these two principle research groups into subgroups in order to tackle my research questions exploring differences in Academic Behavioural Confidence between students with dyslexia, students with no indication of dyslexia and students with potentially non-identified dyslexia.

This StudyBlog post is written mostly to record the process I have used to code my research groups into SPSS so that I can easily differentiate between them and hence use the comprehensive analytical features of SPSS to explore the data in meaningful ways.

As such, I have created TWO variables devoted to the categories of research groups that I have established, one for the coarse differentiation between research groups DI and ND which I have labelled as variable name: ‘RESEARCH GROUP‘, and the other to enable me to differentiate responses within each of these principle research groups according to additional criteria which I have labelled as variable name: ‘RESEARCH SUBGROUP‘. Because this data is categorical in nature, I have assigned ‘code’ values to each dataset to indicate firstly which principle research group the dataset comes from and secondly which research subgroup it subsequently falls into:

RESEARCH GROUP	SPSS CATEGORY CODE	RESEARCH SUBGROUP	SPSS CATEGORY CODE	CRITERIA
ND	**0**	**ND-400**	**10**	students in research group ND who present a Dyslexia Index Dx < 400
		NDx400	**90**	students in research group ND who present a Dyslexia Index 400 < Dx < 600
		DNI	**20**	students in research group ND who present a Dyslexia Index Dx > 600 – this is the group of greatest interest
DI	**1**	**DI-600**	**21**	students in research group DI who present a Dyslexia Index Dx > 600 – this is essentially my ‘control’ group
		DIx600	**91**	students in research group DI who present a Dyslexia Index Dx < 600

Clearly the research subgroups I'm most interested in are RG: DNI and RG: DI-600 as my baseline hypothesis is that students with unknown dyslexia present a higher Academic Behavioural Confidence that students with known dyslexia.

We'll see what transpires!

Using Student's independent-samples t-test in SPSS

Laerd Statistics is an invaluable guide and tutorial about many things statistical but particularly as it provides excellent support for working through statistical processes in SPSS. I am using it as wise counsel for tackling the statistical analysis of my data.

Much of the analysis is looking for significant differences in variables between my research groups. The classical tool to use for exploring this is Student's t-test for differences between means.

I am guided by the 6 primary assumptions that need to be fulfilled in order that the results from the t-test analysis can have proper meaning:

- Dependent variables are continuous in nature: mine are, since I've designed and developed my eQNR to gather data on scale items through the continuous slider devices that respondents adjust to provide their responses to the scale item statements;
- Independent variables are categorical and that two are used in each t-test analysis: the principle categorical variables that will be driving my t-test analysis are the research groups and subgroups;
- Observations are independent: individual respondents' datasets only appear in one group or subgroup;
- Dependent variables should have no significant outliers: it is likely that many of my 80 dependent variables, albeit assembled into 8 scales, will have significant outliers. I can determine these using box-plots in SPSS or they may be easier to spot through correlation scatter diagrams which I can build using HTML5 scripts (as elsewhere on the project webpages);

- Dependent variables should be approximately normally distributed: SPSS recommends the Shapiro-Wilk test for normality; I can establish this through the 'Explore' feature in the

'Descriptive Statistics' section of the 'Analyze' group of functions, where I can also read off values for skewness and kurtosis (which, I have learned, is a measure of the combined sizes of the two tails in the supposed normal distribution);

- The variance is equal in each group of each independent variable: this is the 'homogeneity of variances' and is referring to the POPULATION variances. I am guided that when sample sizes are similar, the t-test copes with this although when sample sizes are quite disparate the t-test goes off the rails a bit. If this happens, I'll find out what to do to fix it.

So the first step is to run the SPSS outputs to explore the extent to which these 6 assumptions are satisfied in order for me to proceed with t-test analysis at a more detailed level.