2.5. Testing whether a Distribution is Normal

2.5.1. *Running the Analysis*

It is all very well to look at histograms, but they tell us little about whether a distribution is close enough to normality to be useful. Looking at histograms is subjective and open to abuse (I can imagine researchers sitting looking at a completely distorted distribution and saying 'yep, well Bob, that looks normal to me', and Bob replying 'yep, sure does'). What is needed is an objective test to decide whether or not a distribution is normal. Fortunately, such tests exist: the Kolmogorov-Smirnov and Shapiro-Wilk tests. These tests compare the set of scores in the sample to a normally distributed set of scores with the same mean and standard deviation. If the test is non-significant ($p > 0.05$) it tells us that the distribution of the sample is not significantly different from a normal distribution (i.e. it is probably normal). If, however, the test is significant ($p < 0.05$) then the distribution in question is significantly different from a normal distribution (i.e. it is non-normal). These tests are great: in one easy procedure they tell us whether our scores are normally distributed (nice!).

The Kolmogorov-Smirnov (K-S from now on) test can be accessed through the *explore* command (**Analyze⇒Descriptive Statistics⇒Explore...**).[2] Figure 2.6 shows the dialog boxes for the *explore* command. First, enter any variables of interest in the box labelled *Dependent List* by highlighting them on the left-hand side and transferring them by clicking on [▸]. For this example, just select the exam scores and numeracy scores. It is also possible to select a factor (or grouping variable) by which to split the output (so, if you select **uni** and transfer it to the box labelled *Factor List*, SPSS will produce exploratory analysis for each group—a bit like the *split file* command). If you click on [Statistics...] a dialog box appears, but the default option is fine (it will produce means, standard deviations and so on). The more interesting option for our purposes is accessed by clicking on [Plots...]. In this dialog box select the option ☑ Normality plots with tests, and this will produce both the K-S test and normal Q-Q plots for all of the variables selected. By default, SPSS will produce boxplots (split according to group if a factor has been specified) and stem and leaf diagrams as well. Click on [Continue] to return to the main dialog box and then click [OK] to run the analysis.

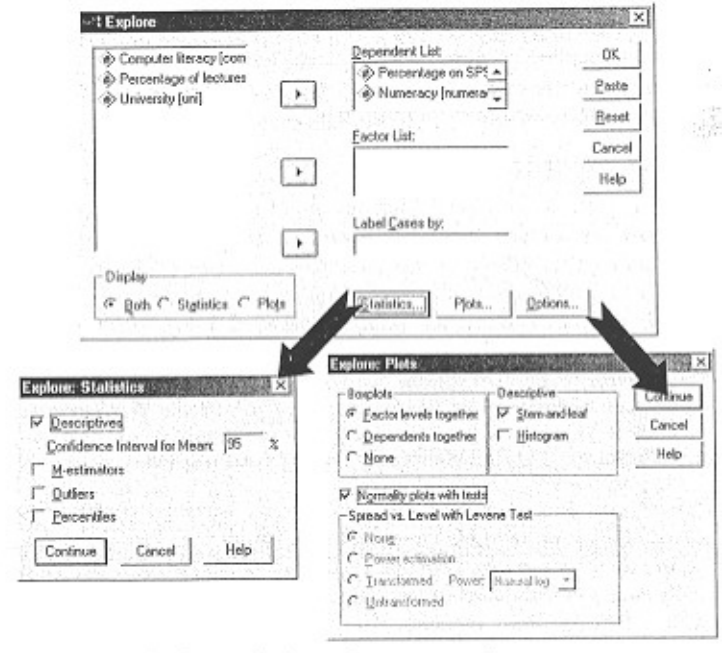

Figure 2.6: Dialog boxes for the *explore* command

2.5.2. *Output*

The first table produced by SPSS contains descriptive statistics (mean etc.) and should have the same values as the tables obtained using the frequencies procedure. The important table is that of the Kolmogorov-Smirnov test. This table includes the test statistic itself, the degrees of freedom (which should equal the sample size) and the significance value of this test. Remember that a significant value (*Sig.* less than 0.05)

[2] This menu path would be **Statistics⇒Summarize⇒Explore...** in version 8.0 and earlier.

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
Percentage on SPSS exam	.102	100	.012
Numeracy	.153	100	.000

a. Lilliefors Significance Correction

indicates a deviation from normality. For both numeracy and SPSS exam, the K-S test is highly significant, indicating that both distributions are not normal. This result is likely to reflect the bimodal distribution found for exam scores, and the positively skewed distribution observed in the numeracy scores. However, these tests confirm that these deviations were *significant*. This finding is important because the histograms tell us only that our sample distributions deviate from normal; they do not tell us whether this deviation is large enough to be important.

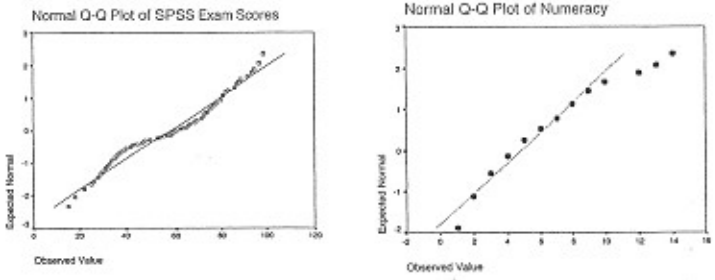

Figure 2.7: Normal Q-Q plots of numeracy and SPSS exam scores

SPSS also produces a normal Q-Q plot for any variables specified (see Figure 2.7). The normal Q-Q chart plots the values you would expect to get if the distribution were normal (expected values) against the values actually seen in the data set (observed values). The expected values are a straight diagonal line, whereas the observed values are plotted as individual points. If the data are normally distributed, then the observed values (the dots on the chart) should fall exactly along the straight line (meaning that the observed values are the same as you would expect to get from a normally distributed data set). Any deviation of the dots from the line represents a deviation from normality. So, if the Q-Q plot looks like a straight line with a wiggly snake wrapped around it then you have some deviation from normality! In both of the variables analysed we already know that the data are not normal, and these plots confirm this observation because the dots deviate substantially from the line. It is noteworthy that the deviation is greater for the numeracy scores, and this is consistent with the higher significance value of this variable on the Kolmogorov-Smirnov test. A deviation from normality such as this

tells us that we cannot use a parametric test, because the assumption of normality is not tenable. In these circumstances we can sometimes turn to non-parametric tests as a means of testing the hypothesis of interest. In the next section we shall look at some of the non-parametric procedures available on SPSS.