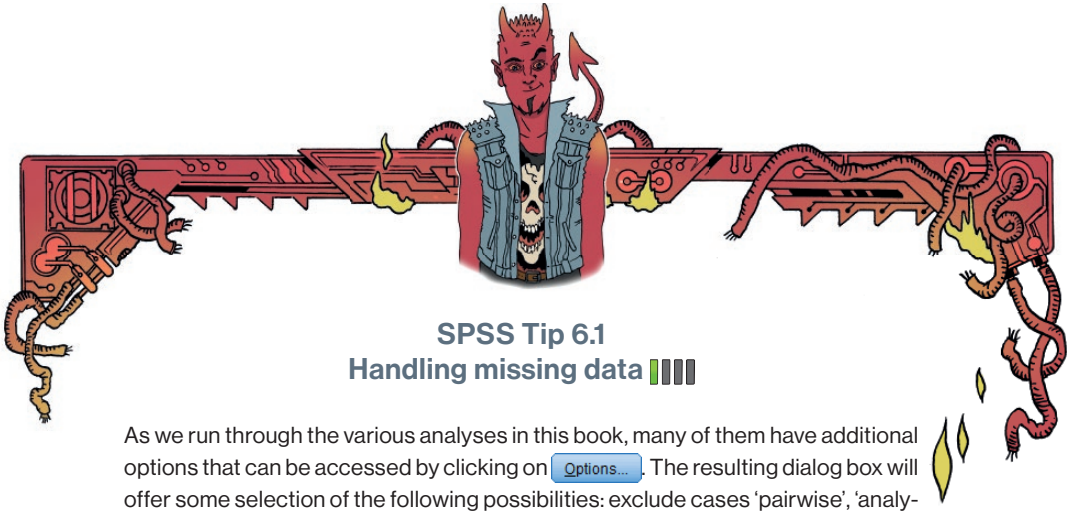

SPSS Tip 6.1
Handling missing data

As we run through the various analyses in this book, many of them have additional options that can be accessed by clicking on Options... . The resulting dialog box will offer some selection of the following possibilities: exclude cases 'pairwise', 'analysis by analysis', or 'listwise', and sometimes 'replace with mean'. Let's imagine we wanted to use our hygiene scores to compare mean scores on days 1 and 2, days 1 and 3, and days 2 and 3. First, we can exclude cases listwise, which means that if a case has a missing value for any variable, then the case is excluded from the whole analysis. So, for example, if we had the hygiene score for a person (let's call her Melody) at the festival on days 1 and 2, but not day 3, then Melody's data will be excluded for all of the comparisons mentioned above. Even though we have her data for days 1 and 2, we won't use them for that comparison – *they would be completely excluded from the analysis*. Another option is to exclude cases on a *pairwise* (a.k.a. *analysis-by-analysis* or *test-by-test*) basis, which means that Melody's data will be excluded only for analyses for which she has missing data: so her data would be used to compare days 1 and 2, but would be excluded for the other comparisons (because we don't have her score on day 3).

Sometimes SPSS will offer to replace the missing score with the average score for this variable and then include that case in the analysis. The problem is that this will likely suppress the true value of the standard deviation (and, more importantly, the standard error). The standard deviation will be suppressed because for any replaced case there will be no difference between the mean and the score, whereas if data had been collected for that missing case there would, almost certainly, have been some difference between the mean and the score. If the sample is large and the number of missing values small then this may not be a serious consideration. However, if there are many missing values this choice is potentially dangerous because smaller standard errors are more likely to lead to significant results that are a product of the data replacement rather than a genuine effect.