Data cleaning

Chris Wild, University of Auckland

In large research projects, much more time may be spent on "cleaning" the data before analysis begins than on analysis itself.

Data collection processes seldom run entirely smoothly. Raw data inevitably arrives "dirty". In all but tiny datasets, there will always be mistakes in the data - values in cells that are wrong - at least until the data has been "cleaned".

Ideally, every entry in a data file would be checked against original sources but this is seldom possible. So cleaning entails checking for values that look suspect. Where possible, suspicious values will be checked back against original sources (the first place where a piece of data got recorded). This may enable us to correct some of them. If we "know" that they are wrong but cannot correct them, we set them to missing. And then there is the middle ground where we do not know whether a suspicious value is right or wrong. That is usually dealt with using some analysis strategy (like analysing with points included and removed to see how much difference it makes to conclusions).

So what are people checking for? It begins with checking that all values of each numeric variable lie within believable (or sometimes allowable) limits, and all values of each categorical variable correspond to what is expected. For example, if for Gender, we are using "Female" and "Male", close variants such as "female", "fem", "F" and "f" would have to be converted to "Female". Otherwise analysis programs treat them as entirely different categories.

How else can we look for values that are suspect? We've already seen what makes points look suspicious in dot plots and scatter plots, for example.

iNZight has a number of tools that are useful during data cleaning. You can find them under **Advanced > Quick Explore**. You may want to try some of them out on one of the NHANES datasets (optional). Start with NHANES-1000.

- With **All 1-variable plots**, each time you click on the plotting window you get a plot from the next variable until you have stepped through all of them.

- **All 1-variable Summaries**, gives a summary for every variable in the dataset.

- **Explore 2-variable Plots** takes a user-selected variable and then steps through plotting every other variable in the data set against it.

Missing data

Just as important as knowing what is wrong, is knowing what is going missing. Failing to keep a close eye on "who is going missing" can lead us to doing some very silly things in analysis. The simplest tools for this job are under **Advanced > Quick Explore > All 1-variable Summaries**, which includes the number of missing values in each variable, and **Advanced > Quick Explore > Missing Values** which gives plots and information about combinations of variables that tend to go missing together. Others are beyond the scope of this course.