Once the data is entered……

Tip! Before you start cleaning your data, save a new copy and work from this.

That way if you make a mistake, you still have the untouched original!

Before analysis can start, we need to verify that the data is clean and reliable.

As the saying goes

RUBBISH IN RUBBISH OUT!

DATA CLEANING IS OFTEN THE MOST TIME CONSUMING PART OF THE DATA ANALYSIS PROCESS. How you do this and what you do depends on whether you have access to a statistical package (e.g. SPSS, Stata and others) or not.

Data management and analysis needs to be viewed as part of the research process itself: everything needs to be documented and replicable by someone else. Statistical packages offer options that are set up to allow easy replication and traceability.

a) **I don't have access to a stats package and I am sending my data off for analysis:**

   You can use Excel to run some preliminary checks on your data

   - Does your dataset meet the criteria outlined in the data entry guidelines manual?
   - Do you have the number of observations you expect?
   - Check the range of variables by using the formulae:
     '= max(cellrange)' and '= min(cellrange)' - are they what you expect?

Note you can do this by sorting the data but you have to be really careful that you don't accidentally sort only 1 column and end up with random numbers……

- Send it off for analysis, but expect some queries before analysis can begin.

b) I have access to a data analysis package and am doing my own analysis:

These are the things you need to do and check, and most statistical packages have ways you can do this relatively easily. We highly recommend using syntax or .do files for all data checking and manipulation, and at the very least log/save your output – if you would like any help with how to do this with SPSS or Stata, please contact us!

- Make sure your data meets criteria outlined in the data entry guidelines manual

- First label all your variables and all the values! (Paola covers this in the SPSS intro manual)

- Look for duplicate cases – remember each row in your dataset should be able to be uniquely identified.

- Are entry/exclusion criteria (if used) actually met?

- Look for missing values – check why they are missing (go back to source data, is this truly missing or was it accidentally not entered?) and recode true missing values to the statistical package specific missing value (e.g. sysmis in SPSS,` .' in Stata)

- Check the distribution of continuous variables (histograms, box plots) - any outliers that need to be queried? Typos? Impossible values? Or just extreme but still plausible? Decide what to do about any anomalous values (hint: don't delete just because they are extreme!) (Paola covers this in SPSS Basics?)

- Check the coding for the categorical values (crosstabs) – do you have the expected number in each group?

- Do the date variables make sense? Calculate any age variables or time intervals – check that you don't have any negative values! (Paola You do the dates in SPSS I think?)

- Do the variables make sense in relation to each other (crosstabs and scatterplots)? E.g. why do I have one male participant listed as having had a previous pregnancy? Or why do I have one person scoring really high on one depression scale but really low on another? (Note that in these examples, the strange values don't look strange unless considered in relation to another variable!)

- Make sure you document any changes that have been made to the dataset as part of this process – syntax or .do files are a great way to do this – we will be happy to show you how!