**BIOS 6621: Data checking - Summary of some steps**

Here are some general steps I use for data examination and checking.

1. Look at each variable individually using the most appropriate tool: histograms or qqplots for continuous variables, tables for categorical, count or ordinal variables.  Histograms are not good for categorical variables, and boxplots are not detailed enough for this kind of examination. Report any unusual things. If investigators supply clinical ranges those can be helpful.
2. Look at missing data for each variable individually.  Many SAS and R procedures will omit cases with any missing variables but may not tell you they have so it's important to know this at the start.  I prefer to deal with it explicitly at the start, usually some combination of imputing missing covariate data and imputing or omitting missing outcome or key exposure data, to create a dataset for analysis.  I prefer not to let the software decide by default how to handle missing data.
3. Look at key variables bivariately as in step 1. For example if the study is to compare CABG and Valve patients I would look at those separately, or if sex is suspected to be a key player I would look by sex, or if variables are known to be closely related graph or table them as x and y.
4. Think about how the data were collected and look specifically at things related to that. For example,

a. Repeat any calculated relationships, e.g. recalculate BMI and compare with listed values, or if components of food are given like grams of fat, carb, protein and alcohol, recalculate the total food weight, or total calories (~9\*fatg+7\*alcg+4\*carbg+4\*protg where g = grams).

b. If data are clustered by hospital look separately by hospital, or if clustered by patient (multiple measurements per patient) look by patient. This includes missing data patterns as well as distribution of values.

c. If data are collected over time, graph or table them over time.

5. Whenever moving data from file types, software packages, formats, etc., be especially careful to check that the number of rows and colums agree, and spot check a few values. This is an easy place for potentially disastrous errors to occur.

6. Also, data examination is a continuous process throughout a project. Each time you do a new kind of analysis you may look at the data in new ways and find new issues.

Finally, remember to examine the data. This may seem obvious, but in the middle of a big confusing analysis while focused on learning the new methods it’s easy to forget.