Basic Dataset Features

Most datasets come in the form of a spreadsheet. You need to identify certain features of the dataset you are using before you can begin statistical analysis. If you have collected your own data the points below also describe how to record it in a spreadsheet for analysis. If your dataset does not seem to be a spreadsheet, go to the page on opening your dataset.

1. Any good quality dataset will come with some information. Generally this is recorded in a separate document to the dataset, although it may be listed at the beginning of a tab or space delimited file. This information should include:
    - What does the data represent?
    - Who collected the data?
    - How was the data collected (the methodology)?
    - Where has the data come from?
    - When was the data collected?
    - Is there more than one spreadsheet? If yes what does each spreadsheet show?

2. Generally, each ROW of data corresponds to a single experimental unit. Depending on the type of study, the experimental units may be individuals, generally referred to as participants, or they may be objects, such as engine parts, or they may be entities, such as hospitals. In addition, these experimental units may be labelled by a name or simply a number. Usually the labels are given in the first column of data. If no experimental unit is provided it may be necessary to analyse the methodology in order to determine what the experimental units are. Furthermore take caution as the experimental unit may not always be listed in the first column, again consult the methodology to determine what the experimental units are and reposition them in the left hand column for ease of use, if this is the case. The number of experimental units is called the sample size of your study or experiment.

3. Generally, each COLUMN of your dataset corresponds to a single variable. Often, the variable name will appear at the top of the column above the entries. However in some datasets this may not hold true and it may be necessary to consult the methodology to determine these variable names, and list them above their relevant columns. Often full names of the variables are not given in the sheet, but rather their names are given by some code. Make sure you understand what each variable stands for and take note of the number of variables listed for each experimental unit. When generating your own data make sure you make it clear what units your variables are measured in and what range of values each variable takes. If you are using existing data, make sure you know what units each of the variables are measured in, and record the range of measurements for each variable. Furthermore it is important to examine if any of your variables have been transformed such as compound variables or derived variables.

- Compound variables: A compound variable is a type of variable which has been generated from combining several pieces of data. For example a researcher may ask 100 participants to fill out a questionnaire composed of 10 yes or no questions. The researcher may not be interested in the candidate's specific responses to every individual question; they may however be only interested in their answer to questions 3 and 5. So they may combine the data from these two questions to generate a single result for example: YES: the candidate answered yes for both questions, NO: the candidate answered no for both questions, MIX: the candidate answered yes for one question and no for one question etc.
- Derived variables: *Derived variables* are variables that are created from other variables using an expression. For example a researcher may record the height of 30 participants in cm, however they are only interested in the value A for each participant which is given by A= $(height - 100)^2$ etc.

4. Each CELL in the spreadsheet represents one piece of data, which is the value of the variable whose name is at the top of the column for the experimental unit whose label or name is at the left side of the row. The cell entry may be a number, or may be some word or code. You need to know what each number, word or code represents. For instance, if the entry is a number, is that number a measurement of something, like height? Sometimes a number is not a measurement, but really a label, such as 1=male, 0=female. If the cells under a variable name are filled with labels (either number labels as in this example, or word labels, such as 2wd, 4wd or Awd for vehicles), note how many different labels there are for your variable. This is the number of LEVELS of the variable. So the variable "gender" has two levels: male and female (or 0 and 1) and the variable "drive type" has three levels: 2wd, 4wd, Awd. When looking at our data cells it is necessary to be aware of data errors such as missing data or miscoded data.
    - Missing data: Look through your dataset and see if every cell is filled, or if some cells are left blank. Blank cells are called missing data. Before you can begin to do anything statistical with your dataset, you will need to deal with missing data. It is important to take note of the proportion of missing data you have for each of your variables.
    - Miscoded data: Miscoded data is just a fancy term for mistakes in data entry. You should look through each column (representing each variable) and identify any cells that don't make sense. For instance, if you have a word in a cell that should have a number, this is a mistake. Similarly, you may have a numerical variable in which most of the entries lie in a certain range, but there is one value far away from the others that may not make any sense. For instance, if you have a variable "height in inches" for a sample of humans, and you saw the value 1000, you would know this was a mistake, as 1000 inches is not a possible human height. If such miscoded data can be corrected, that is the best thing to do, for instance, by going back to original documents where the values were first recorded. However, if it is not possible to determine the correct value for a given entry,

the entry must be simply deleted (just the cell, not the whole column or row!) and treated as missing data.