**Statistical Monitoring Program Instructions.**

**4 a. Digit Preference: Benford’s law and comparison to all other sites**

The function **digit\_preferece** can be used on continuous variables to check if there appears to be any digit preference in each site, i.e. the distribution of leading or tailing digits does not fit Benford’s law or appears different from the data as a whole.

Parameters to give the function:

1. data:

This should be in the form of a data frame with the site number in the first column (must be numeric – if a string is given please recode as numeric) followed by the continuous measurements to check (i.e. columns 2+).

Data frames can be read in with the following code:

**options(stringsAsFactors = FALSE)**

**reg.data<-data.frame(read.table("STUDY12\_REG.txt", row.names=NULL, header=TRUE, sep="\t"))**

(This would read in a text file called *STUDY12\_REG.txt* and store it in the data frame *reg.data*.)

1. trial.name

The name of the trial. This will be used to label the output files. For example:

**trial.name<- “STUDY12”**

1. Digit

This tells the program whether to look at the leading (first) or tailing (last) digit of each value. Set as “leading” or “tailing”.

1. Benford

If this is set as TRUE, then the frequency of the leading digits in each site will be compared to Benford’s distribution. If this is set as FALSE then the frequency of the leading or tailing digits in each site will be compared to the frequency in all of the other sites put together.

Note: Benford’s law only works with leading digits so if Benford is set as TRUE the value of digit will be overridden and taken as “leading” even if “tailing” is entered.

**Calling the function**

Once the program and the parameters above are stored in R’s memory the program can be run using the following command:

**digit\_preferece (data, trial.name, digit, benford)**

Where each parameter is stored as in 1-4

**The output:**

The program outputs 2 text files. The first gives a list of the site numbers and a p-value for the chi squared test comparing them to either Benford’s distribution or to the distribution in all of the other sites put together. If there are not enough observations to perform a chi-squared test, then the p-value will be given as “Not enough observations in at least 1 group”. This file has a name in the form: *“BENFORD\_results\_STUDY12 \_2013-11-06.txt”* where “STUDY12” was given as the trial name and the date is set as the date the program was run.

The second text file gives details of the frequencies for all of the sites. It has a filename in the form:

*“BENFORD\_results\_detail\_STUDY12 \_2013-11-06.txt”*

**Warnings:**

There are no error messages coded into the function. If data is not read in as above, the function may not work as it should, or possibly at all. Please take care when creating the parameters from your data.

Benford’s law should only be used on data where the leading digit can range from 1-9 and variables should not be normally distributed, we found it should not be used in our trial data.

If you wish to test more than one datasheet please give include this in the trial name or save into a different folder or the text files will be over ridden.

The program removes all values which are <0 (the data we used had dummy values set as -9 or -999), if negative values are possible in your dataset please, replace them with the absolute value before running the code.