Data analysis 3
Week 5 practical

In this practical you will be required to input your answers into Blackboard. You should be familiar with the following:

- How to use the NORMDIST function in Excel for calculating probabilities—see http://youtu.be/JZXX231rIlg.
- How to use the NORMINV function in Excel for calculating confidence intervals—see http://youtu.be/5aAlTk-C21c.
- How to use the normcdf function in MATLAB for calculating probabilities—see http://youtu.be/Lo7tkbWIKco.
- How to use the norminv function in MATLAB for calculating confidence intervals—see http://youtu.be/aIi79zNJPTs.

Now start the week 5 Blackboard practical and work through questions 1 to 7.

For questions 8-12 on the *Blackboard practical* you will need to generate some random data by sampling from a distribution. You should be familiar with:

- How to sample randomly from a normal distribution function in Excel—see http://youtu.be/l08MXJs0uz4.
- How to sample randomly from a normal distribution function in MATLAB—see http://youtu.be/p70ScjpZTwA.

Now read and complete the rest of this sheet before attempting questions 8-12

Our example is as follows: *when two consecutive years of 149 measurements of PM10 (pollution aerosols) around Manchester were compared, the reduction in PM loading was 1.0 $\mu g\ m^{-3}$, with a standard deviation of 2.2 $\mu g\ m^{-3}$.*

The following Excel and MATLAB procedures show how to generate normally distributed data and then shift the mean and standard deviation to desired values. This can be useful in a number of situations, but here we just compare the data to a theoretical normal distribution.

We will sample data from a distribution with mean equal to 0 and standard deviation equal to 1 and then shift the mean and standard deviation to desired values. If we multiply every sample value by a constant the standard deviation is multiplied by that constant, and if we then *add a constant to every sample value the same constant is added to the mean but the standard deviation does not change*.
See Section 3.1 for the Excel method and Section 3.2 for the MATLAB method.

3.1 Excel Method

1. Generate a list of sample values from a population with a mean of 0 and a standard deviation of 1. If using Excel 2007, click on Data, then Data Analysis; if using Excel 2003, click on Tools, then Data Analysis. Select Random Number Generation, then click OK. In the dialog box, enter 1 for the number of variables, enter 149 (or any desired sample size) for the number of random values, select a Normal distribution, and use the default parameter values of Mean = 0 and Standard Deviation = 1. Select the default output option of New Worksheet Ply so that the sample data will be put in column A. Click OK.

2. Now find the existing standard deviation of the generated values. Click on any empty cell, say `D1` then type `=STDEV()` and enter the range of cells containing the generated sample data, such as A1:A149 for 149 values in rows 1 through 149 of column A.
3. Transform the sample data so that the standard deviation is changed from the existing standard deviation found in Step 2 to the desired standard deviation (what is the desired standard deviation and why?). To accomplish this, multiply each generated sample value by the desired standard deviation, and divide each result by the existing standard deviation (calculated above). e.g.:
   (a) In cell B1, enter this expression:

   ```
   =A1 *2.2 / $D$1
   ```

   where the 2.2 is the desired standard deviation.
   (b) Press the Enter key, then click and drag the lower right corner of cell B1 down to the same number of rows as in column A.
4. Column B now should contain data with the desired standard deviation. Now transform column B so that the mean is changed to the desired mean. e.g.:
   (a) Click on any empty cell, say `D2`, and find the mean of the values in column B by using `=average()` and enter the range of values for column B, such as B1:B149. Record the existing mean.
   (b) Click on cell C1 and enter this expression:

   ```
   =B1 + 1.0- ($D$2)
   ```

   where, the 1.0 above is the desired mean and `$D$2` is the actual mean found above.
   (c) Press the Enter key, then click and drag the lower right corner of cell C1 down to the same number of rows as in column A.
5. Column C should now contain the desired number of sample values having the desired mean and the desired standard deviation. Verify this by using the Descriptive Statistics module found in the Data Analysis menu.

To check whether the data are indeed normally distributed, complete the following (you may want to watch `http://youtu.be/l08MXJs0uz4` again):

1. Create a histogram of the data, divide the counts in each bin by the bin width and divide again by the sum of all bins.
2. On the same plot, use `NORMDIST` with cumulative flag=`0` to plot a normal distibution with a mean of 1.0 and standard deviation equal to 2.2.
3. Do a similar comparison for the cumulative distribution (i.e. compare the cumulative frequency to the cumulative probability using `NORMDIST` with cumulative flag=1).

3.2 MATLAB method

1. Generate 149 normally distributed random numbers with a mean of 0 and a standard deviation of 1:

```
% Note rand(149,1) generates a column of 149 random numbers between 0 and 1
% norminv is therefore sampled with a random probability
dat=norminv(rand(149,1),0,1);
```

2. Transform these variables so that they have a standard deviation of 2.2

```
% Multiply by the desired standard deviation and divide by the actual sd.
dat2=dat.*2.2./std(dat);
```

   Note the `.` before the `*` and `/`. The reason for this is it allows *all values* in the table (or array) to be multiplied or divided. `dat2`, now contains the 149 variables with a standard deviation of 2.2.

3. Transform `dat2` so that it has a mean of 1.0

```
% Subtract the actual mean and add the desired mean
dat3=dat2+1.0-mean(dat2);
```

4. `dat3` will now have 149 values that have a mean of 1.0 and standard deviation of 2.2. Check this by calculating summary statistics.

To check whether the data are indeed normally distributed, complete the following:

1. Do a normalised histogram:

```
% N is frequency, X is bin centre
[N,X]=hist(dat3);
% Note: 'diff' calculates the difference between values in a list.
bin_spacing=diff(X);
% This is a plot of the normalized frequency density
% i.e. N divided by sum(N) and divided by the bin width
plot(X,N./(sum(N)./bin_spacing(1)),'k')
```

2. Compare to the theoretical normal distribution:

```
% hold on; holds the plot so that we can plot another graph on top
hold on;
% Plot the theoretical normal prob. density.
plot(X,normpdf(X,1.0,2.2),'r');
```

3. Compare the cumulative distribution to the theoretical cumulative distribution:

```
figure;
% Cumulative frequency: cumsum creates a cumulative sum of data
plot(X,cumsum(N)./sum(N),'k');hold on;
% theoretical cumulative distribution:
plot(X,normcdf(X,1.0,2.2),'r');
```

Save your spreadsheet or MATLAB workspace (e.g. type: `save <filename>`).