Data analysis 1
Week 7 practical

There is one exercise to do this week:

1. the *'Week 7 project'* in the computing cluster. It is an exercise looking at what determines whether a data point is an *outlier*.
2. MATLABers can learn how to plot out the landgrav data using the code on the page opposite. Either enter each line into the command window, or put the text into a text file and call it '`landgrav.m`', save it on your path and type '`landgrav`' into the command window.

This will be done in the practical session on Monday. You may wish to do the exercise in a spreadsheet or MATLAB before opening the test.
In order to do this practical you should be familiar with the following:

- How to calculate various descriptive statistics and z-scores using Excel—see `http://youtu.be/Q7RuUnLpXjM`.
- If you want to use MATLAB then see how to calculate descriptive statistics and z-scores using MATLAB—`http://youtu.be/c4yPS1kgcko`.
- To read in the `landgrav_all.csv` file in MATLAB type `dat=csvread('landgrav_all.csv',2,0);` as described on lines 1-6 on the opposite page.

1. Calculate the mean and the standard deviation of the duration times of Eruptions of the Old Faithful geyser (based on the OLDFAITH.xls dataset—see Course content→Spreadsheets and data).
   (a) What is the difference between a duration time of 110 and the mean of the column of data?
   (b) How many standard deviations is that (the difference found in part (a)?);
   (c) Convert the duration of 110 sec to a z-score;
   (d) If we consider *'usual'* duration times to be those that convert to z-scores between -2 and 2, is the duration time 110 usual or unusual?
2. Calculate the mean and the standard deviation of the Bouguer anomally using the data in the csv file `landgrav_all.csv`. Convert the following Bouguer anomallies to z-scores (2 d.p.s) and say whether they are usual or unusual
   (a) 40 mGal;
   (b) -45 mGal.
3. Refer to the VOLTAGE.xls dataset, which includes measured voltage levels from a US home, a generator, and an uninterruptible power supply. Generate histograms for all three columns, with a range of 122 to 125.2V (bin width = 0.2V), and calculate statistics to compare the three sets of data. Are there any outliers? (hint: calculate the z-scores). Do all three power sources appear to provide electricity with properties that are basically the same?

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

1.1 Neat stuff in MATLAB

You can plot out the `landgrav_all.csv` data set using MATLAB. The function to plot out 3-d scattered data is `scatter(x,y,siz,z,'filled')` (where `siz` is the size of the dots on the image), which plots the x and y. locations of the measurements and colours them by the measurement value, z. Try typing `help scatter` for more information.

The code excerpt shows how to do use `scatter` to plot out the landgrav data.

Note, if you get an error that MATLAB cannot find the coast variable or file you will need to download it. It is available on Blackboard: Course content-¿ Semester 1-¿ Matlab resources-¿ coast.mat. Download it and put in your working directory. You could also just comment lines 21, 22, 34 and 46.

```
1  % Note there are two headerlines in the file
2  % (hence the 2 for the second argument)
3  % The columns are: latitude, longitude, elevation (m),
4  % observed gravity -980000 (mGal), Bouguer anomally (mGal),
5  % Bouguer density
6  dat=csvread('landgrav_all.csv',2,0);
7
8  % Locate values that equal 999999 and set to 'NaN'
9  dat(find(dat(:)==999999))=NaN;
10
11 % Open a figure to write the plot to, using the 'openGL' renderer for speed
12 figure('name','Observed gravity','renderer','openGL')
13 scatter(dat(:,2),dat(:,1),2,dat(:,4),'filled');
14 h=colorbar;
15 ylabel(h,'Observed gravity -980000 (mGal)')
16 hold on;
17
18 % 'coast' is a low resolution coastline dataset that is part of matlab.
19 % it contains the arrays 'long' and 'lat', which are the locations of the
20 % coastlines.
21 load coast
22 plot(long(:,1),lat(:,1));
23 axis([-11 2 49 60])
24 xlabel('longitude (^\circ)');
25 ylabel('latitude (^\circ)');
26 grid on;
27
28 figure('name','Bouguer anomaly','renderer','openGL')
29 scatter(dat(:,2),dat(:,1),2,dat(:,5),'filled');
30 h=colorbar;
31 ylabel(h,'Bouguer anomaly (mGal)')
32 hold on;
33
34 plot(long(:,1),lat(:,1));
35 axis([-11 2 49 60])
36 xlabel('longitude (^\circ)');
37 ylabel('latitude (^\circ)');
38 grid on;
39
40 figure('name','Elevation','renderer','openGL')
41 scatter(dat(:,2),dat(:,1),2,dat(:,3),'filled');
42 h=colorbar;
43 ylabel(h,'Elevation (m)')
44 hold on;
45
46 plot(long(:,1),lat(:,1));
47 axis([-11 2 49 60])
48 xlabel('longitude (^\circ)');
```

```
49 ylabel('latitude (^\circ)');
50 grid on;
```