Data analysis 3
Week 9 practical

In this practical you will be required to input your answers into Blackboard. You may wish to do the exercise in the spreadsheet first though before opening the test. *It is designed to give you an appreciation why the hypothesis testing method works.*

In this practical you will use Excel or MATLAB for hypothesis testing. You should be familiar with the following:

- how to do a hypothesis test about a mean in Excel—`http://youtu.be/wb1q4FWOwTY`
- and the same in MATLAB—`http://youtu.be/o7yPcDnrXws`
- how to conduct the sampling proceedure (Section 3.1) in Excel—`http://youtu.be/4Vcdn9gMWWA`
- and the same (Section 3.2) in MATLAB—`http://youtu.be/mwDi8tTkrW0`

This practical considers two ways to *test a hypothesis*, through theory and sampling.

1. Open the file RUBBISH.xls (in either Excel or Matlab—e.g. using `xlsread`), which contains weights in kg of household rubbish discarded each week. Use the sample weights of discarded plastic (PLAS or column 4) to *test the claim that the mean weight of plastic discarded in a week is less than 1.0 kg.* Use a 0.05 significance level ($\alpha = 0.05$).
   Is the mean weight of plastic than 1.0 kg? if so *is it significantly below 1.0 kg?*
2. Estimate how often we would calculate a sample mean less than 0.95 kg, if we were to sample from a population with $\mu = 1.0$ kg and $\sigma = 0.4843$ kg. To do this do the following:
   - Calculate the value of t for $\mu = 1.0$, $\sigma = 0.4843$ and $\bar{x} = 0.95$:
   $$t = \frac{\bar{x} - \mu}{s/\sqrt{N}}$$
   - Use `tdist` in Excel or `tcdf` in MATLAB, with $N-1$ degrees of freedom and the above value of t to calculate the p-value (probability of sampling a value of t, less than the value of t being considered.):
     Excel: `=TDIST(ABS((0.95-1)/(0.4843/SQRT(62))),61,1)` (note we have used '1' for the third input, why?)
     MATLAB: `tcdf((0.95-1)./(0.4843./sqrt(62)),61)`
   - Your answer should be about 0.2 or so. Make sure you understand how these functions work as they will be used in later practicals.
3. Now, lets look into why we actually use these methods. We will generate some data by *sampling* from a normal distribution with $\mu = 1.0\,kg$ and $\sigma = 0.4843$.

- Using the procedure described, in Section 3.1 (Excel users) or 3.2 (MATLAB users), below generate 100 samples of 62 weights from a normally distributed population having the assumed mean of 1.0 kg (as in the null hypothesis, i.e. $H_0 : \mu = 1.0\,kg$). Use a standard deviation, s, of 0.4843, like the in sample data.

4. *Testing the hypothesis*: Based on the sample means that are found from the simulated samples, determine whether a sample mean of 0.95, is 'unusual' or can easily occur by chance (i.e. count them and calculate a fraction that are less than 0.95)
   - How many of the sample means are less than or equal to 0.95 kg? What fraction is less than or equal to 0.95 kg?
   - What do the simulated results suggest about the claim that the mean weight of discarded plastic is less than 1.0 kg? How does your answer compare to the p-value you calculated above?

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

3.1 Excel procedure for generating 100 simulated samples

- If using Excel 2007, click on Data, then select Data Analysis; if using Excel 2003, click on Tools, then select Data Analysis.
- In the Data Analysis window, select Random Number Generation, click OK.
- In the dialog box, enter 100 for the number of variables (or samples), enter a sample size of 62 (as in the sample of weights of plastic), select the distribution option of Normal, enter a mean of 1 (as assumed) and enter a standard deviation of 0.4843 kg, then click OK.
- Find the sample mean of the first generated sample by going to cell A63 and using the Excel command `=AVERAGE(A1:A62)`. Click and drag the lower right corner of cell A63 to the right so that all of the 100 sample means are found. Inspect the 100 sample means to determine how many are less than 0.95 kg. Return back to point 4 above to complete the practical.

3.2 MATLAB procedure for generating 100 simulated samples

- Use the `rand()` function to sample from a normal distribution using `norminv`.

```
% generate a table of 100 by 62 random numbers.
dat_rand=rand(100,62);

% generate 100 lots of 62 normally distributed random samples.
dat=norminv(dat_rand,1,0.4843);
```

- Find the sample means of all 100 samples

```
sample_means=mean(dat,2);
```

  which takes the mean over the 2nd dimension.
- Inspect the 100 sample means to determine how many are less than 0.95 kg:

```
number_less_than=length(find(sample_means < 0.95))
```

which finds the length of all rows that have a sample mean less than 0.95.

- Return to point 4 to complete the practical.

3.3 Fancy things in MATLAB: Function m-file

You can save commands in a text file so that you can run common tasks easily. In this example we will save a *function m-file* that allows us to test the hypothesis of whether a sample is drawn from a distribution with a specified population mean. First, type `edit` on the MATLAB command window. In the editor that opens, type the following:

```
function [accept,pvalue]=hypothesis01(data,alpha,pop_mean,tails)
% Test whether a sample is drawn from a distribution with a population mean
% equal to pop_mean at a significance, alpha. Tails is either 1 or 2 depending
% on whether a one or two tailed test is needed.
% It returns which hypothesis is accepted in 'accept' and the 'pvalue'

% Tails must be 1 or 2
if(tails ~= 1 & tails ~=2)
        disp('tails must be 1 or 2');
        return;
end
xbar=mean(data); % calculate the mean
std1=std(data); % calculate the standard deviation
N=length(data); % sample size
t=abs((xbar-pop_mean)./(std1./sqrt(N))); % t-statistic - absolute value
% adjust the probability depending on how many tails.
if(tails==1)
        alpha1=alpha;
elseif(tails==2)
        alpha1=alpha./2;
end
tcrit=abs(tinv(alpha1,N-1)); % critical value of t
pvalue=tcdf(-t,N-1); % p-value

% If t is less than the critical value, accept the null, etc.
if(t<=tcrit)
        accept='Null';
else
        accept='Alternate';
end
```

Save it as `hypothesis01.m` on your file path and run it for a 1 tailed test at a significance level = 0.05 by typing:

```
[accept,pvalue]=hypothesis01(data,0.05,1,1)
```

on the MATLAB commandline, where `data` contains your data. The function will tell you whether to accept the null or alternate hypothesis and also give a p-value.