8. Statistics

1. The file `atg1year.mat` (see the Exercise 2 on convolution) contains 1 year of temperature data from the roof of ATG sampled every 5 minutes.

   (a) Calculate the mean (mean) and standard deviation (std) of this sample.

   (b) Plot a histogram (hist) of temperatures with a bin spacing of 1.0°C and overlay this with the appropriately normalized (multiply by the product of number of temperature data and the width of the histogram bins) normal distribution. Do the temperatures approximate a normal distribution.

   (c) The commands

   ```matlab
   rand('state',sum(100*clock)); %Seeds the random number generator
   meantemp=zeros(10000,1)
   for j=1:10000;
       i=floor(rand(100,1)*length(temp))+1;
       t=temp(i);
       meantemp(j)=mean(t);
   end
   ```

   will generate a vector `meantemp` of 10000 estimates of the mean temperature each of which is determined from 100 randomly selected samples. Plot a histogram of the mean temperatures and with a bin spacing of 0.1°C and overlay this with the appropriately normalized normal distribution. Do the mean temperatures approximate a normal distribution as the central limit theorem predicts.

2. A geochemist measures the concentration of a pollutant in 10 lakes in the Puget Sound region (sample 1) and 5 lakes in Eastern Washington (sample 2). In arbitrary units the results are

   \[ X_1 = [1.2 \ 1.1 \ 1.9 \ 1.6 \ 2.2 \ 2.0 \ 1.4 \ 1.7 \ 1.4 \ 1.5] \]

   \[ X_2 = [0.9 \ 1.2 \ 0.8 \ 1.0 \ 1.6] \]

   Does this data show statistically that the lakes in the Puget Sound region are more populated to the 95% confidence level? To answer this question, use the \( t \) statistic to determine the error bounds on the difference between these two means at the 95% confidence level (equation 12-13). If a difference of zero lies within these confidence bounds you cannot be sure the samples are from populations with different means at this confidence level.
3. A hydrology graduate working in an arid area measures the water table depth, \( z \) in 15 boreholes. The student likes inverse theory and fits the results to a cubic model of the form

\[
z = c_1 + c_2x + c_3y + c_4x^2 + c_5xy + c_6y^2 + c_7x^3 + c_8x^2y + c_9xy^2 + c_{10}y^3
\]  

(1)

and then estimates the root mean sample misfit according to

\[
\overline{\Delta z} = \sqrt{\frac{\sum_{i=1}^{n} (z_{\text{observed}} - z_{\text{model}})^2}{n-p}} = 2.3 \text{ m}
\]  

(2)

The advisor is skeptical of such a complex model and suggests fitting the data with a planar surface of the form

\[
z = c_1 + c_2x + c_3y
\]  

(3)

The student does this and obtains an estimate of \( \overline{\Delta z} = 6.9 \text{ m} \) using the expression.

Does the more complex model provide a better fit that is statistically significant at the 95% confidence level? To answer this you first have to recognize that the \( \overline{\Delta z} \) is not an estimate of the sample standard deviation for the model. To obtain these estimates, you have to calculate

\[
s^2 = \frac{1}{n-p} \overline{\Delta z}^2
\]  

(4)

where \( n \) is the number of observation (\( n = 15 \)) and \( p \) is number of degrees of freedom in the model (\( p_1 = 10 \) in model 1 and \( p_2 = 3 \) in model 2). After obtaining \( s^2 \) for the two models, you can then take the ratio of the two variances and apply an F-test (equation 12-22) with degrees of freedom \( n - p_1 \) and \( n - p_2 \) (it is important to get the degrees of freedom in the correct order - \( \nu_1 \) is the degrees of freedom for the numerator and \( \nu_2 \) for the denominator).