ADVANCED ENVIRONMENTAL DATA ANALYSIS
HOMEWORK #2

1) Use MATLAB random number generator to produce an N=10000 realizations of a random variable with the following probability density function $P_x(X)$: Gaussian $\left(\mu_1 = 0; \sigma^2 = 2\right)$, Lognormal $\left(\mu_1 = 4; \sigma^2 = 2\right)$, Uniform [-1 1] and Triangular [5 7]. Plot the $P_x(X)$. Check your results by superimposing a plot of the analytical shape of $P_x(X)$.

NOTE: The analytical shape of the uniform and triangular $P_x(X)$ can be obtained using the heavyside function $H(X-a)$. For example a uniform distribution between $a \leq X \leq b$ is expressed as:

$$P_x(X) = \frac{H(X-a) - H(X-b)}{b-a}$$

On the class website I have posted a MATLAB `heavyside` function `utl_H.m`.
To evaluate $H(X-a)$ from MATLAB where X is the x-axis:

```
>> y = H( X, -a );
```

To evaluate $H(-X+a)$:

```
>> y = H( -X, a );
```

2) (a) Produce $M = 30$ random variable $n_m(t)$ each with uniform $P_n(N)$ [-1 1] and length 10000.

a) Show in MATLAB that the variable $y(t) = \sum_{m=1}^{M} n_m(t)$ has a Gaussian $P_y(Y)$ (as predicted by the central limit theorem).

b) Now assume $P_n(N)$ is Gaussian $\left(\mu_1 = 0; \sigma^2 = 1\right)$. Compute in MATLAB $P_z(Z)$ of the variable $z(t) = \sum_{m=1}^{M} n_m^2(t)$. Verify that $P_z(Z)$ is a *chi-square distribution* by superimposing a plot of the analytical shape.

3) Now assume that you have measured a random variable $x(t)$ along with its statistics. You find that $P_x(X)$ is a uniform distribution between [-1 1] and that each value of $x(t_i)$ is independent of each other (in other words $x(t)$ has no autocorrelation in time). What are the statistics of the variable $y(t)$ given by the following equation:

$$y(t) = \frac{dx}{dt}$$

HINT: Remember that you can discretize the equation $y(t_i) = \frac{x(t_{i+1}) - x(t_i)}{dt}$ and that $x(t_{i+1})$ and $x(t_i)$ can be considered independent random variables with known $P_x(X) = P_{x_i}(X_i) = P_{x_{i+1}}(X_{i+1})$. If you have trouble with the math, find the $P_y(Y)$ numerically using MATLAB. Once you know the shape of it try to see if you can derive the same result analytically.

4) In class we have quantified the linear relationship between two variables x' and y' with $\langle x' \rangle = \langle y' \rangle = 0$, using the correlation coefficient. If a linear relationship exists, than one could write a *statistical optimal linear estimator* by imposing the following model (see chap. 3 Davis on class website):

$$\hat{y}' = \alpha x' \quad \text{where} \quad \alpha = \frac{\langle x'y' \rangle}{\langle x'^2 \rangle}$$

This fundamental result can be obtained using the least square approach and minimizing the expected value $\langle J \rangle$ of the cost function $J = \sum_{i=1}^{M} n_i^2 = \sum_{i=1}^{M} \left(y_i' - \alpha x_i' \right)^2$ with respect to α, where $n = y' - \hat{y}' = y' - \alpha x'$ is the residual between the model estimate $\hat{y}'$ and the observed value y'. Show analytically or numerically that when the model errors n_i are normally distributed, the estimate $\hat{y}' = \alpha x'$ is equivalent to the estimate obtained using the conditional probability density function

$$\hat{y}' = \alpha x' = \int_{-\infty}^{+\infty} Y' \; P_{y|x}\left(Y' \,|\, X' = x' \right) dY'$$

Now show that this is no longer true when the errors are lognormally distributed.

5) Download the MATLAB function `hw2_generate_Y.m` and `bs_rand.m`. This function will generate 4 different sets of observations of a variable y along with the standard deviation of the measurement error n.

a) Using the least square approach and *chi-square* statistics, test which of the two models $y(t) = a + bt + n(t)$ and $y(t) = a + bt + ct^2 + n(t)$ is consistent with each of the observational data sets. Compute the uncertainty of the model parameters.

In MATLAB to generate a realization for set number one

```
>> set_number = 1;
>> [y,t,n_std] = hw2_generate_Y(set_number);
```

You can generate many realizations of data set number one by just calling this function over and over again. You will need to do this to test the *chi-square* statistics of the cost function $J = \sum_{i=1}^{M} n(t_i)^2$.

You will do the same test for observational set number 2, 3 and 4.

```
>> set_number = 2;
>> [y,t,n_std] = hw2_generate_Y(set_number);
```

b) Explain why the expected value of J is inferred by *chi-square* statistics.

$$\langle J \rangle = \int_{-\infty}^{+\infty} J \; P_J(J) dJ \quad \text{where} \quad P_J(J) = chi - square(r)$$

and r is the degrees of freedom.

OPTIONAL STUDY PROBLEMS- NOT REQUIRED

1) Prove that the moments of a random variable x with a Gaussian $P_x(X)$ can be determined only by the mean m_1 and the variance σ^2. Compute also the value of m_3.

$$P_x(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(X-m_1)^2}{2\sigma^2}\right] dX$$

2) Random variable X has $P_x(X) = \left\langle \delta\left(X^2 - \alpha^2\right)\right\rangle$. Compute the probability density function of the two new random variables Y and Z defined as:

$$Y = \frac{1}{3}(X_1 + X_2)$$

$$Z = \frac{X_1 X_2}{2\alpha}$$

where X_n are assumed to be independent realizations of the random variable X.