MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Civil and Environmental Engineering

1.017/1.010 Computing and Data Analysis for Environmental Applications / Uncertainty in Engineering

---

Problem Set 4 (Solutions provided at end of each problem)
Due: Thursday, Oct. 16, 2003

Please turn in a hard copy of your MATLAB program as well as all printed outputs (tables, plots, etc.) required to solve the problem.

Problem 1

For Problem 1 please turn in your solution to the Bangladesh arsenic problem discussed in class on Thursday Oct. 9. Here is the problem statement:

Go to the British Geological Survey web site at : http://www.bgs.ac.uk/arsenic/bangladesh/DataDownload.htm and select "Village survey (59kB)" to download water supply well arsenic data for Bangladesh.

When you select the "Village survey (59kB)" link you should see an EXCEL spreadsheet in your browser. The data of interest are the "As arsenator' values in column O ("arsenator" refers to particular field method for measuring total arsenic). Rather than downloading this data with MATLAB commands simply select the values in column O, copy, and then paste into a MATLAB script. You can do this by pasting the data after the following expression:

```
arsenic_data = [
```

and then entering a final `]` after the data appear in your script. The result should be a long column vector called `arsenic_data` containing the values from the spreadsheet. You can insert this statement directly in your program (there are other ways to bring the data in -- feel free to do something else if you want).

In the rest of your script plot the sample CDF and histograms for the arsenic data.

Construct a MATLAB function `cdffit(data,ndist,p1,p2)` that fits the sample arsenic CDF to one of the above common CDFs, indexed by `ndist = 1,2,3,or 4`. The inputs `p1` and `p2` are distributional parameters. Adjust these to achieve the best possible fit.

Display the sample and postulated distributions on the same set of axes (using the MATLAB hold function).

Provide copies of plots for your preferred choice of postulated distribution and include a few sentences describing why you made this choice. Repeat the entire process for the iron data in column AA. Be sure to edit out any non-numeric characters (such as <). Also, compute from your fitted distribution the probability that the arsenic level exceeds the new (lowered) US standard (upper limit) of 10 uG/L.

Some relevant MATLAB functions: `cdfplot,hold,unifcdf, expcdf, normcdf, logncdf`

Problem 1 Solution

```
function cdffit(data,ndist,p1,p2)
close all
% Problem Set 4 -- Question 1 function
% call in command window or from other program
% write (for example):
% load arsenic.txt
% cdffit(arsenic,2,110,0)
figure
cdfplot(data)
hold
x=0:.1:max(data);
if ndist==1
     y=unifcdf(x,p1,p2);
     title('Actual and Theoretical Uniform CDF',…
     ' for Bangladesh Data')
end
if ndist==2
     y=expcdf(x,p1);
     title('Actual and Theoretical Exponential CDF', …
     'for Bangladesh Data')
end
if ndist==3
     y=normcdf(x,p1,p2);
     title('Actual and Theoretical Normal CDF', …
     ' for Bangladesh Data')
end
if ndist==4
     y=logncdf(x,p1,p2);
     title('Actual and Theoretical Lognormal CDF', …
     ' for  Arsenic Data')
end
plot(x,y,'--')
legend('Actual','Theoretical')
xlabel('Contaminant Level [ug/L]')
ylabel('F(x)')
```

```
return
```

Problem 2

Use the definition of the cumulative distribution function (CDF) to find the CDF $F_y(y)$ of the random variable $y = g(x)$, given the CDF $F_x(x)$ for each of the following problems. Also, find the probability density function $f_y(y)$ in each case by differentiating $F_y(y)$. Finally, compute the mean and variance of y from appropriate integrals over $f_x(x)$. Pay particular attention to the inequalities that define the limits on x and, by implication, limits on y.

a) Suppose $y = g(x) = 3x$ and $F_x(x) = 1 - \exp(-x)$; $x \geq 0$ (x has an exponential CDF). The function $g(x)$ in this case is a linear transformation that changes the shape of the exponential distribution. As an example, x could be the waiting time between two service events and y could be the cost incurred by waiting (a linear function of waiting time). The derived distribution problem considers how uncertainty in waiting time translates to uncertainty in cost.

b) Suppose $y = g(x) = \dfrac{x}{x+2}$ and $F_x(x) = 0.2x$; $0 \leq x < 5$ (x has a uniform CDF). The function $g(x)$ in this case is often used to describe saturation phenomena. For example, x could be the concentration of a nutrient and y the growth rate of an organism. This derived distribution problem considers how uncertainty in the nutrient concentration translates to uncertainty in the growth rate.

Problem 2 Solutions:

a.) $x = y/3$ $\qquad F_y(y) = 1 - e^{-y/3}$

$f_y(y) = dF_y(y)/dy = (1/3)e^{-y/3}$

$$E[y] = \int_0^\infty \frac{y}{3} e^{-\frac{y}{3}} dy = 3$$

$$VAR[y] = \int_0^\infty \frac{(y-3)^2}{3} e^{-\frac{y}{3}} dy = 9$$

b.) $x = 2y/(1 - y)$ $\qquad F_y(y) = 0.4y/(1 - y)$

$f_y(y) = dF_y(y)/dy = 0.4/(1 - y)^2$

$$E[g(x)] = \overline{g} = \int_0^5 0.2 \frac{x}{x+2} dx = 0.2\left[x\ln(x+2) - (x+2)[\ln(x+2) - x - 2\right]_0^5 = 0.50$$

$$Var[y] = E\left\{[g(x) - \bar{g}]^2\right\} = \int_0^5 0.2\left[\frac{x}{x+2} - 0.5\right]^2 dx = 0.03$$