1.017/1.010 Class 10
Some Common Probability Distributions

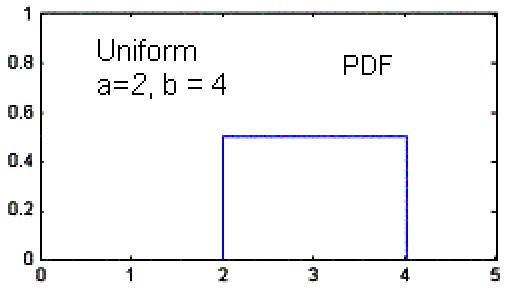
Some common continuous PDFs and CDFs

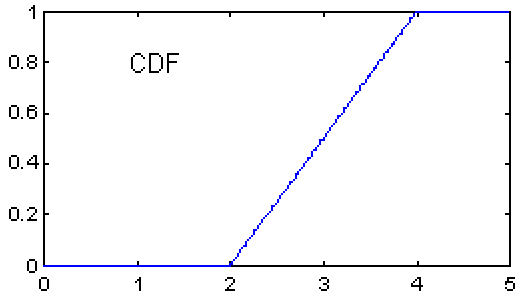
Each depends on a few distributional parameters

1. Uniform (distributional parameters a,b):

$$f_x(x) = \frac{1}{b-a} \qquad F_x(x) = a + \frac{x-a}{b-a} \quad ; \quad a \le x \le b$$

$$E(x) = \frac{a+b}{2} \qquad Var(x) = \frac{(b-a)^2}{12}$$

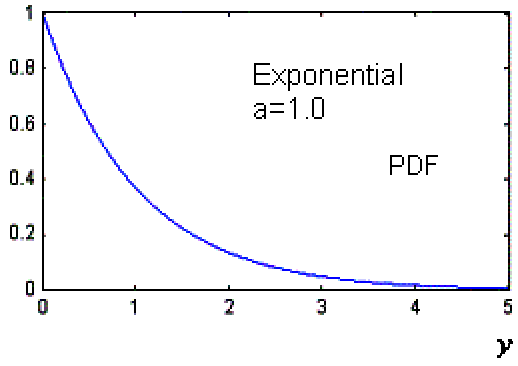


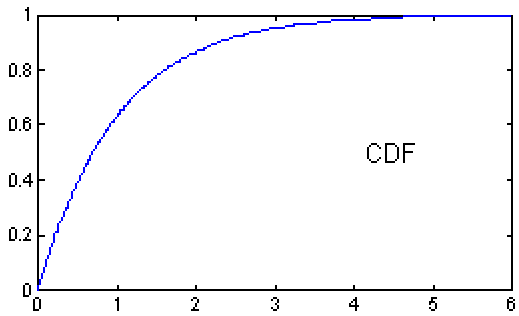


2. Exponential (distributional parameter a):

$$f_x(x) = \frac{1}{a}\exp\left[-\frac{x}{a}\right] \quad F_x(x) = 1 - \exp\left[-\frac{x}{a}\right] \quad ; \quad x \ge 0$$

$$E(x) = a \qquad Var(x) = a^2$$

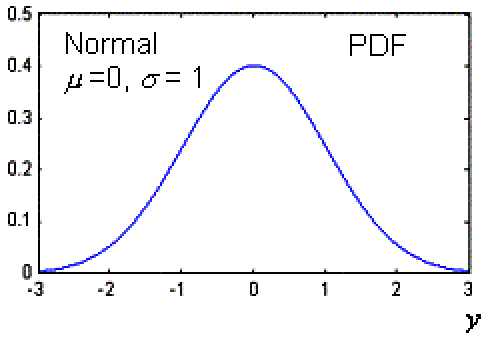


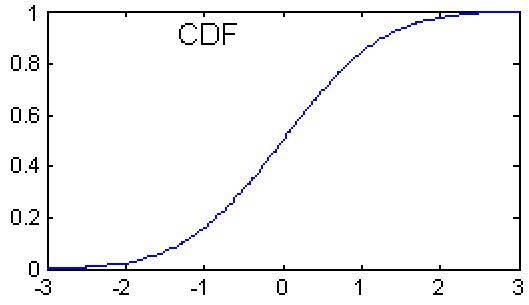


3. Univariate Normal (Gaussian) (distributional parameters μ, σ):

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad F_x(x) = \text{no closed form, tabulated}$$

$$E(x) = \mu \qquad Var(x) = \sigma^2$$



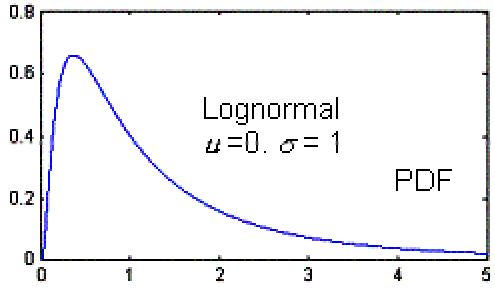


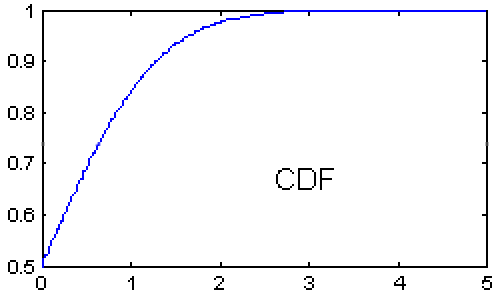
4. Lognormal (distributional parameters $\mu_{\ln x}, \sigma_{\ln x}$):

$$f_x(x) = \frac{1}{\sigma x\sqrt{2\pi}} \exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right] \qquad x \geq 0 \qquad F_x(x) = \text{no closed form, tabulated}$$

$$E(x) = \exp\left[\mu_{\ln x} + \frac{1}{2}\sigma^2_{\ln x}\right]$$

$$Var(x) = \exp\left[2\mu_{\ln x} + \sigma^2_{\ln x}\right]\left[\exp(\sigma^2_{\ln x}) - 1\right]$$





Exercise: Fitting distributions to data

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`

MATLAB function: cdffit.m