

Duke University
Department of Civil and Environmental Engineering
CE 130L. Uncertainty, Design, and Optimization

Spring 2011

Philip Scott Harvey and Henri P. Gavin

Probability Distributions

1 Probability Distributions

Consider a continuous, random variable (rv) X with support over the domain \mathcal{X} . The **probability density function** (PDF) of X is the function $f_X(x)$ such that for any two numbers a and b with $a < b$,

$$P[a < X \leq b] = \int_a^b f_X(x) dx$$

For $f_X(x)$ to be a proper distribution, it must satisfy the following two conditions:

1. $f_X(x)$ is positive-valued; $f_X(x) \geq 0$ for all values of $x \in \mathcal{X}$
2. the rule of total probability; the total area under $f_X(x)$ is 1, $\int_{\mathcal{X}} f_X(x) dx = 1$

Alternately, X may be described by its **cumulative distribution function** (CDF). The CDF of X is a function $F_X(x)$ that gives, for any specified number $x \in \mathcal{X}$, the probability that the random variable X is less than or equal to the number x is written as $P[X \leq x]$. For real values of x , the CDF is defined by

$$F_X(x) = P[X \leq b] = \int_{-\infty}^b f_X(x) dx ,$$

so,

$$P[a < X \leq b] = F_X(b) - F_X(a)$$

Every cumulative distribution function $F_X(x)$ is a monotonic non-decreasing function.

By the first fundamental theorem of calculus, the functions $f_X(x)$ and $F_X(x)$ are related as

$$f_X(x) = \frac{d}{dx} F_X(x)$$

A few important characteristics for the CDF of X are:

1. For any number a , $P[X > a] = 1 - P[X \leq a] = 1 - F_X(a)$
2. For any two numbers a and b with $a < b$, $P[a < X \leq b] = F_X(b) - F_X(a) = \int_a^b f_X(x) dx$

2 Descriptors of random variables

The **expected** or **mean value** of a continuous random variable X with PDF $f_X(x)$ is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

The expected value of an arbitrary function of X , $g(X)$, with respect to the PDF $f_X(x)$ is

$$\mu_{g(X)} = E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

The **variance** of a continuous rv X with PDF $f_X(x)$ and mean μ_X gives a quantitative measure of how much spread or dispersion there is in the distribution of x values. The variance is calculated as

$$\begin{aligned} \sigma_X^2 = V(X) &= \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx \\ &= \\ &= \\ &= \\ &= \end{aligned}$$

The **standard deviation** (s.d.) of X is $\sigma_X = \sqrt{V(X)}$. The **coefficient of variation** (c.o.v.) of X is defined as the ratio of the standard deviation σ_X to the mean μ_X :

$$c_X = \frac{\sigma_X}{\mu_X}$$

for non-zero mean. The c.o.v. is a normalized measure of dispersion (dimensionless).

A **mode** of a probability density function, $f_X(x)$, is a value of x such that the PDF is maximized;

$$\left. \frac{d}{dx} f_X(x) \right|_{x=x_{\text{mode}}} = 0 .$$

The **median** value is the value of x such that

$$P[X \leq x_{\text{median}}] = P[X > x_{\text{median}}] = F_X(x_{\text{median}}) = 1 - F_X(x_{\text{median}}) = 0.5 .$$

3 Some common distributions

A few commonly-used probability distributions are described at the end of this document: the uniform, triangular, exponential, normal, and log-normal distributions. For each of these distributions, this document provides figures and equations for the PDF and CDF, equations for the mean and variance, the names of MATLAB functions to generate samples, and empirical distributions of such samples.

3.1 The Normal distribution

The Normal (or Gaussian) distribution is perhaps the most commonly used distribution function. The notation $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ denotes that X is a normal random variable with mean μ_X and variance σ_X^2 . The **standard normal** random variable, Z , or “ z -statistic”, is distributed as $\mathcal{N}(0, 1)$. The probability density function of a standard normal random variable is so widely used it has its own special symbol, $\phi(z)$,

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Any normally distributed random variable can be defined in terms of the standard normal random variable, through the change of variables

$$X = \mu_X + \sigma_X Z.$$

If X is normally distributed, it has the PDF

$$f_X(x) = \phi\left(\frac{x - \mu_X}{\sigma_X}\right) = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(x - \mu_X)^2}{2\sigma_X^2}}$$

There is no closed-form equation for the CDF of a normal random variable. Solving the integral

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-u^2/2} du$$

would make you famous. Try it. The CDF of a normal random variable is expressed in terms of the **error function**, $\text{erf}(z)$. If X is normally distributed, $P[X \leq x]$ can be found from the standard normal CDF

$$P[X \leq x] = F_X(x) = \Phi\left(\frac{x - \mu_X}{\sigma_X}\right).$$

Values for $\Phi(z)$ are **tabulated** and can be computed, e.g., the MATLAB command ...

`Prob_X_le_x = normcdf(x,muX,sigX)`. The standard normal PDF is symmetric about $z = 0$, so $\phi(-z) = \phi(z)$, $\Phi(-z) = 1 - \Phi(z)$, and $P[X > x] = 1 - F_X(x) = 1 - \Phi((x - \mu_X)/\sigma_X) = \Phi((\mu_X - x)/\sigma_X)$.

The linear combination of two independent normal rv's X_1 and X_2 (with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2) is also normally distributed,

$$aX_1 + bX_2 \sim \mathcal{N}\left(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2\right),$$

and more specifically, $aX - b \sim \mathcal{N}(a\mu_X - b, a^2\sigma_X^2)$.

Given the probability of a normal rv, i.e., given $P[X \leq x]$, the associated value of x can be found from the **inverse standard normal CDF**,

$$\frac{x - \mu_X}{\sigma_X} = z = \Phi^{-1}(P[X \leq x]).$$

Values of the inverse standard normal CDF are **tabulated**, and can be computed, e.g., the MATLAB command ... `x = norminv(Prob_X_le_x,muX,sigX)`.

3.2 The Log-Normal distribution

The Normal distribution is symmetric and can be used to describe random variables that can take positive as well as negative values, regardless of the value of the mean and standard deviation. For many random quantities a negative value makes no sense (e.g., modulus of elasticity, air pressure, and distance). Using a distribution which admits only positive values for such quantities eliminates any possibility of non-sensical negative values. The log-normal distribution is such a distribution.

If $\ln X$ is normally distributed (i.e., $\ln X \sim \mathcal{N}(\mu_{\ln X}, \sigma_{\ln X}^2)$) then X is called a *log-normal* random variable. In other words, if Y is normally distributed, e^Y is log-normal.

$$\mu_Y = \mu_{\ln X}$$

$$\sigma_Y^2 = \sigma_{\ln X}^2$$

The mean and standard deviation of a log-normal variable X are related to the mean and standard deviation of $\ln X$.

$$\mu_{\ln X} = \ln \mu_X - \frac{1}{2} \sigma_{\ln X}^2$$

$$\sigma_{\ln X}^2 = \ln \left(1 + (\sigma_X / \mu_X)^2 \right)$$

If $(\sigma_X / \mu_X) < 0.30$, $\sigma_{\ln X} \approx (\sigma_X / \mu_X) = c_X$

The median, x_m , is a useful parameter of log-normal rv's. By definition of the median value,

$$\Phi \left(\frac{\ln x_m - \mu_{\ln X}}{\sigma_{\ln X}} \right) = 0.5$$

So,

$$\frac{\ln x_m - \mu_{\ln X}}{\sigma_{\ln X}} = \Phi^{-1}(0.5) = 0$$

and, $\ln x_m = \mu_{\ln X} \leftrightarrow x_m = \exp(\mu_{\ln X}) \leftrightarrow \mu_X = x_m \sqrt{1 + c_X^2}$

For the log-normal distribution $x_{\text{mode}} < x_{\text{median}} < \mu_X$.

If $\ln X$ is normally distributed (X is log-normal) then (for $c_X < 0.3$)

$$P[X \leq x] = \Phi \left(\frac{\ln x - \ln x_m}{c_X} \right)$$

If $\ln X \sim \mathcal{N}(\mu_{\ln X}, \sigma_{\ln X}^2)$, and $\ln Y \sim \mathcal{N}(\mu_{\ln Y}, \sigma_{\ln Y}^2)$, and $Z = aX^n/Y^m$ then

$$\ln Z = \ln a + n \ln X - m \ln Y \sim \mathcal{N}(\mu_{\ln Z}, \sigma_{\ln Z}^2)$$

where $\mu_{\ln Z} = \ln a + n\mu_{\ln X} - m\mu_{\ln Y}$ and $\sigma_{\ln Z}^2 = (n\sigma_{\ln X})^2 + (m\sigma_{\ln Y})^2$.

4 Random variable generation using the Inverse CDF method

A sample of a random variable having virtually any type of CDF, $P = F_X(x)$ can be generated from a sample of a uniformly-distributed random variable, U , ($0 < U < 1$), as long as the inverse CDF, $x = F_X^{-1}(P)$ can be computed. There are many numerical methods for generating a sample of uniformly-distributed random numbers. It is important to be aware that samples from some methods are “more random” than samples from others. The MATLAB command `u = rand(1,N)` computes a (row) vector sample of N uniformly-distributed random numbers with $0 < u < 1$.

If X is a continuous rv with CDF $F_X(x)$ and U has a uniform distribution on $[0, 1]$, then the random variable $F_X^{-1}(U)$ has the distribution F_X . Thus, in order to generate a sample of data distributed according to the CDF F_X , it suffices to generate a sample, u , of the rv $U \sim \mathcal{U}[0, 1]$ and then make the transformation $x = F_X^{-1}(u)$.

For example, if X is exponentially-distributed, the CDF of X is given by $F_X(x) = 1 - e^{-\lambda x}$, so $F_X^{-1}(u) = -\ln(1 - F_X(x))/\lambda$. Therefore if u is a value from a uniformly-distributed rv in $[0, 1]$, then $x = -\ln(1 - u)/\lambda$ is a value from an exponentially distributed random variable.

Note that since expressions for $\Phi(z)$ and $\Phi^{-1}(P)$ do not exist, the generation of normally-distributed random variables requires other numerical methods. The MATLAB command `...`

`x = muX + sigX*randn(1,N)` computes a (row) vector sample of N normally-distributed random numbers.

5 Monte Carlo Simulation

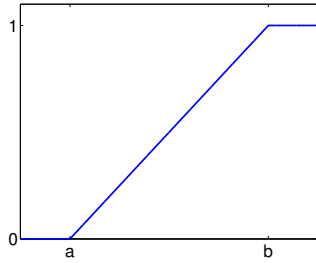
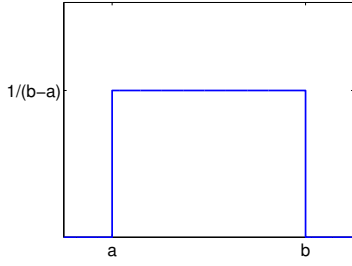
The probability distributions of virtually any function of random variables can be computed using the powerful method of Monte Carlo Simulation (MCS). MCS involves computing values of functions with large samples of random variables.

For example, consider a function of three random variables, X_1 , X_2 , and X_3 , where $X_1 \sim \mathcal{N}(6, 2)$, $\ln X_2 \sim \mathcal{N}(2, 1)$, and $X_3 \sim \mathcal{U}(5, 8)$. The function

$$Y = \sin(X_1) + \sqrt{X_2} - \exp(-X_3) - 5$$

is a function of random variables and is therefore also random. Given samples of N values of X_1 , X_2 and X_3 , a sample of N values of Y can also be computed. The statistics of Y , (mean, variance, PDF, and CDF) can be estimated by computing the average value, sample variance, histogram and ogive (empirical CDF) of the sample of Y . The probability $P[Y > 0]$ can be estimated by counting the number of positive values in the sample and dividing by N . The MATLAB command `P_Y_gt_0 = sum(y>0)/N` may be used to estimate this probability.

Uniform $X \sim \mathcal{U}[a, b]$
 $\mathcal{X} = \mathbb{R}, b > a$



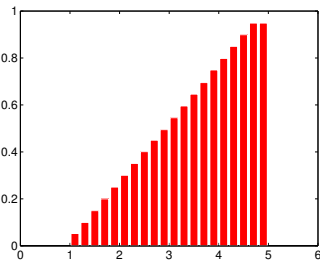
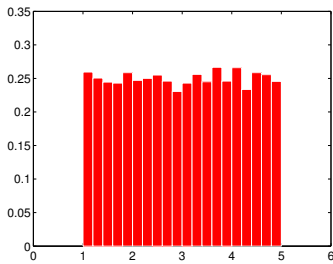
$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

$$F(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & x \in [a, b] \\ 1, & x \geq b \end{cases}$$

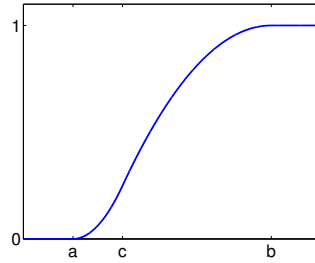
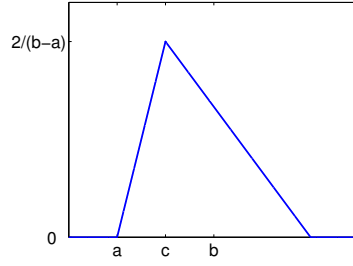
$$\mu_X = \frac{1}{2}(a + b)$$

$$\sigma_X^2 = \frac{1}{12}(b - a)^2$$

`x = a + (b-a).*rand(1,N);`



Triangular $X \sim \mathcal{T}ri(a, b, c)$
 $\mathcal{X} = \mathbb{R}, a \leq c \leq b$



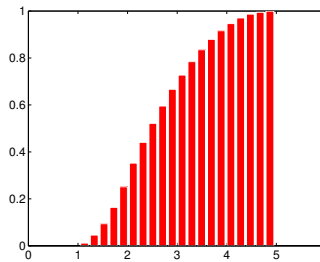
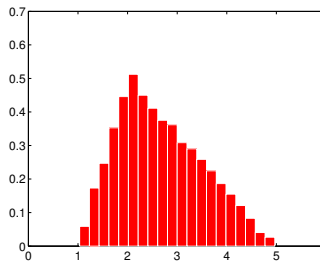
$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)}, & x \in [a, c] \\ \frac{2(b-x)}{(b-a)(b-c)}, & x \in [c, b] \\ 0, & \text{otherwise} \end{cases}$$

$$F(x) = \begin{cases} 0, & x \leq a \\ \frac{(x-a)^2}{(b-a)(c-a)}, & x \in [a, c] \\ 1 - \frac{(b-x)^2}{(b-a)(b-c)}, & x \in [c, b] \\ 1, & x \geq b \end{cases}$$

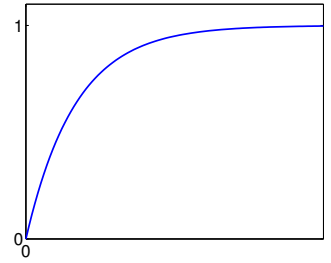
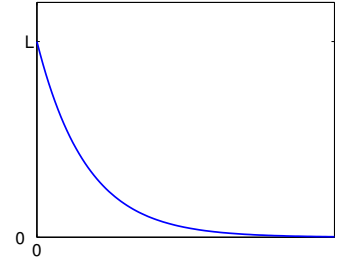
$$\mu_X = \frac{1}{3}(a + b + c)$$

$$\sigma_X^2 = \frac{1}{18}(a^2 + b^2 + c^2 - ab - ac - bc)$$

`x = triangular_rand(a,b,c,1,N);`



Exponential $X \sim \mathcal{E}xp(\lambda)$
 $\mathcal{X} = \mathbb{R}^+, \lambda > 0$



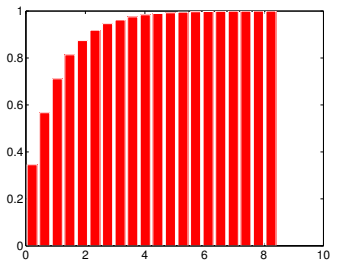
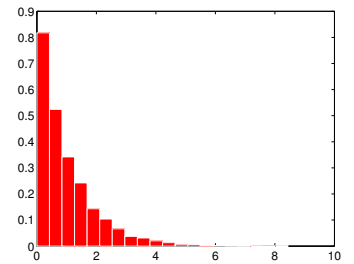
$$f(x) = \lambda e^{-\lambda x}$$

$$F(x) = 1 - e^{-\lambda x}$$

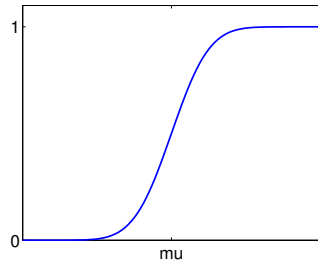
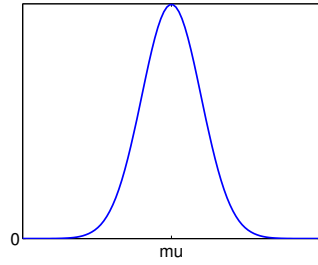
$$\mu_X = \lambda^{-1}$$

$$\sigma_X^2 = \lambda^{-2}$$

`x = exp_rand(lambda,1,N);`



Normal $X \sim \mathcal{N}(\mu, \sigma^2)$
 $\mathcal{X} = \mathbb{R}, \mu \in \mathbb{R}, \sigma^2 > 0$



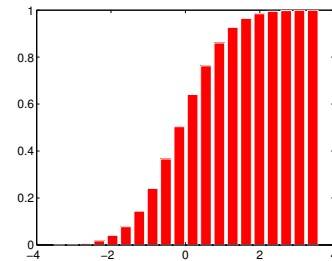
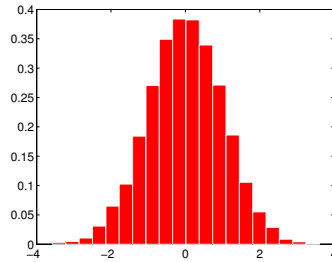
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$F(x) = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x-\mu}{\sqrt{2}\sigma} \right) \right]$$

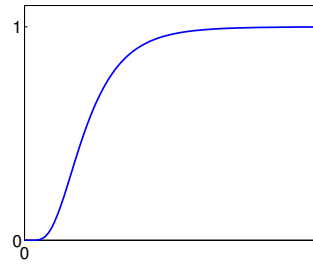
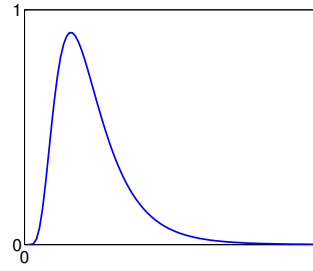
$$\mu_X = \mu$$

$$\sigma_X^2 = \sigma^2$$

`x = muX + sigmaX*randn(1,N);`



Log-Normal $\ln X \sim \mathcal{N}(\mu_{\ln X}, \sigma_{\ln X}^2)$
 $\mathcal{X} = \mathbb{R}^+, \mu_{\ln X} \in \mathbb{R}^+, \sigma_{\ln X}^2 > 0$



$$f(x) = \frac{1}{x\sqrt{2\pi\sigma_{\ln X}^2}} e^{-\frac{(\ln x - \mu_{\ln X})^2}{2\sigma_{\ln X}^2}}$$

$$F(x) = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\ln x - \mu_{\ln X}}{\sqrt{2}\sigma_{\ln X}} \right) \right]$$

$$\mu_X = \exp [\mu_{\ln X} + \sigma_{\ln X}^2/2]$$

$$\sigma_X^2 = (\exp[\sigma_{\ln X}^2] - 1) \exp[2\mu_{\ln X} + \sigma_{\ln X}^2]$$

`x = logn_rnd(muX,sigmaX,1,N);`

