REVIEW OF PROBABILITY THEORY

INTRODUCTION

Most water resources decision problems face the risk of uncertainty mainly because of the randomness of the variables that influence the performance of the systems. The hydrologic variables such as rainfall in a command area, inflow to a reservoir, evapo-transpiration of crops which influence decision making in water resources, are all random variables. Optimization models developed for water resources management must therefore be formulated to give optimal decisions with an indication of the associated hydrologic uncertainty. Two classical approaches to deal with the hydrologic uncertainty in optimization models are: Implicit Stochastic Optimization (ISO) and Explicit Stochastic Optimization (ESO).

In Implicit Stochastic Optimization (ISO) the hydrologic uncertainty is implicitly incorporated. The optimization model itself is a deterministic model, in which the hydrologic inputs are varied with a number of equi-probable sequences and the deterministic optimization model is run once with each of the input sequences. Output set is then statistically analyzed to generate a set of optimal decisions.

In Explicit Stochastic Optimization (ESO), the stochastic nature of the inputs is explicitly included in the optimization model through their probability distributions. Optimization model is a stochastic model and a single run of the model specifies the optimal decisions. Two commonly used ESO techniques are: Chance Constrained Linear Programming (CCLP), and Stochastic Dynamic Programming (SDP). These techniques will be discussed in the following lectures. However, a background of probability theory is essential for ESO, which will be discussed in the present lecture.

CONCEPT OF PROBABILITY

A sample space S is the area containing all possible outcomes of an experiment. An event is one subset of these outcomes. Probability is a measure of the likelihood of occurrence of an event. Probability can be assessed in two ways: (i) Objective or posterior probability which is based on the observation of events and (ii) Subjective or prior probability which is based on experience or judgement. Three basic axioms of probability are:

(i) Totality: P(S) = 1 where S is the sample space

(ii) Nonnegativity: P(A) ≥ 0 where A is an event

(iii) Mutually exclusive: If A and B are two mutually exclusive events, then

$$P(A \cup B) = P(A) + P(B)$$

For mutual exclusive events $P(A \cap B) = 0$. Hence, an extension of the above axiom after relaxing mutual exclusiveness will be

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

RANDOM VARIABLE

A variable whose value is not known or cannot be measured with certainty (or is nondeterministic) is called a random variable (r.v). Examples of random variables of interest in water resources are rainfall, streamflow, time between hydrologic events (e.g. floods of a given magnitude), evaporation from a reservoir, groundwater levels, re-aeration and de-oxygenation rates etc. Any function of a random variable is also a random variable. In this discussion, we use an upper case letter to denote a random variable and the corresponding lower case letter to denote the value that it takes. For example, daily rainfall may be denoted as X. The value it takes on a particular day is denoted as x. We then associate *probabilities* with events such as $X \geq x$, $0 \leq X \leq x$, etc.

Random variable can be essentially classified into two categories: discrete and continuous. If a r.v. X can take on only discrete values $x_1, x_2, x_3, \ldots$, then X is a discrete random variable.

An example of a discrete random variable is the number of rainy days in a year which may take on values such as, 10, 20, 30, ... A discrete random variable can assume a finite number of values. On the other hand, if a r.v. X can take on *all* real values in a range, then it is a continuous random variable. Most variables in hydrology are continuous random variables. The number of values that a continuous random variable can assume is infinite.

PROBABILITY DISTRIBUTIONS

For discrete random variables, the probability distribution is called a *probability mass function* and in case of continuous random variables it is called a *probability density function* (*pdf*). The cumulative distribution function (CDF), $F(x)$, represents the probability that X is less than or equal to x, i.e. $F(x) = P(X \leq x)$.

The probability mass function (PMF) of X is defined as $p(x) = P(X = x)$. The PMF of a discrete random variable and its CDF (appears as a staircase) are shown in figures 1(a) and (b) respectively. For a discrete random variable, there are spikes of probability associated with the values that the random variable assumes.

For a continuous random variable, the probability density function (PDF) is defined as $f(x) = \frac{dF(x)}{dx}$, where $F(x)$ is the CDF of X. The PDF and the corresponding CDF for continuous random variable are shown in figures 2(a) and (b) respectively. Probability distributions of continuous random variables are smooth curves. The cumulative distribution function (CDF) of a continuous random variable denoted by $F(x)$, is a non-decreasing function with a maximum value of 1. The CDF represents the probability that X is less than or equal to x, i.e. $F(x) = P(X \leq x)$.

Any function *f(x)* defined on the real line can be a valid probability density function if and only if (i) $f(x) \geq 0$ for all x, and (ii) $f(x) \leq 1$ for all x.

Given the PMF or PDF, the CDF can be obtained as

$$F(x) = \sum_{i \le i \le n} p(x_i) \qquad \text{for discrete random variables}$$

$$F(x) = \int_{-\infty}^{x} f(x)\,dx \qquad \text{for continuous random variables}$$

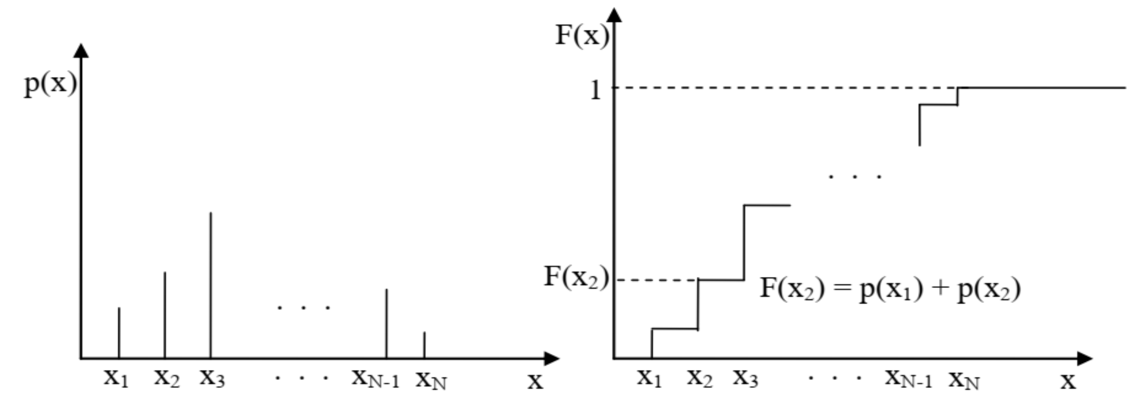

Fig. 1 (a) PMF and (b) CDF of a discrete random variable

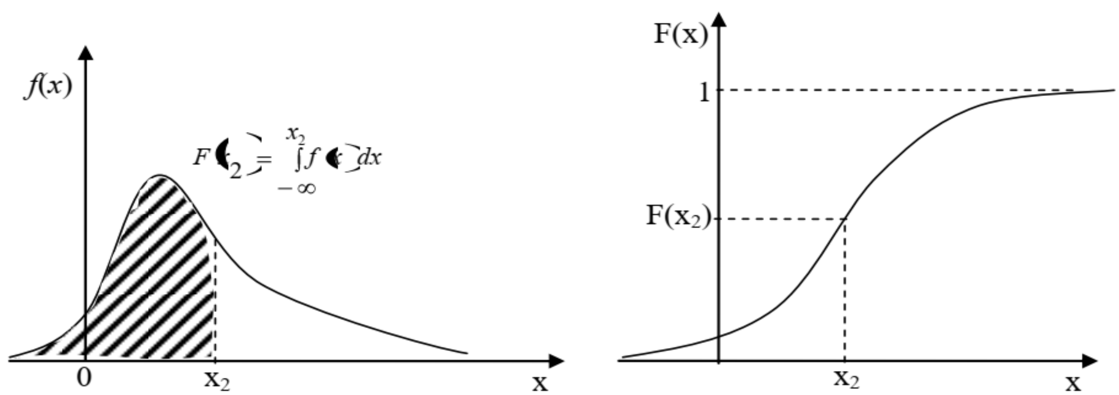

Fig. 2 (a) PDF and (b) CDF of a continuous random variable

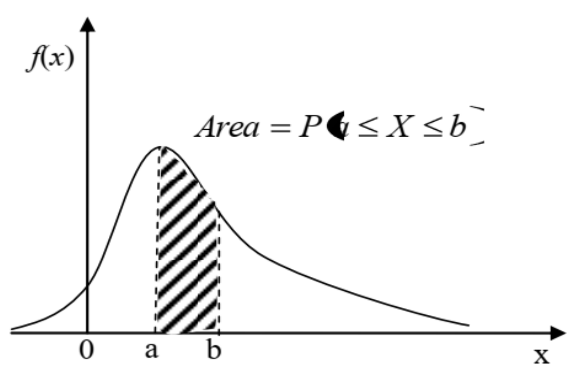

Fig. 3 Probability density function

Referring figure 3,

Area under the curve to the left of $x = a$ is $P(X \leq a)$

Area under the curve to the left of $x = b$ is $P(X \leq b)$

Area between $x = a$ and $x = b$ is $P[a \leq X \leq b]$.

For a continuous random variable, probability of the random variable taking a value exactly equal to a given value is zero because $P(X = d) = P(d \leq X \leq d) = \int_{d}^{d} f(x)\,dx = 0$.

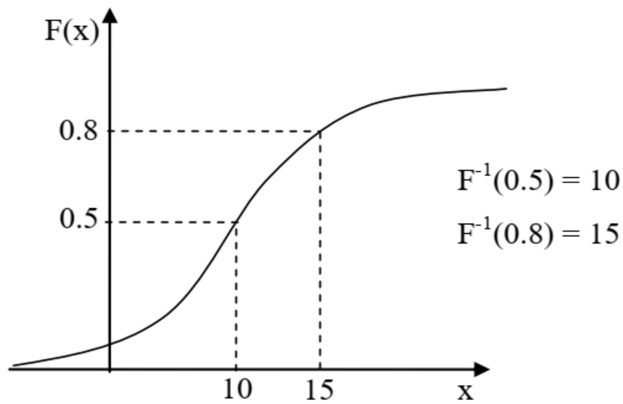

Fig. 4 CDF

Referring figure 4, for any given probability α, $0 \leq \alpha \leq 1$, the value x of the random variable can be determined from the CDF as $x = F^{-1}(\alpha)$.

<u>Statistical properties of random variables</u>

A population represents the set of all the values taken by a random process. A sample is a subset of the population. The expected value of $(X - x_0)^r$ is the r^{th} moment of a random variable X about any reference point $X = x_0$. Mathematically,

$$E\left[(X - x_0)^r\right] = \int_{-\infty}^{\infty} (x - x_0)^r f(x)\,dx \qquad \text{for continuous case}$$

$$E\left[(X - x_0)^r\right] = \sum_{i=1}^{N} (x_i - x_0)^r p(x_i) \qquad \text{for discrete case}$$

where $E[\ ]$ is a statistical expectation operator. The first three moments describe the central tendency, variability and asymmetry of the distribution of a random variable.

Expected Value or Mean

The central tendency is expressed as an expectation as

$$E[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx \qquad \text{for continuous case}$$

$$E[X] = \sum_{i=1}^{N} x_i\; p(x_i) \qquad \text{for discrete case}$$

The mean of a r.v is denoted by μ is equal to the expected value, i.e., $\mu = E[X]$.

Variance

It is the second order central moment. The variance of a continuous r.v. is defined as

$$Var(X) = \sigma^2 = E\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

The positive square root of variance is called the standard deviation, σ. Coefficient of variation is defined as $C_v = \frac{\sigma}{\mu}$.

Skewness

The asymmetry of PDF of a r.v. is measured by skew coefficient defined as

$$\gamma = E\left[(X-\mu)^3\right] / \sigma^3$$

Example:

Probability density function (PDF) of a random variable X is

$f(x) = 6\,x^2 \qquad 0 \le x \le 1$

$\quad = 0 \qquad$ else where

Determine (1) Cumulative distribution function (cdf); (2) Expected value, E(X); (3) Variance, Var (X); (4) P[$X \geq 0.6$]; and (5) P[$0.4 \leq X \leq 0.7$]

<u>Solution:</u>

1. Cumulative distribution function

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

$$= \int_{0}^{x} 6x^2\,dx = 2x^3 \qquad 0 \leq x \leq 1$$

2. Expected value, E(X)

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{0}^{1} x\,6x^2\,dx = 1.5$$

3. Variance, Var (X)

$$Var[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

$$= \int_{0}^{1} (x-3/2)^2\,6x^2\,dx = 1.2$$

4. P[$X \geq 0.6$]

$$P[X \geq 0.6] = 1 - P[X \leq 0.6] = 1 - F(0.6)$$

$$= 1 - 2 \times 0.6^3 = 0.568$$

5. P[$0.4 \leq X \leq 0.7$]

$$P[0.4 \leq X \leq 0.7] = P[X \leq 0.7] - P[X \leq 0.4]$$

$$= F(0.7) - F(0.4)$$

<u>Commonly used probability distributions</u>

Three commonly used distributions in water resources are: Normal, Lognormal and Exponential distributions.

Normal distribution

The normal distribution is also called Gaussian distribution. Two parameters are involved in this distribution: mean and variance. A normal random variable with mean μ and variance σ^2 is denoted as $X \sim N(\mu, \sigma^2)$. The PDF of the normal distribution given by $f(x)$ is expressed as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad for\ -\infty < x < \infty$$

The PDF of normal distribution is bell-shaped and symmetric at $x = \mu$ as shown in figure 5

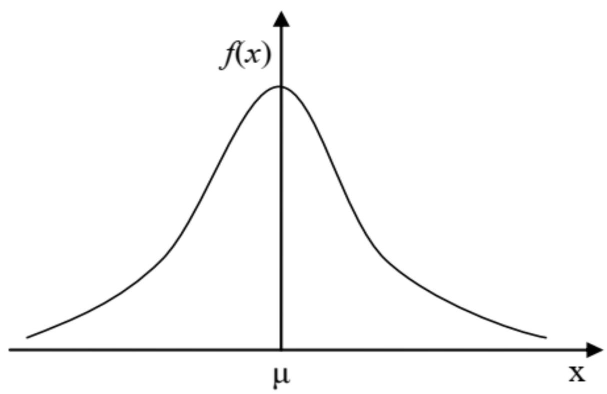

Fig. 5 Normal PDF

The CDF of a normal distribution is

$$f(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \qquad for\ -\infty < x < \infty$$

Normal random variables are usually transformed to standardized variate Z with zero mean and unit variance i.e., $Z = (X - \mu) / \sigma$. Then PDF of Z can be expressed as

$$\phi(z) = \frac{1}{\sqrt{2\pi}} exp\left[-\frac{z^2}{2}\right] \qquad for\ -\infty < z < \infty$$

Values of $\phi(z)$ obtained by numerical integration are used in the computations for normal distributions.

Example:

The monthly streamflow at a reservoir site is represented by a random variable X which follows normal distribution with a mean of 100 units and a standard deviation of 50 units. Find (1) $P[X > 150]$; (2) $P[X \leq 40]$ and (3) The flow value which will be exceeded with a probability of 0.8.

Solution:

The monthly streamflow at a reservoir site is represented by a random variable X which

(1) $P[X > 150]$

$$\begin{aligned} P[X > 150] &= P[(X-\mu)/\sigma \geq (150-100)/50] \\ &= P[Z \geq 1] = 1 - P[Z \leq 1] \\ &= 1 - 0.8413 = 0.1587 \end{aligned}$$

(2) $P[X \leq 40]$

$$\begin{aligned} P[X \leq 40] &= P[(X-\mu)/\sigma \leq (40-100)/50] \\ &= P[Z \leq -1.2] \\ &= 0.1539 \end{aligned}$$

(3) To find $P[X \geq x] = 0.8$

$$\begin{aligned} &P[X \geq x] = 0.8 \\ &P[Z \geq (x-\mu)/\sigma] = 0.8 \\ &1 - P[Z \leq z] = 0.8 \\ &P[Z \leq z] = 0.2 \\ &z = (x-100)/50 = -0.84 \\ &x = 58 \ units \end{aligned}$$

Lognormal distribution

This is used when random variable cannot be negative. A r.v. X is lognormally distributed if its logarithmic transform $Y=ln(X)$ has a normal distribution with mean $\mu_{\ln X}$ and variance $\sigma^2_{\ln X}$. The PDF of lognormal r.v. is

$$f(x) = \frac{1}{\sqrt{2\pi} X \sigma_{\ln X}} \exp\left[-\frac{1}{2}\left(\frac{\ln X - \mu_{\ln X}}{\sigma_{\ln X}}\right)^2\right] \qquad for \ -\infty < X < \infty$$

Exponential distribution

The probability density function (pdf) of an exponential distribution with parameter λ is:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Here $\lambda > 0$ is the parameter of the distribution.

Mean $E[X] = 1 / \lambda$

$\mathrm{Var}(X) = E[X^2] - E[X]^2 = 1 / \lambda^2$.

The cumulative distribution function is given by:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$