Module 7
7 Lectures

Hydrological Statistics

Prof. Subhankar Karmakar
IIT Bombay
Objective of this module is to learn the fundamentals of stochastic hydrological phenomena.
Topics to be covered

- Statistical parameter estimation
- Probability distribution
- Goodness of fit
- Concepts of probability weighted moments & l-moments
Module 7

Lecture 1: Introduction to probability distribution
Introduction

- **Deterministic model**
  - Variables involved are deterministic in nature.
  - No uncertainty for a given set of inputs.
  - E.g. Newtonian model:
    \[ s = ut + \frac{1}{2} ft^2 \] (2nd law of motion)

- **Probabilistic/Stochastic model**
  - Variables involved are random in nature.
  - There is always uncertainty for a given set of inputs.
  - E.g. Rainfall-runoff model.

In a deterministic model, a given input yields the same output always, while a probabilistic model yields different outputs for multiple trials with the same input.
Random Variables

Random Variable (R.V.) is a variable whose value cannot be predicted with certainty before it actually takes over the value. R.V. may be discrete or continuous.

E.g. rainfall, stream flow, soil hydraulic properties (permeability, porosity, etc.), evaporation, diffusion, temperature, groundwater level, etc.

Notations:
- Capital letters X, Y, Z indicates R.V.
- Small letter x, y, z indicates value of the R.V.

\[ X \leq x \]

\[ \uparrow \quad \uparrow \]

Rainfall 30 mm (value)

If X is a R.V. & Z(X) is function of X \( \Rightarrow \) Z is a R.V.

E.g. Rainfall is R.V, so Runoff is R.V.
Random Variables

Discrete random variables

- If the set of values for a R.V. can be assumed to be finite (or countable infinite), then RV is said to be a discrete RV.
  
  E.g. No. of raining days in a month (0,1,2,…,31)
  No. of particles emitted by a radio-active material.

Continuous random variables

- If the set of values for a R.V. can be assumed to be infinite, then RV is said to be a continuous R.V.
  
  E.g. Depth of rainfall in a period of given month

By using ‘class interval’ we can convert continuous random variables to discrete random variables.
Probability Distribution

**Discrete Variable**

Important conditions:

(i) \(0 \leq P[X = x_i] \leq 1\)

(ii) \(\sum_i P[X = x_i] = 1\)

[Here 1 is likelihood of occurrence]

**Cumulative distribution function (CDF)**

- CDF is a step function

\[
P[X \leq x_i] = \sum_{x_i \leq x} P[X = x_i]
\]

\[
P(X = x) = \left\{ \forall X = x_1, x_2, \ldots, x_n \right\}
\]
Continuous Variable

- Probability density function (PDF), $f(x)$ does not indicate probability directly but it indicates probability density.

- CDF is given by

$$ F(a) = P[X \leq a] = \int_{-\alpha}^{a} f(x) dx $$
Piecewise Continuous Distribution

\[
\int_{-\alpha}^{d} f_1(x)dx + P[X=d] + \int_{d}^{\alpha} f_2(x)dx = 1
\]

Then, \( P[X < d] \neq P[X \leq d] \);

Similarly, \( P[X > d] \neq P[X \geq d] \);
CDF within an interval

Here, interval is given by \((x-\Delta x/2, x+\Delta x/2)\)

\[
\left( x_i - \frac{\Delta x}{2} \right) \frac{\Delta x}{2} \quad x_i \quad \left( x_i + \frac{\Delta x}{2} \right)
\]

\[
P[x_i - \frac{\Delta x}{2} \leq X \leq x_i + \frac{\Delta x}{2}]
= F(x_i + \frac{\Delta x}{2}) - F(x_i - \frac{\Delta x}{2}) = \text{Area under the curve in the interval}
\]

\[
\left( x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2} \right)
= f(x) \cdot \Delta x_i
\]
Example Problem

Estimate the expected relative frequencies within each class of interval 0.25.

\[ f(x) = \frac{3x^2}{125}; \quad 0 \leq x \leq 5 \]

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( f(x_i) = \frac{3x^2}{125} )</th>
<th>( f(x_i) \times \Delta x = \text{expected rel. freq.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.0015</td>
<td>0.00075</td>
</tr>
<tr>
<td>0.75</td>
<td>0.0135</td>
<td>0.00675</td>
</tr>
<tr>
<td>1.25</td>
<td>0.0375</td>
<td>0.01875</td>
</tr>
<tr>
<td>1.75</td>
<td>0.0735</td>
<td>0.03675</td>
</tr>
<tr>
<td>2.25</td>
<td>0.1215</td>
<td>--</td>
</tr>
<tr>
<td>2.75</td>
<td>0.066</td>
<td>--</td>
</tr>
<tr>
<td>3.25</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3.75</td>
<td>0.3375</td>
<td>0.16875</td>
</tr>
<tr>
<td>4.25</td>
<td>0.4335</td>
<td>0.21675</td>
</tr>
<tr>
<td>4.75</td>
<td>0.5415</td>
<td>0.27075</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td>0.9975 ≈ 1.0</td>
</tr>
</tbody>
</table>
Bivariate Random Variables

- (X,Y) – two dimensional random vector or random variable
  - e.g. Temperature $\rightarrow$ Evaporation
  - Temperature $\rightarrow$ Radiation Coeff.
  - Rainfall $\rightarrow$ Recharge
  - Rainfall $\rightarrow$ Runoff

- Variables may or may not be dependent

**Case 1:** (X,Y) is a 2-D discrete random vector

The possible values of may be represented as $(X_i, Y_j), \quad i=1,2,\ldots,n \quad \& \quad j=1,2,\ldots,n$

**Case 2:** (X,Y) is a 2-D continuous random variable

Here, (X,Y) can assume all possible values in some non-countable set.
1. Discrete 2-D Random Vector

\[ F(x, y) = P[X \leq x, Y \leq y] = \sum_{-\infty}^{x} \sum_{-\infty}^{y} p(X, Y) \]

\[ F(\infty, \infty) = 1 \]

E.g. Calculate the following using the values from the given table.

1) \( F(3,2) \)

2) \( F(7,4) \)

3) \( F(7,8) \)

\[ \sum_{i=1}^{6} \sum_{j=1}^{4} p(X_i, Y_j) = 1 \]

\[ F(3,2) = P[X \leq 3, Y \leq 2] = 0.35 \]

\[ F(7,4) = F(\infty, \infty) = 1 \]

\[ F(7,8) = 1 \]
2. Continuous 2-D Random Vector

If \( f(X,Y) \rightarrow \) joint pdf of \((X,Y)\)

\[ f(X,Y) \geq 0 \]

\[ \int_{-\alpha}^{\alpha} \int_{-\alpha}^{\alpha} f(x,y)dx\,dy = 1 \]

\( F(x,y) \rightarrow \) joint cdf of \((X,Y)\)

\[ = P[X \leq x, Y \leq y] \]

\[ = \int_{-\alpha}^{y} \int_{-\alpha}^{x} f(x,y)dx\,dy \]

Probability of hatched region on the plane

\[ = \int_{B} f(x,y)dB \]

\( F(\alpha,\alpha) = 1 \)

\( F(-\alpha,y) = F(x,-\alpha) = 0 \)
Example Problem 1

Consider the flows in two adjacent streams. Denote it as a 2-D RV \((X,Y)\), with a joint pdf,

\[
f(X,Y) = C \quad \text{if} \quad 5,000 \leq x \leq 10,000 \quad \text{and} \quad 4,000 \leq y \leq 9,000
\]

Solution: To get \(C\),

\[
\int_{-\alpha}^{\alpha} \int_{-\alpha}^{\alpha} f(x,y) \, dx \, dy = 1 \quad \Rightarrow \quad \int_{-\alpha}^{\alpha} \int_{-\alpha}^{\alpha} C \, dx \, dy = 1
\]

\[
\therefore \quad C = \frac{1}{(5000)^2}
\]

For a watershed region \(B\),

\[
P(B) = \int \int_{B} f(x,y) \, dx \, dy
\]
Example Problem 1

Determine $P[X \geq Y]$

\[
1 \leq P[X \leq Y] = 1 - \int_{5000}^{9000} \int_{5000}^{y} f(x, y) \, dx \, dy
\]

\[
= 1 - \int_{5000}^{9000} \int_{5000}^{y} C \, dx \, dy = 1 - \int_{5000}^{9000} [y - 5000] \, dy
\]

\[
= 1 - C \left[ \frac{y^2}{2} - 5000y \right]_{5000}^{9000}
\]

\[
= 1 - C \left[ \frac{(9000)^2}{2} - 45000000 - \frac{(5000)^2}{2} + 25000000 \right]_{5000}^{9000}
\]

\[
= \frac{17}{25}
\]
Example Problem 1

Calculate $F(10000,9000)$:

$$F(x,y) = \int_{-\alpha}^{x} \int_{-\alpha}^{y} f(x,y) \, dy \, dx = \int_{-\alpha}^{x} \int_{-\alpha}^{y} \frac{1}{(5000)^2} \, dy \, dx$$

$$= \int_{-\alpha}^{x} \left[ \frac{y}{(5000)^2} - \frac{4000}{(5000)^2} \right] \, dx$$

$$= \frac{xy}{(5000)^2} - \frac{4000x}{5 \times (5000)} - \frac{y(5000)}{(5000)^2} + \frac{5000 \times 4000}{(5000)^2}$$

$$= \frac{xy}{(5000)^2} - \frac{4x}{5 \times (5000)} - \frac{y}{(5000)} + \frac{4}{5} \text{ (Ans.)}$$

$F(10000,9000)$

$$= \frac{10000 \times 9000}{(5000)^2} - \frac{4(10000)}{5 \times (5000)} - \frac{9000}{(5000)} + \frac{4}{5}$$

$$= 3.6 - 1.6 - 1.8 + 0.8$$

$$= 1.0$$
Example Problem 2

\[ f(x, y) = x^2 + xy \quad 0 \leq x \leq 1 \, \& \, 0 \leq y \leq 2 \]

\[ = 0 \quad \text{elsewhere} \]

Find, P[x+y \geq 1], verify \( \int_{-\alpha}^{\alpha} \int_{-\alpha}^{\alpha} f(x, y) \, dx \, dy = 1 \)

Solution:

\[
\int_{0}^{2} \int_{0}^{1} (x^2 + \frac{xy}{3}) \, dx \, dy = \int_{0}^{2} \left[ \frac{x^3}{3} + \frac{x^2 y}{6} \right]_{0}^{1} \, dy
\]

\[
= \int_{0}^{2} \left[ \frac{1}{3} + \frac{y}{6} \right] \, dy = \left[ \frac{1}{3} + \frac{y^2}{12} \right]_{0}^{2} = \left[ \frac{2}{3} + \frac{4}{12} \right]
\]

\[ = 1 \, \text{[verified]} \]
Example Problem 2

\[ P[X + Y \geq 1] \]

\[ = 1 - P[X + Y \leq 1] \]

\[ = 1 - \int_0^1 \int_0^{1-x} \left( x^2 + \frac{xy}{3} \right) dy \, dx \]

\[ = 1 - \int_0^1 \left[ x^2 y + \frac{y^2 x}{6} \right]_{0}^{1-x} \, dx \]

\[ = 1 - \int_0^1 \left[ x^2 (1-x) + \frac{x(1-2x+x^2)}{6} \right] \, dx \]

\[ = 1 - \int_0^1 \left[ 6x^2 - 6x^3 + x - 2x^2 + x^3 \right] \, dx \]

\[ = 1 - \frac{1}{6} \left( 4x^3 - 5x^4 + x^2 \right) \bigg|_{0}^{1} \]

\[ = 1 - \frac{1}{6} \left[ \frac{4}{3} - \frac{5}{4} + \frac{1}{2} \right] \]

\[ = 1 - \frac{1}{6} \left[ \frac{7}{12} \right] \]

\[ = \frac{65}{72} \]
Marginal Distribution Function

- **Discrete variables**

Marginal Distribution of X is 
\[ p[x_i] = \sum_{j=1}^{\alpha} p(x_i, y_j), \quad \forall i \]

Marginal Distribution of y is 
\[ q[y_j] = \sum_{i=1}^{\alpha} p(x_i, y_j), \quad \forall j \]

- **Continuous variables**

Marginal distribution function of x is given by 
\[ g(x) = \int_{-\alpha}^{\alpha} f(x, y) dy \]

and of y is given by 
\[ h(y) = \int_{-\alpha}^{\alpha} f(x, y) dx \]

[g(x) & h(y) are also pdf and should satisfy the conditions]
Marginal Distribution Function

\[ P[c \leq X \leq d] = P[c \leq X \leq d, -\alpha \leq y \leq \alpha] \]

\[ = \int_c^d \left[ \int_{-\alpha}^{\alpha} f(x,y) \, dy \right] \, dx \]

\[ = \int_c^d g(x) \, dx \]

\[ f(x,y) = g(x) \times h(y) \quad \text{......stochastically independent} \]

\[ g(x) \text{ is original distribution of the 1-D random variable } X \]

Remember C.D.F = \( F(x) = \int_{-\alpha}^{x} g(x) \, dx \)

\( g(x) \) is same as \( f(x) \), i.e. original density function of \( x \).
Example Problem

Function $P_{x,y}(x,y) = C(5 - \frac{y}{2} - x)$ for $0 \leq x \leq 2$ and $0 \leq y \leq 2$ can serve as a bivariate continuous probability density function.

i) Find the value of $C$

\[
\begin{align*}
\therefore \text{ Probability (x \leq 2 & y \leq 2) } &= \int_{0}^{2} \int_{0}^{2} C \left(5 - \left(\frac{S}{2}\right) - t\right) \, ds \, dt = 1 \\
\int_{0}^{2} C \left[5s - \frac{s^2}{4} - ts\right]_0^2 \, dt &= 1 \\
\int_{0}^{2} C[10 - 1 - 2t] \, dt &= 1 \\
\int_{0}^{2} C[9 - 2t] \, dt &= 1 \Rightarrow C[9t - t^2]_0^2 = 1 \\
C[18 - 4] &= 1 \Rightarrow C = \frac{1}{14}
\end{align*}
\]
ii) Find probability of $x \leq 1$ & $y \leq 1$

$$P_{x,y}(x,y) = \text{prob}(X \leq x, Y \leq y)$$

$$= \int_0^x \int_0^y (5 - \left(\frac{S}{2}\right) - t) / 14 \ ds dt$$

$$= \int_0^x \frac{1}{14} [5y - \frac{y^2}{4} - ty] dt = \frac{1}{14} \left[ 5xy - \frac{xy^2}{4} - \frac{x^2y}{2} \right] = \frac{1}{14} \left[ 5 - \frac{1}{4} - \frac{1}{2} \right] = 0.304$$

iii) Find "marginal densities" $P_x(x)$ & $P_y(y)$

$$P_x(x) = \int_0^2 P_{x,y}(x,s) ds = \int_0^2 (5 - \frac{S}{2} - x) / 14 \ ds$$

$$= \frac{1}{14} \left[ 5s - \frac{s^2}{4} - xs \right]_0^2$$

$$= \frac{1}{14} [10 - 1 - 2x]$$

$$= \frac{1}{14} (9 - 2x)$$

$$P_y(y) = \int_0^2 P_{x,y}(t,y) dt$$

$$= \int_0^2 \frac{1}{14} (5 - \frac{y}{2} - t) dt$$

$$= \frac{1}{14} \left[ 5t - \frac{yt}{2} - \frac{t^2}{2} \right]_0^2$$

$$= \frac{1}{14} [10 - y - 2] = \frac{1}{14} (8 - y)$$
iv) Find cumulative marginal distributions $P_{x,y}(x, \alpha)$ & $P_{x,y}(\alpha, Y)$

\[
\therefore P_{x,y}(x, \alpha) = \text{prob}(X \leq x) = \int_0^x P_x(t)dt = \int_0^x \left( \frac{9 - 2x}{14} \right) dx
\]

\[
= \left[ \frac{9x - x^2}{14} \right]_0^x = \left( \frac{9x - x^2}{14} \right)
\]

\[
P_{x,y}(\alpha, y) = \text{prob}(Y \leq y) = \int_0^y P_y(s)ds = \int_0^y \left( \frac{8 - y}{14} \right) ds = \left( \frac{8y - y^2 / 2}{14} \right)
\]

Check putting $x=2$, \( \frac{9x - x^2}{14} = 1 \)

& $y=1$, \( \frac{16y - y^2}{28} = 1 \)

v) Find "conditional densities"

\[
P_{x/y}(x/y) = \frac{P_{x,y}(x,y)}{P_y(y)} = \frac{(5 - y/2 - x)}{(8 - y)}
\]

and \( P_{x/y}(y/x) = \frac{P_{x,y}(x,y)}{P_x(x)} = \frac{(5 - y/2 - x)}{(9 - 2x)} \)
Conditional Density Functions

Let \((X,Y)\) is a 2-D random vector, with joint density function of \(f(x,y)\). Let \(g(x)\) and \(h(y)\) be the marginal density functions of \(X\) and \(Y\) respectively.

The CDF of \(X\), given \(Y=y\) is defined as,

\[
g(x \mid y) = \frac{f(x,y)}{h(y)} \quad \text{but } h(y) > 0
\]

and the CDF of \(Y\), given \(X=x\) is defined as,

\[
h(y \mid x) = \frac{f(x,y)}{g(x)} \quad \text{but } g(x) > 0
\]

\[
g(x \mid y) \geq 0 \quad \text{as } f(x,y) \text{ and } h(y) \text{ are (+ve)}
\]

\[
\int_{-\alpha}^{\alpha} g(x \mid y) dx = 1 = \int_{-\alpha}^{\alpha} \frac{f(x,y)}{h(y)} dx = \frac{1}{h(y)} \int_{-\alpha}^{\alpha} f(x,y) dx = \frac{1}{h(y)} h(y) = 1
\]

\[
\therefore \int_{-\alpha}^{\alpha} f(x,y) dx = h(y)
\]
Conditional Density Functions

CDF = G(x|y) = P[ X ≤ x | Y = y ] = \int_{-\alpha}^{\alpha} g(x | y) dx

where y belongs to certain region, i.e. y ∈ R (R is a region)

\therefore g(x | y ∈ R) = \frac{\int_{\mathbb{R}} f(x,y) dy}{\int_{\mathbb{R}} h(y) dy}

Now cumulative conditional distribution function

= P [ X ≤ x | y ∈ R] = F( x | y ∈ R)

= F(X | y ∈ R) = \int_{-\alpha}^{\alpha} g(x | y ∈ R) dx
Independence of random variables

\[ g(x \mid y) = g(x) \text{, when } X \& Y \text{ are statistically independent variables} \]

\[ g(x \mid y) = \frac{f(x, y)}{h(y)} = g(x) \]

If \( f(x, y) = g(x) \times h(y) \), then \( x, y \) are independent

For e.g.,

\( f(x, y) = e^{-(x+y)}; \ x>0; \ y>0 \)

\[ \therefore g(x) = \int_0^\alpha e^{-(x+y)} dy = \left[ \frac{e^{-(x+y)}}{-1} \right]_0^\alpha \]

\[ = e^{-x}, \text{ x>0 & h(y) = e}^{-y}, \ y>0 \]

\[ \therefore f(x, y) = g(x) \times h(y) \therefore x \& y \text{ are independent} \]
Functions of random variables

- Simultaneous occurrence \(\rightarrow\) Joint density function
- Distribution of one variable irrespective of the value of the other variables \(\rightarrow\) Marginal density function
- Distribution of one variable conditioned on the other variable \(\rightarrow\) Conditioned distribution

Example 1:

\(x\) : discrete variable, \(p(x) = f(x) = \frac{c}{x}\), \(x=2,3,4,5\), \(y = x^2 - 7x + 12\)

\(y\) takes two values 0 & 2, \(p(y=0) = p(x=3) + p(x=4) = \frac{c}{3} + \frac{c}{4} = \frac{7c}{12}\)

\(p(y=2) = p(x=2) + p(x=5) = \frac{c}{2} + \frac{c}{5} = \frac{7c}{10}\)

<table>
<thead>
<tr>
<th>(x)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>(p(x))</td>
<td>c/2</td>
<td>c/3</td>
<td>c/4</td>
<td>c/5</td>
</tr>
</tbody>
</table>
**Example 2:**

For functions of RVs:

For continuous curve

\[ f(x) = e^{-x}; \quad x > 0 \]

\[ y = 2x + 1 \]

\[ P[y \geq 5] = P[X \geq \frac{y - 1}{2}] \]

\[ = P[X \geq 2] \]

\[ = 1 - P[X \leq 2] \]

\[ = 1 - \int_{0}^{2} e^{-x} \, dx \]

\[ = 1 - \left[ e^{-x} \right]_{0}^{2} \]

\[ = 1 - \left[ e^{-2} - e^{0} \right] \]

\[ = e^{-2} \quad \text{(Ans.)} \]

Alternatively

\[ \int_{2}^{\alpha} e^{-x} \, dx = \left[ \frac{e^{-x}}{-1} \right]_{2}^{\alpha} = e^{-2} \quad \text{(Ans.)} \]
i) \( g(x|y) = \frac{f(x,y)}{h(y)} \)

\[
= \frac{1}{14} \left(5 - \frac{y}{2} - x\right) \times \frac{14}{(8 - y)} = \frac{10 - y - 2x}{2(8 - y)} , \ 0 \leq x \leq 2, \ 0 \leq y \leq 2
\]

Conditional CDF:

\[
g(x | y) = P[X \leq x | Y = y] = \int_{-\alpha}^{x} g(x | y)dx = \frac{(10x - xy - x^2)}{2(8 - y)}
\]

Now, \( P[X \leq 1, Y = \frac{3}{2}] \)

\[
= a[1|\frac{3}{2}] = \frac{(10(1) - (1)(\frac{3}{2}) - 1^2)}{2(8 - \frac{3}{2})} = \frac{10 - \frac{3}{2} - 1}{2(\frac{13}{2})} = \frac{15}{26}
\]

ii) Now, \( g(x|y \leq 1) \ 0 \leq x \leq 2; \ 0 \leq y \leq 2 \)

\[
= \frac{\int_{0}^{1} f(x,y)dy}{\int_{0}^{1} h(y)dy} = \frac{\int_{0}^{1} \frac{1}{14} \left(5 - \frac{y}{2} - x\right)dy}{\int_{0}^{1} (8 - y)/14dy}
\]
\[
\int_0^1 (5 - \frac{y}{2} - x)dy = \int_0^1 (8 - y)dy = \left[ 5y - \frac{y^2}{4} - xy \right]_0^1 = \left[ \frac{5}{4} - x \right] = \frac{(19 - 4x)}{30}
\]

\[
iii \ F[x \mid y \leq 1] = \int_0^x \left( \frac{19 - 4x}{30} \right)dx = \left[ \frac{19x}{30} - \frac{x^2}{15} \right]
\]

Now,

\[
P\left[ X \leq \frac{1}{2} \mid Y \leq 1 \right] = F\left[ \frac{1}{2} \mid Y < 1 \right] = \left. \frac{18x}{15} - \frac{2x^2}{15} \right|_{x=\frac{1}{2}}
\]

\[
= \frac{18x}{2 \times 15} - \frac{2}{15(4)} = \frac{17}{30}
\]
Properties of RV

- **Population**: A complete assemblage of all possible RVs.
- **Sample**: A subset of population.
- **Realization**: A time series of the RVs actually realized. They are continuous. All samples are not realizations.
- **Observation**: A particular value of the RV in the realization.

*Note: ARMA is a synthetically generated realization*