STOCHASTIC HYDROLOGY

Lecture -8

Course Instructor : Prof. P. P. MUJUMDAR
Department of Civil Engg., IISc.

Summary of the previous lecture

- Extreme Value Distributions
  - Extreme Value Type-I Distribution
    (Gumbel's Extreme Value Distribution)
  - Extreme Value Type-III Minimum Distribution
    (Weibull's Distribution)
- Parameter estimation
  - Method of matching points
  - Method of moments
  - Method of maximum likelihood

Method of Maximum Likelihood

- The likelihood function is constructed as,

$L = f(x_1;\theta_1;\theta_2 \ldots\theta_m) \times f(x_2;\theta_1;\theta_2 \ldots\theta_m) \times f(x_n;\theta_1;\theta_2 \ldots \theta_m)$

$$= \prod_{i=1}^{n} f\left(x_i, \theta_1, \ldots\ldots\theta_m\right)$$

- Maximize the likelihood function

$$\frac{\partial L}{\partial \theta_i} = 0 \quad \forall\ i$$

- Solving the 'm' equations, the 'm' parameters are estimated

Example-1

Obtain the maximum likelihood estimates of the parameter 'β' in the pdf

$$f(x) = 2\beta\sqrt{\frac{\beta}{\pi}}\,x^2\,e^{-\beta x^2} \qquad -\infty < x < \infty$$

$$L(\beta) = 2\beta\sqrt{\frac{\beta}{\pi}}\,x_1^2\,e^{-\beta x_1^2} \times 2\beta\sqrt{\frac{\beta}{\pi}}\,x_2^2\,e^{-\beta x_2^2}\,.............2\beta\sqrt{\frac{\beta}{\pi}}\,x_n^2\,e^{-\beta x_n^2}$$

$$= 2^n\,\beta^n\left(\frac{\beta}{\pi}\right)^{n/2}\left(\prod_{i=1}^{n} x_i^2\right)e^{-\sum_{i=1}^{n}\beta x_i^2}$$

$$= 2^n\,\beta^{(n+n/2)}\pi^{-n/2}\left(\prod_{i=1}^{n} x_i^2\right)e^{-\sum_{i=1}^{n}\beta x_i^2}$$

Example-1 (contd.)

$$\ln L(\beta) = n \ln 2 + (n + n/2)\ln \beta - \frac{n}{2}\ln \pi + \ln\left(\prod_{i=1}^{n} x_i^2\right) - \beta \sum_{i=1}^{n} x_i^2$$

$$\frac{\partial \ln L(\beta)}{\partial \beta} = 0$$

$$(n + n/2)\frac{1}{\beta} - \sum_{i=1}^{n} x_i^2 = 0$$

$$\frac{3n}{2} = \sum_{i=1}^{n} x_i^2 \times \beta$$

$$\hat{\beta} = \frac{3n}{2\sum_{i=1}^{n} x_i^2}$$

Chebyshev Inequality

- Chebyshev inequality states that a single observation selected at random from any probability distribution will deviate more than $k\sigma$ from mean μ with a probability less than or equal to $1/k^2$.

$$P\left[\left|X-\mu\right| \geq k\sigma\right] \leq \frac{1}{k^2}$$

(places an upper bound on the probability for deviation from the mean.)

- Irrespective of probability distribution

Example-2

The mean annual stream flow of a river is 135 Mm^3 and standard deviation is 23.8 Mm^3.What is the maximum probability that the flow in a year will deviate more than 45 Mm^3 from the mean.

Applying Chebyshev inequality, $P\left[|X-\mu| \geq k\sigma\right] \leq \frac{1}{k^2}$

$k\sigma = 45$

$k \times 23.8 = 45$

$k = 1.891$

$$P\left[|X-\mu| \geq 45\right] = P\left[|X-\mu| \geq 1.891\sigma\right] \leq \frac{1}{k^2}$$

$\leq 1/1.891^2$

≤ 0.28

Moments and Expectation – Jointly Distributed Random Variables

$$\mu_n = \int_{-\infty}^{\infty} \left(x-\mu\right)^n f(x)dx$$ → n^{th} moment about mean

… Single dimensional RV

X and Y are jointly distributed random variables;

f(x,y) is joint pdf.

r, s^{th} moment of the two dimensional rv (X, Y) is

$$\mu_{r,s} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(x-\mu_x\right)^r \left(y-\mu_y\right)^s f(x,y)dx\,dy$$

Covariance

- Covariance of X and Y

$$\mu_{1,1} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x-\mu_x\right)\left(y-\mu_y\right) f(x,y)\,dx\,dy$$

$$= E\left[\left(x-\mu_x\right)\left(y-\mu_y\right)\right]$$

$$E\left[g(x,y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f(x,y)\,dx\,dy$$

- Also denoted as $\sigma_{X,Y}$ or Cov(X, Y)
- $\sigma_{X,Y}$ = Cov(X, Y) = 0, if X and Y are independent
- The converse may not be necessarily true
- Sample estimate for population covariance is given by

$$s_{X,Y} = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{n-1}$$

Correlation

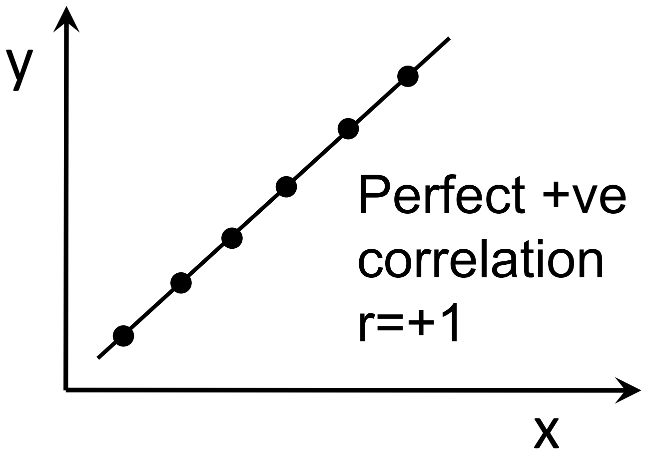
- Correlation is a measure of degree of association between two rvs X and Y

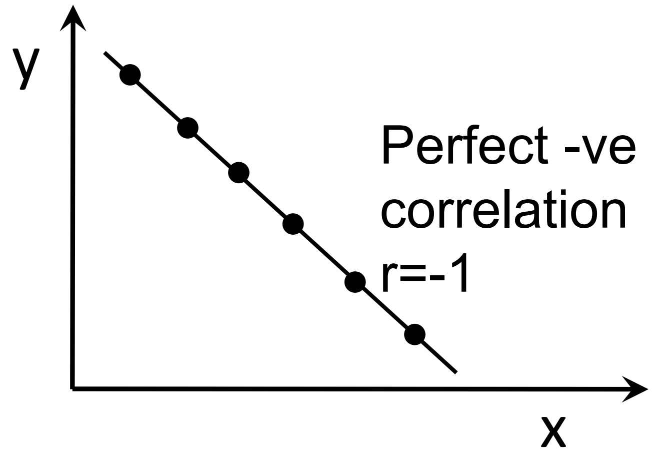
$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \qquad -1 \leq \rho_{X,Y} \leq 1$$

- Correlation is normalized covariance.
- $\rho_{X,Y} = 0$, if X and Y are independent
- The converse may is not necessarily true
- Sample estimate correlation coefficient is given by

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$$

Correlation Coefficient



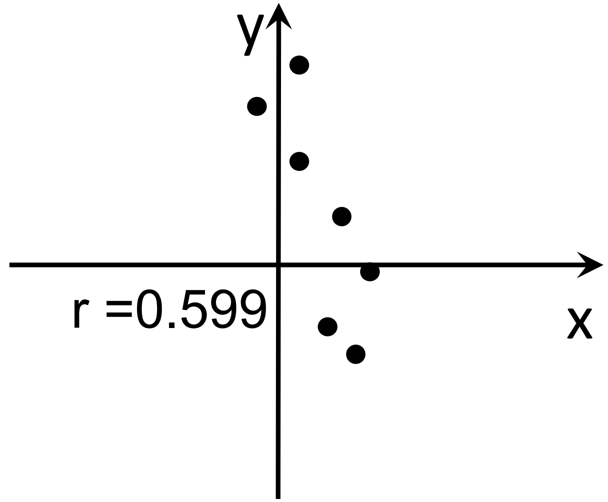


Y = aX+b Perfect linear relation between X and Y

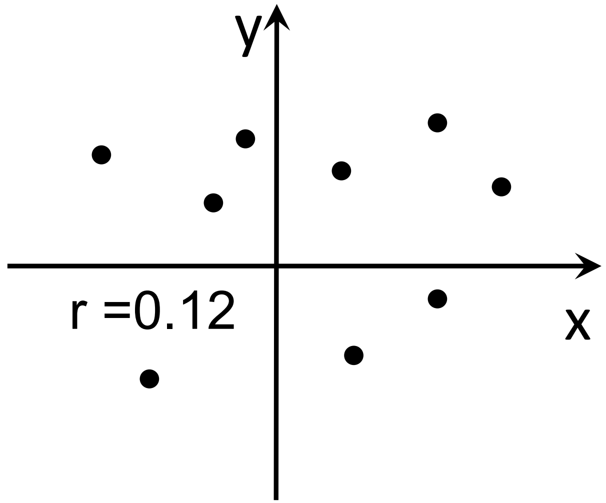
- $\rho_{X,Y}$ is +ve, larger values of X tend to be paired with larger values of Y and vice versa.
- $\rho_{X,Y}$ is -ve, larger values of X tend to be paired with smaller values of Y and vice versa

Correlation Coefficient

- r =0.599
- Points are scattered
- Existence of some stochastic dependence

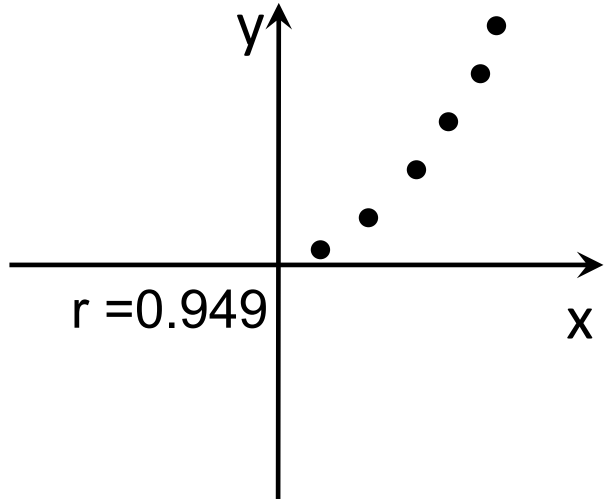


- r =0.12
- Points are scattered
- Lack of strong stochastic linear dependence

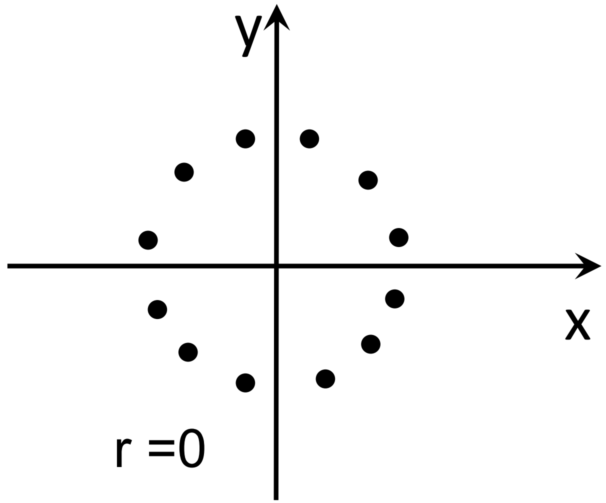


Correlation Coefficient

- r =0.949
- High degree of stochastic dependence
- Even though the dependence is non linear, a high correlation coefficient can result



- r =0
- Although X and Y are functionally related



Correlation Coefficient

$$\text{Cov}(X,Y) = E[(X-\mu_x)(Y- \mu_y)]$$

$$= E[XY – X\mu_y – Y\mu_x + \mu_x\mu_y]$$

$$= E[XY] – \mu_y E[X] – \mu_x E[Y] + \mu_x\mu_y$$

$$= E[XY] – 2\mu_x\mu_y + \mu_x\mu_y$$

$$= E[XY] – \mu_x\mu_y$$

$$= E[XY] – E[X]\, E[Y]$$

Correlation Coefficient

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

Consider Y = aX+b ;   perfect linear relation

$$\rho_{X,Y}^2 = \frac{\left(\sigma_{X,Y}\right)^2}{\sigma_X^2 \sigma_Y^2}$$

$$= \frac{\left(E\left[XY\right] - E\left[X\right]E\left[Y\right]\right)^2}{\sigma_X^2 \sigma_Y^2}$$

Substitute Y = aX+b

$$= \frac{\left(E\left[aX^2 + bX\right] - E\left[X\right]E\left[aX + b\right]\right)^2}{\sigma_X^2 \sigma_Y^2}$$

Correlation Coefficient

$$= \frac{\left(aE\left[X^2\right] + bE\left[X\right] - a\left\{E\left[X\right]\right\}^2 - bE\left[X\right]\right)^2}{\sigma_X^2 \sigma_Y^2}$$

$$= \frac{a^2\left(E\left[X^2\right] - \left\{E\left[X\right]\right\}^2\right)^2}{\sigma_X^2 \sigma_Y^2}$$

$$= \frac{a^2\left(\sigma_X^2\right)^2}{\sigma_X^2 a^2 \sigma_X^2} = 1 \qquad \left(\because\ \sigma_Y^2 = a^2 \sigma_X^2\right)$$

$\rho = \pm 1$ if there is a perfect relationship in between X and Y

Correlation coefficient is a measure of linear dependence

Example-3

Obtain the correlation coefficient for the yearly rainfall and the yearly runoff of a catchment for 15 years.

Year	1	2	3	4	5	6	7	8	9	10
Rainfall (cm)	105	115	103	94	95	104	120	121	127	79
Runoff (cm)	42	46	26	39	29	33	48	58	45	20

Year	11	12	13	14	15
Rainfall (cm)	133	111	127	108	85
Runoff (cm)	54	37	39	34	25

Example-3 (contd.)

Mean, $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

$\sum_{i=1}^{n} x_i$ = 1627

Therefore mean, $\bar{x}$ = 1627/15

= 108.5 cm

Variance, $s_x^2 = \dfrac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1} = \dfrac{3499.73}{15-1} = 250$

Standard deviation, s_x = 15.811 cm

Example-3 (contd.)

Mean, $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$

$\sum_{i=1}^{n} y_i$ = 575

Therefore mean, $\bar{y}$ = 575/15

= 38.33 cm

Variance, $s_y^2 = \dfrac{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}{n-1} = \dfrac{1645.33}{15-1} = 117.5$

Standard deviation, s_y = 10.841 cm

Year	Rainfall cm (x_i)	Runoff cm (y_i)	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x}) \times (y_i - \overline{y})$
1	105	42	-3.47	3.67	12.02	13.44	-12.71
2	115	46	6.53	7.67	42.68	58.78	50.09
3	103	26	-5.47	-12.33	29.88	152.11	67.42
4	94	39	-14.47	0.67	209.28	0.44	-9.64
5	95	29	-13.47	-9.33	181.35	87.11	125.69
6	104	33	-4.47	-5.33	19.95	28.44	23.82
7	120	48	11.53	9.67	133.02	93.44	111.49
8	121	58	12.53	19.67	157.08	386.78	246.49
9	127	45	18.53	6.67	343.48	44.44	123.56
10	79	20	-29.47	-18.33	868.28	336.11	540.22
11	133	54	24.53	15.67	601.88	245.44	384.36
12	111	37	2.53	-1.33	6.42	1.78	-3.38
13	127	39	18.53	0.67	343.48	0.44	12.36
14	108	34	-0.47	-4.33	0.22	18.78	2.02
15	85	25	-23.47	-13.33	550.68	177.78	312.89
Σ	1627	575	0	0	3499.73	1645.33	1974.67

Example-3 (contd.)

$$s_{X,Y} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{n-1}$$

$$= \frac{1974.67}{15-1}$$

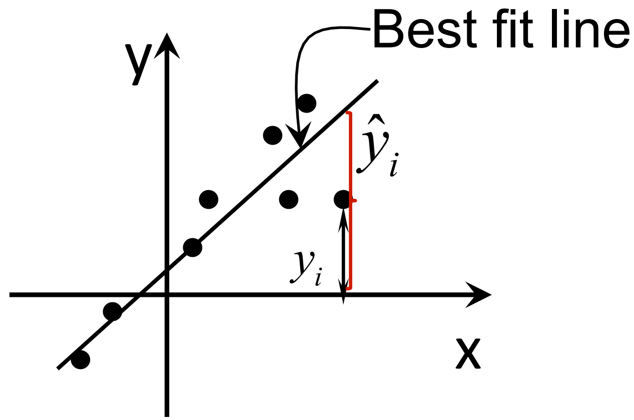
$$= 141.05$$

Correlation coefficient, $r_{X,Y} = \dfrac{s_{X,Y}}{s_X s_Y}$

$$= \frac{141.05}{15.811 \times 10.841}$$

$$= 0.823$$

Simple Linear Regression



(x_i, y_i) are observed values

$\hat{y}_i$ is predicted value of x_i

$$\hat{y}_i = a + bx_i$$

Error, $e_i = y_i - \hat{y}_i$

Estimate the parameters a, b such that the square error is minimum

Sum of square errors $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$

$$M = \sum_{i=1}^{n} \left\{ y_i - \left(a + bx_i \right) \right\}^2$$

Simple Linear Regression

$$M = \sum_{i=1}^{n} \left\{ y_i - a - bx_i \right\}^2$$

$$\frac{\partial M}{\partial a} = 0 \qquad -2\sum_{i=1}^{n} \left\{ y_i - a - bx_i \right\} = 0$$

$$\sum_{i=1}^{n} \left\{ y_i - a - bx_i \right\} = 0$$

$$\sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

$$a = \overline{y} - b\overline{x}$$

Simple Linear Regression

$$\frac{\partial M}{\partial b} = 0 \qquad -2\sum_{i=1}^{n} x_i \left\{ y_i - a - bx_i \right\} = 0$$

$$\sum x_i y_i - a\sum x_i - b\sum x_i^2 = 0$$

$$\sum x_i y_i - \frac{\sum y_i - b\sum x_i}{n} \sum x_i - b\sum x_i^2 = 0$$

$$b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}$$

$$Let \quad \left(x_i - \overline{x} \right) = x_i^{'} \quad and \quad \left(y_i - \overline{y} \right) = y_i^{'}$$

Simple Linear Regression

$$\sum x_i' y_i' = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \overline{xy})$$

$$= \sum \left(x_i y_i - \frac{\sum y_i}{n} x_i - \frac{\sum x_i}{n} y_i + \frac{\sum x_i}{n} \frac{\sum y_i}{n} \right)$$

$$= \sum x_i y_i - \frac{\sum y_i}{n} \sum x_i - \frac{\sum x_i}{n} \sum y_i + n \left(\frac{\sum x_i \sum y_i}{n^2} \right)$$

$$= \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} - \frac{\sum x_i \sum y_i}{n} + \frac{\sum x_i \sum y_i}{n}$$

$$= \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

Simple Linear Regression

$$\sum\left(x_i^{'}\right)^2 = \sum\left(x_i - \overline{x}\right)^2 = \sum\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right)$$

$$= \sum\left(x_i^2 - 2\frac{\sum x_i}{n}x_i + \left\{\frac{\sum x_i}{n}\right\}^2\right)$$

$$= \sum x_i^2 - 2\frac{\sum x_i}{n}\sum x_i + n\left\{\frac{\sum x_i}{n}\right\}^2$$

$$= \sum x_i^2 - 2\frac{\left(\sum x_i\right)^2}{n} + \frac{\left(\sum x_i\right)^2}{n}$$

$$= \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Simple Linear Regression

$$b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}$$

$$= \frac{\sum \left(x_i - \overline{x} \right)\left(y_i - \overline{y} \right)}{\sum \left(x_i - \overline{x} \right)^2}$$

$$= \frac{\sum x_i^{'} y_i^{'}}{\sum \left(x_i^{'} \right)^2}$$

Example-4

Consider the previous example, obtain the regression equation between rainfall (X) and runoff (Y)

$$\bar{x} = 108.5$$

$$\bar{y} = 38.33$$

$$\sum x_i^{'} y_i^{'} = 1974.67$$

$$\sum \left(x_i^{'} \right)^2 = 3499.73$$

$$b = \frac{\sum x_i^{'} y_i^{'}}{\sum \left(x_i^{'} \right)^2} = \frac{1974.67}{3499.73} = 0.564235$$

$$a = \bar{y} - b\bar{x} = 38.33 - (0.564235 \times 108.5) = -22.8895$$

Therefore the equation is

$$Y = 0.564235X - 22.8895$$

Year	Rainfall cm (x_i)	Runoff cm (y_i)	$\left(x_i-\overline{x}\right)$	$\left(y_i-\overline{y}\right)$	$\left(x_i-\overline{x}\right)^2$	$\left(y_i-\overline{y}\right)^2$	$\left(x_i-\overline{x}\right)\times\left(y_i-\overline{y}\right)$
1	105	42	-3.47	3.67	12.02	13.44	-12.71
2	115	46	6.53	7.67	42.68	58.78	50.09
3	103	26	-5.47	-12.33	29.88	152.11	67.42
4	94	39	-14.47	0.67	209.28	0.44	-9.64
5	95	29	-13.47	-9.33	181.35	87.11	125.69
6	104	33	-4.47	-5.33	19.95	28.44	23.82
7	120	48	11.53	9.67	133.02	93.44	111.49
8	121	58	12.53	19.67	157.08	386.78	246.49
9	127	45	18.53	6.67	343.48	44.44	123.56
10	79	20	-29.47	-18.33	868.28	336.11	540.22
11	133	54	24.53	15.67	601.88	245.44	384.36
12	111	37	2.53	-1.33	6.42	1.78	-3.38
13	127	39	18.53	0.67	343.48	0.44	12.36
14	108	34	-0.47	-4.33	0.22	18.78	2.02
15	85	25	-23.47	-13.33	550.68	177.78	312.89
Σ	1627	575	0	0	3499.73	1645.33	1974.67

DATA GENERATION

Data Generation

- Necessity :

Length of Historical Record

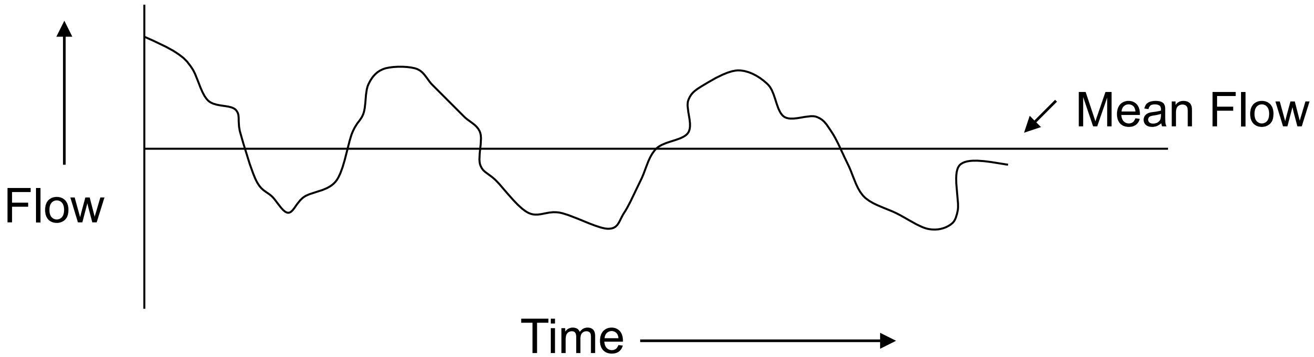
1. Economic Life of the Project

2. Use of Historical Record alone gives no idea of the Risks involved.

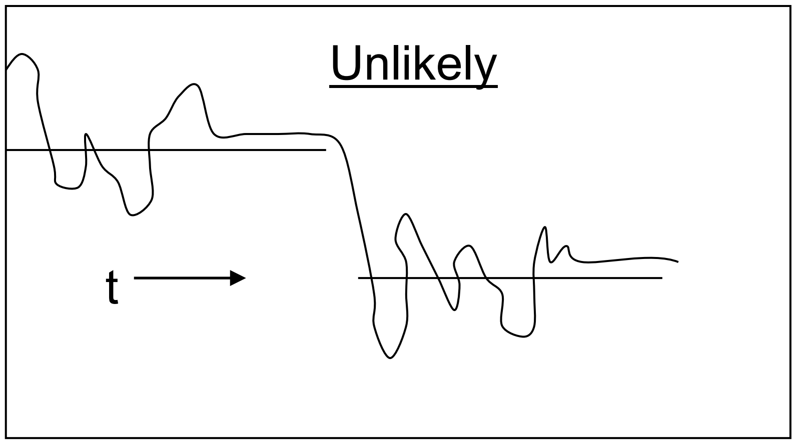
3. Exact Pattern of flows during the Historical Period is extremely unlikely to recur during the Economic Life of the system.

• Motivation for the Generating Models :

• Statistical Regularity of Flows :



Unless drastic changes in the Basin occur, flow tend to maintain their Statistical Distributions over a long period of time.



History provides a valuable clue to the future

- Persistence

  Tendency of the flows to follow the trend of Immediate Past.

  [Low flows follow low flows and high flows follow high flows].

Generating Models: Reproduce the Statistical Distributions and Persistence of Historical Flows

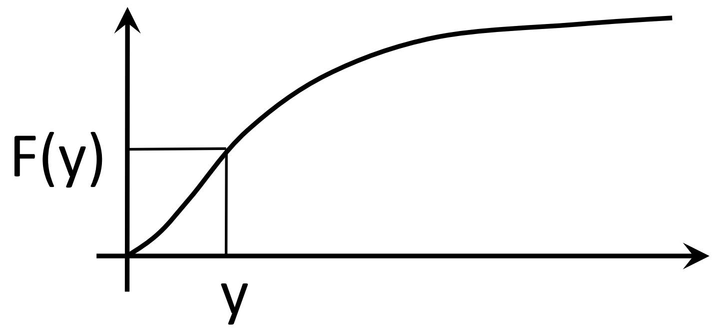
Important Statistics Preserved by Generating Models :

- Mean ………………………. Average Flow
- Std. Deviation……………… Variability of Flows
- Correlation Coefficient……. Dependence on Previous Flows and/ or other Hydrologic Variables (Rainfall)

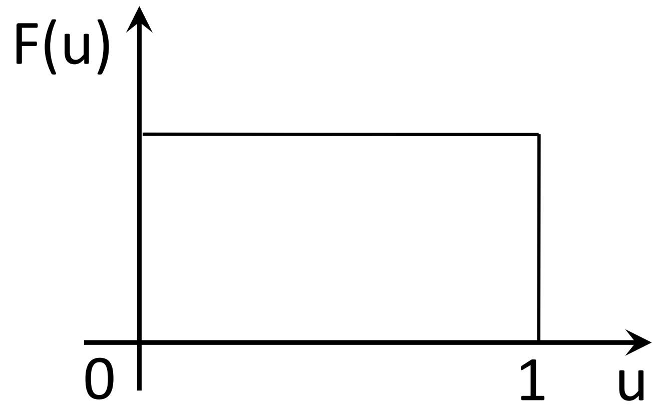
Data Generation

- Given a distribution, generate data belonging to that distribution

Randomly picked up values of F(y) follow a uniform distribution u(0, 1)



Choose a random F(y) from uniform distribution, get corresponding y.



Data Generation

$$F\left(y\right) = \int_{-\infty}^{y} f(y)dy$$

$$F\left(y\right) = R_u = \int_{-\infty}^{y} f(y)dy$$

R_u: uniformly distributed random no.s in the interval (0,1)

Most scientific programs have built-in functions for generating random numbers.

Data Generation

Algorithm for random number (R_u) generation:

X_i=(a+bX_{i-1}) Modulo M

{X_i^s/M} are the random numbers

m Modulo n
Remainder of (m/n)

For e.g., M = 10, a = 5, b = 3

Let X_0 = 2, then X_1 = (3*2+5) Modulo 10

= 11 Modulo 10

= 1

X_1 = (3*1+5) Modulo 10

= 8 Modulo 10

= 8

Data Generation

X_2 = (3*8+5) Modulo 10
= 29 Modulo 10
= 9

X_3 = (3*9+5) Modulo 10
= 32 Modulo 10
= 2

The random numbers are $\frac{2}{10}, \frac{1}{10}; \frac{8}{10}; \frac{9}{10}; \frac{2}{10}$..........

Pseudo random numbers:

If M is very large, then the repetition of numbers occur after a very large set is generated.

Data Generation

Exponential distribution:

$f(y) = \lambda e^{-\lambda y}$ $\lambda > 0$

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

$R_u = 1 - e^{-\lambda y}$

$1 - R_u = e^{-\lambda y}$

Since R_u is a random number, $1 - R_u$ is also a random number.

$R_u = e^{-\lambda y}$

$\ln R_u = -\lambda y$; $y = -\dfrac{\ln R_u}{\lambda}$

Example-5

Generate 10 values from exponential distribution with $\lambda = 5$

S.No.	R_u	y
1	0.026	0.729932
2	0.85	0.032504
3	0.654	0.08493
4	0.805	0.043383
5	0.205	0.316949
6	0.957	0.00879
7	0.035	0.670481
8	0.285	0.251053
9	0.996	0.000802
10	0.549	0.119931
Σ		2.258755

$$y = -\frac{\ln R_u}{\lambda}$$

$$\bar{x} = \frac{2.26}{10} = 0.226$$ … generated values

$$\hat{\lambda} = \frac{1}{\bar{x}} = \frac{1}{0.226} = 4.43$$

Data Generation

- Analytic inverse transform not possible for some distributions (eg., Normal distribution, Gamma distribution)
- Numerically generated tables of standard normal deviate available
- Given R_N, a random no. belonging to standard normal distribution,

$$y = \sigma R_N + \mu$$

- Most scientific programs have built-in functions to generate standard normal deviates.

Example-6

Generate 10 values from $N(10, 15^2)$

S.No.	R_N	y	$(y-\bar{y})^2$
1	0.335	15.025	0.02434
2	-0.051	9.235	31.742
3	1.226	28.39	182.82
4	-0.642	0.37	210.221
5	0.377	15.655	0.618
6	2.156	42.34	754.66
7	0.667	20.005	26.3785
8	-1.171	-7.565	503.284
9	0.28	14.2	0.4476
10	0.069	11.035	14.6996
Σ		148.69	1724.9

$y = \sigma R_N + \mu$

$y = 15\, R_N + 10$

$$\hat{\mu} = \bar{y} = \frac{148.69}{10} = 14.869$$

$$\hat{\sigma}^2 = \frac{1724.9}{10-1} = 191.65$$

$$\hat{\sigma} = 13.8$$

R_N obtained from: Statistical methods in Hydrology by C.T.Haan Iowa State University Press 1994 Table No.-E.11

Data Generation

Gamma Distribution:

$$f(x) = \frac{\lambda^n x^{\eta-1} e^{-\lambda x}}{\Gamma(\eta)} \qquad x, \lambda, \eta > 0$$

$$y = \frac{-\sum_{i=1}^{\eta} \ln R_{u_i}}{\lambda}$$ (for integer values of η)

For e.g., η = 2

$$y = \frac{-\sum_{i=1}^{2} \ln R_{u_i}}{\lambda} = \frac{-\left(\ln R_{u_1} + \ln R_{u_2}\right)}{\lambda}$$

Example-7

Generate 10 values for η = 2 and λ = 3

S.No.	R_{u_1}	R_{u_2}	y
1	0.376	0.005	2.092
2	0.077	0.959	0.869
3	0.323	0.216	0.888
4	0.773	0.544	0.289
5	0.24	0.073	1.348
6	0.597	0.631	0.325
7	0.879	0.614	0.206
8	0.942	0.563	0.211
9	0.213	0.48	0.76
10	0.325	0.112	1.104
Σ			8.092

$$y = \frac{-\sum_{i=1}^{\eta} \ln R_{u_i}}{\lambda}$$

$$y = \frac{-\left(\ln R_{u_1} + \ln R_{u_2}\right)}{\lambda}$$

$$\bar{y} = \frac{8.092}{10} = 0.8092$$

$$\hat{\eta} = 1.95$$

$$\hat{\lambda} = 2.41$$