INDIAN INSTITUTE OF SCIENCE

STOCHASTIC HYDROLOGY

Lecture -9

Course Instructor : Prof. P. P. MUJUMDAR

Department of Civil Engg., IISc.

Summary of the previous lecture

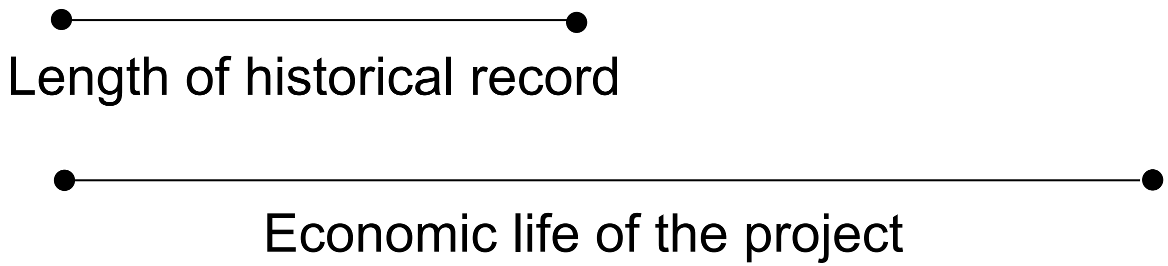
- Parameter estimation
  - Method of maximum likelihood
- Correlation coefficient
- Simple Linear Regression

DATA GENERATION

Data Generation

Necessity :

1.



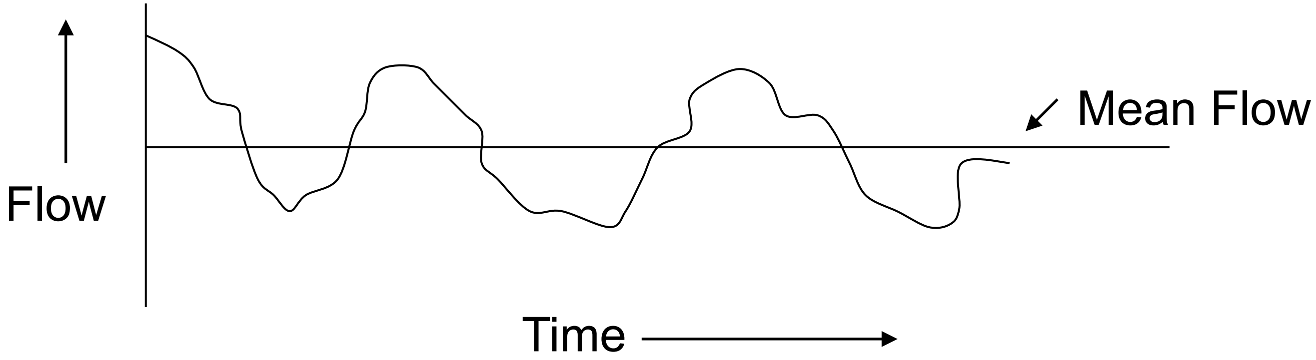
3. Use of historical record alone gives no idea of the risks involved.

4. Exact pattern of flows during the historical period is extremely unlikely to recur during the economic life of the system.

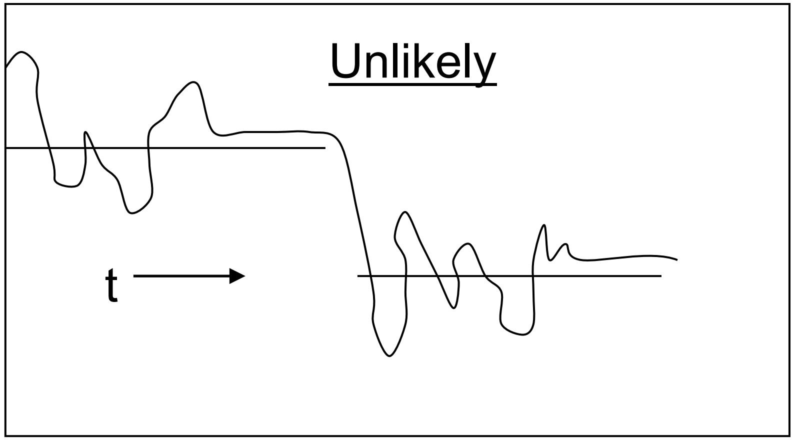
Data Generation

• 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

Data Generation

- 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

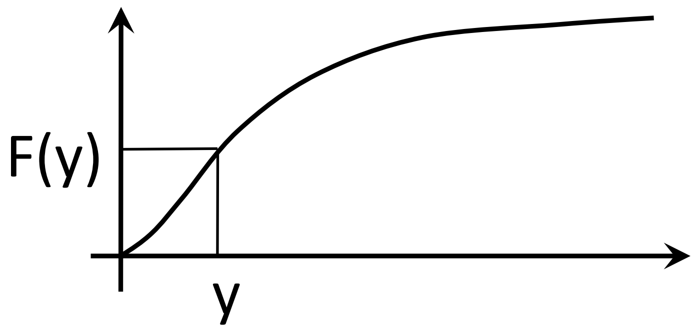
Important statistics normally preserved by generating models :

- Mean ………………………. Average flow
- Std. Deviation……………… Variability of flows
- Correlation Coefficient……. Dependence on previous flows and/ or other hydrologic variables (Rainfall)

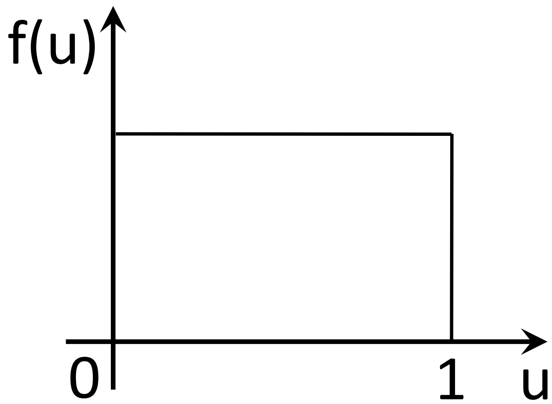
Data Generation

- Given a distribution, to 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 uniformly distributed random numbers.

Data Generation

An algorithm for random number (R_u) generation:

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

$\{X_i/M\}$ are the required random numbers

m Modulo n =
Remainder of (m/n)

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

Data Generation

Exponential distribution:

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

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

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

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

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

$\ln R_u = -\lambda y$

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

Example-1

Generate 10 values from exponential distribution with λ = 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

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

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

$$\hat{\lambda} = \frac{1}{\bar{y}}$$

$$= \frac{1}{0.226}$$

$$= 4.43$$

Data Generation

- Analytic inverse transform not possible for some distributions (eg., Normal distribution, Gamma distribution)
- Numerically generated tables of standard normal deviates (R_N) available
- Given R_N, data is generated by

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

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

Example-2

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

S.No.	R_N	y
1	0.335	15.025
2	-0.051	9.235
3	1.226	28.39
4	-0.642	0.37
5	0.377	15.655
6	2.156	42.34
7	0.667	20.005
8	-1.171	-7.565
9	0.28	14.2
10	0.069	11.035
Σ		148.69

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

$$y = 15\, R_N + 10$$

$$\hat{\mu} = \overline{y} = 14.869$$

$$\hat{\sigma}^2 = 191.65$$

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

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:

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

Gamma variate with integer values of η can be shown to be the sum of η exponential variates each with parameter λ)

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

e.g., $\eta = 2$

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

Example-3

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

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

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

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

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

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

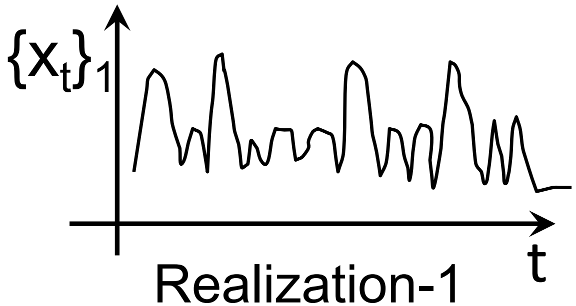
TIME SERIES ANALYSIS

Time Series Analysis

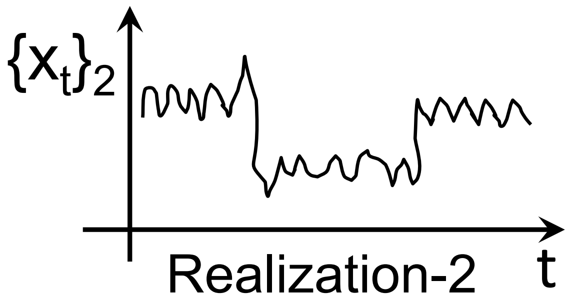
- Sequence of values of a random variable collected over time is time series.
- Discrete time series: measured at discrete time intervals
- Continuous time series: recorded continuously with time
- Single time series : A realization
- Ensemble: collection of all realizations

  $\{x_t\}_1, \{x_t\}_2 \ldots\ldots \{x_t\}_m$

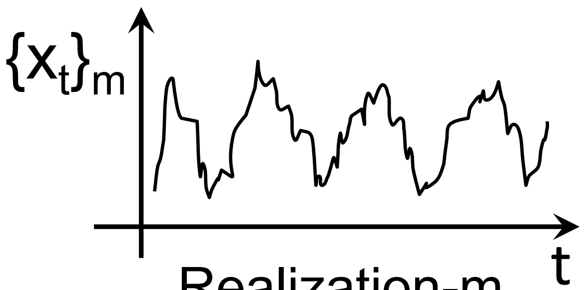
Time Series Analysis



Realization-1



Realization-2

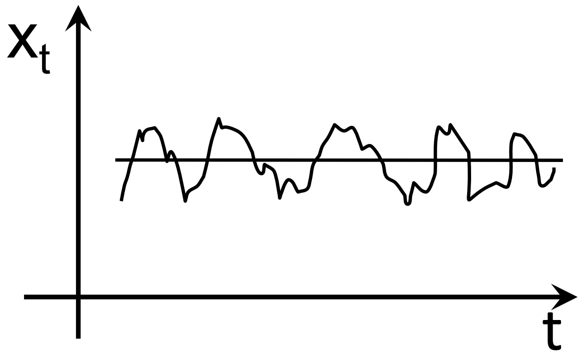


Realization-m

Ensemble:
collection of all realizations
$\{x_t\}_1, \{x_t\}_2 \ldots\ldots \{x_t\}_m$

Time Series Analysis

- Hydrologic time series composed of deterministic and stochastic components

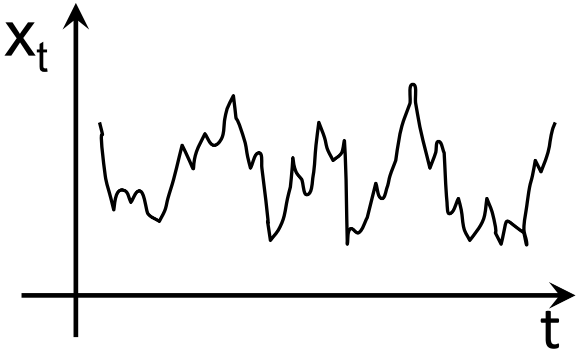


Long term mean

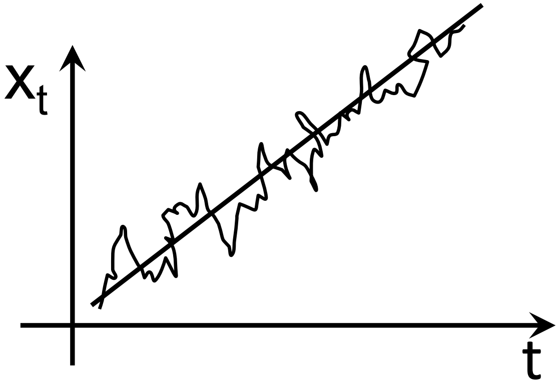
Time Series plot

$$X_t = d_t + \varepsilon_t$$

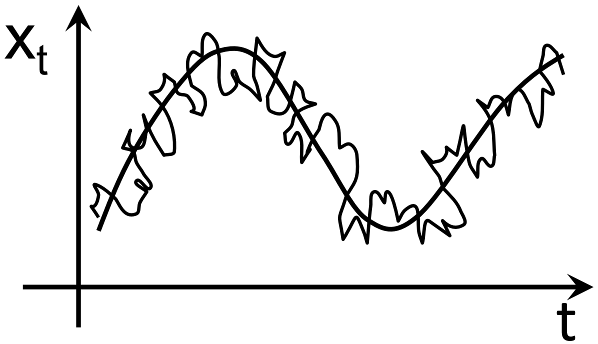
Time Series Analysis



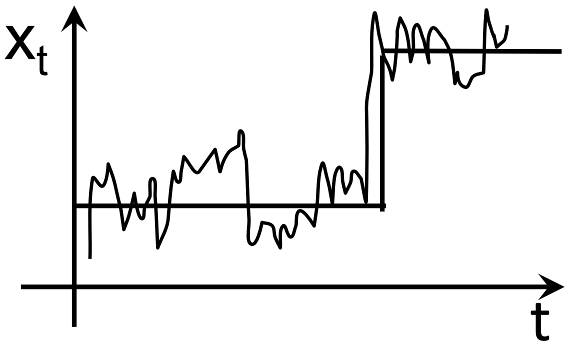
Stochastic



Stochastic + Trend



Stochastic + Periodic



Stochastic + Jump

$$X_t = d_t + \varepsilon_t$$

Time Series Analysis

- Deterministic component is a combination of a long term mean, trend, periodicity and jump.

- Time scale of time series – either discrete or continuous
- Discrete time scale: observations at specific times separated by Δt. (eg., average monthly stream flow, annual peak discharge, daily rainfall etc.)
- Continuous time scale: data recorded continuously with time (eg., turbulence studies, pressure measurements)

Time Series Analysis

- The pdf of a stochastic process $X(t)$ is $f(x; t)$
- $f(x; t)$ describes the probabilistic behavior of $X(t)$ at specified time 't'
- The time series is said to be stationary, if the properties do not change with time.
- $f(x; t) = f(x; t+\tau) \forall t$
- for stationary series, pdf of X_t is same as $X_{t+\tau}$

Time Series Analysis

Time average for a realization

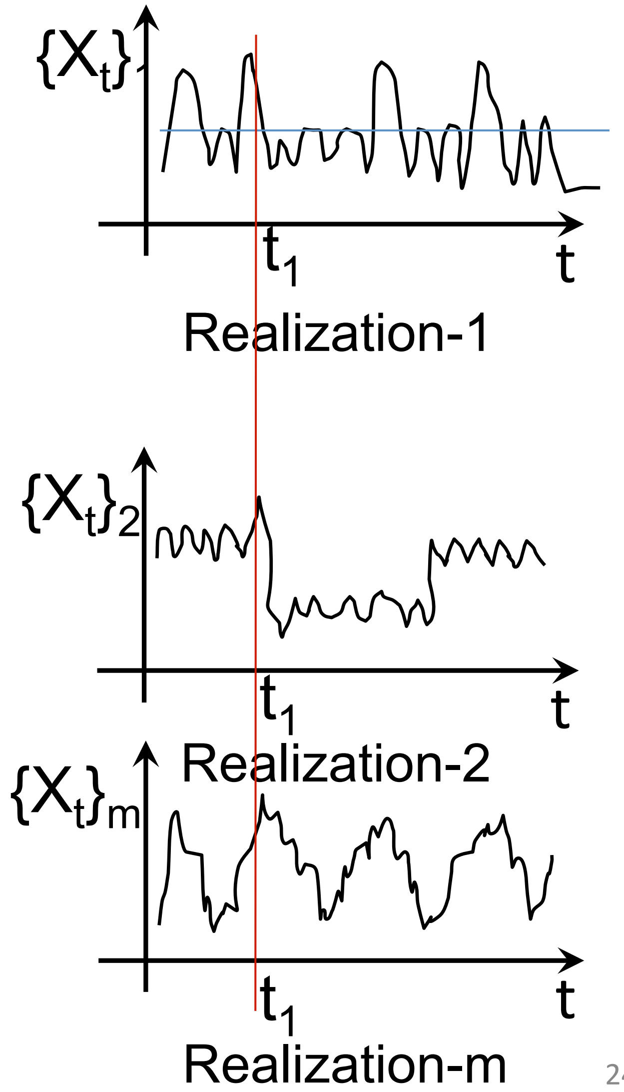
$$\bar{X}_1 = \frac{\sum_{j=1}^{n} \left\{ X_j(t) \right\}_1}{n}$$

n is no. of observations

Ensemble average at time t

$$\bar{X}_t = \frac{\sum_{i=1}^{m} X_i(t)}{m}$$

m is no. of realizations



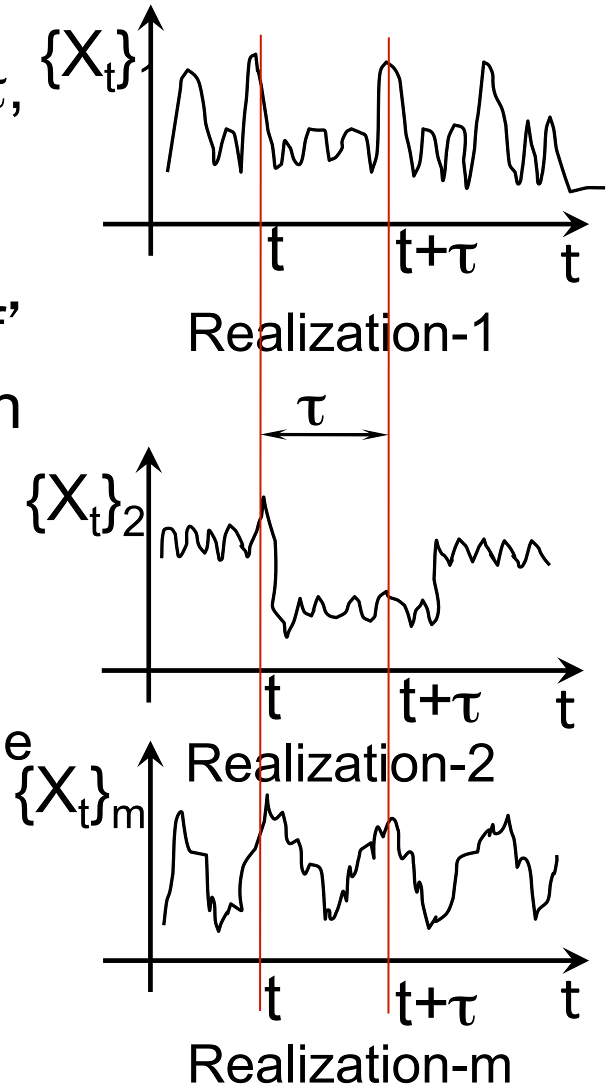
Time Series Analysis

- $\bar{X}_t = \bar{X}_{t+\tau}$ for all values of t and τ, then the process is stationary in mean (first order stationary)
- If all the moments up to order 'f' are same for time t and $t+\tau$, then the time series is weakly stationary of order 'k'

  k = 1 Stationary in mean

  k = 2 mean & covariance are same
- For a strictly stationary time series,

  $f(x_1) = f(x_2) = \dots\dots\dots = f(x)$



Time Series Analysis

- Auto covariance

$$\gamma_k = \text{cov}\left(X_t, X_{t+k}\right)$$
$$= E\left[\left(X_t - \mu\right)\left(X_{t+k} - \mu\right)\right]$$
$$\gamma_0 = \sigma_X^2$$

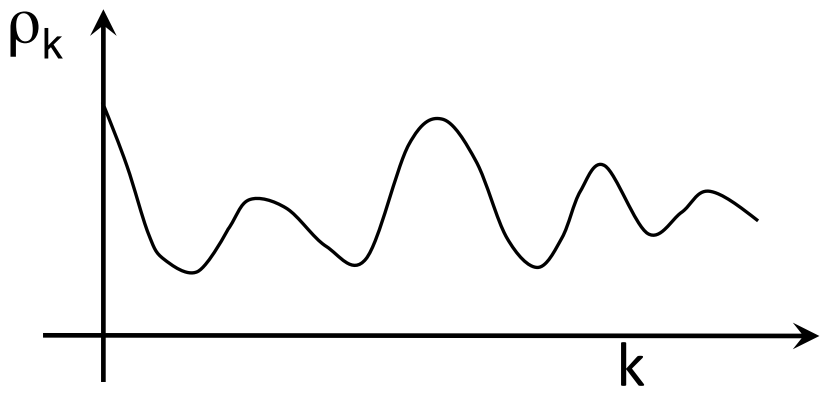
- Auto correlation between X_t and $X_{t+\tau}$,

$$\rho_k = \frac{\text{cov}\left(X_t, X_{t+k}\right)}{\sigma_{X_t}\sigma_{X_{t+k}}}$$

If process is stationary
$\sigma_{X_t} = \sigma_{X_{t+k}}$

$$= \frac{\text{cov}\left(X_t, X_{t+k}\right)}{\sigma_X^2} = \frac{\gamma_k}{\gamma_0}$$
$$\rho_0 = 1$$

Time Series Analysis



Correlogram

- Auto correlation indicates the memory of a stochastic process

Time Series Analysis

- Auto covariance matrix

$$
\Gamma_n = \begin{array}{c c}
 & \begin{array}{cccccc} X_1 & X_2 & X_3 & . & . & X_n \end{array} \\
\begin{array}{c} X_1 \\ X_2 \\ X_3 \\ . \\ . \\ X_n \end{array} &
\begin{bmatrix}
\gamma_0 & \gamma_1 & \gamma_2 & \cdot & \cdot & \gamma_{n-1} \\
\gamma_1 & \gamma_0 & \gamma_1 & \cdot & \cdot & \gamma_{n-2} \\
\gamma_2 & & & & & \\
\cdot & & & & & \\
\cdot & & & & & \\
\gamma_{n-1} & \gamma_{n-1} & & & & \gamma_0
\end{bmatrix}
\end{array}
$$

Γ_n is symmetric and +ve definite matrix

Time Series Analysis

- Dividing the matrix by γ_o,

$$\mathrm{P}_n = \frac{\Gamma_n}{\gamma_0} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & . & . & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & . & . & \rho_{n-2} \\ \rho_2 & & & & & \\ . & & & & & \\ . & & & & & \\ \rho_{n-1} & \rho_{n-2} & . & . & . & 1 \end{bmatrix}$$

P_n is symmetric and +ve definite matrix

Time Series Analysis

- Because P_n is +ve definite

$$\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix} \geq 0$$

$$1 - \rho_1^2 \geq 0$$

$$-1 \leq \rho_1 \leq 1$$

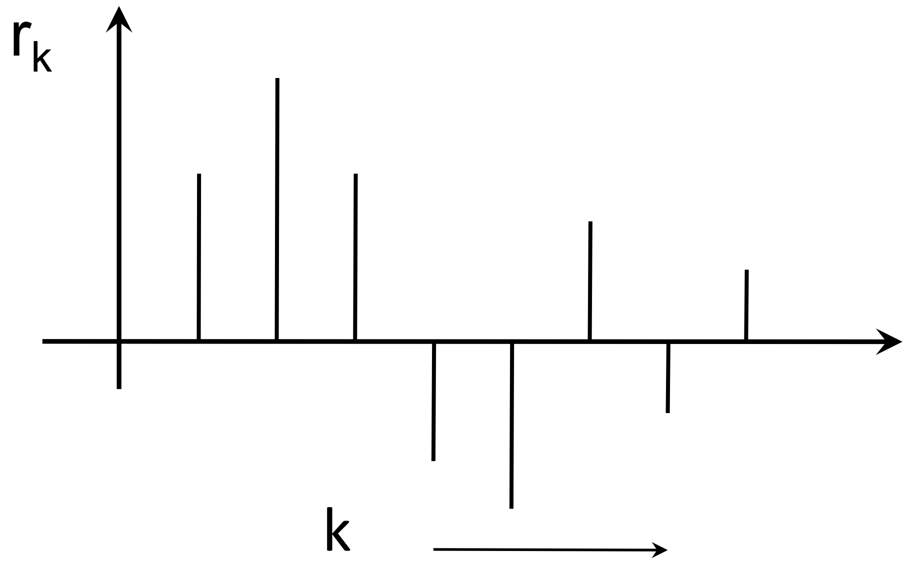
Time Series Analysis

- Sample estimates:

$$r_k = E\left[\left(X_t - \mu\right)\left(X_{t+k} - \mu\right)\right]$$

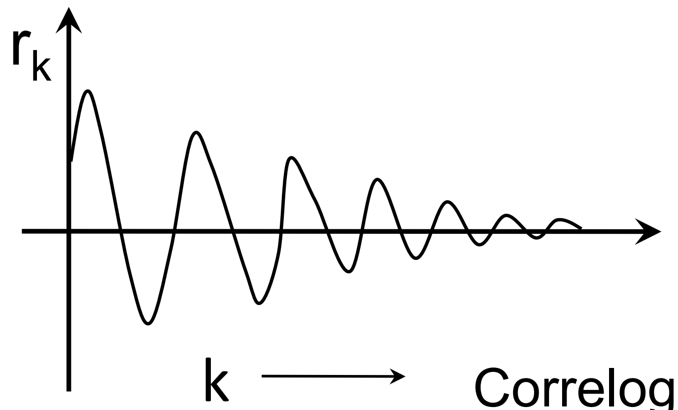
$$c_k = \frac{1}{N}\sum_{i=1}^{n-k}\left(X_t - \bar{X}\right)\left(X_{t+k} - \bar{X}\right)$$

$$r_k = \frac{c_k}{c_0}$$



Time Series Analysis

- Auto correlation function (r_k)



Correlogram

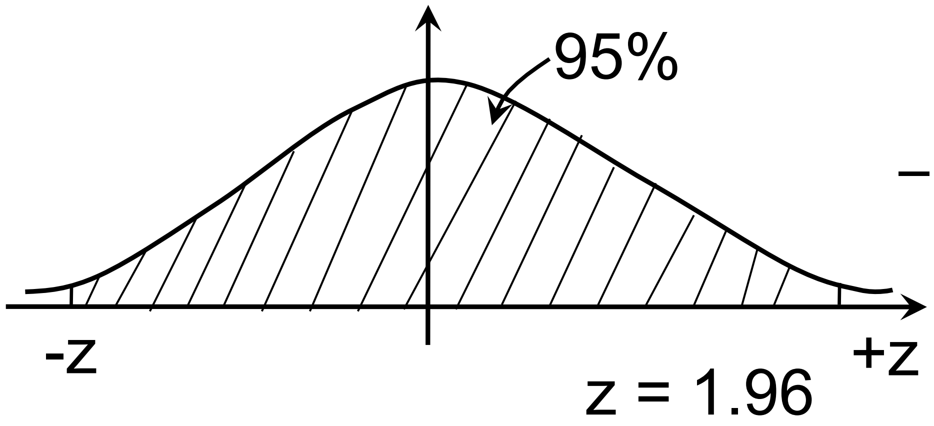
If it is purely stochastic (random) series,

$\rho_k = 0, \quad \forall \quad k = 1, 2, 3, \ldots\ldots..$

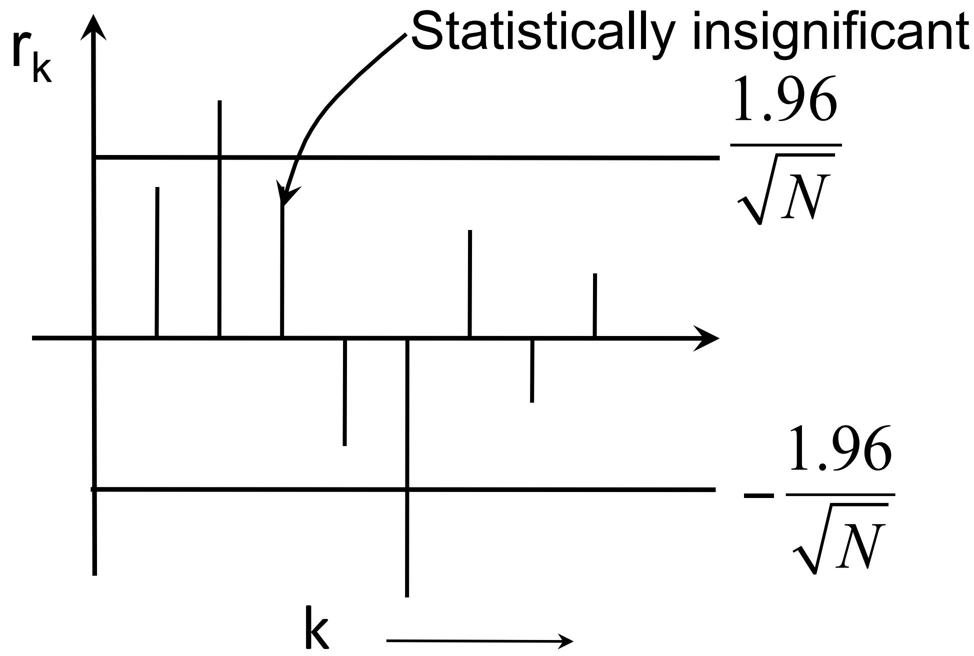
r_k = may not be zero (because r_k is a sample estimate)

$$r_k : \textit{Normal Distribution}\left(0, \frac{1}{\sqrt{N}}\right)$$

Time Series Analysis



$$-\frac{1.96}{\sqrt{N}} \le r_k \le +\frac{1.96}{\sqrt{N}}$$



Example-4

Obtain Auto correlation for k=1

S.No.	X_t	$(x_t - \overline{x})$	X_{t+1}	$(x_{t+1} - \overline{x})$	$(x_t - \overline{x}) \times (x_{t+1} - \overline{x})$
1	97	-10.50	110	2.5	-26.25
2	110	2.50	121	13.5	33.75
3	121	13.50	117	9.5	128.25
4	117	9.50	79	-28.5	-270.75
5	79	-28.50	140	32.5	-926.25
6	140	32.50	75	-32.5	-1056.25
7	75	-32.50	127	19.5	-633.75
8	127	19.50	90	-17.5	-341.25
9	90	-17.50	119	11.5	-201.25
10	119	11.50			
Σ	1075				-3293.75

Example-4 (contd.)

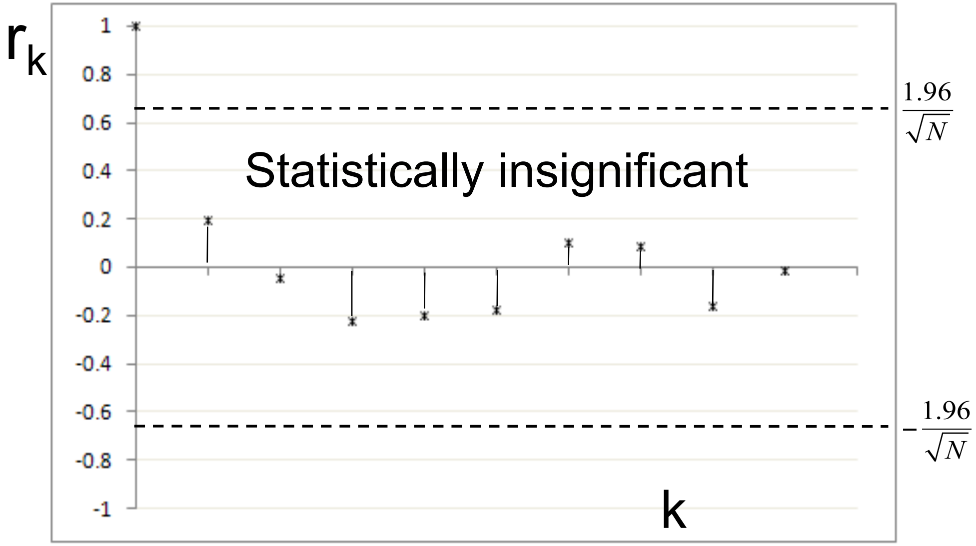
mean $\overline{x}$ = 1075/10

= 107.5

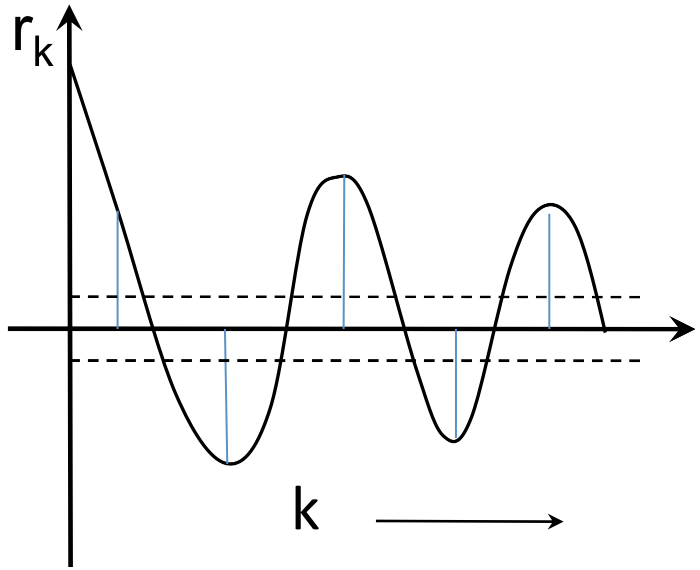
Variance, $c_0 = \dfrac{\sum_{t=1}^{n}\left(x_t - \overline{x}\right)^2}{n-1} = \dfrac{4132.5}{10-1} = 459.2$

$$c_1 = \frac{\sum_{t=1}^{n-1}\left(x_t - \overline{x}\right)\left(x_{t+1} - \overline{x}\right)}{n} = \frac{3293.75}{10} = 329.375$$

$$r_1 = \frac{c_1}{c_0} = \frac{329.375}{459.2} = 0.72$$



Purely stochastic process

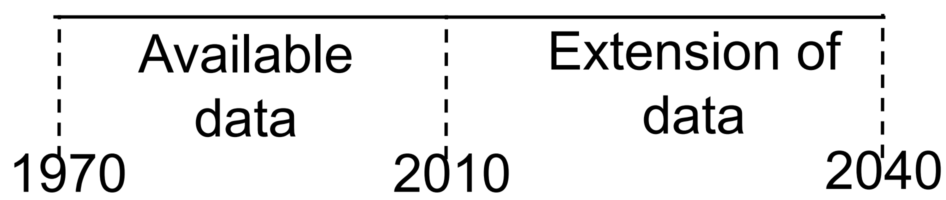


Periodic process

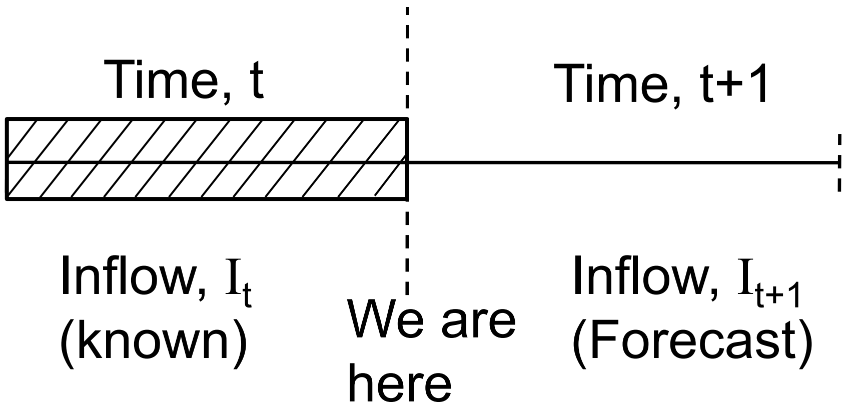
DATA EXTENSION & FORECASTING

Data Extension & Forecasting

e.g., Stream flow records for reservoir planning

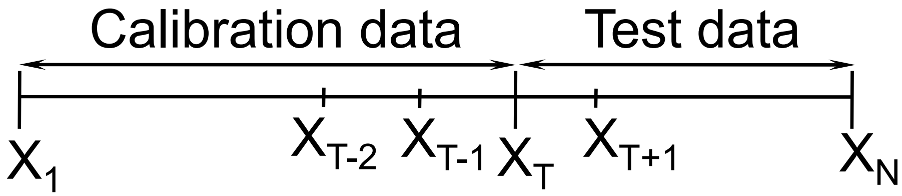


Data forecasting



$$\hat{I}_{t+1} = f\left[I_t, I_{t-1}, \ldots\ldots\ldots\ldots\right] + \text{Random component}$$

Data Extension & Forecasting



Use first 'T' values to build the model, use rest of data to validate it

F_{T+1} F_{T+2} ……… F_N forecasts obtained from the model

$$\left.\begin{array}{l} (X_{T+1} - F_{T+1}) \\ (X_{T+2} - F_{T+2}) \\ . \\ . \\ (X_N - F_N) \end{array}\right\} \text{Forecast errors}$$

Data Extension & Forecasting

Method of simple averages: take the average of all the data in the calibration data as the forecast for period (T+1)

$$\hat{X}_{t+1} = F_{T+1} = \frac{\sum_{t=1}^{T} X_t}{T}$$

$$\hat{X}_{t+2} = F_{T+2} = \frac{\sum_{t=1}^{T+1} X_t}{T+1}$$ and so on

For jumps, trends this is not a good procedure

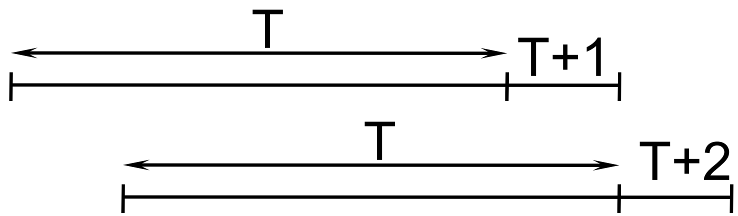
Example-6

Data	Forecast
105	-
115	110
103	107.67
108	107.75
120	110.2
97	108
110	108.28
121	109.87
117	110.67
79	107.5

Data Extension & Forecasting

Smoothening technique:

Moving Average (MA)



•As new observation is available, new average is computed by dropping the oldest observation and including the newest one.

•No. of data points in each average remains constant

•Deals with the latest ‘T’ periods of known data

Example-7

Data	MA (3)	MA (3, 3)
105	-	
115	-	
103	-	
108	107.67	
120	108.67	
97	110.33	108.89
110	108.33	109.11
121	109	109.22
117	109.33	108.89
79	116	111.44