

INDEX

NAME: Susendra Sharma STD: IV/II SEC: 078ACE ROLL NO.: 078ACE178
GH

S. No.	Date	Title	Page No.	Teacher's Sign / Remarks
		- Time Series		
		Ch-1 & 2 → Th. (8+8)		
		Ch-3 → Th. (Chebyshev inequality)		
		Ch-5 → AR & MA derivation		
		Ch-6 → Defn spectral analysis		
		<u>Rem.</u> ↳ Numerical.		

classmate

1. Introduction

classmate

Date _____
Page _____

1.1 Background

• Time series

- Time series is a series of values of variable collected over time. In other words, it is a set of observations x_t , each one being recorded at a specific time t .
- It is event based series of data with corresponding time period. Events data are taken with the sequence of the equal time interval. Time is an independent variable in the series.
- Such type of the data was taken for the future analysis of the events in the various sectors (like business, economic analysis, hydrology, earthquake sequence, geological sequences, Data analysis for censuses, etc.)
- Time series analysis has been done widely in Hydrology. Examples of time series data in hydrology are:
 - ✓ > Temperature
 - ✓ > Precipitation
 - ✓ > Discharge or water level
 - ✓ > Evaporation
 - ✓ > Groundwater level
 - ✓ > Tide level readings

• Rating curve: 4 monsoon months → every week measure } 32 times a year
8 dry months → 2 times a month

↓

Rating equation

- (Now can find discharge for any water level)

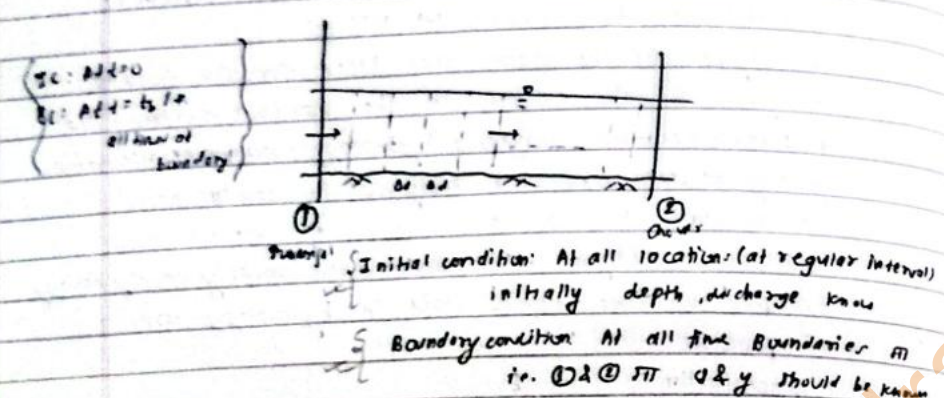
*wading → खोल, किंग बने

If no खोल → twin



→ boundary condition: initial condition
 → initial condition: initial condition
 → initial condition: initial condition
 After every process, (Q, V, A) should be known to see any change.

- Model
 - ↳ simplified representation of reality of the events and it associates with output or result.
 - ↳ type of the model is defined as the concept and analysis characteristic of the model boundary. The model boundary separates the part of reality described by the model from the rest of reality.



- ↳ The real situation of the events is expressed either physical or numerical (mathematical) based model. A mathematical model consists of a set of equations describing the system.

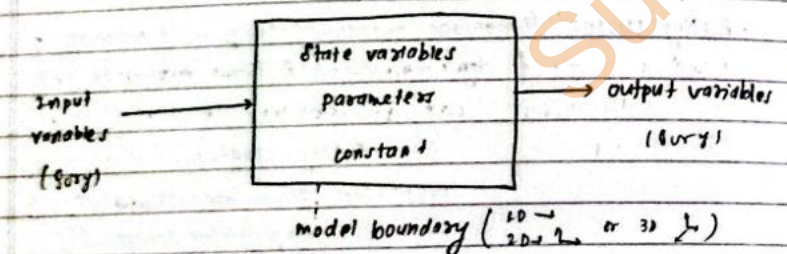


Fig: Model and model properties according to system's theory.



Eg. Manning's equation

$$V = \frac{1}{n} R^{2/3} S^{1/2}$$

S → slope: constant for given condition (straight) (provided other quantities unchanged)
 (eg. g)
 n → parameter: depends on condition of site (roughness) (varies with space)
 ↳ depends on surface roughness

State variables → variable on which the value of other variable depend on them.

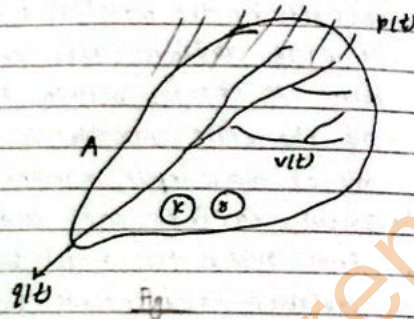
- ↳ The events that describe its reality are called the state variables or the events are the function of state variables (variables that describe current condition of system).
- ↳ Everything that is to know about the past of reality described by the model at a certain time is contained in the state variables. There are variables because their value can change both in space and time. The variation of the state variable is caused by the variation of one or more input variables.
- ↳ Input variables are always observed and originate from outside the model boundary. Consequently, input variables also include boundary conditions and initial conditions such as used when solving differential equations. (eg. P, Q at sundarjal, etc.)
- ↳ If the state variables are known, one or more output variables can be calculated. An output variable traverses
- ↳ output variable traverses the model boundary & thus influences the part of reality not described by model (what comes out of system)

- State variables → related to input variables } both change in space & time
- Parameter → related to output variables } space & time
- Parameter → may change in space but not with time
- As they are constant in time, they represent intrinsic properties of model

Parameter → constant property of system
(doesn't change with time but may change spatially)

Finally, a model may have one or more constants
Constants are properties of a model that do not change
in both space and time (within the confines of the model)

Examples of such constants are the gravity constant and the viscosity of water in density independent groundwater flow at a constant temperature.



- $v(t)$: State variables, surface water storage in catchment [L^3];
- $q(t)$: Output variables, surface runoff from catchment [$L^3 T^{-1}$];
- $p(t)$: Input variable, precipitation [$L T^{-1}$];
- K : Parameter, reservoir constant [T^{-1}];
- z : parameter, infiltration capacity [$L T^{-1}$];
- A : constant, area of the catchment [L^2];

The state variable is modelled with the following water balance equation:

$$v(t) = v(t-1) + A \{ p(t) - z \} - q(t) \Delta t$$

$$\text{If } p(t) < z \text{ we have } [p(t) - z] = 0 \quad \{ \text{bc} \}$$

The output variable 'q' is related to the state variable v at the previous time step with following equation

$$q(t) = K v(t)$$

K → parameter

Some Definitions:

1. Population

A collection of all possible values of a variable.

2. Sample

A group of values chosen from a population.

3. Random Variable

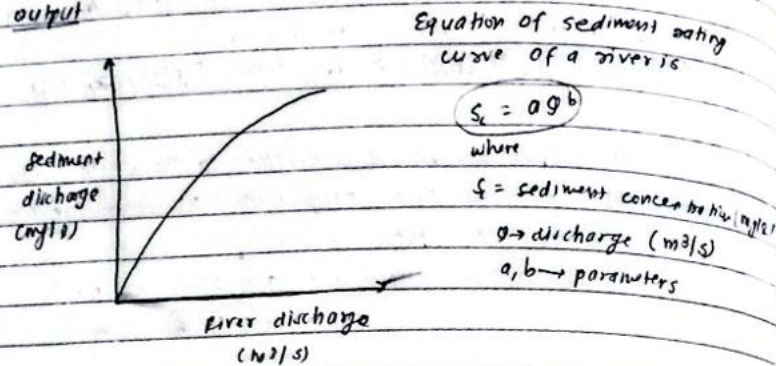
A variable which is associated with chance or random phenomenon

4. Deterministic and Stochastic (random, probabilistic) process

If the value of a variable can be predicted with the unique value referred with its observed sequence, then the process is known as deterministic process

The equations in deterministic process are based on the physical phenomenon and the outputs are found through the constant parameters

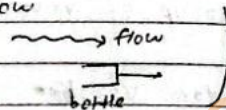
- Each process does not involve any randomness
 Therefore, a given set of input always gives some set of output



This eqⁿ represents deterministic approach

Dry months → 2 times a month } Measure ~~sed~~
 wet months → every week

sample collect if no other equipment available, use bottle
 & move in direction of flow



↳ Above curve gives suspended sediment.

↳ Bed load??

↳ 50-20% of suspended load

↳ If the

- If the value of a variable cannot be predicted uniquely, then process is known as stochastic process
 → The outputs in the stochastic process are governed through inherent randomness of the events
 → Value of random variable is associated with probability distribution

→ Therefore, a given set of input generates different set of outputs in each trial.

For eg, sediment discharge at time 't' is predicted as

$$S_{Qt} = K Q_{t-1} + e_t$$

where, e_t → random component

→ Most of the natural events → stochastic

& perfect prediction of such events are considerably more complicated

Time series Model

- Mathematical model which represents and analysis of the time series data.
- It is also known as stochastic model but it covers the deterministic as well as probabilistic event in natural world (stochastic)
- It has specific mathematical form and a set of parameters
- knowledge of statistics is applicable
- Two components
 - deterministic
 - Stochastic

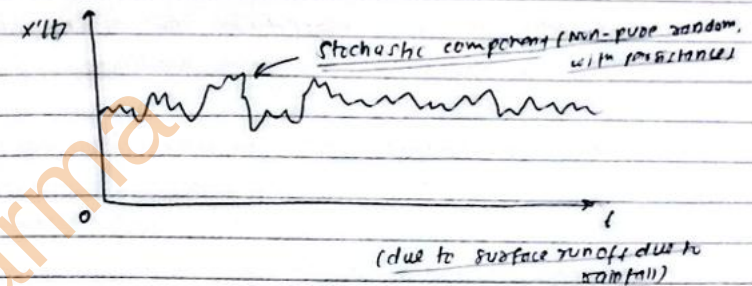
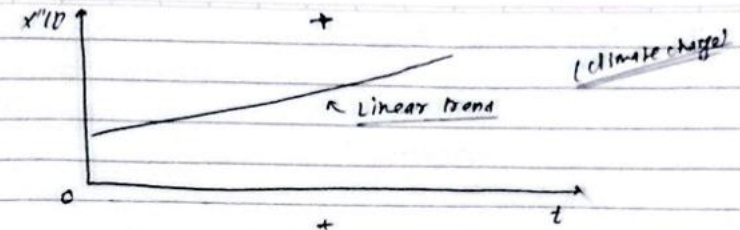
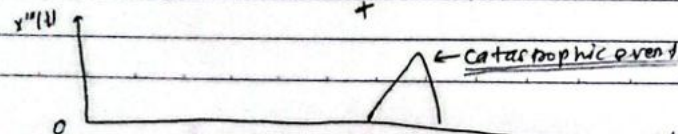
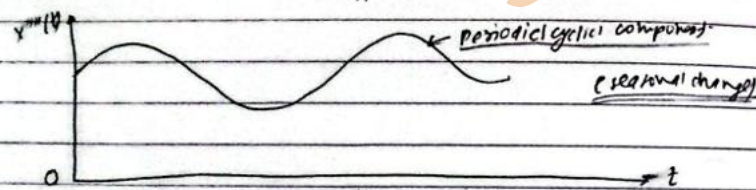
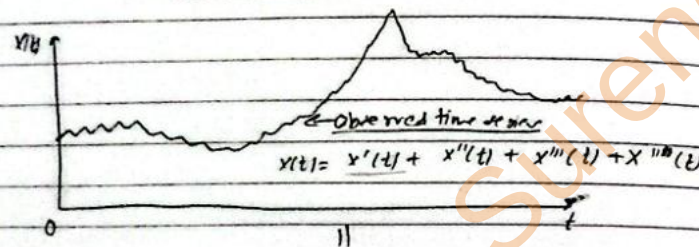
Eg.

1. Let, X_t : normally distributed RV with mean μ and var σ^2

$X_t = f(e_t)$ where $e_t \rightarrow$ independent RV

TS Model: $[X_t = \mu + \sigma e_t]$

parameters μ & σ^2



(w) Trend, cycle, Random walk

(b) $Q_t = \alpha Q_{t-1} + e_t$

Q_t = Discharge at time t

e_t = independent EV

Q_t depends on Q_{t-1}

~ A technique and procedure for fitting time series data is known as time series modelling.

• Steps to be followed for time series modeling.

- (1) Selecting type and form of model among several alternative models (best-fit model for given condition/case)
- (2) Parameter estimation of selected model (Calibration or validation)

eg. Hydrology (hydrological data)

- ~ Find parameters by matching past 15 years data
- ~ Predict and compare the output & actual data from above set parameters for remaining 15 years

(3) Testing the goodness of fit of model data.

* Realization (one observed ts)

→ Outcome of an experiment in which the process is observed

+ This single time series event in a stochastic process (eg. Rainfall etc.)

→ Process of river runoff due to precipitation is a single event with observed data.

eg. Rainfall data from 2000-2020

eg. All possible rain fall pattern under same conditions

* Ensemble (group) (all possible realizations)

→ Collection of all possible realizations of a process (group of the process)

→ The process of river runoff event occurred due to precipitation, temperature, soil type, humidity, etc.

* Ergodic property → If time average = ensemble average

→ Time-average properties are computed from a single realization, whereas ensemble-average properties are computed from several realizations. If the time averages and ensemble averages are equal over the same time period, the time-series is said to be ergodic.

Ergodic $\Rightarrow$ (Realization & Ensemble both $\rightarrow$ statistical parameters are same) ($\mu, \sigma^2, \sigma, \text{Mod}$)

→ A river runoff process is influenced by factors such as ppt, temp², soil type, and humidity. If the statistical characteristics of runoff obtained from a long-term record at one river location are the same as those obtained from many similar runoff events under the same conditions, then the runoff procedure is considered ergodic.

eg. If Avg rainfall over 100 years = avg of many possible rainfall scenarios then ergodic

* Normalization

→ Process of converting time series data into normal distribution

- 12 Application of time series analysis in Hydrology
- ↳ To generate synthetic hydrologic records
(Long weekly data & critical daily data Probabilistic)
 - ↳ To forecast hydrological events (2011-2015 data → 2020 needed)
 - ↳ To fill in missing data
 - ↳ To extend the records
(say we have 15 yrs data now extrapolating to 30 yrs)
(2021 missing data → need 2021)
 - ↳ To detect trend and shift in records

- 13 Physical basis of time series modelling in Hydrology
- Deterministic models → based on physics of system
 - Stochastic or time series models → based on statistical characteristics of historical data
- ↳ The data contain value has considered to generating mechanism of the process, the time series modeling represents some physical basis in an implicit way.
- ↳ The parameters of TS model can be considered to be analogous to physical based parameters by comparing the form of equations of two models.

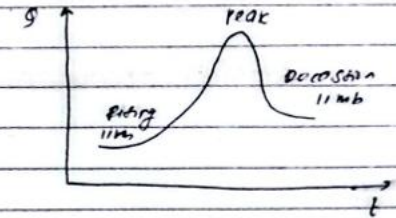
Eg
(a) Equation for recession limb in hydrograph

Eqr of recession curve: $Q_t = Q_0 e^{-kt}$

(Phys: water decreases exponentially)

Q_0 : initial discharge, Q_t : discharge at a time interval of t days.

k : recession const

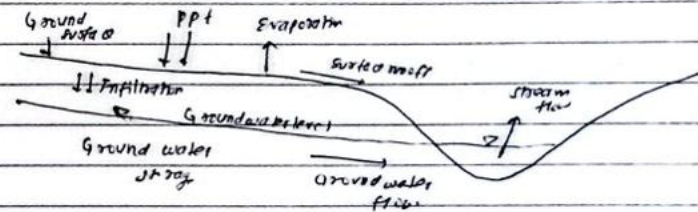


Analogous eqn with time series represented as
(TS form) $Q_i - \bar{Q} = \phi(Q_{i-1} - \bar{Q}) + e_i$

Q_i : discharge at any time i , $\bar{Q}$: mean discharge,

e_i : random component, ϕ : constant, which is equivalent to $(-1/k)$

(b) Conceptual representation of precipitation runoff process



Consider watershed system as shown in figure

Precipitation = X_1 , Evaporation = $b \cdot X_1$,

Infiltration = $a \cdot X_1$

$$\begin{aligned}\text{Surface runoff} &= PPT - \text{Evapo} - \text{Infil} \\ &= (1-a-b) X_t \\ &= d \times X_t\end{aligned}$$

$$GW \text{ storage} = G_{t-1}$$

$$GW \text{ flow} = C \times (G_{t-1}) \quad \cdot (\text{storage } G, \text{ in flow } X_t)$$

$$\text{stream flow } (Z_t) = C \times G_{t-1} + d \times X_t$$

Continuity equation for GW storage:

$$G_t = G_{t-1} + \text{inflow} - \text{outflow} \\ G_t = G_{t-1} + a \times X_t - C \times G_{t-1} = (1-C) \times G_{t-1} + a \times X_t$$

According to Salor and Smith written as,

$$Z_t = Z_{t-1} + d \times X_t - S X_{t-1} \quad \text{--- (1)}$$

This is a conceptual model considering of physical processes.

Analogous model in time series representing this form of equation, $Z_t = \phi_0 + \phi_1 Z_{t-1} + e_t - \theta_1 e_{t-1}$ --- (2)

where, e_t = random variable representing variable X_t (precipitation)

we fix ϕ_1 & θ_1 by matching eqn (1) & (2).

• Part of Reality

↳ you cannot model the whole world, so you select a limited system

⇒ selected system = part of reality

↳ what you include = affects accuracy

↳ what you ignore = becomes uncertainty

Chapter-2

Characteristics of Hydrologic series

2.1 Types of Hydrological Time series:

(a) Full Time series & Partial Duration series

If time series contains all recorded observations overtime the series is known as full time series

eg. daily rainfall

If TS contains selected obs. → partial duration series
eg. 1 day maximum rainfall (for each year)

DHM data Discharge

T → Trace (पानी का) but least count होगा कमी)

DNA → Data Not Available

8AM } Daily (Manually) (Inst. Max)
12PM }
4PM } Max Discharge

all } Daily (Inst. Max)
data } Max Discharge

✓ Maximum Discharge → Flood mark बार चला पड़ेने

✓ Instantaneous Minimum { 8am }
 { 12pm } to minimum discharge.
 { 4pm }

(b) Discrete & Continuous Time series:

→ If Time series data are recorded continuously with time, the series is called continuous TS

eg. water level data recorded continuously

↳ Automatic gauge recorder

→ If time series data are recorded at discrete points in time,

↳ Discrete TS

eg. Daily rainfall, Daily Discharge

→ Discrete TS is often derived from continuous TS

✓ Usually in a hydrograph, a time series is of discrete type

→ A continuous plot of a discrete time series should not be

confused with continuous TS

(c) Univariate (single) & Multivariate (multiple) time series

→ Univariate TS is a TS of one variable at a given site

eg. Rainfall measured at each site, Minimum rainfall at basin

↑
At a catchment, RGT rainfall from station

→ Multivariate TS is a set of two or more TS

eg. discharge data of diff station in river

↳ Discharge RGT time 8AM 12PM 4PM पर station se
every station se match hone jaye hain

Univariate → Only one variable changes with time

↳ eg. Daily rainfall at pulchowk; River discharge at one station

Multivariate → Multiple variables together

↳ eg. Rainfall + temp + humidity; Discharge at multiple stations.

(d) Uncorrelated & correlated TS

- If value of a variable at time 't' does not depend on the past values, the series is called uncorrelated or independent
- Otherwise, the series is called correlated (auto-correlated or serially correlated)
- eg. Autocorrelation of stream flow due to storage effects causing the water to remain in the system (ch1, latching)

(e) Intermittent and counting time series

- If a variable has zero & non-zero value throughout the length of record, the series is called intermittent time series
eg: Rainfall data
- | | |
|---|---|
| <u>Intermittent</u>
→ has many zero values
→ Eg. Rainfall data
0 on many days, > 0 on rainy days | <u>Counting</u>
→ counts no. of events
→ Eg. Number of rainy days per month; No. of floods per year |
|---|---|
- If a time series is derived by counting number of occurrences of event it is counting time series.
eg. count of rainy days in a month

(f) Regularly (evenly) & irregularly (unevenly) spaced time series

- If data is recorded at regular time interval → Regularly spaced
eg. Daily discharge data
- If not observed on regular time interval → Irregularly spaced
eg. Water quality measurements

Regular	Irregular
↳ Equal time interval	↳ Unequal interval
↳ Eg. Daily Rainfall; Monthly flow	↳ Eg. Water quality sampling; Landslide events

↳ AR, MA etc. models require regular spacing

Stationary: Annual rain fall (almost constant mean)
Non-stationary: Monthly rainfall (seasonal variation)
↳ Most models (AR, MA) assume stationarity

Detect by
→ Visual plot
→ Split become test
→ ACF behaviour

(g) Stationary and non-stationary time series

Stationary TS → A data sequence where statistical properties like mean and variance remain constant over time, meaning the pattern of series do not change significantly throughout its duration, such as annual rainfall, annual discharge

Otherwise → Non-stationary

eg. Monthly rainfall

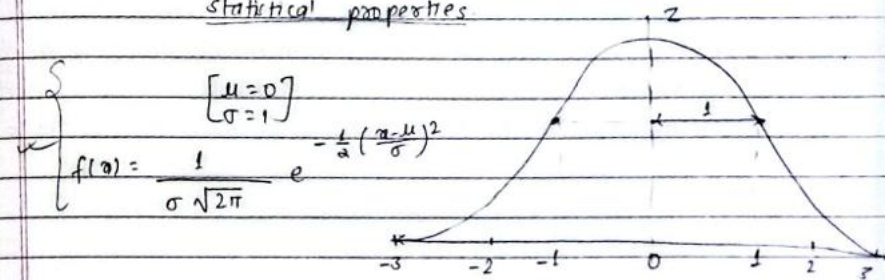
↳ It exhibits changing statistical properties often displaying trends, seasonality, or increasing variance as time progresses, indicating a pattern that evolves over time

First order stationary: only mean remaining constant with time

Second order stationary: mean and variance remaining constant with time

Weak stationary: second order stationary

Strong stationary: stationary in mean, variance and other statistical properties



2.1 General properties of hydrological time series:

(1) Homogeneity:

- If data in series belong to one population or single event of data, the series is homogeneous

- The non-homogeneity in time series is due to multiple set of observations like natural or manmade causes like land use changes, urbanization & deforestation

(2) Inconsistency:

- Time series occurs due to systematic error in observation
 - ↳ Longer time series has greater chances of non-homogeneity

(3) Persistence

- If value of variable depends on past values $\rightarrow$ series is persistent
- Time dependence in hydrological series is due to geophysical process of water changes in basin, their random fluctuation from year to year & periodic fluctuation within a year

Components of Time series:

A hydrological time series contains both deterministic & stochastic components

Deterministic components:

(i) Trend

- ↳ Long term movement in time series
- ↳ Trend in time series exists due to natural or man-made causes such as landscape changes, urbanization, deforestation, climate change
- ↳ describe pattern of the data
- ↳ movement of a data to relatively higher value or lower value over a long period of time
- ↳ general tendency of data to occur in a particular way

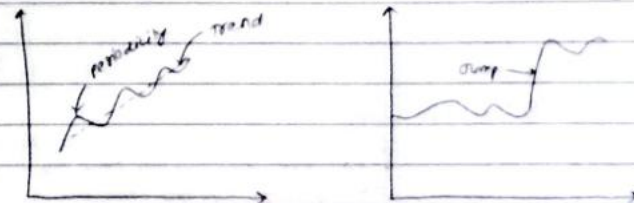
(ii) Periodicity

- ↳ Regular fluctuations (cyclic behaviour) in time series
- ↳ If fluctuation in T is for a period less than a year, the property is called seasonality.

↓
It is due to rotation of earth around the sun

(iii) Jump or shift

- ↳ Abrupt change in time series which occurs due to sudden change in basin characteristics



Test w/o assuming distribution
 Eg: Mann-Whitney Test, Spearman Test
 Use: Hydrologic data not always normal & has randomness

• Non-parametric test:-

→ A non-parametric test in statistics does not assume that the data has been taken from a normal distribution. A normal distribution belongs to a parametrized family of probability distributions and includes parameters such as mean, variance, standard deviation, etc. Thus, a non-parametric test does not make assumptions about the probability distribution's parameters.

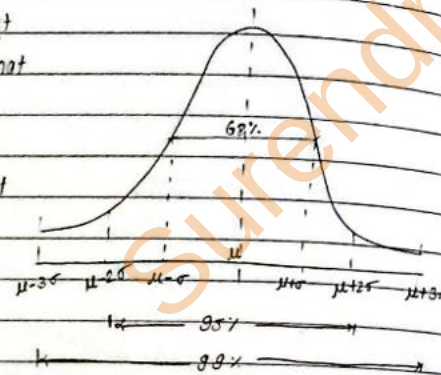
→ In parametric test it is simple, the events follow the normal distribution curve and test like t-test, z-test, Anova, etc.

• Level of significance

Probability of being incorrect
 $P > 0.05$ is the probability that the null hypothesis is true

A statistically insignificant test result ($P < 0.005$) means test hypothesis is false or should be rejected

False positive
 False negative



• False +ve (Type I error)

→ detect trend when no trend exists
 Eg: Data random but test says "trend exists"
 Wrong design (or false detection)

• False -ve (Type II error)

→ Miss real trend
 Eg: Actual increasing rainfall but test says "no trend"
 Undetected trend (or missed detection)

True +ve

Reality → Trend
 Test → Detects trend

1. Identify type of data
2. Check stationarity
3. Choose test: (i) Parametric (if normal)
 (ii) Non-Parametric (if unknown)
4. Detect: trend & periodicity

Detection of trend (for climate change analysis; hydrology design)

(a) Regression test for linear trend

Fit a straight line in the form of $y = at + b$

y = given variable, t = time, a, b = constant

From linear regression analysis

$$a = \frac{N(\sum ty) - (\sum t)(\sum y)}{N(\sum t^2) - (\sum t)^2}$$

$$b = \frac{\sum y - a(\sum t)}{N}$$

If a is not close to zero, then there is presence of linear trend in time series

(b) Spearman's Rank Correlation Test { Check relationship b/w time (t) and data (x) }

It is the statistical measure to show the strength of a relationship between two variables

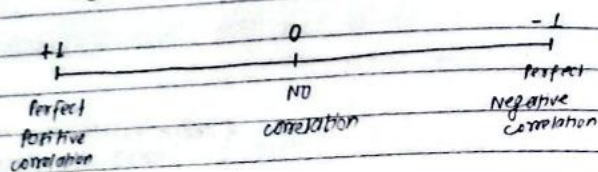
Minimum nos required for test = 10 data

List the data, x_i , $i = 1, 2, \dots, n$ in date order & the rank (r_i) in ascending order of magnitude (with smallest allocated rank 1 & largest n)

Compute Spearman's Rank correlation

$$R_{sp} = 1 - \frac{6 \sum (r_i - i)^2}{n(n^2 - 1)}$$

Date order	Rank order	
i	r_i	$(r_i - i)$
1	5	
2	3	
3	1	
4		
5		
6		
7		
8		
9		
10		$(2 \cdot n - 1)$



Compute t-test statistic

$$t = \frac{\sum (r_i - i)}{\sqrt{\frac{n-2}{1 - k^2}}}$$

Find tabulated value of t for α level of significance.

(say $\alpha = 5\%$) and $(n-2)$ degree of freedom

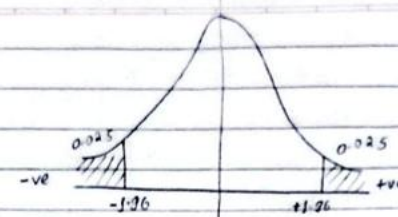
निर्णय करने के लिए तब तक तालिका देखें

Result (Accept or Reject) H_0

क्या स्वीकार करेंगे

(here, 2 data)

If computed $t <$ tabulated t , $\rightarrow$ There is no trend



(c) Mann-Kendall Test (Compare each data point with all future values)

Rank (R) of the time series data in ascending order of magnitude (smallest allocated rank 1 and largest rank n)

Compute Mann-Kendall test statistic 'S' as

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(R_j - R_i)$$

$\text{sgn} \rightarrow \text{sigma}$

(data 1 to 10)

$$\text{sgn}(x) = \sigma(x) = 1 \text{ for } x > 0$$

$$\sigma(x) = 0 \text{ for } x = 0$$

$$\sigma(x) = -1 \text{ for } x < 0$$

Standard deviation of S

$$(i) \text{Var}(S) = \frac{n(n-1)(2n+5)}{18} \text{ (no tied value: all data different)}$$

$$(ii) \text{Var}(S) = \frac{n(n-1)(2n+5) - \sum h_i(h_i-1)(2h_i+5)}{18} \text{ (tied values)}$$

Z-test statistic is

$$Z = \frac{S}{\sqrt{\text{Var}(S)}}$$

$\sqrt{\text{Var}(S)}$

eg: [10, 12, 13, 15]

12 appears twice $\rightarrow$ tied values

Find tabulated value of Z for level of significance

(eg, $\alpha = 5\%$, $Z = 1.96$)

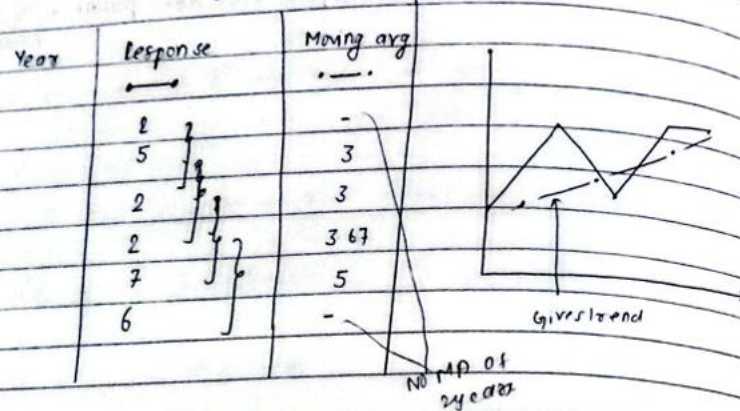
If computed $Z <$ tabulated Z , there is no trend

Removal of data

(1) Moving Average Method

↳ Moving average: Average of n successive interval of data

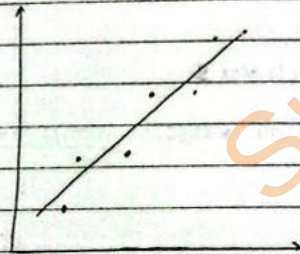
↳ Computed moving average & subtract it from original data



(2) Linear trend

↳ A straight line is fit of form $y = at + b$

↳ At time 't', subtract the values of the line from original data



ACE: compare data with past values

$$r_k = \frac{\sum (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum (x_t - \bar{x})^2}$$

↳ If periodic ACF show repeating peaks

Graph (Correlation)
Log → x-axis
r → y-axis

Detection of periodicity:

- By visual inspection of time series: nothing effect of period in the plot

- By performing autocorrelation & spectral analysis

(ii) Spectral Analysis → convert time series → frequency domain

↳ peak in spectrum → strong periodicity

↳ More precise than ACF for complex periodicity

Detection of Jump

- By visual inspection of TS

Split record test for checking stationarity

- Testing whether mean and variance of subsets of a set are not significantly different

Approach

(1) Divide the records into two non-overlapping subsets

$j = 1, 2, \dots, m$

& $j = m+1, m+2, \dots, n$ (eg. 50 years → 25 + 25)

(2) Compute mean & variance of each subset

Mean → t-test
Variance → F-test

(3) Examine the variance (V) first using F-test

(4) $F = \frac{V(a)}{V(b)}$ where, $V(a) > V(b)$

(5) Compare the computed F with tabulated value of F for $m-1$ and $n-m-1$ degree of freedom at α level

of significance

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$F = \frac{s_1^2}{s_2^2}$$

* If both mean & variance equal
→ Stationarity
also for stationarity

(6) If the computed $F <$ tabulated value of F , then the variances of subsets can be accepted as estimator of the same population variance.

(7) compute pooled estimate of variance, V_p

$$V_p = \frac{(m-1) V(a) + (n-m-1) V(b)}{n-2}$$

(8) Test statistic for mean:

$$t = \frac{M(a) - M(b)}{\sqrt{\frac{V_p}{n-m}}}$$

(9) compare the computed t with the tabulated value of t for $n-2$ degree of freedom at $\frac{\alpha}{2}$ level of significance (two-tailed test).

(10) If the computed $t <$ tabulated value of t , then mean zero difference b/w two subsets is accepted.

ii Double Mass curve analysis for checking consistency:

(1) compute cumulative sum of data (most annual data is sufficient) of base station (A)

(2) compute cumulative sum of mean value of neighbouring stations (B)

(3) plot A vs B & fit points by a straight line. If there is no break in the line, the data is considered to be consistent.

#Spearman's Rank correlation test

Step 1: Arrange data in time order $\rightarrow R(t)$

Step 2: Assign ranks to X (ascending) $\rightarrow R(x)$

Step 3: compute difference $\{[d = R(x) - R(t)] \rightarrow d^2 \rightarrow \sum d^2\}$

Step 4: Calculate Spearman's Rank correlation,

$$R_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Step 5: Significance test (convert to t)

$$t = R_s \sqrt{\frac{n-2}{1-R_s^2}}$$

Step 6: Compare with table value

computed $t <$ tabulated $t \rightarrow$ no trend.

* Interpretation:

$R_s \approx +1 \rightarrow$ strong increasing trend

$R_s \approx -1 \rightarrow$ decreasing trend

$R_s \approx 0 \rightarrow$ no trend

$R(t)$	X	Rank(x)	$d = R(x) - R(t)$	d^2
1	12	1	0	0
2	15	3	1	1
3	14	2	-1	1
4	18	4	0	0
5	20	5	0	0
				$\sum d^2 = 2$

Type →	Spearman Correlation	Mann-Whitney U-Test
Bar chart	Rank	Non-parametric
Sensitivity	Modest	Sign
Wage	Simple	Strong
		Preferred in industry

Mann-Kendall test.

eg: Data: $X = [12, 15, 14, 18, 20]$

• Step 1: Compare pairs

we compute $\text{sgn}(x_j - x_i)$

Rule: +1 if $x_j > x_i$

0 if equal

-1 if $x_j < x_i$

• Step 2: Table

Pair	Difference	Sign
15-12	+	+1
14-12	+	+1
18-12	+	+1
20-12	+	+1
14-15	-	-1
18-15	+	+1
20-15	+	+1
18-14	+	+1
20-14	+	+1
20-18	+	+1
		$\Sigma = 8$

• Step 3: Compute S

$$S = \Sigma \text{sign} = 8$$

• Step 4: Compute variance (substituting in given formula)

✓ tied value

✓ no tied value

• Step 5: Compute Z :

$$Z = \frac{S}{\sqrt{\text{Var}(S)}}$$

* Compare with tabulated Z (eg: $Z = 1.96$ for 5% level)

* Interpretation, computed $Z >$ tabulated $Z \rightarrow$ trend exists.

Chapter-3

Statistical Techniques Principles and Techniques for Time Series Modeling

3.0 Concept of Probability Distribution

A. Frequency

- For discrete RV, the number of occurrences of a variate is generally called frequency.
- When number of occurrences of a variate, or the frequency is plotted against the variate, a pattern of distribution is obtained called frequency distribution.

B. Relative frequency function

If the number of observations n , in interval i is divided by the total number of observations n , the result is called relative frequency function

$$f_i = \frac{n_i}{n} = P(x_i)$$

This is also called probability of a function

C. Probability mass function (PMF) and cumulative distribution function (CDF) of a discrete RV

→ PMF: Probability function or PMF

PMF gives the probability associated with all possible values of a RV

→ For a discrete RV X , PMF ($P(x_i)$) is given by

$$P(x_i) = P[X = x_i]$$

Properties of PMF

$$0 \leq P(x_i) \leq 1$$

$$\sum P(x_i) = 1$$

$$P[a \leq X \leq b] = \sum_{x_i \in [a, b]} P(x_i)$$

CDF:

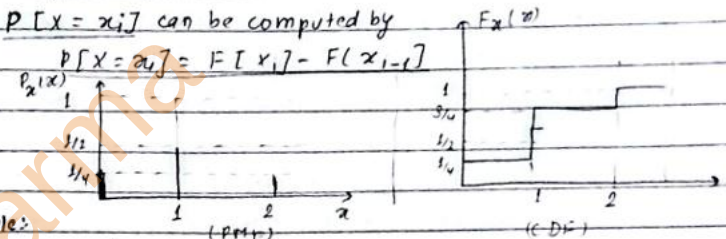
→ the probability of the event that the RV X is less than or equal to x , for every possible value x .

$$F(x) = P[X \leq x]$$

→ For a discrete RV, the CDF is found by summing up the probabilities

$P[X = x_i]$ can be computed by

$$P[X = x_i] = F(x_i) - F(x_{i-1})$$



Example:

Assume that X has the PMF given by

$$P_X(x) = \begin{cases} 1/4 & x=0 \\ 1/2 & x=1 \\ 1/4 & x=2 \\ 0 & \text{otherwise} \end{cases}$$

Find the CDF of X Also plot PMF and CDF

→ Solⁿ:

PMF of the RV X is as follows:

$$P_X(0) = 1/4, P_X(1) = 1/2, P_X(2) = 1/4 \text{ and otherwise } 0$$

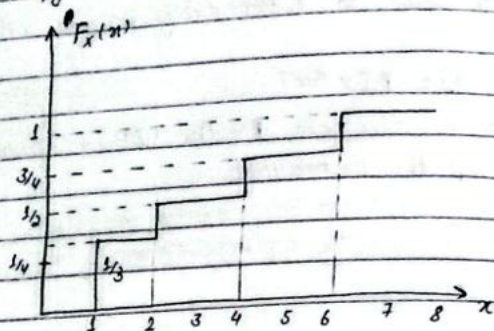
Here, the RV changes its values at $x=0, x=1$ and $x=2$.

The values are $0, 0 + 1/4 = 1/4, 1/4 + 1/2 = 3/4$ and $3/4 + 1/4 = 1$

Thus, CDF of X is given by

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

Q. The plot of the CDF of a discrete random variable X is shown in figure below. Find the PMF of X .



→ Soln:

The random variable takes on values with non-zero probability at $x=1, x=2, x=4$ and $x=6$.

The size of jump at $x=1$ is $1/3$ ✓

$$x=2 \text{ is } 1/2 - 1/3 = 1/6 \checkmark$$

$$x=4 \text{ is } 3/4 - 1/2 = 1/4 \checkmark$$

$$x=6 \text{ is } 1 - 3/4 = 1/4 \checkmark$$

Thus, the PMF of X is given by

$$P_X(x) = \begin{cases} 1/3 & x=1 \\ 1/6 & x=2 \\ 1/4 & x=4 \\ 1/4 & x=6 \\ 0 & \text{otherwise} \end{cases}$$

eg. Find the PMF of a discrete RV X whose CDF is given by.

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/6 & 0 \leq x < 2 \\ 1/2 & 2 \leq x < 4 \\ 5/8 & 4 \leq x < 6 \\ 1 & x \geq 6 \end{cases}$$

→ Solution

• Here, CDF changes values at $x=0, x=2, x=4, x=6$ which means that these are the values of random variables that have non-zero probabilities at these points.

• Next Determine Probability

$$P_X(0) = 1/6$$

$$\text{At, } x=2, \text{ size of jump is } 1/2 - 1/6 = 1/3 = P_X(2) \checkmark$$

$$\text{At, } x=4, \text{ size of jump is } 5/8 - 1/2 = 1/8 = P_X(4) \checkmark$$

$$\text{Finally at } x=6, \text{ size of jump is } 1 - 5/8 = 3/8 = P_X(6) \checkmark$$

• Therefore, PMF of X is given by,

$$P_X(x) = \begin{cases} 1/6 & x=0 \\ 1/3 & x=2 \\ 1/8 & x=4 \\ 3/8 & x=6 \\ 0 & \text{otherwise} \end{cases}$$

D. Probability Density function (PDF) and Cumulative Distribution Function (CDF) of a continuous RV.

→ In case of continuous RV, we cannot assign probability for every value that it can take, where X-axis is divided into a large number of mutually exclusive intervals each of an infinitesimal length dx . We can think of a function $f(x)$ i.e. the probability that x lies in the interval x to $x+dx$ is given by $f(x)dx$. Such a function is called PDF.

$$P[a_1 \leq X \leq a_2] = \int_{a_1}^{a_2} f(x) dx$$

where, $f(x)$ represents probability at point 'x'.

$$f(x) \geq 0 \text{ and } \int_{-\infty}^{\infty} f(x) dx = 1$$

x = specified value of variable x .

$P[a \leq x \leq b]$ is the probability that x lies b/w a and b .

Knowing CDF, $P[a \leq x \leq b]$ can be computed as

$$P[a \leq x \leq b] = F(b) - F(a)$$

Relationship b/w CDF and PDF is:

$$f(x) = \frac{dF(x)}{dx} \text{ where } f(x) = \text{PDF}$$

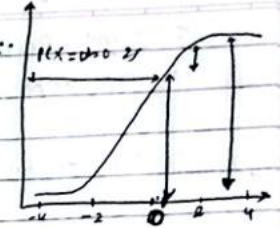
$$\text{and } F(x) = \text{CDF}$$

(B) The PDF of a daily rainfall is given as:

$$\text{Prob}[X=0] = 0.25$$

$$f(x) = \frac{x}{4} \text{ for } 0 < x < 1.732$$

$$f(x) = 0.866 - \frac{x}{4} \text{ for } 1.732 < x < 3.464$$



Is this a proper PDF? If yes, what is $P[X > 0.5]$ and what is $P[X > 0.5 / X \neq 0]$

→ Solution:-

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) dx &= \int_{-\infty}^0 f(x) dx + \int_0^{1.732} f(x) dx + \int_{1.732}^{3.464} f(x) dx \\ &= 0.25 + \int_0^{1.732} \left(\frac{x}{4}\right) dx + \int_{1.732}^{3.464} \left(0.866 - \frac{x}{4}\right) dx \\ &= 0.25 + \left[\frac{x^2}{8}\right]_0^{1.732} + \left[0.866x - \frac{x^2}{8}\right]_{1.732}^{3.464} \\ &= 0.25 + \frac{1.732^2}{8} + \left[0.866x - \frac{x^2}{8}\right]_{1.732}^{3.464} \\ &= 1 \end{aligned}$$

As, $\int_{-\infty}^{\infty} f(x) dx = 1$, the given function is a proper PDF.

$$\begin{aligned} \rightarrow P[X > 0.5] &= 1 - P[X < 0.5] \\ &= 1 - \left[\int_{-\infty}^0 f(x) dx + \int_0^{0.5} \left(\frac{x}{4}\right) dx \right] \\ &= 1 - [0.25 + \dots] \\ &= 0.72 \end{aligned}$$

$$P[X \neq 0] = 1 - P[X=0] = 1 - 0.25 = 0.75$$

$$P[X > 0.5 / X \neq 0] = \frac{0.72}{0.75} = 0.96$$

$X \neq 0$ का case मात्र लिए $P(X > 0.5)$ निकाला जाये

eg: Probability
capture dependence b/w variables

(JPD)

E Joint Probability Distribution (Bivariate or Multivariate)
→ JPD represent probability distribution of more than one RV

→ For two discrete RV X and Y, JPD is -

$$P(x, y) = P[X=x, Y=y]$$

conditions,

$$P(x, y) \geq 0 \text{ and } \sum_x \sum_y P(x, y) = 1$$

The Joint CDF is

$$F(x, y) = P[X \leq x, Y \leq y] = \sum_{s \leq x} \sum_{t \leq y} P(s, t)$$

For $-\infty < x < +\infty, -\infty < y < +\infty$ where $P(s, t)$ is the value of joint probability function of X and Y at (s, t) .

→ For two continuous RV X and Y, joint PDF is

$$P(x, y) = P[X=x, Y=y] \quad \text{Black (Y)}$$

	$f(x, y)$	1	2	3	4	$f_X(x)$
$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2)$	1	$3/16$	$1/16$	$1/16$	$1/16$	$4/16$
	2	$1/16$	$1/16$	$1/16$	$1/16$	$4/16$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$4/16$
	4	$1/16$	$1/16$	$1/16$	$1/16$	$4/16$
		$4/16$	$4/16$	$4/16$	$4/16$	1

$$= \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y) dx dy$$

Conditions

$$f(x, y) \geq 0 \text{ for } -\infty < x < +\infty, -\infty < y < +\infty$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

Joint CDF for two continuous RV

$$F(x, y) = P[-\infty < X \leq x, -\infty < Y \leq y]$$

$$= \int_{-\infty}^x \int_{-\infty}^y f(s, t) ds dt$$

For $-\infty < x < +\infty, -\infty < y < +\infty$, where $f(s, t)$ is the value of joint probability function of X and Y at (s, t) .

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

eg: show that the function $f(x, y) = \begin{cases} \frac{2}{5}(2x+3y), & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases}$

is a joint density function of X and Y.

→ Solution:-

If $f(x, y)$ satisfies the condition

$$(i) f(x, y) \geq 0$$

$$(ii) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

Then, $f(x, y)$ is JDF

(i) Given, $f(x, y) \geq 0$ in the given interval $0 \leq (x, y) \leq 1$

$$\begin{aligned} (ii) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy &= \int_0^1 \int_0^1 \frac{2}{5} (2x+3y) dx dy \\ &= \frac{2}{5} \int_0^1 \left[\frac{2x^2}{2} + 3y \right]_0^1 dy \\ &= \frac{2}{5} \int_0^1 [1 + 3y] dy \\ &= \frac{2}{5} \left[y + \frac{3y^2}{2} \right]_0^1 \\ &= \frac{2}{5} \left[1 + \frac{3}{2} \right] \\ &= 1 \end{aligned}$$

Since $f(x,y)$ satisfies the two conditions (i) and (ii) it is a JDF

probability of one variable ignoring the other

1 F. Marginal distribution:

Given the Joint probability distribution of two RV, marginal distribution is the distribution of one of the variables obtained by integrating out the variable that is no longer of interest.

Given $P(x,y)$ = Joint Probability Distribution of two RV X and Y

Discrete case,

Marginal PMF of X is $g(x) = \sum_{all y} P(x,y)$

Marginal PMF of Y is $h(y) = \sum_{all x} P(x,y)$

		Black (Y)				
		1	2	3	4	$f_x(x)$
Red (X)	1	$1/16$	$1/16$	$1/16$	$1/16$	$4/16$
	2	$1/16$	$1/16$	$1/16$	$1/16$	$4/16$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$4/16$
	4	$1/16$	$1/16$	$1/16$	$1/16$	$4/16$
		$f_y(y)$	$4/16$	$4/16$	$4/16$	$4/16$

Continuous case:

Marginal PDF of X is $g(x) = \int_{-\infty}^{\infty} f(x,y) dy$ for $-\infty < x < \infty$

Marginal PDF of Y is $h(y) = \int_{-\infty}^{\infty} f(x,y) dx$ for $-\infty < y < \infty$

probability of one variable given another

2. Conditional Distribution

A distribution of values for one variable that exists based on the existence of the values of other variables. This type of distribution allows to assess the dispersal of a variable of interest under specific conditions.

If $f(x,y)$ = Joint PDF of two RV X and Y ,

$g(x)$ = Marginal PDF of X ,

$h(y)$ = Marginal PDF of Y ,

Conditional Pdf of X is $f(x|y) = \frac{f(x,y)}{h(y)}$

Conditional Pdf of Y is $f(y|x) = \frac{f(x,y)}{g(x)}$

If $f(x|y) = g(x)$ and $f(y|x) = h(y)$, two random variables are independent.

In this case, $f(x,y) = g(x)h(y)$

X \ Y	0	10	25	35	$f_x(x)$
0	0	0	0	$1/4$	$1/4$
1	0	$1/4$	$1/4$	0	$2/4$
2	$1/4$	0	0	0	$1/4$
$f_y(y)$	$1/4$	$1/4$	$1/4$	$1/4$	1

		Y				
		0	10	25	35	$f_x(x)$
X	0	0	0	0	1	1
	1	0	1	1	0	1
	2	1	0	0	0	1
		$f_y(y)$	1	1	1	1

(Q) If two RV X and Y have the joint PDF given by
 $f(x,y) = \frac{2x+y}{4}$ for $0 < x < 1, 0 < y < 2$.

= 0 elsewhere

Find the marginal density of X , marginal density of Y and verify whether two RV are independent.

→ Solution:

Marginal PDF of X ,

$$\begin{aligned} g(x) &= \int_{-\infty}^{\infty} f(x,y) dy = \int_0^2 \frac{2x+y}{4} dy \\ &= \frac{1}{4} \left(2xy + \frac{y^2}{2} \right)_0^2 \\ &= \frac{1}{4} (4x + 2) \end{aligned}$$

Marginal PDF of Y ,

$$\begin{aligned} h(y) &= \int_{-\infty}^{\infty} f(x,y) dx = \int_0^1 \frac{2x+y}{4} dx \\ &= \frac{1}{4} \left[\frac{2x^2}{2} + x \cdot y \right]_0^1 \\ &= \frac{1}{4} (1 + y) \end{aligned}$$

Now,

$$f(x|y) = \frac{f(x,y)}{h(y)} = \frac{\frac{2x+y}{4}}{\frac{y+1}{4}} = \frac{2x+y}{y+1}$$

$$f(y|x) = \frac{f(x,y)}{g(x)} = \frac{\frac{2x+y}{4}}{\frac{4x+2}{4}} = \frac{2x+y}{4x+2}$$

As, $f(x|y) \neq g(x)$ and $f(y|x) \neq h(y)$,

the given random variables are not independent.

eg. Rainfall → Runoff from rainfall

(Discharge derived from rainfall)

H. Derived distributions:

Given the pdf of independent variable, derived distributions is the distribution obtained for the functionally dependent variable.

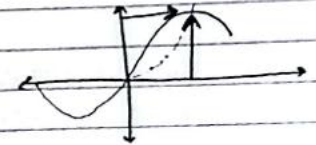
Given: $f(x)$ as pdf of X and

to compute PDF of Y , say $g(y)$.

For small element in PDF

Area of PDF of X = Area of PDF of Y

$$\begin{aligned} \int f(x) dx &= g(y) dy \\ g(y) &= f(x) \frac{dx}{dy} \end{aligned}$$



Since, $g(y)$ cannot be negative.

$$g(y) = f(x) \left| \frac{dx}{dy} \right|$$

Q. A Random variable X has the PDF defined by,

$$f(x) = x e^{-x^2/2} \text{ for } 0 \leq x < \infty$$

= 0 elsewhere

and is related to another RV Y through $y = x^2$. Find the pdf of Y .

→ Solution:

$y = x^2$ (distribution of Y derived from X)

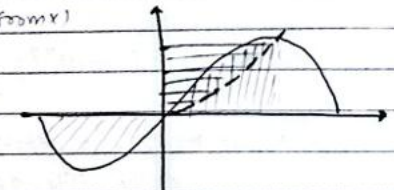
$$\frac{dy}{dx} = 2x$$

$$\left| \frac{dx}{dy} \right| = \frac{1}{2x}$$

Since, $g(y)$ cannot be negative:

$$g(y) = f(x) \left| \frac{dx}{dy} \right|$$

$$= \frac{x e^{-x^2/2}}{2x} = \frac{1}{2} e^{-x^2/2}$$



* Statistical Properties

• Sample statistics

✓ Mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

✓ Variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$, S.D., $\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

✓ CV = $\frac{\sigma}{\bar{x}}$

✓ Coefficient of skewness, $s_3 = \frac{n \sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(n-2)s^2}$; skewness = $\sum_{i=1}^n \frac{(x_i - \bar{x})^3}{n s^3}$

• Population statistics

✓ Mean = $\mu = E(x) = \int_{-\infty}^{\infty} x f(x) dx$ = first moment about origin
(expectation of x)

Median = x such that $\int_0^x f(x) dx = 0.5$

✓ Variance = $\text{var}(x) = E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$ = Second moment about mean

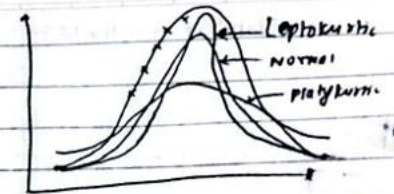
CV = $\frac{\sigma}{\mu}$

✓ Skewness: $E[(x - \mu)^3] = \int_{-\infty}^{\infty} (x - \mu)^3 f(x) dx$ = third moment about mean

✓ Coefficient of skewness: $\gamma = \frac{E[(x - \mu)^3]}{\sigma^3}$

✓ Kurtosis: $E[(x - \mu)^4] = \int_{-\infty}^{\infty} (x - \mu)^4 f(x) dx$ = Fourth moment about mean

↳ Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. In other words, kurtosis identifies whether the tails of given distribution contains extreme values.



Coefficient of kurtosis: $K = \frac{E[(x - \mu)^4]}{\sigma^4}$

(For discrete case, replace $f(x)$ by $p(x)$ and apply summation sign)

If $K = 3$: Curve is having normal peak and it is called Mesokurtic curve

If $K > 3$: more peaked than normal $\Rightarrow$ Leptokurtic curve

If $K < 3$: less peaked than normal $\Rightarrow$ Platykurtic curve

* Measures of Central Tendency

Mean: $\frac{\sum x_i}{n} = \bar{x}$ (It represents the center of mass or the balance point of the distribution if the data points were a physical mass distributed along a line. It is the expected value of a dataset)

Median: middle value in a sorted dataset splits into two halves

Mode: value that occurs most frequently. It identifies the most common or most probable observation in the dataset, representing the peak of the data's distribution when visualized as a graph

* Measures of Dispersion:

Variance: Average of the squared differences from the mean, showing data dispersion (spread)

↳ quantifies the data's overall dispersion or how spread out the data points are from the mean. A high variance means the data is widely spread

Standard deviation (SD): The square root of the variance, in the same units as the data. Long distance of the data points from the mean $\rightarrow$ low SD $\rightarrow$ clustered data points tightly around mean.
 $\rightarrow$ High SD $\rightarrow$ data points are widely spread out.

Coefficient of variance (CV):

Ratio of SD to the mean (SD/Mean), useful for comparing variability between the different datasets or scales.

* Measures of shape (Asymmetry & Tail)

• Skewness:

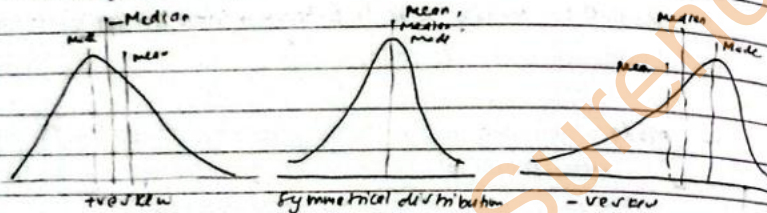
$\rightarrow$ Measures the asymmetry of a distribution's shape around its mean.
 $\rightarrow$ Indicator which side of the center of mass has a 'longer tail' or more extreme values pulling the mean away from the median.

$\rightarrow$ +ve skew: A longer tail to the right; Mean > Median.

$\rightarrow$ -ve skew: A longer tail to the left; Mean < Median.

• Coefficient of skewness

$\rightarrow$ Measures the asymmetry of a distribution's shape around its mean.
 $\rightarrow$ A standardized, unitless measure (the third standardized moment) that provides the direction and the extent of the asymmetry, allowing for comparison b/w different datasets.



• Kurtosis: Measures the "tailedness" or "peakedness" of a distribution

Compared to a normal distribution (mesokurtic)

Leptokurtic: Heavy tail, sharp peak (high kurtosis)

Platykurtic: Light tail, flat peak (low kurtosis)

• Coefficient of kurtosis (Excess Kurtosis): A standardized, unitless measure (the 4th standardized moment) that quantifies the degree of peakedness/tailedness relative to a normal distribution.

Q6 The pdf of a RV is given by:

$$f(x) = kx \text{ for } 0 \leq x \leq 4$$

$$= k(8-x) \text{ for } 4 \leq x \leq 8$$

$$= 0 \text{ elsewhere}$$

Evaluate the constant k . Also find mean and standard deviation.

$\rightarrow$ Solⁿ:

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$\text{on } \int_{-\infty}^0 f(x) dx + \int_0^4 f(x) dx + \int_4^8 f(x) dx + \int_8^{\infty} f(x) dx = 1$$

$$\text{on } 0 + \int_0^4 kx dx + \int_4^8 k(8-x) dx + 0 = 1$$

$$\text{on } \frac{k \cdot 4^2}{2} + \frac{8k[8-4] - k[8^2-4^2]}{2} = 1$$

$$\text{on } \Rightarrow \frac{40k}{16} = 1 \Rightarrow (40-24)k = 1$$

$$\text{Mean, } \mu = E(x) = \int_{-\infty}^{\infty} xf(x) dx = \int_0^4 xf(x) dx + \int_4^8 xf(x) dx + 0$$

$$= \int_0^4 kx^2 dx + \int_4^8 kx(8-x) dx$$

$$= 4$$

$$\text{Var.}(x) = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

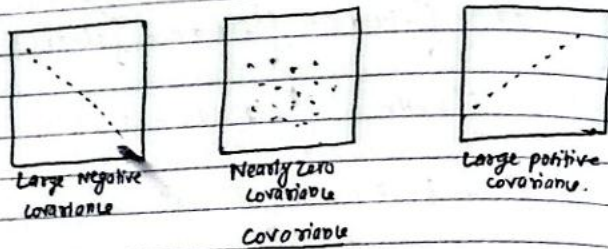
$$= \int_{-\infty}^{\infty} (x-4)^2 f(x) dx$$

$$= \int_0^4 (x-4)^2 \cdot \frac{kx}{16} dx + \int_4^8 (x-4)^2 \cdot \frac{k(8-x)}{16} dx$$

$$= 2.67$$

$$\text{Standard deviation (SD)} = \sqrt{2.67} = 1.63$$

- * Covariance between two random variables X and Y .
- Covariance is the linear relationship b/w two RV
 - Height and wake-up time are uncorrelated, but height and weight are correlated.
 - Covariances:
 - $\text{Cov}(X, Y) = 0$, for $X = \text{ht.}$ & $Y = \text{wake-up time}$
 - $\text{Cov}(X, Y) > 0$, for $X = \text{ht.}$ & $Y = \text{weight}$.



Defⁿ:

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

- Covariance signifies the direction of the linear relationship b/w two variables.

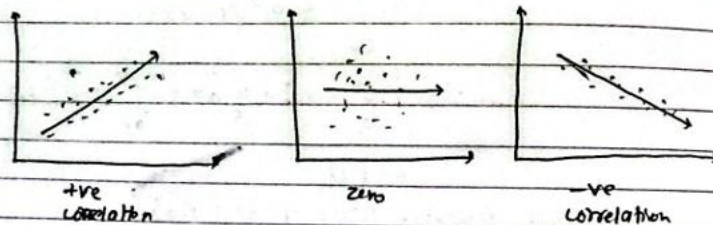
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y) dx dy$$

μ_x and μ_y = mean of variables X & Y

$f(x, y)$ = joint pdf of X and Y

- $\text{Cov}(X, Y)$ is also given as

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$



3.3 Chebyshev's inequality

- Knowing mean and SD of a variable, approx probability can be obtained by using Chebyshev's inequality.
- It states that for any $k > 1$ at least $1 - 1/k^2$ of data from a sample must fall within k standard deviations from the mean.
- Chebyshev's inequality, in probability theory, a theorem that characterizes the dispersion of data away from its mean (avg.)
- $P[(\mu - k\sigma) \leq X \leq (\mu + k\sigma)] \geq [1 - \frac{1}{k^2}]$

where, μ = Mean

σ = SD of a variable

- For $k = 2$, we have, $1 - \frac{1}{k^2} = 75\%$

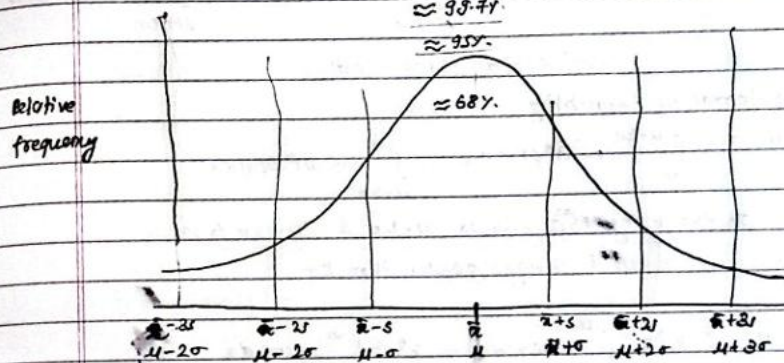
So, Chebyshev's inequality says that at least 75% of the data values of any distribution must be within two standard deviations of the mean.

Similarly, for $k = 3 \Rightarrow 1 - \frac{1}{k^2} = 89\%$

$\approx 98.7\%$

$\approx 95\%$

$\approx 68\%$



→ Chebyshev's inequality is a probability theory that guarantees only a definite fraction of values will be found within a specific distance from the mean of a distribution. The fraction for which no more than a certain number of values can exist is represented by $1/k^2$.

* Proof:

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq \left(1 - \frac{1}{k^2}\right) \text{ where } \mu = \text{Mean} \\ \sigma = \text{SD of a variable}$$

Proof:

$$\Rightarrow \sigma^2 = \text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

$$\sigma^2 = \int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu - k\sigma}^{\mu + k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) dx$$

$\int_{\mu - k\sigma}^{\mu + k\sigma} (x - \mu)^2 f(x) dx \rightarrow \text{non-negative}$

In terms of inequality

$$\sigma^2 \geq \int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) dx$$

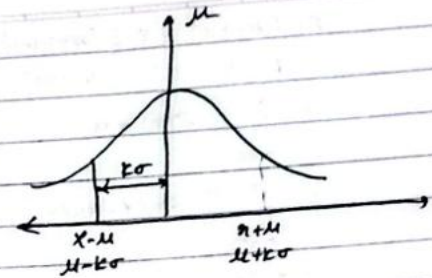
In the range of $-\infty$ to $\mu - k\sigma$ & $\mu + k\sigma$ to ∞

$|x - \mu|$ is always greater than $k\sigma$

$$\sigma^2 \geq k^2 \sigma^2 \left[\int_{-\infty}^{\mu - k\sigma} f(x) dx + \int_{\mu + k\sigma}^{\infty} f(x) dx \right]$$

$$\sigma^2 \geq k^2 \sigma^2 [P(X \leq \mu - k\sigma) + P(X \geq \mu + k\sigma)]$$

$$\left\{ P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq \left(1 - \frac{1}{k^2}\right) \right\}$$



$$(9) f(x) = 6x(1-x) \text{ for } 0 < x < 1 \\ = 0 \text{ elsewhere}$$

Find the probability that x will take on a value within 2 times standard deviation of the mean and compare it with the lower bound provided by Chebyshev's inequality.

→ Soln:

$$\text{Mean} = \mu = E(x) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x f(x) dx$$

$$= \int_0^1 x \cdot 6x(1-x) dx$$

$$= 0.5$$

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

$$= \int_0^1 (x - 0.5)^2 6x(1-x) dx$$

$$= 1/20$$

$$\text{Standard deviation, } (\sigma) = \sqrt{1/20} = 0.2236$$

Now

$$\mu - k\sigma = 0.5 - 2 \times 0.2236 = 0.053$$

$$\mu + k\sigma = 0.5 + 2 \times 0.2236 = 0.947$$

$$P[(\mu - k\sigma) \leq X \leq (\mu + k\sigma)]$$

$$= P[0.53 \leq X \leq 0.94]$$

$$= \int_{0.53}^{0.94} f(x) dx$$

$$= \int_{0.53}^{0.94} \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}(\mu - x)^2} dx$$

$$= 0.91$$

The lower bound from Chebyshev's inequality.

$$P[(\mu - k\sigma) \leq X \leq (\mu + k\sigma)] \geq 1 - \frac{1}{k^2}$$

$$\geq 1 - \frac{1}{2^2} \geq 0.75$$

8. The annual rainfall runoff of a stream has a mean of 200 and standard deviation of 40 m³/s respectively. Find range within which the runoff will lie with a probability of at least 0.5. Use Chebyshev's inequality.

- Solⁿ:

$$\mu = 200 \text{ m}^3/\text{s}$$

$$\sigma = 40 \text{ m}^3/\text{s}$$

$$P[(\mu - k\sigma) \leq X \leq (\mu + k\sigma)] = 0.5$$

From Chebyshev's inequality

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

$$0.5 \geq 1 - \frac{1}{k^2}$$

$$[k = 1.414]$$

$$\mu - k\sigma = 200 - 40k = 143.4 \text{ m}^3/\text{s}$$

$$\mu + k\sigma = 200 + 40k = 256.6 \text{ m}^3/\text{s}$$

Range of runoff: 143.4 m³/s to 256.6 m³/s.

$M_X(t) = E[e^{tx}]$
 find mean, variance & higher moments
 1st derivative = mean; 2nd derivative = variance
 MGF shows all statistical information
 $M'(0) = \text{Mean}$; $M''(0) = \text{Variance}$

3.4 Moment Generating Function (MGF)

- Moment Generating Functions are a way to find moments like the mean (μ) and variance (σ^2)
- Are an alternative way to represent a probability distribution with a single one-variable function

If X is a discrete RV, its moment generating function, when it exists is given by

$$M_X(t) = E[e^{tx}] = \sum_x e^{tx} p(x)$$

If X is a continuous RV, then

$$M_X(t) = E[e^{tx}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

The n^{th} moment about origin is given by n^{th} derivative of $M_X(t)$ w.r.t. t and evaluated at $t=0$

$$\mu_n' = \frac{d^n M_X(t)}{dt^n} \text{ at } t=0$$

Proof:-

Expanding e^{tx} by Maclaurin's series

$$M_X(t) = \sum_x \left[1 + tx + \frac{t^2 x^2}{2!} + \frac{t^3 x^3}{3!} + \dots \right] p(x)$$

$$= \sum p(x) + t \sum x p(x) + \frac{t^2}{2!} \sum x^2 p(x) + \dots + \frac{t^n}{n!} \sum x^n p(x) + \dots$$

$$= \sum p(x) + t \sum x p(x) + \frac{t^2}{2!} \sum x^2 p(x) + \dots + \frac{t^n}{n!} \sum x^n p(x) + \dots$$

$$1 + t \mu_1' + \frac{t^2}{2!} \mu_2' + \dots + \frac{t^n}{n!} \mu_n' + \dots$$

$$\frac{d^n M_X(t)}{dt^n} = \mu_n' + \text{terms of } t$$

$$\mu_n' = \frac{d^n M_X(t)}{dt^n} \text{ at } t=0$$

A Maclaurin series is a Taylor series about 0.
 $f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \dots + \frac{f^{(n)}(0)}{n!}x^n + \dots$

(Q9) The PDF $f(x)$ of a RV is given by

$$f(x) = 2x \text{ for } 0 < x < 2$$

Find moment generating function of this variable

→ Solⁿ:

$$M_x(t) = E[e^{tx}] = \int_0^2 e^{tx} f(x) dx$$

$$= \int_0^2 e^{tx} \cdot 2x dx$$

$$u = 2x \text{ \& } v = e^{tx}$$

$$du = 2 dx \text{ \& } dv = \frac{e^{tx}}{t} dx$$

Then

$$\int u dv = uv - \int v du = \left| \frac{2}{t} \cdot x e^{tx} - \frac{2}{t^2} e^{tx} \right|_0^2$$

$$M_x(t) = \frac{4}{t} e^{2t} - \frac{2}{t^2} e^{2t} + \frac{2}{t^2}$$

$$\int u dv = uv - \int \left(\frac{du}{dx} v \right) dx$$

3.5 Normal (Gaussian) distribution

- The graph of the Gaussian distribution depends on two factors - the mean and the standard deviation
- The mean of the distribution determines the location of the centre of the graph, and the standard deviation determines the height and width of the graph
- Normal (Gaussian) distribution is a probability distribution that is symmetric about the mean, showing the data near the mean are more frequent in occurrence than data far from the mean
- Is a widely used distribution because it fits many natural phenomena and easily describe like, hydrological variables of large intervals such as annual runoff, annual rainfall.
- Random errors associated with any hydrological measurements can be approximated by normal distribution.
- Also useful in hypothesis testing

→ The PDF of normal distribution is given by

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

for $-\infty < x < \infty$, for $-\infty < \mu < \infty$, $\sigma > 0$

μ = location parameter

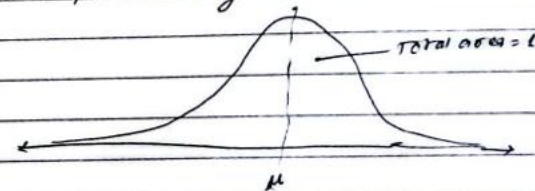
σ = scale parameter

1. The mean, median, and mode are equal

2. The normal curve is bell-shaped & symmetric about the mean

3. The total area under the curve is equal to one

4. The normal curve approaches, but never touches the x-axis as it extends farther and farther away from the mean.



The cdf of the x can be expressed by

$$P[X \leq x] = P(u) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du$$

This exp cannot be evaluated analytically. Alternative way is to transform $N(\mu, \sigma)$ to standardized form. If $Z = \frac{x-\mu}{\sigma}$, then mean of $Z = 0$, variance = 1.

The variable Z is known as standard normal variate ($N(0, 1)$).

The pdf of Z is

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

The cumulative distribution of Z is

$$P[Z \leq z] = F(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

$$P[x_1 \leq x \leq x_2] = P\left[\frac{x_1 - \mu}{\sigma} \leq Z \leq \frac{x_2 - \mu}{\sigma}\right] = P[z_1 \leq Z \leq z_2]$$

$$P[x_1 \leq x \leq x_2] = F(z_2) - F(z_1)$$

$F(z)$ is found from tables

(910)

The annual runoff of a stream follows normal distribution with mean and SD of 190 and 256 m³/s. Find the probability that the runoff exceeds 400 m³/s in any year. Find the probability that the runoff would be b/w 300 and 500 m³/s.

→ Solⁿ

(a) $x = \text{runoff}$

$$Z = \frac{x - \mu}{\sigma} = \frac{x - 190}{256}$$

$$P[X \geq 400] = P\left[Z \geq \frac{400 - 190}{256}\right] = P[Z \geq 0.82]$$

$$= 1 - P[Z \leq 0.82]$$

$$= 1 - F(0.82)$$

$$= 1 - 0.7939 \quad (\text{from table})$$

$$= 0.2061$$

(b)

$$P[300 \leq x \leq 500] = P\left[\frac{300 - 190}{256} \leq Z \leq \frac{500 - 190}{256}\right]$$

$$= P(0.43) - P(-1.21) = F(1.21) - F(0.43)$$

$$= 0.8869 - 0.6664 \quad (\text{from table})$$

$$= 0.22$$

* Joint distribution → relationship

Marginal → individual behaviour

Conditional → dependence

Desired → transformation

MGF → extract properties

CLT → simplify analysis

For large sample size: $X \sim N(\mu, \sigma^2/n)$
 even if data not normal & sample mean becomes normal
 eg: Rainfall distribution - not normal
 But, Avg. rainfall over many years - normal

3.6 Central Limit Theorem (CLT)

It states that, given a sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution.

The CLT states that, "If there is a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample mean will be approximately normally distributed."

It states that if a sequence of RV x_i are independently and identically distributed with mean μ and variance σ^2 , then the distribution of the sum of n such RV follows normal distribution with mean $n\mu$ and variance $n\sigma^2$ as n becomes large.

→ If a daily event such as daily runoff or daily precipitation is considered to be RV, the corresponding annual event is a result of the summation of 365 random variables. Though the daily events are not truly independent, since 365 is a fairly large number, by CLT annual hydrologic variables can be taken to follow normal distribution. Random errors associated with any hydrologic measurement are usually symmetrical distributed and can be well approximated by the normal distribution.

3.7 Estimating the parameters of probability distribution

a. Method of moments (Karl Pearson method)

→ The expected value of RV is population moment whereas the sample moment is computed from the given time series data.

→ The function parameters of time series model are functions of population moments.

→ Therefore, the sample moment is equated to population moment and the resulting equations are solved to get the parameters. For k no. of parameters, first k population and sample moments are equated and solved simultaneously.

b. Method of maximum likelihood (Fisher method)

→ Likelihood of a set of data is the probability of obtaining that particular set of data, given the chosen probability distribution model.

→ This expression contains the unknown model parameters.

→ The values of these parameters that maximize the sample likelihood are known as Maximum Likelihood Estimates.

→ Maximum likelihood estimation is a method that determines values for the parameters of a model.

→ The parameter values are found such that they maximize the likelihood that the process described by the model produced the data that were actually observed.

→ According to this method, the best value of the parameter of a distribution should be that value which maximizes the likelihood function or joint probability of occurrence of the observed sample.

$f(x, \alpha, \beta, \dots)$: pdf of random variable X with parameters $\alpha, \beta, \dots$

$(x_1, x_2, \dots, x_n)$: observed sample

likelihood function = $f(x_1, \alpha, \beta, \dots) f(x_2, \alpha, \beta, \dots) \dots f(x_n, \alpha, \beta, \dots)$

$$O_2, L(\alpha, \beta, \dots) = \prod_{i=1}^n f(x_i, \alpha, \beta, \dots)$$

Find the values of $\alpha, \beta, \dots$ which maximize the likelihood function.

It is convenient to work with log-likelihood.

$$\ln[L(\alpha, \beta, \dots)] = \ln\left[\prod_{i=1}^n f(x_i, \alpha, \beta, \dots)\right]$$

$$= \sum_{i=1}^n \ln[f(x_i, \alpha, \beta, \dots)]$$

$$\frac{\partial \ln[L(\alpha, \beta, \dots)]}{\partial \alpha} = 0$$

$$\frac{\partial \ln[L(\alpha, \beta, \dots)]}{\partial \beta} = 0$$

Ex (9.11) The pdf of a RV is given by $f(x) = \lambda e^{-\lambda x}$ for $x > 0$, find the parameter λ of the distribution by (a) Method of moments (b) Method of maximum likelihood.

→ Sol:

(a) Method of moments

As there is only one parameter, only first moment is calculated

$$\mu = \int_{-\infty}^{\infty} x f(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx$$

$$u = \lambda dx; dv = \lambda dx$$

$$dv = e^{-\lambda x} dx$$

$$v = -\frac{1}{\lambda} e^{-\lambda x}$$

$$\int u dv = uv - \int v du = \left| -x e^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x} \right|_0^{\infty}$$

$$\mu = \frac{1}{\lambda}$$

$$\lambda = \frac{1}{\mu}$$

Method of maximum likelihood

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

log likelihood function

$$\ln L = \sum_{i=1}^n \ln(f(x_i))$$

$$= \sum_{i=1}^n \ln[\lambda e^{-\lambda x_i}]$$

$$= n \ln \lambda - \lambda \sum_{i=1}^n x_i$$

For max value of L ,

$$\frac{d(\ln L)}{d\lambda} = 0$$

$$\text{or } n - \sum_{i=1}^n x_i = 0$$

$$\text{or } \lambda = \frac{1}{\frac{1}{n} \sum_{i=1}^n x_i} = \frac{1}{\bar{x}} = \frac{1}{\mu}$$

Selection of distribution

a. probability plotting

→ Plot b/w RV and probability of exceedance

• Method

→ Arrange the data in descending order and give rank m starting from 1

→ Compute plotting position, $P[X \geq x_m]$ and return period, $T(1/p)$

→ Plot given data versus P or T on probability paper

→ If the plot is made on probability paper of an assumed distribution and if the plot is straight line, then the chosen distribution is valid for given data.

→ Such plot is also useful for interpolation or extrapolation.

→ California formula for plotting position: $P(X \geq x_m) = m/n$
(simplest formula)

→ Weibull's formula for plotting position: $P(X \geq x_m) = \frac{m}{n+1}$

where, $m = \text{rank}$

$n = \text{Total number of values}$

6. Testing goodness of fit

(i) Chi square test

Used to determine whether there is a statistically significant difference b/w the expected frequency and the observed frequency.

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

where O_i = no. of observation in a class interval

E_i = expected no. of observation in a class interval

No. of degree of freedom = $k - h - 1$

where h = no. of parameters to be estimated

k = no. of class intervals

→ Divide a RV into k class intervals

→ Select a probability distribution whose adequacy is to be tested

→ obtain its parameters

→ Determine probability P_i with which the random variable lies in each class interval

$$E_i = n p_i$$

→ compute χ^2

→ Find the critical value of χ^2_{α} for $k-h-1$ degree of freedom and α level of significance from table ($\alpha = 5\%$ or 10%)

→ If computed χ^2 is less than critical value of χ^2_{α} , the assumed distribution is good fit

Q.12 The annual peak discharge of a river basin for a period of 39 years is given below:

2880 3090 4050 3220 1630 1840 450 1010 650 1630 430
380 510 1000 350 630 1450 2670 510

Test the goodness of fit of normal distribution to the observed data at 5% significance level by chi-squared test

→ Soln:

$$n = 39$$

$$\text{Mean: } \bar{x} = \frac{1}{n} \sum_{i=1}^n (x_i) = 1662$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = 1350$$

$F(z_i)$: Obtained from table

Expected no. of observation = $E_i = n p_i$

Probability of class interval
 $= P_i = F(z_i) - F(z_{i-1})$

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = 6.73$$

Dividing the data into class interval of 1000

i	Class Interval	n_i	z_i of upper class	$F(z_i)$	P_i	E_i	χ^2
1	<1000	8	-0.49	0.31	0.31	5.89	0.76
2	1001-2000	6	0.25	0.6	0.29	5.51	0.04
3	2001-3000	2	0.99	0.84	0.24	4.56	1.44
4	3001-4000	1	1.73	0.96	0.12	2.28	0.72
5	4001-5000	1	2.47	0.99	0.03	0.57	0.32
6	>5000	1		1	0.01	0.19	0.22
	Sum	19			1	19	6.73

No. of class interval (k) = 6, No. of parameters (h) = 2,

No. of degree of freedom = $k - h - 1 = 3$

For 5% significance level and 3 degree of freedom, $\chi^2 = 7.81$ (from table)

As computed χ^2 is less than the tabulated value, the normal distribution is appropriate for the data.

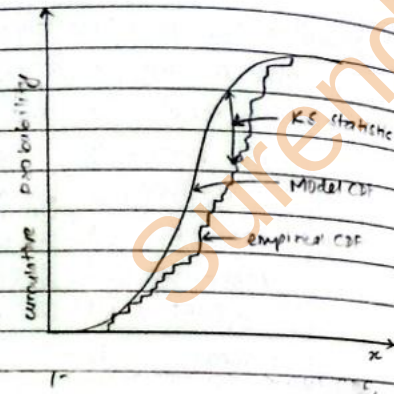
(ii) Kolmogorov-Smirnov test (K-S test or KS test)

- The Kolmogorov-Smirnov test is used to decide if a sample comes from a population with a specific distribution.
- The Kolmogorov-Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution.
- Compute cumulative probability ($P(x_m)$) by Weibull's formula
 $[P(x_m) = 1 - \text{exceedance prob.}]$
- Compute theoretical cum. probability $F(x_m)$ for the assumed distribution.
- Compute the Smirnov-Kolmogorov test statistic.

$$\Delta = \text{maximum of } |P(x_m) - F(x_m)|$$

For 5% or 10% level of significance, obtain the value of Smirnov-Kolmogorov statistic Δ_0 for sample size n from table.

If $\Delta < \Delta_0$, the assumed distribution is a good fit for the data.



13 Test the goodness of fit of normal distribution to the data of QN 12 at 10% significance level by Smirnov-Kolmogorov test.

Solution:

$$n = 19$$

$$\text{Mean, } \bar{x} = \frac{1}{n} \sum_{i=1}^n (x_i) = 1662$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = 1350$$

$\Delta = \text{maximum of } |P(x_m) - F(x_m)|$ ← from Z-table

$$Z = \frac{(X - \text{mean})}{\text{SD}}$$

$$P(x_m) = 1 - \text{exceedance probability}$$

Given data	x	$\frac{i}{n+1}$	Rank (m)	Exceedance prob.	$P(x)$	Z	$F(x) = P(Z)$	delta
2860	5140	1	1	0.05	0.95	2.58	0.99	0.04
3090	4050	2	2	0.1	0.9	1.77	0.96	0.06
4050	3090	3	3	0.15	0.85	1.06	0.85	0
1220	2860	4	4	0.20	0.8	0.89	0.81	0.01
1630	2670	5	5	0.25	0.75	0.75	0.77	0.02
1840	1840	6	6	0.3	0.7	0.13	0.55	0.15
450	1630	7	7	0.35	0.65	-0.02	0.5	0.15
1610	1630	8	8	0.4	0.6	-0.02	0.5	0.15
650	1610	9	9	0.45	0.55	-0.04	0.48	0.08
1630	1450	10	10	0.5	0.5	-0.16	0.44	0.06
430	1220	11	11	0.55	0.45	-0.33	0.37	0.08
360	1000	12	12	0.60	0.4	-0.49	0.31	0.09
510	650	13	13	0.65	0.35	-0.75	0.23	0.02
1000	630	14	14	0.70	0.3	-0.76	0.23	0.03
350	510	15	15	0.75	0.25	-0.85	0.2	0.05
630	450	16	16	0.80	0.20	-0.90	0.19	0.03
1450	430	17	17	0.85	0.15	-0.91	0.18	0.03
2670	360	18	18	0.90	0.10	-0.96	0.17	0.07
5140	350	19	19	0.95	0.05	-0.97	0.17	0.12

Smirnov-Kolmogorov test static (D) = $\max |P(X_m) - F(X_m)| = 0.15$
 For $n=19$ and 10% level of significance, D from table = 0.244
 As calculated $D <$ tabulated D , the normal distribution is good fit for the data.

3.9 Some commonly used probability distributions

Discrete

a. Binomial

→ Probability of occurrence of event x times in n successive years is found by Binomial distribution.

$$P_{x,n} = C(n, x) p^x q^{n-x} = \frac{n!}{(n-x)! x!} p^x q^{n-x}$$

where, p = probability of occurrence, $q = 1-p$

b. Geometric

$$p(x) = q^{x-1} p$$

n = no. of trials till the first success is met.

c. Poisson

When n tends to be very large and p tends to be very small,

$$P_{x,n} = C(n, x) p^x q^{n-x} \rightarrow \frac{\lambda^x e^{-\lambda}}{x!} \text{ where, } np \rightarrow \lambda$$

Continuous distributions

1. Uniform

$$f(x) = \frac{1}{b-a}$$

2. Normal

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

Parameters: μ, σ

3. Log-normal

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu_y)^2}{2\sigma_y^2}\right) \text{ where } y = \log x$$

Parameters: μ_y, σ_y

4. Exponential

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

Parameter: λ

5. Gamma

$$f(x) = \frac{\lambda^\beta x^{\beta-1} e^{-\lambda x}}{\Gamma(\beta)}$$

Parameters: λ, β

6. Pearson type III

$$f(x) = \frac{\lambda^\beta (x-\epsilon)^{\beta-1} e^{-\lambda(x-\epsilon)}}{\Gamma(\beta)}$$

Parameters: λ, β, ϵ

7. Log-Pearson type III

$$f(x) = \frac{\lambda^\beta (y-\epsilon)^{\beta-1} e^{-\lambda(y-\epsilon)}}{\Gamma(\beta)} \text{ where, } y = \log x$$

Parameters: λ, β, ϵ

8. Extreme value I (Gumbell)

$$f(x) = \frac{1}{\alpha} \exp\left[-\frac{x-u}{\alpha} - \exp\left(\frac{x-u}{\alpha}\right)\right]$$

parameter: α, u

Ch 4

Auto correlation Analysis

Correlation

↳ correlation determines the degree of association b/w two variables.

The association b/w two variables is called covariation.

Correlation measures the degree of linear dependence b/w two variables.

Population correlation coefficient (r) = $\frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$

Sample correlation coefficient (r) = $\frac{S_{xy}}{S_x S_y}$

where, S_{xy} = sample covariance, S_x, S_y = sample SD of X and Y.

If (x_i, y_i) is the i th observation of random variables X and Y,

and $\bar{x}$ and $\bar{y}$ are the sample means, then r is given by

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}$$

Higher value of one variable associated with higher value of other or lower value of one variable associated with lower value of other gives +ve correlation.
Independent variables: 0 correlation.

Regression

↳ Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation b/w A and B is same as the correlation b/w B and A.

↳ Regression Analysis produces a regression equation where the coefficients represent the relationship between each independent variable and the dependent variable.

↳ It is the relationship b/w dependent and independent variables.

It represents the cause and effect relationship, which is described by polynomial equation.

↳ Multiple regression consists of more than one independent variable.

$$Y = a + bX$$

where,

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$y = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

$$b_1 = \frac{(\sum x_1^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

* Autocorrelation

↳ correlation of a variable with its past and future values (correlation in itself)

↳ correlation is b/w two values of the same variable at different times x_t and x_{t+k} . It is called the Autocorrelation.

↳ When autocorrelation is used to detect non-randomness, it is usually only the first (lag 1) autocorrelation that is of interest. When the autocorrelation is used to identify the appropriate time series model, the autocorrelations are usually many lags according to requirements.

↳ Autocorrelation coefficient or Serial correlation coefficient is given by

$$r_k = \frac{\frac{1}{n-k} \sum_{t=1}^{n-k} (x_t - \bar{x}_t)(x_{t+k} - \bar{x}_{t+k})}{\sqrt{\frac{1}{n-k} \sum_{t=1}^{n-k} (x_t - \bar{x}_t)^2} \sqrt{\frac{1}{n-k} \sum_{t=1}^{n-k} (x_{t+k} - \bar{x}_{t+k})^2}}$$

$$\text{or, } r_k = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x}_t)(x_{t+k} - \bar{x}_{t+k})}{\sqrt{\sum_{t=1}^{n-k} (x_t - \bar{x}_t)^2} \sqrt{\sum_{t=1}^{n-k} (x_{t+k} - \bar{x}_{t+k})^2}}$$

$$\bar{x}_t = \frac{1}{n-k} \sum_{i=1}^{n-k} x_t \quad \text{and} \quad \bar{x}_{t+k} = \frac{1}{n-k} \sum_{i=1}^{n-k} x_{t+k}$$

* Correlogram (plot of $\log(k)$ vs r_k)

Case I: Independent process $\rightarrow r_k \approx 0$ for all $k > 0$

Case II: AR(1) $\rightarrow$ gradual decay

Case III: MA(1) $\rightarrow$ cut-off after lag 2

Case IV: Periodic $\rightarrow$ oscillating pattern

* Independence check:

$r_k \approx 0$ for all $k \neq 0$
Use confidence limit $\pm \frac{1.96}{\sqrt{n}}$ $\rightarrow$ All r_k within band $\rightarrow$ independent
if any outside band $\rightarrow$ dependent

If $n \gg k_{\max}$ (for large n), mean and variance of whole series can be used then simplified formula is

$$r_k = \frac{\frac{1}{n-k} \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\frac{1}{n} \sum_{t=1}^n (x_t - \bar{x})^2}$$

where, k = time lag,

n = no of data

r_k can also be computed using,

$$r_k = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}$$

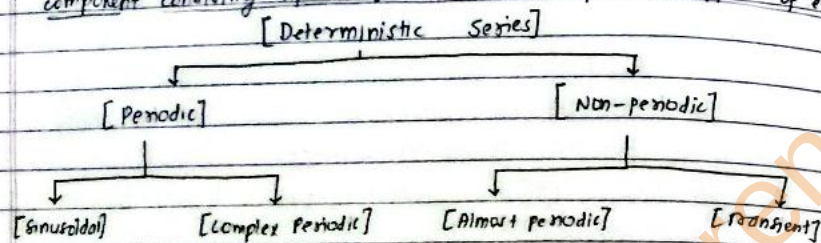
* The plot of serial correlation coefficient with time lag is called correlogram or Autocorrelation function (ACF).

* Partial Autocorrelation function (PACF) is the plot of serial correlation by neglecting some values

Partial Autocorrelation function (PACF) is the plot of serial correlation by neglecting some values.

4.2 Classification of time series considering deterministic and stochastic process

4. A time series is an observed realization of the stochastic process, and is the basis for analysis, identification and description of the process.
1. Any time series of a random phenomenon can generally be considered to consist of a deterministic component which is the one that can be determined uniquely and a stochastic component consisting of chance and chance dependent effects of event.



a. Sinusoidal series

Sinusoidal data can be represented by

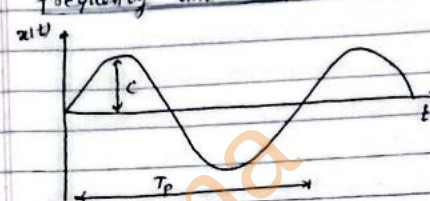
$$x(t) = C \sin(2\pi f_0 t + \Phi)$$

where,

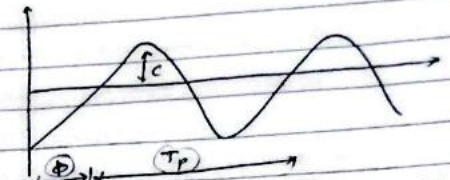
- ✓ $x(t)$ = instantaneous value at t ,
- ✓ C = amplitude,
- ✓ f_0 = frequency
- ✓ Φ = initial phase angle with respect to time origin

eg voltage dp of generator.

The time interval required for one full fluctuation or for one full cycle is called the frequency denoted by f_0 . The frequency and the period is related by



Sinusoidal series (with phase angle)



Sinusoidal series (with phase angle)

b. Complex periodic series

Series which repeats itself at regular intervals is the periodic series.

$$x(t) = x(t + nT_p);$$

$$n = 1, 2, 3, \dots$$

where $x(t)$ = instantaneous value of t and

T_p = period

Such series can be represented by Fourier series

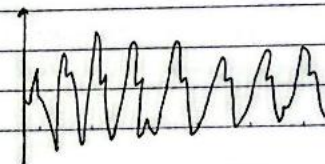
$$x(t) = \bar{x} + \sum_{n=1}^{\infty} C_n \sin(2\pi f_1 t + \Phi_n)$$

$\bar{x}$ = mean of series

f_1 = fundamental frequency

C_n and Φ_n = coefficient

Complex periodic series = static component + infinite no. of sinusoidal components with



C_n = amplitude,

Φ_n = phase angle

II Non-periodic $\rightarrow$ No fixed repeating pattern
eg: Flood events; sudden rainfall bursts

a. Almost periodic series

↳ there is no common factor for frequency and no period (T_p)
↳ physical phenomena producing almost periodic series frequently occur in practice when the effect of two or more unrelated periodic phenomena are mixed

↳ Eg: the tides composed of almost periodic and stochastic elements are tidal levels along coasts and bays. The three cycles which affect the tides namely the earth's daily rotation, the moon's revolution around the earth and the earth's revolution around the sun don't have equal frequency.

↳ Such series can be represented by:

$$x(t) = \bar{x} + c_1 \cos(2\pi f_1 t + \phi_1) + c_2 \cos(2\pi f_2 t + \phi_2) + \dots$$

→ ✓ shows repeating behaviour
✓ but not perfectly periodic.

→ Features:

- Multiple frequencies
- Irregular repetition
- Approximate cycles

→ Eg: Monsoon varies each year; pattern exists but not exact

→ 100% periodic but peak timing changes; Amplitude varies

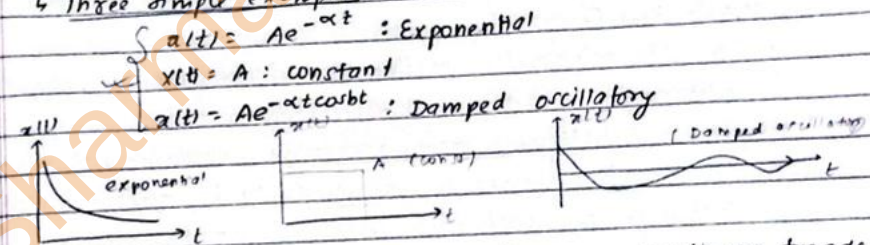
Feature	Almost periodic	Temporary short-term sudden change
Pattern duration	Repeating long-term cyclic	
Nature		

b. Transient series

↳ All non-periodic data other than almost periodic is called transient series. Such series is represented by time varying function. The physical phenomena which produce transient series have numerous process and drivers

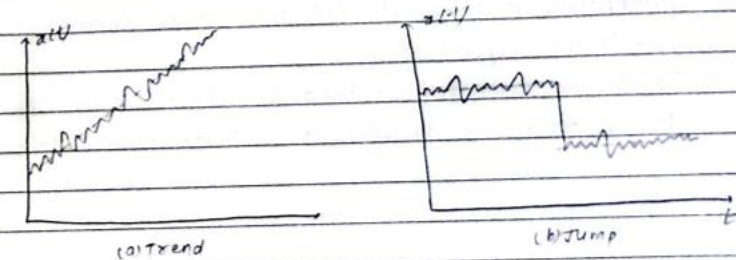
↳ Such series are formed causally in the case of hydrological series, superimposed on the other such as periodic and stochastic components

↳ Three simple examples:



↳ Another simple examples are linear or nonlinear trends and jumps in the hydrological series

↳ Trends are linear or non-linear and created by the slow changes. Jumps are created by the sudden changes in the basin.



→ ✓ series showing temporary change.

→ Types: Trend; sudden jump; structural change

→ Eg: sudden flood spike; climate shift; dam construction effect

→ Exists for short time then disappears

2. Stochastic series

- ↳ Is described by the probability distribution
- ↳ The most widely used distribution in time series analysis for describing random variable is the normal distribution.
- ↳ The observed series are fit in the normal distribution and analyzed it.

Stationary series

→ Stationary stochastic series are time invariant (does not depend upon time change)

→ In Hydrology, the mathematical techniques for analyzing stationary series are well developed.

→ After identifying and removing deterministic components, the resulting series is considered to be approximately stationary. The stationary series is divided into dependent and independent series.

→ The independent stationary series is called the pure random series and easily described through the single probability distribution function.

→ Dependent series is again divided into linear dependent and non-linear dependent.

→ The linear dependent stationary stochastic series are analyzed based on moving average and autoregressive series.

Non-stationary series

It is time variant series for a long duration so the statistical parameters are changed with the long run like river runoff, climate change, etc. events.

Autoregressive Integrated moving average series represents a kind of non-stationary series for which the analysis techniques have been developed.

Components of time series

Observed time series = Stochastic component + Trend component + Periodic component + Catastrophic event

$$x(t) = x'(t) + x''(t) + x'''(t) + x^{(4)}(t)$$

Autocorrelation analysis for the investigation of independent process
→ It is very extensively used to determine the linear dependence in the existing series

The method consists of estimating the autocorrelation coefficients for various lags and plotting them against the respective lags resulting in what is known as an empirical correlogram.

The empirical correlogram thus obtained is then compared with the theoretical correlogram of known process to identify a suitable mathematical model to describe the process under consideration.

- (a) The theoretical Autocorrelation function (population data) γ_k for an independent and random process is zero for all lags except zero lag. (for zero lag, it is 1.)
- (b) The ACF for observed data (sample data) will not be exactly equal to zero due to sampling effects. The ACF fluctuates around zero for independent sample data.

→ To verify this fact the sampling distribution of γ_k is required. Anderson has proposed a test of significance for the autocorrelation coefficients of the time series.

In a random, normally distributed series of length n , if n is large in comparison to k ($n \gg k$) γ_k is found to be normally distributed with a mean of $-(n-k)^{-1}$ and a variance of $(n-k-1)/(n-k)^2$.

The 95% confidence limits (+ for upper limit and - for lower limit)

$$CL(\gamma_k) = \frac{-1 \pm 1.96\sqrt{n-k-1}}{n-k}$$

If all the autocorrelation coefficients for various lags fall within the tolerance band, the sample data is considered to be independent.

Q. The 78 years annual runoff data is used to obtain the first 10 autocorrelation coefficients which are given below: Test the series for independence at 5% level of significance

lag k in years	1	2	3	4	5	6	7	8	9	10
r_k	0.27	0.18	0.22	0.00	0.12	0.24	0.22	0.05	0.00	0.00

$$r_{upper} = \frac{-1 + 1.96\sqrt{n-2}}{(n-1)} \text{ and } n=78$$

$$r_{upper} = \frac{-1 + 1.96\sqrt{76}}{77} = \frac{-1 + 17.087}{77} = 0.209$$

$$r_{lower} = \frac{-1 - 17.087}{77} = -0.235$$

Upper and lower tolerance limits of r_k are computed to be 0.209 and -0.235. The autocorrelation coefficients are falling outside the tolerance band the annual runoff series being analyzed is not independent.

r_k falls outside of band $\Rightarrow$ not independent

ii. correlogram of different type of series

- complex periodic with a period T_p : correlogram is periodic with same T_p and unit amplitude
- linear trend: horizontal line
- combination of periodic sine component and independent random component: correlogram is periodic with same frequency or one component, but the amplitude is not unity
- combination of complex periodic and independent random component: correlogram is periodic with frequency equal to fundamental frequency of complex periodic series

5. Time Series Models

The most natural phenomenon of a time series model is a linear regression approach on the past instances, i.e. a regression on the series itself.

The process term as autoregressive.

An autoregressive model of order m , abbreviated as $AR(m)$, is based on a linear combination of past observations according to the following general equation.

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_m X_{t-m} + \dots + e_t \quad (1)$$

where, the value of variable at time t is equal to weighted sum of past values plus random component (white noise).

where, $\alpha_1, \alpha_2, \dots, \alpha_m$ = parameters of model,

e_t = random component with zero mean

Above equation is written as $AR(m)$ where

m = order of model. (no. of lag times considered)

If the process is stationary with the mean and variance are same in the process

The alternative form of above equation is

$$X_t - \mu = \alpha_1 (X_{t-1} - \mu) + \alpha_2 (X_{t-2} - \mu) + \dots + \alpha_m (X_{t-m} - \mu) + \dots + e_t \quad (1)$$

where, μ = expected value of the process or mean value of the process

Yule-Walker equations

Multiplying both sides of eqn. (1) by $(X_{t-k} - \mu)$ and taking expectations:

$$E[(X_t - \mu)(X_{t-k} - \mu)] = \alpha_1 E[(X_{t-1} - \mu)(X_{t-k} - \mu)] + \alpha_2 E[(X_{t-2} - \mu)(X_{t-k} - \mu)] + \dots + \alpha_m E[(X_{t-m} - \mu)(X_{t-k} - \mu)] + \dots +$$

$$E[e_t (X_{t-k} - \mu)] \quad (11)$$

$$E[e_t (X_{t-k} - \mu)] = 0, \text{ since } e_t \text{ and } X_{t-k} \text{ are independent whenever}$$

$$k > 0.$$

Yule-Walker eq.

$E[(X_t - \mu)(X_{t-k} - \mu)]$ is nothing but $\text{cov}(X_t - \mu, X_{t-k} - \mu) = \gamma(k-j)$
 $\gamma(k) = \alpha_1 \gamma(k-1) + \alpha_2 \gamma(k-2) + \dots + \alpha_m \gamma(k-m) \quad \text{--- (iv)}$

Dividing eq. (iv) by $\gamma(0)$ (variance of X_k).

$$\rho_k = \alpha_1 \rho_{k-1} + \alpha_2 \rho_{k-2} + \dots + \alpha_m \rho_{k-m} \quad \text{--- (v)}$$

$$\frac{\gamma(k)}{\gamma(0)} = \rho_k$$

where, ρ = population autocorrelation coefficient

For $k = 1, 2, \dots, m$, above eq. can be expanded as
 (with $\rho_0 = 1$ and $\rho_{-j} = \rho_j$ for stationary process)

$$\rho_1 = \alpha_1 + \alpha_2 \rho_1 + \dots + \alpha_m \rho_{m-1}$$

$$\rho_2 = \alpha_1 \rho_1 + \alpha_2 + \dots + \alpha_m \rho_{m-2}$$

$$\dots$$

$$\rho_m = \alpha_1 \rho_{m-1} + \alpha_2 \rho_{m-2} + \dots + \alpha_m$$

Dividing eq. (iv) by $\gamma(0)$ (variance of X_k)
 $\rho_k = \alpha_1 \rho_{k-1} + \alpha_2 \rho_{k-2} + \dots + \alpha_m \rho_{k-m}$

These are known as Yule-Walker equations. The parameters can be found by simultaneously solving above equations taking population autocorrelation coefficient equal to the sample autocorrelation coefficient.

Relationship b/w variance of variable X and random components e .

For $k = 0$ in eq. (iii)

$$\gamma(0) = \alpha_1 \gamma(1) + \alpha_2 \gamma(2) + \dots + \alpha_m \gamma(m) + E[e_t (X_t - \mu)] \quad \text{--- (vi)}$$

$$E[e_t (X_t - \mu)] = E[e_t \{ \alpha_1 (X_{t-1} - \mu) + \alpha_2 (X_{t-2} - \mu) + \dots + \alpha_m (X_{t-m} - \mu) + e_t \}]$$

$$= \alpha_1 E[e_t (X_{t-1} - \mu)] + \alpha_2 E[e_t (X_{t-2} - \mu)] + \dots + \alpha_m E[e_t (X_{t-m} - \mu)] + E[e_t e_t]$$

Substituting value of $(X_t - \mu)$ from eq. (i)

$$E[e_t (X_t - \mu)] = 0 + 0 + \dots + 0 + \sigma_e^2 \quad \text{--- (vii)}$$

from eq. (vi) and (vii)
 $\gamma(0) = \alpha_1 \gamma(1) + \alpha_2 \gamma(2) + \dots + \alpha_m \gamma(m) + \sigma_e^2$ where,

σ_e^2 = variance of independent process

Dividing it by $\gamma(0)$

$$1 = \alpha_1 \rho_1 + \alpha_2 \rho_2 + \dots + \alpha_m \rho_m + \frac{\sigma_e^2}{\gamma(0)} \quad \text{where,}$$

$\gamma(0) = \sigma_x^2$ = variance of AR(m) process

$$\sigma_x^2 = \frac{\sigma_e^2}{(1 - \alpha_1 \rho_1 - \alpha_2 \rho_2 - \dots - \alpha_m \rho_m)} \quad \text{--- (viii)}$$

This is the relationship between of variance b/w AR(m) process and independent process

5.1.1 AR(1) Model: First order AR Model (Thomel - Fiering model)

- This is the most widely used model in Hydrology. AR(1) model is also called Thomel - Fiering model

$$(X_t - \mu) = \alpha_1 (X_{t-1} - \mu) + e_t$$

$$S_k = \alpha_1 S_{k-1} + \alpha_2 S_{k-2} + \dots + \alpha_m S_{k-m}$$

For $m=1$, ACF of AR(1) model is

$$S_k = \alpha_1 S_{k-1}$$

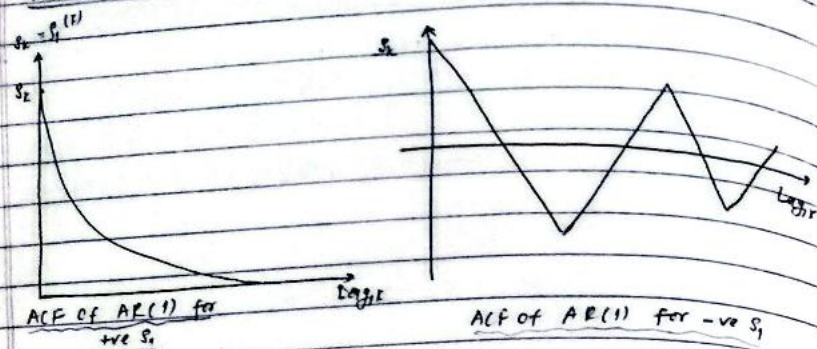
$$S_1 = \alpha_1, S_2 = \alpha_1^2$$

$$S_3 = \alpha_1 S_2 = \alpha_1^3, S_4 = \alpha_1^4$$

$$S_k = \alpha_1^k$$

α_1 = first serial correlation coefficient

- ACF of AR(1) process is completely described by ρ_1 . As, $\rho_k = \rho_1^k$ for stationary process ACF is 'symmetrical' about vertical axis.



- For +ve ρ_1 , ACF of AR(1) process decays exponentially to zero.
- For -ve ρ_1 , ACF of AR(1) oscillates about horizontal axis, decaying exponentially to zero.
- Hydrological data such as annual flow series with recession post decaying exponentially can be described by AR(1) with +ve ρ_1 .

AR(1) model with -ve ρ_1 cannot be used to model annual flow series, because for -ve ρ_1 , the series has to be zig-zag and the adjacent values in series should be negatively correlated, this situation is not found in annual series.

Relationship of variance in AR(1) model

$$\sigma_x^2 = \frac{\sigma_e^2}{(1 - \alpha_1 \rho_1 - \alpha_2 \rho_2 - \dots - \alpha_m \rho_m)}$$

For $m=1$,

$$\sigma_x^2 = \frac{\sigma_e^2}{(1 - \alpha_1 \rho_1)} = \frac{\sigma_e^2}{(1 - \rho_1^2)}$$

S1.2 Applicability of AR(1) model in Hydrology for generation and forecasting of annual flow (non-seasonal)

- If the first autocorrelation coefficient is significant and observed correlogram follows the trend of theoretical correlogram for AR(1) model, AR(1) model would be suitable.

$$X_t - \mu = \alpha_1 (X_{t-1} - \mu) + e_t, \text{ where } e_t \text{ has mean zero}$$

$$\alpha_1 = \rho_1$$

$$\sigma_e = \sigma_x \sqrt{(1 - \rho_1^2)}$$

If we consider another independent process Z_t with mean=0 and variance=1, then

$$e_t = Z_t \sigma_x \sqrt{(1 - \rho_1^2)}$$

Hence, AR(1) model can be written as

$$X_t - \mu = \rho_1 (X_{t-1} - \mu) + Z_t \sigma_x \sqrt{(1 - \rho_1^2)}$$

→ This is Thomas-Fiering model ★

- For forecasting annual flows, the population parameters are replaced by sample parameters. The annual flow model is

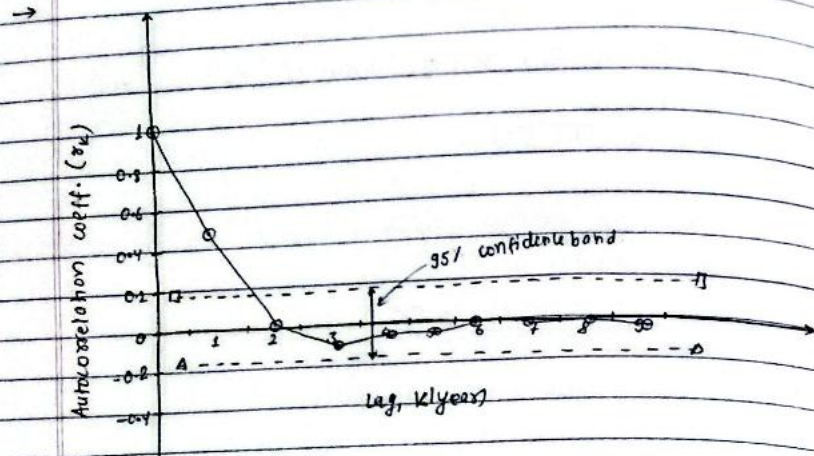
$$\star \left[\Phi_t = \bar{\Phi} + \rho_1 (\Phi_{t-1} - \bar{\Phi}) + Z_t \sigma_x \sqrt{(1 - \rho_1^2)} \right]$$

where, Φ_t = flow at time t $(X_t - \mu) = \rho_1 (X_{t-1} - \mu) + Z_t \sigma_x \sqrt{(1 - \rho_1^2)}$
 $\bar{\Phi}$ = mean of flow
 ρ_1 = first correlation coefficient

Z_t = random process with mean=0 and variance=1

The annual flow is assumed to be normally distributed

- Q The first 9 autocorrelation coefficients computed from 150 years of annual flow record at a river site are:
 $0.461, 0.005, -0.035, -0.053, -0.048, -0.01, -0.018, -0.016, -0.055$
 Test the flow series for independent at 95% confidence level.
 Also, suggest a suitable model for the series



From the plot of autocorrelation-lag, it is seen that autocorrelation coefficients for different lags do not fluctuate around zero. Hence, we can say that the series is not independent. It is further verified by computing confidence level.

Computing the confidence limit where $n=150$,

$$CL(r_k) = \frac{-1 \pm 1.96\sqrt{n-k-1}}{n-k}$$

for r_1 ,

$$CL(r_1) = \frac{-1 \pm 1.96\sqrt{n-2}}{n-1} = \frac{-1 \pm 1.96\sqrt{150-2}}{150-1} = 0.153, -0.167$$

$r_1 = 0.461$ (outside confidence band) → series is not independent.
 Remaining r_k values fluctuate close to zero.
 Only the first autocorrelation coefficient is significant.
 After lag 1, autocorrelation die out rapidly.
 Correlogram shows an exponentially decaying pattern.
 for $r_9 (k=9)$
 $CL(r_9) = \frac{-1 \pm 1.96\sqrt{150-9-1}}{150-9} = 0.157, -0.152$
 AR(1) model: ACF decays exp. After lag 1, ACF die rapidly.
 MA(1) model: ACF cuts off abruptly after lag 1, not decaying.

The dotted line shows the confidence limits.

Since r_1 is outside the confidence band, it is concluded that the series is not independent.

As the first autocorrelation coefficient is significant and observed correlogram follows the trend of theoretical correlogram for AR(1) model, AR(1) model would be suitable.
 $(r_k = r_1^k \text{ for theoretical correlogram})$

- Q A hydrologic variable is described by an AR(1) model with mean=100, variance=900, first serial correlation coeff. = 0.5. Compute the standard deviation of the random process.

→ Solution.

$$\sigma_k = \sqrt{900} = 30$$

$$\rho_1 = 0.5$$

$$\alpha_1 = \rho_1 = 0.5$$

For AR(1) process,

$$\sigma_e^2 = \sigma_x^2 (1 - \alpha_1 \rho_1 - \alpha_2 \rho_2 - \dots - \alpha_m \rho_m)$$

For $m=1$,

$$\begin{aligned} \sigma_e &= \sigma_x \sqrt{1 - \alpha_1 \rho_1} \\ &= 30 \sqrt{1 - 0.5 \times 0.5} \\ &= 26 \end{aligned}$$

* AR(2) model: Second order AR Model

This is also used in Hydrology

$$(X_t - \mu) = \alpha_1 (X_{t-1} - \mu) + \alpha_2 (X_{t-2} - \mu) + \epsilon_t$$

$$S_t = \alpha_1 S_{t-1} + \alpha_2 S_{t-2} + \dots + \alpha_m S_{t-m}$$

For $m=2$, ACF of AR(2) model is

$$S_t = \alpha_1 S_{t-1} + \alpha_2 S_{t-2}$$

For $t=1$ and 2 ,

$$S_1 = \alpha_1 + \alpha_2 S_1$$

$$S_2 = \alpha_1 S_1 + \alpha_2$$

Solving above equations, the parameters of AR(2) model are

$$\alpha_1 = \frac{S_1 - S_1 S_2}{1 - S_1^2}$$

$$\alpha_2 = \frac{S_2 - S_1^2}{1 - S_1^2}$$

Using sample serial correlation coefficient r_1 and r_2 , instead of S_1 and S_2 , parameters α_1 and α_2 can be estimated

For AR(2) to be stationary,

The conditions to be satisfied by parameters α_1 and α_2 ,

$$\alpha_1 + \alpha_2 < 1$$

$$\alpha_2 - \alpha_1 < 1$$

$$-1 < \alpha_2 < 1$$

The conditions to be satisfied by S_1 and S_2 ,

$$-1 < S_1 < 1$$

$$-1 < S_2 < 1$$

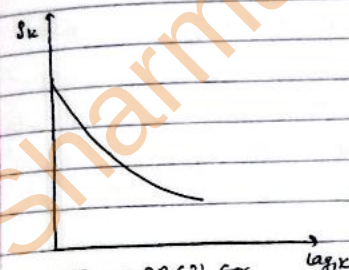
$$S_1^2 < \frac{1}{2} (S_2 + 1)$$

* Autocorrelation Function (ACF)

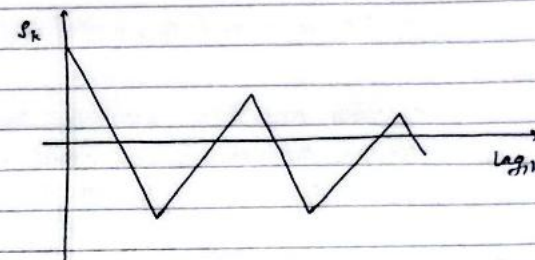
The ACF of AR(2) depends on the roots of the characteristic function

$$1 - \alpha_1 z - \alpha_2 z^2 = 0$$

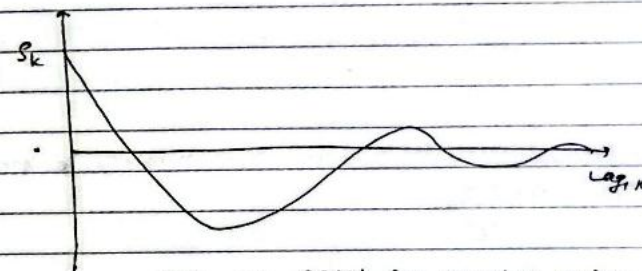
For real roots, the ACF consists of mixture of damped exponentials. When both roots are positive, the ACF remains positive as it damps out. When one of the roots is negative, the ACF alternates in sign as it damps out. If roots are complex, the ACF will be in the form of damped sine wave.



ACF of AR(2) for both +ve roots



ACF of AR(2) for one -ve root



ACF of AR(2) for complex roots

Applicability of AR(2) model in hydrology for generation and forecasting of annual flow (non-seasonal)

$$X_t - \mu = \alpha_1 (X_{t-1} - \mu) + \alpha_2 (X_{t-2} - \mu) + e_t$$

If we consider another independent process Z_t with mean = 0 and variance = 1, then

$$e_t = Z_t \sigma_e \sqrt{1 - R^2}$$

where, R = coeff. of Multiple correlation

$$X_t - \mu = \alpha_1 (X_{t-1} - \mu) + \alpha_2 (X_{t-2} - \mu) + Z_t \sigma_e \sqrt{1 - R^2}$$

Replacing population parameter by their corresponding sample estimates the annual flow model is

$$Q_t = \bar{Q} + \alpha_1 (Q_{t-1} - \bar{Q}) + \alpha_2 (Q_{t-2} - \bar{Q}) + Z_t \sigma_e \sqrt{1 - R^2}$$

where, Q_t = flow at time t

$\bar{Q}$ = mean of flow

Z_t = random process with mean = 0 & variance = 1

$$\sigma_e^2 = \frac{\sigma_e^2}{(1 - \alpha_1 R_1 - \alpha_2 R_2 - \dots - \alpha_m R_m)}$$

The annual flow is assumed to be normally distributed.

$$\alpha_1 = \frac{r_1 - r_1^2}{1 - r_1^2}$$

$$\alpha_2 = \frac{r_2 - r_1^2}{1 - r_1^2}$$

$$R^2 = \frac{r_1^2 + r_2^2 - 2r_1 r_2}{1 - r_1^2}$$

r_1, r_2 = first and second autocorrelation coeff.

Moving Average (MA) model:

In time series analysis, the moving-average model (MA model), also known as moving-average process, is a common approach for modeling univariate time series. The moving-average model specifies that the output variable depends linearly on the current and various past values of a stochastic (imperfectly predictable) term.

In MA model, the value of variable at time t is equal to weighted sum of present and past values of random component (noise, shock)

$$X_t = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_m e_{t-m} \quad \text{--- (I)}$$

where, $\beta_1, \beta_2, \dots, \beta_m$ = parameters of model,

e = random component with zero mean

Above eqⁿ is written as MA(m) where,

m = order of model (no. of lag times)

taking expectation,

$$E[X_t] = E[e_t] - \beta_1 E[e_{t-1}] - \beta_2 E[e_{t-2}] - \dots - \beta_m E[e_{t-m}]$$

$$\mu_x = \mu_e - \beta_1 \mu_e - \beta_2 \mu_e - \dots - \beta_m \mu_e$$

$$\mu_x = \mu_e \left[1 - \sum_{j=1}^m \beta_j \right] \quad \text{--- (II)}$$

If the process X_t is to have the same expected value as e_t then, $\mu_x = \mu_e$. Required condition is:

$$\sum_{j=1}^m \beta_j = 0$$

The alternative form of above equation is,

$$(X_t - \mu) = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_m e_{t-m} \quad \text{--- (III)}$$

Relationship of variance,

Squaring

$$(X_t - \mu)^2 = (e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_m e_{t-m})$$

$$(e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_m e_{t-m})$$

$$(x_t - \mu)^2 = e_t^2 + \beta_1^2 e_{t-1}^2 + \beta_2^2 e_{t-2}^2 + \dots + \beta_m^2 e_{t-m}^2$$

Taking expectations

$$\sigma_{x^2} = \sigma_e^2 + \beta_1^2 \sigma_e^2 + \beta_2^2 \sigma_e^2 + \dots + \beta_m^2 \sigma_e^2$$

$$\gamma(0) = \sigma_e^2 (1 + \beta_1^2 + \beta_2^2 + \dots + \beta_m^2) \quad \text{--- (iv)}$$

Multiplying eq (iii) by $(x_{t-k} - \mu)$ on both sides

$$(x_t - \mu)(x_{t-k} - \mu) = (e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_m e_{t-m})(x_{t-k} - \mu)$$

$$(x_t - \mu)(x_{t-k} - \mu) = (e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_m e_{t-m})(e_{t-k} - \beta_1 e_{t-k-1} - \beta_2 e_{t-k-2} - \dots + \beta_m e_{t-m-k})$$

Taking expectations on both sides $[E[e_t, e_{t-k}] = 0; \text{ independent}]$
 $E[e_{t-1}, e_{t-1-k}] = \sigma_e^2$

$$\text{cov}(k) = \gamma(k) = \sigma_e^2 (-\beta_k + \sum_{j=1}^m \beta_j \beta_{j+k})$$

When $j > m-k$, the term in the summation = 0, so above expression can be written as,

$$\gamma(k) = \sigma_e^2 (-\beta_k + \sum_{j=1}^{m-k} \beta_j \beta_{j+k})$$

$$\gamma(k) = \sigma_e^2 (-\beta_k + \beta_1 \beta_{k+1} + \beta_2 \beta_{k+2} + \dots + \beta_{m-k} \beta_m) \quad \text{--- (v)}$$

Dividing eq (v) by eq (iv)

$$s_k = \frac{\gamma(k)}{\gamma(0)} = \frac{(-\beta_k + \beta_1 \beta_{k+1} + \beta_2 \beta_{k+2} + \dots + \beta_{m-k} \beta_m)}{(1 + \beta_1^2 + \beta_2^2 + \dots + \beta_m^2)}$$

$s_k = 0$ for $k > m$. So, the ACF has cutoff at lag m .
 i.e. ACF of MA(m) process is zero beyond m .
 This is the expression for ACF.

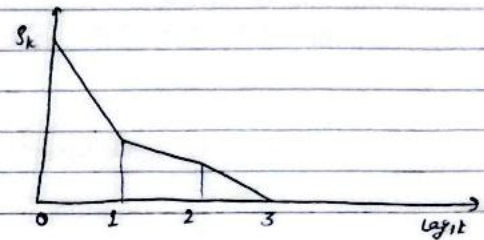
For MA(1)

$$s_1 = \frac{-\beta_1}{1 + \beta_1^2}$$

For MA(2)

$$s_1 = \frac{-\beta_1 + \beta_1 \beta_2}{1 + \beta_1^2 + \beta_2^2}$$

$$s_2 = \frac{-\beta_2}{1 + \beta_1^2 + \beta_2^2}$$



ACF of MA(2)

8) The first serial correlation coefficient of MA(2) process is 0.3. Compute the parameter of the process

$$s_1 = 0.3$$

For MA(m) process

$$s_k = \frac{(-\beta_k + \beta_1 \beta_{k+1} + \beta_2 \beta_{k+2} + \dots + \beta_{m-k} \beta_m)}{(1 + \beta_1^2 + \beta_2^2 + \dots + \beta_m^2)}$$

For MA(1)

$$s_1 = \frac{-\beta_1}{1 + \beta_1^2}$$

$$\text{or } 0.3 = \frac{-\beta_1}{1 + \beta_1^2}$$

$$\text{or } 0.3\beta_1^2 + \beta_1 + 0.3 = 0$$

Solving,

$$[\beta_1 = -0.33, -0.33]$$

Q4 A moving average process is given by $(X_t - \mu) = e_t - 1.2e_{t-1} + 0.5e_{t-2}$.
Determine the variance of X_t if the variance of e_t is 1.5.
→ $\beta_1 = 1.2 \quad \beta_2 = -0.5$

$$\sigma_x^2 = \sigma_e^2 (1 + \beta_1^2 + \beta_2^2) \\ = 1.5^2 (1 + 1.2^2 + (-0.5)^2) \\ = 6.05$$

H Autoregressive moving average (ARMA) model:

→ In ARMA model, the value of variable at time t is equal to weighed sum of past values of variables and present and past values of random component. The combined AR and MA process ARMA(p,q) model is given as:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_p X_{t-p} + e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_q e_{t-q}$$

Alternative expression is

$$(X_t - \mu) = \alpha_1 (X_{t-1} - \mu) + \alpha_2 (X_{t-2} - \mu) + \dots + \alpha_p (X_{t-p} - \mu) + e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_q e_{t-q}$$

e = random process with mean zero.

$\alpha_1, \alpha_2, \dots, \alpha_p, \beta_1, \beta_2, \dots, \beta_q$ = parameters of the model.

• ARMA(1,1) model

$$(X_t - \mu) = \alpha_1 (X_{t-1} - \mu) + e_t - \beta_1 e_{t-1}$$

variance relationship.

$$\sigma_x^2 = \frac{\sigma_e^2 (1 + \beta_1^2 - 2\alpha_1 \beta_1)}{(1 - \alpha_1^2)} \quad \text{--- (i)}$$

$$S_1 = \frac{(1 - \beta_1 \alpha_1) (\alpha_1 - \beta_1)}{(1 + \beta_1^2 - 2\alpha_1 \beta_1)} \quad \text{--- (ii)}$$

$$S_k = \alpha_1 S_{k-1} \text{ for } k \geq 2 \quad \text{--- (iii)}$$

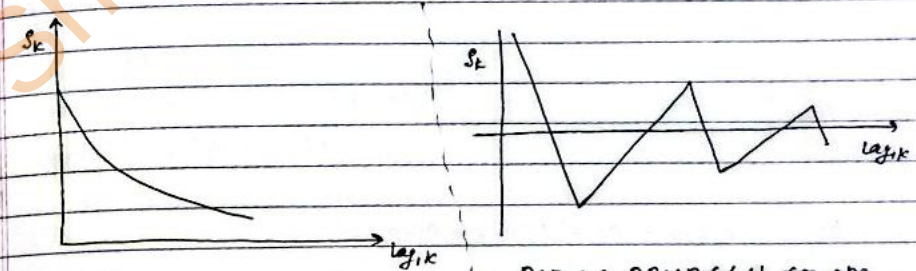
The ACF of ARMA(1,1) decays exponentially from a starting value of S_1 . If S_1 and S_2 are known, α_1 can be obtained from eqⁿ (iii) and substituting the value on eqⁿ (ii) gives the value of β_1 .

For ARMA(1,1) to be stationary

$$|S_2| < |S_1|$$

$$S_2 > S_1 (2S_1 + 1) \text{ if } S_1 < 0$$

$$S_2 > S_1 (2S_1 - 1) \text{ if } S_1 > 0$$



ACF of ARMA(1,1) for both the parameters

ACF of ARMA(1,1) for one -ve parameter

Q.5) Classify each of the following processes as AR, MA or ARMA verify whether they are stationary and compute first five serial correlation coefficients.

a. $(X_t - \mu) = 0.5(X_{t-1} - \mu) + e_t$
 AR, $m=1$, and present value depends on past values and random process, it is an AR(1) process

$$\rho_1 = 0.5$$

Computing serial correlation coeff

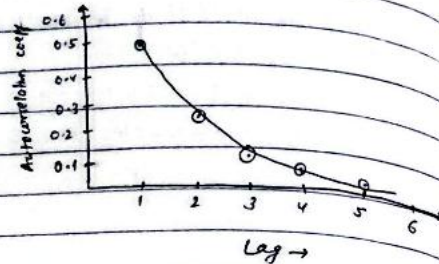
$$\rho_k = \rho_1^k$$

$$\rho_2 = (0.5)^2 = 0.25$$

$$\rho_3 = (0.5)^3 = 0.125$$

$$\rho_4 = (0.5)^4 = 0.0625$$

$$\rho_5 = (0.5)^5 = 0.03125$$



(As ACF is symmetrical about vertical axis, AR(1) model is stationary)

b. $(X_t - \mu) = e_t + 1.3e_{t-1} + 0.4e_{t-2}$

As the model has only 2 random processes and $m=2$, it is MA(2) model. It is always stationary.

$$\beta_1 = -1.3, \beta_2 = -0.4$$

Serial correlation coefficients,

$$\rho_k = \frac{(-\beta_1 + \beta_1\beta_2 + \beta_2\beta_1\beta_2 + \dots + \beta_{m-k}\beta_m)}{(1 + \beta_1^2 + \beta_2^2 + \dots + \beta_m^2)}$$

$$\rho_1 = \frac{-\beta_1 + \beta_1\beta_2}{1 + \beta_1^2 + \beta_2^2} = \frac{-(-1.3) + (-1.3)(-0.4)}{1 + (-1.3)^2 + (-0.4)^2} = 0.64$$

$$\rho_2 = \frac{-\beta_2}{1 + \beta_1^2 + \beta_2^2} = \frac{-(-0.4)}{1 + (-1.3)^2 + (-0.4)^2} = 0.14$$

$$\text{For } k > 3, \rho_k = 0$$

c. $(X_t - \mu) = (X_{t-1} - \mu) + e_t + 0.5e_{t-1}$

As present value depends on past values of variable and present and past random process, it is ARMA process

$$p=1, q=1$$

$$\text{ARMA}(1,1)$$

$$\alpha_1 = 1, \beta_1 = -0.5$$

$$\rho_1 = \frac{(1 - \alpha_1\beta_1)(\alpha_1 - \beta_1)}{(1 + \beta_1^2 - 2\alpha_1\beta_1)} = \frac{(1 - 1 \times -0.5)(1 - (-0.5))}{(1 + (-0.5)^2 - 2 \times 1 \times -0.5)} = 1$$

$$\rho_k = \alpha_1 \rho_{k-1} \text{ for } k \geq 2$$

$$\rho_2 = \alpha_1 \rho_1 = 1 \times 1 = 1$$

$$\rho_3 = \alpha_1 \rho_2 = 1 \times 1 = 1$$

$$\rho_4 = \alpha_1 \rho_3 = 1 \times 1 = 1$$

$$\rho_5 = \alpha_1 \rho_4 = 1 \times 1 = 1$$

For ARMA(1,1) to be stationary,

$$|\beta_1| < |\alpha_1|$$

$$\rho_1 > \rho_1(2\rho_1 - 1) \text{ if } \rho_1 > 0$$

As these conditions are satisfied, the process is stationary.

5.3.2

Applicability of ARMA(1,1) model in Hydrology for generation and forecasting of annual flow (non-seasonal model)

$$(X_t - \mu) = \alpha(X_{t-1} - \mu) + e_t - \beta e_{t-1}$$

If we consider another independent process Z_t with mean = 0 and variance = 1, then

$$e_t = Z_t \sqrt{\frac{\sigma_2^2(1-\alpha^2)}{1+\beta^2-2\alpha\beta}}$$

Replacing population parameters by their corresponding sample estimates, the annual flow model is

$$\hat{Q}_t = \bar{Q} + \alpha(\hat{Q}_{t-1} - \bar{Q}) + T(Z_t - bZ_{t-1})$$

where, $\hat{Q}_t$ is the flow at time t ; $\bar{Q}$ is the mean of flows, and Z_t is the random process with mean = 0 and variance = 1.

The annual flow is assumed to be normally distributed.

$$\sigma = \frac{\sigma_2}{\sigma_1}$$

$$r_1 = \frac{(1-\alpha b)(\alpha-b)}{1+b^2-2\alpha b} \quad \text{compute } b \text{ from this.}$$

σ_1, σ_2 = first and second autocorrelation coeff.

$$T = \sigma_2 \sqrt{\frac{(1-\alpha^2)}{1+b^2-2\alpha b}}$$

* Autoregressive integrated moving average (ARIMA) model

In this model, the current value of difference in series = weighted sum of past values of difference in series + present and past random component

ARIMA (p,d,q) model:

$$\Delta^d X_t = W_t = \alpha_1 W_{t-1} + \alpha_2 W_{t-2} + \dots + \alpha_p W_{t-p} + e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_q e_{t-q}$$

eg. first difference at t is $X_t - X_{t-1}$

second difference at t is $(X_t - X_{t-1}) - (X_{t-1} - X_{t-2})$

ARIMA(p,d,q) model is a non-stationary process. However, the d^{th} difference of X_t is stationary. ARMA(p,q) is a special case of ARIMA(p,d,q) with no differencing.

ARIMA

✓ For non-stationary data

✓ Make data stationary by differencing.

✓ Then apply ARMA

✓ trend present → use ARIMA

5.5 Time series modeling procedure

Identification of form of model

a. Visual inspection of time series data

b. ACF for different orders

• Finding the dependence

• If the ACF of given data resembles with that of theoretical ACF of assumed model, then the assumed model is a good choice

• Testing goodness of fit: computing confidence limit (eg: 95%) of ACF and checking whether all autocorrelation coefficient lies within the band

* Akaike Information Criteria (AIC)

The AIC is defined as the log-likelihood term penalized by the number of model parameters. In general, AIC is given as:

$$AIC = -N \ln(\text{likelihood}) + 2k$$

where

N = sample size

k = number of parameters

For AR(m) model,

$$AIC(m) = -N \ln(\sigma_p^2) + 2m$$

N = Number of observations

m = number of parameters

σ_p^2 = maximum likelihood estimate of variance of residuals

✓ AIC balances Model accuracy & Model simplicity

✓ Lower AIC = better model

• Determination of parameters of time series models (AR, MA, ARMA, ARIMA)

a. Method of moments

In this method, for 'k' number of parameters, first k sample moments are computed. The sample moments are equated to population moments, and the resulting equations are solved simultaneously to obtain the required parameters.

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

b. Method of maximum likelihood

In this method, the likelihood function is first determined. The likelihood function is function of parameters given the observation, which is represented by joint probability distribution of error term. The likelihood function (or its logarithm) is maximized by computing partial derivative with respect to each parameter and equating the partial derivative to zero.

For maximum value of L

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

c. Method of least squares

In this method, residual is computed by comparing observed and computed value. The sum of square of residuals (SSR) is minimized to obtain the parameters. In the minimization process, the partial derivative of SSR with respect to each parameter is set to zero and the resulting equations are solved simultaneously to obtain the value of parameter.

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_m X_{t-m}$$

Testing the assumptions of model (AR, MA, ARMA, ARIMA model)

a. Testing the Normality of residuals

I. Graphical method:

Plot empirical distribution of residuals on a normal probability paper and verify whether the plotted points fall on a straight line.

II. Statistical method

Skewness test: The coefficient of skewness for normal variable is zero.

$$\text{Coefficient of skewness } C_s = \frac{n \sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(n-2)s^3}$$

Kurtosis test: The coefficient of kurtosis for normal variable is three.

$$\text{Coefficient of kurtosis } K = \frac{E[(x - \mu)^4]}{\sigma^4}$$

Chi-square test: Chi-squared test is computed for the residuals. The tabulated value of chi-square is obtained for the assumed level of significance and $(k-h-1)$ degree of freedom (k = number of parameters, h = number of data interval). If the computed chi-square less than tabulated value, the residuals are considered to be normal.

If computed χ^2 is less than critical value of χ^2 , the assumed distribution is good fit.

Testing the independence of residuals

1. Plot of ACF:

Compute ACF for different lags and prepare correlogram.

If the computed ACF are close to zero and within the confidence limits, the residuals are considered to be independent.

check if residuals are independent (white noise)

If model is correct: Residuals should have no autocorrelation

2. Porte Manteau lack of fit test:

The test statistic (Q) in this test is given by:

$$Q = N \sum_{k=1}^L r_k^2(e)$$

where, $r_k^2(e)$ = autocorrelation of residuals

or:

L = number of lags (gen. 30% of sample size)

N = sample size

Obtain chi-squared value with $L-p-q$ degree of freedom at assumed level of significance (p, q = order of ARMA model).

If $Q <$ Chi-squared value, the residuals are considered to be independent (Good fit)

use minimum number of parameters

1. Testing of Parsimony of parameter

Parsimony refers to the lesser number of parameters to represent a system/process

Fewer parameters can be easily determined by optimization and there is also less parameters uncertainty in prediction

Parsimony of parameters or selection of optimum order of model can be determined by Akaike information criteria (AIC)

The larger the likelihood, the better the model.
The more parameters, the worse the model

AIC for AR(m) model

In general, AIC is given as

$$AIC = -N \ln(\text{likelihood}) + 2k$$

where

N = sample size

k = number of parameters

σ_e^2 = maximum likelihood estimate of variance of residual

N = number of observations

m = number of parameters

$$AIC(m) = -N \ln(\sigma_e^2) + 2m$$

AIC for different values of m are computed. A plot of AIC- m is prepared and the order of model which gives minimum AIC is selected.

✓ Why parsimony?

- ↳ Avoid overfitting
- ↳ Improve prediction
- ↳ Simpler interpretation

Thus

1. Fit model
2. Check residuals → Portmanteau
3. Compare Models → AIC
4. Choose simplest → Parsimony

(S) A hydrologic variable is characterized by an autoregressive AR(m) model. The standard deviation of the series is $\sigma_x = 4.85$, and the autocorrelation coefficients are $\gamma_1 = 0.430$ and $\gamma_2 = -0.077$. The sample size is $N=35$

- (a) Determine the parameters of AR(1) and AR(2) model using the Yule-Walker equations.
- (b) Compute the residual (white noise) variance for each model.
- (c) Using Akaike Information Criterion (AIC), Identify the optimal order m of the AR model.

Given,

$$\sigma_x = 4.85 \Rightarrow \sigma_x^2 = 23.5225$$

$$\gamma_1 = 0.430, \gamma_2 = -0.077$$

$$N = 35$$

(a) Parameter estimation

For AR(1)

$$\alpha_1 = \gamma_1 = 0.430$$

For AR(2)

$$\gamma_1 = \alpha_1 + \alpha_2 \gamma_1$$

$$\gamma_2 = \alpha_1 \gamma_1 + \alpha_2$$

Solving,

$$\alpha_2 = -0.321$$

$$\alpha_1 = 0.568$$

(b) Residual (white noise) variance

$$\begin{aligned} \text{AR(1): } \sigma_e^2 &= \sigma_x^2 (1 - \alpha_1^2) \\ &= 23.5225 (1 - 0.430^2) \\ &= 19.17 \end{aligned}$$

$$\begin{aligned} \text{AR(2): } \sigma_e^2 &= \sigma_x^2 (1 - \alpha_1 \gamma_1 - \alpha_2 \gamma_2) \\ &= 23.5225 [1 - (0.568)(0.430) - (-0.321)(-0.077)] \\ &= 17.19 \end{aligned}$$

(c) AIC calculation: (Direct variance is given, so we use)

$$AIC = N \ln(\sigma_e^2) + 2m \quad \text{--- (i)}$$

$$\begin{aligned} \text{AR(1)}: AIC &= 35 \ln(19.12) + 2 \\ &= 35 \times 2.95 + 2 \\ &= 105.32 \end{aligned}$$

$$\begin{aligned} \text{AR(2)}: AIC &= 35 \ln(17.19) + 4 \\ &= 35 \times 2.84 + 4 \\ &= 103.54 \end{aligned}$$

Lower AIC is better. Therefore, AR(2) is optimal

★ If Likelihood was given

$$AIC = -n \ln(\hat{\sigma}_m^2) + 2m \quad \text{--- (ii)}$$

- (i) → Minimizing the variance criterion → better fit
→ larger variance → worse fit.
→ so no need -ve sign.

- (ii) → From likelihood perspective → higher likelihood = better
→ so we negate it to minimize
→ Maximizing log-likelihood ⇒ Minimizing $\ln(\hat{\sigma}_m^2)$

ACF pattern

Cut-off → MP

Decay → AR

Mixed → ARMA

classmate

Date

Page

6. Spectral Models

6.1. Spectral Analysis (SA)

→ Spectral analysis or spectrum analysis in terms of a spectrum of frequencies or related quantities.

Many time series show periodic behaviour. This periodic behaviour can be very complex. Spectral analysis is a technique that allows us to discover underlying periodicities.

To perform spectral analysis, we first must transform data from time domain to frequency domain.

In other words, spectral analysis deals with the variation of variance of a stochastic process with frequency. The frequency is used in harmonic sense.

→ SA provides a powerful means of extracting information from hydraulic data. This type of analysis can reveal processes that may be obscured in direct time series analysis by providing data not just on temporal fluctuations, but also on the spectrum of frequencies with those fluctuations.

* Fourier series for spectral analysis

Fourier analysis involves the resolution of signals into sinusoidal components to describe the spectral contents of the signals in the frequency domain.

There are two major techniques of Fourier analysis:

→ Fourier series (FS)

→ Fourier transform (FT)

FS analysis resolves the periodic signal (power signal) into an infinite sum of sinusoidal wave components, while FT analysis performs a similar role in the analysis of the unperiodic signal (energy signal) which are more general use in the signal processing.

→ FT can also be used, in a limiting sense, to represent the periodic signal as an infinite sum of sinusoidal wave components

The finite FS representation for discrete RV & IV written as

$$x_t = \bar{x} + \sum_{m=1}^q [a_m \cos(2\pi f_m t) + b_m \sin(2\pi f_m t)] \quad \text{--- (c)}$$

where, x_t = Random variable (RV)
 $t = 1, 2, \dots, n$ spaced at apert

Total length, $T = n \Delta t$

If n is odd, $n = 2q+1$ where

q = no. of harmonics to be considered

$$q = \frac{(n-1)}{2}$$

f_m = frequency of m^{th} harmonic = $\frac{m}{n}$

a_m & b_m = coefficients of the m^{th} harmonic

$f_1 = \frac{1}{n}$ = fundamental frequency

With the value of f_m , Fourier series is

$$x_t = \bar{x} + \sum_{m=1}^q [a_m \cos\left(\frac{2\pi m t}{n}\right) + b_m \sin\left(\frac{2\pi m t}{n}\right)] \quad \text{--- (b)}$$

The values of coefficients are given by,

$$a_m = \frac{2}{n} \sum_{t=1}^n x_t \cos\left(\frac{2\pi m t}{n}\right) \quad \text{--- (c)}$$

$$b_m = \frac{2}{n} \sum_{t=1}^n x_t \sin\left(\frac{2\pi m t}{n}\right) \quad \text{--- (d)}$$

Fourier series can also be written as

$$x_t = \bar{x} + \sum_{m=1}^q c_m \sin\left(\frac{2\pi m t}{n} + \phi_m\right) \quad \text{--- (e)}$$

where, c_m is amplitude of m^{th} harmonic,
 ϕ_m is the phase angle of the m^{th} harmonic

$$c_m^2 = a_m^2 + b_m^2 \quad \text{--- (f)}$$

$$\phi_m = \tan^{-1}\left(\frac{-b_m}{a_m}\right) \quad \text{--- (g)}$$

* Line Spectrum:

A plot of c_m^2 versus the frequency f_m is called line spectrum. It is also known as discrete spectrum or variance spectrum or periodogram.

The variance of the series (S_x^2) is

$$S_x^2 = \sum_{m=1}^q \left(\frac{c_m^2}{2}\right) \quad \text{--- (h)}$$

where, $\frac{c_m^2}{2}$ is the variance of each harmonic

Sum of variances of all the harmonics = Variance of the series

The line spectrum consists of $q = \frac{(n-1)}{2}$ values and

displays the variance explained by each harmonic.

If n is even, then $q = \frac{n}{2}$ and eqn (c) and (d) are used to compute coefficients for $m = 1, 2, \dots, q-1$ and a_q and b_q are given by

$$a_q = \frac{1}{n} \sum_{t=1}^n (-1)^t x_t$$

and, $b_q = 0$

Used in monthly water planning, combined seasonality, correlation, Randomness etc.
 Reservoir design; flow simulation
 T-F = seasonal AR model

6.2 Univariate Seasonal model

6.2.1 Thomas-Fiering model for generating seasonal flows (T-F)

→ Thomas-Fiering model is an AR(1) model which can be used to forecast/generate seasonal flow data such as monthly flows or biannual flows. The flow at time t is a weighted sum of immediate past flows plus random component.

→ The form of equation in Thomas-Fiering model is

$$X(i, j) = \mu_j + \frac{s_j s_j}{s_{j-1}} [X(i, j-1) - \mu_{j-1}] + z_t s_j \sqrt{(1 - s_j^2)}$$

where, $i = 1, 2, \dots = \text{year}$

$j = 1, 2, \dots, m = \text{season of a year}$

$X(i, j) = \text{flow in the } j^{\text{th}} \text{ season of } i^{\text{th}} \text{ year}$

μ_j and $s_j = \text{mean and SD of flow in } j^{\text{th}} \text{ season}$

$s_j = \text{correlation coefficient b/w flow of } j^{\text{th}} \text{ and } (j-1)^{\text{th}} \text{ season}$

$z_t \sim N(0, 1)$

→ In practice, the population parameters of the equation are replaced by sample estimates.

$$\hat{\phi}(i, j) = \bar{\phi}_j + \frac{s_j s_j}{s_{j-1}} [\hat{\phi}(i, j-1) - \bar{\phi}_{j-1}] + z_t s_j \sqrt{(1 - s_j^2)}$$

where, $i = 1, 2, \dots = \text{year}$

$j = 1, 2, \dots, m = \text{season of a year}$

$\hat{\phi}(i, j) = \text{flow in } j^{\text{th}} \text{ season of } i^{\text{th}} \text{ year}$

$\bar{\phi}_j$ and $s_j = \text{mean and SD of flow in } j^{\text{th}} \text{ season}$

$s_j = \text{correlation coefficient b/w flows of } j^{\text{th}} \text{ and } (j-1)^{\text{th}} \text{ season}$

$b_j = \text{regression coefficient of flow of } j^{\text{th}} \text{ and } (j-1)^{\text{th}} \text{ season}$

$z_t \sim N(0, 1)$

→ Computation of statistical parameters for any season j

$$\bar{\phi}_j = \frac{1}{n} \sum_{i=1}^n \phi(i, j)$$

$$s_j = \frac{s_j s_j}{s_{j-1}} \text{ for } j = 2, 3, 4, \dots, m$$

$$s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n [\phi(i, j) - \bar{\phi}_j]^2}$$

$$s_j = \frac{1}{n} \sum_{i=1}^n [\phi(i, j) - \bar{\phi}_j][\phi(i, j-1) - \bar{\phi}_{j-1}] \text{ for } j = 2, 3, 4, \dots, m$$

For $j=1$

$$s_1 = \frac{1}{n-1} \sum_{i=1}^n [\phi(i, 1) - \bar{\phi}_1][\phi(i-1, m) - \bar{\phi}_m]$$

$$b_1 = \frac{s_1 s_1}{s_m}$$

→ In this model, seasonal flows are considered to be normally distributed. If negative values are generated during the computation, replace them by zero or small number.

6.2.2 Daily-flow Model

- * Yevjevich and Pulpo model for generating daily flows
- Numbers of parameters in T-F model is $8m$, where $m = \text{no. of seasons}$. If this model is applied for daily flow estimation, we have to estimate a large number of parameters (eg. 365×3), which is not practicable. Therefore, T-F model is practically not applicable for daily flow simulation. Another model developed by Yevjevich and Pulpo is suitable for such simulation.

→ For daily flow model, the cyclicity for a period of 365 days is considered. The variation in mean and variance of daily flows is described by harmonic fitting. After removing cyclicity, an appropriate stationary model is fitted for residual series.

→ If $x(i, j)$ = observed flow in the j th day of the i th year, $j = 1, 2, \dots, 365$ and $i = 1, 2, \dots, n \text{ years}$.

- Mean of flows in j th day ($\bar{x}_j$) = $\frac{1}{n} \sum_{i=1}^n x(i, j)$

- Standard deviation of flows in j th day (s_j)

$$s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n [x(i, j) - \bar{x}_j]^2}$$

- Mean of daily means ($\bar{x}$) = $\frac{1}{365} \sum_{j=1}^{365} \bar{x}_j$

- Variance of daily means (v_x) = $\frac{1}{365} \sum_{j=1}^{365} (\bar{x}_j - \bar{x})^2$

- Mean of daily standard deviations ($\bar{s}$) = $\frac{1}{365} \sum_{j=1}^{365} s_j$

- Variance of the daily standard deviations (v_s) = $\frac{1}{365} \sum_{j=1}^{365} (s_j - \bar{s})^2$

- Number of harmonics required for exact fitting = $\frac{(n-1)}{2}$
 $= \frac{(365-1)}{2} = 182$,

but six harmonics are usually sufficient for reasonable result.

→ Harmonic coefficients to fit for mean: a_k and b_k ($k = 1, 2, \dots, 6$)

- Harmonic coefficients to fit for standard deviation: c_k and d_k ($k = 1, 2, \dots, 6$)

$$a_k = \frac{2}{365} \sum_{j=1}^{365} \bar{x}_j \cos\left(\frac{2\pi jk}{365}\right)$$

$$b_k = \frac{2}{365} \sum_{j=1}^{365} \bar{x}_j \sin\left(\frac{2\pi jk}{365}\right)$$

$$c_k = \frac{2}{365} \sum_{j=1}^{365} s_j \cos\left(\frac{2\pi jk}{365}\right)$$

$$d_k = \frac{2}{365} \sum_{j=1}^{365} s_j \sin\left(\frac{2\pi jk}{365}\right)$$

- Harmonic fitting for mean is

$$\bar{x}_j' = \bar{x} + \sum_{k=1}^{2m} \left[a_k \cos\left(\frac{2\pi jk}{365}\right) + b_k \sin\left(\frac{2\pi jk}{365}\right) \right]$$

- Harmonic fitting for standard deviation is

$$s_j' = \bar{s} + \sum_{k=1}^{2s} \left[c_k \cos\left(\frac{2\pi jk}{365}\right) + d_k \sin\left(\frac{2\pi jk}{365}\right) \right]$$

where, $2m$ and $2s = 6$

→ The residual series is

$$e_t = \frac{a(i,j) - \bar{x}_j}{s_j}$$

where $t = (i-1)365 + j$

→ Next, a stationary model such as AR(1) or AR(2) is fitted for y_t .

Using AR(1) model

$$y_t = \bar{y} + \alpha_1 (y_{t-1} - \bar{y}) + z_t s_y \sqrt{1 - \alpha_1^2}$$

$$z_t \sim N(0,1)$$

where

α_1 = first autocorrelation coefficient

No. of parameters to estimate:

24 harmonic coefficient, $\bar{x}$, $\bar{y}$, s_y and α_1

The daily flow is obtained by introducing cyclicity

$$a(i,j) = y_t s_j + \bar{x}_j$$

- A random number is a number chosen from a pool of limited or unlimited numbers that has no discernible pattern for prediction. The pool of numbers is ^{almost} always independent from each other.
- However, the pool of numbers may follow a specific distribution. For example, the height of the students in a school tends to follow a normal distribution around the median height. If the height of ~~the~~ student is picked at random, the picked number has a higher chance to be closer to the median height than being classified as very tall or very short.
- Random numbers are numbers that occur in a sequence such that two conditions are met:
 - (1) The values are uniformly distributed over a defined interval or set.
 - (2) It is impossible to predict future values based on past or present ones.
- Random numbers are important in statistical analysis and probability theory.

7.1 Methods for generating uniformly distributed random numbers (RN)

- For performing stochastic simulations, random numbers that are distributed in accordance with certain probability distributions are needed. These random numbers are obtained by various distribution-dependent methods from random numbers that are uniformly distributed on the interval $[0, 1]$.
- In general, no truly random numbers can be generated by a deterministic working computer, but only pseudo-RN. These pseudo-RN are indeed nearly uniformly distributed on the interval $[0, 1]$ and meet a number of criteria for real RN, but still have some irregularities, real RN do not have.

A. Mid-square technique

- This technique starts with an initial number or seed and then squared and the middle digit of this square become the RN after placement of the appropriate decimal. The middle digit are then squared to generate the second RN.
- In the beginning of the generation process select an initial (seed) of 2d digits.
- Square the $(k-1)^{th}$ RN of 2d digits to produce a number of 4d digits and select the middle 2d digits and repeat the procedure.
- Limitations: The process terminates if 0000 appears in the chain.

Eg:

$$\text{Seed} = 4538$$

$$4538^2 = 20593444$$

$$\text{Next RN} = 5834$$

$$5834^2 = 34032356$$

$$\text{Next RN} = 2123$$

B. Mid-product technique

- The mid-product technique starts by selecting two seeds X_0 and X_0' each containing the same number of digits D . Now, multiply X_0 by X_0' to get a number U_1 . Set X_1 equal to the middle D digits of U_1 with the placement of appropriate decimals.
- In the beginning of the generation process select two seeds of 2d digit.
- Multiply the $(k-1)^{th}$ and $(k-2)^{th}$ random number of 2d digit and select the central 2d digit and repeat the procedure.
- Limitations: The process terminates if 0000 appears in the chain.

Eg:

$$\text{Seeds: } 6728 \text{ and } 4591$$

$$6728 \times 4591 = 30865293$$

$$\text{Next RN} = 8652$$

$$8652 \times 4591 = 39721882$$

$$\text{Next RN} = 7213$$

$$7213 \times 8652 = 62406876$$

$$\text{Next RN} = 4068$$

(C) Mixed Congruential Method (MCM)

- A mixed congruential method or linear congruential generator (LCG) is an algorithm that yields a sequence of pseudo-randomized numbers calculated with a discontinuous piecewise linear equation.
- The method represents one of the oldest and best-known pseudo-random number generator algorithms. The theory behind them is relatively easy to understand, and they are easily implemented and fast, especially on computer hardware which can provide modular arithmetic by storage-bit truncation.

→ The recursive formula used in this method is

$$u_k = (au_{k-1} + c) \bmod m \quad \text{--- (1)}$$

where,

u_k = current RN

u_{k-1} = previous RN

parameter a = multiplier

parameter c = increment

parameter m = modulus

mod refers to integer division (remainder left)

u_0 = seed.

⇒ Choice of parameters a, c , and m

• Binary computer

$$m = 2^b$$

c = any odd integer

$$a \bmod 4 = 1$$

u_0 = any integer seed so that $0 \leq u_0 \leq m$

• Decimal computer

$$m = 10^d$$

c = any integer not divisible by either 2 or 5

$$a \bmod 20 = 1$$

u_0 = any integer seed so that $0 \leq u_0 \leq m$

- According to above selection criteria, all the integers less than m appear before the seed appears. To make the generated RN free from initial internal serial correlation, the conditions are

$$a \approx \sqrt{m}$$

$$c \approx m/2$$

- to generate uniformly distributed RN in the range of 0 to 1, divide the sequence by m .

- Perform autocorrelation analysis for verifying the generated RN to be independent.

(Q) Generate a chain of 12 uniformly distributed RN in range 0 to 0.999 using MCM. Take $a = 361$, $c = 56377$, $m = 10^5$ and $u_0 = 42675$.

→ Solution:-

$$\text{eqn (1): } u_k = (361 u_{k-1} + 56377) \bmod 10^5 \text{ \& } u_k(0:1) = u_k/m$$

S.N.	u_{k-1}	u_k	$u_k(0:1)$
1	42675	62052	0.62052
2	62052	57149	0.57149
3	57149	87166	0.87166
4	87166	23303	0.23303
5	23303	68760	0.68760
6	68760	78737	0.78737
7	78737	80434	0.80434
8	80434	93051	0.93051
9	93051	47788	0.47788
10	47788	7845	0.07845
11	7845	88422	0.88422
12	88422	76719	0.76719

eg:

$$A = 361 \times 42675 + 62052$$

$$A = 15462052$$

$$A \div 10^5 = 154.62052$$

$$u_k = 62052$$

$$u_k(0:1) = 0.62052$$

$$\textcircled{1} a = 361 \approx \sqrt{10^5} \approx \sqrt{m}$$

$$\rightarrow a \bmod 20 = 1$$

$$\textcircled{2} c = 56377 \approx \frac{10^5}{2} \left(\frac{m}{2} \right)$$

$$\rightarrow \text{not divisible by 2 or 5}$$

$$\textcircled{3} u_k = 42675 (0 \leq u_k < 10^5)$$

$$\textcircled{4} m = 10^4 : 10^5 \text{ (at 5 digits)}$$

- # 7-2 Generation of normal random numbers
- As many hydrological variables follow normal distribution, the generation of normal RN is required in hydrological time series modeling.
 - If Z is a standard normal variable, then $x = \mu + \sigma Z$ follows normal distribution.

* Methods of generating normal RN

(A) Inverse transformation method

- If Z is a standard normal variable, its CDF is given by

$$F(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^Z e^{-t^2/2} dt$$

- If U = new random variable such that $F(Z) = U$ and if $h(u)$ is the PDF of U , then

$$h(u) du = f(z) dz$$

- For $-\infty < z < \infty$, the corresponding value of U is 0 to 1, i.e., U is uniformly distributed with 0 to 1.

Procedure

- First generate uniformly distributed RN (u_i) and then obtain normal RN (z_i) correspondingly to u_i such that $u_i = F(z_i)$ or $z_i = F^{-1}(u_i)$ where F^{-1} is called Inverse function.
- Using CDF table of normal distribution, value of z correspondingly to $F^{-1}(u_i)$ (treated as $F(z)$ in table) can be found by interpretation. This method is called table look up method.
- Limitation: time consuming.

- Q) Obtain the chain of normal RN using the table-look up method corresponding to the chain of following uniform RN: 0.5597, 0.8815, 0.6457

Solⁿ

Interpolation (for two $F(z)$ values) within which the given value lies; the corresponding Z values are obtained from tables

From Z-table,

$u = F(z)$

0.55960 $\rightarrow$ 0.15
0.56360 $\rightarrow$ 0.16

$$\left\{ 0.15 + \frac{0.16 - 0.15}{0.5636 - 0.5596} \times (0.5597 - 0.5596) \right\}$$

0.5597 $\rightarrow$ 0.15025

or, (calculator $\rightarrow$ $\oplus$ $\rightarrow$ Inv. Normal)
Area = 0.5597
 $\therefore Z = 0.15025$

$F(z)$ z

0.8810 $\rightarrow$ 1.18
0.88298 $\rightarrow$ 1.19

0.8815 $\rightarrow$ 1.183

$F(z)$

0.64430 $\rightarrow$ 0.37
0.64800 $\rightarrow$ 0.38

0.6457 $\rightarrow$ 0.3737

S.N.	$u_i = F(z_i)$	$z_i = F^{-1}(u_i)$
1	0.5597	0.15025
2	0.8815	1.183
3	0.6457	0.3737

(B) Central limit theorem method

→ $U_1, U_2, \dots, U_m = m$ independent and uniformly distributed variables each with range 0 to 1.

→ Mean of each variable = 0.5

→ variance of each variable = $1/12$

→ Let, $X = U_1 + U_2 + \dots + U_m$

→ Acc. to CLT, for large m , X will follow normal distribution with mean $\mu_x = \mu_{u_1} + \mu_{u_2} + \dots + \mu_{u_m}$ and variance $\sigma_x^2 = \sigma_{u_1}^2 + \sigma_{u_2}^2 + \dots + \sigma_{u_m}^2$

→ Mean of $X = 0.5m$

→ Variance of $X = m/12$

→ If Z is defined as, $Z = \frac{(X - 0.5m)}{\sqrt{m/12}}$, then Z will follow normal distribution with mean = 0 and variance = 1.

→ Using this concept, a chain of standard normal random numbers Z_i can be obtained from a chain of uniformly distributed random numbers U_i from the following equation:

$$Z_j = \frac{\sum_{k=1}^m (U_k - 0.5m)}{\sqrt{m/12}} \quad \text{where, } k = (j-1)m + i$$

→ To generate n normal RN, $n \times m$ uniformly distributed RN are required.

→ In practice, $m=12$ or 6 is used.

For $m=6$,

$$Z_j = \left(\sum_{k=1}^6 U_k - 3 \right) \sqrt{2} \quad \text{where, } k = 6(j-1) + i$$

For $m=12$

$$Z_j = \left(\sum_{k=1}^{12} U_k - 6 \right) \quad \text{where, } k = 12(j-1) + i$$

(a) obtain two normal RN from the sequence of following uniform RN using CLT method

0.62829, 0.96977, 0.81039, 0.32496, 0.47399, 0.91900,

0.68682, 0.88705, 0.9475, 0.42545, 0.50203, 0.47676

→ Solⁿ:

According to central limit theorem method,

$$Z_j = \frac{\sum_{k=1}^m (U_k - 0.5m)}{\sqrt{m/12}} \quad \text{where, } k = (j-1)m + i$$

For $m=6$

$$Z_j = \left(\sum_{k=1}^6 U_k - 3 \right) \sqrt{2} \quad , \quad k = 6(j-1) + i$$

For $j=1$, $[k=i]$

$$Z_1 = \left(\sum_{k=1}^6 U_k - 3 \right) \sqrt{2} = \left[\left(0.62829 + 0.96977 + 0.81039 + 0.32496 + 0.47399 + 0.91900 \right) - 3 \right] \sqrt{2}$$

$$= [4.1264 - 3] \sqrt{2}$$

$$= 1.593$$

For $j=2 \rightarrow k = 6(2-1) + i = 6+i$

$$Z_2 = \left(\sum_{k=1}^6 U_{6+i} - 3 \right) \sqrt{2} = \left[\left(0.68682 + 0.88705 + 0.9475 + 0.42545 + 0.50203 + 0.47676 \right) - 3 \right] \sqrt{2}$$

$$= [3.9256 - 3] \sqrt{2}$$

$$= 1.309$$

(C) Box-Muller Method (BMM)

→ Let z_1 and z_2 be two independently and normally distributed RV with mean = 0 and variance = 1. Then the joint pdf for z_1 and z_2 at any point (z_1, z_2) is given by

$$\phi(z_1, z_2) = f(z_1) \cdot f(z_2) = \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-z_2^2/2}$$

$$= \frac{1}{2\pi} e^{-(z_1^2 + z_2^2)/2}$$

→ The outcome is the random vector (z_1, z_2) in 2D, which is represented in polar co-ordinate as $z_1 = R \cos \theta$ and $z_2 = R \sin \theta$.

→ If U_1 and U_2 are two uniformly distributed RN in the range 0 to 1, then

$$z_1 = R \cos \theta = [-2 \ln(1 - U_1)]^{1/2} \cdot \cos(2\pi U_2)$$

$$z_2 = R \sin \theta = [-2 \ln(1 - U_1)]^{1/2} \cdot \sin(2\pi U_2)$$

→ Here, z_1 and z_2 will be two independently distributed standard normal variables. Hence, a chain of standard normal RN z_i can be obtained from a chain of uniformly distributed random numbers u_i in pairs from the following equation.

$$z_i = [-2 \ln(1 - u_i)]^{1/2} \cdot \cos(2\pi u_{i+1}) \quad \text{---(a)}$$

$$z_{i+1} = [-2 \ln(1 - u_i)]^{1/2} \cdot \sin(2\pi u_{i+1}) \quad \text{---(b)}$$

(D) Using the sequence of uniform RN in pairs obtain the sequence of normal RN using BMM

0.6778	0.1851
0.865	0.2425
0.6332	0.2542
0.5873	0.1804
0.1646	0.8608

→ Soln:

$$z_i = [-2 \ln(1 - u_i)]^{1/2} \cdot \cos(2\pi u_{i+1})$$

$$z_{i+1} = [-2 \ln(1 - u_i)]^{1/2} \cdot \sin(2\pi u_{i+1})$$

A	B	C	D	
u_i	u_{i+1}	z_i	z_{i+1}	
0.6778	0.1851	0.5969	1.3816	$C_1 = [-2 \ln(1 - u_1)]^{1/2} \cdot \cos(2\pi u_2)$
0.865	0.2425	0.0943	1.9390	$D_1 = [-2 \ln(1 - u_1)]^{1/2} \cdot \sin(2\pi u_2)$
0.6332	0.2542	-0.0374	1.4158	
0.5873	0.1804	0.5634	1.2052	
0.1646	0.8608	0.3846	-0.4602	

Application of Thomas-Fiering Model

Q) Generate synthetic flows for 4 seasons for a year using the following statistical parameters and T-F Model.

Season	Mean flow (m ³ /s)	S.D	Correlation with previous season	$z_t: N(0,1)$
1	10	8.8	0.63	0.25
2	20	10	0.62	-1.71
3	60	57	0.20	0.31
4	9	8	0.05	0.11

→ Solⁿ:-

T-F model:

$$Q(i,j) = \bar{Q}_i + b_j [Q(i,j-1) - \bar{Q}_{j-1}] + z_t S_j \sqrt{(1 - r_j^2)}$$

$i \rightarrow$ year ; $j \rightarrow$ season ; $Q \rightarrow$ flow

$r_j =$ corr. coeff. b/w flows of j^{th} and $(j-1)^{\text{th}}$ season

$b_j =$ regression coeff. of flows of j^{th} and $(j-1)^{\text{th}}$ season

$z_t: N(0,1)$

$$b_j = \frac{r_j S_j}{S_{j-1}} \text{ for } j = 2, 3, 4.$$

$$\text{For } j=1 \quad b_j = \frac{r_j S_j}{S_4}$$

$$Q(1,0) = \bar{Q}_4 \text{ and } \bar{Q}_0 = \bar{Q}_4 \text{ (take last season's value)}$$

Computation in tabular form

year(i)	Season(j)	Mean flow ($\bar{Q}_j$)	Standard deviation (S_j)	Correlation with previous season (r_j)	$z_t: N(0,1)$	b_j	$Q(i,j)$ (m ³ /s)
1	1	10	8.8	0.63	0.25	0.693	11.709
1	2	20	10	0.62	-1.71	0.704	7.786
1	3	60	57	0.20	0.31	1.14	63.390
1	4	9	8	0.05	0.11	0.01	9.903

$$[10 + 0.693 (9 - 9) + 0.25 \times 8.8 \sqrt{1 - 0.63^2}]$$

$$[20 + 0.704 (11.709 - 10) + (-1.71) \times 10 \sqrt{1 - 0.62^2}]$$

Application of Fourier analysis for line spectrum

(Q) Following are the mean bimonthly flow (m³/s) of a river: 10, 13, 25, 60, 43, 18. Fit Fourier series which explains more than 90% of variance and plot line spectrum.

→ Solⁿ:

Fourier series:

$$x_t = \bar{x} + \sum_{m=1}^q [a_m \cos\left(\frac{2\pi mt}{n}\right) + b_m \sin\left(\frac{2\pi mt}{n}\right)]$$

$$n=6$$

(*) For even number no. of harmonics = $n/2 = 3 = q$

The values of coefficients are given by

$$a_m = \frac{2}{n} \sum_{t=1}^n x_t \cos\left(\frac{2\pi mt}{n}\right) \text{ for } m=1, 2$$

$$b_m = \frac{2}{n} \sum_{t=1}^n x_t \sin\left(\frac{2\pi mt}{n}\right) \text{ for } m=1, 2$$

For $m=3$, ($m=2$)

$$a_3 = \frac{1}{n} \sum_{t=1}^n (-1)^t x_t ; b_3 = 0$$

Now,

$$\text{Mean of data } (\bar{x}) = 28.17 \text{ m}^3/\text{s}$$

$$\phi_m = \tan^{-1}\left(\frac{-b_m}{a_m}\right)$$

Computation in tabular form using above equations

(m) Harmonics	a_m	b_m	c_m	c_m^2	ϕ_m	$c_m^2/2$ (Var)	% of variance
1	-5.7	-23.1	23.79	566.1	-76	283.1	89.64
2	-6.7	4	7.80	60.9	31	30.4	9.64
3	2.166	0	2.17	4.7	0	2.3	0.74

$$\therefore \text{Var} = \sum c_m^2/2 = 315.8$$

→ Sample calculation

$$a_1 = \frac{2}{6} \left[10 \cos \frac{2\pi \times 1 \times 1}{6} + 13 \cos \frac{2\pi \times 1 \times 2}{6} + 25 \cos \frac{2\pi \times 1 \times 3}{6} + 60 \cos \frac{2\pi \times 1 \times 4}{6} + 43 \cos \frac{2\pi \times 1 \times 5}{6} + 18 \cos \frac{2\pi \times 1 \times 6}{6} \right]$$

$$= -5.667 \approx -5.7$$

$$b_1 = \frac{2}{6} \left[10 \sin \frac{2\pi \times 1 \times 1}{6} + 13 \sin \frac{2\pi \times 1 \times 2}{6} + 25 \sin \frac{2\pi \times 1 \times 3}{6} + 60 \sin \frac{2\pi \times 1 \times 4}{6} + 43 \sin \frac{2\pi \times 1 \times 5}{6} + 18 \sin \frac{2\pi \times 1 \times 6}{6} \right]$$

$$= -23.091 \approx -23.1$$

$$c_1 = \sqrt{a_1^2 + b_1^2} = \sqrt{(-5.7)^2 + (-23.1)^2} = 23.79$$

$$c_1^2 = a_1^2 + b_1^2 = (-5.7)^2 + (-23.1)^2 = 566.1$$

$$\phi_m = \tan^{-1}\left(\frac{-b_1}{a_1}\right) = \tan^{-1}\left(\frac{23.1}{-5.7}\right) = -1.329 \text{ rad} = -1.823 \times \frac{180}{\pi} = -76^\circ$$

$$\text{var} = \frac{c_1^2}{2} = \frac{566.1}{2} = 283.1$$

$$\% \text{ of variance} = \frac{283.1}{283.1 + 30.4 + 2.3} \times 100\% = 89.64\%$$

→ The first and second harmonics account for more than 90% of variance. Discarding other harmonics, Fourier equation is

$$x_t = \bar{x} + \sum_{m=1}^2 \left[a_m \cos\left(\frac{2\pi mt}{n}\right) + b_m \sin\left(\frac{2\pi mt}{n}\right) \right]$$

$$x_t = 28.17 - 5.7 \cos\left(\frac{\pi t}{3}\right) - 23.1 \sin\left(\frac{\pi t}{3}\right) - 6.7 \cos\left(\frac{2\pi t}{3}\right) + 4 \sin\left(\frac{2\pi t}{3}\right)$$

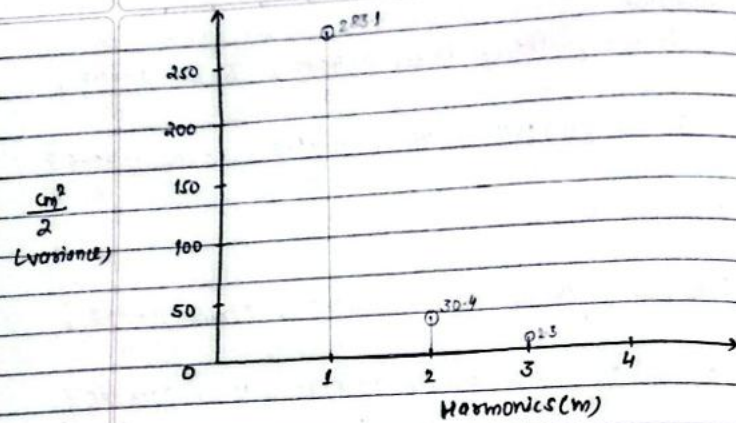


Fig: Line Spectrum

Time series Modeling

Q1. The first 9 autocorrelation coefficients computed from 150 years of annual flow record at a river site are:-

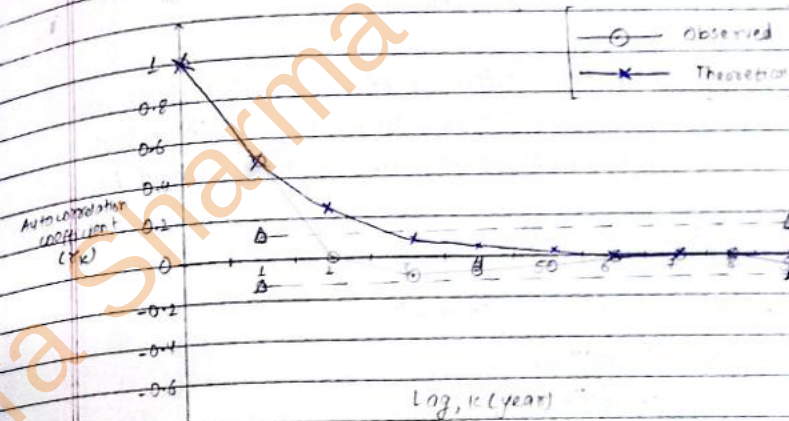
0.461, 0.005, 0.095, -0.057, -0.048, -0.01, -0.018, -0.016, -0.055

Test the flow series for independence at 95% confidence level.

Also, suggest a suitable model for the series.

→ Solⁿ:-

Plotting the given autocorrelation coefficients (τ_k)



From the plot of autocorrelation-lag, it is seen that AC coefficients for different lags do not fluctuate around zero. Hence, we can say that the series is not independent. It is further verified by computing confidence limit

Computing the confidence limit

$$n = 150$$

$$CL(\tau_k) = \frac{-1 \pm 1.96\sqrt{n-k-1}}{n-k}$$

for $x_k (k=1)$

$$CL(x_1) = \frac{-1 \pm 1.96 \sqrt{150-1-1}}{150-1} = 0.153, -0.167$$

for $x_g (k=9)$

$$CL(x_9) = \frac{-1 \pm 1.96 \sqrt{150-9-1}}{150-9} = 0.157, -0.152$$

The dotted lines show the confidence limits

→ Since x_1 is outside the confidence band, it is concluded that the series is not independent.

→ As the first autocorrelation coefficient is significant and observed correlogram follows the trend of theoretical correlogram for AR(1) model, AR(1) model would be suitable.

$x_k = x_1^k$ for theoretical correlogram

$$x_1 = 0.461 \Rightarrow x_k = (0.461)^k$$

$$x_0 = (0.461)^0 = 1$$

$$x_1 = (0.461)^1 = 0.461$$

$$x_2 = (0.461)^2 = 0.212$$

$$x_3 = (0.461)^3 = 0.0979$$

$$x_9 = (0.461)^9 = 9.4 \times 10^{-4} = 0.00094 \approx 0$$

(Q.2) The first serial correlation coefficient of MA(1) process is 0.3. compute the parameter of the process.

Sol:-

$$S_1 = 0.3$$

For MA(m) process

$$S_k = \frac{(-B_k + B_1 B_{k+1} + B_2 B_{k+2} + \dots + B_{m-k} B_m)}{1 + B_1^2 + B_2^2 + \dots + B_m^2}$$

Now,

For MA(1)

$$S_1 = \frac{-B_1}{1 + B_1^2} \Rightarrow 0.3 = \frac{-B_1}{1 + B_1^2}$$

$$\text{or } 0.3B_1^2 + B_1 + 0.3 = 0$$

Solving

$$B_1 = -0.33, -3$$

(Q.3) The hydrologic variable is described by an AR(1) model with mean = 100, variance = 900, first serial correlation coefficient = 0.5. compute standard deviation of random process.

Sol:-

$$\sigma_x = \sqrt{900} = 30$$

$$S_1 = 0.5$$

$$\alpha_1 = S_1 = 0.5$$

For AR(m) process

$$\sigma_e^2 = \sigma_x^2 (1 - \alpha_1 S_1 - \alpha_2 S_2 - \dots - \alpha_m S_m)$$

For m=1,

$$\sigma_e = \sigma_x \sqrt{1 - \alpha_1 S_1}$$

Thus

$$\sigma_e = 30 \sqrt{1 - 0.5 \times 0.5} = 26$$

(Q4) A moving average process is given by $(X_t - \mu) = e_t - 1.2e_{t-1} + 0.5e_{t-2}$

Determine the variance of X_t if the variance of e_t is 1.5.

→ Soln:

$$(X_t - \mu) = e_t - 1.2e_{t-1} + 0.5e_{t-2}$$

$$\beta_1 = 1.2, \beta_2 = -0.5$$

Thus,

$$\sigma_m^2 = \sigma_e^2 (1 + \beta_1^2 + \beta_2^2)$$

$$= 1.5 (1 + 1.2^2 + (-0.5)^2)$$

$$= 4.035$$

(Q5) Classify each of the following processes as AR, MA or ARMA, verify whether they are stationary and compute first five serial correlation coefficients.

(a) $(X_t - \mu) = 0.5(X_{t-1} - \mu) + e_t$

As $m=1$, and present value depends on past values and random process, it is AR(1) process.

$$S_1 = 0.5$$

Computing serial correlation coefficient.

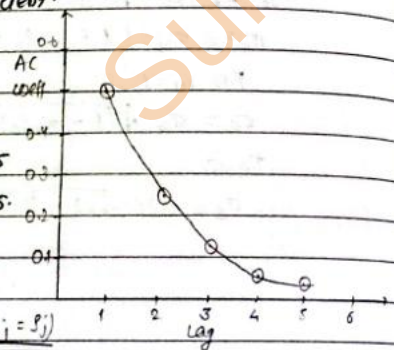
$$S_k = S_1^k$$

$$S_2 = (0.5)^2 = 0.25$$

$$S_3 = (0.5)^3 = 0.125$$

$$S_4 = (0.5)^4 = 0.0625$$

$$S_5 = (0.5)^5 = 0.03125$$



(?) As ACF is symmetrical about vertical axis, AR(1) model is stationary.

(b) $(X_t - \mu) = e_t + 1.3e_{t-1} + 0.4e_{t-2}$

✓ only random processes → MA

✓ $m=2$ ⇒ MA(2) model

✓ I+U always stationary

$$\beta_1 = -1.3; \beta_2 = -0.4$$

serial correlation coefficients

$$S_k = \frac{(-\beta_k + \beta_1\beta_{k+1} + \beta_2\beta_{k+2} + \dots + \beta_{m-k}\beta_m)}{(1 + \beta_1^2 + \beta_2^2 + \dots + \beta_m^2)}$$

$$S_1 = \frac{-\beta_1 + \beta_1\beta_2}{1 + \beta_1^2 + \beta_2^2} = \frac{-(-1.3) + (-1.3)(-0.4)}{1 + (-1.3)^2 + (-0.4)^2} = 0.64$$

$$S_2 = \frac{-\beta_2}{1 + \beta_1^2 + \beta_2^2} = \frac{-(-0.4)}{1 + (-1.3)^2 + (-0.4)^2} = 0.14$$

$$\text{For } k \geq 3, S_k = 0$$

(c) $(X_t - \mu) = 1.5(X_{t-1} - \mu) - 0.5(X_{t-2} - \mu) + e_t$

→ past values and random process & $m=2$

AR(2) process

→ For AR(2) to be stationary,

$$\alpha_2 + \alpha_1 < 1$$

$$\alpha_2 - \alpha_1 < 1$$

$$-1 < \alpha_2 < 1$$

Here, $\alpha_1 = 1.5$ & $\alpha_2 = -0.5$

$$1.5 - 0.5 = 1 \leq 1$$

$$-0.5 - 1.5 = -2 < 1$$

$$-1 < -0.5 < 1$$

As these conditions are satisfied, the process is stationary.

Serial correlation coefficients

$$S_k = \alpha_1 S_{k-1} + \alpha_2 S_{k-2} + \dots + \alpha_m S_{k-m}$$

$$\text{For } m=2, S_k = \alpha_1 S_{k-1} + \alpha_2 S_{k-2}$$

$$S_1 = \alpha_1 + \alpha_2 S_0 = 1.5 + (-0.5)S_0 \Rightarrow S_1 = 1$$

$$S_2 = \alpha_1 S_1 + \alpha_2 = 1.5 \times 1 - 0.5 = 1$$

$$S_3 = \alpha_1 S_2 + \alpha_2 S_1 = 1.5 \times 1 - 0.5 \times 1 = 1$$

$$S_4 = \alpha_1 S_3 + \alpha_2 S_2 = 1.5 \times 1 - 0.5 \times 1 = 1$$

$$S_5 = \alpha_1 S_4 + \alpha_2 S_3 = 1.5 \times 1 - 0.5 \times 1 = 1$$

$$(a) (X_t - \mu) = (X_{t-1} - \mu) + e_t + 0.5e_{t-1}$$

present value depends on

past values

$p=1$

present & past random processes $q=1$

ARMA process.

$$p=1, q=1 \Rightarrow \text{ARMA}(1,1)$$

$$\alpha_1 = 1, \beta_1 = -0.5$$

$$k=1 \quad S_1 = \frac{(1-\alpha_1\beta_1)(\alpha_1-\beta_1)}{(1+\beta_1^2-2\alpha_1\beta_1)} = \frac{(1-1 \times -0.5)(1-(-0.5))}{(1+(-0.5)^2-2 \times 1 \times -0.5)} = 1$$

$$S_k = \alpha_1 S_{k-1} \text{ for } k \geq 2$$

$$S_2 = \alpha_1 S_1 = 1 \times 1 = 1$$

$$S_3 = \alpha_1 S_2 = 1 \times 1 = 1$$

$$S_4 = \alpha_1 S_3 = 1 \times 1 = 1$$

$$S_8 = \alpha_1 S_7 = 1 \times 1 = 1$$

For ARMA(1,1) to be stationary

$$|\beta_2| < |\beta_1|$$

$$1 < 1$$

$$S_2 > S_1(2\beta_1 - 1) \text{ if } S_1 > 0 \quad \left. \begin{array}{l} 1 < 1 \\ 1 > 1(2 \times 1 - 1) \end{array} \right\}$$

As these conditions are satisfied, the process is stationary.

6. For MA(2) process, following data are given

$$\sigma_1 = 0.4, \sigma_2 = -0.19$$

compute parameters β_1 and β_2 .

Soln:-

For MA(m), autocorrelation function is:-

$$\sigma_k = \frac{(-\beta_k + \beta_1\beta_{k+1} + \beta_2\beta_{k+2} + \dots + \beta_{m-k}\beta_m)}{(1 + \beta_1^2 + \beta_2^2 + \dots + \beta_m^2)}$$

For MA(2)

$$\sigma_1 = \frac{-\beta_1 + \beta_1\beta_2}{1 + \beta_1^2 + \beta_2^2}$$

$$\text{or, } 0.4 = \frac{-\beta_1 + \beta_1\beta_2}{1 + \beta_1^2 + \beta_2^2} \quad \text{--- (a)}$$

$$\sigma_2 = \frac{-\beta_2}{1 + \beta_1^2 + \beta_2^2}$$

$$\text{or, } -0.19 = \frac{-\beta_2}{1 + \beta_1^2 + \beta_2^2} \quad \text{--- (b)}$$

$$\text{eq}^n (a) \div \text{eq}^n (b)$$

$$\frac{0.4}{-0.19} = \frac{-\beta_1 + \beta_1\beta_2}{-\beta_2}$$

$$\text{or, } \beta_1 = \frac{2.1\beta_2}{1 - \beta_2} \quad \text{--- (c)}$$

Substituting in (b)

$$\text{or, } -0.19 = \frac{-\beta_2}{1 + \left(\frac{2.1\beta_2}{1 - \beta_2}\right)^2 + \beta_2^2}$$

Solving,

$$\beta_2 =$$

- (9) A second order AR process is given by
 $(X_t - \mu) = (X_{t-1} - \mu) - 0.5(X_{t-2} - \mu) + e_t$
 Compute first 15 correlation coefficients and plot the correlogram. From the plotted correlogram, find the frequency of the damped sine wave. Compare this with the value from the equation of frequency computation for the damped sine wave.

→ Solⁿ:-

$$(X_t - \mu) = (X_{t-1} - \mu) - 0.5(X_{t-2} - \mu) + e_t$$

$$\alpha_1 = 1, \alpha_2 = -0.5$$

Serial correlation coefficients,

$$S_k = \alpha_1 S_{k-1} + \alpha_2 S_{k-2} + \dots + \alpha_m S_{k-m}$$

For $m=2$, ACF of AR(2) model is

$$S_k = \alpha_1 S_{k-1} + \alpha_2 S_{k-2}$$

At $k=1$ $S_1 = \alpha_1 \times 1 + \alpha_2 \times S_0 = 1 + (-0.5) \times 1$ i.e. $S_1 = 0.667$

$$S_2 = \alpha_1 S_1 + \alpha_2 \times 1 = 1 \times 0.667 + (-0.5) \times 1 = 0.167$$

$$S_3 = \alpha_1 S_2 + \alpha_2 S_1 = 1 \times 0.167 + (-0.5) \times 0.667 = -0.167$$

Similarly $S_4 = \alpha_1 S_3 + \alpha_2 S_2 = -0.25$

$$S_5 = -0.167$$

$$S_6 = -0.042$$

$$S_7 = 0.0415$$

$$S_8 = 0.0626$$

$$S_9 = 0.0105$$

$$S_{10} = 0.0105$$

$$S_{11} = 0.0104$$

$$S_{12} = -0.0157$$

$$S_{13} = -0.0522$$

$$S_{14} = -0.044$$

$$S_{15} = \alpha_1 S_{14} + \alpha_2 S_{13} = 1 \times (-0.044) - 0.5 \times (-0.0522) = -0.018$$

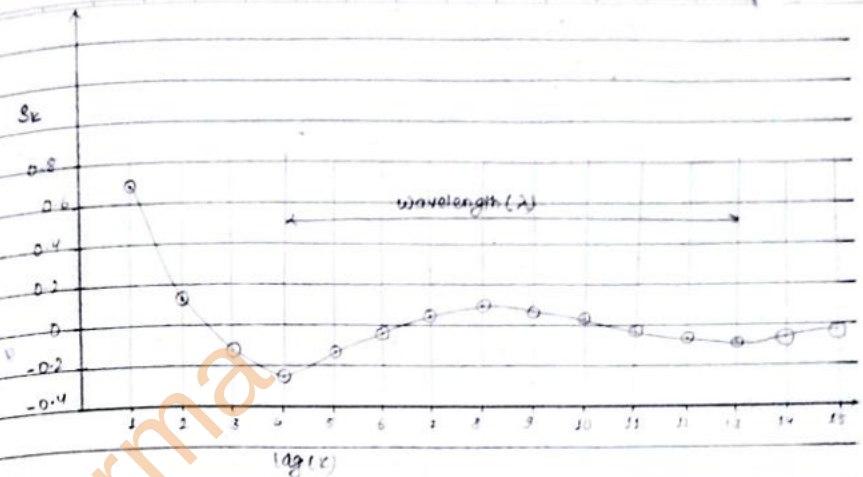


Fig: Plot of ACF

Wavelength (λ) = 9
 $f_0 = 1/\lambda = 0.11$

From equation,

$$\cos(2\pi f_0) = \frac{|\alpha_1|}{2\sqrt{-\alpha_2}} = \frac{1}{2\sqrt{0.5}}$$

$$\Rightarrow [f_0 = 0.125]$$

- (9) Following are the annual flow values (m^3/s) of a river for 10 years:

300, 250, 415, 723, 525, 600, 910, 375, 550, 823

Assuming AR(1) model is good fit, find the parameter of the model. If MA(1) model is fitted for the data, what will be the parameter?

→ Solⁿ:-

Parameter of AR(1) model (α) = AC coeff. of lag(1)

Autocorrelation coeff. or serial correlation coefficient is given by

$$r_k = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sqrt{\sum_{t=1}^{n-k} (x_t - \bar{x})^2 \sum_{t=1}^{n-k} (x_{t+k} - \bar{x})^2}}$$

where, $k = \text{lag}$ $[k=1]$
computing $\bar{x}_t$,

	x_t	x_{t+k}	$x_t - \bar{x}_t$	$(x_{t+k} - \bar{x}_{t+k})$	$(x_t - \bar{x}_t)$	$(x_t - \bar{x}_t)^2$	$(x_{t+k} - \bar{x}_{t+k})^2$
1	300	250	-216.4	-324.5	70221.8	46828.96	105300.2
2	250	415	-266.4	-159.5	42490.8	70968.96	25440.25
3	415	723	-101.4	148.5	-15057.9	10281.96	22052.25
4	723	525	206.6	-49.5	-10226.7	42683.56	2450.25
5	525	600	8.6	25.5	219.3	73.96	650.25
6	600	910	83.6	335.5	28047.8	63889.6	112580.2
7	910	875	383.6	-139.5	-78523.2	154921	39800.25
8	375	550	-141.4	-24.5	3464.3	19993.96	600.25
9	550	823	38.6	248.5	8349.6	11289.6	61752.25
					48985.8	353870.2	370606.3

Mean of $x_t = \bar{x}_t = 516.44$

Mean of $x_{t+k} = \bar{x}_{t+k} = 574.96$

Substituting values in formula of $\bar{x}_k$

$$\bar{x}_1 = \frac{48985.8}{\sqrt{353870.2 \times 370606.3}} = 0.1353 \approx 0.135$$

Hence, $\alpha = 0.135$ (for AR(1) model)

For MA(1)

$$\sigma_1 = -\beta_1$$

$$1 + \beta_1^2$$

$$\text{On } 0.135 = \frac{-\beta_1}{1 + \beta_1^2}$$

$$\text{On } 0.135\beta_1^2 + \beta_1 + 0.135 = 0$$

Solving,

$$\omega [\beta_1 = -0.137, -7.27]$$

$$\text{Confidence limit (CL)} = \frac{-3 \pm 0.596 \sqrt{n-k-1}}{n-k}$$

$n \rightarrow$ no. of obs. $k \rightarrow$ given lags

AR

Linear combination of past observations

$$\text{① AR}(m) \quad x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \dots + \alpha_m x_{t-m} + \dots + e_t \quad \text{--- (1)}$$

$\alpha_1, \alpha_2, \dots, \alpha_m \rightarrow$ parameters

$e_t =$ random component with zero mean

$m =$ order of model

(no. of time lags considered)

'or'

$$\text{② } (x_t - \mu) = \alpha_1 (x_{t-1} - \mu) + \alpha_2 (x_{t-2} - \mu) + \dots + \alpha_m (x_{t-m} - \mu) + \dots + e_t \quad \text{--- (2)}$$

$\mu =$ expected value or mean of process

$\rightarrow \rho =$ population autocorrelation coeff

$$[\rho_k = \alpha_1 \rho_{k-1} + \alpha_2 \rho_{k-2} + \dots + \alpha_m \rho_{k-m}]$$

($\rho_0 = 1$ and $\rho_{-j} = \rho_j$ for stationary process)

$$\rightarrow \sigma_a^2 = \frac{\sigma_e^2}{1 - \alpha_1 \rho_1 - \alpha_2 \rho_2 - \dots - \alpha_m \rho_m}$$

$$e_t = Z_t \sigma_a \sqrt{1 - \rho_1^2}$$

* AR(1) or ~~AR~~ Thomas-Fiering model

$$(x_t - \mu) = \alpha_1 (x_{t-1} - \mu) + e_t$$

$$\rho_k = \alpha_1 \rho_{k-1} + \alpha_2 \rho_{k-2} + \dots + \alpha_m \rho_{k-m}$$

For $m=1$,

$$\rho_k = \alpha_1 \rho_{k-1}$$

$$\rho_1 = \alpha_1, \rho_0 = \alpha_1$$

$$\rho_2 = \alpha_1^2, \rho_1 = \alpha_1, \rho_0 = \alpha_1^2$$

$$[\rho_k = \alpha_1^k]$$

$$\sigma_a^2 = \frac{\sigma_e^2}{(1 - \rho_1^2)} = \frac{\sigma_e^2}{(1 - \alpha_1^2)}$$

* AR(2) Model

$$(x_t - \mu) = \alpha_1 (x_{t-1} - \mu) + \alpha_2 (x_{t-2} - \mu) + e_t$$

$$S_t = \alpha_1 S_{t-1} + \alpha_2 S_{t-2} + \dots + \alpha_m S_{t-m}$$

for $k \geq 2$, AR(2)

AR(2) to be stationary,

$$\begin{bmatrix} \alpha_1 + \alpha_2 < 1 \\ \alpha_2 - \alpha_1 < 1 \\ -1 < \alpha_2 < 1 \end{bmatrix} \rightarrow 3 \text{ conditions}$$

MA $\rightarrow$ linearly dependent on present & past stochastic value,

$$x_t = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_m e_{t-m} \quad (1)$$

$\beta_1, \beta_2, \dots, \beta_m \rightarrow$ parameters

'OR'

$$(x_t - \mu) = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_m e_{t-m} \quad (2)$$

$$\rightarrow \sigma_y^2 = \sigma_e^2 (1 + \beta_1^2 + \beta_2^2 + \dots + \beta_m^2)$$

$$\rightarrow S_k = \frac{-\beta_k + \beta_1 \beta_{k+1} + \beta_2 \beta_{k+2} + \dots + \beta_{m-k} \beta_m}{(1 + \beta_1^2 + \beta_2^2 + \dots + \beta_m^2)}$$

($S_k = 0$ for $k > m$)

(ACF has cutoff at lag m)

ARMA

4 part values of variables & present & past values of random comp

ARMA(p, q):

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \dots + \alpha_p x_{t-p} + \dots + e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_q e_{t-q}$$

'OR'

$$(x_t - \mu) = \alpha_1 (x_{t-1} - \mu) + \alpha_2 (x_{t-2} - \mu) + \dots + \alpha_p (x_{t-p} - \mu) - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_q e_{t-q}$$

ARMA(1,1)

$$\sigma_y^2 = \sigma_e^2 \left(\frac{1 + \beta_1^2 - 2\alpha_1\beta_1}{(1 - \alpha_1^2)} \right)$$

$$S_1 = \frac{(1 - \beta_1\alpha_1)(\alpha_1 - \beta_1)}{(1 + \beta_1^2 - 2\alpha_1\beta_1)}$$

$$S_k = \alpha_1 S_{k-1} \text{ for } k \geq 2$$

ACF of ARMA(1,1) decays exponentially from a starting value of S_1 .

For stationary,

$$\begin{cases} |\beta_2| < |\beta_1| \\ S_2 > S_1(2S_1 + 1) & \text{if } S_1 < 0 \\ S_2 > S_1(2S_1 - 1) & \text{if } S_1 > 0 \end{cases}$$

Random Number

① Mid Square Method

Seed = 4538

$$4538^2 = 20593444 \rightarrow \text{next RN} = 5934$$

$$5934^2 = 35212356 \rightarrow \text{next RN} = 2123$$

② Mid product technique

$$a_1 = 6723, a_2 = 4591$$

seeds: 6723 and 4591

$$6723 \times 4591 = 30865933 \rightarrow a_3 = 8652$$

$$4591 \times 8652 = 39721332 \rightarrow a_4 = 7213$$

$$8652 \times 7213 = 62406876 \rightarrow a_5 = 4068$$

③ Mixed Congruential method

Recursive formula:

$$u_k = (au_{k-1} + c) \bmod m$$

where u_k = current random number

u_{k-1} = previous " "

a = Multiplier, c = increment, m = modulus

$u \bmod$ refers to integer division (remainder left)

Choice of a, c , and m

• Binary computer

$$m = 2^b$$

c = any odd integer

$$a \bmod 4 = 1$$

u_0 = any integer seed

$$\text{s.t. } 0 \leq u_0 \leq m$$

• Decimal comp

$$m = 10^d$$

c = any integer divisible by either 2 or 5.

$$a \bmod 20 = 1$$

u_0 = any integer seed

$$\text{s.t. } 0 \leq u_0 \leq m$$

NOTES

② Central limit theorem method

mean of each var = 0.5

var. of each var = $1/12$

$$X = U_1 + U_2 + \dots + U_m$$

mean of $X = 0.5m$

var of $X = m/12$

$$Z_j = \frac{\sum_{i=1}^m (u_k - 0.5m)}{\sqrt{m/12}} \quad \text{where, } k = (j-1)m + i$$

In practice, $m = 12$ or 6 is used.

⑥ Box-Muller method

$$Z_i = [-2 \ln(1 - u_i)]^{1/2} \cos(2\pi u_{i+1})$$

$$Z_{i+1} = [-2 \ln(1 - u_i)]^{1/2} \sin(2\pi u_{i+1})$$

③ Inverse transformation method

z = standard normal variate

$$\text{CDF} = F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt$$

$$[u = \text{new RV} = F(z)]$$

z found from table by interpolation

$U_1, U_2, \dots, U_m$
are m independent
and uniformly distributed
variable with range
0 to 1

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

Detection of trend

(i) Regression test for linear trend:

Fit st. line: $y = at + b$

For linear regression analysis,

$$a = \frac{N(\sum ty) - (\sum t)(\sum y)}{N(\sum t^2) - (\sum t)^2}$$

$$b = \frac{\sum y - a(\sum t)}{N}$$

(ii) Spearman's Rank correlation Test

Min = data seq $\rightarrow 10$

↳ Arrange data in time order $\rightarrow R(t)$

↳ Assign ranks to x (ascending) $\rightarrow R(x)$

↳ difference: $d = R(x) - R(t) \rightarrow d^2 \rightarrow \sum d^2$

↳ Spearman's Rank correlation:

$$R_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

↳ significance test (convert to t)

$$t = R_s \sqrt{\frac{n-2}{1-R_s^2}}$$

↳ compare: computed $t <$ tabulated $d \Rightarrow$ No trend

(iii) Mann-Kendall test:

↳ Rank of data in ascending order of magnitude

↳ compare pairs:

$$\text{sgn}(x_j - x_i)$$

Rule: +1 if $x_j > x_i$

0 if equal

-1 if $x_j < x_i$

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(R_i - R_j)$$

↳ variance of S:

(i) No tied value: $\text{Var}(S) = \frac{n(n-1)(2n+5)}{18}$

(ii) Tied values: $\text{Var}(S) = \frac{n(n-1)(2n+5) - \sum t_i(t_i-1)(2t_i+5)}{18}$

↳ z-test statistic: $z = \frac{|S|}{\sqrt{\text{Var}(S)}}$

↳ Tabulated value of z

↳ computed $z <$ tabulated $z, \Rightarrow$ no trend

Removal of data

- (i) Moving average Method
- (ii) Linear trend (fit st-line)

Detection of periodicity.

- (i) Visual inspection
- (ii) ACF (if periodic ACF shows repeating peaks)

$$r_k = \frac{\sum (x_t - \bar{x})(x_{t-k} - \bar{x}_{t-k})}{\sum (x_t - \bar{x})^2}$$

- (iii) Spectral analysis (time series $\rightarrow$ frequency domain)
Peak in spectrum $\Rightarrow$ strong periodicity

Detection of jump: visual inspection

Checking stationarity: (Split Record Test)

- ↳ Divide records into two non-overlapping subsets $\overbrace{(1, 2, \dots, m)}^{S_1}$ & $\overbrace{(m+1, \dots, n)}^{S_2}$
- ↳ Compute mean and variance of each subset ($\bar{x}_1, \bar{x}_2$ & s_1^2, s_2^2)
- ↳ Examine variance (V) using F-test:

$$F = \frac{s_1^2}{s_2^2} \quad \{s_1^2 > s_2^2\}$$

- ↳ ~~Examine~~ compare computed F with tabulated F
(at $m-1$ & $n-m-1$ DOF at $\alpha/2$ level of significance)

- a) $\hookrightarrow$ computed F < tabulated F $\Rightarrow$ variances subsets can be accepted as estimator of same population variance.

- ↳ Compute pooled estimate of variance,

$$V_p = \frac{(m-1)(s_1^2) + (n-m-1)(s_2^2)}{n-2}$$

- ↳ Test statistic for mean

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \left. \vphantom{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right\} \text{or } t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{V_p}{n-m}}}$$

- ↳ compare computed t with tabulated t

(for $n-2$ DOF at $\alpha/2$ level of significance)

- b) $\hookrightarrow$ Computed t < tabulated t $\Rightarrow$ mean zero difference b/w two subsets is accepted.

Both (a) & (b) satisfied $\Rightarrow$ stationary series.

② # Checking consistency: Double Mass curve

- ↳ Compute cumulative sum of data of base station (A)
- ↳ " " " " mean value of neighbouring station (B)
- ↳ Plot A vs B & fit points by straight line.
- ↳ If no break in line, data is considered to be consistent.

Chapter-3 (Statistical Principles)

* Frequency $\rightarrow$ No. of occurrences of a variate

* Relative frequency function ($f_r(x)$) $= \frac{n_i}{n} = P(x)$
(probability of a function)

* PMF $\rightarrow$ gives probability associated with all possible values of RV

CDF $\rightarrow$ probability of event that the RV X is less than or equal to a , for every possible value of a .

Discrete RV

- PMF $\left\{ \begin{array}{l} P(x_i) = P(X = x_i) \\ 0 \leq P(x_i) \leq 1 \\ \sum P(x_i) = 1 \\ P(a \leq X \leq b) = \sum_{x_i \leq b}^{x_i \geq a} P(x_i) \end{array} \right.$

- CDF $\left\{ \begin{array}{l} F(x) = P[X \leq x] \\ \text{Found by summing of probabilities} \end{array} \right.$

$$P(X = x_i) = F(x_i) - F(x_{i-1})$$

* PDF & CDF (for continuous RV)

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} f(x) dx$$

$f(x) \rightarrow$ probability at pt. x .

$$f(x) \geq 0 \quad \& \quad \int_{-\infty}^{\infty} f(x) dx = 1$$

$$P[a \leq x \leq b] = F(b) - F(a)$$

$$f(x) = \frac{dF(x)}{dx} ; f(x) = \text{pdf} \& F(x) = \text{CDF}$$

Joint Probability distribution (JPD).

↳ probability distribution of more than one RV.

↳ For two discrete RV X & Y , JPD:-

$$\text{Condition} \left\{ \begin{array}{l} \rightarrow P(x, y) = P[X=x, Y=y] \\ \rightarrow P(x, y) \geq 0 \\ \rightarrow \sum_x \sum_y P(x, y) = 1 \end{array} \right.$$

• Joint CDF $\rightarrow F(x, y) = P[X \leq x, Y \leq y]$

$$= \sum_{s \leq x} \sum_{t \leq y} P(s, t)$$

for $-\infty < x < \infty$ & $-\infty < y < \infty$.

$P(s, t) \rightarrow$ JPD of X & Y at (s, t) .

↳ For two continuous RV.

$$\text{Condition} \left\{ \begin{array}{l} \rightarrow P(x, y) = P[X=x, Y=y] \\ \rightarrow P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) \\ \quad = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y) dx dy \\ \rightarrow f(x, y) \geq 0 \text{ for } -\infty < x < \infty \text{ & } -\infty < y < \infty \\ \rightarrow \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1. \end{array} \right.$$

• Joint CDF = $P[-\infty < X \leq x, -\infty < Y \leq y]$

$$= \int_{-\infty}^x \int_{-\infty}^y f(s, t) ds dt$$

• $f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$

③ # Marginal distribution

↳ Distribution of one of the variables obtained by integrating out the variable that is no longer of interest
(prob. of one variable ignoring the other)

↳ Given: $P(x, y)$ = JPD of two RV X & Y .

↳ Discrete

$$\begin{aligned} \text{Marginal PMF of } X &= g(x) = \sum_{all y} P(x, y) \\ \text{" " " } Y &= h(y) = \sum_{all x} P(x, y) \end{aligned}$$

↳ Continuous

$$\begin{aligned} \text{Marginal PDF of } X &= g(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{for } -\infty < x < \infty \\ \text{" " " } Y &= h(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad \text{" " " } \end{aligned}$$

Conditional Probability

↳ Probability of one variable given other

↳ Given $f(x, y)$ = JPD
 $g(x)$ & $h(y)$.

$$\text{Conditional prob. of } X = f(x|y) = \frac{f(x, y)}{h(y)}$$

$$\text{Conditional prob. of } Y = f(y|x) = \frac{f(x, y)}{g(x)}$$

$$\text{If } f(x, y) = g(x) \cdot f(y|x) \text{ \& } f(y|x) = h(y) \Rightarrow f(x, y) = g(x) \cdot h(y).$$

↳ two RV are independent

Derived Distribution

↳ New variable formed from existing variable.

↳ Given, $f(x)$ as PDF of x

Find: PDF of Y say $g(y)$

↳ Area of PDF of X = Area of PDF of Y

$$f(x) \cdot dx = g(y) \cdot dy$$

$$\Rightarrow g(y) = f(x) * \left| \frac{dx}{dy} \right|$$

Sample statistics

$$\rightarrow CV = \frac{\sigma}{\bar{x}}$$

$$\rightarrow \text{coeff. of skewness, } C_s = \frac{n \sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(n-2)s^3}$$

$$\rightarrow \text{skewness} = \sum_{i=1}^n \frac{(x_i - \bar{x})^3}{n s^3}$$

Population statistics

$$\hookrightarrow \text{Mean} = \mu = E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

$$\hookrightarrow \text{Median} = a \cdot s.t. \int_0^x f(x) dx = 0.5$$

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

$$\hookrightarrow CV = \frac{\sigma}{\mu}$$

$$\hookrightarrow \text{Skewness} = E[(x - \mu)^3] = \int_{-\infty}^{\infty} (x - \mu)^3 f(x) dx$$

$$\hookrightarrow C_s = \frac{E[(x - \mu)^3]}{\sigma^3}$$

$$\hookrightarrow \text{Kurtosis} = E[(x - \mu)^4] = \int_{-\infty}^{\infty} (x - \mu)^4 f(x) dx$$

$$\hookrightarrow C_k = \frac{E[(x - \mu)^4]}{\sigma^4}$$

For discrete replace $f(x)dx$ by $p(x)$ & apply $\sum$ sign

Covariance b/w two RV X & Y

$$\hookrightarrow \text{cov}(X, Y) = C_{xy} = E[(X - \mu_x)(Y - \mu_y)]$$

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

$$f(x, y) \rightarrow \text{JPD}$$

$$\hookrightarrow \text{cov}(x, y) = E(xy) - E(x)E(y)$$

Chebyshev's inequality

For any $k > 1$ at least $1 - 1/k^2$ of data from the sample must fall within k standard deviations from mean.

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

④

Moment Generating Function

↳ way of finding moments like mean, variance & other moments.

↳ Discrete RV: $M_X(t) = E[e^{tX}] = \sum_{\alpha} e^{t\alpha} p(\alpha)$

↳ Continuous RV: $M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{t\alpha} f(\alpha) d\alpha$

↳ r th moment about origin:

$$\mu_r' = \frac{d^r M_X(t)}{dt^r} \text{ at } t=0$$

$$u \int v da = u \int v da - \int \left(\frac{du}{da} \int v da \right) da$$

Estimating parameters

(i) Method of moments (Karl Pearson method)

✓ For k no. of parameters, first k pop^l & sample moments are equated and solved simultaneously.

eg: $\mu = \int_{-\infty}^{\infty} \alpha f(\alpha) d\alpha$

(ii) Method of maximum likelihood (Fisher method)

↳ $f(\alpha, \beta, \dots)$ → pdf of RV X with parameter $\alpha, \beta, \dots$

↳ $(x_1, x_2, \dots, x_n)$ → observed sample

↳ Likelihood function = $f(x_1, \alpha, \beta, \dots) f(x_2, \alpha, \beta, \dots) \dots f(x_n, \alpha, \beta, \dots)$

Or, $L(\alpha, \beta, \dots) = \prod_{i=1}^n f(x_i, \alpha, \beta, \dots)$

↳ Find value of $\alpha, \beta, \dots$ which maximize likelihood function

↳ log-likelihood.

$$\begin{aligned} \ln [L(\alpha, \beta, \dots)] &= \ln \left[\prod_{i=1}^n f(x_i, \alpha, \beta, \dots) \right] \\ &= \sum_{i=1}^n \ln [f(x_i, \alpha, \beta, \dots)] \end{aligned}$$

↳ $\frac{\partial}{\partial \alpha} \ln(L) = 0, \frac{\partial}{\partial \beta} \ln(L) = 0 \dots$

to find parameters.

(a) Chi-square test

↳ to determine whether there is statistically significant diff. b/w expected and observed frequencies.

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}$$

n_i → no. of observation in given CI

e_i → expected no. of obs. in given CI.

k → no. of CI.

h → no. of parameters

No. of DOF = ~~k~~ $k - h - 1$

i	CI	n_i	7 th of upper class	$F(z_i)$	P_i	e_i	χ^2
	$<$		$Z_i = \frac{x_{upper} - \bar{x}}{\sigma}$	↓ Obtained from table	$F(z_i) - F(z_{i-1})$		
	$-$						
	$-$						
	$>$						
		$\sum n_i = n$					
							$\sum \Rightarrow (\chi^2)_{\text{computed}}$

Computed $\chi^2 < \text{Tabulated } \chi^2 \Rightarrow$ normal distribution appropriate for data.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

(b) K-S test

$\bar{x} \checkmark \quad \sigma \checkmark$

Given data	x	Rank (m)	Exceedence prob	$P(x)$	z	$F(x) = F(z)$	Delta
		1					
	↓ descending order	↓ n	Weibull's formula " $\frac{m}{m+1}$ "	1 - exceed. prob.	$\frac{x - \bar{x}}{\sigma}$	calc ✓ Normal CD ✓ Lower = -1000 ✓ upper = z ✓ $F(z)$	$ P(x)_n - F(x_m) $

$$\Delta = \max \Delta f |P(x_m) - F(x_m)|$$

~ K-S statistic Δ_0 for sample size n & 5% or 10% significance

~ $\Delta < \Delta_0 \Rightarrow$ assumed distribution good fit for data.

⑤ Ch-5 Auto correlation Analysis.

Correlation (r) = $\frac{COV(X,Y)}{\sigma_x \sigma_y} = \frac{S_{x,y}}{S_x S_y}$ } degree of association b/w two variables

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}$$

Regression

For $Y = a + bX$

$$a = \frac{(\sum Y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum Y)}{n(\sum x^2) - (\sum x)^2}$$

Autocorrelation

↳ correlation of a variable b/w past & future values.

Autocorrelation coeff. or serial correlation coeff. (r_k)

$$r_k = \frac{\frac{1}{n-k} \sum_{t=1}^{n-k} (x_t - \bar{x}_t)(x_{t+k} - \bar{x}_{t+k})}{\sqrt{\left(\frac{1}{n-k} \sum_{t=1}^{n-k} (x_t - \bar{x}_t)^2 \right) \left(\frac{1}{n-k} \sum_{t=1}^{n-k} (x_{t+k} - \bar{x}_{t+k})^2 \right)}}$$

If $n \gg k_{\max}$ (large n)

↳ Mean & variance of whole series can be used

Simplified form

$$r_k = \frac{\frac{1}{n-k} \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\frac{1}{n} \sum_{t=1}^n (x_t - \bar{x})^2}$$

(k = time lag, n = no. of data)

~ Plot of r_k v/s $k \Rightarrow$ correlogram or ACF

~ PACF $\rightarrow$ plot of r_k by neglecting some values.

For independent process

Confidence limits (CL)

$$CL(r_k) = \frac{-1 \pm 1.96 \sqrt{n-k-1}}{n-k}$$

(for 95% confidence)

If actual r_k for various lags fall within tolerance band, the sample data is considered to be independent.

Ch-5 Time series Models

Autoregressive model (AR(m))

$$(X_t - \mu) = \alpha_1 (X_{t-1} - \mu) + \alpha_2 (X_{t-2} - \mu) + \dots + \alpha_m (X_{t-m} - \mu) + \dots + e_t \quad \text{---(i)}$$

✓ depends on past values.

• Yule-Walker equation :

$$s_k = \alpha_1 s_{k-1} + \alpha_2 s_{k-2} + \dots + \alpha_m s_{k-m} \quad \text{---(ii)}$$

$s \rightarrow$ population autocorrelation coeff.

$$k = 1, 2, \dots, m$$

$$s_0 = 1 \text{ \& } s_{-j} = s_j \text{ (for stationary process)}$$

($\alpha \rightarrow$ parameters)

$$\sigma_\alpha^2 = \frac{\sigma_e^2}{(1 - \alpha_1 s_1 - \alpha_2 s_2 - \dots - \alpha_m s_m)} \quad \text{---(iii)}$$

$\gamma(0) = \sigma_\alpha^2 \rightarrow$ variance of AR(m) process
 $\sigma_e^2 \rightarrow$ variance of random/independent process.

• AR(1)

$$(X_t - \mu) = \alpha_1 (X_{t-1} - \mu) + e_t \quad \text{---(i)}$$

$$s_k = \alpha_1 s_{k-1} \quad \text{---(ii)}$$

$$s_k = (s_1)^k \quad \text{---(iii)}$$

$$\sigma_\alpha^2 = \frac{\sigma_e^2}{(1 - s_1^2)} \quad \text{---(iv)}$$

• If e_t written as,

$$e_t = z_t \sigma_\alpha \sqrt{1 - s_1^2}$$

$$(X_t - \mu) = s_1 (X_{t-1} - \mu) + z_t \sigma_\alpha \sqrt{1 - s_1^2}$$

$\propto$ Population parameter replaced by sample parameter

$$g_t = \bar{g} + \alpha_1 (g_{t-1} - \bar{g}) + z_t s_\alpha \sqrt{1 - \alpha_1^2}$$

Thomas-Fiering Model

AR(2)

$$(X_t - \mu) = \alpha_1 (X_{t-1} - \mu) + \alpha_2 (X_{t-2} - \mu) + e_t$$

$$s_k = \alpha_1 s_{k-1} + \alpha_2 s_{k-2}$$

$$k=1: s_1 = \alpha_1 + \alpha_2 s_1$$

$$k=2: s_2 = \alpha_1 s_1 + \alpha_2$$

Solving

$$\alpha_1 = \frac{s_1 - s_1 s_2}{1 - s_1^2}; \alpha_2 = \frac{s_2 - s_1^2}{1 - s_1^2}$$

• For stationary,

$$\alpha_2 + \alpha_1 < 1 \quad \left. \begin{array}{l} -1 < s_1 < 1 \\ -1 < s_2 < 1 \\ s_1^2 < \frac{1}{2}(s_2 + 1) \end{array} \right\}$$

$$\alpha_2 - \alpha_1 < 1$$

$$-1 < \alpha_2 < 1$$

• Applicability

$$g_t = \bar{g} + \alpha_1 (g_{t-1} - \bar{g}) + \alpha_2 (g_{t-2} - \bar{g}) + z_t s_\alpha \sqrt{1 - \rho^2}$$

$$\alpha_1 = \frac{\sigma_1 - \sigma_1 \sigma_2}{1 - \sigma_1^2}; \alpha_2 = \frac{\sigma_2 - \sigma_1^2}{1 - \sigma_1^2}$$

$$\rho^2 = \frac{\sigma_1^2 + \sigma_2^2 - 2\sigma_2 \sigma_1^2}{1 - \sigma_1^2}$$

$\rho \rightarrow$ Coeff. of multiple correlation.

Moving Average Model MA(m).

$$(X_t - \mu) = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_m e_{t-m} \quad (i)$$

$$\gamma(0) = \sigma_x^2 = \sigma_e^2 (1 + \beta_1^2 + \beta_2^2 + \dots + \beta_m^2) \quad (ii)$$

$$s_k = \frac{(-\beta_k + \beta_1 \beta_{k+1} + \beta_2 \beta_{k+2} + \dots + \beta_{m-k} + \beta_m)}{(1 + \beta_1^2 + \beta_2^2 + \dots + \beta_m^2)} \quad (iii)$$

$$s_k = 0 \text{ for } k > m.$$

Eg. MA(1).

$$s_1 = \frac{-\beta_1}{1 + \beta_1^2}$$

MA(2)

$$s_1 = \frac{-\beta_1 + \beta_1 \beta_2}{1 + \beta_1^2 + \beta_2^2}; \quad s_2 = \frac{-\beta_2}{1 + \beta_1^2 + \beta_2^2}$$

ARMA model.

$$(X_t - \mu) = \alpha_1 (X_{t-1} - \mu) + \alpha_2 (X_{t-2} - \mu) + \dots + \alpha_p (X_{t-p} - \mu) + e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \dots - \beta_q e_{t-q} \quad (i)$$

ARMA(p, q) model.

• ARMA(1, 1).

$$(X_t - \mu) = \alpha_1 (X_{t-1} - \mu) + e_t - \beta_1 e_{t-1} \quad (i)$$

$$\sigma_x^2 = \frac{\sigma_e^2 (1 + \beta_1^2 - 2\alpha_1 \beta_1)}{(1 - \alpha_1^2)}$$

$$s_1 = \frac{(1 - \beta_1 \alpha_1)(\alpha_1 - \beta_1)}{(1 + \beta_1^2 - 2\alpha_1 \beta_1)}$$

$$s_k = \alpha_1 s_{k-1} \text{ for } k \geq 2$$

• For stationary:

$$\checkmark |s_2| < |s_1|$$

$$\checkmark s_2 > s_1 (2s_1 + 1) \text{ for } s_1 < 0$$

$$s_2 > s_1 (2s_1 - 1) \text{ for } s_1 > 0.$$

• Applicability in hydrology forecasting:

$$\hat{Q}_t = \bar{Q} + \alpha (\hat{Q}_{t-1} - \bar{Q}) + T (z_t - b z_{t-1}) \quad (ii)$$

$$\checkmark e_t = z_t \sqrt{\frac{\sigma_x^2 (1 - \alpha^2)}{1 + \beta^2 - 2\alpha\beta}}$$

$$\checkmark \alpha = \frac{\sigma_2}{\sigma_1} \quad \checkmark \sigma_1 = \frac{(1 - \alpha b)(\alpha - b)}{1 + b^2 - 2\alpha b}$$

$$\checkmark T = s_x \sqrt{\frac{1 - \alpha^2}{1 + b^2 - 2\alpha b}} \quad (b) \checkmark$$

Akaike Information Criteria (AIC)

$$AIC = N \ln(\text{likelihood}) + 2k$$

$N \rightarrow$ sample size

$k \rightarrow$ no. of parameters.

For AR(m) model,

$$AIC(m) = -N \ln(\sigma_e'^2) + 2m$$

$N \rightarrow$ no. of obs.

$m =$ no. of parameter

$\sigma_e' =$ maximum likelihood estimate of variance of residual.

✓ Lower AIC = better model

$$AIC(m) = N \ln(\sigma_e^2) + 2m$$

$\sigma_e^2 \rightarrow$ variance of residual.

Testing Independence of residuals

I. Plot of ACF

II. Porte Monteau lack of fit test:

✓ Test statistic (Q)

$$Q = N \sum_{k=1}^L r_k^2(e)$$

$r_k(e) \rightarrow$ autocorrelation of residual

$L =$ no. of lags (gen. 30% of sample size)

$N =$ sample size.

✓ obtain χ^2 for $(L-p-q)$ DOF.

$p, q \rightarrow$ order of ARMA model.

✓ $Q < \chi^2 \rightarrow$ residuals are independent.

III. Testing of parsimony of parameter

✓ use minimum number of parameters

✓ selecting optimum order of model

✓ Larger likelihood $\rightarrow$ better model

✓ more parameter $\rightarrow$ worse model

✓ Minimum AIC(m) is selected.

Ch-6 Spectral Modes.

- Fourier Series $\rightarrow$ resolve ^(power signal) periodic signal into an infinite sum of sinusoidal wave components
- Fourier Transform $\rightarrow$ does similar analysis for aperiodic signal (energy signal)
- Finite FS representation of discrete RV:

$$x_t = \bar{x} + \sum_{m=1}^q \left[a_m \cos\left(\frac{2\pi m t}{n}\right) + b_m \sin\left(\frac{2\pi m t}{n}\right) \right] \quad \text{---(i)}$$

$x_t \rightarrow$ RV

$t = 1, 2, \dots, n$ spaced at a part

Total length, $T = n \Delta t$

If n is odd, $q = \frac{(n-1)}{2}$

$q \rightarrow$ no. of harmonics to be considered.

• a_m & $b_m \rightarrow$ coefficients of m^{th} harmonic

$$a_m = \frac{2}{n} \sum_{t=1}^n x_t \cos\left(\frac{2\pi m t}{n}\right) \quad \text{---(ii)}$$

$$b_m = \frac{2}{n} \sum_{t=1}^n x_t \sin\left(\frac{2\pi m t}{n}\right) \quad \text{---(iii)}$$

$c_m =$ amplitude of m^{th} harmonics

$\phi_m =$ phase angle of m^{th} harmonic

$$c_m^2 = a_m^2 + b_m^2 \quad \text{---(iv)}$$

$$\phi_m = \tan^{-1} \left(-\frac{b_m}{a_m} \right) \quad \text{---(v)}$$

• $\frac{c_m^2}{2} \rightarrow$ variance of m^{th} harmonic

$$s_a^2 = \sum_{m=1}^q \left(\frac{c_m^2}{2} \right) \quad \text{---(vi)}$$

$\uparrow$ variance of series.

If n is even,

$$q = \frac{n}{2}$$

• a_m & b_m as (ii) & (iii) for $m \neq q$

• For $m = q$,

$$a_q = \frac{1}{n} \sum_{t=1}^n (-1)^t x_t$$

$$b_q = 0$$

Univariate Seasonal Model

(a) Thomas-Fiering model for generating seasonal flows.

$$X(i,j) = \mu_j + \frac{\rho_j}{\sigma_{j-1}} [X(i,j-1) - \mu_{j-1}] + z_t \sigma_j \sqrt{1 - \rho_j^2} \quad (i)$$

$i = 1, 2, \dots = \text{year}$

$j = 1, 2, \dots, m = \text{season}$

$X(i,j)$ = flow in j^{th} season of i^{th} year

μ_j & σ_j = mean & SD of flows in j^{th} season

ρ_j = correlation coefficient b/w flows of j^{th} & $(j-1)^{\text{th}}$ season.

$z_t : N(0,1)$

• If replaced by sample parameters:

$$Q(i,j) = \bar{Q}_j + b_j [Q(i,j-1) - \bar{Q}_{j-1}] + z_t s_j \sqrt{1 - r_j^2} \quad (ii)$$

b_j = regression coefficient of flows of j^{th} and $(j-1)^{\text{th}}$ season

$$\bar{Q}_j = \frac{1}{n} \sum_{i=1}^n (Q(i,j))$$

$$b_j = \frac{s_j r_j}{s_{j-1}} \quad \text{for } j = 2, 3, 4, \dots, m$$

$$s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n [Q(i,j) - \bar{Q}_j]^2}$$

$$r_j = \frac{\frac{1}{n} \sum_{i=1}^n [Q(i,j) - \bar{Q}_j][Q(i,j-1) - \bar{Q}_{j-1}]}{s_j \cdot s_{j-1}} \quad \text{for } j = 2, 3, 4, \dots, m.$$

$$r_1 = \frac{\frac{1}{n} \sum_{i=1}^n [Q(i,1) - \bar{Q}_1][Q(i,m) - \bar{Q}_m]}{s_1 s_m}$$

$$b_1 = \frac{s_m r_1}{s_1}$$

If $Q(i,j) \rightarrow -ve$
(replace by 0 or small values)