Civinnovate

Discover, Learn, and Innovate in Civil Engineering

Time series Analysis

(Elective III)

Lecture notes

TU, IOE
B.E. (Civil), Fourth year/second part

Prepared by Dr. K.N. Dulal
Year: 2014

1. Introduction

1.1 Background

<u>Time series</u>

Time series is a series of values of variable collected over time. Time is an independent variable in the series. 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

<u>Model</u>

A model is a simplified representation of reality. A mathematical model consists of a set of equations describing the system.

<u>Population</u>

A population is a collection of all possible values of a variable.

<u>Sample</u>

A sample is a group of values chosen from a population.

<u>Random variable</u>

A random variable is a variable which is associated with chance or random phenomenon.

<u>Deterministic and stochastic (random, probabilistic) process</u>

If the value of a variable can be predicted uniquely, then the process is known as deterministic process. The equations in deterministic process are based on the physics of the phenomenon. Such process does not involve any randomness. Therefore, a given set of input always generates same set of outputs. For example, the equation for rating curve is $Q = aH^n$ where Q = discharge, H = stage, a and b = constants. This equation represents deterministic process.

If the value of a variable cannot be predicted uniquely, then the process is known as stochastic process. The equations in stochastic process are formulated from the historical properties of data (data-driven model). Such process involves randomness in a variable and the value of random variable is associated with probability distribution. Therefore, a given set of input generates different set of outputs in each trial. For example, discharge at time t is predicted as $Q_t = kQ_{t-1}+e_t$ where e_t is a random component.

Time series model

It is a mathematical model which represents the time series data. It is also known as stochastic model. It has specific mathematical form and a set of parameters. The knowledge of statistics is widely applicable in time series modeling.

Example

a. Let X_t = normally distributed random variable with mean μ and variance σ^2

$X_t = f(e_t)$, where e_t = independent random variable, normally distributed with mean=0 and variance = 1

A form of time series model: $X_t = \mu + \sigma e_t$

Parameters: μ, σ

b. Dependent

$Q_t = aQ_{t-1} + e_t$

Q_t = Discharge at time t, e_t = same meaning as above

Q_t depends on Q_{t-1}

Time series modeling

A technique and procedure for fitting time series data is known time series modeling. The following are the steps to be followed in time series modeling.

- Selecting types of model among several alternative models
- Identifying form of model
- Parameter estimation
- Testing the goodness of fit

Realization

Realization is an outcome of an experiment in which the process is observed. It is a single time series for a stochastic process.

Ensemble

Ensemble is a collection of all possible realizations of a process.

Ergodic property

Time average properties are based on single realization, whereas ensemble average properties are based on several realizations. If time average properties and ensemble average are same, the time series is said to be ergodic.

Normalization

Normalization is the process of converting time series data into normal distribution

1.2 Application of time series analysis in Hydrology

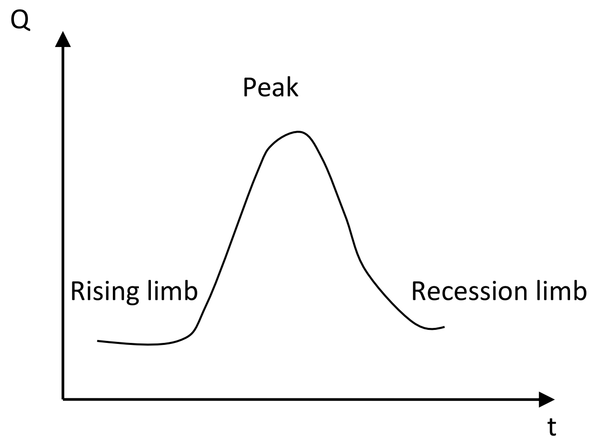
- To generate synthetic hydrologic records
- To forecast hydrologic events
- To fill in missing data
- To extend the records
- To detect trend and shift in records

1.3 Physical basis of time series modeling in Hydrology

Deterministic models are based on the physics of the system, but stochastic or time series models are based on the statistical characteristics of historical data. As data is considered to contain the generating mechanism of the process, the time series modeling represents some physical basis in an implicit way. The parameters of time series model can be considered to be analogous to physical based parameters by comparing the form of equations of two models.

Examples

a. Equation for Recession limb of hydrograph



Equation for recession curve

$Q_t = Q_0 e^{-Kt}$

Q_0: initial discharge, Q_t: discharge at a time interval of t days, K : recession constant

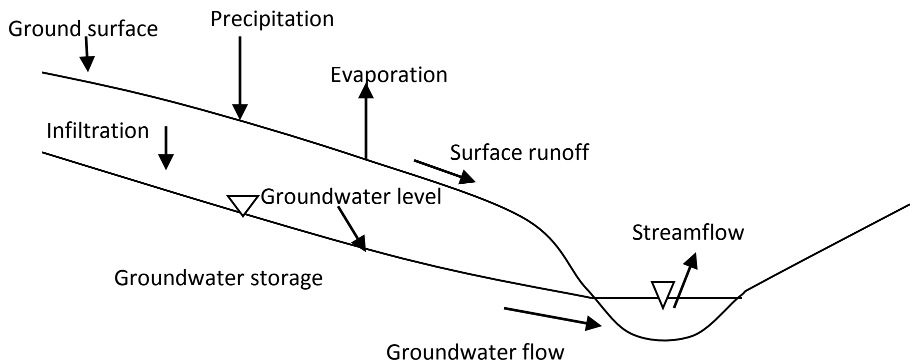
Analogous equation with time series is represented as

$Q_i - \bar{Q} = \phi(Q_{i-1} - \bar{Q}) + e_i$

Q_i =discharge at any time i, $\bar{Q}$= mean discharge, e_i = random component

ϕ = constant, which is equivalent to $\exp(-K)$.

b. Precipitation-runoff process



Conceptual representation of Precipitation-runoff process

Consider the watershed system as shown in the figure. Consider annual values of variables.

Precipitation $=X_t$

Evaporation = b X_t

Infiltration = a X_t

Surface runoff = Precipitation- Evaporation- Infiltration = (1-a-b) X_t = dX_t

Groundwater storage = G_{t-1}

Groundwater flow = c G_{t-1}

Total streamflow (Z_t) = c G_{t-1}+ dX_t (I)

Continuity equation for Groundwater storage is

$G_t = G_{t-1} + a X_t - cG_{t-1} = (1-c) G_{t-1} + a X_t$ (II)

According to Salas and Smith, the combined equation of I and II can be written as

$Z_t = Z_{t-1} + dX_t - [d(1-c)-ac] X_{t-1}$ (III)

This is a conceptual model considering conceptual representation of physical processes. Analogous model in time series representing this form of equation is

$Z_t = \phi_1 Z_{t-1} + e_t - \theta_1 e_{t-1}$

Where e = random variable representing variable X (precipitation).

2. Characteristics of hydrological time series

2.1 Types of hydrological time series

a. Full series and partial duration series

If a time series contains all recorded observations over time, the series is known as full time series, e.g. daily rainfall. If a time series contains selected observations, the series is known as partial duration series, e.g. 1 day maximum rainfall (for each year).

b. Continuous and discrete time series

If time series data are recorded continuously with time, the series is called continuous time series, e.g. water level data recorded continuously. If time series data are recorded at discrete points in time, the series is called discrete time series, e.g. daily rainfall, daily discharge. Such time series is obtained by integrating continuous time series over successive intervals.

c. Univariate (single) and multivariate (multiple) time series

Univariate time series is a time series of one variable at a given site, e.g. rainfall measured at each site, mean rainfall of a basin. Multivariate time series is a set of two or more time series, e.g. discharge data of different stations in a river.

d. Uncorrelated and correlated time series

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), e.g. autocorrelation of streamflow due to storage effects causing the water to remain in the system.

e. Intermittent and counting time series

If a variable has zero and non-zero value throughout the length of the record, the series is called intermittent time series, e.g. rainfall data. If a time series is derived by counting the number of occurrences of event, it is called counting time series, e.g. count of rainy days in a month.

f. Regularly (evenly) and irregularly (unevenly) spaced time series

If data is recorded on regular time interval, the series is known as regularly spaced, e.g. daily discharge data. If data is not observed on regular time interval, the series is known as irregularly spaced, e.g. water quality measurements

g. Stationary and non-stationary time series

If statistical properties of a time series, such as mean, variance remain constant with time, the series is called stationary series, such as annual rainfall, annual discharge. Otherwise, the time series is non-stationary, e.g. monthly rainfall.

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

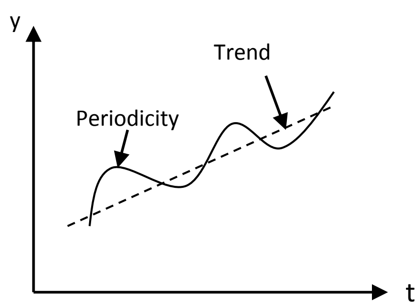
2.2 General properties of hydrological time series

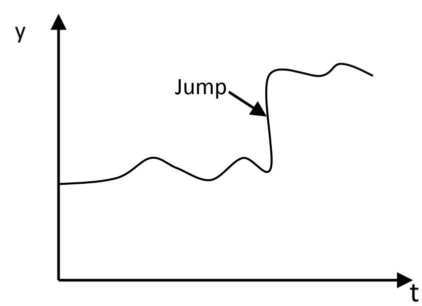
Homogeneity: If data in series belong to one population, the series is homogeneous. The non-homogeneity in a time series is due to natural or manmade causes, such as landuse changes, urbanization, deforestation. Inconsistency in time series occurs due to systematic error in observation. The longer the time series, the greater the chances of non-homogeneity.

Persistent: if value of variable depends on past values, the series is persistent. Time dependence in hydrological series is due to the geophysical process of water storages in basin, their random fluctuation from year to year and periodic fluctuation within a year.

<u>Components of time series</u>
A hydrological time series contains both deterministic and stochastic components. The deterministic components are: trend, jump and periodicity. The random nature of process is the cause of stochastic component.

Trend is a long term movement in time series. The trend in time series exists due to natural or manmade causes, such as landuse changes, urbanization, deforestation. Periodicity is regular fluctuations (cyclic behavoiur) in time series. If the fluctuation in times series is for a period of less than a year, the property is called seasonality. The seasonality is due to the rotation of the earth around the sun. The abrupt change in time series is referred to as jump or shift, which occurs due to sudden change in basin characteristics.





Non-parametric test: no particular probability distribution is considered.
level of significance: probability of being incorrect

Detection of trend

<u>1. Regression test for linear trend</u>
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

- List the data, X_i, i = 1,2,………,n, in date order, and rank (r_i) them in ascending order of magnitude (with smallest allocated rank 1 and largest rank n)
- Compute spearman's rank correlation (R_{sp})

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

  $t = R_{sp}\sqrt{\frac{n-2}{1-R_{sp}^2}}$
- Find tabulated value of t for $\alpha/2$ level of significance (say $\alpha = 5\%$) and n-2 degree of freedom.
- If computed t<tabulated t, then there is no trend.

C. Mann -Kendal test

- Rank (R) the time series data in ascending order of magnitude (with smallest allocated rank 1 and largest rank n).
- Compute S as

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

  Sgn (x) = 1 for x>0, sgn (x) = 0 for x = 0, sgn (x) = -1 for x<0
- Standard deviation of s is (σ) = $\frac{n(n-1)(2n+5)}{18}$
- Z- test statistic is $Z = \frac{|S|}{\sqrt{\sigma}}$
- Find tabulated value of Z for α level of significance (e.g. $\alpha = 5\%, Z = 1.96$).
- If computed Z<tabulated Z, then there is no trend.

Removal of trend

1. Moving average method

Moving average is the average of n successive intervals of data.

Compute moving average and subtract it from original data.

2. Linear trend

Fit a straight line of the form y = at+b. At time t, subtract the values of the line from the original data.

Detection of periodicity

- By visual inspection of time series: noticing effect of periods in the plots
- By performing autocorrelation and spectral analysis (discussed in subsequent chapters)

Detection of jump

- By visual inspection of time series

Split record test for checking stationarity

Testing whether mean and variance of subsets of a series are not significantly different.
Approach

- Divide the record into two non-overlapping sub-sets, j=1,2,…,m and j = m+1, m+2,…..,n.
- Compute mean and variance of each sub-set.
- Examine the variance (V) first using F test
  F=V(a)/V(b) where V(a)>V(b)
- Compare the computed F with the tabulated value of F for m-1 and n-m-1 degree of freedom at $\alpha/2$ level of significance. If the computed F < tabulated value of F, then the variances of subsets can be accepted as estimates of the same population variance.
- Compute pooled estimate of variance (V_p) as
  $V_p = \frac{(m-1)V(a)+(n-m-1)V(b)}{n-2}$
- Test statistics for mean is
  $t = \frac{|M(a)-M(b)|}{\sqrt{\frac{V_p}{n-m}}}$
- Compare the computed t with the tabulated value of t for n-2 degree of freedom at $\alpha/2$ level of significance (two tailed). If the computed t < tabulated value of t, then the mean zero difference between two subsets is accepted.

Double mass curve analysis for checking consistency

- Compute cumulative sum of data (mostly annual data is sufficient) of base stations (A).
- Compute cumulative sum of mean value of neighbouring stations (B).
- Plot A versus B and fit points by a straight line. If there is no break in the line, the data is considered to be consistent.

3. Statistical principles for time series modeling

3.1 Concept of probability distribution

a. Frequency

For discrete random variable, the number of occurrences of a variate is generally called frequency. When the number of occurrences of a variate, or the frequency, is plotted against the variate as the abscissa, a pattern of distribution is obtained. The pattern is called the frequency distribution.

b. Relative frequency function

If the number of observations n_i in interval i is divided by the total number of observations, the result is called relative frequency function $f_s(x)$

$f_s(x) = n_i/n$

This is also called the probability of a function.

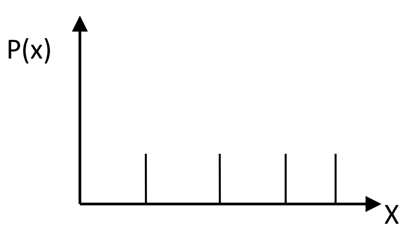
c. Probability mass function (PMF) and Cumulative Distribution Function (CDF) of discrete random variable

PMF: PMF gives the probability associated with all possible values of random variable. For a discrete random variable X, PMF ($P(x_i)$) is given by

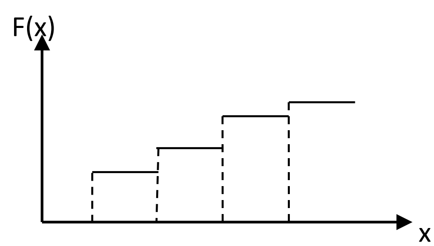
$P(x_i) = P[X=x_i]$

Properties of PMF

- $0 \leq P(x_i) \leq 1$
- $\sum P(x_i) = 1$
- $P[a \leq X \leq b] = \sum_{x_i \leq a}^{x_i \leq b} P(x_i)$



PMF



CDF

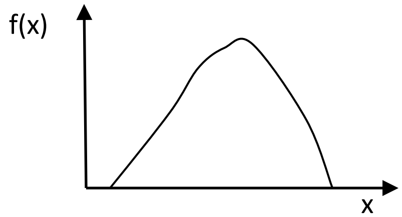
CDF: CDF is the probability of the event that the random variable X is less than or equal to x, for every value x.

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

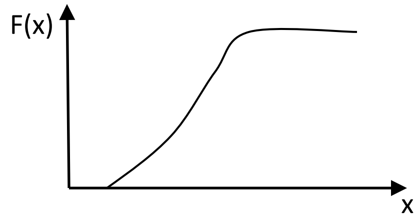
For a discrete random variable, the CDF is found by summing up the probabilities. $P[X=x_i]$ can be computed by

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

d. Probability Density Function (PDF) and Cumulative Distribution Function (CDF) for continuous random variable



PDF



CDF

PDF represents the probability associated with the random variable, which is given by

$$P(a \le X \le b) = \int_a^b f(x)dx$$

conditions

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

X = specified value of variable X, f(x) = frequency of occurrence of x, $P(a \le X \le b)$ is the probability that x lies between a and b.

The CDF for continuous random variable is given by

$$F(x) = P[X \le x] = \int_{-\infty}^{x} f(x)dx$$

Knowing CDF, $P(a \le X \le b)$ can be computed as

$$P(a \le X \le b) = F(b) - F(a)$$

Relationship between PDF and CDF is

$f(x) = \frac{dF(x)}{dx}$ where f(x) = PDF and F(x) = CDF

The magnitude of a variable corresponding to a probability of exceedence is called quantiles.

e. Joint probability distribution

(Bivariate or multivariate)

Joint probability distribution represents probability distribution of more than one random variables.

For two discrete random variable X and Y, joint probability function is

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

Conditions

$$P(x, y) \ge 0$$

$$\sum_x \sum_y p(x, y) = 1$$

Joint CDF is

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

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 random variable X and Y, Joint pdf is

$$P(x_1 \le X \le x_2,\ y_1 \le Y \le y_2\,) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y)dxdy$$

Conditions

$f(x, y) \ge 0$ for $-\infty < x < \infty, -\infty < y < \infty$

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

Joint CDF for two continuous random variable

$$F(x, y) = P[-\infty < X \le x, -\infty < Y \le y\,] = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)dsdt$$

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).
Relationship between PDF and CDF

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

f. Marginal distribution

Given the joint probability distribution of two random variables, 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 random variables X and Y

Discrete case

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

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

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 $-\infty < x < \infty$

g. Conditional distributions

f(x,y) = joint pdf of two random variables 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)

h. Derived distributions

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

Given: f(x) as pdf of X

To compute pdf of Y, say g(y)

For small element in pdf

Area of pdf of X = area of pdf of Y

f(x)dx=g(y)dy

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

Since g(y) cannot be negative

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

3.2 Statistical properties

Sample statistics

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

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

Coefficient of variation: $CV = \frac{\sigma}{\bar{x}}$

Coefficient of skewness: $C_s = \frac{n\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n-1)(n-2)s^3}$

Population statistics

Mean = μ = E(x)= $\int_{-\infty}^{\infty} xf(x)dx$ = first moment about origin (expectation of X)

Median = x such that $\int_{0}^{x} f(x)dx$ =0.5

Variance = Var(X) = $E[(x-\mu)^2] = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx$ = second moment about mean

Coefficient of variation: $CV = \frac{\sigma}{\mu}$

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

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

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

Covariance between two random variables X and Y

Covariance is the linear relationship between two random variables.

Cov(X,Y)=$E\left[(X-\mu_x)\left(Y-\mu_y\right)\right]$

$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)\left(y-\mu_y\right) f(x,y)dxdy$

μ_x and μ_y = mean of variables X and Y

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

Cov (x,y) is also given as

Cov(X,Y)=E[XY]-E[X]E[Y]

3.3 Chebyshev's inequality

Knowing mean and standard deviation of a variable, approximate probability can be obtained by using Chebyshev's inequality. It states that for any k>1

$P[(\mu - k\sigma) \le X \le (\mu + k\sigma)] \ge \left(1 - \frac{1}{k^2}\right)$ where μ = mean and σ = standard deviation of a variable

Proof:

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

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

($(x-\mu)^2 f(x)dx$: 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 $\mu - k\sigma$ and $\mu + k\sigma$ to ∞

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

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

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

$\frac{1}{k^2} \geq P[|x-\mu| \geq k\sigma\,]$

$P[|x-\mu| \leq k\sigma\,] \geq 1 - \frac{1}{k^2}$

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

3.4 Moment generating function

- Alternative method for finding moment

If X is a discrete random variable, its moment -generating function, when it exits is given by

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

If x is a continuous random variable, then

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

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

$\mu_r' = \frac{d^r M_X(t)}{dt^r}$ at t = 0

Proof:

Expanding e^{tx} by Macalaurin's series

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

$= \sum p(x) + t\sum xp(x) + \frac{t^2}{2!}\sum x^2 p(x) + \cdots. + \frac{t^r}{r!}\sum x^r p(x) + \cdots$

$= 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \cdots + \frac{t^r}{r!}\mu_{r+\cdots}'$

$\frac{d^r M_X(t)}{dt^r} = \mu_r' + terms\ of\ t$

$\mu_r' = \frac{d^r M_X(t)}{dt^r}$ at t=0

3.5 Normal (Gaussian) distribution

Normal (Gaussian) distribution is widely used distribution to describe hydrological variables of large intervals such as annual runoff, annual rainfall. Random errors associated with any hydrological measurements can be approximated by normal distribution. This type of distribution is also useful in hypothesis testing.

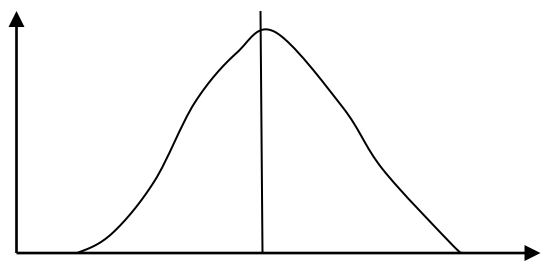
The PDF of normal distribution is given by

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

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

μ =location parameter

σ =scale parameter



Properties

Bell shaped

Symmetric about mean

Mean, median and mode same

All odd central moment = 0

For N(μ, σ^2)

$P[X \leq x] = F(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du$. This expression cannot be evaluated analytically. The alternative way is to transform N(μ, σ^2) 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

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

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

$= P[z_1 \leq Z \leq z_2]$

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

F(z) is found from tables.

3.6 Central Limit theorem

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

No need to know individual distribution of contribution of variables
For large no. of variables, even if they are not independent, the summation tends to follow normal distribution.
For independent random variables, the convergence to normal distribution is faster.

Hydrological variables: annual runoff, annual rainfall
Outcome = no. of additive causative factor: normally distributed

3.7 Estimating the parameters of probability distribution

a. Method of moments (Karl Pearson method)
The expected value of random variable is population moment, whereas the sample moment is computed from the given time series data. The 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)
According to this method, the best value of the parameters 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 $\alpha, \beta, \dots$
$(x_1, x_2, \dots\dots x_n)$: observed sample
Likelihood function = $f(x_1, \alpha, \beta, \dots) f(x_2, \alpha, \beta, \dots) \dots . f(x_n, \alpha, \beta, \dots)$
Or, $L(x, \alpha, \beta, \dots) = \prod_{i=1}^{n} f(x_i, \alpha, \beta, \dots)$
Find the values of $\alpha, \beta, \dots$ which maximizes the likelihood function.
It is convenient to work with log-likelihood.
$ln[L(x, \alpha, \beta, \dots)] = ln[\prod_{i=1}^{n} f(x_i, \alpha, \beta, \dots)] = \sum_{i=1}^{n} ln\,[f(x_i, \alpha, \beta, \dots)]$

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

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

3.8 Selection of distribution

a. Probability plotting
Plot between random variable and probability of exceedence

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 (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 ($X \geq x_m$) = m/n+1

where m = rank, n = total number of values

b. Testing goodness of fit

<u>χ^2 test</u>

Test statistics is

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

Where n_i = no. of observations in a class interval

e_i = expected no. of observations in a class interval

No. of degree of freedom = k-h-1 where h = no. of parameters to be estimated and k=no. of class intervals

Divide a random variables 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_0^2 for k-h-1 degree of freedom and α level of significance from table.(α = 5% or 10%)

If computed χ^2 is less than critical value of χ_0^2, the assumed distribution is good fit.

<u>Smirnov-Kolmogorov test</u>

Compute cumulative probability $P(x_m)$ by Weibull's formula ($P(x_m)$=1-exceedence prob).

Compute theoretical cumulative probability $F(x_m)$ for the assumed distribution.

Compute the Smirnov-Kolmogorov test statistic

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

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.

3.9 Some commonly used probability distributions

Discrete

a. Binomial

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

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

Where p = probability of occurrence, q = 1-p

b. Geometric

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

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,

$P_{r,n} = C(n,r)P^r q^{n-r} = \frac{\lambda^r e^{-\lambda}}{r!}$ where $np \to \lambda$

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

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

Parameters: μ_y, σ_y

4. Exponential

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

Parameters: 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

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

Parameters: λ, β, ϵ

8. Extreme value I (Gumbel)

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

Parameters: α, u

4. Autocorrelation analysis

4.1 Correlation and regression

Correlation

Correlation determines the degree of association between variables. The association between two variables is called cross-correlation. Correlation measures the degree of linear dependence between two variables.

Population correlation coefficient $(r) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$

Sample correlation coefficient $(r) = \frac{S_{x,y}}{s_x\, s_y}$ where $S_{x,y}$ = sample covariance, s_x, s_y =sample standard deviation of X and Y

If (x_i, y_i) is the ith 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 positive correlation. Otherwise, the correlation is negative.

Regression

Regression is the relationship between 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 variables.

Autocorrelation

Autocorrelation is the correlation of a variable with its past and future values (correlation in itself). Autocorrelation coefficient or Serial correlation coefficient is given by

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

Or, $r_k = \frac{\sum_{t=1}^{n-k}(x_t-\overline{x_t})(x_{t+k}-\overline{x_{t+k}})}{\sqrt{\sum_{t=1}^{n-k}(x_t-\overline{x_t})^2 \sum_{t=1}^{n-k}(x_{t+k}-\overline{x_{t+k}})^2}}$

$\overline{x_t} = \frac{1}{n-k}\sum_{i=1}^{n-k} x_t$ and $\overline{x_{t+k}} = \frac{1}{n-k}\sum_{i=1}^{n-k} x_{t+k}$

If n>>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.

4.2 Classification of time series considering deterministic and stochastic process

1. Deterministic series

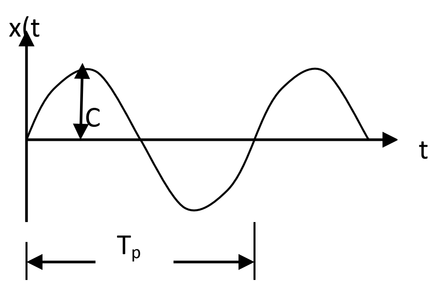
I. Periodic series

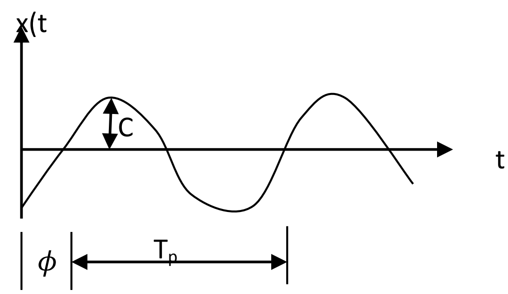
a. Sinusoidal series

Sinusoidal data can be represented by

$x(t) = CSin(2\pi f_0 t + \phi)$

Where x(t) = instantaneous value at t, C = amplitude, f_0 = frequency, ϕ = initial phase angle with respect to time origin. Example: voltage output of generator





Sinusoidal series (without phase angle) sinusoidal series (with phase angle)

Period (T_p): time for one full fluctuation

f_0 = frequency = no. of cycles per unit time

$T_p = 1/ f_0$

b. Complex periodic series

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

$x(t)=x(t+nT_p)$; n = 1,2,3,.. where x(t) = instantaneous value at t and T_p = period

Such series can be represented by Fourier series

$x(t) = \bar{x} + \sum_{n=1}^{\infty} c_n Sin(2n\pi f_1 t + \phi_n)$

$\bar{x}$ = mean of series

f_1 = fundamental frequency

C_n and ϕ_n =coeff.

Complex periodic series = static component ($\bar{x}$)+infinite no. of sinusoidal components (harmonics) with C_n =amplitude and ϕ_n = phase angle

Example: daily and monthly rainfall, discharge, groundwater levels, water quality

II. Non-periodic

a. Almost periodic series

In almost periodic series, there is no common factor for frequency and no period.

Such series can be represented by

$x(t) = \bar{x} + c_1 Cos(2\pi f_1 t + \phi_1) + c_2 Cos(2\pi f_2 t + \phi_2) + \cdots \ldots$

This type of series is found where two unrelated periodic phenomena are mixed. E.g. tidal levels along coast (earth's rotation, moon's rotation: unrelated)

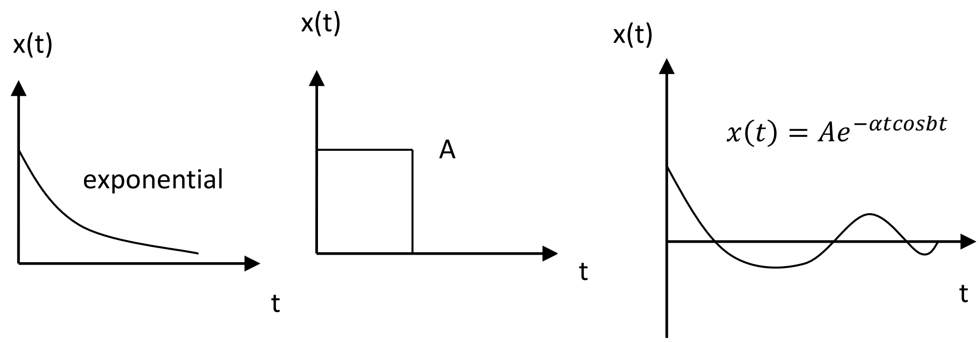
b. Transient series

All non-periodic data other than almost periodic is called transient series. Such series is represented by time varying function.

$x(t) = Ae^{-\alpha t}$: Exponential

X(t) = A : Constant

$x(t) = Ae^{-\alpha t cosbt}$



Common types of transients: trend and jump

<u>2. Stochastic series</u>

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.

In Hydrology, the mathematical techniques for analyzing stationary series are well developed. After identifying and removing deterministic component, the resulting series is considered to be approximately stationary.

Components of time series

Observed time series = stochastic component+ trend component+ periodic component + catastrophic event

Stochastic component: dependent and independent (pure random component)

4.3 Autocorrelation analysis for the investigation of independent process

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

Test of significance for Autocorrelation coefficient (r_k)
In a random, normally distributed series of length n, rk 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 are

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

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

<u>Correlogram of different types of series</u>
Complex periodic with a period T_p: The correlogram is periodic with same T_p and unit amplitude.
Linear trend: The correlogram is a horizontal line.
Combination of periodic sine component and independent random component: The correlogram is periodic with same frequency as sine component, but the amplitude is not unity.
Combination of complex periodic and independent random component: The correlogram is periodic with frequency equal to fundamental frequency of complex periodic series.

5. Time Series Models

5.1 Autoregressive (AR) model

5.1.1 Basic equations of AR model

In AR model, the value of variable at time t is equal to weighed sum of past values plus random component.

$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots \ldots \ldots + \alpha_m X_{t-m} + \cdots \ldots \ldots + e_t$ (I)

Where $\alpha_1, \alpha_2, \ldots \ldots, \alpha_m$ = 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).

The alternative form of above equation is

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

Where μ =expected value of the process

Yule-Walker equations

Multiplying both sides of eq. II by $(x_{t-k} - \mu)$ and taking expectations

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

Here, $E\left[\left(X_{t-j} - \mu\right)(X_{t-k} - \mu)\right]$ =Cov.(X_{t-j}, X_{t-k}) =$\gamma(k - j)$ = covariance function and $E[e_t (X_{t-k} - \mu)] = 0$ (independent)

$\gamma(k) = \alpha_1 \gamma(k-1) + \alpha_2 \gamma(k-2) + \cdots \ldots + \alpha_m \gamma(k-m)$ (IV)

Dividing eq. IV by $\gamma(0)$ (variance of X_t)

$\rho_k = \alpha_1 \rho_{k-1} + \alpha_2 \rho_{k-2} + \cdots . + \alpha_m \rho_{k-m}$ (V)

Where ρ = population autocorrelation coefficient

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

$\rho_1 = \alpha_1 + \alpha_2 \rho_1 + \cdots . + \alpha_m \rho_{k-1}$

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

…………………………………………………..

…………………………………………………..

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

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 between variance of variable X and random component e

For k = 0 in eq. (III)

$\gamma(0) = \alpha_1 \gamma(1) + \alpha_2 \gamma(2) + \cdots \ldots + \alpha_m \gamma(m) + E[e_t (X_t - \mu)]$ (VI)

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

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

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

$E[e_t(X_t - \mu)] = 0 + 0 + \cdots. + 0 + \sigma_e^2$ (VII)

From eq. (VI) and (VII)

$\gamma(0) = \alpha_1\gamma(1) + \alpha_2\gamma(2) + \cdots.. + \alpha_m\gamma(m) + \sigma_e^2$ where σ_e^2 =variance of independent process

Dividing eq. IV by $\gamma(0)$

$1 = \alpha_1\rho_1 + \alpha_2\rho_2 + \cdots. + \alpha_m\rho_m + \frac{\sigma_e^2}{\gamma(0)}$ where $\gamma(0) = \sigma_x^2$ = variance of AR(m) process.

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

This is the relationship of variance between AR(m) process and independent process.

5.1.1 AR(1) model: First order AR model (Thomas-Fiering model)

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

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

From eq. (V)

$\rho_k = \alpha_1\rho_{k-1} + \alpha_2\rho_{k-2} + \cdots. + \alpha_m\rho_{k-m}$

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

$\rho_k = \alpha_1\rho_{k-1}$

$\rho_1 = \alpha_1\rho_0 = \alpha_1$

So, parameter α_1 = first serial correlation coeff.

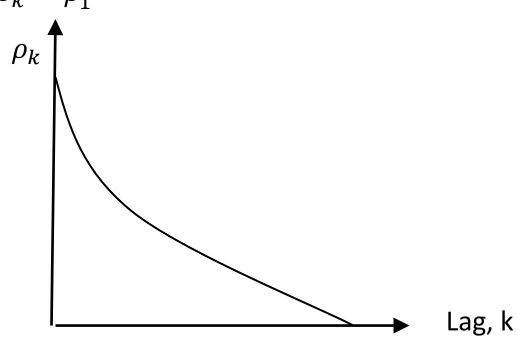
$\rho_2 = \alpha_1\rho_1 = \rho_1\rho_1 = \rho_1^2$

$\rho_k = \rho_1^k$

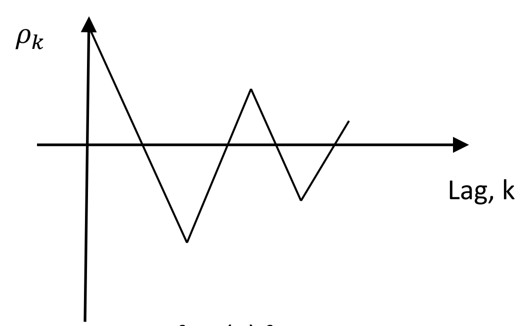
ACF of AR(1) process is completely described by ρ_1. As $\rho_{-k} = \rho_k$ for stationary process, ACF is symmetrical about vertical axis.

General form

$\rho_k = \rho_1^{|k|}$



ACF of AR(1) for +ve ρ_1



ACF of AR(1) for -ve ρ_1

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 part 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 flow series.

Relationship of variance in AR(1) model

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

For m=1

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

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

5.1.2 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.

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

$Q_t = \bar{Q} + r_1(Q_{t-1} - \bar{Q}) + z_t s_x\sqrt{(1-r_1^2)}$

Where Q_t = flow at time t

$\bar{Q}$ = mean of flows

r_1 = first autocorrelation coeff.

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

The annual flow is assumed to be normally distributed.

5.1.3 AR(2) model: Second order AR model

This is also used in Hydrology.

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

$\rho_k = \alpha_1\rho_{k-1} + \alpha_2\rho_{k-2} + \cdots . + \alpha_m\rho_{k-m}$

For m = 2, ACF of AR(1) model is

$\rho_k = \alpha_1\rho_{k-1} + \alpha_2\rho_{k-2}$

For k = 1 and 2

$\rho_1 = \alpha_1 + \alpha_2\rho_1$

$\rho_2 = \alpha_1\rho_1 + \alpha_2$

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

$\alpha_1 = \frac{\rho_1 - \rho_1\rho_2}{1-\rho_1^2}$

$\alpha_2 = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2}$

Using sample serial correlation coefficients r_1 and r_2 instead of ρ_1 and ρ_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

$\alpha_2 + \alpha_1 < 1$

$\alpha_2 - \alpha_1 < 1$

$-1 < \alpha_2 < 1$

The conditions to be satisfied by ρ_1 and ρ_2

$-1 < \rho_1 < 1$

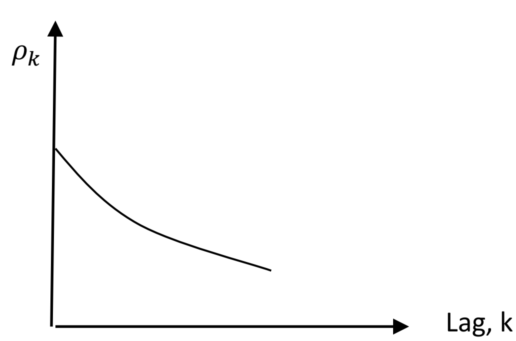
$-1 < \rho_2 < 1$

$\rho_1^2 < \frac{1}{2}(\rho_2 + 1)$

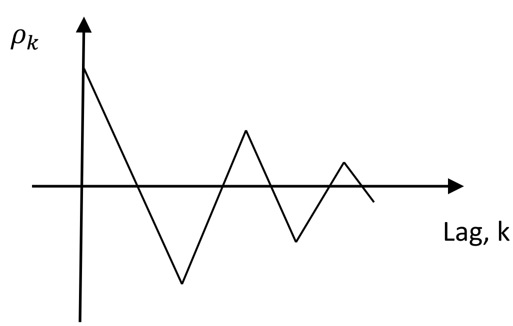
ACF

The ACF of AR(2) depends on the roots of the characteristics function $1 - \alpha_1 x - \alpha_2 x^2 = 0$.

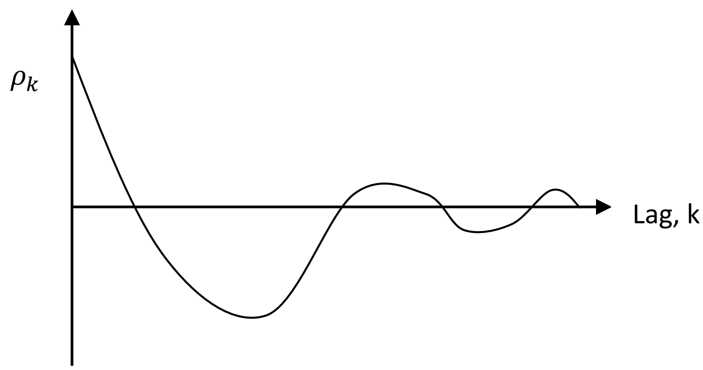
For real roots, the ACF consists of mixture of damped exponentials. When both roots are positive, the ACF remains positive as damps out. When one of the roots is negative, the ACF alternates in sign as it damps out. If the roots are complex, the ACF will be in the form of damped sine wave.



ACF of AR(2) for both +ve roots



ACF of AR(2) for one -ve root



ACF of AR(2) for complex roots

5.1.4 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

$e_t = Z_t \sigma_x \sqrt{(1 - R^2)}$ where R = coeff. Of multiple correlation

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

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

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

Where Q_t = flow at time t

$\bar{Q}$ = mean of flows

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

The annual flow is assumed to be normally distributed.

$\alpha_1 = \frac{r_1 - r_1 r_2}{1 - r_1^2}$

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

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

r_1, r_2 = first and second autocorrelation coeff.

5.2 Moving average (MA) model

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

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

Where $\beta_1, \beta_2, \dots .., \beta_m$ = parameters of model, e = random component with zero mean. Above equation is written as MA(m) where m = order of model (no. of lag times).

Taking expectations

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

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

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

The alternative form of above equation is

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

Relationship of variance

Squaring

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

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

Taking expectations, we get

$\sigma_x^2 = \sigma_e^2 + \beta_1^2 \sigma_e^2 + \beta_2^2 \sigma_e^2 + \cdots + \beta_m^2 \sigma_e^2$

$\sigma_x^2 = \sigma_e^2 (1 + \beta_1^2 + \beta_2^2 + \cdots + \beta_m^2)$ (IV)

Multiplying eq. (III) by $(X_{t-k} - \mu)$ on both sides

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

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

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

Taking expectations on both sides ($E[e_t, e_{t-k}]$ = 0: independent, $E[e_{t-lag1}, e_{t-lag2}] = \sigma_e^2$)

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

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

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

$\gamma(k) = \sigma_e^2 (-\beta_k + \beta_1 \beta_{k+1} + \beta_2 \beta_{k+2} + \cdots + \beta_{m-k} \beta_m)$ (V)

Dividing eq. (V) by eq.(IV)

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

$\rho_k = 0$ 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.

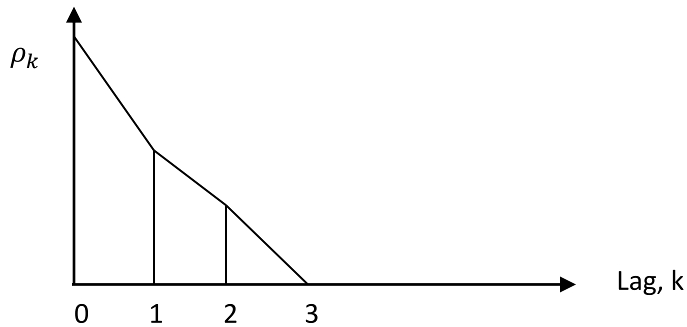
For MA(1)

$\rho_1 = \frac{-\beta_1}{1 + \beta_1^2}$

For MA(2)

$\rho_1 = \frac{-\beta_1 + \beta_1 \beta_2}{1 + \beta_1^2 + \beta_2^2}$

$\rho_2 = \frac{-\beta_2}{1 + \beta_1^2 + \beta_2^2}$



ACF of MA(2)

5.3 Autoregressive moving average (ARMA) model

In MA 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. This is combined Autoregressive and moving average process. ARMA(p,q) model is given as

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

Alternative expression is

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

e = random process with mean zero, $\alpha_1, \alpha_2, \dots .., \alpha_m, \beta_1, \beta_2, \dots .., \beta_m$ = parameters of the model

5.3.1 ARMA(1,1) model

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

Variance relationship

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

ACF

$$\rho_1 = \frac{(1 - \alpha_1 \beta_1)(\alpha_1 - \beta_1)}{(1 + \beta_1^2 - 2\alpha_1 \beta_1)} \qquad \text{(II)}$$

$\rho_k = \alpha_1 \rho_{k-1}$ for $k \geq 2$ (III)

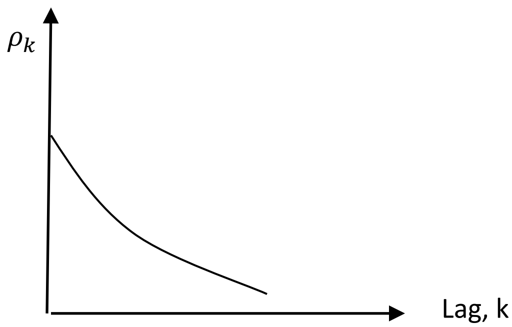
The ACF of ARMA(1,1) decays exponentially from a starting value of ρ_1. If ρ_1 and ρ_2 are known, α_1 can be obtained from eq. (III) and substituting the values on eq. (II) gives the value of β_1.

For ARMA(1,1) to be stationary

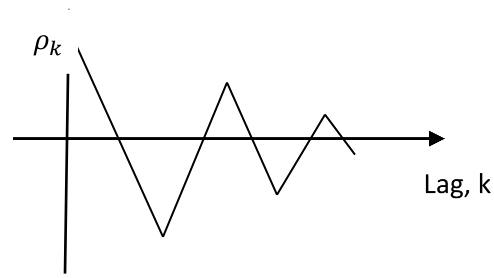
$|\rho_2| < |\rho_1|$

$\rho_2 > \rho_1 (2\rho_1 + 1)$ if $\rho_1 < 0$

$\rho_2 > \rho_1 (2\rho_1 - 1)$ if $\rho_1 > 0$



ACF of ARMA(1,1) for both +ve parameters



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

5.3.2 Applicability of 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

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

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

$Q_t = \bar{Q} + \alpha(Q_{t-1} - \bar{Q}) + T(z_t - bz_{t-1})$

Where Q_t = flow at time t

$\bar{Q}$ = mean of flows

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

The annual flow is assumed to be normally distributed.

$\alpha = \frac{r_2}{r_1}$

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

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

r_1, r_2 = first and second autocorrelation coeff.

5.4 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

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

e.g. 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 dth difference of X_t it stationary. ARMA(p,q) is a special case of ARIMA (p,d,q) with no differencing.

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 (e.g. 95%) of ACF and checking whether all autocorrelation coefficient lies within the band.

c. 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=-Nln(likelihood)+2k where N = sample size, k = number of paraneters.

For AR(m) model,

$AIC(m) = -Nln\left({}_{\sigma_e}'2\right) + 2m$

N = number of observations, m = number of parameters

${}_{\sigma_e}'2$ = maximum likelihood estimate of variance of residuals.

The value of m which gives minimum AIC is selected as order of model.

<u>Determination of parameters of time series models (AR, MA, ARMA, ARIMA model)</u>

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.

b. Method of maximum likelihood

In this method, the likelihood function is first determined. The 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.

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 parameters.

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.
Kurtosis test: The coefficient of kurtosis for normal variable is three.
Chi-squared 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 class interval). If the computed chi-square less than tabulated value, the residuals are considered to be normal.

b. Testing the independence of residuals
I. 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.
II. Porte Manteau lack of fit test: The test static (Q) in this test is given by
$Q = N\sum_{k=1}^{L} r_k^2\,(e)$
Where $r_k^2(e)$ =autocorrelation of residual e
L = number of lags
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.

c. Testing the parsimony of parameters
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 parameter 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=-Nln(likelihood)+2k where N = sample size, k = number of paraneters.
For AR(m) model,
$AIC(m) = -Nln\left({}_{\sigma_e}'2\right) + 2m$
N = number of observations, m = number of parameters
${}_{\sigma_e}'2$ = maximum likelihood estimate of variance of residuals.
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.

6. Seasonal models

6.1 Spectral analysis

Spectral analysis deals with the variation of variance of a stochastic process with frequency. The frequency is used in the harmonic sense.

Fourier series for spectral analysis

The finite Fourier series representation for discrete random variables is written as

$$x_t = \bar{x} + \sum_{m=1}^{q} [a_m Cos(2\pi f_m t) + b_m Sin(2\pi f_m t)] \quad \text{(a)}$$

Where x_t = random variable, t = 1,2,…,n spaced Δt apart

T=total length = $n\Delta t$

If n is odd, n = 2q+1 where q = no. of harmonics to be considered

q=(n-1)/2

f_m = frequency of the mth harmonic = m/n

a_m and b_m = coefficients of the mth harmonic

f_1 =1/n = fundamental frequency

With the value of f_m, Fourier series is

$$x_t = \bar{x} + \sum_{m=1}^{q} \left[a_m Cos\left(\frac{2\pi mt}{n}\right) + b_m Sin\left(\frac{2\pi mt}{n}\right)\right] \quad \text{(b)}$$

The values of coefficients are given by

$$a_m = \frac{2}{n}\sum_{t=1}^{n} x_t Cos\left(\frac{2\pi mt}{n}\right) \quad \text{(c)}$$

$$b_m = \frac{2}{n}\sum_{t=1}^{n} x_t Sin\left(\frac{2\pi mt}{n}\right) \quad \text{(d)}$$

Fourier series can also be written as

$$x_t = \bar{x} + \sum_{m=1}^{q} c_m Sin\left(\frac{2\pi mt}{n} + \phi_m\right) \quad \text{(e)}$$

c_m = amplitude of the mth harmonic

ϕ_m = phase angle of the mth 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 $c_m^2/2$ versus the frequency fm 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

$$s_x^2 = \sum_{m=1}^{q}(c_m^2/2) \quad \text{(h)}$$

Variance of each harmonic = $c_m^2/2$

Sum of variances of all the harmonics = variance of the series

The line spectrum consists of q=(n-1)/2 values and displays the variance explained by each harmomic.

If n is even, then q=n/2 and eq. (c) and (d) are used to compute coefficients for m = 1,2,….,q-1 while a_q and b_q are given by

$a_q = \frac{1}{n}\sum_{t=1}^{n}(-1)^t x_t$ and $b_q = 0$

6.2 Univariate seasonal model

6.2.1 Thomas-Fiering model for generating seasonal flows

Thomas-Fiering model is an AR(1) model which can be used to forecast/generate seasonal flow data such monthly flows or biannual flows. The flow at time t is a weighted sum of immediate past flow plus random component. The form of equation in Thomas-Fiering model is

$$X(i,j) = \mu_j + \frac{\rho_j \sigma_j}{\sigma_{j-1}}\left[X(i,j-1) - \mu_{j-1}\right] + z_t \sigma_j \sqrt{\left(1-\rho_j^2\right)}$$

Where i = 1, 2,…. = year

j = 1,2,….,m =season of a year

X(i,j) =flow in the jth season of ith year

μ_j and σ_j = mean and standard deviation of flows in jth season

ρ_j = correlation coeff. between flows of jth and (j-1)th season

z_t: N(0,1)

In practice, the population parameters of the equation are replaced by sample estimates.

$$Q(i,j) = \overline{Q_J} + b_j\left[Q(i,j-1) - \overline{Q_{J-1}}\right] + z_t s_j \sqrt{\left(1-r_j^2\right)}$$

Where i = 1, 2,…. = year

j = 1,2,….,m =season of a year

Q(i,j) =flow in the jth season of ith year

$\overline{Q_J}$ and s_j = mean and standard deviation of flows in jth season

r_j = correlation coeff. between flows of jth and (j-1)th season

b_j = regression coefficient of flows of jth and (j-1)th season

z_t: N(0,1)

Computation of statistical parameters for any season j

$$\overline{Q_J} = \frac{1}{n}\sum_{i=1}^{n} Q(i,j)$$

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Q(i,j) - \overline{Q_J}\right)^2}$$

$r_j = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[Q(i,j)-\overline{Q_J}\right]\left[Q(i,j-1)-\overline{Q_{J-1}}\right]}{s_j\, s_{j-1}}$ for j = 2,3,4,…,m

$b_j = \frac{r_j s_j}{s_{j-1}}$ for j = 2,3,4,…,m

For j = 1

$$r_1 = \frac{\frac{1}{n-1}\sum_{i=2}^{n}[Q(i,1)-\overline{Q_1}][Q(i-1,m)-\overline{Q_m}]}{s_1\, s_m}$$

$$b_j = \frac{r_1 s_1}{s_m}$$

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 Quimpo model for generating daily flows

Numbers of parameters in Thomas-Fiering model is 3m, where m = no. of seasons. If this model is applied for daily flow estimation, we have to estimate a large number of parameters (e.g. 365x3), which is not practicable. Therefore, Thomas-Fiering model is practically not applicable for daily flow simulation. Another model developed by Yevjevich and Quimpo 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 the cyclicity, an appropriate stationary model is fitted for residual series.

x(i,j) =observed flow in the jth day of the ith year, j = 1,2,….,365 and i=1,2,…..n

Mean of flows in jth day $\left(\overline{x_j}\right) = \frac{1}{n}\sum_{i=1}^{n} x(i,j)$

Standard deviation of flows in jth day $\left(s_j\right) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x(i,j) - \overline{x_j}\right)^2}$

Mean of daily means $(\bar{X}) = \frac{1}{365}\sum_{j=1}^{365} \overline{x_j}$

Variance of daily mean $(VX) = \frac{1}{365}\sum_{j=1}^{365}\left(\overline{x_j} - \bar{X}\right)^2$

Mean of the daily standard deviation$(\bar{S}) = \frac{1}{365}\sum_{j=1}^{365} s_j$

Variance of the daily standard deviation$(VS) = \frac{1}{365}\sum_{j=1}^{365}\left(s_j - \bar{S}\right)^2$

Number of harmonics required for exact fitting = (n-1)/2 = (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,…..6

Harmonic coefficients to fit for standard deviation: c_k, and d_k, k=1,2,…..6

$$a_k = \frac{2}{365}\sum_{j=1}^{365} \overline{x_j}\, Cos\left(\frac{2\pi jk}{365}\right)$$

$$b_k = \frac{2}{365}\sum_{j=1}^{365} \overline{x_j}\, Sin\left(\frac{2\pi jk}{365}\right)$$

$$c_k = \frac{2}{365}\sum_{j=1}^{365} s_j\, Cos\left(\frac{2\pi jk}{365}\right)$$

$$d_k = \frac{2}{365}\sum_{j=1}^{365} s_j\, Sin\left(\frac{2\pi jk}{365}\right)$$

Harmonic fitting for mean is

$$\overline{x_j'} = \bar{X} + \sum_{k=1}^{qm}\left[a_k Cos\left(\frac{2\pi jk}{365}\right) + b_k Sin\left(\frac{2\pi jk}{365}\right)\right]$$

Harmonic fitting for standard deviation is

$$\overline{s_j'} = \bar{S} + \sum_{k=1}^{qs}\left[c_k Cos\left(\frac{2\pi jk}{365}\right) + d_k Sin\left(\frac{2\pi jk}{365}\right)\right]$$

qm and qs = 6

The residual series is

$$y_t = \frac{x(i,j) - \overline{x_j'}}{\overline{s_j'}}$$ 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

$$y_t = \bar{y} + r_1(y_{t-1} - \bar{y}) + z_t s_y\sqrt{(1 - r_1^2)}$$

z_t: N(0,1)

r_1 = first autocorrelation coeff.

No. of parameters to estimate: 24 harmonic coeff., $\bar{X}, \bar{S}, \bar{y}, s_y$ and r_1

The daily flow is obtained by introducing cyclicity.

$$x(i,j) = y_t s_j + \overline{x_j}$$

7. Generation of random numbers

7.1 Methods for generating uniformly distributed random numbers

a. Mid-square technique

In the beginning of the generation process select an initial value (seed) of 2d digit.
Square the (k-1)th random number of 2d digit 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.
Example
Seed = 4538
$4538^2 = 20593444$
Next random number = 5934
$5934^2=35212356$
Next random number = 2123

b. Mid-product technique

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.
Example
Seeds: 6723 and 4591
6723x4591=30865293
Next random number = 8652
8652x4591=39721332
Next random number = 7213
7213x8652=62406876
Next random number = 4068

c. Mixed congruential method

The recursive formula used in this method is
$u_k = (a u_{k-1}+c) \bmod m$
u_k = current random number
u_{k-1} = previous random number
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 mod 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 mod 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 appears before the seed appears. To make the generated random numbers free from internal serial correlation, the conditions are
$a \approx \sqrt{m}$ and $c \approx m/2$

To generate uniformly distributed random numbers in the range of 0 to 1, divide the sequence by m.

Perform autocorrelation analysis for verifying the generated random numbers to be independent.

7.2 Generation of normal random numbers

As many hydrological variables follow normal distribution, the generation of normal random number 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 random number

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 random numbers u_i and then obtain normal random numbers z_i corresponding 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 the CDF table of normal distribution, value of z corresponding to $F^{-1}(u_i)$ (treated as F(z) in table) can be found by interpolation. This method is called table look up method. The limitation of this method is that it is time consuming.

b. Central limit theorem method

$U_1, U_2, \ldots U_m$ = 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 $X= U_1+U_2+ \ldots+U_m$

According to central limit theorem, for large m, X will follow normal distribution with mean $\mu_x = \mu_{u1} + \mu_{u2} + \cdots + \mu_{um}$ and variance $\sigma_x^2 = \sigma_{u1}^2 + \sigma_{u2}^2 + \cdots . + \sigma_{um}^2$.

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.

$z_j = \frac{\sum_{i=1}^{m}(u_k-0.5m)}{\sqrt{m/12}}$ where k =(j-1)m+i

To generate n normal random numbers, nxm uniformly distributed random numbers are required.

In practice, m=12 or 6 is used.

For m = 6

$z_j = \left(\sum_{i=1}^{6} u_k - 3\right)\sqrt{2}$ where k=6(j-1)+i

For m = 12

$z_j = \left(\sum_{i=1}^{12} u_k - 6\right)$ where k=12(j-1)+i

c. Box-Muller method

Let z_1 and z_2 be two independently and normally distributed random variables 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

$\phi(z_1, z_2) = f(z_1)f(z_2)$

$= \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \frac{1}{\sqrt{2\pi}} e^{-z_2^2/2} = \frac{1}{2\pi} e^{-(z_1^2+z_2^2)/2}$

The outcome is the random vector (z_1, z_2) in 2D, which is represented in polar co-ordinate as

$z_1 = Rcos\theta$ and $z_2 = Rsin\theta$

If U_1 and U_2 are two uniformly distributed random numbers in the range 0 to 1, then

$z_1 = Rcos\theta = [-2\ln(1-U_1)]^{1/2} cos(2\pi U_2)$

$z_2 = Rsin\theta = [-2\ln(1-U_1)]^{1/2}sin(2\pi U_2)$

Here, z_1 and z_2 will be two independently distributed standard normal variables. Hence, a chain of standard normal random numbers z_i can be obtained from a chain of uniformly distributed random numbers u_i in pairs from the following equation.

$$z_i = [-2\ln(1-u_i)]^{1/2}cos(2\pi u_{i+1})$$
$$z_{i+1} = [-2\ln(1-u_i)]^{1/2}sin(2\pi u_{i+1})$$

Civinnovate

Discover, Learn, and Innovate in Civil Engineering