Chapter Four
Stochastic Analysis of Stream flow

Stochastic and empirical models

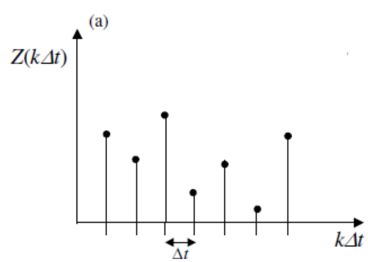
- Time Series Analysis, and Hydrological Forecasting
    - Markov Processes
    - Markov Chains
- Multivariate Regression Analysis and Hydrological Forecasting
- Monte Carlo Simulation in Hydrological Modeling
    - Random numbers and variables generation

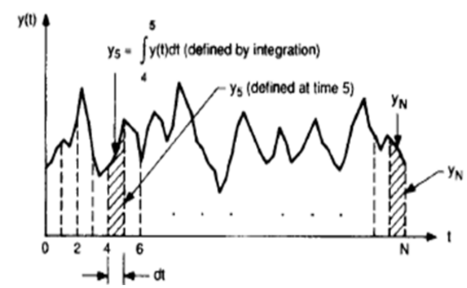
Time-Series Analysis : Definitions

- Many dynamic variables in hydrology are observed at more or less regular time intervals.
  - rainfall,
  - surface water stage or flow
  - groundwater levels.

- Successive observations from a particular monitoring station observed at regular intervals are called **a time series.**

- In the context of stochastic hydrology we should look at a time series as a realization of a random function.
  - **real-valued discrete-time random function or**
  - **a real-valued continuous-time random function**





- Irrespective of this view of reality, hydrologists have been using techniques specially designed to analyze and model hydrological time series, Which collectively know as **"time series analysis".**

Time-Series Analysis: Reasons

- The main reasons for analyzing hydrological time series are:

1. **Characterization:-** to analyze seasonal behavior and trend values .

2. **Prediction and forecasting:-** to estimate the value of the time series at non-observed points in time. This can be
   - a prediction at a time in future (forecasting), or
   - a prediction at a non-observed point in time in the past (fill in gaps in the observed series due to missing values).

3. **Identify and quantify input-response relations:-**
   - Many hydrological variables are the result of a number of natural and man-induced influences. To quantify the effect of an individual influence and to evaluate water management measures, the observed series is split into components which can be attributed to the most important influences

AAiT

Addis Ababa University
Institute of Technology (AAiT)

Time-Series Analysis: Classifications

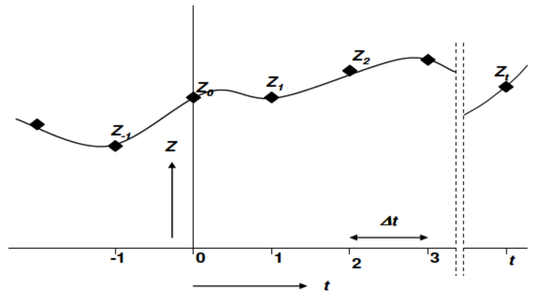
- **By Number variables/stations involves**
  - Single/univariate /time series
  - Multiple time series
- **By the relation with time**
  - Autocorrelation/correlation/correlation with time /dependent
  - Uncorrelation /independent
- **By the time Space**
  - Regular time series
  - irregular time series
- **By the trends and shifts involved**
  - Stationary
  - Non-stationary
- **Others**
  - Intermittent time series
  - Counting time series

AAiT

Addis Ababa University
Institute of Technology (AAiT)

Time-Series Analysis: Terminologies and rules

- **Discrete stationary time series :-**
  - most hydrological variables are continuous in time. However, if we consider the variable Z(t) at regular intervals in time Δt , we can define a discrete time series



$$Z_k = Z(k\Delta t) \quad k = -\infty,\ldots,-1,0,1,\ldots\ldots,\infty .$$

$$Z_t \quad t = -\infty,\ldots,-1,0,1,\ldots\ldots,\infty .$$

- **Moments and Expectation**
  - A single time series is considered to be a stochastic process that can be characterized by its (central) statistical moments. In particular the first and second order moments are relevant: **the mean value, the variance and the autocorrelation function.** For a statistical stationary process the mean value and the variance are

$$\mu_z = E[Z_t] \qquad \sigma_z^2 = VAR[Z_t] = E[\{Z_t - \mu_z\}\{Z_t - \mu_z\}]$$

  - The autocovariance is a measure of the relationship of the process at two points in time. For two points in time k time steps apart **(often k is called the time lag)** , the autocovariance is defined by

$$COV[Z_t, Z_{t+k}] = E[\{Z_t - \mu_z\}\{Z_{t+k} - \mu_z\}] \quad k = -\infty,\ldots,-1,0,1,\ldots,\infty$$

Time-Series Analysis: Terminologies and rules

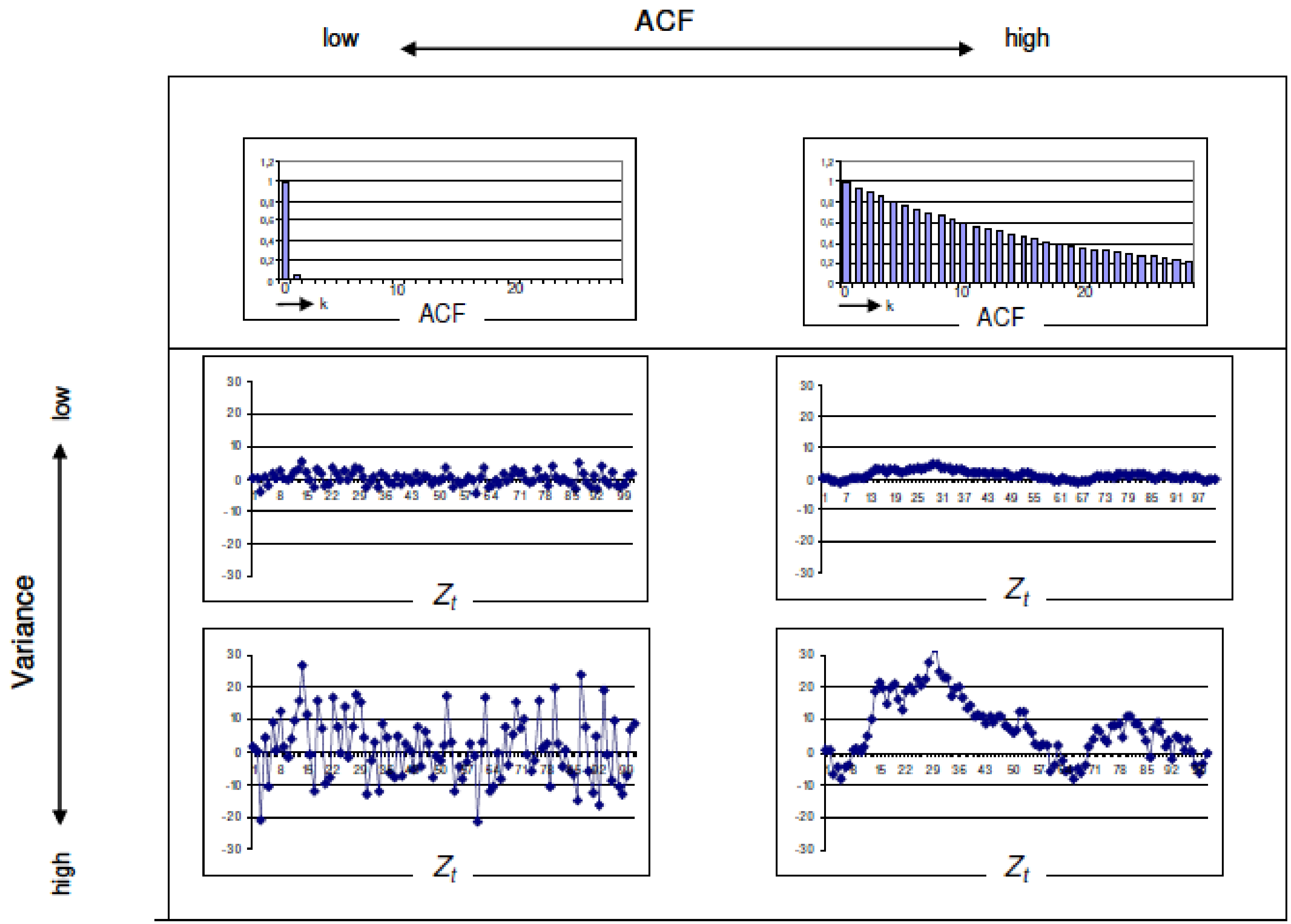
- **Autocorrelation function (ACF):-**
  - In time series analysis we often use the function is called Autocorrelation function (ACF), that defined by

$$\rho_{zz,k} = \frac{COV[Z_t, Z_{t+k}]}{COV[Z_t, Z_t]} = \frac{E[\{Z_t - \mu_z\}\{Z_{t+k} - \mu_z\}]}{\sigma_z^2} \quad k = -\infty, \ldots, -1, 0, 1, \ldots, \infty$$

  - It can be proven that the value of the ACF is always between 1 and -1. A value of 1 or -1 means a perfect correlation, while a value 0 indicates the absence of correlation. From the definition it follows that the ACF is maximum for k=0.
  - The graphical representation of the ACF is called the **autocorrellogram** Because the ACF is symmetrical around k=0, only the right (positive) side is shown.
  - The dynamic behavior of a time series is characterized by its variance and ACF. This is visualized in figure for zero mean time series.

Time-Series Analysis: Terminologies and rules

- **Autocorrelation function (ACF):-**



Time-Series Analysis: Terminologies and rules

- **Discrete white noise process :-**
  - An important class of time series is the discrete white noise process a_t. This is **a zero mean time series with a Gaussian probability distribution and no correlation in time**.

$$E[a_t]=0$$

$$E[a_t a_{t+k}]=\begin{cases}\sigma_a^2 & if \quad k=0 \\ 0 & if \quad k\neq 0\end{cases}$$

  - Because of the absence of correlation in time, the discrete white noise process at time step t does not contain any information about the process at other time steps

- **Rules of calculus**
  - Calculation rules with expectations are summarized as:

$$E[c]=c$$

$$E[cZ]=cE[Z]$$

$$E[Z+X]=E[Z]+E[X]$$

$$E[ZX]=E[Z]E[X]+COV[Z,X]$$

  - Where X and Z are discrete time series and c is a constant.

Time series Modeling

- All the concepts and principles discussed in the previous sections are for the purpose representing the hydrological time series by mathematical models

- A number of stochastic models are presented with their parameter estimation methods and model testing procedures

- The models in this section belongs to
  - a purely random process,
  - an autoregressive (AR) process
  - a moving average (MA) process,
  - an autoregressive moving average (ARMA) process, and
  - an autoregressive integrated moving average (ARIMA)process.

Time-Series Models: Principle

- **Principle of linear univariate time series models :-**
  - The general concept of (linear) time series models is to capture as much information as possible in the model. This information is characterized by **the mean value, the variance and the ACF**. We consider the time series Z_t as a linear function of a white noise process a_t

  - Because the ACF(k) of the white noise process equals zero for any $k \neq 0$, all information of the autocorrelation in Z_t is captured in the time series model.

Time-Series Models: Autoregressive (AR)

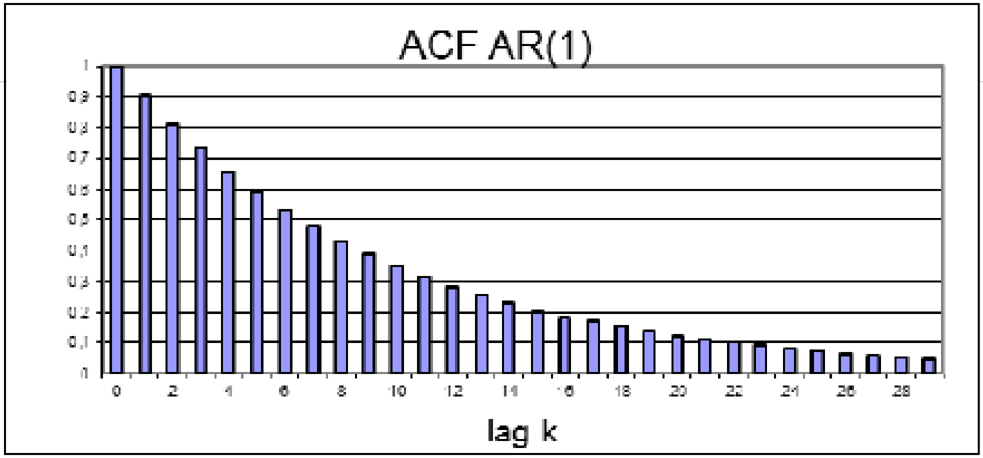
- **AR(1) Model:-**
  - ***Definition:*** A zero mean AR(1) process (Zt) is defined as $Z_t = \phi_1 Z_{t-1} + a_t$
  - ***Parameter Determination*** : since the white noise process is a zero mean, uncorrelated process, therefore the AR(1) process contain two unknowns:
    - the first order auto regressive parameter φ_1 and
    - the variance of the white noise process σ^2
    - These unknowns have to be determined from the characteristics of the time series Z_t, in particular the variance and the ACF

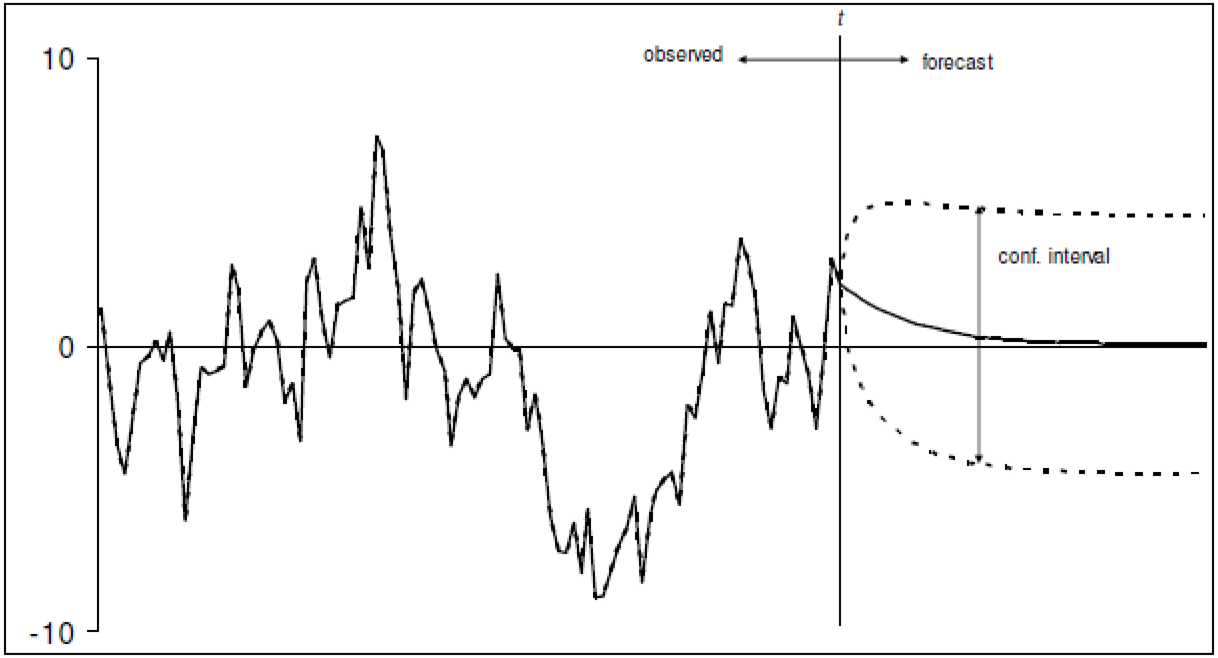
$$\frac{E[Z_{t-1}Z_t]}{\sigma_Z^2} = \phi_1 \frac{E[Z_{t-1}^2]}{\sigma_Z^2} \qquad \phi_1 = \rho_{ZZ,1} \qquad \sigma_Z^2 = \phi_1^2 \sigma_Z^2 + \sigma_a^2 \rightarrow \sigma_Z^2 = \frac{\sigma_a^2}{1-\phi_1^2}$$

  - ***Properties of an AR(1) model*** is **stationary** the absolute value of the model parameter should be smaller than 1. $|\phi_1| < 1$ if this condition is not fulfilled, it follows that the variance of the process Z_t does not exist. In this case, the process Z_t is said to be **non-stationary.**
  - ***Example of an AR(1) process :-*** Let Zt be a zero mean AR(1) process, with $\phi_1 = 0.9$ and $\sigma_a^2 = 1$:

$$Z_t = 0.9 Z_{t-1} + a_t \qquad \rho_{ZZ,k} = 0.9^k \qquad \hat{Z}_{t+\ell|t} = 0.9^\ell \quad \text{and} \quad \sigma^2_{e_{t+\ell|t}} = 1\sum_{i=1}^{\ell} 0.9^{2(i-1)}$$

Time-Series Models: Autoregressive (AR)





Time-Series Models: Autoregressive (AR)

- **AR(p) Model:-**
  - ***Definition:*** $Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + a_t$
  - ***Parameter Determination*** : Similar to the parameter determination of an AR(1) process, the parameters of the AR(p) model can be expressed in terms of the variance and auto correlation of the process

$$\begin{aligned}
Z_{t-1}Z_t &= \phi_1 Z_{t-1} Z_{t-1} + \phi_2 Z_{t-1} Z_{t-2} + \cdots + \phi_p Z_{t-1} Z_{t-p} + Z_{t-1} a_t \\
Z_{t-2}Z_t &= \phi_1 Z_{t-2} Z_{t-1} + \phi_2 Z_{t-2} Z_{t-2} + \cdots + \phi_p Z_{t-2} Z_{t-p} + Z_{t-2} a_t \\
&\vdots \\
Z_{t-p}Z_t &= \phi_1 Z_{t-p} Z_{t-1} + \phi_2 Z_{t-p} Z_{t-2} + \cdots + \phi_p Z_{t-p} Z_{t-p} + Z_{t-p} a_t
\end{aligned}$$

$$\begin{aligned}
\rho_{ZZ,1} &= \phi_1 \qquad + \phi_2\, \rho_{ZZ,1} + \cdots + \phi_p\, \rho_{ZZ,p-1} \\
\rho_{ZZ,2} &= \phi_1\, \rho_{ZZ,1} + \phi_2 \qquad + \cdots + \phi_p\, \rho_{ZZ,p-1} \\
&\vdots \\
\rho_{ZZ,p} &= \phi_1\, \rho_{ZZ,p-1} + \phi_2\, \rho_{ZZ,p-2} + \cdots + \phi_p
\end{aligned}$$

$$\begin{bmatrix} 1 & \rho_{ZZ,1} & \cdots & \rho_{ZZ,p-1} \\ \rho_{ZZ,1} & 1 & \cdots & \rho_{ZZ,p-2} \\ \vdots & & \ddots & \vdots \\ \rho_{ZZ,p-1} & \rho_{ZZ,p-2} & \cdots & 1 \end{bmatrix} \bullet \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} \rho_{ZZ,1} \\ \rho_{ZZ,2} \\ \vdots \\ \rho_{ZZ,p} \end{bmatrix}$$

$$\phi_1 = \frac{\rho_{ZZ,1}(1-\rho_{ZZ,2})}{1-\rho_{ZZ,1}^2} \qquad \phi_2 = \frac{\rho_{ZZ,2}-\rho_{ZZ,1}^2}{1-\rho_{ZZ,1}^2}$$

Time-Series Models: Moving average (MA)

- **MA(1) Model:-**
  - ***Definition:*** A zero mean MA(1) process (Z_t) is defined as $Z_t = a_t - \theta_1 a_{t-1}$
  - ***Parameter Determination*** : since the white noise process is a zero mean, uncorrelated process, therefore the AR(1) process contain two unknowns:
    - the moving average parameter Θ_1 and
    - - the variance of the white noise process σ^2
    - These unknowns have to be determined from the characteristics of the time series Z_t, in particular the variance and the ACF

$$E[Z_t Z_{t-1}] = -\theta_1 \sigma_a^2 \qquad \rho_{zz,1} = \frac{-\theta_1 \sigma_a^2}{\sigma_Z^2} \qquad \rho_{zz,1} = \frac{-\theta_1 \sigma_a^2}{(1+\theta_1^2)\sigma_a^2} = \frac{-\theta_1}{(1+\theta_1^2)} \qquad \sigma_a^2 = \frac{\sigma_Z^2}{1+\theta_1^2}.$$

  - ***Properties of an AR(1) model*** is always **stationary and invertible.** This can be seen by subtracting two successive values of the process Zt.

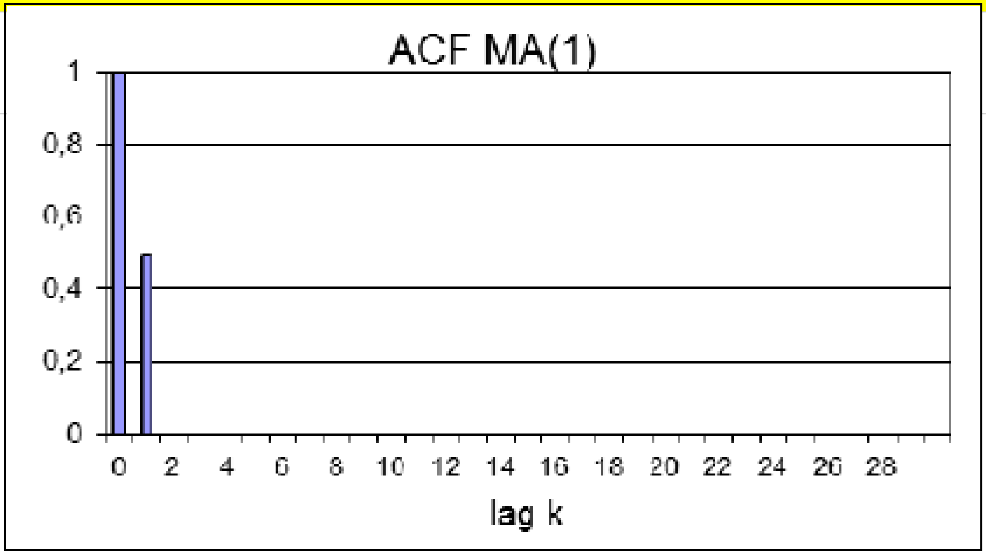
  - ***Example of an AR(1) process :-*** Let Zt be a zero mean MA(1) process, with $\theta_1 = -0.9$ and $\sigma_a^2 = 1$.

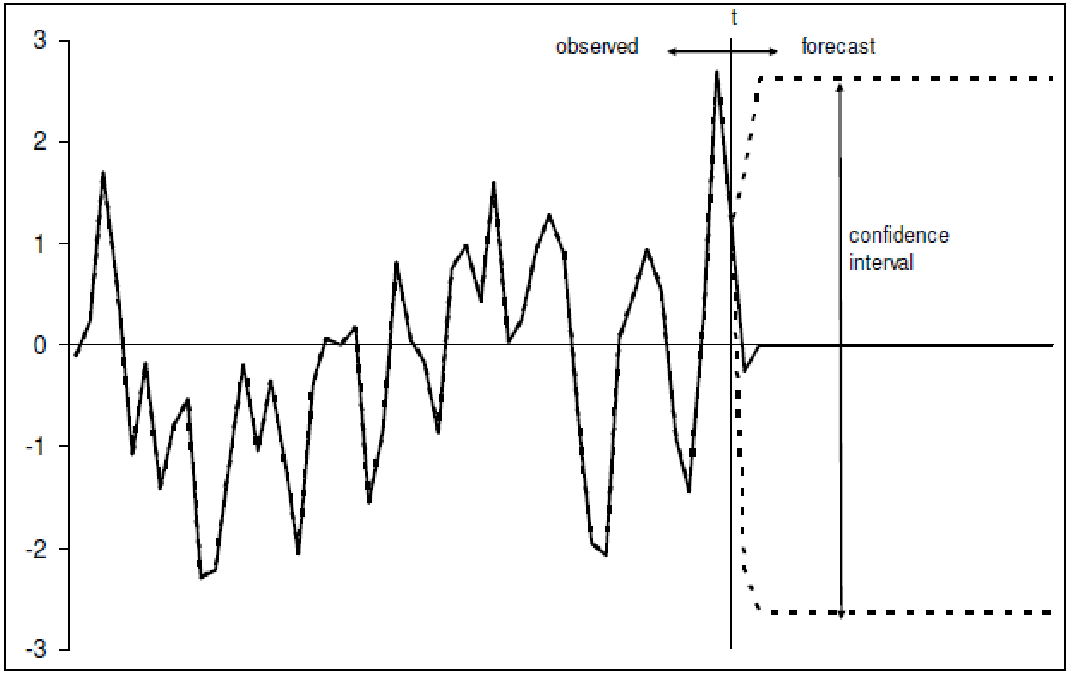
$$Z_t = a_t + 0.9 a_{t-1} \qquad \rho_{zz,0} = 1$$

$$\rho_{zz,1} = 0.497$$

$$\rho_{zz,k} = 0 \qquad |k| > 1$$

Time-Series Models: Moving average (MA)





Time-Series Models: Moving average (MA)

- **MA(p) Model:-**
  - ***Definition:*** $Z_t = a_t - \theta_1 a_{t-1} \cdots - \theta_q a_{t-q}$
  - ***Parameter Determination*** : Similar to the parameter determination of an MR(1) process, the parameters of the MR(p) model can be expressed in terms of the variance and auto correlation of the process

$$E[Z_t Z_{t-k}] = E[(a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q})(a_{t-k} - \theta_1 a_{t-k-1} - \cdots - \theta_q a_{t-k-q})] \quad (6.74)$$

For $k=0$ equation (6.74) yields:

$$E[Z_t Z_t] = \sigma_Z^2 = E[(a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q})^2]$$
$$\sigma_Z^2 = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma_a^2 \quad (6.75)$$

For $k>0$ equation (6.74) gives:

$$E[Z_t Z_{t-k}] = \begin{cases} (-\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q)\sigma_a^2 & \text{for } k = 1,2,\ldots,q \\ 0 & \text{for } k > q \end{cases} \quad (6.76)$$

Combining (6.75) and (6.76) yields the set of equations:

$$\rho_{ZZ,k} = \begin{cases} \dfrac{(-\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q)}{(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)} & \text{for } k = 1,2,\ldots,q \\ 0 & \text{for } k > q \end{cases} \quad (6.77)$$

AAiT

Autoregressive Moving Average Model

We will now discuss models that are combinations of the AR and MA models. These are called autoregressive moving average (ARMA) models. An ARMA(p, q) model is defined as

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}$$

Using the lag operator L, we can write this as

$$\Phi(L) X_t = \theta(L) \varepsilon_t$$

where $\Phi(L)$ and $\theta(L)$ are polynomials of orders p and q, respectively, defined as

$$\Phi(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p$$

$$\theta(L) = 1 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q$$

For stationarity we require that the roots of $\Phi(L) = 0$ lie outside the unit circle. For invertibility of the MA component, we require that the roots of $\theta(L)$ lie outside the unit circle. For instance, for the ARMA(2, 2) process these conditions are given by equations (13.1) and (13.4). The *acvf* and *acf* of an ARMA model are more complicated than for an AR or MA model.

ARMA Model….

We will derive the *acf* for the simplest case: the ARMA(1, 1) process

$$X_t = \alpha_1 X_{t-1} + \varepsilon_t + \beta_1 \varepsilon_{t-1}$$

In terms of the lag operator L this can be written as

$$X_t - \alpha_1 X_{t-1} = \varepsilon_t + \beta_1 \varepsilon_{t-1}$$

or

$$(1 - \alpha_1 L) X_t = (1 + \beta_1 L) \varepsilon_t$$

or

$$X_t = \frac{1 + \beta L}{1 - \alpha L} \varepsilon_t$$

$$= (1 + \beta L)(1 + \alpha L + \alpha^2 L^2 + \cdots) \varepsilon_t$$

$$= [1 + (\alpha + \beta) L + \alpha(\alpha + \beta) L^2 + \alpha^2 (\alpha + \beta) L^3 + \cdots] \varepsilon_t$$

ARMA Model….

Since ε_t is a pure random process with variance σ^2 we get

$$\mathrm{var}(X_t) = [1 + (\alpha+\beta)^2 + \alpha^2(\alpha+\beta)^2 + \cdots]\sigma^2$$

$$= \left(1 + \frac{(\alpha+\beta)^2}{1-\alpha^2}\right)\sigma^2 = \frac{1+\beta^2+2\alpha\beta}{1-\alpha^2}\sigma^2$$

Also

$$\mathrm{cov}(X_t, X_{t-1}) = [(\alpha+\beta) + \alpha(\alpha+\beta)^2 + \alpha^2(\alpha+\beta)^2 + \cdots]\sigma^2$$

$$= \left(\alpha+\beta+\frac{(\alpha+\beta)^2\alpha}{1-\alpha^2}\right)\sigma^2$$

$$= \frac{(\alpha+\beta)(1+\alpha\beta)}{1-\alpha^2}\sigma^2$$

ARMA Model….

$$\rho(1) = \frac{\text{cov}(X_t, X_{t-1})}{\text{var}(X_t)} = \frac{(\alpha + \beta)(1 + \alpha\beta)}{1 + \beta^2 + 2\alpha\beta}$$

Successive values of $\rho(k)$ can be obtained from the recurrence relation $\rho(k) = \alpha\rho(k-1)$ for $k \geq 2$. For the AR(1) process of $\rho(1) = \alpha$, it can be verified that $\rho(1)$ for the ARMA(1, 1) process is $> \alpha$ or $< \alpha$ depending on whether $\beta > 0$ or < 0, respectively.

Autoregressive Integrated Moving Average Model

Autoregressive Integrated Moving Average Process

In practice, most time series are nonstationary. One procedure that is often used to convert a nonstationary series to a stationary series is successive differencing. Let us define the operator $\Delta = 1 - L$, so that $\Delta X_t = X_t - X_{t-1}$, $\Delta^2 X_t = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2})$, and so on. Suppose that $\Delta^d X_t$ is a stationary series that can be represented by an ARMA(p, q) model. Then we say that X_t can be represented by an autoregressive integrated moving average (ARIMA) model ARIMA(p, d, q). The model is called an *integrated* model because the stationary ARMA model that is fitted to the differenced data has to be summed or "integrated" to provide a model for the nonstationary data. Actually, even if there is no need for a moving average component in modeling X_t, the procedure of differencing X_t will produce a moving average process (the Slutsky effect mentioned in our discussion of the MA process).

GAR Models

GAR Models. Skewed hydrologic processes must be transformed into normal processes before AR and ARMA models are applied. However, a direct modeling approach which does not require a transformation may be a viable alternative. The *gamma autoregressive process* (GAR process) offers such an alternative. It is defined as[103]

$$y_t = \phi y_{t-1} + \varepsilon_t \tag{19.3.15}$$

where ϕ is the autoregressive coefficient, ε_t is a random component, and y_t has a three-parameter gamma marginal distribution. The noise ε_t can be obtained as a function of ϕ and the parameters of the gamma distribution λ, α, and β (the location, scale, and shape parameters, respectively) as

$$\varepsilon = \lambda(1 - \phi) + \eta \tag{19.3.16}$$

Multivariate AR and ARMA models

Multivariate AR and Multivariate ARMA Models. Consider a multiple time series $\mathbf{Y}_t$, a column vector with elements $y_t^{(1)}, \ldots, y_v^{(n)}$ in which n is the number of series (number of sites or number of variables) under consideration. The *multivariate AR(1) model* suggested by Matalas[114] is defined as

$$\mathbf{Z}_t = \mathbf{A}_1 \mathbf{Z}_{t-1} + \mathbf{B}\boldsymbol{\varepsilon}_t \qquad (19.3.42)$$

in which $\mathbf{Z}_t = \mathbf{Y}_t - \boldsymbol{\mu}$, $\mathbf{A}_1$ and $\mathbf{B}$ are n- by n-parameter matrices and $\boldsymbol{\mu}$ is a column parameter vector with elements $\mu^{(1)}, \ldots, \mu^{(n)}$. The noise term $\boldsymbol{\varepsilon}_t$ is also a column vector of noises $\varepsilon_t^{(1)}, \ldots, \varepsilon_t^{(n)}$, each with zero mean such that $\mathbf{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^T) = \mathbf{I}$, where T denotes the transpose of the matrix and $\mathbf{I}$ is the identity matrix, and $\mathbf{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_{t-k}^T) = 0$ for $k \neq 0$. In addition, it is assumed that $\boldsymbol{\varepsilon}_t$ is uncorrelated with $\mathbf{Z}_{t-1}$ and $\boldsymbol{\varepsilon}_t$ is normally distributed. Model (19.3.42) has been widely used in operational hydrology. Higher-order multivariate AR models are also available.[134,155] Likewise, the *multivariate ARMA(1, 1) model* is written as[130,155]

$$\mathbf{Z}_t = \mathbf{A}_1 \mathbf{Z}_{t-1} + \mathbf{B}\boldsymbol{\varepsilon}_t - \mathbf{C}_1 \boldsymbol{\varepsilon}_{t-1} \qquad (19.3.43)$$

in which $\mathbf{C}_1$ is an additional n- by n-parameter matrix.

Estimation of AR, MA, and ARMA Models

- Testing Goodness of Fit
  - When an AR, MA, or ARMA model has been fitted to a given time series, it is advisable to check that the model does really give an adequate description of the data
  - There are two criteria often used that reflect the closeness of fit and the number of parameters estimated.
    - Akaike information criterion (AIC), and
    - Schwartz Bayesiancriterion (SBC) or Bayesian information criterion (BIC).

Estimation of AR, MA, and ARMA Models

$$\text{AIC}(p) = n \log \hat{\sigma}_p^2 + 2p$$

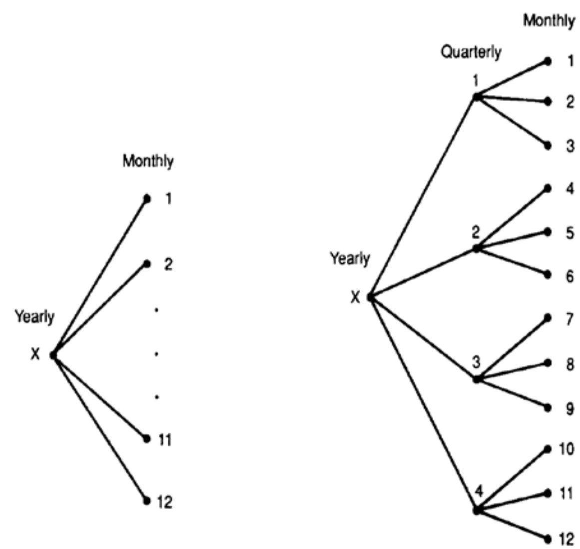
$$\text{BIC}(p) = n \log \hat{\sigma}_p^2 + p \log n$$

Here n is the sample size. If RSS is the residual sum of squares, $\sum \hat{\varepsilon}_t^2$, then $\hat{\sigma}_p^2 =$ RSS$/(n - p)$. If we are considering several ARMA models, we choose the one with the lowest AIC or BIC. (The two criteria can lead to different conclusions.) These goodness of fit criteria are more like the $\bar{R}^2$ or minimum $\hat{\sigma}^2$-type criterion. In addition, we have to check the serial correlation pattern of the residuals—that is, we need to be sure that there is no serial correlation. One can look at the first-order autocorrelation among the residuals. However, as discussed in Chapter 6, one cannot use the Durbin–Watson statistic. With autoregressive models, we have to use Durbin's h-test, or the LM test discussed in Section 6.8.

Disaggregation Models

Disaggregation of Annual to Seasonal Series

Generally, modeling of seasonal hydrologic time series is geared to preserving seasonal statistics only, while statistics at other levels of aggregation, such as annual statistics, may not be preserved. For instance, if the PAR(1) model is used to generate monthly flows, the historical monthly statistics are usually preserved, yet if such generated monthly flows are aggregated to obtain the corresponding annual flows, there is no assurance that the historical annual statistics will be preserved. *Disaggregation models* have been developed for reproducing statistics at more than one level of aggregation. Disaggregation models can be used for both temporal and spatial disaggregation; however, the models in this section are mostly described in terms of temporal disaggregation.



Disaggregation Models

Traditional Valencia-Schaake Model.

$$\mathbf{Y} = \mathbf{AX} + \mathbf{B}\boldsymbol{\varepsilon}$$

where **X** is an n vector of annual values at n sites, **Y** is an $n\omega$ vector of seasonal value in which ω is the number of seasons in the year, **A** and **B** are $n\omega$- by $n\omega$-paramete matrices, and $\boldsymbol{\varepsilon}$ is an $n\omega$ vector of independent standard normal variables. A desir able property of disaggregation models is additivity, i.e., the sum of the seasona values must add up to the annual values. The parameters **A** and **B** may be estimate by[189]

$$\mathbf{A} = \mathbf{S}_{YX}\mathbf{S}_{XX}^{-1}$$

$$\hat{\mathbf{B}}\hat{\mathbf{B}}^{T} = \mathbf{S}_{YY} - \mathbf{A}\mathbf{S}_{XY}$$

Addis Ababa University
Institute of Technology (AAiT)

Disaggregation Models

Markov Chains

The models included in the previous sections are applicable for continuous variables. However, various processes in hydrology can be formulated as discrete-valued processes or continuous processes can be discretized for computational convenience. In these cases, the theory of *Markov chains* may be applicable. Markov chains have been used in hydrology for modeling processes such as precipitation, stream flow, soil moisture, and water storage in reservoirs.

Definition and Properties. Consider $X(t)$ to be a discrete-valued process which started at time 0 and developed through time t. The values that the $X(t)$ process takes on are denoted by x_t, $t = 0, 1, \ldots$. Then

$$P[X(t) = x_t | X(0) = x_0, X(1) = x_1, \ldots, X(t-1) = x_{t-1}] \quad (19.3.68)$$

is the probability of the process being equal to x_t at time t, given its entire history. If the foregoing probability simplifies to

$$P[X(t) = x_t | X(t-1) = x_{t-1}] \quad (19.3.69)$$

Disaggregation Models : Markov Chains

Transition Probability Matrix. A simple Markov chain is defined by its *transition probability matrix* $\mathbf{P}(t)$, which is a square matrix with elements $p_{ij}(t)$ given by

$$p_{ij}(t) = P[X(t) = j | X(t-1) = i] \qquad (19.3.70)$$

for all i, j pairs. Figure 19.3.4 shows that the chain may go from state i at time $t-1$ to states $1, \ldots, r$ at time t, with corresponding transition probabilities $p_{i1}(t), \ldots, p_{ir}(t)$. Then,

$$\sum_{j=1}^{r} p_{ij}(t) = 1 \qquad i = 1, \ldots, r$$

Furthermore, if the transition probability matrix $\mathbf{P}(t)$ does not depend on time, the Markov chain is a *homogeneous chain* or a *stationary chain*. In this case, the notations P and p_{ij} are used. For the rest of this section, a homogeneous Markov chain is assumed.

n-Step Probability. Assume that the chain is now in state i and after n time steps it is in state j. The transition probability from i to j in n steps, denoted by $p_{ij}^{(n)}$, is given by[133]

$$p_{ij}^{(n)} = \sum_{k=1}^{r} p_{ik}^{(n-1)} p_{kj} \qquad n > 1 \qquad (19.3.71)$$

and $p_{ij}^{(1)} = p_{ij}$. Thus, $p_{ij}^{(n)}$, $i, j = 1, \ldots, r$ are elements of the n-step *transition probability matrix* $\mathbf{P}^{(n)}$. It may be shown that $\mathbf{P}^{(n)}$ can be found by multiplying the one-step transition probability matrix $\mathbf{P}$ by itself n times.

Disaggregation Models : Markov Chains

Steady-State Probabilities. The steady-state probability vector $\mathbf{q}^*$ with elements $q_1^*, \ldots, q_r^*$ represents the average fraction of time the chain is in states $1, \ldots, r$, respectively. It can be found by estimating $\mathbf{P}^{(t)}$ for large t until it converges. Also, the elements q_i^*, $i = 1, \ldots, r$ can be found by solving the system of equations

$$q_i^* = \sum_{k=1}^{r} q_k^* p_{ki} \qquad i = 1, \ldots, r \tag{19.3.73a}$$

$$\sum_{i=1}^{r} q_i^* = 1 \tag{19.3.73b}$$

Disaggregation Models : Markov Chains

Example. Assume that daily rainfall for a given site is represented by a simple Markov chain with two states, $j = 1$ for dry and $j = 2$ for wet, and a transition probability matrix **P** with elements $p_{11} = 0.6$, $p_{12} = 0.4$, $p_{21} = 0.3$, and $p_{22} = 0.7$. Assume also that initially the day is dry or $j = 1$ at $t = 0$. This also means that the initial marginal state probability vector is $\mathbf{q}(0) = [1, 0]$. Find: (1) the probability that the next day will be a dry day, (2) the probability that after 2 days, the day will be wet, (3) the probability of states dry and wet after 3 days, and (4) the probability of states dry and wet at any given day (regardless of the initial state). Since initially the day is dry, then $p_{11}^{(1)} = p_{11} = 0.60$ and Eq. (19.3.71) gives $p_{12}^{(2)} = p_{11}p_{12} + p_{12}p_{22} = 0.6 \times 0.4 + 0.4 \times 0.7 = 0.52$. The probabilities of states dry and wet after 3 days are determined by

$$\mathbf{q}(3) = \mathbf{q}(0)\begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}^3 = [1 \quad 0]\begin{bmatrix} 0.444 & 0.556 \\ 0.417 & 0.583 \end{bmatrix} = [0.444 \quad 0.556]$$

Finally, the probabilities of states dry and wet regardless of the initial state (long-run probabilities) are obtained by solving the system of Eqs. (19.3.73). Alternatively, it may be obtained from $\mathbf{P}^{(t)}$ where t is large. For example, for $t = 8$, it may be shown

$$\mathbf{P}^{(8)} = \begin{bmatrix} 0.429 & 0.571 \\ 0.429 & 0.571 \end{bmatrix}$$

Therefore, $\mathbf{q}^* = [0.429 \quad 0.571]$ with approximation to the third decimal figure.

Tools use for Time series Analysis

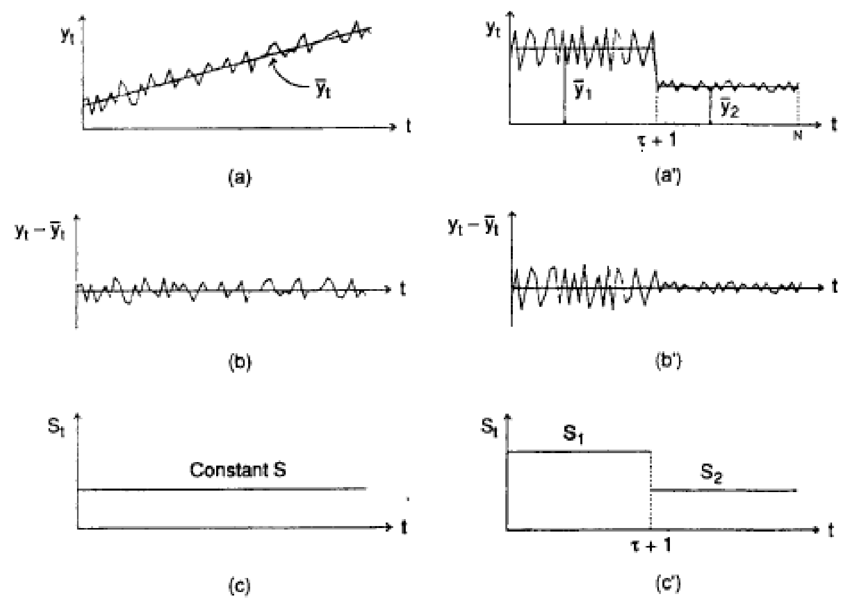
- Excel
- Matlab
- R
- SPSS
- Mintab
- etc

Time Series Structure Tests

Trends and Shifts

Natural and human factors may produce gradual and instantaneous trends or shifts (jumps)

- Examples
  - Effect of a large forest fire in a basin on runoff
  - Large land slides sediment transport on water quality
  - Changes of land use or reservoir construction on stream flows
  - Effects of global warming ar climate changes



Addis Ababa University
Institute of Technology (AAiT)

Test for Trends

Turning Point test

- It is a method, which identify how many turning points are there in a sample data.
- the procedure is
  - Arrange the data in order of their occurrence
  - Apply either of the conditions

    $x_{i-1} < x_i > x_{i+1}$ or $x_{i-1} > x_i < x_{i+1}$
  - Let the total number of turning point be P
  - Expected number of turning points in the series is $E(p) = \frac{2(N-2)}{3}$ where N is the total number of data
  - Variance of P is $Var(p) = \frac{(16N-29)}{90}$
  - Expressing P in standard normal form $Z = \frac{(P - E(P))}{Var(p)^{0.5}}$
  - Test it at 5% level of significance, that is take the value of Z as $\pm$ 1.96 at 5% level of significance
  - If Zcal < Ztab there is no trend

Determination and testing of Trends

Kendal's Rank-correlation Test

- Pick up the first value of the series xi and compare it with the rest of the series x2, x3, ….xn. And find out how many times it is greater than others, assign all the great values with one suffix (P1ex = all expected values of X1)
- Repeat it for all other values
- Find P= P1ex +P2ex +……..Pnex
- Maximum value of P can be $$P_{\max} = \frac{n(n-1)}{2}$$
- $$E(P) = \frac{n(n-1)}{4}$$
- Kendal's Ʈ is computed as $\tau = \left[\left\{\frac{4P}{n(n-1)}\right\} - 1\right]$ E(Ʈ) should be zero
- Variance of Ʈ = $\operatorname{var}(\tau) = \left[\frac{\{2(2n+5)\}}{9n(n-1)}\right]$
- Standard test for Statistics of $Z = \left[\frac{\tau}{Var(\tau)^{1/2}}\right]$
- Test the hypothesis at 5% level of significance of Z, i.e. Z= ±1.96

AAiT

Addis Ababa University
Institute of Technology (AAiT)

Test for Shifts (Jumps)

Test for Shift in the Mean. Suppose that y_t, $t = 1, \ldots, N$ is an annual hydrologic series which is uncorrelated and normally distributed with mean μ and standard deviation σ and N = sample size. The series is divided into two subseries of sizes N_1 and N_2 such that $N_1 + N_2 = N$. The first subseries y_t, $t = 1, 2, \ldots, N_1$, has mean μ_1 and standard deviation σ, and the second subseries y_t, $t = N_1 + 1, N_1 + 2, \ldots, N$, is assumed to have mean μ_2 and standard deviation σ. The simple t test can be used to test the hypothesis $\mu_1 = \mu_2$ when the two subseries have the same standard deviation σ. Rejection of the hypothesis can be considered as a detection of a shift. The test statistic in this case is given by[106,173]

$$T_c = \frac{|\bar{y}_2 - \bar{y}_1|}{S\sqrt{\dfrac{1}{N_1} + \dfrac{1}{N_2}}} \qquad (19.2.26)$$

$$S = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N - 2}} \qquad (19.2.27)$$

where $\bar{y}_1$ and $\bar{y}_2$ and s_1^2 and s_2^2 are the estimated means and variances of the first and the second subseries, respectively. The hypothesis $\mu_1 = \mu_2$ is rejected if $T_c > T_{1-\alpha/2,\nu}$ where $T_{1-\alpha/2,\nu}$ is the $1 - \alpha/2$ quantile of the Student's t distribution with $\nu = N - 2$ degrees of freedom and α is the significance level of the test. Modifications of the test are available when the variances in each group are different[173] and when the data exhibit some significant serial correlation.[106]

Test for Shifts (Jumps)

Mann-Whitney Test for Shift in the Mean. Suppose that y_t, $t = 1, \ldots, N$ is an annual hydrologic series that can be divided into two subseries $y_1, \ldots, y_{N_1}$ and $y_{N_1+1}, \ldots, y_N$ of sizes N_1 and N_2, respectively, such that $N_1 + N_2 = N$. A new series, z_t, $t = 1, \ldots, N$, is defined by rearranging the original data y_t in increasing order of magnitude. One can test the hypothesis that the mean of the first subseries is equal to the mean of the second subseries by using the statistic[173]

$$u_c = \frac{\sum_{t=1}^{N_1} R(y_t) - N_1(N_1 + N_2 + 1)/2}{[N_1 N_2 (N_1 + N_2 + 1)/12]^{1/2}} \qquad (19.2.28)$$

where $R(y_t)$ is the rank of the observation y_t in ordered series z_t. The hypothesis of equal means of the two subseries is rejected if $|u_c| > u_{1-\alpha/2}$, where $u_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution and α is the significance level of the test. Equation (19.2.28) can be modified for the case of groups of values that are tied.[53]

Test of Stationary

One way of describing a stochastic process is to specify the joint distribution of the variables X_t. This is quite complicated and not usually attempted in practice. Instead, what is usually done is that we define the first and second moments of the variables X_t .

These are

1. The mean $\mu(t) = E(X_t)$.
2. The variance $\sigma^2(t) = \text{var}(X_t)$.
3. The autocovariances $\gamma(t_1, t_2) = \text{cov}(X_{t1}, X_{t2})$.

When $t_1 = t_2 = t$, the autocovariance is just $\sigma^2(t)$.