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
- Geostatistics

Multiple Regression Analysis (MRA): Definition

Multiple Regression Analysis (MRA):

- Method for studying the relationship between a dependent variable and two or more independent variables.

Purposes:

- Prediction
- Explanation
- Theory building

MRA: Design Requirements

- One dependent variable (criterion)

- Two or more independent variables (predictor variables).

- Sample size: >= 50 (at least 10 times as many cases as independent variables)

MRA: Assumptions

- **Independence**: the scores of any particular subject are independent of the scores of all other subjects
- **Normality**: in the population, the scores on the dependent variable are normally distributed for each of the possible combinations of the level of the X variables; each of the variables is normally distributed
- **Homoscedasticity**: in the population, the variances of the dependent variable for each of the possible combinations of the levels of the X variables are equal.
- **Linearity**: In the population, the relation between the dependent variable and the independent variable is linear when all the other independent variables are held constant.

Simple vs. Multiple Regression

Simple	Multiple
• One dependent variable Y predicted from one independent variable X	• One dependent variable Y predicted from a set of independent variables (X1, X2 ….Xk)
• One regression coefficient	• One regression coefficient for each independent variable
• **R^2**: proportion of variation in dependent variable Y predictable from X	• **R^2**: proportion of variation in dependent variable Y predictable by set of independent variables (X’s)

Formulation of Multiple Regression

y – response variable

$x_1, x_2, \ldots, x_k$ -- a set of explanatory variables

- *Multiple regression equation* (population)*:*

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots. + \beta_k x_k$$

$\alpha = E(y)$ when $x_1 = x_2 = \ldots = x_k = 0.$

$\beta_1, \beta_2, \ldots, \beta_k$ are called *partial regression coefficients.*

- Controlling for other predictors in model, there is a linear relationship between *E(y)* and x_1 with slope $\beta_{1.}$ i.e., consider case of *k* = 2 explanatory variables,

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

- If x_1 goes up 1 unit with x_2 held constant, the change in *E(y)* is

$$[\alpha + \beta_1(x_1 + 1) + \beta_2 x_2] - [\alpha + \beta_1 x_1 + \beta_2 x_2] = \beta_1$$

Prediction Equation

- With sample data, software finds "least squares" estimates of parameters by minimizing

SSE = sum of squared prediction errors (residuals)

= Σ(observed *y* – predicted *y*)^2

Denote the sample prediction equation by

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Different Ways of Building Regression Models

- **Simultaneous**: all independent variables entered together

- **Stepwise**: independent variables entered according to some order
  - By size or correlation with dependent variable
  - In order of significance

- **Hierarchical**: independent variables entered in stages

Various Significance Tests

- Testing R^2
  - Test R^2 through an F test
  - Test of competing models (difference between R^2) through an F test of difference of R^2s

- Testing b
  - Test of each partial regression coefficient (b) by t-tests
  - Comparison of partial regression coefficients with each other - t-test of difference between **standardized** partial regression coefficients (β)

  - H0: $\beta = 0$,
  - with N-k-1 df

$$t_{\text{observed}} = \frac{b - \beta}{\text{standard error of } b}$$

Geostatistics

- Geostatistics is the part of statistics that is concerned with geo-referenced data, i.e. data that are linked to spatial coordinates.

- Geostatistics is concerned with the unknown value z0 at the non-observed location x0. In particular, geostatistics deals with:
  - **spatial interpolation and mapping:** predicting the value of Z_0 at x_0 as accurately as possible, using the values found at the surrounding locations;
  - **local uncertainty assessment:** estimating the probability distribution of Z_0 at X_0 given the values found at the surrounding locations,
  - **simulation:** generating realizations of the conditional RF $Z(\mathbf{x}|z(\mathbf{x}_i), i = 1,..,4)$ at many non-observed locations xi simultaneously (usually on a lattice or grid) given the values found at the observed locations;

Geostatistics

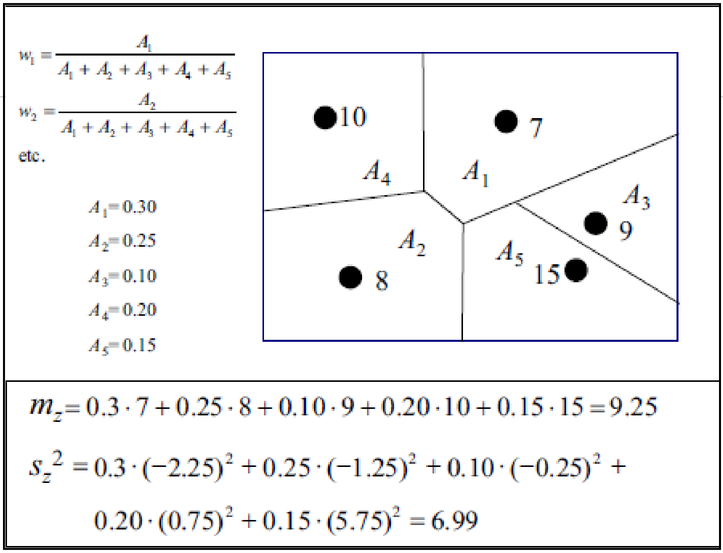
- Some important hydrological problems that **have been tackled using geo-statistics** are among others:
  - spatial interpolation and mapping of rainfall depths and hydraulic heads;
  - estimation and simulation of representative conductivities of model blocks used in groundwater models;
  - simulation of subsoil properties such as rock types, texture classes and geological faces;
  - uncertainty analysis of groundwater flow and -transport through heterogeneous formations (if hydraulic conductivity, dispersivity or chemical properties are spatially varying and largely unknown)

Descriptive spatial statistics

- **Declustering**

$$m_z = \sum_{i=1}^{n} w_i z_i$$

$$s_z^2 = \sum_{i=1}^{n} w_i (z_i - m_z)^2$$



$$w_1 = \frac{A_1}{A_1 + A_2 + A_3 + A_4 + A_5}$$

$$w_2 = \frac{A_2}{A_1 + A_2 + A_3 + A_4 + A_5}$$

etc.

$A_1 = 0.30$

$A_2 = 0.25$

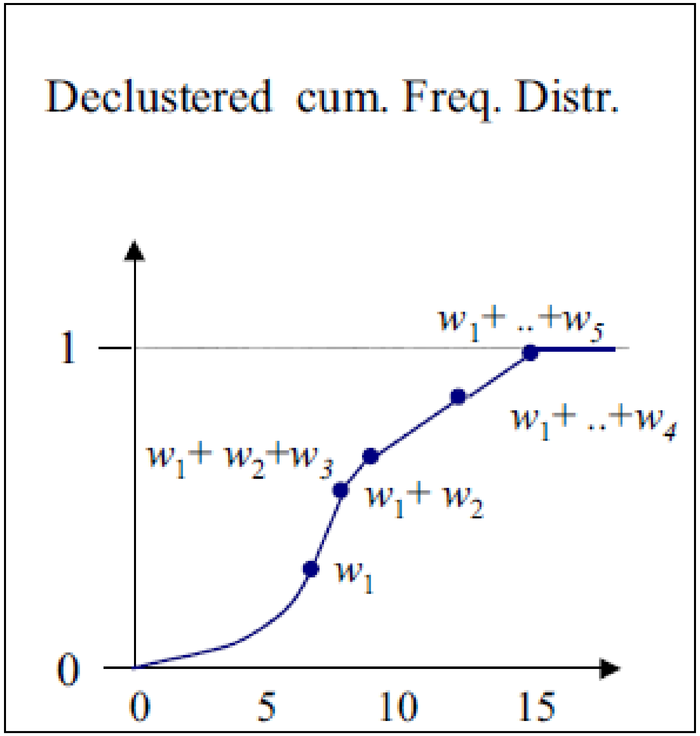
$A_3 = 0.10$

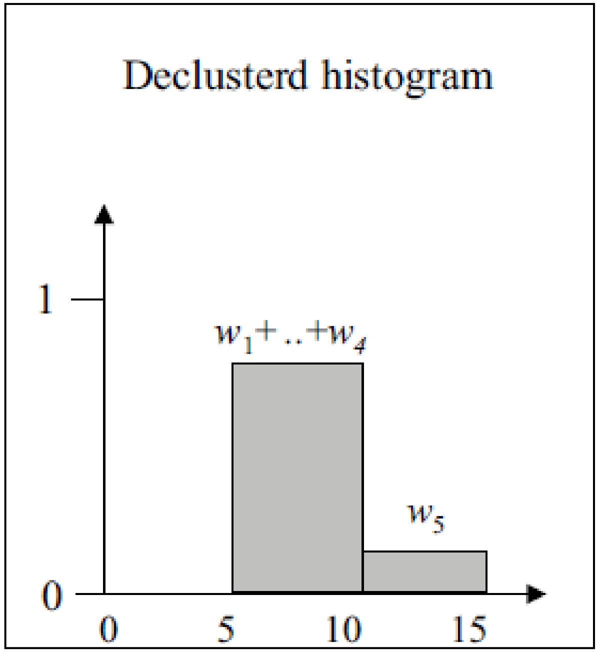
$A_4 = 0.20$

$A_5 = 0.15$

$$m_z = 0.3 \cdot 7 + 0.25 \cdot 8 + 0.10 \cdot 9 + 0.20 \cdot 10 + 0.15 \cdot 15 = 9.25$$

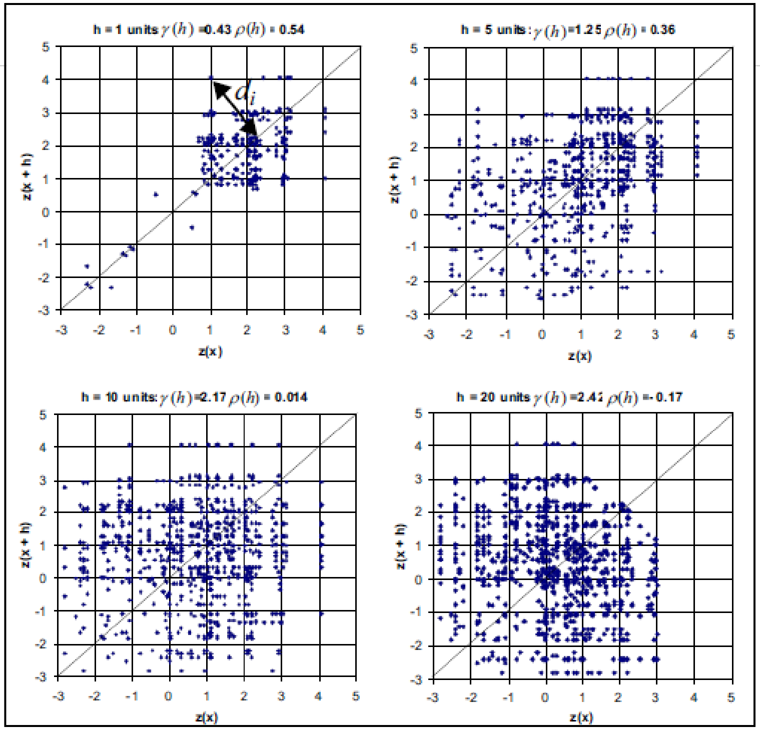
$$s_z^2 = 0.3 \cdot (-2.25)^2 + 0.25 \cdot (-1.25)^2 + 0.10 \cdot (-0.25)^2 + 0.20 \cdot (0.75)^2 + 0.15 \cdot (5.75)^2 = 6.99$$





Descriptive spatial statistics

- **Semivariance and correlation**
  - For each pair of points the distance *di* to the
  - one-to-one line is can be calculated. The semivariance of a given distance is given by (with *n*(*h*) the number of pairs of points that are a distance *h* = |**h**| apart):



$$\hat{\gamma}(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} d_i^2 = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} [z(\mathbf{x}+\mathbf{h}) - z(\mathbf{x})]^2$$

  - and the correlation coefficient:

$$\hat{\rho}(h) = \frac{1}{n(h)} \sum_{i=1}^{n(h)} \frac{z(\mathbf{x}+\mathbf{h})z(\mathbf{x}) - m_{z(\mathbf{x}+\mathbf{h})} m_{z(\mathbf{x})}}{s_{z(\mathbf{x}+\mathbf{h})} s_{z(\mathbf{x})}}$$

Spatial interpolation by kriging

- Kriging is a collection of methods that can be used for spatial interpolation.
- Kriging provides optimal linear predictions at non-observed locations by assuming that the unknown spatial variation of the property is a realization of a random function that has been observed at the data points only.
- **Types of Kriging**
  - Simple kriging
  - Ordinary kriging
  - Block kriging
- **Co-Kriging**
- **Estimating the local conditional distribution**
  - **Multivariate Gaussian random functions**

Estimating the local conditional distribution

- Kriging can also be used to estimate for each non-observed location the probability distribution

- We have different methods that can be used to estimate the conditional pdf
  - **Multivariate Gaussian random functions**
  - **Log-normal kriging**
  - **Kriging normal-score transforms**

Geostatistical simulation

- The third field of application of geostatistics is simulating realizations of the conditional random function

- The aim of geostatistical simulation is to generate in the individual conditional realizations.

- There are two important reasons for individual realizations of the conditional RSF are preferred over the interpolated map that is provided by kriging
  - kriging provides a so called best linear prediction (it produces values that minimize the variance of the prediction error

  - multiple realizations as input for a model can be used for uncertainty analysis and ensemble prediction

More Geostatistics

- More advanced geostatistical methods are concerned with:
  - kriging in case of non-stationary random functions;
  - kriging using auxiliary information;
  - estimating conditional probabilities of non-Gaussian random functions;
  - simulating realizations of non-Gaussian random functions (e.g. positively skewed
  - variables such a s rainfall; categorical data such as texture classes);
  - geostatistical methods applied to space-time random functions;
  - geostatistics applied to random functions defined on other metric spaces such as a sphere or river networks;
  - Bayesian geostatistics, i.e using various forms of a priori information about the random
  - function and formally updating this prior information with observations

Tools use for Time series Analysis

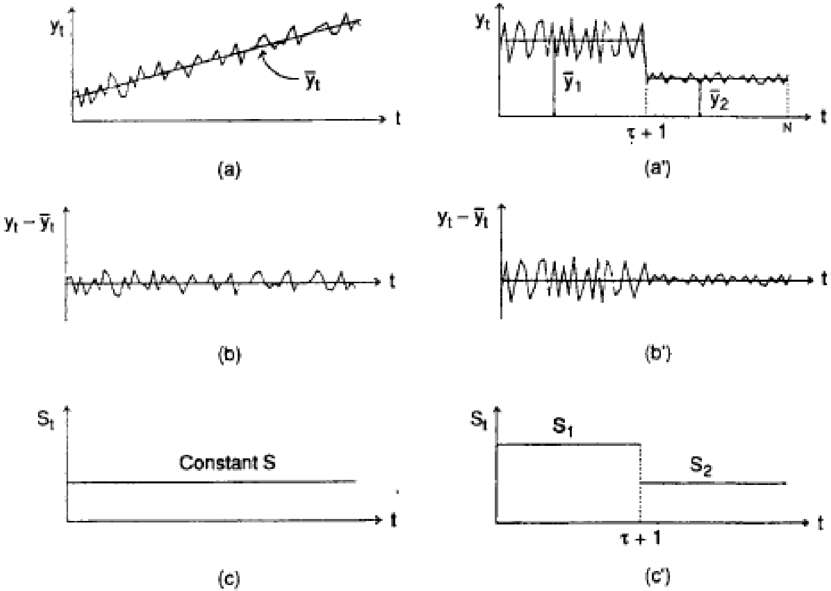
- Excel
- Matlab
- R
- SPSS
- Mintab
- etc

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



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 ± 1.96 at 5% level of significance

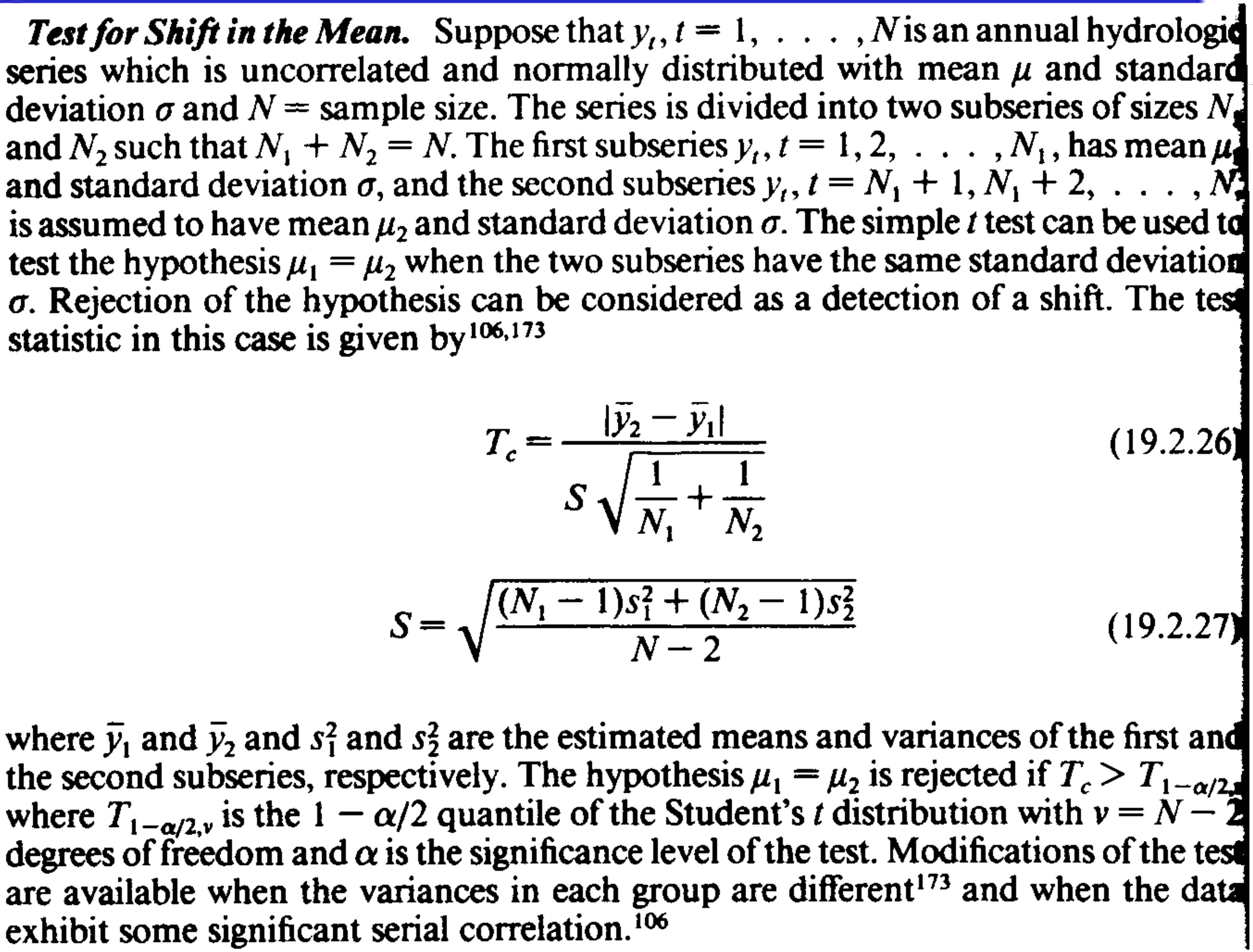
•If Zcal < Ztab there is no trend

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

Test for Shifts (Jumps)



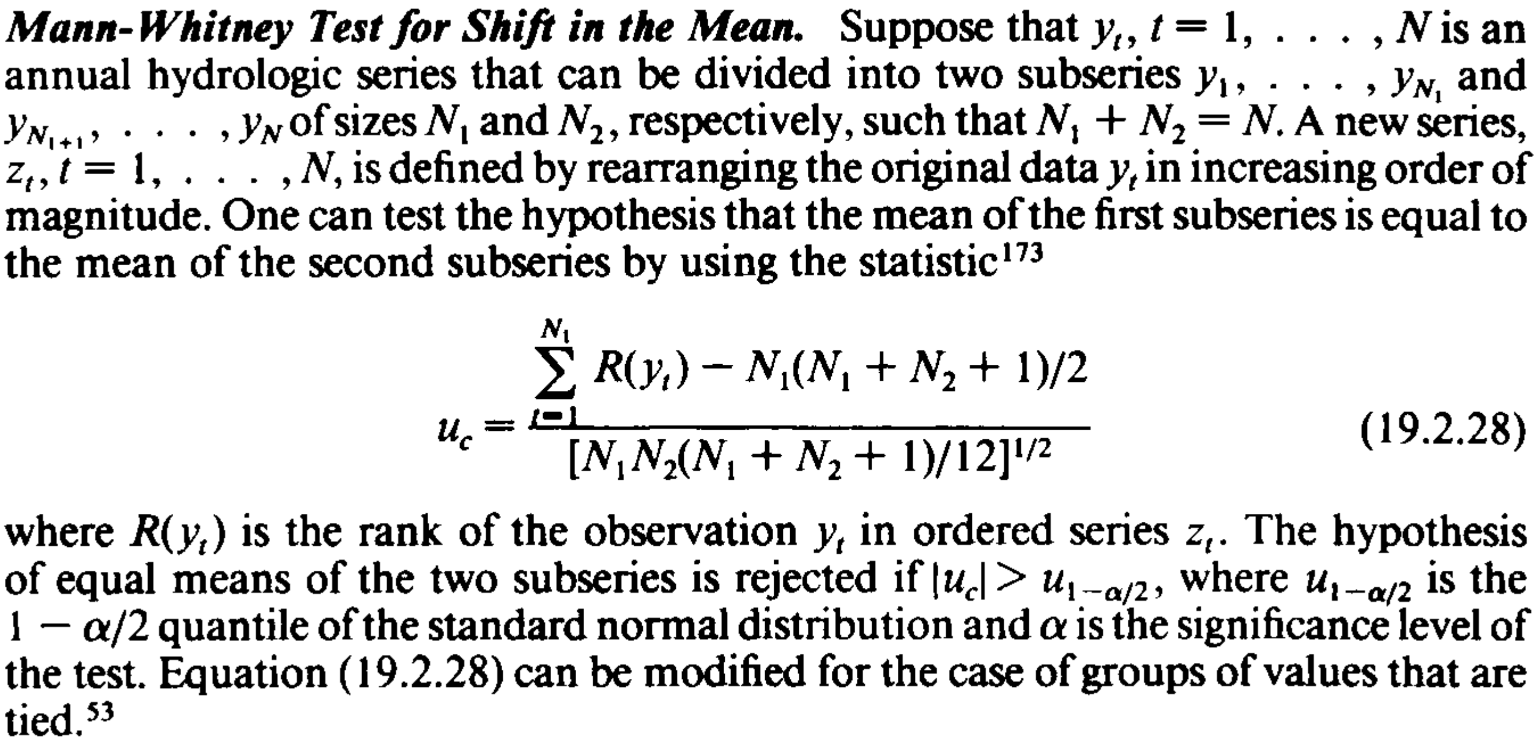
Test for Shift in the Mean. Suppose that y_t, $t = 1, \ldots, N$ is an annual hydrologic series which is uncorrelated and normally distributed with mean μ and standard deviation σ and N = sample size. The series is divided into two subseries of sizes N_1 and N_2 such that $N_1 + N_2 = N$. The first subseries 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]

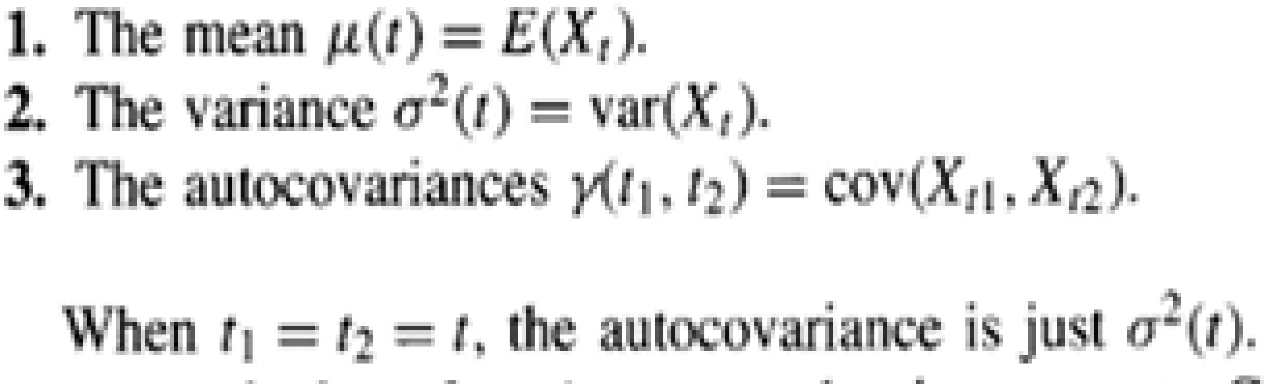
$$u_c = \frac{\sum_{t=1}^{N_1} R(y_t) - N_1(N_1 + N_2 + 1)/2}{[N_1 N_2(N_1 + N_2 + 1)/12]^{1/2}} \qquad (19.2.28)$$

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 $\sigma^2(t)$.