TRAINING COURSE

ON

SOFTWARE FOR GROUNDWATER DATA MANAGEMENT

UNDER

WORLD BANK FUNDED HYDROLOGY PROJECT

LECTURE NOTES
ON

PROCESSING OF GROUNDWATER DATA

BY

VIJAY KUMAR

ORGANISED BY

NATIONAL INSTITUTE OF HYDROLOGY
ROORKEE - 247 667
INDIA

PROCESSING OF GROUNDWATER DATA

1.0 INTRODUCTION

The evaluation, rational development and management of groundwater resources, which is very essential to feed the growing population, requires a through knowledge of the subsurface environment and an understanding of the hydrological processes that governs the occurance, movement and yield of groundwater. To do this various types of groundwater data is collected and stored in various forms. The processing of groundwater data is necessary to represent the data in more informative and useful form so that the data becomes meaningful to make some preliminary inferences and for further use.

The important features of the groundwater data can be captured by a few characteristics of the data. It is known as summary statistics and includes measure of location, measure of spread and measure of shape. The basic measure of location of data is some type of average value. Various measures like mode, median and arithmetic average exist. The measure of spread or dispersion about the mean is given by variance and the standard deviation. Standard deviation is often used instead of variance since its units are the same as the units of the variable being described. A small value of standard deviation indicates that observations are clustered tightly around a central value. On the other hand, a large standard deviation indicates that values are scattered widely about the mean and the tendency for central clustering is weak.

The shape of the distribution is described by the coefficient of skewness and the coefficient of variation. Coefficient of skewness provides information on the symmetry while the coefficient of variation provides information on the length of the tail for certain type of distribution.

Statistical analysis of water table data is also carried out to determine any underlying trends in the data i.e. to determine whether the groundwater is rising of falling in a particular aquifer. Trend indicates a long term growth or decline in the time series of groundwater owing to man's activities. The man's activities are related to artificial discharge and artificial recharge. Artificial discharge usually refers to pumping whereas artificial recharge refers to introduction of water into ground by wells, pits, excavations or by irrigation.

2.0 ESTIMATION OF THE MISSING DATA

In preparing data for analysis, some records are often found incomplete. To fill gaps in a time series or blank spaces on a map, and thus exploit partial records, the missing portions of a record may be estimated by such methods as interpolating from simultaneous records at nearby stations. Judgement is required in deciding how far to go in estimating missing data. If too few gaps are estimated, large quantities of nearly complete records may be ignored. If too many data are estimated, the aggregate data may be too diluted by interpolation. The methods described below under Data Interpolation can be used for this purpose.

3.0 DATA INTERPOLATION

Groundwater parameters (like groundwater levels, hydraulic conductivity, etc.) are measured generally as point values at scattered points which are generally nonuniformily distributed over an area. To get a complete picture of the spatial characteristics of these parameters, it is necessary to interpolate the values. Also, the distributed groundwater models requires the estimated/measured value of parameters at the nodes of a prespecified grid.

Different methods of interpolation of spatial variables are available and are used. In all these methods, the estimated value is represented by a weighted linear combination of observed values i.e.

$$Z_0^* = \sum_{i=1}^{N} \lambda_i Z_i \qquad (1)$$

where, Z_0^* = Estimated value of true value Z_0
λ_i = Weight assigned to a value Z_i
Z_i = Observed value at point i
N = No. of points

Different approaches of assigning the weights to the observed value give rise to different interpolation methods.

3.1 Arithmetic Mean Method

In this method it is considered that the variable is constant over the region and can be estimated by the average of all sample values. In this method

$$\lambda_i = \frac{1}{N} \qquad (2)$$

3.2 Nearest Neighbour Method

This is the simplest possible interpolation method. In this method, the estimated value at any given point is taken as the measure of value at the nearest data point. It is based on the assumption that the influence of an observed point extends halfway to the next observed point. The Thiessen polygon method is also based on this assumption.

The graphical procedure for this method (as used in Thiessen polygon) is to draw lines joining every pair of neighbouring points and to construct the perpendicular bisectors of these lines. The so formed polygons are assumed to have the properties of the observed point which it contains.

This method can also be defined as weighted linear combination that gives all of the weight to the closest sample value. This method has a main drawback that it does not provide a continuous representation of the process involved.

3.3 Distance Weighting Method

As the nearer points have more influence on the estimated point, this fact is taken into consideration by the distance weighting method. In this method the weights are inversely proportional to the distance or any power of distance from the estimated point.

$$\lambda_i = \frac{f(d_{0i})}{\sum_{i=1}^{N} f(d_{0i})} \tag{3}$$

Where,

$f(d_{0i})$ represents a given function of the distance d_{0i}.

A commonly used form is

$$f(d_{0i}) = \frac{1}{(d_{0i})^{\beta}} \tag{4}$$

As β decreases, the weights given to the sample becomes more similar. As β approaches 0, the distance weighting function approaches the arithmetic average method and as β approaches ∞, the method approaches the Thiessen polygon method. When $\beta=1$, the method is known as inverse distance interpolation and when $\beta=2$, it is known as the inverse square distance interpolation. The mostly used method is the latter one.

The drawbacks of this method are the arbitrariness in choosing the value of β and failure to discriminate redundant information as it does not take into account the position of sample points with each other.

3.4 Polynomial Interpolation

In this, a polynomial equation is fitted to the area of interest. Polynomial equation consists of monomials in terms of spatial coordinates of data points. The general form of the polynomial equation is written as :

$$Z_i = \sum_{k=1}^{m} a_k \phi_k(x_i, y_i) \tag{5}$$

where,

a_k = kth polynomial coefficient
$\phi_k (x_i, y_i)$ = kth monomial
m = total number of monomials fitted

The polynomial coefficients can be evaluated either using least square approach or the lagrange approach. The lagrange approach is an exact interpolation technique. In this case, the number of monomials are equal to the number of data points (i.e. m=N) and the fitted

equation passes through all the data points. Least square approach requires that the number of monomials must be less than the number of data points (m < N).

3.5 Triangulation Method

This method overcomes the problem of discontinuities as present in nearest neighbour method. This is done in this method by fitting a plane through three sample points that surround the point being estimated. The equation of plane can be expressed generally as

$$Z = a*x + b*y + c \qquad (6)$$

The three vertex of the plane give rise to three equations which can be solved for the three unknown parameters a, b, and c of the plane. Using the solved equation, the values can be calculated for any point within the triangle.

The estimated value at any point depends on which three nearby sample points are selected to fit a plane. One method of selection is known as Delaunary triangulation in which triangles are made in such a way that they are close to equilateral.

The main draw back of the triangulation method is that for different points, different triangles may have to be constructed and so many equations to be solved. Another drawback is that extrapolation is not possible.

Most of the geophysical variables exhibit to some extent some structural pattern which are not directly taken into account in the above mentioned approaches. G. Matheron developed the theory of regionalized variables and proposed a method of estimation which he called Kriging.

3.6 Basics of Kriging

Kriging is an interpolation technique based on the theory of regionalized variables as proposed by Matheron. It is a linear unbiased minimum variance estimator. The key arguments are that a geophysical process is generally continuous and it exhibits a structure, part of which may be decipherable in terms of its geographic location. A part of the structure may also be decipherable in terms of the degree of association to the local neighbourhood, independent of geographic setting. Further, measured values are corrupted by errors in measurement, analysis etc. The theory derives strength on account of the inclusion of the second component in a statistically acceptable manner.

In many a process, it is observed that high values tend to be near other high values and low values near other low values. A quantitative evaluation of this degree of association is obtained through the analysis of the average variance observed at various lag or separation distances. The plot of variance - called variogram when used in the spatial context-against separation distance is called the variogram. Generally, it is half the variogram that is plotted and the plot is accordingly called semivariogram. A sample of the semivariogram is shown in Fig. 1.

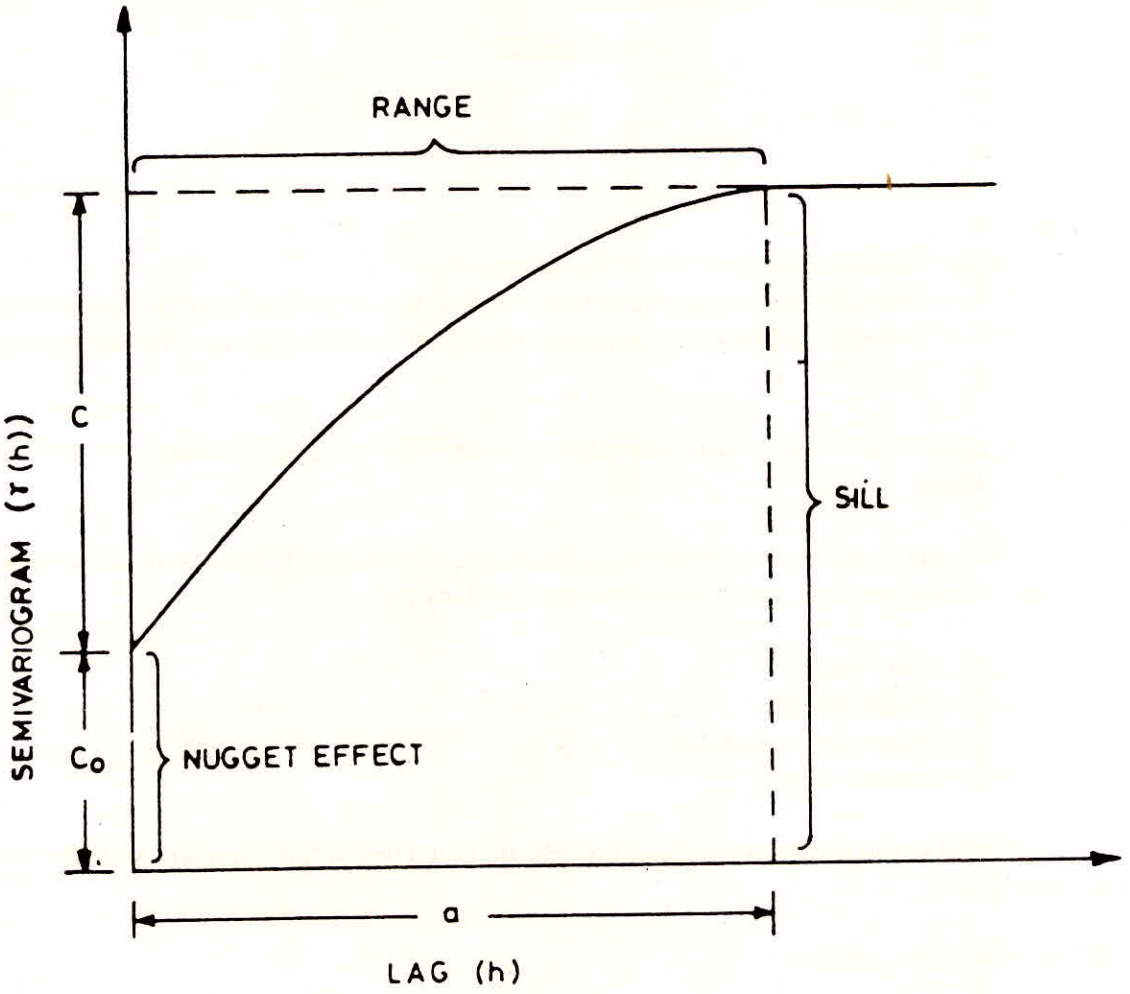


FIG. 1 PLOT OF SEMIVARIOGRAM

The semi-variogram is defined as

$$\gamma(h) = \frac{1}{2N}\sum_{i=1}^{N}[Z(x_i+h) - Z(x_i)]^2 \qquad (7)$$

Where,

$\gamma(h)$ = Semi-variogram value for a distance h
h = Average distance between pairs of data points belonging to a distance class.
N = Number of pairs of data points belonging to the distance interval represented by h

A plot of the semi-variogram against various values of h is the structural information used in kriging.

The commonly used function for representing the structural information as represented in a semi-variogram fall under the following categories :

(i) Linear model
(ii) Spherical model
(iii) Exponential model and
(iv) Gaussian model.

The function representing these models are given below and the shape of these models is given in Fig. 2.

- SPHERICAL MODEL

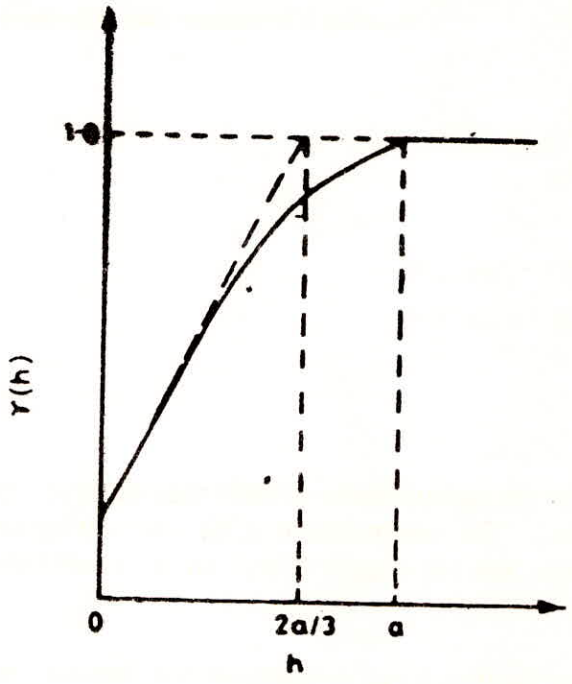
$$\gamma(h) = \begin{cases} C_0[1-\delta(h)] + C\left[\frac{3}{2}\frac{h}{a} - \frac{1}{2}\frac{h^3}{a^3}\right] & h \le a \\ C_0+C & h > a \end{cases} \qquad (8)$$

- EXPONENTIAL MODEL

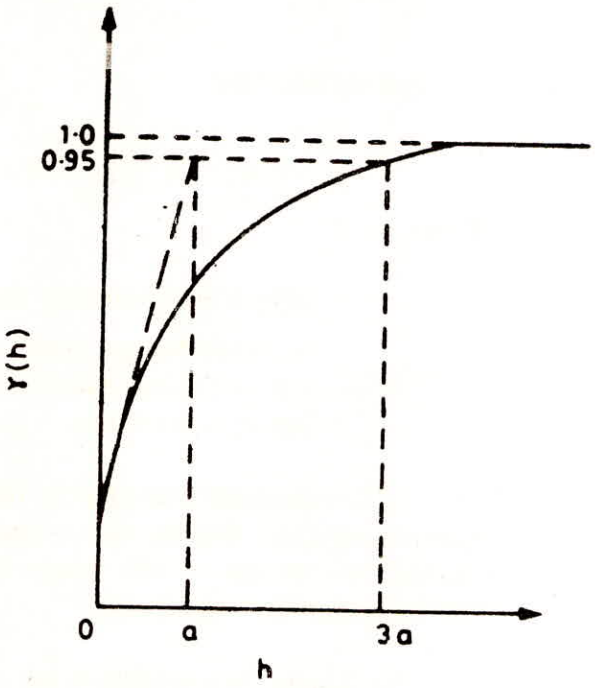
$$\gamma(h) = C_0[1 - \delta(h)] + C\left[1 - \exp\left(-\frac{h}{a}\right)\right] \qquad (9)$$

- GAUSSIAN MODEL

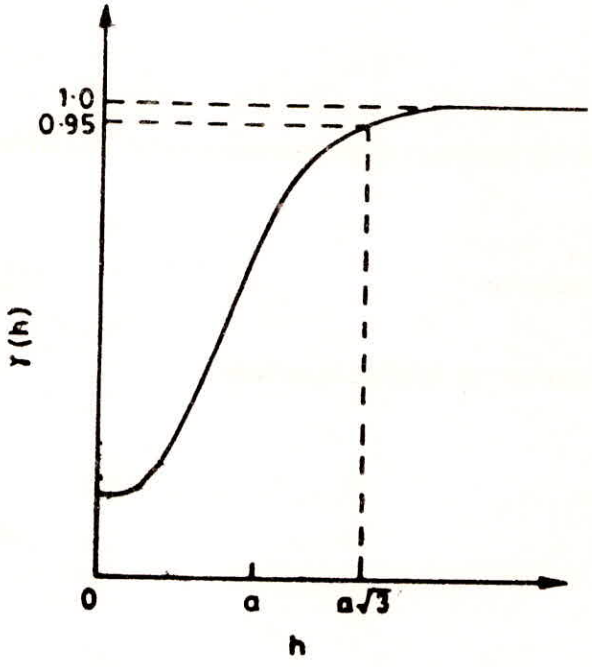
$$\gamma(h) = C_0[1 - \delta(h)] + C\left[1 - \exp\left(-\frac{h^2}{a^2}\right)\right] \qquad (10)$$



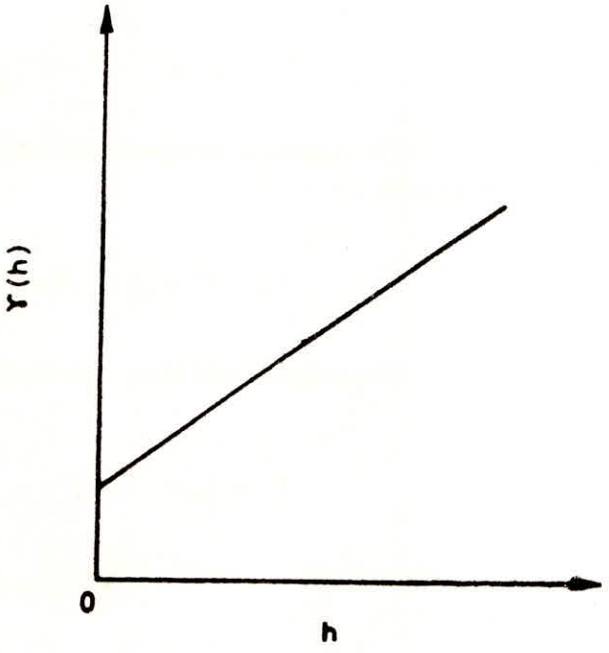
(a) SPHERICAL



(b) EXPONENTIAL



(c) GAUSSIAN



(d) LINEAR

FIG.2. THEORETICAL MODELS OF SEMIVARIOGRAM

LINEAR MODEL

$$\gamma(h) = C_0[\,1 - \delta(h)\,] + bh \tag{11}$$

Where,

$\delta(h)$ is the Kronecker delta $= \begin{cases} 1 & h = 0 \\ 0 & h \neq 0 \end{cases}$

C_0 is the Nugget effect
C0 + C is the Sill
and a is the Range

The variogram may also be obtained in an appropriately transformed domain; for example, log Z(x), may be used in place of Z(x). The interpretation of the semi-variogram requires the concepts of sill, range of influence and the nugget effect, all of which have physical interpretations as well.

In kriging, the weights λ_i are calculated so that $Z^*(x_0)$ is unbiased and optimal. An estimate is said to be unbiased if there is no systematic over or under estimation of the estimate. It means that the expected value of the estimation error (the difference between estimated $Z^*(x_0)$ and true (unknown $Z(x_0)$) value should on the average be zero.

$$E\{Z^*(x_0) - Z(x_0)\} = 0 \tag{12}$$

The condition of optimality means that the variance of the estimation error should be minimum i.e.

$$Var\,\{Z^*(x_0) - Z(x_0)\} = minimum \tag{13}$$

The solution of above two equation leads to the kriging equations.

$$\sum_{j=1}^{N} \lambda_j \gamma(x_i, x_j) + \mu = \gamma(x_i, x_0) \qquad i = 1, 2, 3 \ldots N$$
$$\sum_{j=1}^{N} \lambda_j = 1 \tag{14}$$

where

$\gamma(x_i,x_j)$ is the value of semi-variogram between two points x_i and x_j.

The solution of above kriging equations leads to the quantification of weights to be assigned to each observation point. The more detailed theory and some of the computer programmes are available in the book by Journal and Huijbregts (1978).

4.0 CORRELATION AND REGRESSION

Correlation attempts to measure the strength of relationship between two quantitative variables by means of a single number called Correlation coefficient(r). Correlation coefficient gives the measure of how the two variables X and Y vary together. Two variables are positively correlated if the large values of one variable tend to be associated with large values of the other variable, and similarly the smaller values of each variable. Two variables are negatively correlated if the large values of one variable tend to be associated with the smaller values of the other. The final possibility is that the variables are not related.

Correlation coefficient is the statistic that is most commonly used to summarize the relationship between two variables. It provides a measure of the linear relationship between two variables. It is actually a measure of how much close the observed values come to falling on a straight line. For a sample, It is given as

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y} \tag{15}$$

where $s_{X,Y}$ is the sample covariance between X and Y and s_X and s_Y are the sample standard deviation of X and Y respectively i.e.

$$s_{X,Y} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(Y_i - \bar{y})}{(n-1)} \tag{16}$$

$$s_X = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{(n-1)}} \quad \text{and} \quad s_Y = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{(n-1)}} \tag{17}$$

It can be shown that $-1 \leq r_{X,Y} \geq +1$. If $r=+1$, then the scatterplot will be straight line with a positive slope; if $r=-1$, then the scatterplot will be a straight line with a negative slope. A value of $r=0$ implies a lack of linearity and not necessarily a lack of association. For $|r| < 1$ the scatterplot appears as a cloud of points that becomes fatter and more diffuse as $|r|$ decreases from 1 to 0.

It is important to note that r provides a measure of the linear relationship between two variables. If the relationship is not linear, the correlation coefficient may be very poor summary statistics.

Fig.3. demonstrates some typical values for $r_{X,Y}$. In Fig.3(a) all of the points lie on the line Y=X-1 and consequently there is perfect linear dependence between X and y and the correlation coefficient is unity. In Fig.3(b) the points are either on or slightly off the line Y=X-1, and $r_{X,Y}$=0.986. In Fig.3(c) the correlation coefficient has dropped to -0.671. In fig.3(d), $r_{X,Y}$ = 0.211, the scatter of the points is very large and with a corresponding lack of strong dependence. Fig.3(e), the relationship is $Y = X^2/4$ for $X \geq 0$ and the $r_{X,Y}$ = 0.963. It shows that even through the dependence between X and Y is nonlinear, a high correlation coefficient can result. Fig.3(f) shows the converse. In this the X any Y are perfectly functionally related as $Y = \pm\sqrt{9 - X^2}$ for $-3 \leq X \geq 3$, but the correlation coefficient is zero.

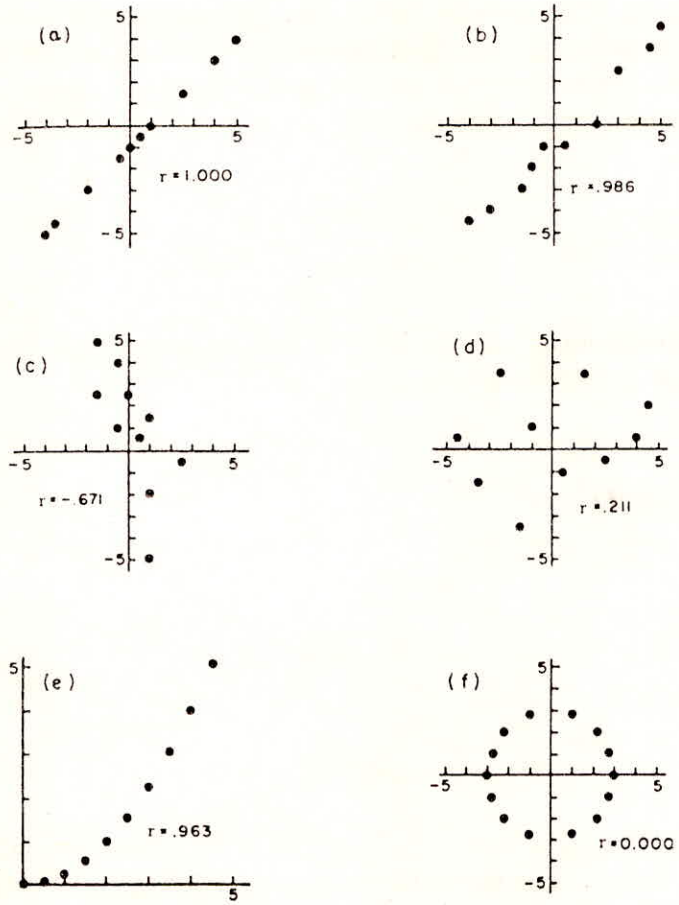


Fig. 3. Examples of the correlation coefficient.

In some of time series of hydrologic data, an observation at one time period is correlated with the observation in the preceding time period. Such correlation is termed as serial correlation or autocorrelation.

Correlation coefficient is useful, in determining whether a linear relationship between two variables exists and, second, in providing a measure of the strength of this relationship. It only expresses association and by itself tells nothing of the causal relationships of the variates.

A more useful approach to the study of simultaneous variation of two characterists when a relationship exists is the study of regression. The underlying relation between the expected value y and x in a bivariate population can be expressed in the form of a mathematical equation known as regression equation and said to represent the regression of the variate y on the variate x. The relationship can be written as

$$Y = a + b X \tag{18}$$

It should be noted that fitting a straight line to a set of observations does not imply that the data actually follow a straight line relationship.

The regression coefficients a and b in the above equation are determined by the least square procedure. The least square technique uses the criterion of minimization of square of error between the approximated and observed value i.e.

$$Z = Min.\ e^2 = \sum (\hat{Y}_i - Y_i)^2 \tag{19}$$

Where, Y_i= Approximated value
Y_i= Observed value

The minimisation of Z can be obtained by equating to zero the partial derivatives of the above equation with respect to the coefficients.

$$\frac{\partial Z}{\partial a} = -2 \sum (Y_i - a - b X_i) = 0 \tag{20}$$

$$\frac{\partial Z}{\partial b} = -2 \sum X_i (Y_i - a - b X_i) = 0 \tag{21}$$

These equations can be written as

$$\sum (Y_i - a - b X_i) = 0 \tag{22}$$

$$\sum X_i (Y_i - a - b X_i) = 0 \tag{23}$$

The solution of the above equations in terms of the regression coefficients a and b are given as

$$b = \frac{\sum X_i Y_i - \frac{(\sum Y_i \sum X_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} \tag{24}$$

$$a = \frac{\sum Y_i - b \sum X_i}{n} \tag{25}$$

The above procedure of determining a and b is known as simple regression. 'Simple' because it uses only one variable i.e. X.

The significance of a regression can be tested by performing an analysis of variance. It gives the goodness of fit i.e. how well does the regression equation accounts for variations in the dependent variable. The ratio of the sum of squares due to regression to the total sum of squares corrected by the mean can be used as a measure or ability of the regression line to explain variations in the dependent variable. Numerically it is given by coefficient of determination (R^2).

$$R^2 = RSS/TSS \tag{26}$$

where

RSS = Regression (explained) sum of squared deviation

$$= \sum (\hat{Y}_i - \bar{Y})^2$$

TSS = Total sum of squared deviation

$$= \sum (Y_i - \bar{Y})^2$$

R^2 indicates the explanatory power of the regression model. The possible values of the measure range from '+1' to '0'. When R^2 is near a value of +1, it shows a good fit of data.

4.1 Inferences on Regression Coefficients

The confidence intervals on the regression coefficients represent the range of the values which these coefficients may take for a specified confidence level. The confidence limits on a and b can be estimated from

For a

$$L_a = a - t_{1-\alpha/2,n-2} \cdot S_a$$

$$U_a = a + t_{1-\alpha/2,n-2} \cdot S_a$$

For b

$$L_b = b - t_{1-\alpha/2,n-2} \cdot S_b$$

$$U_b = b + t_{1-\alpha/2,n-2} \cdot S_b$$

where, L stands for lower limit and U for upper limit. α is the confidence level, and $(1-\alpha/2)(n-2)$ represents t value corresponding to $1-\alpha/2$ confidence limit and n-2 degree of freedom. S_a and S_b is the variance of a and b respectively and given as

$$S_a = S\sqrt{\frac{1}{N} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}} \tag{27}$$

$$S_b = \frac{S}{\sqrt{\sum (X_i - \bar{X})^2}} \tag{28}$$

and

$$S = \sqrt{\frac{\sum (\hat{Y}_i - Y_i)^2}{n-2}} \tag{29}$$

Test of hypotheses concerning a and b can be made by noting that $(a-a_0)/S_a$ and $(b-b_0)/S_b$ both have t distributions with N-2 degree of freedom. The hypothesis H_0: $a=a_0$ versus H_a: $a \neq a_0$ is tested by computing

$$t = (a-a_0)/S_a$$

H_0 is rejected if $|t| > t_{1-\alpha/2,n-2}$. Similarly H_0: $b=b_0$ versus H_a: $b \neq b_0$ is tested by computing

$$t = (b-b_0)/S_b$$

H_0 is rejected if $|t| > t_{1-\alpha/2,n-2}$.

The significance of overall regression equation can be evaluated by testing the hypothesis that b=0. If this hypothesis is accepted, the regression line does not explain a significant amount of the variation in Y.

4.2 Confidence Intervals on Regression Line

The confidence limits on the regression line are computed using the following relationships:

$$L = \hat{y}_k - t_{(1-\alpha/2),(n-2)} \cdot S_{\hat{y}_k} \tag{30}$$

$$U = \hat{y}_k + t_{(1-\alpha/2),(n-2)} \cdot S_{\hat{y}_k} \tag{31}$$

where,

$$\bar{Y}_k = a + bx_k \tag{32}$$

$$S_{\hat{y}_k} = S\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]^{1/2} \tag{33}$$

L and U represent lower and upper confidence limits.

4.3 Confidence Intervals on an Individual Predicted Value

The confidence limits for an individual predicted value of y (dependent variable) are computed using the following relationships:

$$L' = \hat{y}_k - S_{\hat{y}_k} \cdot t_{(1-\alpha/2),(n-2)} \tag{35}$$

$$U' = \hat{y}_k + S_{\hat{y}} \cdot t_{(1-\alpha/2),(n-2)} \tag{35}$$

$$S'\hat{y}_k = S\left(1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right)^{1/2} \tag{36}$$

Many a times a dependent variable is dependent on several other quantities. The equation in this case is of the form of

$$Y = a + b*X_1 + c*X2 + \ldots\ldots\ldots \tag{37}$$

where,

Y = Dependent variable

X_1, X_2 = Independent variables

a,b,c = Coefficients of the polynomial regression

The regression coefficients in the above equation are also determined by the least square procedure. This type of regression is known as multiple regression.

The above equation can be written in the matrix form as below

$$[Y] = [X] [B] \quad (38)$$

where,

[Y] = column matrix of dependent variable of order N*1

[B] = column matrix of coefficients to be calculated M*1

[X] = square matrix of independent variables of order n*m

N = no. of data points

M = no. of coefficients to be calculated

The general solution of the above matrix is given by

$$[B] = [X^T X]^{-1} [X^T Y] \quad (39)$$

5.0 PROBABILISTIC DISTRIBUTIONS

A distribution is an attribute of a statistical population. The distribution describes their location on the axis; tells whether they are bunched together or spread out; and whether they are symmetrically disposed on the X axis or not. It also tells the relative frequency or proportion of various X values in the population in the same way that a histogram gives that information about a sample. These relative frequencies are also probabilities and hence the distribution tells us the probability, $Pr(X<x)$, that the X value on an element drawn randomly from the population would be less than a particular value x. Knowing $Pr(X<x)$ for all x values , the laws of probability may then be used to deduce the probability of any proposition about the behaviour of a random sample of X values drawn from the population.

The probability distributions occurring in hydrologic processes can be divided into two types known as discrete and continuous types. The variables in the process are discrete if they are restricted to specific incremental values. The variables in the process are continuous if they can take on all values in the range of occurance, including figures differing by an infinitesimal amount. Distributions for continuous variables present a different picture. In practice, they are generally studied and analysed by grouping total length N of the data into a number of different intervals. As the number of observations N increases, the continuous distribution develops as a result of reducing the size of class interval. Fig.4 is a typical plot of the distribution of probability associated with the values that a discrete random variable can assume. This information would constitute a probability distribution or probability function for a discrete random variable. The cumulative probability distribution for this discrete random variable is shown in Fig.5. Fig.6 and Fig.7 illustrate a possible probability density function and its corresponding cumulative distribution function for a continuous variable.

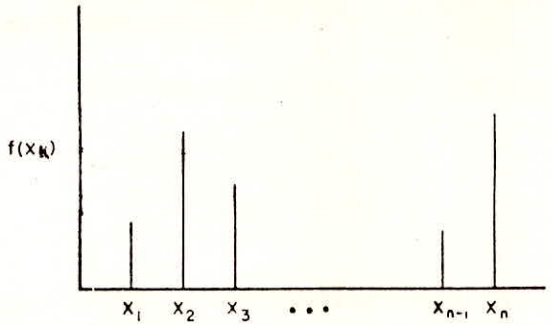


Fig. 4 A discrete probability distribution.

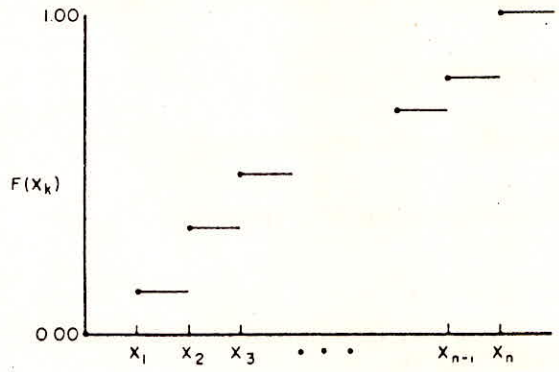


Fig. 5 A discrete cumulative probability distribution.

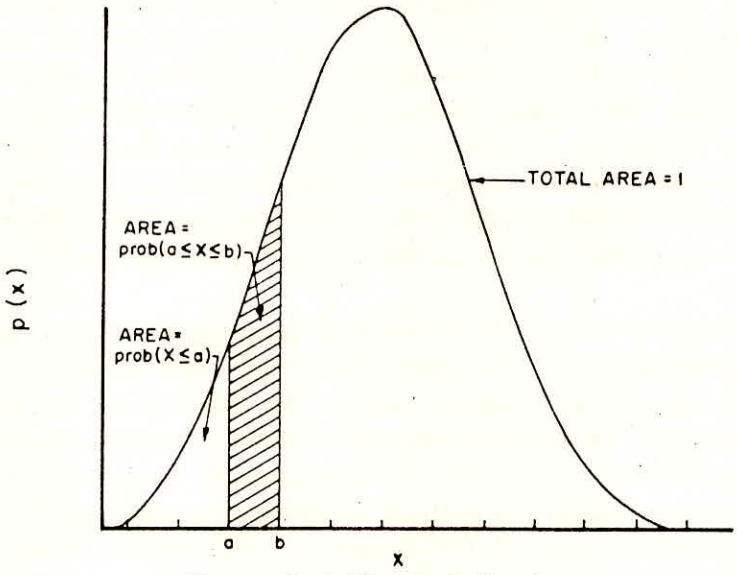


Fig. 6 Probability density function.

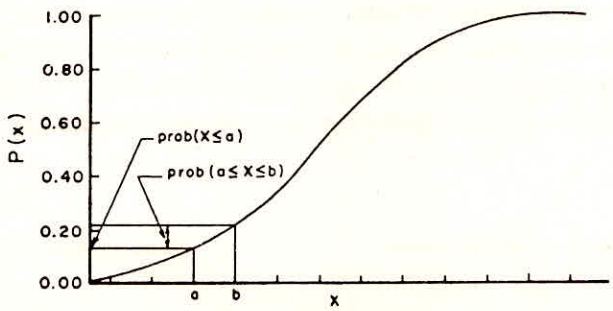


Fig. 7 Cumulative probability distribution function.

The use of discrete probability distributions is restricted generally to those random events in which the outcome can be described as success or failure. Furthermore, the successive trails are independent and the probability of success remains constant from trail to trail. But, most hydrological variables are assumed to come from a continuous random process, and the historical sequences thereof are fitted with some common continuous distributions.

For evaluating the future performance of water resources projects, the statistics involves the analysis of historical data to give observed distributions, which is then approximated with one of the known distributions. To do this, the study and understanding of some of the theoretical distributions is a must before these can be used for comparing with the observed distributions obtained from historical data. The underlying assumption in this evaluation is that the future samples will have the same of similar properties as that of the available sample.

5.1 Discrete Frequency Distributions

Binomial distribution

This applies to populations that have only two discrete but complementary events. For example, rainy and nonrainy days. If X is a random variable representing the number of rainy days in n days and the probability of rainy days is p, then the probability of occurance of event x times in n days is given as

$$P(X=x) = \frac{n!}{x!\,(n-x)!} p^{x} (1-p)^{n-x} \qquad (40)$$

Where,

p = probability of exceedance/ success,

x = number of exceedance/ successes,

n = total number of events.

The mean and variance of the binomial distribution are

$E(X) = np$

$Var(X) = npq$

Poisson distribution

If in the binomial distribution n gets large while p gets small with their product np tending to a constant λ, the Binomial distribution approaches the Poisson distribution given by

$$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{41}$$

Any extreme value problem in which the occurance of an event has a probability p proportional to the period of observation, n can be a Poisson variate.

The mean and variance of the Poisson distribution are

$E(X)=\lambda$

$Var(X)=\lambda$

The conditions for this approximations are:

(i) The number of events is discrete.
(ii) Two events can not coincide.
(iii) The mean number of events in unit time is constant.
(iv) Events are independent.

5.2 Continuous Frequency Distributions

Normal Distribution

The most widely used and most important continuous probability distribution is the Gaussian or normal distribution. This is used to fit empirical distributions with symmetrical histograms or with skewness coefficient close to zero. It is a bell-shaped probability density function(PDF) of a random continuous variable and is given as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(x-\mu)^2/2\sigma^2\right) \tag{42}$$

The two parameters of the distribution are the mean, μ and the standard deviation, σ. By a simple transformation, the distribution can be written as a single parameter function only. Thus when $t=(x-\mu)/\sigma$, the PDF becomes

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right) \tag{43}$$

The Cumulative Density Function(CDF) is given as

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-t^2/2\right) dt \tag{44}$$

Lognormal Distribution

Many hydrological variables show a marked right skewness, i.e. the observed frequency distribution is not symmetric. In such cases frequencies will not follow the normal distribution, and instead, variables are often functionally normal and their logarithms follow a normal distribution.

If Y = ln(X) follows normal distribution, then X follows lognormal distribution. The PDF is given as

$$f(y) = \frac{1}{\sigma_y \sqrt{2\pi}} \exp\left(-(y-\mu_y)^2 / 2\sigma_y^2\right) \tag{45}$$

The following relationship have been found to hold good between the characteristics of the untransformed variate X and the transformed variate Y.

$$\mu = \exp\left(\mu_y + \sigma_y^2 / 2\right) \tag{46}$$

$$\sigma^2 = \mu^2 \left[\exp\left(\sigma_y^2\right) - 1\right] \tag{47}$$

Gamma distribution

The gamma distribution has wide applications in hydrological studies. A special case of Gamma distribution, known as Pearson type III, is more useful. This distribution has been widely used as the standard method for flood frequency analysis in a form known as the log-Pearson type III in which the transformation Y=log(X) is used to reduce skewness.

The PDF is given as

$$f(x) = \frac{x^{\alpha} \exp\left(-x/\beta\right)}{\beta^{\alpha+1}\Gamma(\alpha+1)} \tag{48}$$

Γ is the gamma function.

BIBLIOGRAPHY

1. Davis, J.C. (1973). Statistics and Data Analysis in Geology. John Wiley, New York.

2. Hann, C.T. (1977). Statistical Methods in Hydrology. The Iowa State University Press. Ames.

3. Mutreja, K.N.(1986). Applied Hydrology. Tata McGraw-Hill Publishing Company limited, New Delhi.

4. National Institute of Hydrology (1995). Processing of Groundwater data. CS (AR)-171.

5. National Institute of Hydrology (1987). Surface Fitting of Groundwater Table by Means of Least Square Approach. TR-25.

6. Journal, A.G. and C.J. Huijbregts (1978). Mining Geostatistics. Academic Press, London, England.

7. Matheron, G. (1963). "Principles of Geostatistics". Economic Geology, 58, pp. 1246-1266.