LECTURE-5

METHODS OF PARAMETER ESTIMATION

OBJECTIVES

This lecture presents the theoretical basis of techniques for estimating the parameters of the probability distribution using the available sample data. The limitations of some techniques based on theoretical consideration and overcoming of these limitations by other techniques have been indicated with reference to normal, log normal, Gumbel EV type-I, and Pearson and log Pearson type III distributions.

5.1 INTRODUCTION

This lecture deals with the statistical methods, in general use, for estimating the parameters of the distribution when streamflow records are available at a site under consideration. These methods will be discussed with reference to the annual maximum series throughout because this series is most often used in practice. Normal, log normal, Pearson type-III, log Pearson type-III and Gumbel EV type-I distributions are considered here for parameter estimation purposes. Since no deductive method exists for deciding on the form of a distribution the data would follow, the suitability of each possible candidate distribution must be examined. Each distribution is considered in turn to be the correct distribution and some numerical index is calculated expressing the agreement or lack of fit between the assumption and the information about the distribution contained in flow records. In this case the information in the flow record is that contained in the appropriate series above. First the parameters of the distribution must be estimated from the data. A numerical measure of the difference between the two distributions, fitted theoretical and observed, should be used to make a decision between different forms of distributions. Thus in choosing between the different forms of distribution at a given site of river having historical data, there are two steps :

1. estimation of parameters, and
2. calculation of the numerical measure of agreement

Only step 1 is covered here.

There are four well known parameters estimation techniques, viz.,

1. Graphical
2. Least squares
3. Method of moments and
4. Method of maximum likelihood.

The four methods have been listed in ascending order of efficiency. The first two methods are discussed here only briefly as the graphical method has been discussed in an earlier lecture and the least square method has been found less usefull in practice due to its dependence on plotting position formula and for giving equal weight to the errors of data. More emphasis would be given for the other two methods as they are entirely objective.

5.2 GRAPHICAL ESTIMATION

In graphical estimation, as discussed earlier, the variate under consideration is regarded as a function of a standardised or reduced variate of a known distribution. The sample data is plotted as a series of N discrete points on an ordinary graph paper with abscissa being the reduced variate of the probability distribution under consideration and the ordinate being the variate. The variates are plotted against the corresponding probability or reduced variate or return period determined using the appropriate plotting position formula. The suitable plotting position formulae were discussed earlier. The plotted points on the graph paper represent the sample distribution and a line drawn through these is considered as an estimate of the population relation. Then this straight line is projected to arrive at the flood magnitudes of desired return period.

In graphical estimation the line is subjectively placed and could vary with analyst. This subjectivity is regarded as a major drawback by hydrologists.

5.3 LEAST SQUARES ESTIMATION

In the least squares estimation technique a simple linear regression equation is fitted between the variate under consideration and the corresponding frequency factor K. The form of Chow's general equation is used as the linear regression equation and it is written as :

$$X_i = a + b\,K_i + \epsilon_i \qquad (5.1)$$

in which,

a and b are the intercept and slope of the linear regression equation.

X_i = the i^{th} variate.

K_i = the frequency factor corresponding to the i^{th} variate.

ϵ_i = error term with mean = 0 and standard deviation $\sigma\epsilon$

It is not correct to interpret a and b as mean and standard deviation of the X_i series as these parameters are estimated in the least squares sense and as such they can never become equal to mean and standard deviation of the sample data

This method has not been accepted as a standard method in practice as it involves the use of plotting position formula to determine the frequency factor K_i and due to the assumption that the error variance $\sigma^2\epsilon$ remain same for all observations. The defect due to the former assumption could be eliminated by using the appropriate plotting positions discussed in an earlier lecture.

However the later assumption makes the method more, defective as the higher events recorded have more error variance than the recorded lower events. All these assumptions affect the correct parameter estimation of a and b.

As method of moments and method of maximum likelihood are objective methods of determining parameters they would be dealt in detail for normal, log normal, Gumbel EV type–I, Pearson Type III and log Pearson type III distributions.

5.4 METHOD OF MOMENTS

The method of moments makes use of the fact that if all the moments of a distribution are known then everything about the distribution is known. For all the distributions in common usage, four moments or fewer are sufficient to specify all the moments. For instance, two moments, the first together with any moment of even order are sufficient to specify all the moments of the normal distribution and therefore the entire distribution. Similarly, in the Gumbel EV type–I distribution, the first two moments are sufficient to specify all the moments and hence the distribution. For Pearson type III distribution three moments, always taken as the first three are required to specify all the moments. In these cases the number of moments needed to specify all the moments and hence the distribution equals the number of parameters.

The method of moments estimation is dependent on the assumption that the distribution of variate values in the sample is representative of the population distribution. Therefore, a representation of the former provides an estimate of the later. Given that the form of the distribution is known or assumed, the distribution which the sample follows is specified by its first two or three moments calculated from the data.

Having estimated the mean and standard deviation for two parameter distributions, the magnitude of the required return period flood is computed using Chow's general frequency equation as :

$$X_T = \mu + K_T \sigma \tag{5.2}$$

in which,

X_T = the magnitude of flood at required return period T

K_T = the frequency factor corresponding to T.

μ and σ = mean and standard deviation of the population which would be replaced by the sample statistics.

5.4.1 Normal Distribution

The parameters of the normal distribution, which describe the characteristics of the given data set consisting of N values, are computed as :

$$\mu \simeq \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{5.3}$$

$$\sigma^2 \cong S^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{5.4}$$

$$\sigma \simeq S = \left[\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \right]^{\frac{1}{2}} \tag{5.5}$$

If the sample size is small, then eq. 5.4 is modified as :

$$\sigma^2 \cong S^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{5.6}$$

5.4.2 Log Normal Distribution

Conventional approach :

In the conventional approach the variates are transformed to the log domain by taking log of each variate, and then the mean and standard deviation of the transformed variates are computed as given by eqs. 5.3 and 5.5.

Eq. 5.2 is employed to compute the flood at required interval with K being the standard normal deviate. Then the computed magnitude of flood in log domain is transformed to the original domain.

Chow's approach :

Chow (1964) theoretically related the mean and standard deviation of the given series with the mean and standard deviation of the log transformed series and thus enabling to find the expression for the frequency factor K of log domain in terms of orginal domain statistics. Using this K, the flood of required magnitude is computed from eq. 5.2. Note that Chow s approach does not require the transformation of the original series to log domain in order to transform the series to log normal distribution. In this approach the original series characteristics are not distorted in the log domain.

Chow's relationships needed for the computation of frequency factor are given below :

$$\sigma = \mu \, (e^{\sigma_y^2} - 1)^{1/2} \tag{5.7}$$

$$K = \frac{e^{\sigma_y K_y - \sigma_y^{2/2}} - 1}{(e^{\sigma_y^2} - 1)^{1/2}} \tag{5.8}$$

in which

μ = the population mean of the flood series

σ = the population standard deviation of the flood series.

y = the log transformed series

K_y = the standard normal deviate

K = the frequency factor for the log normal distribution.

σ_y = the population standard deviation of the y series.

To compute flood magnitude of required return period using log normal distribution, the population mean and standard deviation of eq. 5.2 are replaced by the sample statistics, and the frequency factor K is computed using eq. 5.8.

5.4.3 Gumbel's Extreme Value Type–I Distribution

The form of the probability density function and the distribution function of this distribution are given as :

$$f(x) = \frac{1}{\alpha} e^{-(x-u)/\alpha - e^{-(x-u)/\alpha}} \qquad (5.9)$$

$$F(x) = e^{-e^{-y}} \qquad (5.10)$$

in which,

$y = (x-u)/\alpha$, and

u and α are the parameters of the distribution. The parameter estimation by method of moments, of this distribution, has been presented indirectly in the lecture on 'Frequency analysis by graphical techniques'. For the sake of clarity, the relationship between the distribution parameters and the statistical moments of the data are reproduced here :

$$\mu = u + 0.5772\,\alpha \qquad (5.11)$$

$$\sigma = \alpha \frac{\pi}{\sqrt{6}} \qquad (5.12)$$

Solving eqs. 5.11 and 5.12, for u and α, we get

$$\alpha = 0.7797/\sigma \qquad (5.13)$$

$$u = \mu - 0.45\,\sigma \qquad (5.14)$$

The population statistics μ and σ would be replaced by the sample statistics while computing u and α from a given set of data. The frequency factor K required to be used in eq. 5.2 is given as :

$$K = -\,[0.45 + 0.7797 \ln - \ln\ (1-1/T)] \qquad (5.15)$$

5.4.4 Pearson Type-III Distribution

The probability density function and distribution function are respectively :

$$f(x) = \frac{(x - x_o)^{\gamma-1}\ e^{-(x - x_o)/\beta}}{\beta^{\gamma}\ \Gamma\gamma} \qquad (5.16)$$

$$F(x) = \int_{x_o}^{x} \frac{(x - x_o)^{\gamma-1}}{\beta^{\gamma}\ \Gamma\gamma}\ e^{-(x - x_o)/\beta} dx \qquad (5.17)$$

in which x_o, γ and β are the distribution parameters and the notation 'Γ' stands for gamma function.

The reduced variate is given as

$$y = (x - x_o)/\beta \qquad (5.18)$$

Substituting the probability density function of the Pearson type III distribution into the general equation for moments about the origin yields

$$\mu'_r = \int_0^{\infty} x^r \frac{(x - x_o)^{\gamma-1}}{\beta^{\gamma}\ \Gamma\gamma}\ e^{-(x - x_o)/\beta} dx \qquad (5.19)$$

Substituting eq. 5.18 in eq. 5.19 yields

$$\mu'_r = \frac{1}{\Gamma\gamma} \int_0^{\infty} (\beta y + x_o)^r\ y^{(\gamma-1)}\ e^{-y}\ dy \qquad (5.20)$$

which can be evaluated by noting that

$$\int_0^{\infty} y^{\gamma-1}\ e^{-y}\ dy = \Gamma\gamma\text{, the gamma function.} \qquad (5.21)$$

As an example, for $r = 1$, the first moment about the origin, μ'_1 is given by

$$\mu'_1 = \frac{1}{\Gamma\gamma} \int_0^{\infty} (\beta y^{\gamma}\ e^{-y} + x_o\ y^{\gamma-1}\ e^{-y})\ dy \qquad (5.22)$$

$$\mu'_1 = \frac{\beta\,\Gamma(\gamma+1)}{\Gamma(\gamma)} + x_o\,\frac{\Gamma(\gamma)}{\Gamma(\gamma)} = \beta\gamma + x_o \tag{5.23}$$

For $r = 2$

$$\mu'_2 = \beta^2\gamma\,(\gamma+1) + 2\beta\gamma\,x_0 + x_0^2 \tag{5.24}$$

and since

$$\mu_2 = \mu^1{}_2 - \mu^2{}_1, \tag{5.25}$$

the second central moment is given by

$$\mu_2 = \sigma^2 = \beta^2\gamma \tag{5.26}$$

Similarly higher order central moments can be calculated as :

$$\mu_3 = 2\,\beta^3\gamma \tag{5.27}$$

The coefficient of skewness is computed as :

$$g = \frac{\mu_3}{\mu_2^{3/2}} = \frac{2}{\sqrt{\gamma}} \tag{5.28}$$

From eqs. 5.23, 5.26 and 5.28 one can solve for x_0, β and γ and they are given as :

$$\gamma = 4/g^2 \tag{5.29}$$

$$\beta = \frac{\sigma g}{2} \tag{5.30}$$

$$x_0 = \mu - 2\sigma/g \tag{5.31}$$

Substituting these relationships in eq. 5.18 and solving for x gives :

$$x = \mu + \sigma\,(-2/g + g\gamma/2) \tag{5.32}$$

Therefore the frequency factor K is given as :

$$K = -2/g + g\gamma/2 \tag{5.33}$$

The K values are listed in APPENDIX—IV for a given g and for various probabilities of exceedances. Use of eq. 5.2 is made to compute floods at required return period.

5.4.5 Log Pearson Type-III Distribution

The parameters relationships for log Pearson type-III distribution are same as that of Pearson type-III distribution except that the sample statistics are computed from the log transformed flood series. The US Water Resources Council has recommended its use for federal agencies design flood estimation works based on frequency analysis.

5.5 METHOD OF MAXIMUM LIKELIHOOD (M.M.L.)

The principle of maximum likelihood states that for a distribution with probability density function p $(x; \alpha, \beta, \ldots)$ where $\alpha, \beta, \ldots$ are the distribution parameters to be estimated, the probability of obtaining a given value of x_i is proportional to p $(x_i; \alpha, \beta, \ldots)$ and the joint probability, L, of obtaining a sample of N values $x_1, x_2, \ldots x_n$ is proportional to the product.

$$L = \prod_{i=1}^{n} p\,(x_i; \alpha, \beta, \ldots) \tag{5.34}$$

This is called the likelihood. The method of maximum likelihood is to estimate $\alpha, \beta, \ldots\ldots$ such that L is maximised. This is obtained by partially differentiating L with respect to each of the parameters and equating to zero. Frequently ln (L) is used instead of L to simplify computations.

In this lecture we deal only with the parameter estimation of normal and Gumbel's EV type-I distribution by MML procedure. The estimation of parameters of Pearson type-III or log Pearson type-III by MML is far from easy. Therefore only the method of moments is adopted for estimating the parameters of this distribution.

5.5.1 Normal Distribution

The probability density of normal distribution is written with parameters μ and σ as :

$$p\,(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} \tag{5.35}$$

so that likelihood function is expressed as :

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N e^{-\frac{1}{2}\sum_{i=1}^{N}(x-\mu)^2/\sigma^2} \tag{5.36}$$

Taking logarithms

$$\ln L = \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \left(\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}\right) \tag{5.37}$$

Differentiating with respect to the parameters μ and σ^2 and equating to zero

$$\frac{\partial\,(\ln L)}{\partial \mu} = \frac{\sum_{i=1}^{N} (x_i - \mu)}{\sigma^2} = 0 \tag{5.38}$$

so that

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \mu = 0 \tag{5.39}$$

but

$$\sum_{i=1}^{N} \mu = N\,\mu \tag{5.40}$$

so that

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{5.41}$$

Also

$$\frac{\partial\,(\ln L)}{\partial\,\sigma^2} = -\frac{N}{2\sigma^2} + \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{2\sigma^4} = 0 \tag{5.42}$$

so that

$$\sigma^{`} = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \tag{5.43}$$

(i.e.) parameters μ and σ are the mean and standard deviation of the distribution. Note that variance corresponds to that of population and so for small sample data it is a biased estimate.

5.5.2 Gumbel's EV Type-I Distribution

The maximum likelihood solution of Gumbel's EV type-I distribution was first proposed by Kimball (1946) but was not practical until the advent of computers.

The maximum likelihood method of estimating the parameters of EV-1 distribution described by the probability density function given in eq. 5.9, is based on the concept that the probability of N individual maximum events $X_1, X_2......X_n$ actually being observed as N annual peaks should be a maximum. The likelihood function is written as :

$$L = p\ (x_1)\ p\ (x_2)......p\ (x_n) = \pi_{i=1}^{N}\ p\ (x_i) \tag{5.44}$$

$$\pi_{i=1}^{N}\ p\ (x_i) = (1/a)^N e^{-1/a \sum_{i=1}^{N} (x_i-u) - \sum_{i=1}^{N} e^{-(x_i-u)/a}} \tag{5.45}$$

The method of maximum likelihood then takes the logarithm of eq. 5.45, partially differentiates with respect to a and u and equates to zero :

$$\ln\ L = N \ln\ (1/\alpha) - (1/a) \sum_{i=1}^{N} (x_i-u) - \sum_{i=1}^{N} e^{-(x_i-u)/a} \tag{5.46}$$

$$\text{Let } 1/a = a' \tag{5.47}$$

Therefore eq. 5.46 gets modified as :

$$\ln\ L = N \ln \alpha' - a' \sum_{i=1}^{N} (x_i - u) - \sum_{i=1}^{N} e^{-a'(x_i - u)} \tag{5.48}$$

$$\frac{\partial\ (\ln L)}{\partial\ a'} = \frac{N}{\alpha'} - \sum_{i=1}^{N} (x_i - u) + \sum_{i=1}^{N} (x_i - u)\ e^{-a'(x_i - u)} \tag{5.49}$$

$$\frac{\partial\ (\ln L)}{\partial u} = Na' - a' \sum_{i=1}^{N} e^{-a'(x_i - u)} \tag{5.50}$$

Setting eq. 5.50 equal to zero

$$\sum_{i=1}^{N} e^{-a'(x_i-u)} = N \tag{5.51}$$

so that

$$e^{a'u} = N / \sum_{i=1}^{N} e^{-a' x_i} \tag{5.52}$$

or

$$u = 1/a' \ln\ (N/ \sum_{i=1}^{N} e^{-a' x_i}) \tag{6.53}$$

If the arithmetic mean of the series $x_i, x_i......x_n$ is denoted by μ, then eq. 5,49 can be written as :

$$\frac{\partial\,(\ln L)}{\partial\,\alpha'} = \frac{N}{\alpha'} - N\,(\mu - u) + e^{\alpha' u} \sum_{i=1}^{N} (x_i - u)\, e^{-\alpha' x_i} \tag{5.54}$$

Substituting for $e^{\alpha' u}$ from eq. 5.52

$$\frac{\partial\,(\ln L)}{\partial\alpha'} = \frac{N}{\alpha'} - N\,(\mu - u) + \frac{N \sum_{i=1}^{N} (x_i - u)\, e^{-\alpha' x_i}}{\sum_{i=1}^{N} e^{-\alpha' x_i}} \tag{5.55}$$

Equating this to zero and simplifying :

$$F\,(\alpha') = \sum_{i=1}^{N} x_i \;\; e^{-\alpha' x_i} - (\mu - 1/\alpha') \sum_{i=1}^{N} e^{-\alpha' x_i} = 0 \tag{5.56}$$

Eq. 5.56 can not be solved for α' explicitly and so the Taylor series expansion has been used by Panchang and Agarwal (1962). Writing

$$F(\alpha'_{j+1}) = F\,(\alpha'_j + h_j) \tag{5.57}$$

where, h_j is the increment in α at j^{th} iteration :

$$F(\alpha'_{j+1}) = F(\alpha'_j) + h_j \;\; F'(\alpha'_j) \tag{5.58}$$

where $F'\,(\alpha_j')$ is the first order derivative of $F\,(\alpha')$ with the respect to α' :

$$F'(\alpha') = -\sum_{i=1}^{N} x_i^2 \; e^{-\alpha' x_i} + (\mu - 1/\alpha') \sum_{i=1}^{N} x_i \, e^{-\alpha' x_i} - \frac{1}{(\alpha')^2} \sum_{i=1}^{N} e^{-\alpha' x_i} \tag{5.59}$$

and α'_j and α'_{j+1} are successive approximations to α'. The procedure adopted by Panchang and Aggarwal is to estimate α_j' from the method of moments. By evaluating $F\,(\alpha_j')$ and $F'(\alpha_j')$ from equations 5.56 and 5.59 then :

$$h_j = -\,F\,(\alpha_j)\,/\,F'\,(\alpha_j) \tag{5.60}$$

and, $\alpha'_{j+1} = \alpha_j + h_j$ (5.61)

This procedure is repeated until sufficiently small value of $F\,(\alpha_j')$ is obtained and then u can be obtained from eq. 5.53. In most cases only 3 or 4 steps will be required.

5.6 GENERAL COMMENTS ON THE MM AND MML ESTIMATES

The method of moments and the method of maximum likelihood do not always produce the same estimates for the parameters. The accuracy of the method of moments is severely affected if the data contains errors in the tails of the distribution where the moment arms are long. This is especially troublesome with highly skewed distributions.

Maximum likelihood estimators are sufficient and consistent. An estimator θ is said to be a sufficient estimator for θ if θ uses all the information relevent to θ that is contained in the sample. An estimator θ of a parameter θ is said to be consistent if the probability that θ differs from θ by more than an arbitrary constant ϵ approaches zero as the sample size approaches infinity. If an efficient estimator exists, maximum likelihood estimators, adjusted for bias. will be efficient. In view of the properties of the maximum likelihood estimators this method is generally preferred over the method of moments.

REFERENCES :

1. Chow, V.T., 1964. Handbook of Applied Hydrology, McGraw-Hill, New York.

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

3. Kimball, B.F., 1946. Sufficient Statistical Estimation Functions for the Parameters of the Distribution of Maximum Values, Ann. Maths. Statist., Vol. 17, No. 3.

4. Kite, G.W., 1977, Frequency and Risk Analysis in Hydrology, Water Resources Publications, Colorado.

5. National Environment Research Council, 1975, Flood studies Report-Vol. I, London.

6. Panchang, G.M. and Aggarwal, V.P., 1962, Peak Flow Estimation by Method of Maximum Likelihood, Tech. Memo. HLO2, Government of India, Central Water and Power Research Station, Poona, India.