LECTURE-7

DETERMINATION OF CONFIDENCE INTERVALS

OBJECTIVES

The objective of this lecture is to highlight the concept and the use of confidence interval or significance levels in statistical modelling of hydrologic variables.

7.1 INTRODUCTION

Hydrologic variables such as annual peak floods or rainfalls do not occur in a set pattern and are mostly random. In modelling these events, the help of frequency analysis is taken such that the estimate of these hydrological variables for a desired return period can be estimated with a reasonable accuracy. The estimates, usually, arrived from a single set of sample data are vaiable because of randomness associated with these events and the size of the sample used for arriving at the estimates.

Moreover, the sample under consideration is assumed to have resulted from a specific parent population and is random. This results in the fact that there are many equally likely possible samples that can originate from this assumed population. If estimates of the variables for all such sample for the desired return period are plotted against the return period, they seem to follow a normal or t-distribution with its mean as the expected value of the variables at that return period (Fig. 7.1). This therefore indicates that due to sampling variation there can be many estimates and therefore, should be defined through a continuous run of estimates rather than single or point value of the estimate. This range is defined as confidence interval and can be written as

$$\text{Prob}\ (x_L \leqslant x_T \leqslant x_{TU}) = 1-\alpha \qquad (7.1)$$

where x_{TL} and x_{TU} are lower and upper confidence limits of the estimate x_T so that the interval x_{TL} to x_{TU} is the confidence interval and $1-\alpha$ is the confidence level (α = significant level). This can also be graphically represented as in Figure 7.2.

However, the confidence level based on probability values give rise to the limits on either side of curve developed by frequency analysis to indicate the reliability of the estimates as well as the fit.

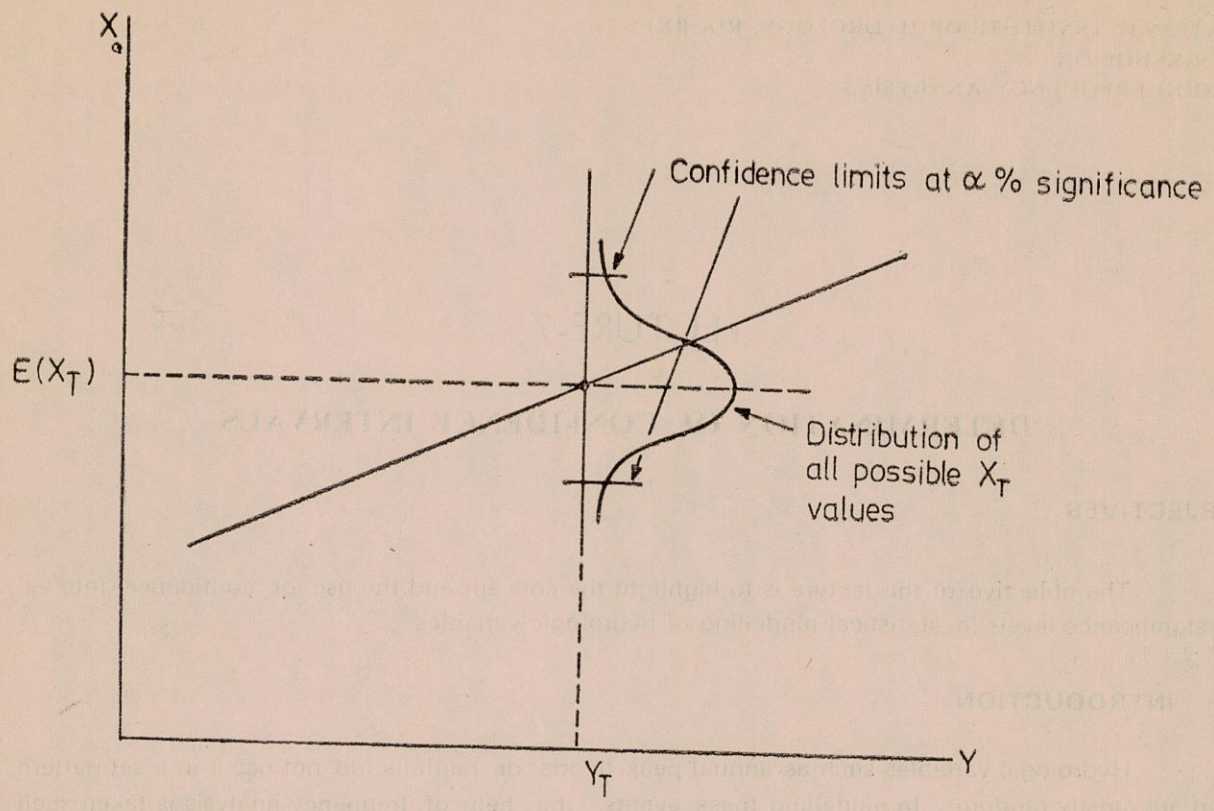

Fig. 7.1 Distribution of all possible quantiles for a given return period

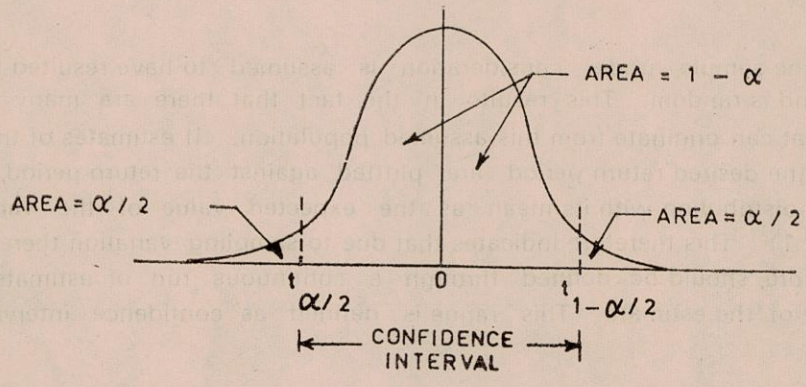

Fig. 7.2 Illustration of confidence intervals using the t – distribution

7.2 DEVELOPMENT OF CONFIDENCE BAND

The confidence limits are computed using the following steps :

(1) Choose a statistical distribution for modelling the annual peak flood data at the given station.

(2) Estimate the parameters.

(3) Compute the quantiles for the desired return period using the estimated parameters.

(4) Compute the standard error [$S_e(x_T)$] of the estimates as given in Sec. 7.3.

(5) Compute the t—statistics for the desired confidence level (i.e. $1-\alpha/2$, considering $\alpha/2\%$, significance on both sides) and (N-n) degrees of freedom where N = sample size and n = no. of parameters in the distribution selected.

(6) Compute the upper and lower confidence limits of the quantiles (x_T) as

$$x_{TU} = x_T + t_{(N-n)\ (1-\alpha/2)} \cdot S_e(x_T) \tag{7.2}$$

$$x_{TL} = x_T - t_{(N-n)\ (1-\alpha/2)} \cdot S_e(x_T) \tag{7.3}$$

(7) Plot them on either side of the plot of quantiles and join the points on the upper and lower region to give the confidence band.

7.3 STANDARD ERRORS AND CONFIDENCE LIMITS OF ESTIMATES

Computation of standard error of quantile estimates for some of the commonly used probability distributions as well as estimates from simple regression analysis are presented in this section.

7.3.1 Normal Distribution

The quantile x_T is written as :

$$x_T = \bar{x} + s\, y_T \tag{7.4}$$

where $\bar{x}$ = sample mean

s = sample stand and deviation

y_T = normal reduced variate for return period T.

Standard error of the quantile estimate is given by :

$$S_e(x_T) = \frac{\sigma}{\sqrt{N}} [1 + 0.5\, y_T^{\,2}]^{0.5} \tag{7.5}$$

The upper and lower confidence limit of the estimate x_T for a sample size of N and α % significance is given by :

$$x_{TU} = x_T + t_{(1-\alpha/2),\ (N-2)} \cdot S_e(x_T) \tag{7.6}$$

$$x_{TL} = x_T - t_{(1-\alpha/2),\ (N-2)} \cdot S_e\ (x_T) \qquad (7.7)$$

where, x_T = the quantile estimate for the return period T

y_T = reduced variate

$\bar{x}$ = mean of the sample

s = sample standard deviation

$S_e\ (x_T)$ = standard error of the quantile

7.3.2 Log-Normal Distribution

The quantile in the log-transormed domain is written as :

$$x'_T = \bar{x}' + s'\, y_T \qquad (7.8)$$

where, $\bar{x}'$ = mean of log transformed x_i' series

s' = sample standard deviation of the log-transformed series

Standard error of quantile estimate in log-transformed domain

$$S_e\ (x_T') = \frac{s'}{\sqrt{N}}\ [1 + 0.5\ y_T^{\,2}]^{0.5} \qquad (7.9)$$

The quantile estimate in natural domain is given by $x_T = e^{x_T'}$

The average standard error in natural domain is obtained from

$$S_e\ (x_T) = [x_T\ \{e^{S_e\ (x_T')} - 1\} - x_T\ \{e^{-S_e\ (x_T')} - 1\}]/2 \qquad (7.10)$$

The upper and lower confidence limits of the estimate x_T for α % significance level can be computed using Equations 7.6 and 7.7.

7.3.3 Gumbel (EV-1) Distribution

The quantile x_T is written as

$$x_T = u + \alpha\, y_T \qquad (7.11)$$

If the parameters u and α have been estimated by method of maximum likelihood then

$$S_e\ (x_T) = \frac{\alpha}{\sqrt{N}}\ (1.11 + 0.52\ y_T + 0.61\ y_T^{\,2})^{1/2} \qquad (7.12)$$

If the parameters have been estimated using the method of moments then

$$S_e(x_T) = \frac{\alpha}{\sqrt{N}} (1.170 + 0.196\, y_T + 1.099\, y_T^{\,2})^{0.5} \qquad (7.13)$$

The upper and lower confidence limits of the quantiles for α % significance level are given by Eqns 7.6 and 7.7 respectively.

7.3.4 Log Pearson Type III Distribution

The location and scale parameter $\bar{x}'$ and s' are computed from log-transformed x_i' series.

Frequency factor K_T which is dependent on the shape parameter is computed using Wilson-Hilferty Transformation as

$$K_T = \frac{2}{C_s}\left[1 + \frac{Y_T \cdot C_s}{6} - \frac{C_s^2}{36}\right]^3 - \frac{2}{C_s} \qquad (7.14)$$

where,

C_s = co-efficient of Skewness of x_i' series, and

Y_T = standard normal derivate corresponding return period T.

The quantile estimate (x_T') for any return period T is computed from $x_T' - x' = + s' \cdot K_T$ and its value x_T in natural domain is $x_T = \exp(x_T')$

Standard error estimate of x_T' is computed from

$$S_e(x_T') = \frac{s'}{\sqrt{N}}\left[1 + K_T \cdot C_s + \frac{K_T^{\,2}}{2}\left(\frac{3C_s^2}{4} + 1\right) + 3K_T \cdot v_T\left(C_s + \frac{C_s^3}{4}\right)\right.$$

$$\left. + 3v_T^{\,2}\left(2 + 3C_s^2 + \frac{5C_s^4}{8}\right)\right]^{1/2} \qquad (7.15)$$

where

$$v_T = \frac{Y_T - 1}{6} + \frac{4\,(Y_T^{\,3} - 6Y_T^{\,4})}{6^3} C_s - \frac{3\,(Y_T^{\,2} - 1)}{6^3} C_s^2 + \frac{4Y_T}{6^4} C_s^3$$

$$- \frac{10}{6^6} C_s^4 \qquad (7.16)$$

Y_T = standard normal deviate corresponding to return period of T years.

Average standard error $S_e(x_T)$ in linear units

$$= \frac{x_T\left[e^{S_e(x_T')} - e^{-S_e(x_T')}\right]}{2} \qquad (7.17)$$

Confidence bounds at $\alpha\%$ significance level are calculated from

$$x_{TU} = x_T + t_{(1-\alpha/2),\,(N-3)} \cdot S_e(x_T) \tag{7.18}$$

$$x_{TL} = x_T - t_{(1-\alpha/2),\,(N-3)} \cdot S_e(x_T) \tag{7.19}$$

7.3.5 Least Squares Estimates :

If dependent variables Y_i are related to independent variables X_i by the regression equation with two parameters a and b;

$$Y_i = a + b\,X_i \tag{7.20}$$

Then standard error of an estimate Y_K for a given value of X_K may be calculated as

$$S_e(Y_K) = s\left(\frac{1}{N} + x_k^2 / \Sigma x_i^2\right)^{1/2} \tag{7.21}$$

where, s = sample standard deviation; N is the sample size, and

$$x_k = X_K - \overline{X} \tag{7.22}$$

The confidence limits on the regression line are given by

$$Y_{KU} = Y_K + t_{(1-\alpha/2),\,(N-2)} \cdot S_e(Y_K) \tag{7.23}$$

$$Y_{KL} = Y_K - t_{(1-\alpha/2),\,(N-2)} \cdot S_e(Y_K) \tag{7.24}$$

7.4 SHAPE OF CONFIDENCE BAND

The upper and lower cenfidence limits computed as per Sec 7.3 when plotted against various return periods in case of frequency analysis or against independant variables in case of regression analysis show the minimum difference near the mean values with a diverging trend away from it.

The interval between them for a particular return period or an independent variable however increases with the decrease of sample size. This is mainly because of large sampling variance and hence in large standard error.

7.5 LIMITATIONS

In hydrology, the use of confidence interval is mainly limited to graphical frequency analysis and regression analysis. In graphical frequency analysis these limits are useful in ascertaining the choice of proper distribution i.e. narrower the interval the better is the choice of that distribution.

It can also be used to ascertain the proper fit of a frequency curve to the observed data, to the extent whether the fit which has been done by eye estimation is acceptable or not.

The upper confidence limit is sometimes used to arrive at conservative estimates of design parameters.

7.6 EXAMPLES

Example 1 : Compute the confidence limits at 5% significance for the 100 year and 1000 year design flood computed from 20 years of annual peak flood data at station D having a mean annual peak flood of 200 cumecs ($\bar{x}$) and a standard deviation of 20 cumecs (s).

Given N = 20, $\bar{x}$ = 200 cumecs and s = 20 cumecs

Assuming the data to follow the Gumbel distribution, the parameters u and α can be computed by the method of moments as :

α = scale parameter = s/1.28

u = location parameter = $\bar{x} - 0.5772\,\alpha$

α = 20/1.28 = 15.625

u = 200 — 0.5772 × 15.625 = 190.98

For T = 100 years

reduced variate $y_T = -\ln [-\ln (1 - 1/T)]$

= 4.60

Using Eqn. 7.11 the design flood for T = 100 years

X_{100} = 200 + 15.625 × 4.60

= 271.87 Say 272 cumecs

Using Eqn. 7.13 standard error of X_{100}

$$S_e (X_{100}) = \frac{15.625}{\sqrt{20}} (1.170 + 0.196 \times 4.6 + 1.099 \times 4.6^2)^{0.5}$$

= 17.591 cumecs

Since the significance level α = 5 % then $t_{(1-\alpha/2),\,(N-2)}$

[from the t-distribution Table in Appendix V]

= 2.101

Using Eqns. 7.6 and 7.7, the upper and lower confidence limits are computed as :

$$X_{TU} = 272 + 2.101 \times 17.591$$

$$= 308.95 \text{ Say } 310 \text{ cumecs}$$

$$X_{TL} = 272 - 2.101 \times 17.591$$

$$= 235.04 \text{ Say } 235 \text{ cumecs}$$

Similarly for T = 1000 years or $y_T = 6.91$

$X_{1000} = 307.97$ Say 308 cumecs

$S_e (X_{1000}) = 25.92$ cumecs

and the upper and lower confidence limits for X_{1000}

$$X_{TU} = 362.46 \text{ Say } 365 \text{ cumecs}$$

$$X_{TL} = 253.54 \text{ Say } 253 \text{ cumecs}$$

Summary

	T = 100 years	T = 1000 years
Estimate	272 cumecs	308 cumecs
Standard Error	17.59 cumecs	25.92 cumecs
X_{TU}	310 cumecs	365 cumecs
X_{TL}	235 cumecs	253 cumecs

It is evident from the results that the confidence interval broadens from 75 to 112 cumecs with change of estimates from T = 100 years to T = 1000 years.

Example 2 : Compute the effect of sample size on the confidence limits of 1000 years quantile estimate for data given in Example 1.

As already computed in Example 1 for a sample size of 20,

$X_{1000} = 308$ cumecs

$X_{TU} = 365$ cumecs

$X_{TL} = 253$ cumecs

Using the same parameters and assuming a sample size of 10.

$X_{1000} = 308$ cumecs

$S_e (X_{1000}) = 36.64$ cumecs

The upper and lower confidence limits for X_{1000} with sample size of 10.

($t_{0.075, 8} = 2.306$ vide Appendix V)

$X_{TU} = 308 + 2.306 \times 36.64 = 392.5$ cumecs

Say 393 cumecs

$X_{TL} = 308 - 2.306 \times 36.64 = 223.5$ cumecs

Say 223 cumecs

Summary

T = 1000 years	N = 20	N = 10
Estimate	308 cumecs	308 cumecs
Standard Error	25.92 cumecs	36.64 cumecs
X_{TU}	365 cumecs	393 cumecs
X_{TL}	253 cumecs	223 cumecs

It is evident from the above results, that the confidence interval increases from 112 cumecs to 170 cumecs with the decrease in sample size from 20 to 10 under other similar conditions.

Bibliography

1. Hean, C.T. 1977. Statistical mathods in Hydrology, Iowa State University Press, Ames., U.S.A.

2. Kite, G.W. 1977. Frequency and Risk Analysis in Hydrology, Water Resources Publications, Colorado 80161.

3. Parida, B.P., S·K. Jain and D. Chalisgaonkar. 1987. Graphical Represention of Information related with Floods, National Institute of Hydrology, Report No. UM-22, pp- 21.