LECTURE - 4

FREQUENCY ANALYSIS BY GRAPHICAL TECHNIQUES

OBJECTIVES

The objective of this lecture is to explain theoretical basis behind the development of probability papers, in general, and present the procedures for the development of probability papers for normal, log normal, Gumbel EV–I, Pearson type–III and log Pearson type–III distributions. The unbiased plotting position formulae corresponding to various well known distributions used in practice have also been identified.

4.1 INTRODUCTION

A primary purpose in plotting a set of ordered observations on probability paper is to simplify the inspection of their distribution. Probability paper for a given cumulative distribution P (x) of a random variable X say, the annual peak flood series, is so designed that a plot of P (x) against X is a straight line. If a set of data follows a straight line on probability paper then the data is said to follow the distribution based on which the probability paper has been developed. Thus the process of extrapolating the probability plot to find the flood magnitudes of desired higher return period is simplified since straight line can be extended. As we will see later the probability paper can be prepared for any distribution based only on two parameters of the distribution. Any additional parameter such as coefficient of skewness must be constant. This is the reason why we find the probability papers for normal (log normal) and Gumbel extreme value type-I distribution are commercially available. As the Pearson type III distribution (log Pearson type III) has got the coefficient of skewness varying as a third parameter, in addition to mean and standard deviation, we generally don't find its probability paper commercially available. Thus there exists a unique probability paper corresponding to each value of skewness.

The aim of this lecture is to explain how to construct a probability paper for the purpose of frequency analysis. Further this lecture also deals with plotting position formulae required for frequency analysis by graphical techniques.

The frequency distributions of normal (log normal), Gumbel extreme value type-I and Pearson type III (log Pearson type III) distributions which are most commonly used for flood frequency analysis would be considered for graphical analysis.

Before studying the details of probability paper development, it is essential to survey the different plotting position formulae available for plotting the data on a probability paper.

4.2 PLOTTING POSITION

Determining the probability to assign to a data point is commonly referred to as determining the plotting position. Probability plotting of hydrologic data requires that individual observations or data points be independent of each other and that the sample data be representative of the population. Four common types of sample data can be classified, viz., complete duration series, annual series, partial duration series and extreme value series. The complete duration series consists of all available data. An example would be all the available daily flow data for a stream. However this particular data set would most likely not have independent observations and thus renders itself unfit for frequency analysis. The annual series consists of one value per year such as the maximum peak flow each year. The partial duration series consists of all values above a certain base. The extreme value series consists of the largest observations in a given time interval. The annual series is a special case of the extreme value series with the time interval being one year. Regardless of the type of sample data used, the plotting position can be determined in the same manner.

Various plotting position formulae have been recommended for assigning probability to a given set of flood peak data by different investigators from time to time. Table-1 shows a list of such formulae, some of which are used in current practice and some have become obsolete.

TABLE—I

VARIOUS PLOTTING POSITION FORMULAE

S. No.	Plotting position formula	Year	$P(X \geqslant x)$	Values of a and b in the formula $P(X \geqslant x) = \frac{(m - a)}{(N + b)}$	Value of a in the formula $P(X \geqslant x) = \frac{m - a}{N + 1 - 2a}$	Value of P (X⩾x) and T for m=1 and N=10	
						$P(X \geqslant x)$	$T = \frac{1}{P(X \geqslant x)}$
1.	California	1923	m/N	a=0.0, b=0.0	—	0.1	10
2.	Hazen	1930	(m-0.5)/N	a=0.5, b=0.0	0.5	0.05	20.0
3.	Weibull	1939	m/(N+1)	a=0.0, b=1.0	0.0	0.091	11.0
4.	Beard	1953	(m-0.31)/(N+0.38)	a=0.31, b=0.38	0.31	0.066	15.0
5.	Chegodajew	1955	(m-0.3)/(N+0.4)	a=0.3, b=0.4	0.3	0.067	14.9
6.	Gringorton	1963	(m-0.44)/(N+0.12)	a=0.44, b=0,12	0.44	0.055	18.1
7.	Blom	1958	(m-3/8)/(N+1/4)	a=3/8, b=1/4	3/8	0.0605	16.4
8.	Tukey	1962	(m-1/3)/(N+1/3)	a=1/3, b=1/3	1/3	0.0645	15.5
9.	Benard	1953	(m-0.3)/(N+0.2)	a=0.3, b=0.2	—	0.068	14.7
10.	Cunnane	1978	(m-0.4)/(N+0.2)	a=0.4, b=0.2	0,4	0.053	17
11.	Adamowski	1981	(m-0.25)/(N+0.5)	a=0.25, b=0.5	0.25	0.071	14.0

Gumbel (1958) states the following criteria for plotting position relationship :

1. The plotting position must be such that all observations can be plotted.
2. The plotting position should lie between the observed frequencies of $(m-1)/n$ and m/n where m is the rank of the observation beginning with $m = 1$ for the largest value and n is the number of years of record (if applicable) or the number of observations.
3. The return period of a value equal to or larger than the largest observation should approach n, the number of observations.
4. The observations should be equally spaced on the frequency scale (i.e.) the difference between the plotting positions of the $(m+1)^{th}$ and the m^{th} observations should be a function of n only and be independent of m.
5. The plotting position should have an intuitive meaning and should be analytically simple.

One of the most commonly used plotting position is due to Weibull. The form of Weibull's plotting position formula is given as :

$$P\,(X \leqslant x) = \frac{m}{n+1} \tag{4.1}$$

where, P (x) is the probability of non-exceedance of a given event x when the data are ranked from the smallest $(m = 1)$ to the largest $(m = n)$ in ascending order. However the same plotting position formula gives the probability of exceedance of a given event when the data are ranked from the largest $(m = 1)$ to the smallest $(m = n)$ in descending order.

The Weibull plotting position formula meets all the requirements specified by Gumbel as : i) all of the observations can be plotted since the plotting positions range from $\frac{1}{(n+1)}$ which is greater than zero to $\frac{n}{(n+1)}$ which is less than one; 2) the relationship $\frac{m}{n+1}$ lies between $(m-1)/n$ and m/n for all values of m and n; 3) the return period of largest value is $(n+1)/1$ which approaches n as n gets large and the return period of the smallest value is $(n+1)/n$ which approaches 1 as n gets large; 4) the difference between the plotting position of the $(m+1)$ and mth value is $1/(n+1)$ for all values of m and n; and 5) the fact that condition 3 is met plus the simplicity of weibull relationship fulfill condition 5.

Benson (1962) and Chow (1964) theoretically justify the use of Weibull's plotting position formula for annual maximum and partial duration series.

Although Weibull's plotting position is widely used in our country and U.S.A., its use for graphical frequency analysis has been discouraged by Cunnane (1978) who had extensively studied various plotting position formulae based on theoretical consideration. He argued against some of the requirements of plotting position formula put forward by Gumbel (1958) and indicated that these requirements were responsible for the adoption of Weibull's formula in practice even today. His views on those requirements put forward by Gumbel are given below :—

Requirement 1 : This is necessary

Requirement 2 : This is most misleading and also appears to have played a major role in the adoption of Weibull formula.

Requirement 3 : This condition is neither necessary nor desirable

Requirement 4 : This condition is also not necessary

Requirement 5 : Although desirable it is not reconciable with any mathematical derivation, and consequently can play no part in the rational development of a formula.

Cunnane concluded that the use of Weibull's plolting position formula leads to biased estimate and recommended the use of following unbiased plotting position formulae for fitting various distributions. A brief discription of these unbiased plotting position formulae is given below :

4.2.1. Unbiased Plotting Position Formulae

For normal and log normal distributions :

Cunnane recommends the use of Blom's (Blom, 1958) plotting position formula for plotting the data in a normal or log normal probability paper. The Blom's plotting position formula which gives the probability of non-exceedance of an event is given as :

$$P\ (X \leqslant x) = \frac{m - 0.375}{n + 0.25} \quad m = 1,2,\ldots\ldots\ldots n \tag{4.2}$$

in which, m = 1 corresponds to the smallest observation.

The Blom's plotting position formula is proved to be unbiased i.e., the average value of an ordered event, say x_i, considered over many number of samples of the same size would lie on the population line when plotted against the Blom's plotting position value.

For Gumbel's EV-type I distribution :

Cunnane recommends the use of Gringorton plotting position formula (Gringorton, 1963) for plotting the data in a Gumbel probability paper. It is proved to be unbiased and it is written for probability of non-exceedance as :

$$P\ (X \leqslant x) = \frac{m - 0.44}{n + 0.12} \quad m=1,2,\ \ldots\ldots\ n \tag{4.3}$$

where m = 1 for the smallest observation.

For Pearson type III and log pearson type III distributions ;

The plotting position formula for this distribution, strictly speaking, depends on the skewness as it is characterised by three parameters. As it would be cumbersome to deal with various plotting positions corresponding to each coefficient of skewness, Cunnane has recommended a single plotting position formula regardless of the coefficient of skewness. It is given as :

$$P\ (X \leqslant x) = \frac{m - 0.4}{n + 0.2} \quad \ldots \tag{4.4}$$

It is important to note that all the above recommended plotting position formulae are of the form,

$$P\ (X \leqslant x) = \frac{m - a}{n + 1 - 2a} \quad 0 \leqslant a \leqslant 1 \quad \ldots \tag{4.5}$$

4.3 PREPARATION OF PROBABILITY PAPER

The probability paper can be developed for the required distribution based on a linear relationship between the discharge and the variable characteristic of the distribution. The form of the such relationship is given as :

$$x_i = a + by \tag{4.6}$$

in which,

x_i = the discharge

y = the reduced variate of the distribution to be plotted.

Eq. 4.6 indicates that a probability paper can be developed using an ordinary graph paper. Indeed the probability paper is developed by taking linear scale on abscissa corresponding to the reduced variate y and then assign the probability of exceedance or non-exceedance or return period at intervals given as in the commercial probability paper corresponding to the appropriate reduced variate. Tne methods of developing the probability paper for normal, log normal, Gumbel EV–type I, Pearson type III and log pearson type III distributions are explained below :

4.3.1 Development of Normal Probability Paper

The normal or Gaussian distribution is described by the probability density function as :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)/\sigma^2} \quad \text{... (4.7)}$$

in which,

x = the variate, say the peak discharge

μ = the population mean of x

σ = the population standard deviation of x.

The probability of non-exceedance of x is given by the distribution function, F (x) as :

$$P(X \leqslant x) = F(x) = \int_{-\infty}^{x} f(x)\,dx \quad \text{... (4.8)}$$

However eq. 4.8 cannot be expressed in the form of any simple mathematical function, when f (x) is substituted from eq. 4.7. Hence for solutions of practical problems eq. 4.8 has to be solved using the table given in Appendix. If access to computer is available, one can use the standard subroutines like that of NDTRI of IBM scientific subroutine package (1968) which gives the reduced variate of normal distribution corresponding to the given probability of non-exceedance. Eq. 4.8 can be modified to the standard form based on which the relationship between the area under the normal curve and the standard normal deviate has been formulated.

Eq. 4.8 is modified by inserting eq. 4.7 as :

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)/\sigma^2}\,dx \quad \text{... (4.9)}$$

Substituting $\frac{(x-\mu)}{\sigma} = y$, and inserting it in eq. 4.9 leads to :

$$F(x) = G(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \quad \text{... (4.10)}$$

The above integral is the improper integral and corresponding to a given y, eq. 4.10 gives the non-exceedance probability. The variable y is the standard normal deviate or the reduced variate. There is a unique relationship exists between y and G (y), therefore with F (x) Appendix– shows this relationship. From the standard normal deviate relationship it can be seen that the variable x, say, the flood peak is related to y in a linear form and it is given as :

$$X = \mu + \sigma y \quad \text{... (4.11)}$$

It is this relationship which is used for developing a normal or log normal probability paper.

In developing the normal probability paper the following steps are followed :

1. Take an ordinary graph paper.
2. Decide the required scale of abscissa for marking the y_i, the standard normal deviate.
3. Mark the value of $y = 0$ at the center of the abscissa. The values of $y < 0$ are marked on the left of this and $Y > 0$ are marked on the right of this at equidistant intervals. It is sufficient to take the range of y between — 4.0 and 4.0.
4. Now the graph is ready for probability plotting. However to develop it as commercially available, the abscissa is marked with the probability of non-exceedance or exceedance at the required intervals corresponding to the appropriate value of y.

The development of log normal probability paper is same in all aspects as that of normal probability paper, except the ordinate is marked in log-scale i.e., log normal probability paper is developed on a semi-log paper having the ordinate in log-scale. The commercially available normal and log normal probability papers are shown in figures 4.1 and 4.2 respectively.

4.3.2 Development of Gumbel EV Type-I Distribution Probability Paper

In developing the Gumbel's probability paper again the use of eq. 4.6 is made. The form of eq. 4.6 is derived from Gumbel's EV type-I distribution as follows :

The Gumbel's EV type-I distribution is described by the probability density function as :

$$f(x) = \frac{1}{\alpha} e^{-(x-u)/\alpha - e^{-(x-u)/\alpha}} \quad (4.12)$$

in which,

x = the variate under study.

u and α are the parameters of the distribution.

The probability of non-exceedance of x is given by the distribution function F (x) as :

$$F(x) = \int_{-\infty}^{x} \frac{1}{\alpha} e^{-(x-u)/\alpha - e^{-(x-u)/\alpha}} dx$$

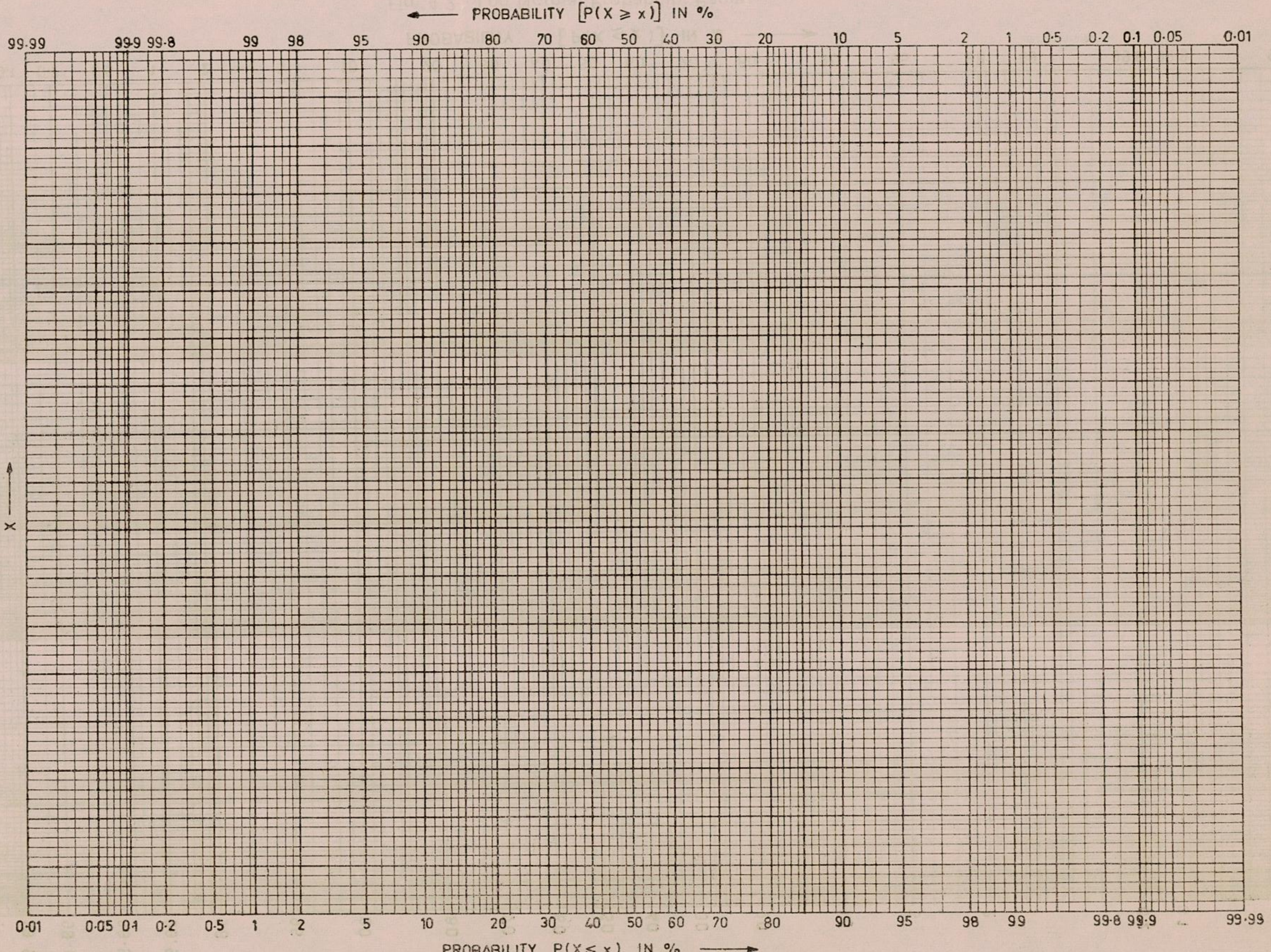

Fig. 4.1 Normal Probability Paper

(L-4/7)

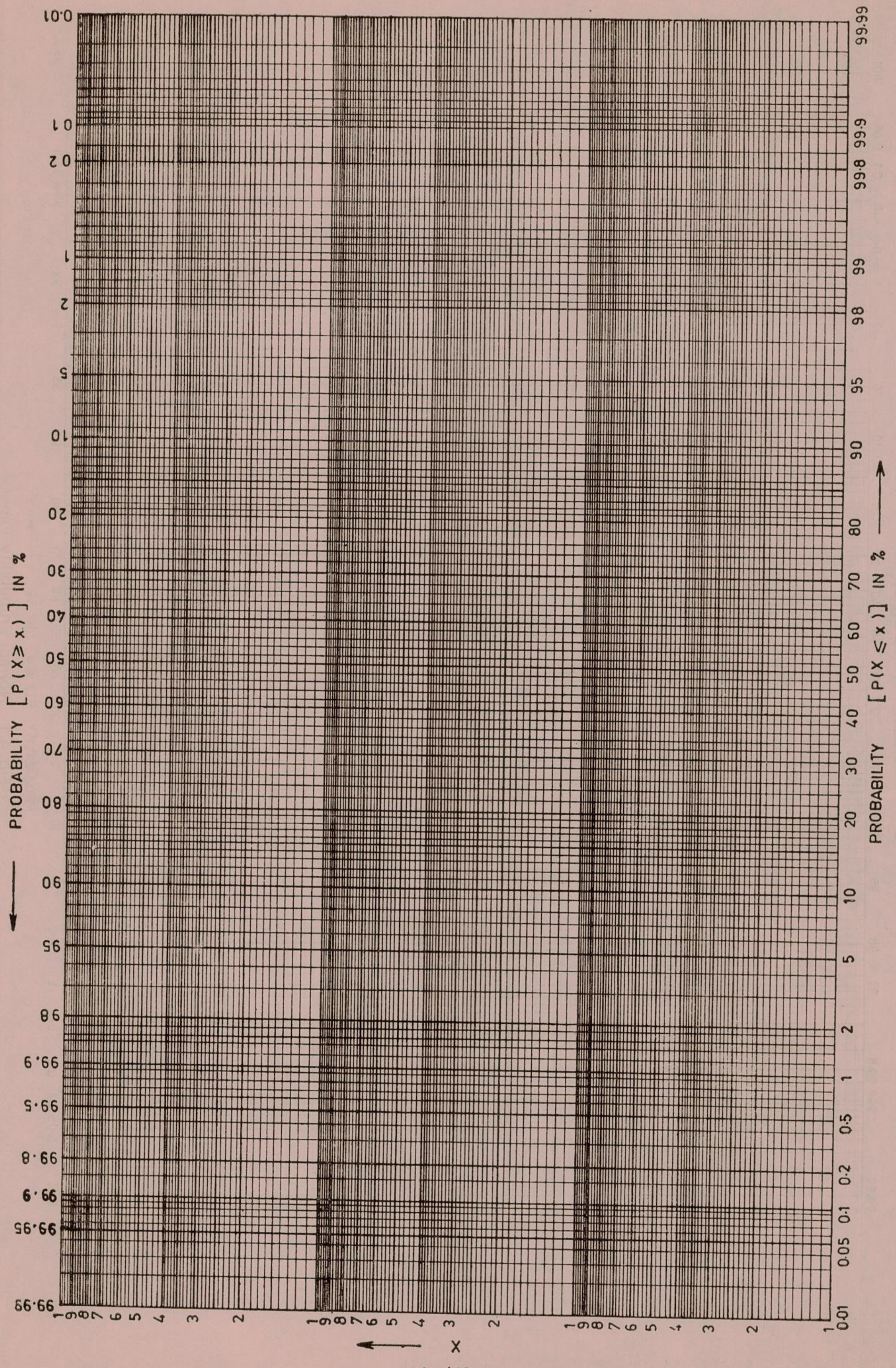

Fig. 4.2 Log Normal Probability Paper

Substituting $\frac{x-u}{\alpha} = y$ and inserting it in eq. 4.13 leads to :

$$F(x) = G(y) = \int_{-\infty}^{y} e^{-y-e^{-y}} dy \qquad \text{........ (4.14)}$$

$$= e^{-e^{-y}} \qquad \text{........ (4.15)}$$

G (y) can also be expressed in terms of return period corresponding to the variate x as :

$$G(y) = 1 - \frac{1}{T} \qquad (4.16)$$

From eq. 4.15 and eq. 4.16 one gets :

$$y = -\ln\left(-\ln\left(1 - \frac{1}{T}\right)\right) \qquad (4.17)$$

Eq. 4.17 defines the relationship between the Gumbel's reduced variate y and the corresponding return period, T.

In developing the Gumbel's probability paper using the reduced variate, y the following steps are followed :

1. Take an ordinary graph paper.
2. Decide the required scale on abscissa for plotting the y, the Gumbel's reduced variate.
3. Mark the value of y at regular intervals on abscissa in the range between — 3 and + 10.
4. Now the Gumbel's probability paper is developed. However to develop it as commercially available, the abscissa is marked with the probability of non-exceedance or exceedance or return period at the required intervals corresponding to the appropriate value of y as defined by eq. 4.17.

4.3.3 Probability Plot Based on Frequency Factor

Chow (1964) proposed a general equation for relating the flood magnitude with the mean and standard deviation of the data series. The form of this general equation is given as :

$$X = \mu + \sigma K \qquad (4.18)$$

in which μ and σ are the mean and standard deviation of the population series and K is called the frequency factor. As we have only the sample data set, the population mean and standard deviation are replaoed by sample mean and sample standard deviation respectively.

The meaning of K is that it is the variate value of return period T in a x distribution having the same form as that of x but which has zero mean and unit variance. In hydrology eq. 4.18 has been considered as the prototype to be copied for use with all distributions.

The form of the relationship used for developing normal probability paper is same (eq. 4.11) as this relationship. In the case of normal distribution the parameters of the distribution are same as that of statistical estimates viz., mean and standard deviation derived from the data, and therefore the reduced variate y is same as that of frequency factor K.

But in the case of Gumbel EV type-I distribution, the parameters of the distribution are not the same as that of the statistical estimates, viz., mean and standard deviation, and therefore $K \neq y$. Nevertheless the same form as given in eq. 4.18 can be deduced from the reduced variate relationship which is given as :

$$x = u + \alpha y \tag{4.19}$$

For this purpose the relationship between the parameters of the distribution and that of the mean and standard deviation of the data has to be established. The derivation of such relationship is described below :

Taking expectation of eq. 4.19, we get

$$E(x_i) = u + \alpha E(y) \tag{4.20}$$

in which,

E (y) is the notation for expectation.

The term $E(x_i)$ is the population mean of x_i and E (y) is given as :

$$E(y) = \int_0^{\infty} y\, e^{-y-e^{-y}}\, dy \tag{4.21}$$

$$= 0.5772, \text{ the Euler's constant.}$$

Therefore eq. 4.20 reduces to :

$$\mu = u + 0.5772\,\alpha \tag{4.22}$$

The variance of Eq. 4.19 is given as :

$$E(x - E(x))^2 = \alpha^2 E(y - 0.5772)^2 \tag{4.23}$$

The term $E(y-0.5772)^2$ is given as :

$$E(y - 0.5772)^2 = \frac{\pi^2}{6} \tag{4.24}$$

Therefore eq. 4.23 simplifies to :

$$\sigma^2 = \alpha^2 \frac{\pi^2}{6} \tag{4.25}$$

and

$$\sigma = \alpha \frac{\pi}{\sqrt{6}} \tag{4.26}$$

Now the distribution parameter values given in eq. 4.19 can be replaced by eqs. 4.22 and 4.26 to yield :

$$x = \mu + \frac{\sqrt{6}}{\pi}(-0.5772 + y)\,\sigma \tag{4.27}$$

The expression $\frac{\sqrt{6}}{\pi}(-0.5772 + y)$ is the required frequency factor K. The frequency factor K is rewritten as

$$K = -[0.45 + 0.7797 \ln(-\ln(1 - 1/T))] \tag{4.28}$$

Therefore for developing the Gumbel's probability paper either the form of relationship given by eq. 4.19 or by that eq. 4.27 can be used. The commercially available Gumbel probability paper is shown in Fig. 4.3.

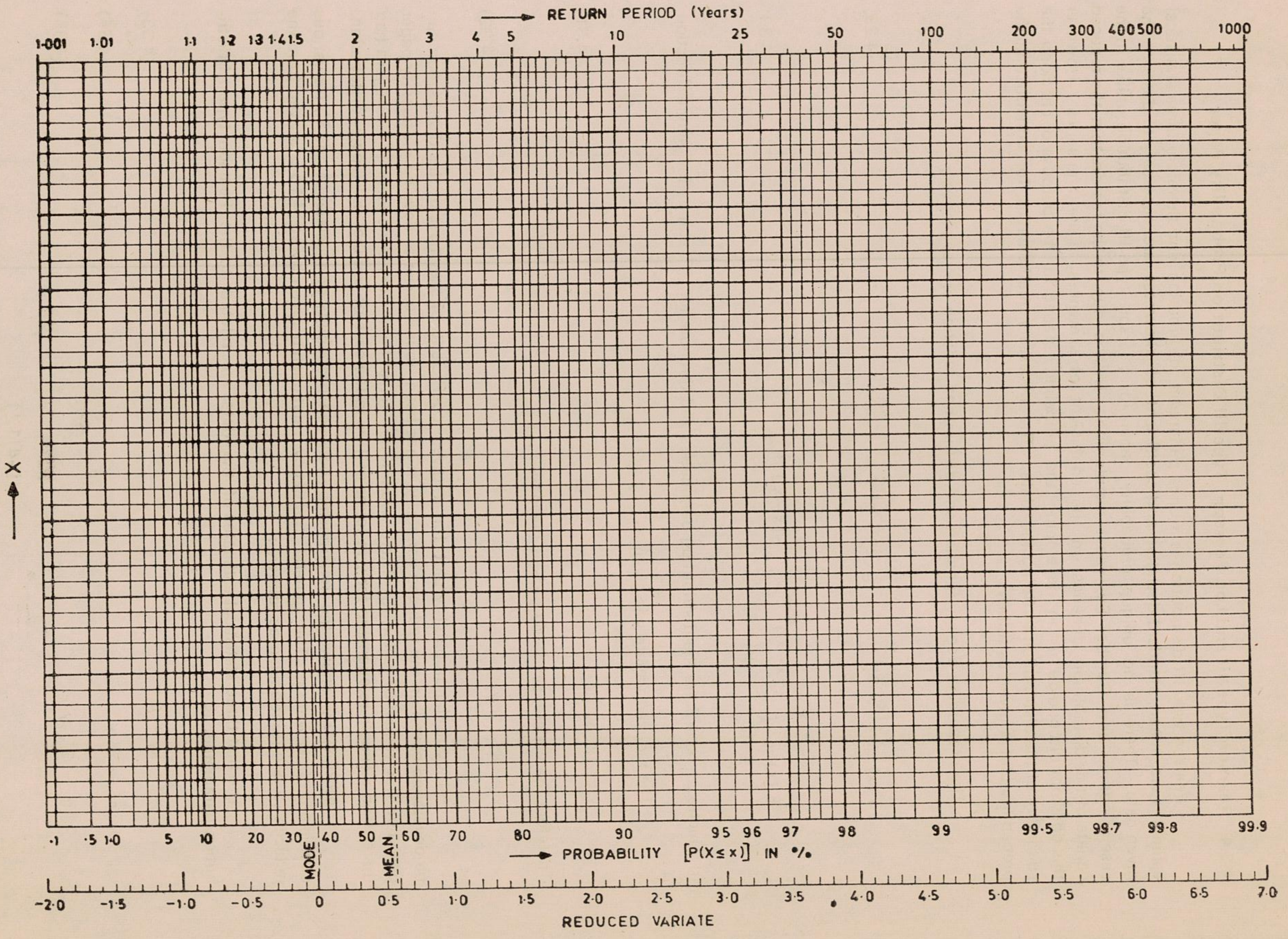

Fig. 4.3 Gumbel Probability Paper

4.3.4 Pearson Type-III and Log Pearson Type-III Distribution Probability Paper

Pearson type III distribution is a three parameter distribution and therefore three statistical estimates viz., mean, standard deviation and coefficient of skewnsss are necessary to describe the given data. There is no unique probability paper for Pearson type-III distribution unlike in the case of normal and Gumbel EV type-I distribution. This is due to the reason that the reduced variate of this distribution depends on the coefficient of skewness of the data series. However this does not discourage us in developing the probability paper for the purpose of checking the suitability of the distribution in fitting the data. The necessary mathematics for constructing the probability plot is given as follows :

The probability density function and distribution function are respectively :

$$f(x) = \frac{(x-x_0)^{\gamma-1}}{\beta^{\gamma}\,\Gamma\gamma}\, e^{-(x-x_0)/\beta} \qquad (4.29)$$

$$F(x) = \int_{x_0}^{x} \frac{(x-x_0)^{\gamma-1}}{\beta^{\gamma}\,\Gamma\gamma}\, e^{-(x-x_0)/\beta}\, dx \qquad \ldots (4.30)$$

in which x_0, γ and β are the distribution parameters and the notation 'Γ' stands for gamma function.

The reduced variate y is given as :

$$y = (x-x_0)/\beta \qquad \ldots (4.31)$$

Therefore the variate, say flood peak is given as :

$$x = x_0 + \beta\, y \qquad \ldots (4.32)$$

Although the probability plot for a given coefficient of skewness of this distribution can be developed based on the reduced variate, y, it has been found easier to develop probability paper based on the frequency factor, K. There would not have been any preference in this regard had there been a simple relationship for y in terms of T as in the case of Gumbel EV type-I distribution.

APPENDIX—shows the relationship between probability of exceedance and K, for various coefficient of skewness. The general frequency relationship given by eq. 4.18 is arrived from the reduced variate relationship as follows :—

The first three statistical estimates of the data are related to the Pearson type-III distribution parameters as :

$$\mu = x_0 + \beta\gamma \qquad \ldots (4.33)$$

$$\sigma^2 = \beta^2\gamma \qquad \ldots (4.34)$$

$$g = 2/\sqrt{\gamma} \qquad \ldots (4.35)$$

in which, g is the coefficient of skewness. Substituting these relationships in eq. 4.32,

$$x = \mu + \sigma \left(- \frac{2}{g} + \frac{gy}{2} \right) \qquad \text{... (4.36)}$$

Therefore the frequency factor is given as :

$$K = \frac{gy}{2} - \frac{2}{g} \qquad \text{... (4.37)}$$

Taking K values on the abscissa and the flood values on the ordinate of the ordinary graph paper, one can develop the Pearson type-III probability paper for a given coefficient of skewness.

4.4 REMARKS ON PROBABILITY PLOT

In the past probability plots played a central part in hydrological investigations. Since then the scope of hydrological modelling has widened enormously and the statistical methods employed have become more sophisticated with the result that probability plots have faded into background. However, few hydrologists who have to make engineering decisions which are based on analytical fitting of distributions of data would do so without the use of a graphical display.

Graphical plotting enables one to assess the suitability of the distribution selected for fitting the data in hand. Further, it enables one to assess the presence of 'outliers' in the data which, when plotted in the desired probability plot, appear to be from a different population because they plot a off of the line defined by the other points. Thus the probability plot is still a necessary part of hydrological practice.

Of course, the correct use of probability plot is important. The use of incorrect plotting position formula may distort the inference regarding the actual distribution followed by the given set of data. Therefore, the use of unbiased plotting position formula is essential. Further different hydrologist may arrive at different quantile estimates when there exists subjectivity in drawing the population line through these plotting points.

References :

1. Benson, M.A. (1962), 'Evolution of Methods for Evaluating the Occurrence of Floods', Water Supply Paper 1580-A, U.S. Geological Survey, Washington, D.C.

2. Blom, G. (1958), 'Statistical Estimates and Transformed Beta Variables', John Wiley, New York.

3. Chow, V.T. (1964), 'Handbook of Applied Hydrology', McGraw Hill, New York.

4. Cunnane, C. (1978), 'Unbiassed Plotting Positions—A Review', Journal of Hydrology Vol. 37, pp, 205-222.

5. Gringorton, I.I. (1963), 'A plotting Rule for Extreme Probability Paper', Journal of Geophysical Research, 68, 813–814.

6. Gumbel, E.J. (1958), 'Statistics of Extreme Values', Columbia University Press', New York.

7. Haan, C.T. (1977), 'Statistical Methods in Hydrology', The Iowa State University Press, Ames, Iowa.

8, IBM, (1968), 'System/360 Scientific subroutine Package (360 A-CM-03X) Version III', Programmer's Manual, White Plains, New York.

9. National Environment Research Council (1975), 'Flood Studies Report–Vol. 1', London.