R.J. Oosterbaan

International Institute for Land Reclamation and Improvement (ILRI)
P.O. Box 45, 6700 AA Wageningen, The Netherlands.

FREQUENCY PREDICTIONS AND THEIR BINOMIAL CONFIDENCE LIMITS

ABSTRACT

Graphs are presented by which the 80, 90 and 95% confidence intervals of a frequency or return period, found from frequency distributions based on records of limited size, can be determined independently of the type or distribution used. The graphs are constructed from binomial probability distributions

Application of the binomial confidence intervals and the resulting confidence belts shows that the kind of frequency distribution (e.g the exponential, log-normal and Gumbel distribution) used for the prediction of frequencies of extreme values is relatively immaterial. The same holds for the method (e.g. the plotting and parametric method) by which the distribution is determined.

RESUME

Dans les graphiques presentés les intervalles de fiabilité d'une fréquence ou d'une période de retour, trouvée à partir des distributions de fréquence basées sur une série de données limitée, peuvent être déterminés indépendamment du type de distribution utilisé. Les graphiques sont construits à partir de distributions binômiales.

L'application des intervalles binômiales de fiabilité et les zones de fiabilité résultantes montre que le type de distribution (p.ex. l'exposant, le log-normal ou Gumbel) utilisé pour la prédiction des fréquences de valeurs extrêmes est d'importance relative. Ceci est également valable pour la méthode (p. ex. la méthode graphique ou paramétrique) avec laquelle on ajuste la distribution.

1. INTRODUCTION

Frequency predictions from records of rainfalls, riverstages or floodlevels need to be accompanied by confidence statements, because the predictions can be quite insecure, especially when the records are short or when extreme values are considered.

For example, Figure 1 (after Benson 1960) shows the various frequency distributions obtained from different samples, each with 50 observations taken randomly from 1,000 values obeying a fixed distribution (base curve). The figure makes it clear that each sample yields a different curve. The highest and lowest curves indicate a confidence belt of the sample curves.

In this article, graphs are given for the estimation of confidence belts for frequency distributions obtained from a record of limited size. These belts are based on binomial probability distributions and indicate the area in which the (unknown) base curve may be situated.

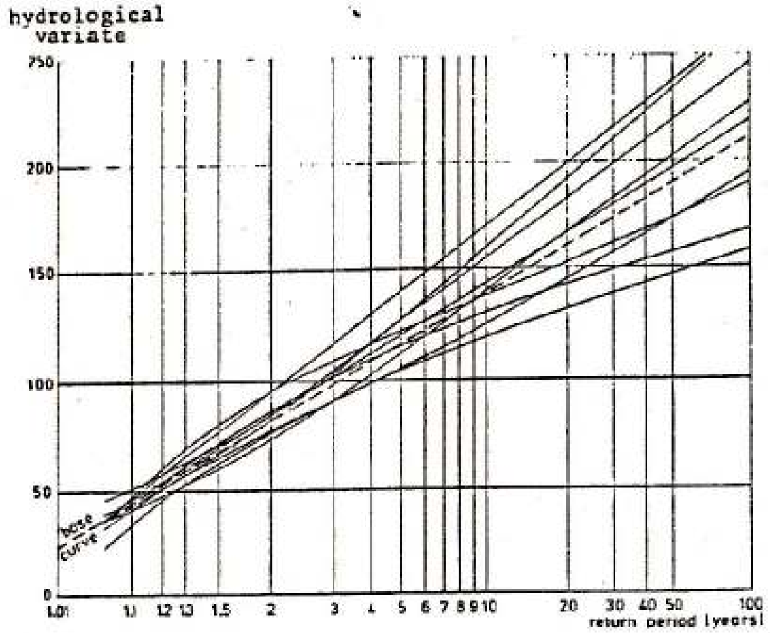

Figure 1. Frequency curves representing different 50-year sample periods derived from the same base curve (after Benson 1960).

In hydrologic literature, numerous methods for confidence estimates of cumulative frequency distributions have been discussed (e.g. Kite 1975), but they do generally not involve the binomial distribution, even though it is quite suitable for this purpose. This article, therefore, is an attempt to open the discussion on the possibility of more frequent use of the binomial distribution for estimates of confidence belts.

The use of the binomial confidence belts given in the graphs is illustrated for three different types of frequency distributions, viz. the exponential, the log-normal, and the Gumbel or Fischer Tippett Type I distribution. In addition, these distributions are determined by two different methods based on respectively:

- Estimation of the frequencies from an interval count or a ranking of the data, followed by a fitting of the distribution (*plotting method*);
- Estimation of the parameters (mean and variance) of the distribution from average and standard deviation of the data (*parametric method*).

The aim of the above illustrations is to investigate whether the type of frequency distribution used and its method of determination lead to significant differences in frequency prediction compared to the width of the confidence intervals.

2. CONFIDENCE ANALYSIS OF CUMULATIVE FREQUENCIES.

In hydrological practice one uses often cumulative (non-exceedance) or exceedance frequency distributions for frequency prediction. These distributions are dealing with two possibilities only: there is exceedance or there is no exceedance. Then the binomial probability distribution can be used to estimate confidence limits of the predicted frequencies, because it deals with the probabilities of only two, mutually exclusive, events.

In hydrologic literature, confidence limits are usually derived from a probability distribution of an event at a certain return period or frequency. For example, Kite (1975) assumes for this a normal probability distribution with parameters determined from the 1st, 2nd and 3rd moments of the frequency distribution concerned. The binominal distribution is used differently: it gives confidence limits of a frequency at a certain magnitude of the event. It's advantages are that it is independent of the frequency distribution used, that it needs no assumption about the probability distribution of the event at at a certain return period, and that it, logically, considers the estimated frequency as the uncertain factor rather than the actually measured event.

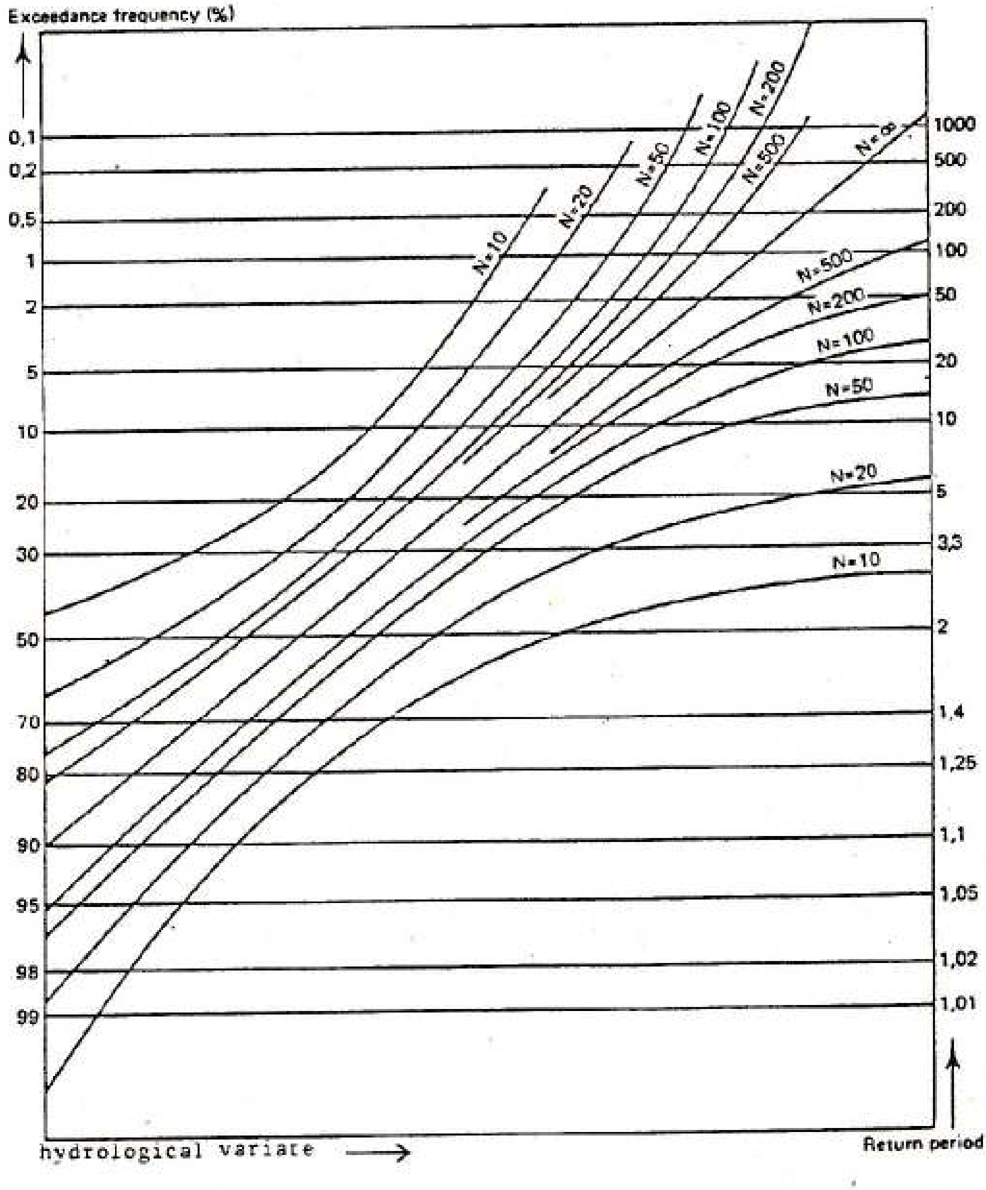

Figure 2a. 95% Binominal confidence belts for different values of sample size N

Application of the binomial distribution requires a large set of binomial tables or computer facilities. Instead, Figures 2a, b and c have been prepared (on normal probability scale) to facilitate the determination of the confidence belts. Similar belts have been presented by Pearson and Hartley (1956), but they were drawn on a linear scale, so that they can not be read for extreme situations.

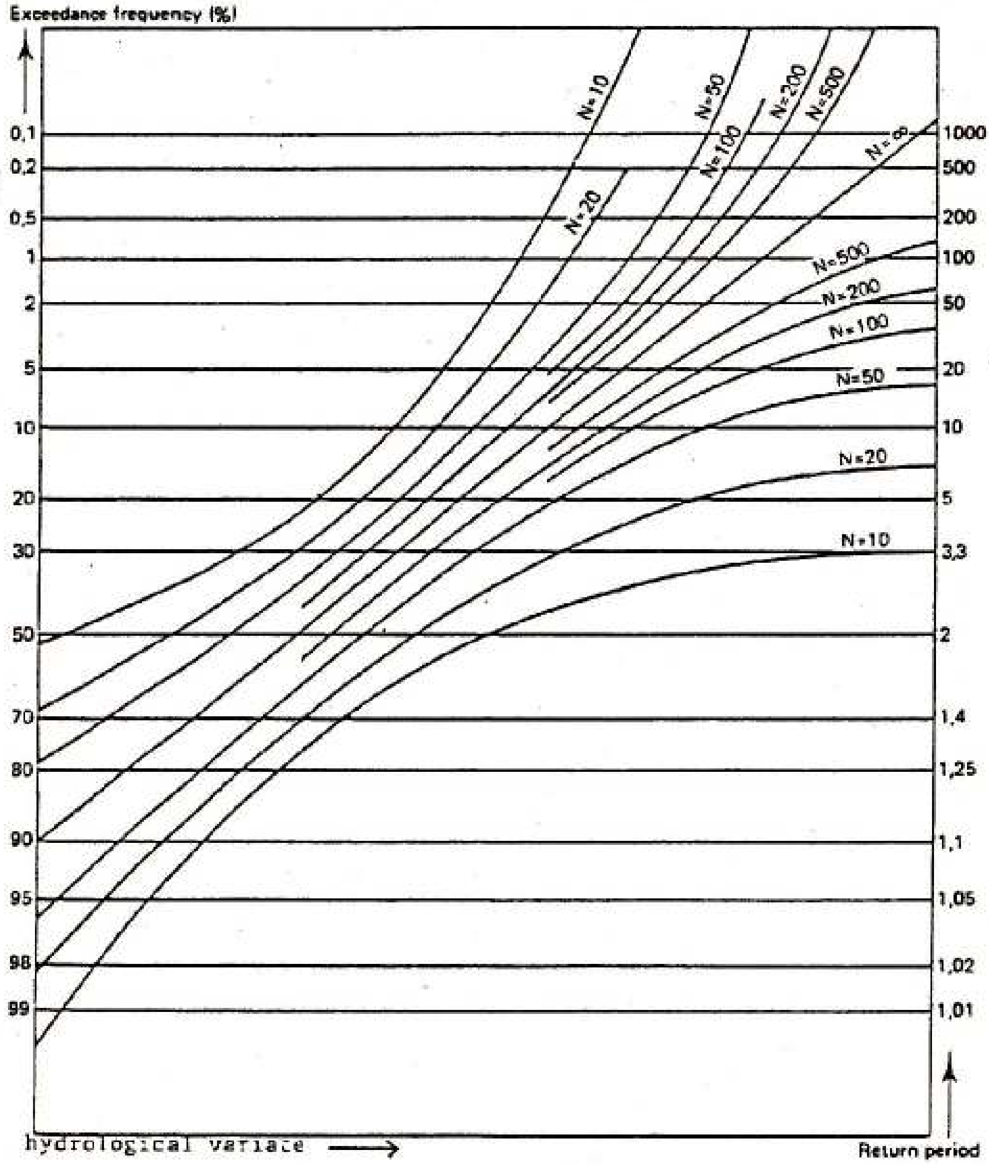

Figure 2b. 90% Binominal confidence belts for different values of sample size N

The Figures 2a, b and c were made with the help of the binomial tables prepared by the Staff of the Computation Laboratory (1955). Use has also been made of the graphs of the binomial distribution prepared by Eisenhart et al. (1947).

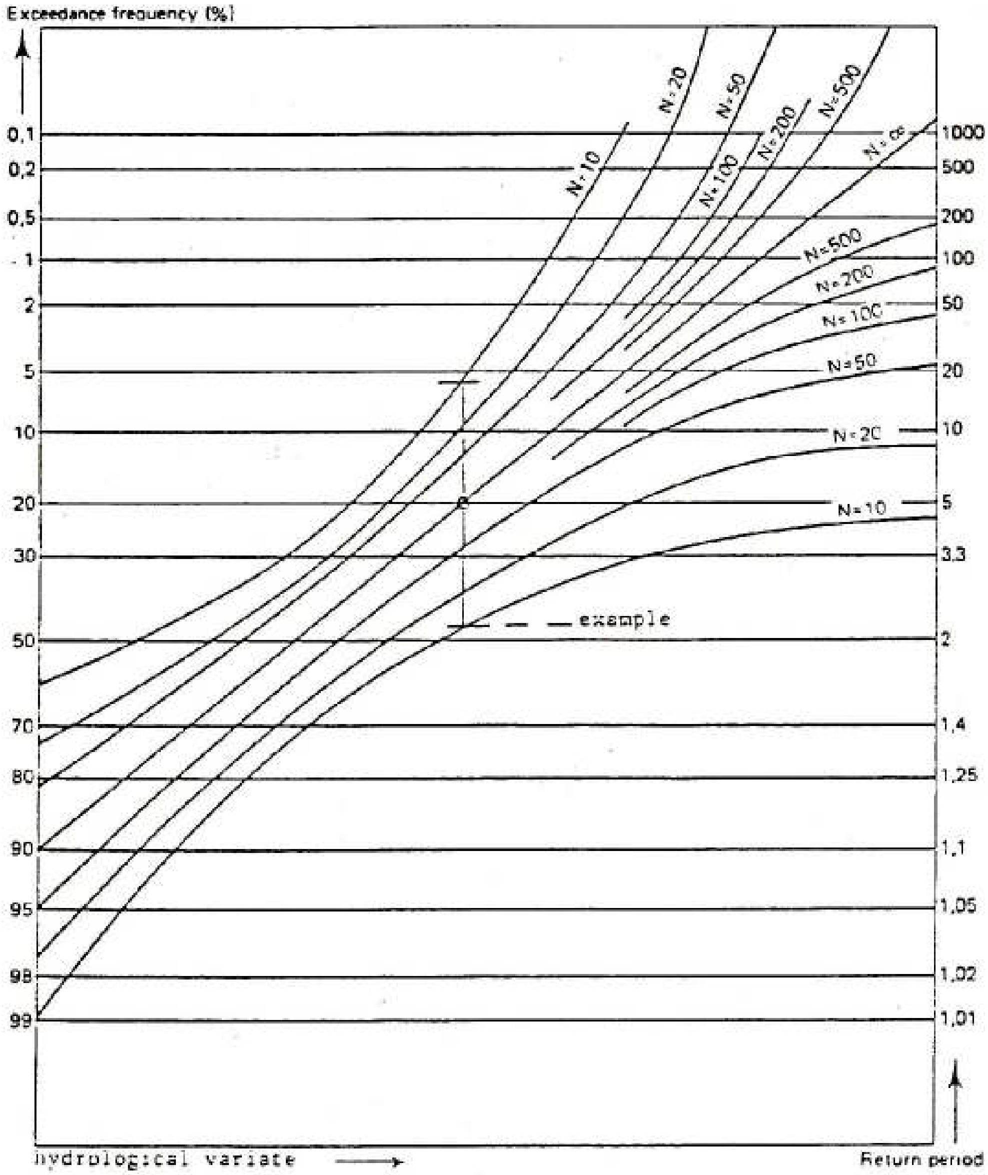

Figure 2c. 80% Binominal confidence belts for different values of sample size N

The use of the figures is as follows. The observed cumulative frequency (F_c) is converted into an exceedance frequency ($F_e = 1 - F_c$) or return period of exceedance ($T_e = 1/F_e$). The value of F_e or T_e is plotted on the middle line of the graph, indicated by $N = \infty$. From there, one goes vertically up or down until reaching the curve corresponding to the number of data (N) use for the estimation of the frequency or the return period (see the example in Figure 2c). Thus one reads respectively:

- An upper confidence limit (T_u) of the return period, or a lower confidence limit (F_{ev}) of the exceedance frequency;
- A lower confidence limit (T_v) of the return period, or an upper confidence limit (F_{eu}) of the exceedance frequency.

The resulting confidence intervals $T_u - T_v$, $F_{eu} - F_{ev}$ or $F_{cu} - F_{cv}$ are determined with 95, 90 or 80% certainty depending on whether one uses Figure 2a, b or c respectively. This means that there is still 5, 10 respectively 20% chance that the frequencies or return periods to be predicted are even higher or lower than indicated by the confidence limits.

The example in Figure 2c shows that a 5 year return period has an 80% confidence interval ranging from 2.2 to 18 years when N = 10.

In theory, the exact confidence intervals are somewhat wider than shown in the graphs, because the mean values and standard deviations of the binomial distributions used are estimated from a data series of limited length. Hence, the true means and standard deviations may be either smaller or larger than the estimated ones. However, the error made in the estimation of the confidence intervals is small compared to their width.

Since the confidence intervals refer to the frequencies expected for a very long future without systematic changes in hydrologic conditions, the confidence intervals for shorter future periods are, again, wider than indicated in the graphs. The same is true when a change of hydrological conditions occurs.

In conclusion it can be said that the intervals presented in the graphs are the narrowest possible.

2. FREQUENCY ESTIMATES BY RANKING METHODS

For direct frequency estimation, the data can be ranked in either ascending or descending order. For a descending order the suggested procedure is:

- Rank the (n) data (x) in a descending order: the highest value first, the lowest last;
- Attach a serial rank number (r) to each value x (x_r, r = 1, 2, 3, ..., n), the highest value being x_1, the lowest x_n;
- Divide the rank number (r) by the total number of observations plus 1 to obtain the frequency of exceedance as

  $$F_e = F(x > x_r) = \frac{r}{n+1}$$

- Determine the cumulative frequency (frequency of non-exceedance) as:

  $$F_c = F(x \leq x_r) = 1 - F(x > x_r) = 1 - \frac{r}{n+1}$$

- Find the return period of exceedance as $T_e = 1/F_e = 1/(1 - F_c)$

The return period gives the long term average interval in which the defined phenomenon will occur once, but there is always the risk that the phenomenon does occur more than once within the return period. According to the binominal distribution, the chance that an event with a return period T_e occurs one or more times in the coming n years is

$$C = 1 - (1 - 1/T_e)^n = 1 - (1 - F_e)^n = 1 - F_c{}^n$$

For example, if T_e is 10 years, then the chance that the exceedance occurs at least once in the coming 5 years is $C = 1 - (1 - 1/10)^5 = 0.41$ or 41 %.

If the ranking procedure is the reverse of the one explained before (ascending instead of descending order), the same relations as above are obtainable interchanging F_e and F_c. An advantage of using the denominator n+1 is that the results are identical when ascending or descending ranking orders are used. In literature, other expressions for the assignment of frequencies have been proposed, but they do not have the advantage of the reversibility.
Most frequency assignments are not unbiased, especially for the extremes. It has been proved (Gumbel, 1954) that the assignment proposed above gives only an unbiased estimate for the mode of the Gumbel distribution.

Table 1 shows the application of the ranking method to maximum 3-day rainfalls (in mm) selected from the April records of 1954 to 1963 from the meteorological station W1 in the Wageningen Polder, Surinam. The data of Table 1 will be used in all following examples

Table 1. Example of the ranking method

Rank	Values		Frequencies			R.P.
r	x	y=log x	F_c	F_e	ln F_e	T_e
1	137	2.14	0.91	0.09	-2.41	11
2	110	2.04	0.82	0.18	-1.71	5.6
3	82	1.91	0.73	0.27	-1.31	3.7
4	71	1.85	0.64	0.36	-1.02	2.8
5	62	1.79	0.55	0.45	-0.80	2.2
6	42	1.62	0.45	0.55	-0.60	1.8
7	39	1.59	0.36	0.64	-0.45	1.6
8	29	1.46	0.27	0.73	-0.31	1.4
9	26	1.41	0.18	0.82	-0.20	1.2
10	5	0.70	0.09	0.91	-0.09	1.1

From the above table we find that

$$\bar{x} = 60 \qquad s_x = 40.7 \qquad \bar{y} = 1.65 \qquad s_y = 0.41$$

where $\bar{x}$ and $\bar{y}$ are the average values of x and y
s_x and s_y are the standard deviations of x and y

3. THE EXPONENTIAL FREQUENCY DISTRIBUTION

The cumulative exponential frequency distribution reads

$F_c = 1 - e^{-a\,x}$, so that $F_e = 1 - F_c = e^{-a\,x}$, and $\ln F_e = -a\,x$

where $a = 1/\mu$, and μ is the mean of the distribution, estimated by $\bar{x}$.

In Figure 3, the data of Table 1 (x and $\ln F_e$, obtained by the ranking method) are plotted in a graph with an eye estimate of the best fitting straight line (plotting method). Further, using the paramatric method, the line $\ln F_e = -a\,x$, with the slope $a = 1/\bar{x} = 1/60 = 0.017$, is plotted through the points ($x = 0$, $\ln F_e = 0$) and ($x = 60$, $\ln F_e = -1$), together with its 80% confidence belt. Also, the 80% confidence interval of a 5 year return period, ranging from 2.2 to 18 years (as shown in the example of Figure 2c), is specifically indicated in Figure 3 by a dotted line.

It is seen that the line determined by the plotting method - for which the expression $\ln F_e = -0.20\,(x - 20)$ holds - falls within the confidence belt (except at very low rainfall values, because this line does not permit values less than 20 mm). Also, the difference between the two lines is small compared to the confidence belt. Hence it is difficult to say which of the lines is the better one.

From the figure it is seen that, for example, a 3-day April rainfall with a return period of 5 years ($T_e = 5$, $F_e = 0.2$ and $\ln F_e = -1.6$) is, with 80% certainty, between the limits of 55 and 170 mm.

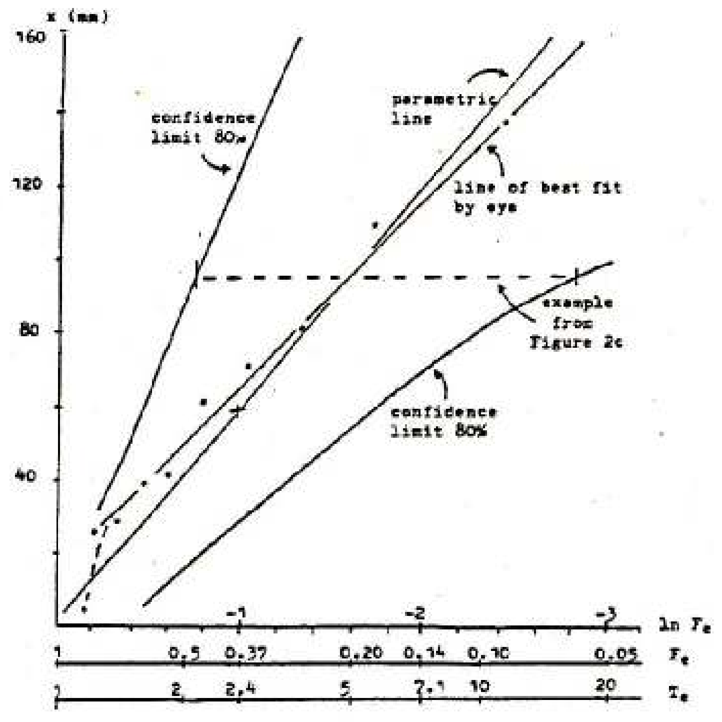

Figure 3. The exponential frequency distribution fitted to the data of Table 1.

4. THE LOG-NORMAL FREQUENCY DISTRIBUTION

The normal frequency distribution, also known as the Gauss or the De Moivre distribution, cannot be written directly in terms of cumulative frequencies. The same holds for the log-normal distribution. The mathematical expression of a log-normal distribution is therefore given as a frequency density function

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

where

- $f(y)$ = the normal frequency density function of y
- y = log x, the log-normal variate ($0 < x < \infty$, $-\infty < y < \infty$)
- μ = the mean of the distribution, estimated by $\bar{y}$
- σ^2 = the variance of the distribution, estimated by s^2, where s is the standard deviation of y

The factor $(y - \mu)/\sigma$ is called the reduced log-normal variate (y').

The frequency of exceedance of y can be found from $F_e(y) = \int_y^{+\infty} f(y)\,dx$, the solution of which must be found by numerical methods. The solutions are given in most statistical handbooks in the form of tables. One finds, for example, $F_e(\mu) = 0.5$, $F_e(\mu+\sigma) = 0.16$ and $F_e(\mu-\sigma) = 0.84$.

In Figure 4, the data of Table 1 (y and F_e, obtained by the ranking method) are plotted on normal probability paper with an eye estimate of the best fitting straight line (the plotting method). Further, using the parametric method, a straight line is drawn through the points ($\bar{x}$, $F_e = 0.5$) and ($\bar{x} + s_y$, $F_e = 0.16$), together with its 80% confidence interval and a specific indication of the example in Figure 2c.

It is seen that the line determined by the plotting method falls entirely within the confidence belt. Also, the difference between the two lines is small compared to the width of the confidence belt. Hence it is difficult to say which of the lines is the better one.

A 3-day April rainfall with a return period of 5 years will be between the limits of 58 and 190 mm, with 80% confidence. This range is not much different for the corresponding range estimated before by the exponential distribution.

5. THE GUMBEL FREQUENCY DISTRIBUTION

The Gumbel or Fischer-Tippett Type I distribution of extreme values (Gumbel, 1954) can be written as a cumulative frequency distribution

$$F_c = \exp\{-\exp\{-\alpha(x - u)\}\}$$

where

- u = $\mu - c/\alpha$, the mode of the distribution
- μ = the mean of the distribution, estimated by $\bar{x}$
- c = Euler's constant = 0.577
- α = $\pi/\sigma\sqrt{6}$
- σ = standard deviation of x, estimated by s_x

The factor $\alpha(x - u)$ is called the reduced Gumbel variate (x').

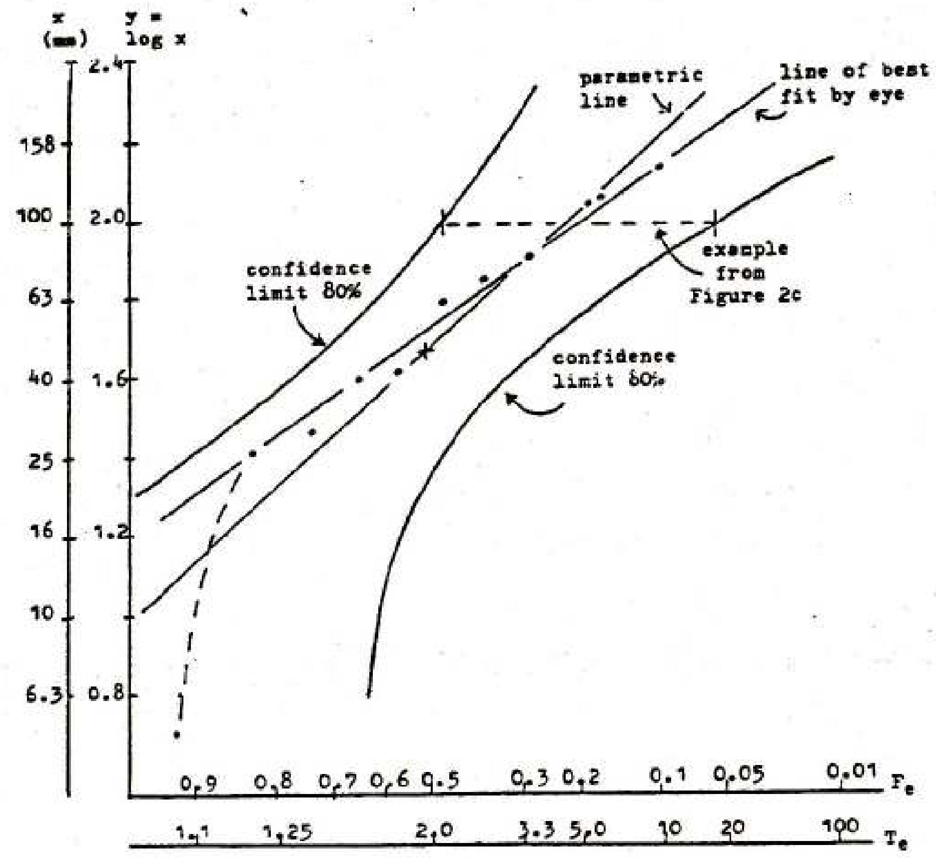

Figure 4. The log-normal frequency distribution fitted to the data of Table 1.

It can be proved that the cumulative probability distribution of the maximum value of a sample of size m approaches asymptotically the Gumbel distribution with increasing m, if the samples are drawn from a distribution of the exponential type. In hydrological practice it is assumed that the asymptotic approach is already realized for m > 10. Therefore, the Gumbel distribution is often used in hydrology for annual or monthly maxima of floods or rainfalls of relatively short duration.

The Gumbel distribution can also be written as

$$x' = -\ln\{-\ln F_c\} = -\ln\{-\ln(1 - F_e)\} = -\ln\{-\ln(1 - 1/T_e)\}$$

From the data of Table 1 it is found that $u = 42$ and $\alpha = 0.032$. Also, it is found that for $x = u$ (where $x' = 0$): $F_c = 0.37$, and for $x = 137$ (where $x' = 3$): $F_c = 0.95$.

In Figure 5, the data of Table 1 (x and F_c, obtained by the ranking method) are plotted on Gumbel probability paper with an eye estimate of the best fitting straight line (the plotting method). Further, using the parametric method, a straight line is drawn through the points ($x = u$, $F_c = 0.37$) and ($x = 137$, $F_c = 0.95$), together with its 80% confidence interval and a specific indication of the example in Figure 2c.

It is seen that the line determined by the plotting method falls entirely within the confidence belt. Also, the difference between the two lines is small compared to the width of the confidence belt. Hence it is difficult to say which of the two lines is the better one.

From Figure 5 it is seen that a 3-day April rainfall with a return period of 5 years will be between the limits of 57 and 180 mm, with 80% confidence. This range does not differ much from the ranges estimated previously by the exponential and log-normal frequency distributions.

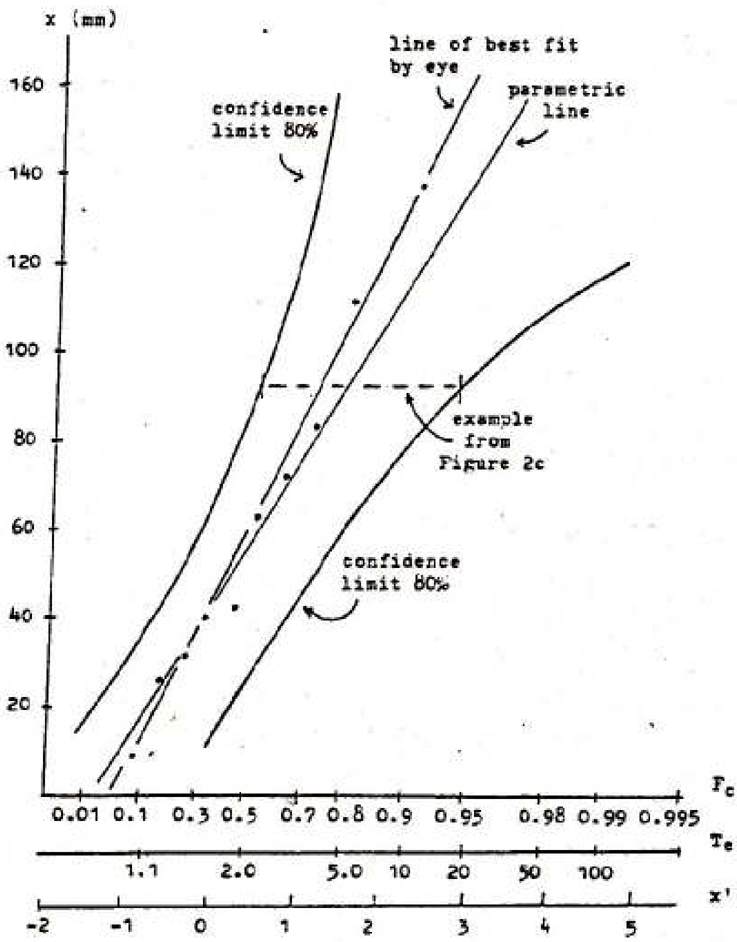

Figure 5. The Gumbel frequency distribution fitted to the data of Table 1.

6. CONCLUSIONS

From the foregoing considerations the following conclusions can be drawn:

- Frequency predictions need to be accompanied by confidence statements;
- The binomial probability distribution can be used for the assessment of the confidence intervals for cumulative and exceedance frequencies, as well as for exceedance return periods, independent of the type of frequency distribution used;
- The differences between frequency predictions, using different frequency distributions and different methods for their determination, are small compared to the width of the confidence intervals and, therefore, these differences are insignificant;
- The variation of the plotting positions around the best fitting lines is not indicative of the reliability of the frequency distribution because, owing to the application of the ranking procedure, a very high correlation between frequencies and values of the hydrologic event is automatically introduced;
- The return period, as a tool for the economic evaluation of hydraulic works designed to cope with hydrologic events not exceeding the corresponding value, has a limited significance because:
  - The assessment of the exact value of the return period and the corresponding value of the hydrologic event is subject to a large uncertainty;
  - There is a considerable risk that the hydrologic event, on which the design is based, is exceeded more than once within the return period.

REFERENCES

Benson, M.A. 1960. Characteristics of frequency curves based on a theoretical 1000 year record. In: T. Dalrymple, Flood frequency analysis. U.S. Geol. Survey Water Supply Paper 1543-A.

Eisenhart, C., M.W. Hastey and W.A. Wallis (eds.), 1947. Selected techniques of statistical analysis. McGraw-Hill, New York.

Gumbel, E.J. 1954. Statistical theory of extreme values and some practical applications. Applied mathematics series 33. U.S. Dep. of Commerce, National Bureau of Standards.

Kite, G.W. 1975. Confidence limits for design events. Water Resources Research Vol. II no. 1, p.48-53.

Pearson, E.S. and H.D. Hartley 1956. Biometrika. Tables for statisticians, Vol. I. Cambridge Univ. Press.

Staff of the Computation Laboratory, 1955. Tables of the cumulative binomial probability distribution. Harvard University Press.