519.2:519.6

ROYAL OBSERVATORY, HONG KONG

TECHNICAL NOTE NO. 76

# COMPUTER PROGRAMS FOR FREQUENCY DISTRIBUTION ANALYSIS IN THE ROYAL OBSERVATORY HONG KONG

by

B.Y. LEE

DECEMBER 1987

CONTENTS

## SYMBOLS

| | |
|---|---|
| A | Observed number of events (used in chi-square statistics) |
| $a, \gamma$ | Lower bound of sample, if any |
| E | Expected number of events (used in chi-square statistics) |
| K | Frequency factor |
| k | Number of class intervals (used in chi-square statistics) |
| P(x) | Cumulative probability of an event being less than or equal to x |
| p(x) | Probability density function |
| S(K) | Standard error corresponding to the frequency factor K |
| SE | Standard error of a fitted frequency distribution |
| T | Return period |
| x | Event magnitude |
| $x_T$ | Event magnitude at a given return period T |
| $\alpha, \beta, \gamma$ | Parameters in a frequency distribution |
| $\mu$ | Population mean |
| $\mu_2$ | Second central moment of the sample |
| $\mu_3$ | Third central moment of the sample |
| $\sigma$ | Population standard deviation estimated by sample moments |

# 1. INTRODUCTION

The problem commonly faced in engineering design is the estimation of a design event from fairly short records. The available data are generally used to fit a curve graphically or to fit a mathematical frequency distribution.

Graphical methods are simple, visually effective and make no assumption of distribution type. These advantages are outweighed, however, by the high degree of subjectivity in the methods, which is not compatible to the uniformity and objectivity required in other phases of engineering design. It is generally stated that mathematical fitting of a standard frequency distribution is preferable to plotting a graph and fitting a curve (Chow 1964).

The available data are usually used to fit a frequency distribution by estimating the parameters of the frequency distribution. The distribution in turn is used to extrapolate the design event from the recorded events. Various frequency distributions have found their application in the estimation of the magnitude of a design event in hydrology, meteorology, air pollution, oceanography as well as in other fields of science and engineering. Applications to meteorology and related atmospheric sciences are discussed by Essenwanger (1976).

In fitting a distribution, the statistical parameters have to be estimated from the sample data. Since the sample data are themselves subject to error, the method of fitting must be as efficient as possible in order to minimize these errors. Yevjevich (1972) listed four techniques of parameter estimation,

in the order of increasing efficiency, as follows :-

(1) graphical,

(2) least sum of squares,

(3) method of moments, and

(4) method of maximum likelihood.

This report describes a number of standard frequency distributions and the computer programs used in the Royal Observatory Hong Kong. The distributions are, namely, the normal, 2-parameter lognormal, 3-parameter lognormal, Pearson type III, log-Pearson type III, type I extremal (Gumbel) and type III extremal (Weibull) distributions. Parameter estimation by the method of moments and the method of maximum likelihood (see, for example, Yevjevich 1972) is based on computer programs (Fortran V) given in Kite (1977). Apart from the computation of the standard error of each of the distributions, the programs have also been modified to incorporate the chi-square and Kolmogorov-Smirnov tests in order to assess the goodness of fit. Program listings and sample outputs are available on request to the Royal Observatory Hong Kong.

## 2. FREQUENCY DISTRIBUTIONS

In this section, the frequency distributions are briefly described. To study them in greater detail, the reader can refer to Chow (1964), Yevjevich (1972) or Kite (1977).

Once the parameters of a distribution have been estimated the question is how to use the distribution in frequency analysis. Chow (1964) proposes a general equation :

$$x_T = \mu + K\sigma \qquad (1)$$

where $x_T$ is the event magnitude at a given return period T, $\mu$ and $\sigma$ are, respectively, the population mean and the standard deviation estimated by sample moments. K is a frequency factor which is a function of the return period and the parameters of the distribution.

Conventionally when it is desired to plot observed data on a graph in order to interpret the data, detect errors, or to have an idea of which frequency distribution should be used to describe the data, the ordinate of such a graph usually contains the event magnitudes (on a linear or logarithmic scale) while the abscissa will be some measure of the probability of occurrence of each event or the return period. It is common to use the following plotting position on the abscissa : m/(N+1), where m is the rank of the recorded event and N is the total number of years of data. Then the cumulative probability of an event being less than or equal to x is :

$$P(x) = 1-1/T = 1-m/(N+1) \qquad (2)$$

Discussions on the use of other plotting positions can be found

in Chow (1964), Yevjevich (1972) and Sevruk and Geiger (1981). In particular, Sevruk and Geiger (1981) note that the differences between estimates due to different plotting positions are generally small.

The skewness coefficient, given by $\mu_3 / \mu_2^{3/2}$ where $\mu_2$ is the second central moment and $\mu_3$ the third central moment, is usually used as a measure of asymmetry.

## 2.1 Normal distribution

The probability density function of the normal distribution is defined as :

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2} \qquad (3)$$

where x is the event magnitude, $\mu$ and $\sigma$ are parameters which can be shown to be the population mean and the standard deviation of the variable respectively. The probability corresponding to any interval in the range of the variate x is represented by the area under the probability density curve,

$$P(x) = \int_{-\infty}^{x} p(x)\, dx \qquad (4)$$

Sometimes, the standard normal deviate t, defined as $(x-\mu)/\sigma$, is used in the place of x. The areas under the standard normal curve for different values of t are available from standard statistical tables. Comparison with Eqn. 1 shows that the frequency factor K is equal to t for the case of normal distribution.

## 2.2 2-parameter lognormal distribution

Sometimes, the causative factors for a variable may act multiplicatively rather than additively. The variable will then be the product of the causative factors. Chow (1964) has provided theoretical justification for the use of the lognormal distribution, in which the logarithm, ln x, is used in the place of the variable x in the normal distribution. The probability density function becomes

$$p(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} \exp \frac{-(\ln x - \mu_y)^2}{2\sigma_y^2} \qquad (5)$$

where $\mu_y$ and $\sigma_y$ are the mean and standard deviation of the natural logarithms of x.

Records may contain zero values and, when taking logarithms, this becomes negative infinity and cannot be processed. Kilmartin and Peterson (1972) proposed several ways of treatment, but all of these affect the parameter estimates. The following two techniques have been tested by the Office of Water Data Coordination, U.S. Department of the Interior (1982) and have less effect on the parameter estimates:

(a) Add 1% of the mean magnitude to all values for computation purposes and subtract that amount from subsequent estimates; or

(b) Remove all zeroes and multiply estimated exceedance frequencies of the remaining by the ratio of the number of non-zero values to the total number of values.

This second 'conditional probability' approach has been discussed in WMO Technical Note No. 98 (1969).

## 2.3 3-parameter lognormal distribution

Where it is likely that for the variable x there is a lower boundary, a, but which is unknown, the 3-parameter lognormal distribution can be used to represent the normal distribution of the logarithms of the reduced variable (x-a) :

$$p(x) = \frac{1}{(x-a)\,\sigma_y\sqrt{2\pi}} \exp \frac{-\left[\ln(x-a)-\mu_y\right]^2}{2\sigma_y^2} \qquad (6)$$

where $\mu_y$ and $\sigma_y$ are the mean and standard deviation of the natural logarithms of (x-a).

This distribution has been found suitable for flood frequency analysis in Ontario (Automated Business and Engineering Ltd. 1980).

## 2.4 Pearson type III distribution

The probability density distribution is of the form

$$p(x) = \frac{1}{\alpha\Gamma(\beta)} \left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} \exp - \left(\frac{x-\gamma}{\alpha}\right) \qquad (7)$$

where $\alpha$, $\beta$ and $\gamma$ are parameters to be estimated and $\Gamma(\beta)$ is the gamma function. If the substitution $y=(x-\gamma)/\alpha$ is made, Eqn. 7 simplifies to

$$p(y) = \frac{y^{\beta-1} e^{-y}}{\Gamma(\beta)} \qquad (8)$$

which is a one parameter gamma function.

## 2.5 Log-Pearson type III distribution

The logarithms, ln x, is used in the place of the variable x in the Pearson type III distribution. The resulting probability density distribution is

$$p(x) = \frac{1}{\alpha x \Gamma(\beta)} \left(\frac{\ln x - \gamma}{\alpha}\right)^{\beta-1} \exp - \left(\frac{\ln x - \gamma}{\alpha}\right) \qquad (9)$$

The U.S. Federal Water Resources Council (1967) recommends the use of the log-Pearson type III distribution in the standard flood frequency analysis. Flood studies by the Natural Environment Research Council of the United Kingdom (1975) also indicate that this distribution gives good fit to the data. Kite (1977) notes, however, that this choice is subjective to some extent as no rigorous statistical criteria exists on which a comparison of distributions can be made.

For treatment of occasional zero values of x, the reader should refer to Section 2.2.

## 2.6 Type I extremal (Gumbel) distribution

Suppose from N samples each containing n events the maximum (minimum) event in each sample is selected. As n increases, the distribution of the N maxima (minima) approaches a limiting form. The type of the limiting form would depend on the type of the initial distribution of the n.N values. The distribution type of the maxima (minima) is given by the functional equation:

$$p_n(x) = P(a_n x + b_n) \qquad (10)$$

where $a_n$ and $b_n$ depend on n.

Fisher and Tippett (quoted in Yevjevich 1972) have shown that there are three possible solutions to Eqn. 10. These are known as types I, II and III extremal distributions. The type I distribution is unbounded, the type II has an upper limit and the type III has a lower limit.

The type I distribution results from any initial distribution of exponential type which converges to an exponential function as x increases. Examples of such initial distributions are the normal and the lognormal distributions. Logically, the form for type I distribution satisfying Eqn. 10 is

$$P(x) = \exp\left[-\exp - \alpha(x-\beta)\right] \tag{11}$$

and

$$p(x) = \alpha \exp\left[-\alpha(x-\beta) - \exp - \alpha(x-\beta)\right] \tag{12}$$

## 2.7 Type III extremal (Weibull) distribution

In contrast to the type I distribution (Section 2.6), the type III distribution results from a type of initial distribution in which x is limited by a lower bound, $\gamma$ . The following form satisfies Eqn. 10 :

$$P(x) = 1 - \exp\left[-\left(\frac{x-\gamma}{\beta-\gamma}\right)^{\alpha}\right] \tag{13}$$

and

$$p(x) = \frac{\alpha}{\beta-\gamma}\left(\frac{x-\gamma}{\beta-\gamma}\right)^{\alpha-1} \exp\left[-\left(\frac{x-\gamma}{\beta-\gamma}\right)^{\alpha}\right] \tag{14}$$

where $\alpha$, $\beta$ and $\gamma$ are parameters to be estimated.

Expressed in a slightly different way, this distribution can deal with minimum events :

$$P(x) = 1 - \frac{1}{T} = \exp\left[-\left(\frac{x-\gamma}{\beta\ \gamma}\right)^{\alpha}\right] \qquad (15)$$

where P(x) is the cumulative probability of an event being less than or equal to x.

## 3. MOMENTS, VARIANCE, FREQUENCY FACTOR AND CONFIDENCE LIMITS

Derivation of analytical expressions for the sample mean, higher order moments, skewness, frequency factor and the variance of a T-year event can be found in Kite (1977) and in Yevjevich (1972).

In particular, the variance of a T-year event $x_T$ is derived from moment estimates by assuming $x_T$ a function of the first three moments of a distribution and the return period T.

To compute the confidence limits, an empirical method is often used which computes the standard error of x(K), S(K), and then assumes that the T-year event is normally distributed with mean x(K) and standard deviation S(K) so that the confidence interval is given by

$$x(K) + tS(K) \qquad (16)$$

where t is the standard normal deviate at the required confidence level. This method has been tested (Kite 1975).

The treatment of suspected outliers and broken record is suggested in Chapter V of Guidelines for Determining Flood Flow Frequency published by the U.S. Department of Interior Geological Survey (1982). In the case of broken record, it is suggested that the different record segments be analysed as a continuous record with length equal to the sum of both records, unless there is some physical change in the observation site between segments which may make the total record non-homogeneous.

## 4. GOODNESS OF FIT

To test the fit of a distribution to a particular sample, the most commonly-used methods are to evaluate the standard error, the chi-square and the Kolmogorov-Smirnov statistics.

### 4.1 Standard error

The standard error of each of the distributions is computed as

$$SE^2 = \frac{\sum_{i=1}^{N} (x_i - y_i)^2}{N-h} \tag{17}$$

where $x_i$, i=1,...,N are the recorded events, $y_i$, i=1,...,N are the computed event magnitudes and h is the number of parameters of the distribution.

### 4.2 Chi-square test

Before applying this test, the sample is usually classified into a number of class intervals (k) which are defined in such a way that each interval corresponds to an equal probability. The expected number of events in each class interval is thus equal to E=N/k, N being the sample size. The chi-square statistics (see Yevjevich 1972)

$$\chi^2 = \sum_{j=1}^{k} \frac{(A_j - E_j)^2}{E_j} \tag{18}$$

is therefore simplified into

$$\chi^2 = \frac{k}{N} \sum_{j=1}^{k} A_j^2 - N \tag{19}$$

where $A_j$ is the observed number of events in the jth class

interval, $E_j$ is the number of events expected from the theoretical distribution and in this case is equal to N/k.

Yevjevich (1972) suggested that both the number of class intervals (k) and the expected number of events in each class interval (N/k) should be at least 5. If the sample size is of the order 30-60, 6 or 7 class intervals are acceptable numbers for k. For the test, the number of degree of freedom is given by k-h-1, where h is the number of parameters in the theoretical distribution. Yevjevich (1972) has provided chi-square distribution values for 3 and 4 degrees of freedom.

The class intervals computed for the various distributions are as follows :

(a) NORMAL : $\bar{x} + tS$

where $\bar{x}$ and S are the sample mean and standard deviation and t is the standard normal deviate corresponding to the probability of exceedance, P. In the case of 7 class intervals (k=7), P is equal to 1/7, 2/7, 3/7, 4/7, 5/7, 6/7 and 1.

(b) 2-PARAMETER LOGNORMAL : $\exp(\bar{x}_n + tS_n)$

where $\bar{x}_n$ and $S_n$ are the mean and standard deviation of the logarithms of the recorded events.

(c) 3-PARAMETER LOGNORMAL : $a + \exp(\bar{x}_{na} + tS_{na})$

where a is the lower boundary of the distribution and $\bar{x}_{na}$ and $S_{na}$ are the mean and standard deviation of the logarithms of the sample x-a.

(d) PEARSON TYPE III : $\bar{x} + (\frac{\chi^2\gamma_1}{4} - \frac{2}{\gamma_1})S$

where $\chi^2$ is the value of chi-square at probability P and $8/\gamma_1^2$ degrees of freedom; $\gamma_1$ is the sample skew coefficient.

(e) LOG-PEARSON TYPE III : $\exp(\bar{x}_n + [\frac{\chi^2\gamma_n}{4} - \frac{2}{\gamma_n}]S_n)$

where $\gamma_n$ is the coefficient of skew of ln x.

(f) TYPE I EXTREMAL (GUMBEL) : $\bar{x} + (\frac{y_m - \mu}{\sigma})S$

where $y_m$ is $-\ln(-\ln P)$ and $\mu$ and $\sigma$ are the mean and standard deviation of the plotting positions.

(g) TYPE III EXTREMAL (WEIBULL) : $\gamma + y_m^{1/\alpha}(\beta - \gamma)$

where $y_m$ is $-\ln(1-P)$; $\alpha$, $\beta$ and $\gamma$ are estimated using the method of moments.

### 4.3 Kolmogorov-Smirnov test

This test is designed to avoid the loss of information due to grouping suffered by the chi-square test and is based on deviations of the sample distribution function P(x) from the completely specified hypothetical distribution function $P_o(x)$ (Yevjevich 1972). The Kolmogorov-Smirnov parameter is given by

$$D_n = \max|P(x) - P_o(x)| \qquad (20)$$

The test requires that the computed parameter be less than the tabulated value at the required confidence level. The tabulated value can be found in standard statistical tables.

## 5. STATISTICAL RESTRICTIONS

Statistical restrictions exist on some of the distributions (Yevjevich 1972, Kite 1977) and are summarized as follows :

NORMAL - the sample coefficient of skew should be very small (< 0.05) and the sample data should be such that P is very small;

2- and 3-PARAMETER LOG-NORMAL - the sample coefficient of skew of the reduced data (ln x or ln(x-a)) should be very small (< 0.05) and positive;

PEARSON TYPE III - this is only unbounded at the upper end for positive coefficient of skew;

LOG-PEARSON TYPE III - this is only unbounded at the upper end for positive coefficients of skew of the logarithms and when the parameter $1/\alpha$ is greater than zero and $\beta$ is greater than 1;

TYPE I EXTREMAL (GUMBEL) - the sample coefficient of skew should be very close to 1.13; and

TYPE III EXTREMAL (WEIBULL) - the skewness coefficient for $\ln(x-\gamma)$ should be very close to 1.13 for the distribution to be considered as log-Gumbel, i.e. when $\ln\left(\frac{x-\gamma}{\beta-\gamma}\right)$ approaches (x-a), a = constant.

## 6. COMPUTER PROGRAMS

Computer programs for parameter estimation of the various distributions have been presented by Kite (1977). The programs are written in FORTRAN V and have been adapted for use on an Eclipse S/130 mini-computer at the Royal Observatory. Outputs from these programs have been checked against those produced by Kite (1977) using the annual maximum daily discharges recorded during 1915-1974 at St. Marys River at Stillwater, Nova Scotia.

The names of the programs are :

SOHNOR, SOHLN2, SOHLN3, SOHPT3, SOHLP3, SOHT1E and SOHT3E the last three characters of each being self-explanatory. Double precision is employed for all variables except integer variables. The programs accept card images as input. The first card consists of a title statement of no more than 80 characters. The second card defines the number of years of data in format I5. This is followed by the input data expressed in format 8D10.0.

In each of the above programs, modifications have been made to include the computation of the chi-square and the Kolmogorov-Smirnov parameters. The computation of the standard error of all the distributions is done by a separate program (SOHSER) and is based on parameters estimated by the method of maximum likelihood. However, this method does not always provide a solution for type III extremal distribution, and in these cases, the method of moments is used to obtain the parameter estimates.

The programs have been applied to annual maximum hourly rainfall recorded at the Royal Observatory during 1947-1982.

Statistical restrictions on the use of a certain distribution, as described in the last section, are included in the computer printout as a reminder to the user.

A separate program, T3EMIN, which is based on Eqn. 15 for the type III extremal distribution, can be used to fit minimum events. Included in this program is the method of smallest observed drought, as described in Gumbel (1958). The program has been tested using the annual minimum daily discharges during 1915-1974 at St. Marys River at Stillwater, Nova Scotia.

## 7. DISCUSSIONS AND CONCLUSIONS

Various common frequency distributions have been discussed. The computer software to fit these distributions statistically makes use of the method of moments and the maximum likelihood method. It also takes note of various statistical restrictions on the distributions. Goodness-of-fit tests are offered to aid the selection of an adequate distribution for a given data base and purpose.

In selecting a distribution, a word of caution is perhaps appropriate. It is normally not known which distribution, if any, the events naturally follow. Sevruk and Geiger (1981) and Kite (1977, Chapter 15) review various applications of frequency distributions to hydrological data and find that no single distribution is acceptable to all hydrologists. There are no rigorous statistical criteria on which a comparison of distributions can be made. Goodness of fit is a necessary but not a sufficient condition for acceptance of a certain distribution. This applies equally well to the verification of estimated design magnitude with actual observation.

Notwithstanding these, the following points can be considered when a choice has to be made between the distributions:

(a) whether there is theoretical justification -- e.g. whether the causative factors for the variate x are additive or multiplicative, and in the case of extremal distributions whether the original assumptions listed in Gumbel (1954) are satisfied; without theoretical justification, a perfect statistical fit may

leave the basic physics unrecognized;
(b) three-parameter distributions generally offer greater flexibility than two-parameter distributions -- e.g. the normal distribution (two parameters) is completely and uniquely specified once the mean and standard deviation are estimated; and (c) whether the distribution of interest complies with the statistical restriction(s) described in Section 5.

The computer programs described in Section 6 are available for use by anyone who is interested in their applications to analyses of meteorological and other data for design purposes. Enquiries about details of these programs should be directed to the Director of the Royal Observatory Hong Kong.

## REFERENCES


| | | |
|---|---|---|
| Automated Business and Engineering Ltd. | 1980 | Choice of Statistical Distributions for Flood Flow Frequency Analysis in Southern Ontario, prepared for Ontario Ministry of Natural Resources. |
| Chow, V.T. | 1964 | Handbook of Applied Hydrology, McGraw-Hill Inc. |
| Essenwanger, O. | 1976 | Applied Statistics in Atmospheric Science Part A: Frequencies and Curve Fitting, Elsevier Scientific Publishing Co. |
| Gumbel, E.J. | 1954 | Statistical Theory of Extreme Values and Some Practical Applications, National Bureau of Standards, Applied Mathematics Series 33. |
| Gumbel, E.J. | 1958 | Statistics of Extremes, Columbia University Press. |
| Kilmartin, R.F. and Peterson, J.R. | 1972 | Rainfall-Runoff Regression with Logarithmic Transforms and Zeros in the Data, Water Resources Research, Vol. 8, No. 4, pp 1096-1099. |
| Kite, G.W. | 1975 | Confidence Limits for Design Events, Water Resources Research, Vol. 11, No. 1, pp 48-53. |
| Kite, G.W. | 1977 | Frequency and Risk Ananlysis in Hydrology, Water Resources Publications. |
| Natural Environment Research Council | 1975 | Flood Studies Report, Vol. 1 : Hydrological Studies, Whitefriars Press Ltd., London. |
| Sevruk, B. and H. Geiger | 1981 | Selection of Distribution Types for Extremes of Precipitation, WMO Operational Hydrology Report No. 15. |

REFERENCES (continued)


| | | |
|---|---|---|
| U.S. Department of the Interior Geological Survey | 1982 | Guidelines for Determining Flood Flow Frequency, Bulletin No.17B, Hydrology Subcommittee, Office of Water Data Coordination. |
| U.S. Water Resources Council | 1967 | A Uniform Technique for Determining Flood Flow Frequencies, Hydrology Committee, Bulletin 15. |
| World Meteorological Organization | 1969 | Estimation of Maximum Floods, WMO Technical Note No. 98. |
| Yevjevich, V. | 1972 | Probability and Statistics in Hydrology, Water Resources Publications. |