Chapter One

11 Frequency Analysis

Extreme rainfall events and the resulting floods can take thousands of live and cause billions of dollars in damage. Flood plain management and design of flood control works, reservoirs, bridges, and other investigations need to reflect the likelihood or probability of such events. Hydrological studies also need to address the impact of unusually low rainfalls causing low stream flows which affects for example water quality and water supply.

The term frequency analysis refers to the techniques whose objective is to analyze the occurrence of hydrologic variable within statistical framework, by using measured data and basing predictions on statistical laws.

Frequency analyses try to answer the following problems:

(1) Given n years of daily streamflow record for stream X, what is the maximum (or minimum) flow Q that is likely to recur with a frequency of once in T years on average?

(2) What is the return period associated with a maximum (or minimum) flow Q. In more general term, the preceding questions can be stated as follows: given n years of streamflow data for stream X and L years design life of a certain structure, what is the probability P of a discharge Q being exceeded at least once during the design life L?.

Frequency analysis is made using appropriate probability distribution function to the random variable under consideration. The next section briefly summarizes probability distribution functions commonly used in hydrology.

11.1 Concepts of statistics and probability

Hydrological processes evolve in space and time in a manner that is partly predictable, or deterministic and partly random, and such processes are called stochastic processes. In this chapter, pure random processes are discussed using statistical parameters and functions.

Let *X* is a random variable that is described by a probability distribution function and represent for example annual rainfall amount at a specified location. Let a set of observations $x_1, x_2, \ldots, x_n$ of this random variable sample be drawn from a hypothetical infinite population possessing constant statistical properties that is stationarity (having no significant trend and variation in variance).

Defining sample space as a set of all possible samples that could be drawn from the population, and an event as a subset of the sample space; the probability of an event *A*, *P(A)*, is the chance that it will occur when an observation of the random variable is made.

If a sample of *n* observation has n_A values in the range of event *A*, then the relative frequency of *A* is

$$f_s = n_A / n \qquad \textbf{(11.1)}$$

and the *P(A)* is given by

$$P(A) = \lim_{n\to\infty} n_A / n \qquad (11.2)$$

The basic three probability laws are:

(1) Total probability law: If the sample space Ω is completely divided into m non overlapping areas or events $A_1, A_2, \ldots, Am$ then

$$P(A_1) + P(A_2) + \ldots + A_m = P(\Omega) = 1 \qquad (11.3)$$

(2) Complementarity: It follows that if $A^{|}$ is the complement of A, that is $A^{|} = \Omega - A$, then

$$P(A^{/}) = 1 - P(A) \qquad (11.4)$$

(3) Conditional probability:

$$P(B/A) = \frac{P(A \cap B)}{P(A)} \qquad (11.5)$$

The above equation being read as P(B/A) the conditional probability that event B will occur given that event A has already occurred is P(A $\cap$ B) the joint probability that events A and B will both occur divided by P(A) the probability of event A occurrence.

Example 11.1. The values of annual rainfall at Addis Ababa from 1900 to 1990 are given in Table E11.1. Plot the time series and find the probability that the annual rainfall R in any year is less than 1000 mm, greater than 1400 mm and between 1000 and 1400.

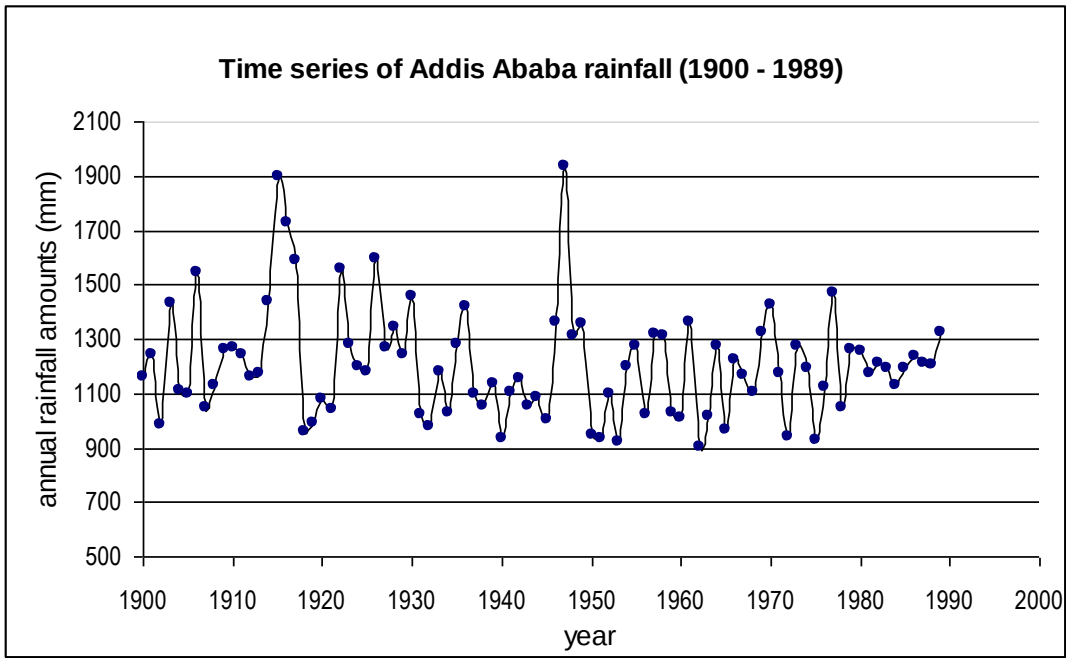
Solution. The annual rainfall R at Addis Ababa over 90 years from 1900 to 1989 is plotted in Figure E11.1. We see that there was extreme rainfalls in years 1947 (1939 mm) and in year 1962 (903 mm).

Table E11.1: Annual rainfall amounts (mm) at Addis Ababa, Ethiopia, 1900 to 1989.

year	*1900*	*1910*	*1920*	*1930*	*1940*	*1950*	*1960*	*1970*	*1980*
0	1164	1270	1077	1460	937	946	1009	1423	1255
1	1241	1244	1041	1023	1105	935	1365	1175	1175
2	986	1162	1560	976	1154	1101	904	938	1209
3	1433	1175	1282	1181	1055	922	1015	1274	1192
4	1112	1439	1200	1027	1083	1199	1275	1192	1128
5	1101	1901	1179	1283	1006	1277	963	930	1190
6	1545	1729	1595	1419	1362	1025	1225	1124	1234
7	1047	1590	1271	1099	1939	1318	1167	1473	1212
8	1133	962	1343	1054	1313	1311	1102	1045	1203
9	1265	992	1245	1134	1354	1028	1328	1262	1324

There are n = 1989 – 1900 + 1 = 90 data points. Let A be the event R < 1000 mm, B is the event R > 1400 mm. The number of events falling in these ranges are $n_A = 12$, $n_B = 13$.

So the P(A) ≈ 12 / 90 = 0.133
P(B) ≈ 13 / 90 = 0.144, and
P(1000 < R < 1400) = 1 – P(A) – P(B)
= 1- 0.133 – 0.144 = 0.723 = 65/90.



11.1.1 Frequency and probability functions.

Relative frequency function $f_s(x)$ is given by

$$f_s(x_i) = n_i / n \qquad (11.6)$$

The number of observations n_i in interval I, covering the range $[x_i - \Delta x, x_i]$. Equation 11.6 is an estimate of $P(x_i - \Delta x, \leq X \leq x_i)$, that is the probability that the random variable X will lie in the interval $[x_i - \Delta x, x_i]$.

Frequency histogram is used to display the distribution of frequencies over selected intervals. For example, the frequency histogram for the rainfall at Addis Ababa with Δx = 50 mm is given in Figure E11.2.

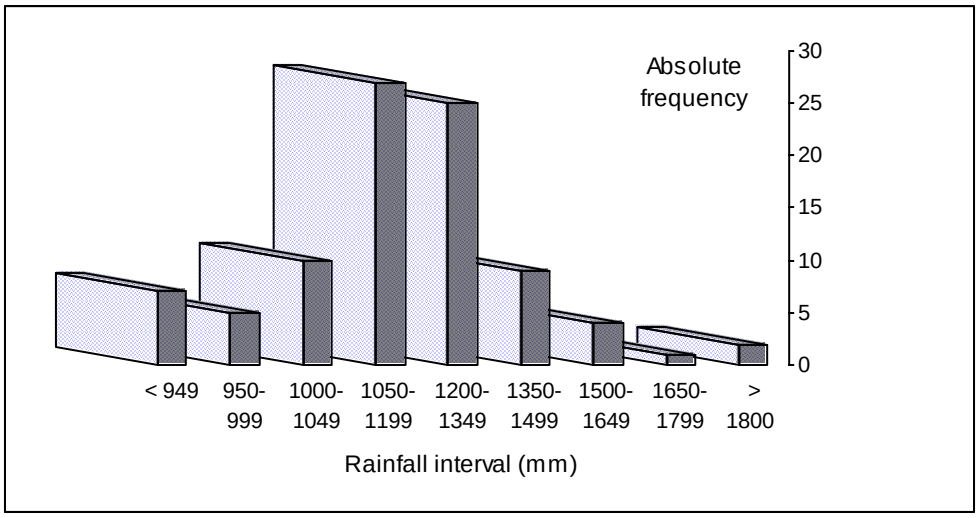


Figure E11.1: Frequency histogram of the Addis Ababa annual rainfall (1900 - 1989)

Cumulative frequency function $F_s(x)$ is given by

$$F_s(x_i) = \sum_{j=1}^{i} f_s(x_j) \tag{11.7}$$

This is an estimate of $P(X \leq x_i)$, that is the cumulative probability of x_i.

Probability density function us defined as:

$$f(x) = \lim_{\substack{n\to\infty \\ \Delta x\to 0}} \frac{f_s(x)}{\Delta x} \tag{11.8}$$

and the probability distribution function is defined as

$$F(x) = \lim_{\substack{n\to\infty \\ \Delta x\to 0}} \frac{F(x)}{\Delta x} \tag{11.9}$$

and

$$f(x) = \frac{dF(x)}{dx} \tag{11.10}$$

Note that the relative frequency, cumulative frequency and probability distribution functions are all dimensionless function varying over the range [0, 1].

However the probability density function *f(x)* has a dimension $[x]^{-1}$ and varies over the range [0, ∞) and has the property of:

$$\int_{-\infty}^{\infty} f(x)dx = 1 \qquad (11.11)$$

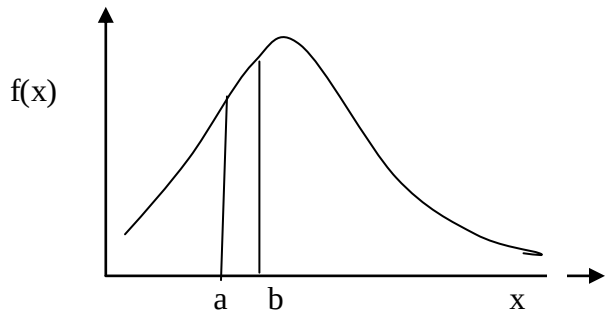


Figure 11.2: A probability density function

$$P(a \le X \le b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$$

$$P(X = a) = \int_{d}^{d} f(x)dx = 0$$

One of the best-known probability density functions is that forming the familiar bell-shaped curve for the normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(x-\mu)^2}{2\sigma^2}] \qquad (11.12)$$

Where mean μ and standard deviation σ are the parameters of the normal distribution.

Defining standardized normal variable *z* as

$$z = \frac{x-\mu}{\sigma}$$

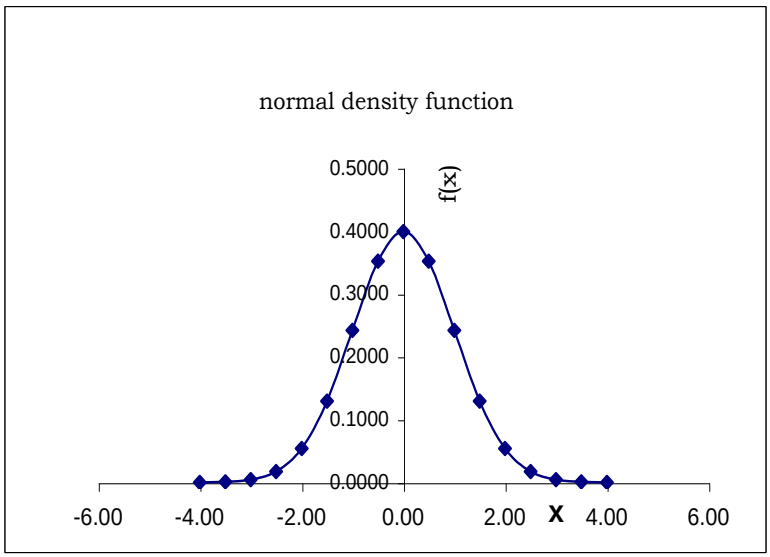
Then

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{z^2}{2}] \qquad -\infty \le z \le \infty \tag{11.13}$$

The standard normal probability distribution function is then

$$F(x) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp[-\frac{u^2}{2}] \quad du \tag{11.14}$$

x	f(x)
-4.0	0.0001338
-3.5	0.0008727
-3.0	0.0044318
-2.5	0.0175283
-2.0	0.0539910
-1.5	0.1295176
-1.0	0.2419707
-0.5	0.3520653
0.0	0.3989423
0.5	0.3520653
1.0	0.2419707
1.5	0.1295176
2.0	0.0539910
2.5	0.0175283
3.0	0.0044318
3.5	0.0008727
4.0	0.0001338



Example 11.2. The annual mean flows of a certain stream have been found to be normally distributed with mean 90 m^3/s and standard deviation 30 m^3/s. Calculate the probability that a flow larger than 150 m^3/s will occur.

Solution. Let X be the random variable describing annual mean flow of the river given above. The standardized variable

$$z = \frac{x - \mu}{\sigma} = \frac{x - 90}{30}$$

z value for flow equal to 150 m^3/s is (150-90)/30 = 2.00

The required probability is that $P(X > 150\ m^3/s) = P(z > 2.0)$

It is known that $P(z > 2.0) = 1- P(z < 2.0) = 1- F(2). = 1 – (0.5 + 0.4772) = 0.0228.$

So the probability that a flow larger than 50 m^3/s will occur is 0.0228.

11.3 Statistical parameters

The objective of statistics is to extract the essential information from a set of data. Statistical parameters are characteristics of a population, such as μ and σ. A statistical parameter is the expected value *E* of some function of a random variable.

$$E(X) = the\ mean\ X = \mu = \int_{-\infty}^{\infty} xf(x)dx \qquad (11.15)$$

The sample mean $\bar{x}$ is calculated from

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (11.16)$$

The variability of data is measured by the variance σ^2 and is defined by:

$$E((X-\mu)^2) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \qquad (11.17)$$

The sample variance s^2 is estimated by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (11.18)$$

Coefficient of variation CV is defined by

$$CV = \frac{\sigma}{\mu} \qquad (11.19)$$

And sample CV is estimated by s / $\bar{x}$

The symmetry of a distribution about the mean is measured by the coefficient of skewness γ:

$$\frac{E[(X-\mu)^3]}{\sigma^3} = \gamma = \int_{-\infty}^{\infty} (x-\mu)^3 f(x)dx \qquad (11.20)$$

Sample estimate

$$Cs = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s^3} - \qquad (11.21)$$

Example 11.3: Calculate the sample mean, sample standard deviation, and sample coefficient of skewness of the Addis Ababa rainfall given in Example 11.1.

Solution: Sample mean is calculated from Eq. (11.16)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{90}\sum_{i=1}^{90} x_i = 1206 \quad mm$$

Sample standard deviation is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{90-1}\sum_{i=1}^{90}(x_i - 1206)^2 = 203\ mm$$

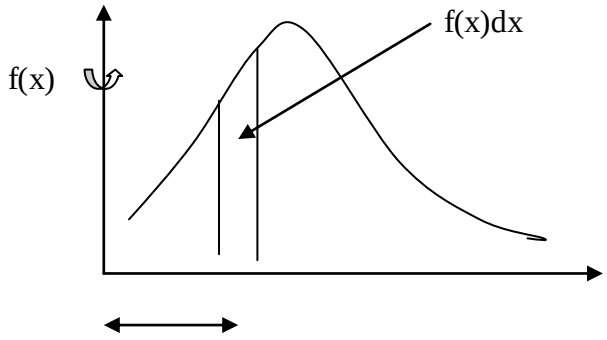
Sample skewness is

$$Cs = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s^3} = \frac{90\sum_{i=1}^{90}(x_i - 1206)^3}{(90-1)(90-2)203^3} = 1.204$$

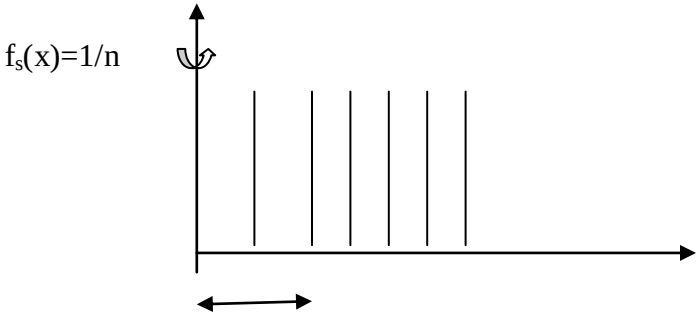
11.4 Fitting data to a probability distribution

As discussed in the previous section, a probability distribution is a function representing the probability of occurrences of a random variable. By fitting a distribution to a set of hydrologic data, a great deal of the probabilistic information in the sample can be compactly summarized in the function and its associated parameters. Two methods can be used for fitting a probability distribution: the first is the method of moment and the second is the method of maximum likelihood.

The method of moment. The principle in the method of moment is to equate the moments of the probability density function about the origin to the corresponding moments of the sample data.



x = moment arm



x_i = moment arm

Example11.4: The exponential distribution can be used to describe various kinds of hydrologic data, such as the interval times between rainfall events. The probability density function is given by

$$f(x) = \lambda e^{-\lambda x}$$

Determine the relationship between the parameter λ and the first moment about the origin μ.

Solution:

$$\mu = E(X) = \int_0^\infty x\lambda e^{-\lambda x} = \frac{1}{\lambda} = \bar{x}$$

The method of maximum likelihood. The central principle in the method of maximum likelihood is that **t**he best value of a parameter of a probability should be that value which maximizes the likelihood or joint probability of occurrence of the observed sample.

Let a sample of independent and identically distributed observations $x_1, x_2, \ldots, x_n$ of interval dx be taken. *$P(X= x_i) = f(x_i)$ = the value of the probability density for $X = x_i$ if $f(x_i)$,* and the probability that the random variable will occur in the interval including *x_i is $f(x_i)dx$*. Since it is assumed that the observations are independent, the joint probability of occurrence is simply the product of the probability of the observations, thus the joint probability of occurrence is

$$[\prod_{i=1}^{n} f(x_i)]dx^n \qquad (11.22)$$

The likelihood function L is given by

$$L = \prod_{i=1}^{n} f(x_i) \qquad (11.23)$$

Or ln(L) is

$$\ln(L) = \sum_{i=1}^{n} \ln[f(x_i)] \qquad (11.24)$$

Example 11.5: The following data are the observed times between rainfall events at a given location. Assuming that the inter-arrival time of rainfall events follows an exponential distribution; determine the parameter λ for this process by the method of maximum likelihood. The time between rainfall events (days) are: 2.2, 1.5, 0.6, 3.4, 2.1, 1.3, 0.8, 0.5, 4.0, and 2.5.

Solution:

The log-likelihood function is

$$\ln(L) = \sum_{i=1}^{n} \ln[\, f(x_i)] = \sum_{i=1}^{n} \ln[\lambda e^{-\lambda x_i}\,]$$

$$= \sum_{i=1}^{n} (\ln\lambda - \lambda x_i) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

The maximum value of ln(L) occurs when

$$\frac{\partial(\ln L)}{\partial \lambda} = 0$$

Thus

$$\frac{\partial(\ln L)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \lambda = \frac{1}{\bar{x}} = \frac{1}{1.89}(1/hr)$$

Testing goodness of fit. By comparing the theoretical and sample values of the relative frequency or the accumulative frequency function, one can test the goodness of fit of a probability distribution. In the case of the relative frequency function, the χ^2 test is used. The sample value of the relative frequency of interval *i* is calculated using Eq. (11.6)

$$f_s(x_i) = n_i / n$$

and the theoretical value is estimated from

$$p(x_i) = F(x_i) - F(x_{i-1})$$

The χ^2 test statistic χ_c^2 is given by

$$\chi_c^2 = \sum_{i=1}^{m} \frac{n[f_s(x_i) - p(x_i)]^2}{p(x_i)} \tag{11.25}$$

Where m = the number of intervals.

To describe the χ^2 test, the χ^2 probability distribution must be defined. A χ^2 distribution with *v* degrees of freedom is the distribution for the sum of squares of *v* independent normal random variables z_i; this sum is the random variable

$$\chi_v^2 = \sum_{i=1}^{v} z_i^2 \tag{11.26}$$

The χ^2 distribution function is tabulated in Annex 1. In the χ^2 test, *v = m - p- 1, where m* is the number of intervals as before, and *p* is the number of parameters used in fitting the proposed distribution. A confidence level is chosen for the test; it is often expressed as 1 - α, where α is termed the significant level. A typical value for the confidence level is 95 percent. The null hypothesis for the test is that the proposed probability distribution fits the data adequately. This hypothesis is rejected (i.e., the fit is deemed inadequate) if χ_c^2 value of in Eq. (11.25) is larger than a limiting value,

$$\chi_{v,\ 1-\alpha}^2$$

as determined from the χ^2 distribution with *v* degrees of freedom as the value having cumulative probability 1 - α.

11.5. Common probabilistic models

Many discrete probability mass functions and continuos probability density functions are used in Hydrology. The most common are the binomial, exponential, normal, gamma (Pearson Type 3), log-normal, log-gamma (log-

Pearson Type III) and Gumbel (extreme value type I). A description of some commonly used probability distribution in hydrology is given below.

11.5.1 The Binomial distribution

It is common to examine a sequence of independent events for which the outcome of each can be either a success or a failure; e.g., either the T-yr flood occurs or it does not. Such a sequence consists of Bernoulli trials, independent trials for which the probability of success at each trial is a constant *p*. The binomial distribution answers the question, what is the probability of exactly *x* successes in *n* Bernoulli trials?

The probability that there will be *x* successes followed by *n-x* failures is just the product of the probability of the *n* independent events: p^x *(1-p)* $^{n-x}$. But this represents just one possible sequence for *x* successes and *n-x* failures; all possible sequences must be considered, including those in which the successes do not occur consecutively. The number of possible ways (combinations) of choosing *x* events out of *n* possible events is given by the binomial coefficient

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{11.27}$$

Thus, the desired probability is the product of the probability of any one sequence and the number of ways in which such a sequence can occur is

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{11.28}$$

Where *x = 0, 1, 2, 3, …, n.*

The mean and the variance of *x* are given by

$$E(x) = np \tag{11.29}$$

$$Var(x) = np(1-p) \tag{1130}$$

The skewness is

$$Cs = \frac{1-2p}{\left(np(1-p)\right)^{0.5}} \tag{11.31}$$

Example 11.6: Consider the 50-yr flood, that is a flood having a return period of 50 years, T = 50 years, and then the probability of exceedence is given by P(the flood > x value) = p = 1/T = 0.02.

(a) What is the probability that at least one 50-yr flood occur during the 30-yr life time of a flood control project?

(b) What is the probability that the 100-yr flood will not occur in 10-yr? In 100 yr?

(c) In general what is the probability of having no floods greater than the T-yr flood during a sequence of T yr?

Solution: (a) The probability of occurrence in any one year (event) is p = 1/T . The probability (at least one occurrence in n events) is called the risk. Thus the risk is the sum of the probabilities of 1 flood, 2 floods, 3 floods, …, n floods occurring during the n-yr period. In other words, risk is 1- probability of no occurrence in n yr [1-P(0)].

$$P(0) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{30}{0} p^0 (1-p)^{30-0}$$

$$\begin{aligned} Risk &= 1 - P(0) \\ &= 1 - (1-p)^n \\ &= 1 - (1 - 1/T)^n \end{aligned}$$

$$Reliability = 1 - Risk$$

For the problem at hand, $p = 1/T = 1/50 = 0.02$

$$\begin{aligned} Risk &= 1 - (1 - 1/T)^n \\ &= 1 - (1 - 0.02)^{30} \\ &= 0.455 \end{aligned}$$

(b) Here p = 1/100 = 0.01, for n =10 yr, P(x =0) = 0.92. For n =100, P(x=0) =0.37.

(c) P (x =0) = $(1 - 1/T)^T$ as T gets larger, P (x=0) approaches 1/e = 0.368. The Risk of flooding in T –yrs is then 1- 0.368 = 2/3.

Example 11.7: A cofferdam has been built to protect homes in a floodplain until a major channel project can be completed. The cofferdam was built for the 20 –yr flood event. The channel project will require 3-yr to complete. What are the probabilities that:

(a) The cofferdam will not be overtopped during the 3 yr (the reliability)?
(b) The cofferdam will be overtopped in any one year?
(c) The cofferdam will be overtopped exactly once in 3 yr?
(d) The cofferdam will be overtopped at least once in 3 yr (the risk)
(e) The cofferdam will be overtopped only in the third year?

Solution:

(a) *Reliability* $= (1 - 1/T)^n = (1 - 1/20)^3 = 0.86$

(b) *Prob* $= 1/T = 0.05$

(c) *for* $p = 0.05, P(x = 1) = \binom{3}{1} p^1 (1-p)^{3-1} = 0.135$

(d) *Risk = 1 – Reliability = 0.14*

(e) *Prob* $= (1-p)(1-p)p = 0.95^2 . (0.05) = 0.045$

11.5.2 The exponential distribution:

Consider a process of random arrivals such that the arrivals (events) are independent, the process is stationary, and it is not possible to have more than one arrival at an instant in time. If the random variable *t* represents the inter-arrival time (time between events), it is found to be exponentially distributed with probability density function of

$$f(t) = \lambda\, e^{-\lambda t}, \quad t \geq 0. \tag{11.32}$$

The mean and the variance of *t:*

$$E(t) = 1/\lambda \tag{11.33}$$
$$Var(t) = 1/\lambda^2 \tag{1134}$$

The skewness is 2.

$$F(t) = 1 - e^{-\lambda t}, \quad t \geq 0.$$

(11.35)

Example 11.8. During the course of a year, about 110 independent storm events occur at a given location, and their average duration is 5.3 hours. Ignoring seasonal variations, a year of 8760 hours, calculate the storm average inter-arrival time. What is the probability that at least 4 days = 96 hr elapse between storms? What is the probability that the separation between two storms will be less than or equal to 12 hrs.

Solution: the average inter-event time is estimated by subtracting the total rainfall periods from the total hours (rainy and non-rainy) and dividing by the number of rainfall events.

The inter-event time = (8760 –110*5.3)/110 = 74.3 hr. and $\lambda = 1/74.3 = 0.0135$.

The probability that at least 4 days = 96 hr elapse between storms is Prob ($t \geq 96$) = 1 – F (96) = $e^{(-0.0135*96)}$.

The probability that the separation between two storms will be less than or equal to 12 hrs is Prob ($t < 12$) = F (12) = 1- $e^{(-0.0135*12)}$. = 0.15

11.5.3 Extreme value distribution

Many time interests exist in extreme events such as the maximum peak discharge of a stream or minimum daily flows. The extreme value of a set of random variables is also a random variable. The probability distribution of this extreme value random variable will in general depend on the sample size and the parent distribution from which the sample was obtained.

The study of extreme hydrological events involves the selection of a sequence of the largest or smallest observations from sets of data. For example, the study of peak flows uses just the largest flow recorded each year at a gauging station - for 30 years of data only 30 points are selected.

Gumbel distribution: Extreme Value Type I

The extreme Value Type I (EVI) probability distribution is

$$F(x) = \exp(-\exp(-\frac{x-u}{\alpha})) \quad -\infty \leq x \leq \infty \tag{11.36}$$

The parameters are estimated by:

$$\alpha = \frac{\sqrt{6}s}{\pi}, \quad u = \bar{x} - 0.5772\alpha \tag{11.37}$$

A reduced variate y can be defined as:

$$y = \frac{x-u}{\alpha} \tag{11.38}$$

Substituting the reduced variate into Eq.(11.37) yields

$$F(x) = exp(-exp(-y)) \tag{11.39}$$

Solving for y:

$$y = -\ln\left[\ln\left(\frac{1}{F(x)}\right)\right] \tag{11.40}$$

The return period and the cumulative probability function is related by

$P(X \geq x_T) = 1/T$

$= 1- P(X < x_T)$

$= 1- F(x_T)$

Also

$F(x_T) = (T-1)/T$

And finally we get *y* in terms of return period for EVI distribution:

$$y_T = -\ln\left[\ln\left(\frac{T}{T-1}\right)\right] \tag{11.41}$$

and related to x_T by

$$x_T = u + \alpha y_T \qquad (11.42)$$

Heavy storm are most commonly modeled by the EVI distribution.

Example 11.9. Annual maximum values of 10-minutes-duration rainfall at Chicago, Illinois, from 1913 to 1947 are given in Table E11.9. Develop a model for storm rainfall frequency analysis using Extreme value Type I distribution and calculate the 5-, 10-, and 50 – year return period maximum values of 10-minute rainfall at Chicago.

Table E11.2. Annual maximum 10-minutes rainfall (mm) at Chicago, Illinois, 1913-1947

Year	10-min R (mm)	Year	10-min R (mm)
1913	12	1930	8
1914	17	1931	24
1915	9	1932	24
1916	15	1933	20
1917	10	1934	16
1918	12	1935	18
1919	19	1936	28
1920	13	1937	16
1921	19	1938	13
1922	14	1939	16
1923	20	1940	9
1924	17	1941	18
1925	17	1942	14
1926	17	1943	23
1927	15	1944	17
1928	22	1945	17
1929	12	1946	16
		1947	15

Solution: The sample moments calculated from the data in Table E11.2. The mean is 16.5 mm and the standard deviation is 4.5 mm.

The parameters of the EVI distribution is then:

$$\alpha = \frac{\sqrt{6}s}{\pi} = \frac{\sqrt{6}4.5}{\pi} = 3.5$$

$$u = \bar{x} - 0.5772\alpha = 16.5 - 0.5772 * 3.5 = 14.4$$

The probability model is

$$F(x) = \exp(-\exp(-\frac{x-u}{\alpha})) \quad -\infty \le x \le \infty$$

$$F(x) = \exp(-\exp(-\frac{x-14.4}{3.5})) \quad 0 \le x \le \infty$$

To determine the values of x_T for various of return period T, it is convenient to use the reduced variate y_T given by Eq. (11.41). For T = 5 years

$$y_T = -\ln\left[\ln\left(\frac{T}{T-1}\right)\right] = -\ln\left[\ln\left(\frac{5}{5-1}\right)\right] = 1.500$$

And Eq. (11.42) yields

$$\begin{aligned} x_T &= u + \alpha y_T \\ &= 14.4 + 3.5 \times 1.500 \\ &= 19.6 \text{ mm.} \end{aligned}$$

By the same method, the 10-, and 50- year values are estimated to be 22.4 mm and 28.2 mm respectively.

It may be noted from the data in Table E11.2 that the 50-year return period rainfall was equaled once in the 35 years data (in 1936), and that the 10-year return period rainfall was exceeded four times during this period.

Extreme Value Type III (Weibull) distribution – Low flow analysis:

Weibull distribution has found greatest use in hydrology as the distribution of low stream flows. It is defined as:

$$f(x) = \alpha(x-\varepsilon)^{\alpha-1}(\beta-\varepsilon)^{-\alpha} \exp(-((x-\varepsilon)/(\beta-\varepsilon))^{\alpha}), \quad x \ge 0, \quad \alpha,\beta > 0. \tag{11.43}$$

The cumulative Weibull, F(x), is given by

$$F(x) = 1 - \exp(-((x-\varepsilon)/(\beta-\varepsilon))^{\alpha}) \tag{11.44}$$

The mean and the variance of the distribution are:

$$E(X) = \varepsilon + (\beta - \varepsilon)\Gamma(1+1/\alpha) \tag{11.45}$$

$$Var(X) = (\beta - \varepsilon)^2[\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)] \tag{11.46}$$

Where ε is a displacement parameter to create 0 as the lower bound of the parameter x.

The estimate of the parameters is done using:

$$\beta = \mu + \sigma\, A(\alpha) \tag{11.47}$$

$$\varepsilon = \beta - \sigma\, B(\alpha) \tag{11.48}$$

Where A(α) and B(α) is taken from Table 11.2.

Example 11.10: The minimum annual daily discharge on a stream are found to have an average of 4.6 m^3/s, a standard deviation of 1.8 m^3/s and a coefficient of skew of 1.4. Evaluate the probability of the annual mean flow being less than 3.69 m^3/s.

Solution: Weibull distribution is used here for low flow analysis

Using estimated coefficient of skewness value γ = 1.4 , then the corresponding parameters are read from Table 11.2.

$1/\alpha = 0.79$, $A(\alpha) = 0.098$, $B(\alpha) = 1.36$

estimates of $\alpha = 1.266$,

Table 11.2. Values of A(α) and B(α)

γ (Skewness)	1/α	A(α)	B(α)
-1.000	0.02	0.446	40.005
-0.971	0.03	0.444	26.987
-0.917	0.04	0.442	20.481
-0.867	0.05	0.439	16.576
-0.638	0.10	0.425	8.737
-0.254	0.20	0.389	4.755
0.069	0.30	0.346	3.370
0.359	0.40	0.297	2.634
0.631	0.50	0.246	2.159
0.896	0.60	0.193	1.815
1.160	0.70	0.142	1.549
1.430	0.80	0.092	1.334
1.708	0.90	0.044	1.154
2.000	1.00	0.000	1.000
2.309	1.10	-0.040	0.867
2.640	1.20	-0.077	0.752
2.996	1.30	-0.109	0.652
3.382	1.40	-0.136	0.563
3.802	1.50	-0.160	0.486
4.262	1.60	-0.180	0.418
4.767	1.70	-0.196	0.359
5.323	1.80	-0.208	0.308
5.938	1.90	-0.217	0.263
6.619	2.00	-0.224	0.224
7.374	2.10	-0.227	0.190
8.214	2.20	-0.229	0.161

$$\beta = \mu + \sigma A(\alpha) = 4.6 + 1.8(0.098) = 4.8$$

$$\varepsilon = \beta - \sigma B(\alpha) = 4.8 - 1.8(1.36) = 2.4$$

Prob ($X \leq 3.7$) = $F(3.7)$ is give by

$$F(3.7) = 1 - \exp(-((3.7 - 2.4)/(4.8 - 2.4))^{1.266})$$

$$= 0.368$$

11.5.4 Frequency Analysis using Frequency Factor

Calculating the magnitude of extreme events by the method outlined in Example 11.9 requires that the probability distribution function be invertible, that is, for a value for *T* or *[F(x_T) = T/(T-1)]*, the corresponding value of x_T can be determined. Some probability distribution functions are not readily invertible, including the Normal and Pearson Type III distributions, and an alternative method of calculating the magnitudes of extreme events is required for these distributions.

The magnitude of x_T of a hydrological event may be expressed as:

$$x_T = \mu + K_T \sigma \tag{11.49}$$

which may be approximated by

$$x_T = \bar{x} + K_T s \tag{11.50}$$

in the event that the variable analyzed is $y = log\ x$, then the same method is applied to the statistics for the logarithms of the data, using

$$y_T = \bar{y} + K_T s \tag{11.51}$$

and the required value of x_T is found by taking antilog of y_T.

The Frequency factor for Normal Distribution

The frequency factor can be expressed from Eq. (11.50) as

$$K_T = \frac{x_T - \mu}{\sigma} \tag{11.52}$$

This is the same as the standard normal variable *z* defined in this chapter.

The value of *z* corresponding to an exceedence probability of $p = 1/T$ can be calculated by finding the value of an intermediate variable *w*:

$$w_T = \left[\ln\left(\frac{1}{p^{2+}} \right) \right]^{0.5} \quad (0 < p \le 0.5) \tag{11.53}$$

then calculating *z* using the approximation

$$z = w - \frac{2.515517 + 0.802853w + 0.010328w^2}{1 + 1.432788w + 0.189269w^2 + 0.001308w^3} \qquad (11.54)$$

When p > 0.5, 1-p is substituted for p in Eq. (11.53) and the value of z computed by Eq.(11.54).

The Frequency factor for Extreme value distribution Type I (EVI)

For the EVI distribution the frequency factor is given by

$$K_T = -\frac{\sqrt{6}}{\pi}\left(0.5772 + \ln\left[\ln\left(\frac{T}{T-1}\right)\right]\right) \qquad (11.55)$$

Extreme value distribution Type II (EVII):

For the Extreme value distribution Type II (EVII) the logarithm of the variate follows the EVI distribution. For this case Eq.(11.51) is used to calculate y_T, using the value of K_T from Eq.(11.55).

Log-Pearson Type III distribution.

Log-Pearson Type III distribution the first step is to take the logarithms of the hydrologic data, y = log x. Then the mean $\bar{y}$, the standard deviation s_y and coefficient of skewness C_s are calculated for the logarithms of the data. The frequency factor depends on the return period T and the coefficient of skewness C_s. When C_s= 0, the frequency factor is equal to the standard normal variable z. When $C_s \neq 0$, K_T is approximated by

$$K_T = z - (z^2 - 1)k + \frac{1}{3}(z^3 - 6z)k^2 - (z^2 - 1)k^3 + zk^4 + \frac{1}{3}k^5 \qquad (11.56)$$

where $k = C_s/6$.

The value of z for a given return period can be calculated using Eq.(11.53) & (11.54).

Example11.11. The annual maximum daily discharge measured on the Beressa river at Debere Birhan gauging site are given in Table E11.3. The Beresa river is a tributary of Jemma River which is lying in Abay basin and has watershed area of 220 km^2 . Calculate the 5- and 50- year return period annual maximum discharge of the Beressa river at Deberebirhan using lognormal, EVI, and log-Pearson Type III distributions.

Table E11.3. Maximum daily discharge of the Beresa River (m^3/s)

Year	Q (m^3/s)	Year	Q (m^3/s)
1961	60.4	1980	84.4
1962	59.5	1981	missed
1963	82.5	1982	180.0
1964	90.0	1983	107.1
1965	32.8	1984	66.8
1966	75.0	1985	92.0
1967	58.0	1986	89.4
1968	112.5	1987	17.9
1969	151.4	1988	67.7
1970	80.7	1989	37.4
1971	144.0	1990	53.5
1972	63.1	1991	56.1
1973	81.3	1992	54.5
1974	163.0	1993	56.6
1975	83.7	1994	252.2
1976	140.0	1995	148.7
1977	58.0	1996	126.0
1978	74.5	1997	91.9
1979	101.0		

Solution: Let X be the maximum annual discharge, then the mean $\bar{x}$ = 91.48 m^3/s , the standard deviation s_x = 46.89 m^3/s, and coefficient of skewness C_s = 1.4. For the log 10 data Y = Log(X), then the mean $\bar{y}$ = 1.91 m^3/s , the standard deviation s_y = 0.22 m^3/s, and coefficient of skewness C_s = -0.3971.

Lognormal distribution.. The frequency factor can be obtained from Eq.(11.55). For T = 50 years, K_T = 2.054

Then

$$y_T = \bar{y} + K_T s$$

$$y_{50} = 1.91 + 2.054 * 0.22$$

$$= 2.36 \quad m^3 / s$$

$$x_{50} = 10^{2.36} = 229\ m^3/s$$

Similarly for T = 5 years, K_T = 0.842.

$$y_T = \bar{y} + K_T s$$

Then

$$y_5 = 1.91 + 0.842 * 0.22$$
$$= 2.095 \quad m^3 / s$$
$$x_5 = 10^{2.095} = 124\ m^3/s$$

EVI distribution. The frequency factor can be obtained from Eq.(11.54). For T = 50 years, $K_T = 2.592$

Then
$$x_T = \bar{x} + K_T s$$

$$x_{50} = 91.48 + 2.592 * 46.89$$
$$= 213 \quad m^3 / s$$

Similarly for T = 5 years, $K_T = 0.719$.

Then
$$x_T = \bar{x} + K_T s$$

$$x_0 = 91.48 + 0.719 * 46.89$$
$$= 125 \quad m^3 / s$$

Log-Pearson Type III distribution. For $C_s = -0.3971$, the value of K_{50} is obtained using Eq. (11.56), $K_{50} \cong 1.834$,

$$y_T = \bar{y} + K_T s$$
$$y_{50} = 1.91 + 1.834 * 0.22$$
$$= 2.313$$
$$x_{50} = 10^{2.313} = 205\ m^3/s$$

Similarly for T = 5 years, $K_T = 0.855$, and $x_5 = 125\ m^3/s$.

In summary:

	Return period	
	5 years	50 years
Lognormal	124	229
EVI	125	213
Log-Pearson Type III	125	205

In this example the values of Beressa flood at Deberebirhan estimated by the lognormal, EVI and log-pearson Type III distribution are very close to each other.

Commonly accepted practice first the data has to be fitted to candidate distributions and select the model that describes the observed data very well and apply the selected model in estimating the required flood of a given return period.

11.6 Probability plot

As a check that a probability distribution fit a set of hydrological data, the data may be plotted on specially designed probability paper, or using a plotting scale that linearizes the distribution function. The plotted data are then fitted with a straight line for interpolation purposes.

Plotting position refers to the probability value assigned to each piece of data to be plotted. Numerous methods have been proposed for the determination of plotting positions. Most plotting position formulas are represented by :

$$P(X \geq x_m) = \frac{m-b}{n+1-2b} = \frac{1}{T} \qquad (11.57)$$

Where m = is the rank 1 for the maximum, and n is for the minimum value
n = the number of data points used in the analysis.
b = 0.5 Hazen's plotting position
b = 0.3 Chegodayev's plotting position
b = 3/8 Blom's plotting position
b = 1.3 Tukey's plotting position
b = 0.44 Gringorten's plotting position

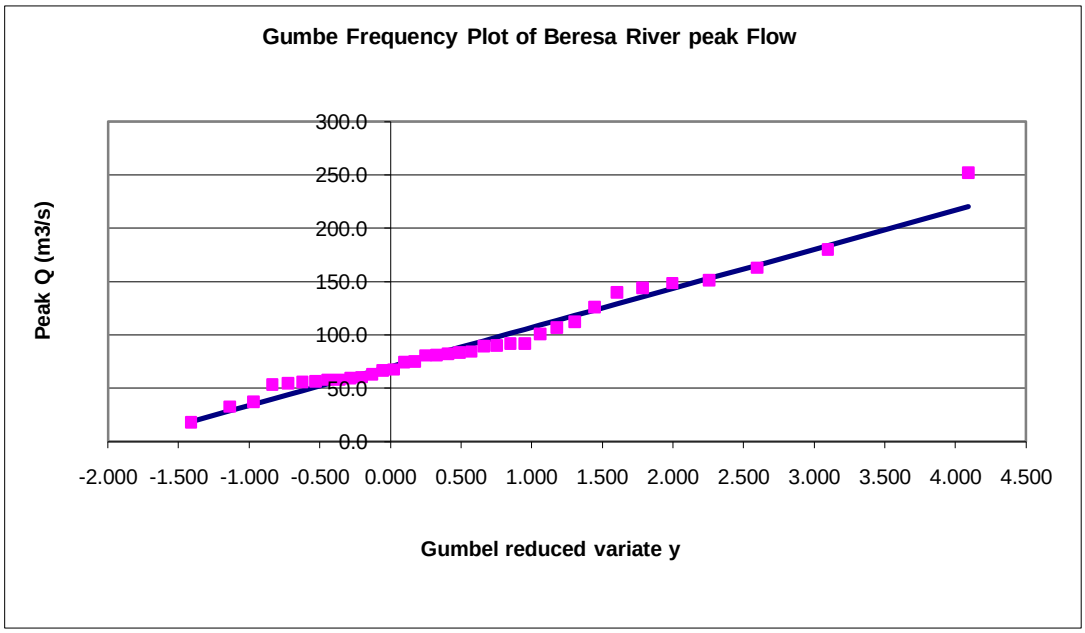
Example 11.12. Considering that the probability distribution of the maximum flow of the Berassa river used in Example 11.11 follows the Gumbel distribution, plot the values on Gumbel paper.

Solution: The data is ranked first col [4], and the plotting position method is chosen, b = 0.4 Gringorten's plotting position, and the return period is calculated for the data Col [5]. Then the reduced variate y_T for the Gumbel distribution is calculated for the T associated

[1]	*[2]*	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Year	*Flow*	Rank	Flow	Empirical	Emprical CDF	Gumbel distribution		
				T = (n+1-2a)/(m-a)	1-1/T	Reduced	Predicted	Observed

						variate y	flow	Flow
1961	*60.4*	1	252.2	60.3	0.983	4.091	220.1	252.2
1962	*59.5*	2	180.0	22.6	0.956	3.097	183.7	180.0
1963	*82.5*	3	163.0	13.9	0.928	2.597	165.4	163.0
1964	*90.0*	4	151.4	10.1	0.901	2.256	153.0	151.4
1965	*32.8*	5	148.7	7.9	0.873	1.996	143.4	148.7
1966	*75.0*	6	144.0	6.5	0.845	1.783	135.7	144.0
1967	*58.0*	7	140.0	5.5	0.818	1.603	129.1	140.0
1968	*112.5*	8	126.0	4.8	0.790	1.445	123.3	126.0
1969	*151.4*	9	112.5	4.2	0.762	1.305	118.2	112.5
1970	*80.7*	10	107.1	3.8	0.735	1.177	113.5	107.1
1971	*144.0*	11	101.0	3.4	0.707	1.060	109.2	101.0
1972	*63.1*	12	92.0	3.1	0.680	0.951	105.2	92.0
1973	*81.3*	13	91.9	2.9	0.652	0.849	101.5	91.9
1974	*163.0*	14	90.0	2.7	0.624	0.753	97.9	90.0
1975	*83.7*	15	89.4	2.5	0.597	0.661	94.6	89.4
1976	*140.0*	16	84.4	2.3	0.569	0.573	91.4	84.4
1977	*58.0*	17	83.7	2.2	0.541	0.489	88.3	83.7
1978	*74.5*	18	82.5	2.1	0.514	0.407	85.3	82.5
1979	*101.0*	19	81.3	1.9	0.486	0.327	82.4	81.3
1980	*84.4*	20	80.7	1.8	0.459	0.249	79.5	80.7
1982	*180.0*	21	75.0	1.8	0.431	0.172	76.7	75.0
1983	*107.1*	22	74.5	1.7	0.403	0.096	73.9	74.5
1984	*66.8*	23	67.7	1.6	0.376	0.021	71.2	67.7
1985	*92.0*	24	66.8	1.5	0.348	-0.054	68.4	66.8
1986	*89.4*	25	63.1	1.5	0.320	-0.129	65.7	63.1
1987	*17.9*	26	60.4	1.4	0.293	-0.206	62.9	60.4
1988	*67.7*	27	59.5	1.4	0.265	-0.283	60.0	59.5
1989	*37.4*	28	58.0	1.3	0.238	-0.363	57.1	58.0
1990	*53.5*	29	58.0	1.3	0.210	-0.445	54.1	58.0
1991	*56.1*	30	56.6	1.2	0.182	-0.532	50.9	56.6
1992	*54.5*	31	56.1	1.2	0.155	-0.624	47.6	56.1
1993	*56.6*	32	54.5	1.1	0.127	-0.724	43.9	54.5
1994	*252.2*	33	53.5	1.1	0.099	-0.836	39.8	53.5
1995	*148.7*	34	37.4	1.1	0.072	-0.968	35.0	37.4
1996	*126.0*	35	32.8	1.0	0.044	-1.138	28.8	32.8
1997	*91.9*	36	17.9	1.0	0.017	-1.411	18.8	17.9

with the plotting position and the flow data col [7]. The Gumbel predicted flow is done in Col. [8] using Eq.(11.42). Then plot the predicted col [8] and observed col [9] flows on the Gumbel scale with x- axis being col [7]. It is seen that the data fits well the Gumbel distribution except at the extreme value.



11.7. Testing for outliers

Outliers are data points that depart significantly from the trend of the remaining data. The retention and deletion of these outliers significantly affect the magnitude of the statistical parameters computed from the data, especially small sample size.

Water Resources Council (1981) recommends that if the station skew is greater than +0.4, tests for high outliers are considered first; if the station skew is less than –0.4, test for low outliers are considered first. Where the station skew is between ± 0.4, test for both high and low outliers should be applied before eliminating any outliers from the date set.

The following frequency equation can be used to detect high outliers:

$$y_H = \bar{y} \pm K_n s_y \qquad (11.58)$$

Where; y_H = the high (+) / low (-) outlier threshold in log units

K_n = values are Given in Table 11.3 for one sided test that detect outlier at the 10-percent level of significance in normally distributed data.

If the logarithms of the values in a sample are greater / less than y_H in the above

equation, then they are considered high / low outlier.

Table 11.3 Outlier test Kn value

Sample size n	K_n	Sample size n	K_n	Sample size n	K_n	Sample size n	K_n
10	2.036	24	2.467	38	2.661	60	2.837
11	2.088	25	2.486	39	2.671	65	2.866
12	2.134	26	2.502	40	2.682	70	2.893
13	2.175	27	2.519	41	2.692	75	2.917
14	2.213	28	2.534	42	2.700	80	2.940
15	2.247	29	2.549	43	2.710	85	2.961
16	2.279	30	2.563	44	2.719	90	2.981
17	2.309	31	2.577	45	2.727	95	3.000
18	2.335	32	2.591	46	2.736	100	3.017
19	2.361	33	2.604	47	2.744	110	3.049
20	2.385	34	2.616	48	2.753	120	3.078
21	2.408	35	2.628	49	2.760	130	3.104
22	2.429	36	2.639	50	2.768	140	3.129
23	2.448	37	2.650	55	2.804		

Example 11.13. Check the data given in Example 11.11 for outliers?

Solution: The mean and standard deviation of log transformed peak flow with sample size n = 36 are 1.9089 m^3/s and 0.2217 m^3/s respectively. For n = 36 the value of K_n = 2.639.

$$y_H = \bar{y} \pm K_n s_y$$

$$y_H = 1.9 \pm 2.639 * 0.2217$$

$$= 2.485 \ \& \ 1.31933$$

Corresponding to Q = 305 m^3/s and 21 m^3/s.

The maximum value is 257 m^3/s and the minimum is 17 m^3/s. It is seen that low outlier is found but it is very near to the boundary of 21 m^3/s. So the data may be acceptable in a sense that no outlier is found. However, one should check the reason behind the low outlier, by comparing to the rainfall in the rainy months of June, July, and August of the year 1987.

11.8 Practice problems

11.1 The values of annual rainfall at Addis Ababa from 1900 to 1990 are given in Table Find the mean, standard deviation, coefficient of variation, and skewness for two period: (a) for data from 1900 – 1945, and 1946 –1990. Fit the data using normal distribution and check its goodness of fit over the two periods indicated. Plot the data normal probability paper to check its fitness.

11.2 Fit the data of the peak flood of the Beresa river (given in Example 11.11) using log-Pearson Type III distribution. Plot is in log-Pearson paper

11.3 The record of annual peak discharge at a stream gaging station is as follows:

year	1961	1962	1963	1964	1965	1966	1967	1968	1969
Q (m^3/s)	45.3	27.5	16.9	41.1	31.2	19.9	22.7	59.0	35.4

Determine using the lognormal distribution:

(a) The probability that an annual flood peak of 42.5 m^3/s will not be exceeded.
(b) The return period of the dischrge of 42.5 m^3/s
(c) The magnitude of a 20-year flood