Technical Assistance
HYDROLOGY PROJECT
World Bank & Government of The Netherlands funded

Training module # SWDP - 43

Statistical Analysis with Reference to Rainfall and Discharge Data

New Delhi, February 2002

CSMRS Building, 4th Floor, Olof Palme Marg, Hauz Khas,
New Delhi – 11 00 16 India
Tel: 68 61 681 / 84 Fax: (+ 91 11) 68 61 685
E-Mail: hydrologyproject@vsnl.com

DHV Consultants BV & DELFT HYDRAULICS

with
HALCROW, TAHAL, CES, ORG & JPS

Table of contents

1. Module context

While designing a training course, the relationship between this module and the others, would be maintained by keeping them close together in the syllabus and place them in a logical sequence. The actual selection of the topics and the depth of training would, of course, depend on the training needs of the participants, i.e. their knowledge level and skills performance upon the start of the course.

2. Module profile

Title	:	Statistical Analysis with Reference to Rainfall and Discharge Data
Target group	:	HIS function(s): ……
Duration	:	x session of y min
Objectives	:	After the training the participants will be able to*:*
Key concepts	:	•
Training methods	:	Lecture, exercises
Training tools required	:	Board, flipchart
Handouts	:	As provided in this module
Further reading and references	:	

3. Session plan

No	Activities	Time	Tools
1	***Preparations***		
2	***Introduction***:	min	OHS x
	Exercise	min	
	Wrap up	min	

4. Overhead/flipchart master

5. Handout

Add copy of the main text in chapter 7, for all participants

6. Additional handout

These handouts are distributed during delivery and contain test questions, answers to questions, special worksheets, optional information, and other matters you would not like to be seen in the regular handouts.

It is a good practice to pre-punch these additional handouts, so the participants can easily insert them in the main handout folder.

7. Main text

Contents

Statistical Analysis with Reference to Rainfall and Discharge Data

1 Introduction

Terminology

A **hydrologic process** is defined as any phenomenon concerning the occurrence and movement of water near the earth's surface continuously changing in time and/or space. If these phenomena are observed at intervals or continuously, **discrete**, respectively, **continuous series** are created, $\rightarrow$ with time: discrete and continuous time series. One single series element is an **outcome** of the process. A set of outcomes is called a **realisation**, while the set of all possible outcomes is the **ensemble**.

The variation within hydrological processes may be deterministic or stochastic. In a **deterministic process** a definite relation exists between the hydrologic variable and time (or space). The functional equation defines the process for the entire time (or space) of its existence. Each successive observation does **not** represent new information about the process. This, in contrast to a **stochastic process**, which evolves, entirely or in part, according to a random mechanism. It means that future outcomes of the process are not exactly predictable. The hydrologic variable in such cases is called a **stochastic variable**, i.e. a variable whose values are governed by the laws of chance. Its behaviour is mathematically described by probability theory.

The elements, creating a stochastic process, may be **dependent** or **independent**, resulting in **a non-pure random**, respectively, a **pure random** process.

A stochastic process can either be **stationary** or **non-stationary**, i.e. homogeneous or non-homogeneous in time and/or space. Stationary processes are distinguished into **strictly** and **weakly** stationary processes.

A process is said to be **strictly stationary** if all its statistical properties which characterise the process, are unaffected by a change in the origin (time and or space). For a time-process this reads: the joint probability distribution of $x(t_1)$, $x(t_2)$, …, $x(t_n)$ is identical to the joint probability distribution of $x(t_1+\tau)$, $x(t_2+\tau)$, …, $x(t_n+\tau)$ for any n and τ, where τ is a time lag. If instead of the joint probability density function only the first m-moments of that function are independent of time (space) the process is called m^{th} order stationary.

Weak stationarity means that only the lower order moments of the distribution function (order ≤ 2, i.e. the mean and the covariance function) fulfil the property of being independent of time. This is also called stationarity in a **wide sense**. (Note that the terminology stationary/non-stationary is used when dealing with homogeneity or non-homogeneity in **time**).

In practice only a limited set of outcomes, a limited series, is available. Based on this sample set the behaviour of the process is estimated: sample versus population. The elements in a hydrological series may be analysed according to **rank of magnitude** and according to the **sequence of occurrence**. Ranking of elements forms the basis of statistics, the classical **frequency analysis**, thereby ignoring the order of occurrence. In contrast to ranking, the study of the sequence of occurrence presumes that past outcomes of the process may influence the magnitude of the present and the future outcomes. Hence the dependency between successive elements in the series is not ignored: **time series analysis**.

About this module

In this module a review is presented of statistics as applied to hydrology to analyse e.g. rainfall or discharge data. With statistics one describes rather than explains features of hydrological processes. Statements are made based on a sample from the entire population of the hydrological variable of concern. With statistics one describes variables only in probabilistic terms for reasons that the cause and effect relation of the physical process is insufficiently known and also because our description is based on a small part of the entire range of outcomes on the variable.

Statistics provides powerful tools to describe hydrological variables, but one should apply it with care. An important condition the series to be subjected to statistical analysis should fulfil is **stationarity**. To judge whether this condition is fulfilled, knowledge is required of the nature of the hydrological variable(s) of concern. The following components are generally distinguished in hydrological time series, see also Figure 1.1:

- **Deterministic components,** including:
  - Transient component, due to natural or man made changes, which can be a jump, in case of a sudden change in the conditions or a trend, linear or non-linear, due to a gradual change
  - Periodic component, e.g. due to the annual solar cycle

- **Stochastic component:**
  - Stochastic dependent part, where the new value is related to one or more predecessors, e.g. due to storage effects
  - Stochastic independent or random part.

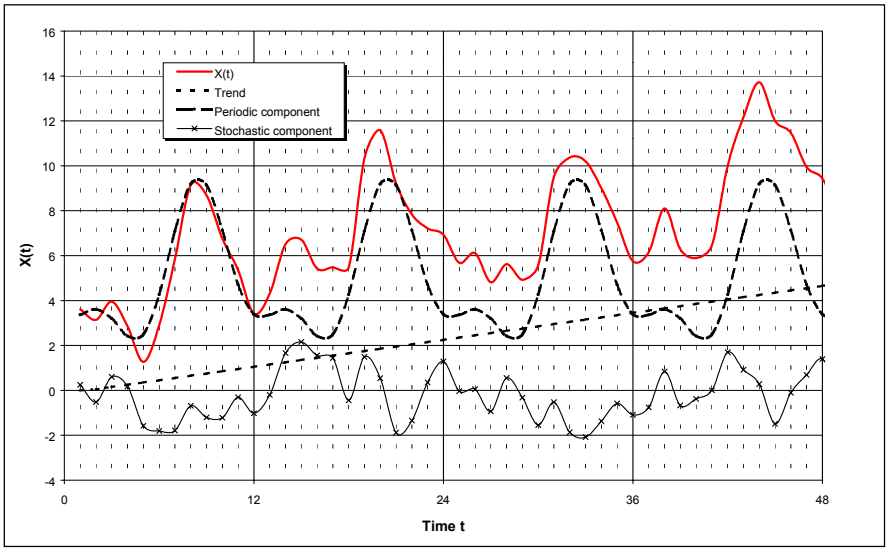

Figure 1.1: Components of a hydrological time series

Figure 1.1 displays a monthly time series, with a clear linear trend and a strong periodic component, repeating itself every year. It will be clear that a series as shown in Figure 1.1 does not fulfil the stationarity condition, the mean value gradually shifts due to the trend. Even with the trend removed the probability distribution changes from month to month due to the existence of the periodic component, again not fulfilling the stationarity condition. If one also eliminates the periodic component in the mean value a process with a stationary mean value is obtained, but still this may not be sufficient as generally also second or higher order moments (variance, covariance, etc.) show periodicity. Therefore, hydrological time series

with time intervals less than a year should not be subjected to statistical analysis. Annual values generally do not have the problem of periodicity (unless spectral analysis shows otherwise due to some over-annual effect) and are fit for statistical analysis, provided that transient components are not present or have been eliminated.

Now, returning to our monthly series, periodicity is avoided if the months are considered separately, that is e.g. if only the values of July of successive years are considered. Similarly, if seasonal series are available, one should consider one season at a time for statistical analysis, i.e. the same season for a number of years.

To illustrate the above considerations monthly rainfall and its statistics of station Chaskman are shown in Figures 1.2 and 1.3. As can be observed from Figure 1.3, there is a strong periodic component in the time series; the mean and standard deviation vary considerably from month to month.

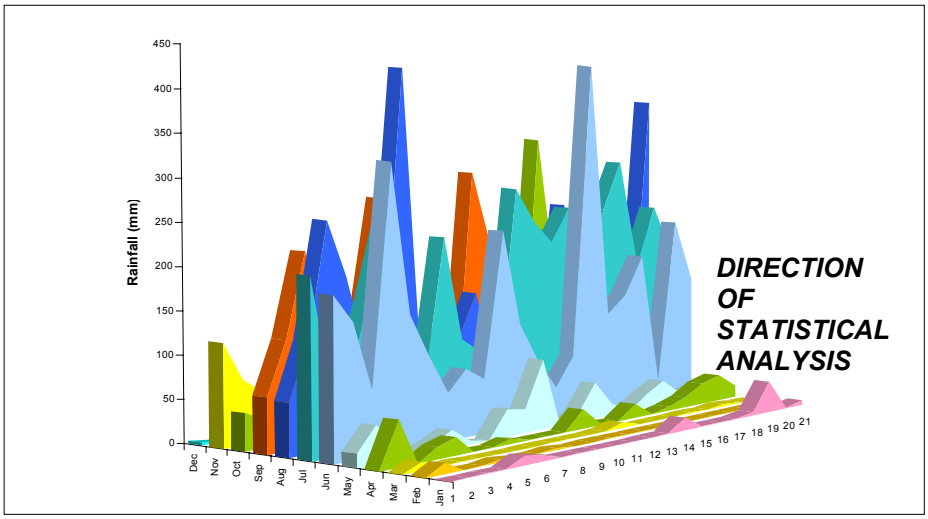

Figure 1.2: 3-D plot of monthly rainfall of station Chaskman

If one would combine the rainfall values of all months one assumes that their probability distribution is the same, which is clearly not so. To fulfil the stationarity condition, statistics is to be applied to each month separately, see Figure 1.2

A series composed of data of a particular month or season in successive years is likely to be serially uncorrelated, unless over-annual effects are existent. Hence, such series will be fully random. Similar observations apply to annual maximum series. It implies that the time sequence of the series considered is unimportant. Above considerations are typical for statistical analysis.

In this module statistics is discussed and the following topics will be dealt with:

- Description of data sets
- Probabilistic concepts
- Discrete and continuous probability distributions
- Estimation of distribution parameters
- Making statistical inference

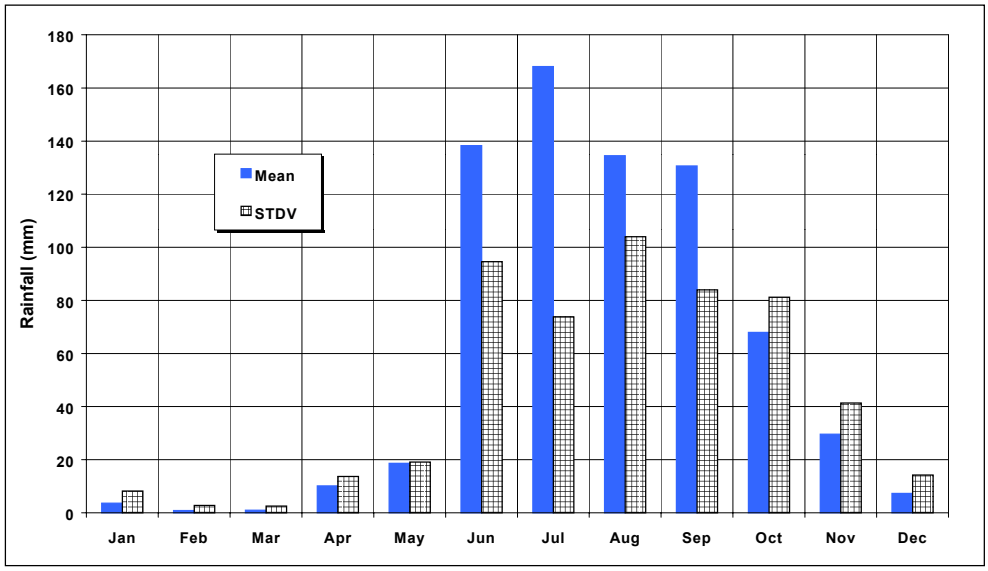

Figure 1.3: Mean and standard deviation of monthly rainfall series of station Chaskman, period 1977 - 1998

2 Description of Datasets

2.1 General

In this sub-section on basic statistics attention will be given to:

- Graphical presentation of data
- Measures of central tendency
- Measures of dispersion
- Measure of asymmetry: skewness
- Measure of peakedness: kurtosis
- Percentiles
- Box plots
- Covariance and correlation coefficient

2.2 Graphical representation

For graphical presentation of the distribution of data the following options are discussed:

- Line diagram or bar chart
- Histogram
- Cumulative relative frequency diagram
- Frequency and duration curves

Note: prior to the presentation of data in whatever frequency oriented graph, it is essential to make a time series plot of the data to make sure that a strong trend or any other type of inhomogeneity, which would invalidate the use of such presentation, does not exist.

Line Diagram or Bar Chart

The occurrences of a **discrete** variate can be classified on a line diagram or a vertical bar chart. In this type of graph, the horizontal axis gives the values of the discrete variable, and the occurrences are represented by the heights of vertical lines. The horizontal spread of

these lines and their relative heights indicate the variability and other characteristics of the data. An example is given in Figure 2.1, where the number of occurrences that in one year the monthly rainfall at Chaskman will exceed 100 mm is presented. The period presented refers to the years 1978 – 1997.

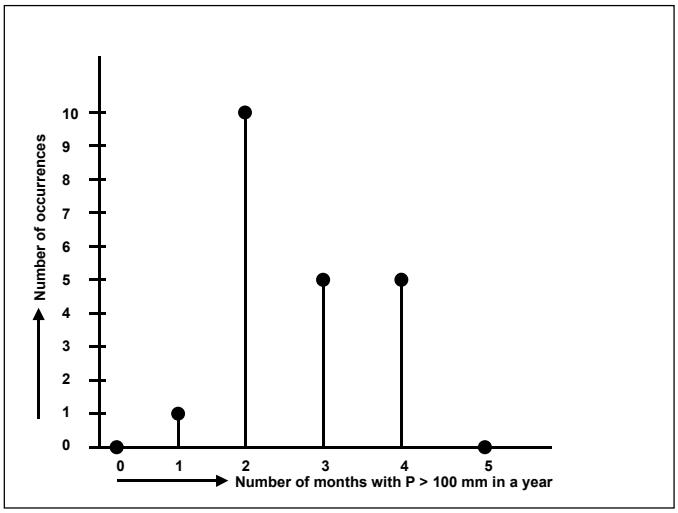

Figure 2.1:
Line diagram of number of months in a year with rainfall sum > 100 mm for period 1978 - 1997

If the number of entries on the horizontal axis would have been small, it means that the variability in the number of months in a year with P > 100 mm is small.

Histogram

If the range of outcomes on the variable is becoming large, then the line diagram is not an appropriate tool anymore to present the distribution of the variable. Grouping of data into classes and displaying the number of occurrences in each class to form a histogram will then provide better insight, see Figure 2.2. By doing so information is lost on the exact values of the variable, but the distribution is made visible. The variability of the data is shown by the horizontal spread of the blocks, and the most common values are found in blocks with the largest areas. Other features such as the symmetry of the data or lack of it are also shown. At least some 25 observations are required to make a histogram.

An important aspect of making a histogram is the selection of the number of classes n_c and of the class limits. The following steps are involved in preparing a histogram:

- The number of classes is determined by one of the following options (see e.g. Kottegoda and Rosso (1997):

$$n_c = \sqrt{N} \tag{2.1}$$

$$n_c = \frac{R\sqrt[3]{n}}{2R_{iq}} \tag{2.2}$$

where: n_c = number of classes
n = number of observations
R = range of observations: $X_{max} – X_{min}$
R_{iq} = interquartile range, defined by: $R_{iq} = M_{up} – M_{low}$
M_{up} = median of highest 50% of the data, i.e. 75% of the data is less
M_{low} = median of lowest 50% of the data, i.e. 25% of the data is less

- To obtain rounded numbers for the class limits convenient lower and upper limits below X_{min} and above X_{max} respectively the lowest and highest value have to be selected.
- Count the occurrences within each class: class frequency
- Present the results in a histogram

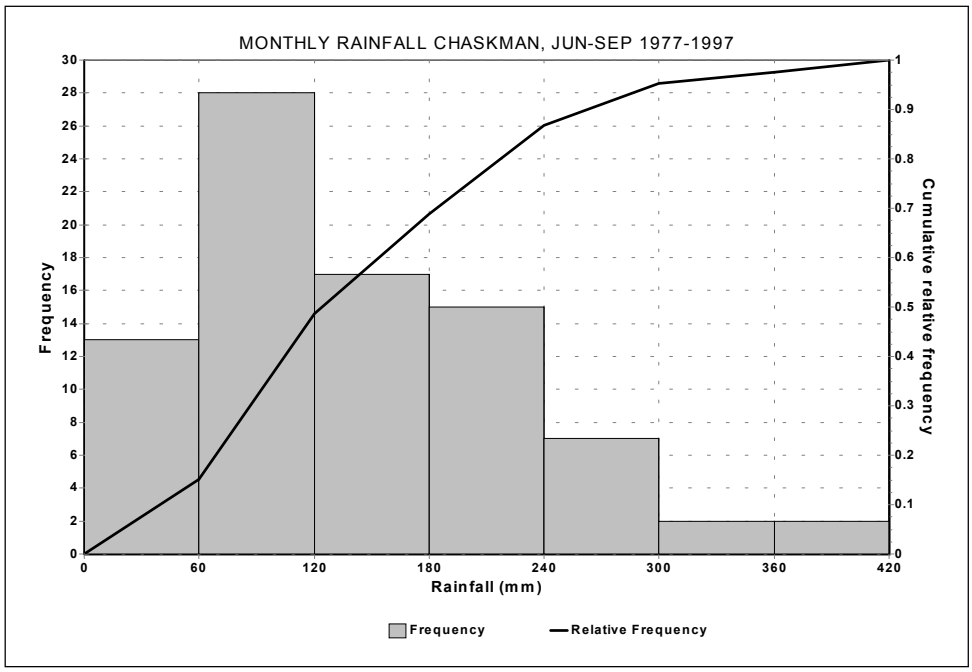

Figure 2.2: Histogram and cumulative relative frequency diagram of monthly rainfall at Chaskman, months June-September, period 1977 – 1997.

The application is shown for monthly rainfall of Chaskman. From Figure 1.2 it is observed, that rainfall in the months June to September behave more or less like a homogeneous group of data. A histogram is made of these monthly values for the years 1977-1997, i.e. 21 years of data. Hence in total the data set comprises 21 x 4 = 84 data points. The data are ranked in ascending order and displayed in Table 2.1

	1	2	3	4	5	6	7	8	9
1	12.1	55.4	*71.8*	92.8	118.1	152.2	196.3	229.0	326.2
2	19.6	55.8	72.2	97.8	124.8	154.4	201.2	234.6	342.6
3	20.8	55.8	74.8	100.2	127.2	158.0	*202.8*	237.2	404.6
4	26.6	61.2	75.4	101.4	128.0	160.2	206.4	258.0	418.7
5	35.4	61.8	75.8	101.4	130.2	161.0	207.0	258.8	
6	37.2	62.8	76.6	103.0	132.8	166.8	221.2	268.2	
7	48.8	64.6	77.4	103.8	136.0	169.2	221.4	268.4	
8	52.4	65.0	77.6	105.2	136.6	172.8	225.7	281.4	
9	52.8	65.6	78.9	105.7	144.0	188.0	227.6	281.8	
10	53.4	69.8	87.2	112.4	148.0	193.4	228.4	282.3	

Table 2.1: June-September monthly rainfall at Chaskman 1977-1997 ranked in ascending order

The values for X_{min} and X_{max} are respectively 12.1 mm and 418.7 mm, hence for the range it follows R = 418.7 – 12.1 = 406.6 mm. Since 84 data points are available 42 data are available in the lowest as well as in the highest group, so the values at positions 21 and 63 in the sorted data vector will give the medians for the lowest and highest 50% of the data M_{low} and M_{up}. These values are respectively 71.8 mm and 202.8 mm, hence the interquartile range is R_{iq} = 202.8 – 71.8 = 131.0 mm. According to (2.1) the number of classes in the histogram should be

$$n_c = \frac{R\sqrt[3]{n}}{2R_{iq}} = \frac{406.6x(84)^{1/3}}{2x131.0} = 6.8 \approx 7$$

Now, with 7 classes, R = 406.6 mm a class interval should be $\geq$ R/7 $\approx$ 58 mm, which is rounded to 60 mm. Using this class-interval and since X_{min} = 12.1 mm and X_{max} = 418.7 mm appropriate overall lower and upper class limits would be 0 mm and 420 mm. The result is displayed in Figure 2.2. The data points in a class are > the lower class limit and $\leq$ the upper class limit, with the exception of the lowest class, where the lowest value may be = lower class limit.

Note that if one uses (2.1) the result would have been $\sqrt{84} \approx 9$ classes, which is a slightly higher value. It follows that the guidelines given in (2.1) and (2.2) are indicative rather than compulsory. In general, at least 5 and at maximum 25 classes are advocated. Equation (2.2) has preference over equation (2.1) as it adapts its number of classes dependent on the peakedness of the distribution. If the histogram is strongly peaked then the inter-quantile range will be small. Consequently, the number of classes will increase, giving a better picture of the peaked zone.

Cumulative Relative Frequency Diagram

By dividing the frequency in each class of the histogram by the total number of data, the relative frequency diagram is obtained. By accumulating the relative frequencies, starting off from the lower limit of the lowest class up to the upper limit of the highest class the cumulative relative frequency diagram is obtained. For the data considered in the above example, the cumulative relative frequency diagram is shown with the histogram in Figure 2.2. The computational procedure is shown in Table 2.2.

Class	LCL	UCL	Freq.	Rel. Freq.	Cum.R. Fr.
1	0	60	13	0.155	0.155
2	60	120	28	0.333	0.488
3	120	180	17	0.202	0.690
4	180	240	15	0.179	0.869
5	240	300	7	0.083	0.952
6	300	360	2	0.024	0.976
7	360	420	2	0.024	**1.000**

Table 2.2:
Computation of cumulative relative frequencies

On the vertical axis of the graph, this line gives the cumulative relative frequencies of values shown on the horizontal axis. Instead of deriving this plot via the histogram, generally it is made by utilising and displaying every item of data distinctly. For this purpose, one ranks the series of size N in ascending order. The cumulative frequency given to the observation at rank *m* then becomes *m*/N, i.e. there are m data points less than or equal to the data point at rank *m.* This is shown in Figure 2.3.

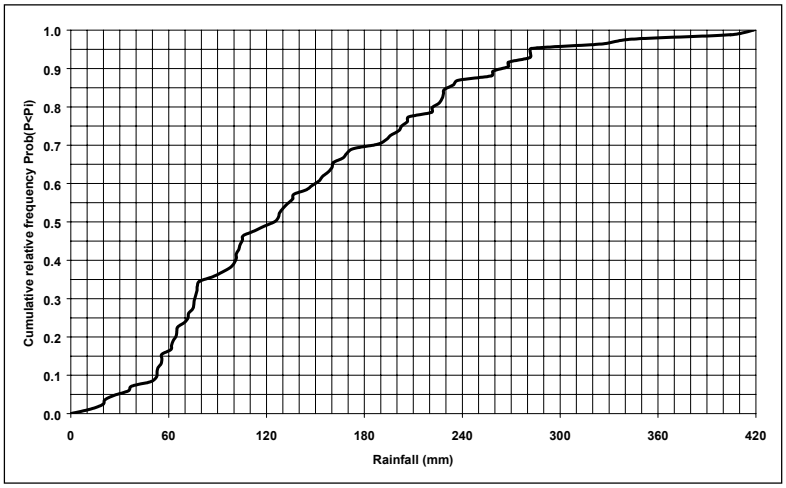

Figure 2.3:
Cumulative relative frequency distribution for Chaskman June–September data in the period 1977-1997

In Figure 2.3 the highest ranked data point (m = N) gets a cumulative relative frequency of m/N = N/N = 1. To describe the distribution of the data in that particular sample series this statement is correct. No observation exceeded the maximum value. However, in statistics one wants to say something about the distribution of data in the population of which the N observations are just one of many possible samples series. The cumulative relative frequency (crf) is then replaced by the non-exceedance probability. A non-exceedance probability of 1 for the maximum observed in the sample series would then imply that all possible outcomes would be less than or equal to that maximum. Unless there is a physical limit to the data such a statement is not justified. The non-exceedance probability of the maximum in the sample series will be less than 1. The non-exceedance probability to be given to the data point with rank m can be determined by viewing the series of ranked observations as order statistics: X(1), X(2), X(3), …,,X(m),…X(N). The expected value of order statistic X(m) depends first of all on the rank of X(m) relative to X(N). Furthermore is the expected value of X(m) a function of the probability distribution of the process from which the sample points are drawn. This will be discussed in more detail in Section 4.

Frequency Curves

Considering again the monthly rainfall series of Chaskman, for each month one can make a cumulative frequency distribution. Distinct crf's are the identified, say e.g. 10%, 50% and 90%, for each month. By displaying the rainfall having say a crf = 10% for all months in the year in a graph a frequency curve is obtained. Similarly for other crf's such a curve can be made. This is shown in Figure 2.4.

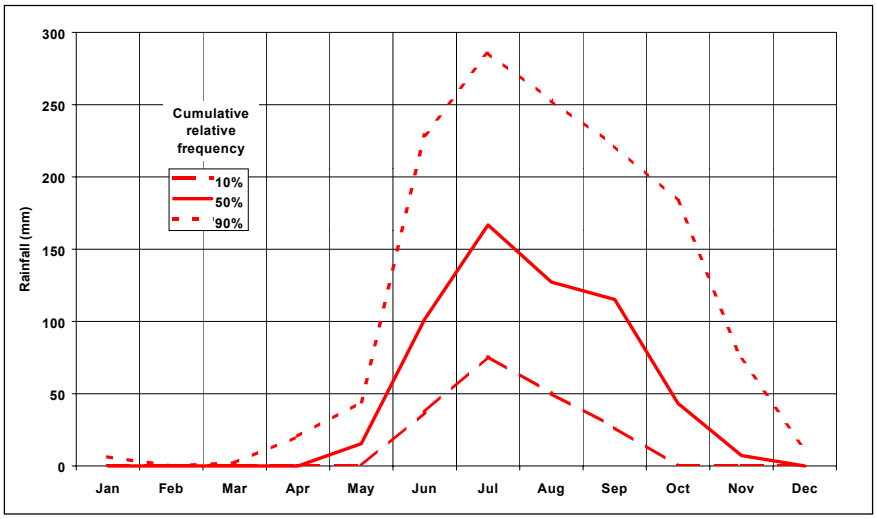

Figure 2.4: Frequency curves of monthly rainfall at station Chaskman, period 1968-1997

The computational procedure to arrive at the frequency curves is presented in the Tables 2.3 and 2.4. In Table 2.3 the actual monthly rainfall for a 30-year period is displayed. Next, the data for each month are put in ascending order, see Table 2.4, with the accompanying crf presented in the first column. The rows with crf = 0.1 (10%), 0.5 (50%) and 0.9 (90%) are highlighted and displayed in Figure 2.4.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
1968	0	0	0	0	0	49.8	144.3	60.5	162.1	43.4	47	0	507.1
1969	0	0	0	0	0	10.5	320.2	267.1	81.2	0	50.2	0	729.2
1970	0	0	0	0	60.8	80	124.9	140.5	30	162.6	0	0	598.8
1971	0	0	0	0	44.4	159.2	85.4	197.8	212.6	12	0	0	711.4
1972	0	0	0	3.2	31.4	46	229.7	38.3	0	0	0	0	348.6
1973	0	0	8.6	0	25	85	312.6	134.6	109.2	101.6	0	0	776.6
1974	0	0	0	0	132.2	72	150.8	175.2	206.2	183.4	0	0	919.8
1975	0	0	0	0	8	123.2	146.2	139.4	191.8	111.6	0	0	720.2
1976	0	0	0	0	0	494.8	323.8	208.6	115.2	3	139.2	0	1284.6
1977	0	0	0	0	16.4	188	207	61.8	64.6	44.2	119.4	3.6	705
1978	0	12.6	10.4	53	43.4	154.4	77.4	127.2	124.8	32.8	73.6	0.2	709.8
1979	0	0	0	0	21.4	75.4	93.8	252.8	221.4	13.4	57	0	735.2
1980	0	0	1	14.2	1.4	325.8	169.2	192.4	136.6	17.8	57.6	11.8	927.8
1981	10.8	2.2	0	20.6	9.2	152.2	258	101.4	160.2	53	7.4	2.2	777.2
1982	6.4	0	0	0	21.2	101.4	71.8	144	132.8	47.8	39.2	0	564.6
1983	0	0	0	0	4.2	55.8	75.8	418.4	268.2	2.8	0	0	825.2
1984	0	1.8	0	5.2	0	78.9	225.45	55.4	104.2	0	7	0	477.95
1985	0	0	0	0	31.6	62.8	105.7	74.8	26.6	91.3	0	0	392.8
1986	0	0	0	0	26.8	229	87.2	97.8	105.2	5.1	3.6	46.6	601.3
1987	0	0	0	0	80.2	118.1	65	148	12.1	89.4	8	11.9	532.7
1988	0	0	0.4	22.2	0	72.2	268.4	53.4	282.3	18.1	0	0	717
1989	0	0	6.4	1.4	9.2	37.2	227.6	61.2	190.7	7.6	0	0	541.3
1990	13.8	0	0	0	33.2	66.6	212	161.4	32.4	195.8	18	3.2	736.4
1991	0	0	0	16.2	12.8	404.6	235.4	50.2	48.6	21	9.4	0.6	798.8
1992	0	0	0	0	10.6	112.4	102	235.2	202.8	13.8	20	0	696.8
1993	0	0	1	3.8	15.8	130.2	226.4	66.6	53.4	304	7.2	31.6	840
1994	3.8	0	0	11.4	26	169	285.8	92.2	85	130.8	40.6	0	844.6
1995	35.2	0	2.2	17	15.4	20.8	157.8	19.8	262	87.8	2.2	0	620.2
1996	0.6	0	0	29.4	10.2	206.4	221.2	55.8	128	217.4	2.4	0	871.4
1997	5.4	0	0	12.4	10	136	166.8	342.6	77.6	66.3	148.7	41	1006.8

Table 2.3: Monthly and annual rainfall at station Chaskman, period 1968-1997

crf	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
0.033	0	0	0	0	0	10.5	65	19.8	0	0	0	0	348.6
0.067	0	0	0	0	0	20.8	71.8	38.3	12.1	0	0	0	392.8
0.100	**0**	**0**	**0**	**0**	**0**	**37.2**	**75.8**	**50.2**	**26.6**	**0**	**0**	**0**	**478.0**
0.133	0	0	0	0	0	46	77.4	53.4	30	2.8	0	0	507.1
0.167	0	0	0	0	0	49.8	85.4	55.4	32.4	3	0	0	532.7
0.200	0	0	0	0	1.4	55.8	87.2	55.8	48.6	5.1	0	0	541.3
0.233	0	0	0	0	4.2	62.8	93.8	60.5	53.4	7.6	0	0	564.6
0.267	0	0	0	0	8	66.6	102	61.2	64.6	12	0	0	598.8
0.300	0	0	0	0	9.2	72	105.7	61.8	77.6	13.4	0	0	601.3
0.333	0	0	0	0	9.2	72.2	124.9	66.6	81.2	13.8	0	0	620.2
0.367	0	0	0	0	10	75.4	144.3	74.8	85	17.8	2.2	0	696.8
0.400	0	0	0	0	10.2	78.9	146.2	92.2	104.2	18.1	2.4	0	705.0
0.433	0	0	0	0	10.6	80	150.8	97.8	105.2	21	3.6	0	709.8
0.467	0	0	0	0	12.8	85	157.8	101.4	109.2	32.8	7	0	711.4
0.500	**0**	**0**	**0**	**0**	**15.4**	**101.4**	**166.8**	**127.2**	**115.2**	**43.4**	**7.2**	**0**	**717.0**
0.533	0	0	0	0	15.8	112.4	169.2	134.6	124.8	44.2	7.4	0	720.2
0.567	0	0	0	0	16.4	118.1	207	139.4	128	47.8	8	0	729.2
0.600	0	0	0	1.4	21.2	123.2	212	140.5	132.8	53	9.4	0	735.2
0.633	0	0	0	3.2	21.4	130.2	221.2	144	136.6	66.3	18	0	736.4
0.667	0	0	0	3.8	25	136	225.45	148	160.2	87.8	20	0	776.6
0.700	0	0	0	5.2	26	152.2	226.4	161.4	162.1	89.4	39.2	0.2	777.2
0.733	0	0	0	11.4	26.8	154.4	227.6	175.2	190.7	91.3	40.6	0.6	798.8
0.767	0	0	0	12.4	31.4	159.2	229.7	192.4	191.8	101.6	47	2.2	825.2
0.800	0.6	0	0.4	14.2	31.6	169	235.4	197.8	202.8	111.6	50.2	3.2	840.0
0.833	3.8	0	1	16.2	33.2	188	258	208.6	206.2	130.8	57	3.6	844.6
0.867	5.4	0	1	17	43.4	206.4	268.4	235.2	212.6	162.6	57.6	11.8	871.4
0.900	**6.4**	**0**	**2.2**	**20.6**	**44.4**	**229**	**285.8**	**252.8**	**221.4**	**183.4**	**73.6**	**11.9**	**919.8**
0.933	10.8	1.8	6.4	22.2	60.8	325.8	312.6	267.1	262	195.8	119.4	31.6	927.8
0.967	13.8	2.2	8.6	29.4	80.2	404.6	320.2	342.6	268.2	217.4	139.2	41	1006.8
1.000	35.2	12.6	10.4	53	132.2	494.8	323.8	418.4	282.3	304	148.7	46.6	1284.6

Table 2.4: Monthly and annual rainfall at station Chaskman, period 1968-1997 ordered in ascending order per column

By plotting the rainfall of a particular year with the frequency curves one has a proper means to assess how the rainfall in each month in that particular year behaved compared to the long term rainfall in that month. However, the say 10% curve should not be considered as a 10%-wet year. To show this in the last column of Table 2.4 the ranked annual values are presented as well. The rainfall in 10%-wet year amounts 478 mm, whereas the sum of the 10% monthly rainfall amounts add up to 189.8 mm only. Similar conclusions can be drawn for other crf's. This is shown in Figure 2.5 a, b, c.

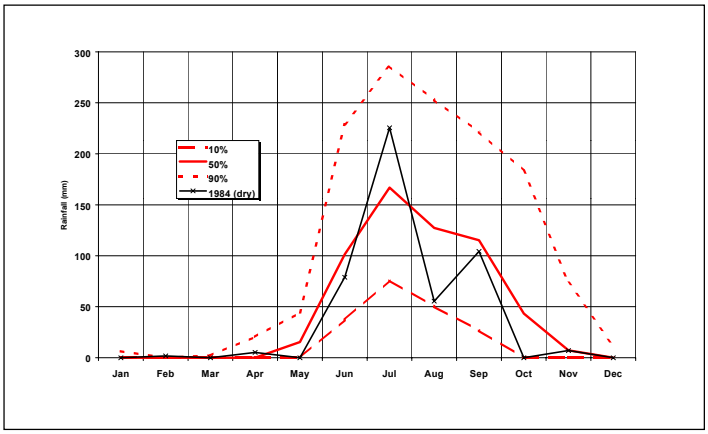

Figure 2.5a

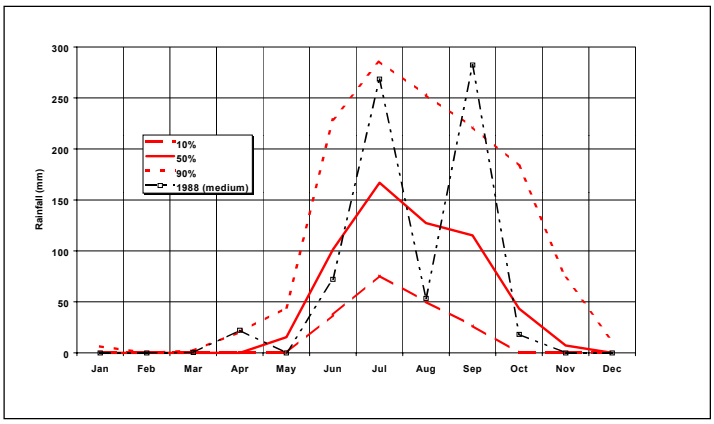

Figure 2.5b

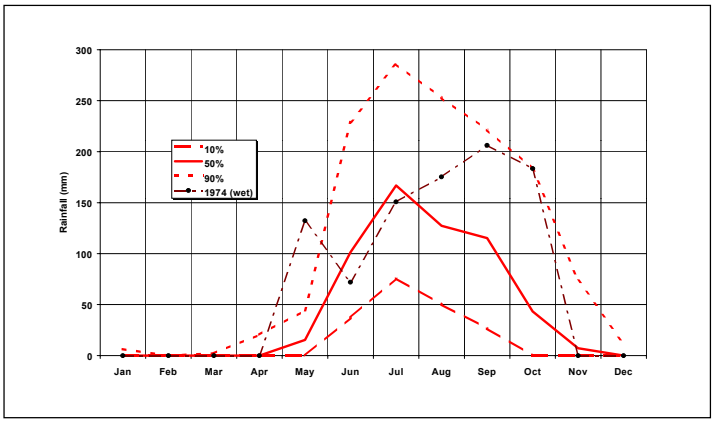

Figure 2.5c

Figure 2.5 a, b, c: Frequency curves of crf = 10, 50 and 90% with 10%, 50% and 90% wet year records.

In the above text frequency curves were discussed for monthly rainfall data. Basically, the technique can be applied to any hydrological variable and the interval may also be day, 10 days, season, etc. Generally, say, we have M observations in a year for N years. Let the observation on the hydrological variable in interval *m* in year *n* be denoted by $X_{m,n}$. Then for *n* = 1, N the X_m's are put in ascending order: $X_{m,k}$, where *k* is the rank of $X_{m,n}$, with *k* running from 1 to N. De crf attributed to $X_{m,n}$ is *k*/N (or *k*/(N+1) or some other estimate for the probability of non-exceedance as discussed earlier). By selecting a specific value for *k* = *k1* corresponding to a required crf the sequence of $X_{m,k1}$ for *m* = 1,M will give us the required frequency curve. In case a required crf, for which a frequency curve is to be made, does not correspond with the k^{th} rank in the sequence of N values, linear interpolation between surrounding values is to be applied.

Duration Curves

For the assessment of water resources, navigational depths, etc. it may be useful to draw duration curves. When dealing with flows in rivers, this type of graphs is known as a flow duration curve. It is in effect a cumulative frequency diagram with specific time scales. On the horizontal axis the percentage of time or the number of days/months per year or season during which the flow is not exceeded may be given. The volume of flow per day/month or flow intensity is given on the vertical axis. (The above convention is the display adopted in HYMOS; others interchange the horizontal and vertical axis.) Similarly, duration curves may be developed for any other type of variable. In Figure 2.6 the duration curve for the monthly rainfall at Chaskman for the period 1968-1997 is presented.

Figure 2.6 tells us that there is no rain during at least four months in a year, and **on average** there is only one month in a year with a monthly total larger than 200 mm. However, from Table 2.3 it can be observed that during 8 years out of 30 the 200 mm threshold was exceeded during two months. So the curve only displays average characteristics. The curve is obtained by multiplying the cumulative relative frequency associated with an observation with the number of intervals one has considered in a year or a season.

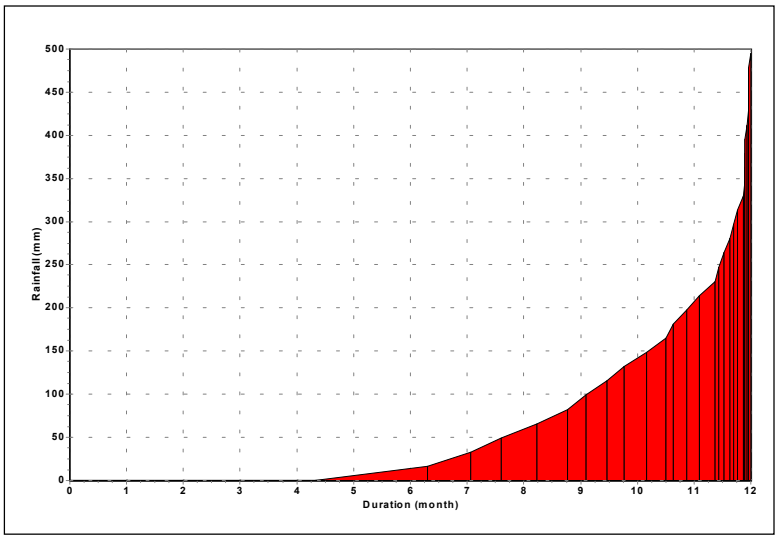

Figure 2.6:
Duration curve of monthly rainfall for station Chaskman

2.3 Measures of Central Tendency

Measures of the central tendency of a series of observations are:

- Mean
- Median
- Mode

Mean

The mean of a sample of size N is defined by

$$m = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (2.3)$$

where x_i = individual observed value in the sample
N = sample size i.e. total number of observed values
m = mean of the sample size n.

When dealing with catchment rainfall determined by Thiessen method, the mean is weighted according to the areas enclosed by bisectors around each station. The sum of the weights is 1:

$$m_w = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \qquad (2.4)$$

Median

The median M of a sample is the middle value of the ranked sample, if N is odd. If N is even it is the average of the two middle values. The cumulative relative frequency of the median is 0.5. For a symmetrical distribution the mean and the median are similar. If the distribution is skewed to the right, then M < m, and when skewed to the left M > m.

Mode

The mode of a sample is the most frequently occurring value and hence corresponds with the value for which the distribution function is maximum. In Figure 2.2 the mode is in the class 60-120 mm and can be estimated as 90 mm.

2.4 Measures of Dispersion

Common measures of dispersion are:

- the range,
- the variance,
- the standard deviation, and
- coefficient of variation.

Range

The range of a sample is the difference between the largest and smallest sample value. Since the sample range is a function of only two of the N sample values it contains no information about the distribution of the data between the minimum and maximum value.

The population range of a hydrological variable is in many cases, the interval from 0 to ∞, and as such displays no information about the process.

In hydrology the word 'range' is also used to quantify the range of accumulative departures from the mean (also indicated as partial sums). That value has important implications when dealing with water storage. It is a measure for the required storage when the average flow is to be drawn from a reservoir.

Variance

The most common measure of dispersion used in statistical analysis is the variance, estimated by s^2:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2 \qquad (2.5)$$

The reason for using the divisor N-1 instead of N is that it will result in an unbiased estimate for the variance. The units of the variance are the same as the units of x^2.

Standard deviation

The standard deviation s is the root of the variance and provides as such a measure for the dispersion of the data in the sample set in the same dimension as the sample data. It is estimated by:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2} \qquad (2.6)$$

Coefficient of Variation

A dimensionless measure of dispersion is the coefficient of variation C_v defined as the standard deviation divided by the mean:

$$C_v = \frac{s}{m} \qquad (2.7)$$

Note that when m = 0 the coefficient of variation C_v becomes undefined; hence for normalised distributions this measure cannot be applied.

From Figure 1.3 it is observed that the coefficient of variation of the monthly rainfall at Chaskman is > 1 for the dry period, but < 1 during the monsoon.

2.5 Measure of Symmetry: Skewness

Distributions of hydrological variables are often skewed, i.e. non-symmetrical. The distributions are generally skewed to the right, like daily rainfall. By aggregation of data, the distribution of the aggregate will approach normality, i.e. will become symmetrical. Positively and negatively skewed distributions and symmetrical distributions are shown in Figure 2.7.

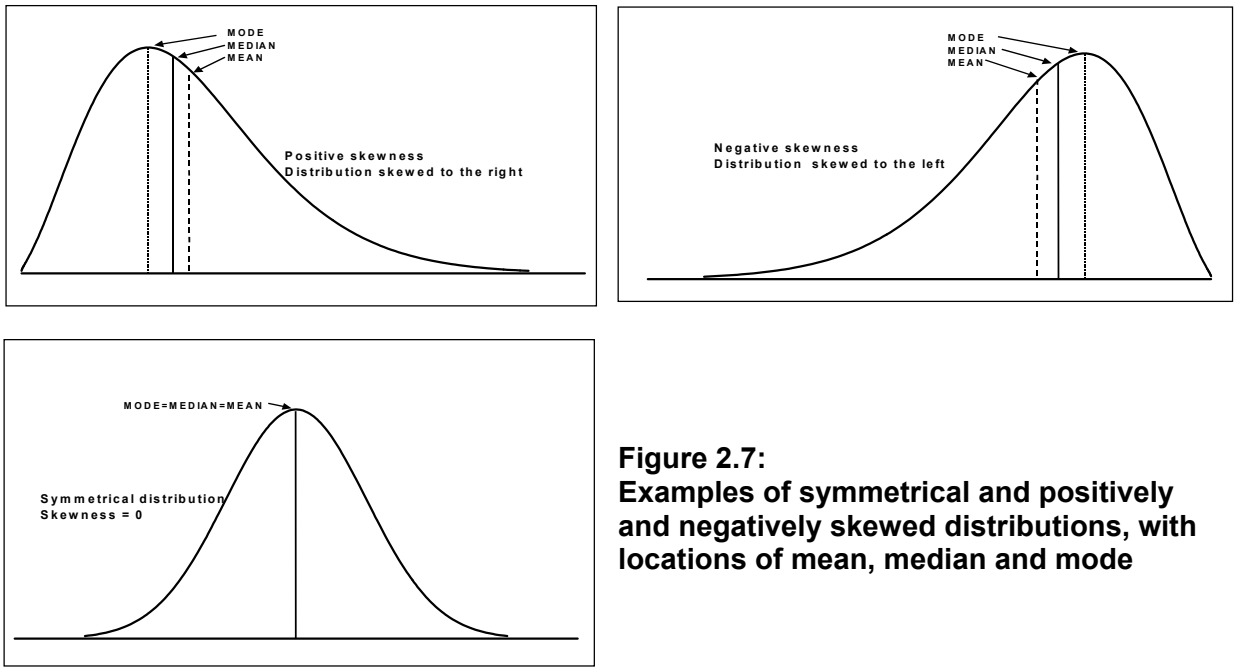

Figure 2.7:
Examples of symmetrical and positively and negatively skewed distributions, with locations of mean, median and mode

The skewness is derived from the third central moment of the distribution, scaled by the standard deviation to the power 3. An unbiased estimate for the coefficient of skewness can be obtained from the following expression:

$$g_1 = \frac{N}{(N-1)(N-2)} \frac{\sum_{i=1}^{N} (x_i - m)^3}{s^3} \qquad (2.8)$$

In Figure 2.7 the relative position of the mean, median and mode for symmetrical and positively and negatively skewed distributions is presented.

2.6 Measure of Peakedness: Kurtosis

Kurtosis refers to the extent of peakedness or flatness of a probability distribution in comparison with the normal distribution, see Figure 2.8. The sample estimate for kurtosis is:

$$g_2 = \frac{N^2 - 2N + 3}{(N-1)(N-2)(N-3)} \frac{\sum_{i=1}^{N} (x_i - m)^4}{s^4} \qquad (2.9)$$

The kurtosis is seen to be the 4^{th} moment of the distribution about the mean, scaled by the 4^{th} power of the standard deviation. The kurtosis for a normal distribution is 3. The normal distribution is said to be **mesokurtic**. If a distribution has a relatively greater concentration of probability near the mean than does the normal, the kurtosis will be greater than 3 and the distribution is said to be **leptokurtic**. If a distribution has a relatively smaller concentration of a probability near the mean than does the normal, the kurtosis will be less than 3 and the distribution is said to be **platykurtic**.

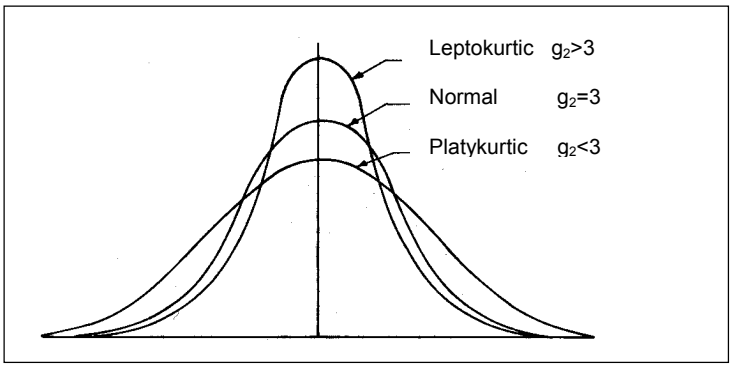

Figure 2.8:
Illustration of Kurtosis

The **coefficient of excess** e is defined as $g_2 - 3$. Therefore for a normal distribution e is 0, for a leptokurtic distribution e is positive and for a platykurtic distribution *e* is negative.

2.7 Quantiles, percentile, deciles and quartiles

The cumulative relative frequency axis of the cumulative relative frequency curve running from 0 to 1 or from 0 to 100% can be split into equal parts. Generally, if the division is in *n* equal parts, one will generate (n-1) **quantiles**. The pth quantile, x_p, is the value that is larger than 100p% of all data. When n = 100, i.e. the division is done in 100 equal parts (percents), then the value of the hydrological variable read from the x-axis corresponding with a crf of p% is called the p^{th} **percentile**. If the frequency axis is divided into 10 equal parts then the corresponding value on the x-axis is called a **decile**. Thus the 10^{th} percentile (also called the first decile) would mean that 10% of the observed values are less than or equal to the percentile. Conversely, the 90^{th} percentile (or 9^{th} decile) would mean that 90% of the observed values are lying below that or 10% of the observed values are lying above that. The median would be the 50^{th} percentile (or fifth decile). Similarly, if the frequency axis is divided in 4 equal parts then one speaks of **quartiles**. The first quartile corresponds with the 25^{th} percentile, i.e. 25% of the values are less or equal than the first quartile; the second quartile is equal to the median and the third quartile equals the 75^{th} percentile.

2.8 Box plot and box and whiskers plot

A **box plot** displays the three **quartiles** of a distribution in the form of a box, see Figure 2.9. If in addition also the **minimum and the maximum** values are displayed by bars extending the box on either side, the plot is called a **box and whiskers plot**. Sometimes also the mean is indicated in the plot. Hence the plot is a 5 or 6 points summary of the actual frequency distribution. Such plots are made for the data in a season or a year or any other selected time interval.

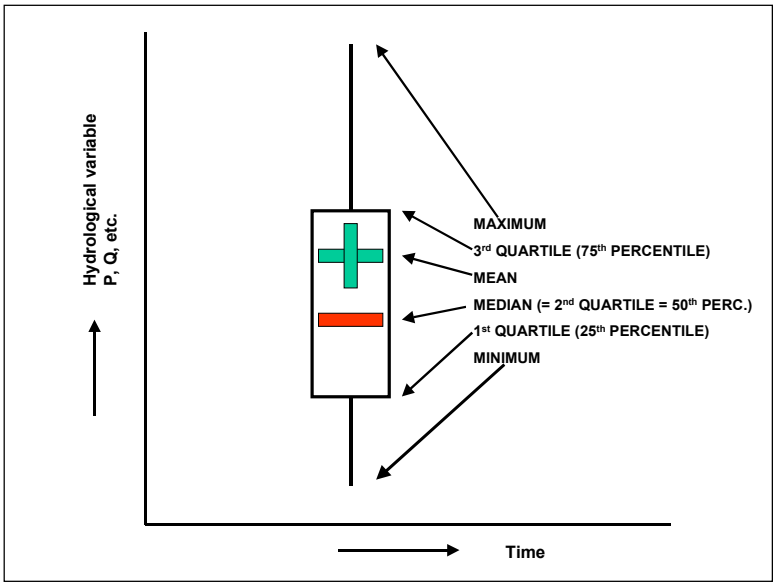

Figure 2.9:
Features of a box and whiskers plot

By displaying the box and bars for successive years a quick insight is provided into the variation of the process from year to year. This form is very popular for displaying the behaviour of water quality variables. For that purpose the plot is extended with threshold values on a particular water quality variable.

In Figure 2.10 an example is given of a box and whiskers plot applied to discharge measurements at station Rakshewa in Bhima basin, where the statistics of the measurements from 1995 to 1998 are shown for each year separately.

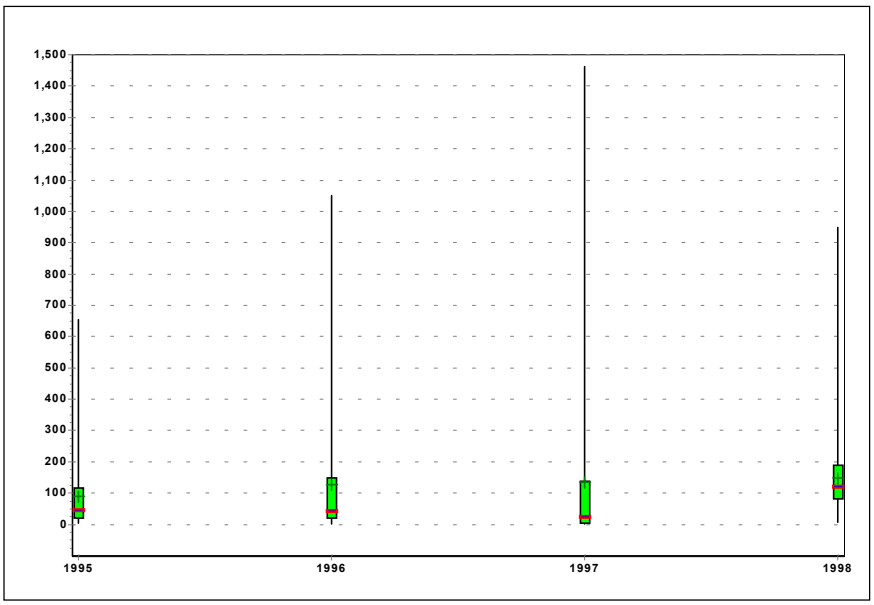

Figure 2.10:
Box and whiskers plot of discharge measurements at station Rakshewa in Bhima basin, period 1995 – 1998.

It is clearly observed from the boxes and bars in Figure 2.10 that the distribution of the measured discharges in a year is skewed to the right. Generally, a large number of discharge measurements are available for the very low stages and only a few for the higher stages. Hence the extent of the box (which comprises 50% of the measurements) is very small compared to the range of the data. The mean is seen to be always larger than the median.

2.9 Covariance and Correlation Coefficient

When simultaneous observations on hydrological variables are available then one may be interested in the linear association between the variables. This is expressed by the covariance and correlation coefficient.

If there are N pairs of observations (x_1, y_1), (x_2, y_2), …, (x_N, y_N), of two variables X and Y, the sample covariance is obtained from the following expression:

$$\hat{C}_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)(y_i - m_Y) \tag{2.10}$$

where: m_X, m_Y = sample means of X and Y respectively:

The correlation coefficient r_{XY} is obtained by scaling the covariance by the standard deviations of X and Y:

$$r_{XY} = \frac{\hat{C}_{XY}}{s_X s_Y} = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)(y_i - m_Y)}{s_X s_Y} \tag{2.11}$$

where: s_X , s_Y = sample standard deviations of X and Y.

To get the limits of r_{XY} consider the case that X and Y have a **perfect** linear correlation. Then the relationship between X and Y is given by :

$Y = a + bX$

and hence:

$m_Y = a + bm_X$ and: $s_Y^2 = b^2 s_X^2$ or: $s_Y = |b| s_X$

Substituting above relations in (2.11) gives:

$$r_{XY} = \frac{\frac{1}{N-1}\sum(x_i - m_X)(a + bx_i - (a + bm_X))}{s_X |b| s_X} = \frac{b}{|b|}\frac{\frac{1}{N-1}\sum(x_i - m_X)^2}{s_X^2} = \frac{b}{|b|} \tag{2.12}$$

If Y increases for increasing X, i.e. they are positively correlated, then $b > 0$ and r_{XY} is seen to be 1. If on the other hand Y decreases when X is increasing, they are negatively correlated; then $b < 0$ and r_{XY} is –1. So r_{XY} is seen to vary between ± 1:

$-1 \le r_{XY} \le 1$.

If there is no linear association between X and Y then r_{XY} is 0. If r_{XY} is 0 it does not mean that X and Y are independent or that there is no association between X and Y. It only means that the linear association is not existing. Still, there may be for example a circular association.

A convenient means to investigate the existence of linear association is by making a XY-scatter plot of the samples. Typical examples of scatter plots are shown in Figure 2.11.

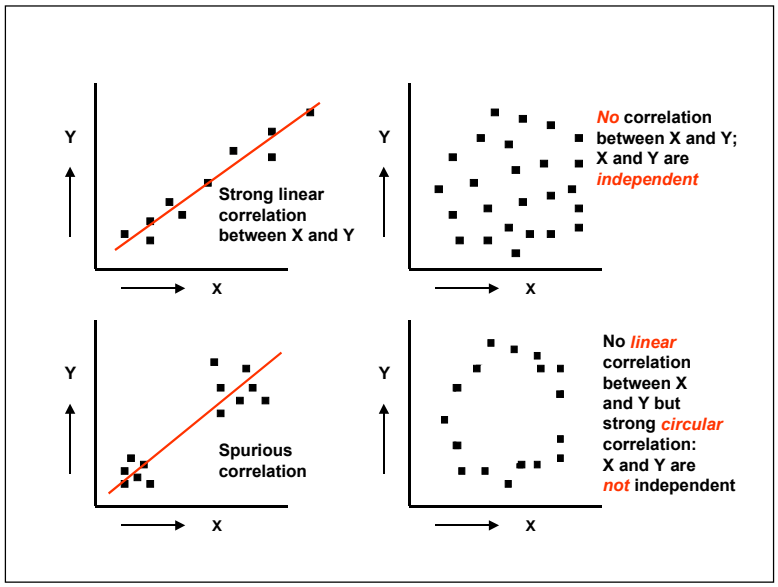

Figure 2.11:
Examples of scatter plots

In some cases the scatter plot may indicate a non-linear type of relationship between the two variables. In such cases some transformation, e.g. a logarithmic, square root, negative reciprocal, or other appropriate transformation to one or both variables may be applied before analysis.

Spurious correlation

The lower left plot in Figure 2.11 gives an example of **spurious** correlation, which is easily obtained in hydrology, when blindly data are being compared. For example if there is a distinct wet and dry period and the discharges of two sites in different regions, but both subjected to monsoonal variation, are plotted in an XY-plot, a situation like the one displayed will occur. In the wet period the data at X and Y may be completely uncorrelated, but simply by the fact of the existence of a dry and wet period, which clusters observations in the low and the high regions, the correlation is seemingly very high. This effect is due to the acceptance of **heterogeneous** data, see also Figure 1.2 and 1.3. By taking the low and high flow values in the same data set, the overall mean value for X and Y will be somewhere between the low and the high values. Hence entries in the wet period on either side will be positive relative to the mean and so will be their products. In the same way, entries in the dry period will both be negative relative to the mean, so their product will be positive as well, ending up into a large positive correlation.

Similarly, wrong conclusions can be drawn by comparing data having the same denominator. If X, Y and Z are **uncorrelated** and X/Z and Y/Z are subjected to correlation analysis, a non-zero correlation will be found (see e.g. Yevjevich, (1972)):

$$r = \frac{C_{v,Z}^2}{\left(C_{v,X}^2 + C_{v,Z}^2\right)^{1/2}\left(C_{v,Y}^2 + C_{v,Z}^2\right)^{1/2}} \qquad (2.13)$$

From (2.13) it is observed, that when all coefficients of variation are equal, it follows that r = 0.5!!!

It indicates that one has to select the sample sets to be subjected to correlation and regression analysis carefully. Common divisors should be avoided. Also, the direction of analysis as indicated in Figure 2.2 is of utmost importance to ensure homogeneous data sets.

3 Fundamental Concepts of Probability

3.1 3.1 Axioms and Theorems

Sample and Space Events

The **sample space** denoted by Ω, is defined here as the collection of all possible outcomes of sampling on a hydrological variable.

An **event** is a collection of sample points in the sample space Ω of an experiment. An event can consists of a single sample point called a simple or elementary event, or it can be made up of two or more sample points known as a compound event. An event is (denoted by a capital letter A (or any other letter)) is thus a subset of sample space Ω.

The Null Event, Intersection and Union

Two events A_1 and A_2 are **mutually exclusive** or **disjoint** if the occurrence of A_1 excludes A_2, i.e. none of the points contained in A_1 is contained in A_2. and vice versa.

The **intersection** of the events A_1 and A_2 is that part of the sample space they have in common. This is denoted by $A_1 \cap A_2$ or A_1A_2.

If A_1 and A_2 are mutually exclusive then their intersection constitutes a **null event**: $A_1 \cap A_2 = A_1A_2 = \varnothing$.

The **union** of two events A_1 and A_2 represents their joint occurrences, and it comprises the event containing the entire sample in A_1 and A_2. This is denoted by $A_1 \cup A_2$, or simply $A_1 + A_2$. With the latter notation one has to be careful as the sum of the two has to be corrected for the space in common (i.e. the intersection.

The intersection is equivalent to the "**and**" logical statement, whereas the union equivalent to "**and/or**".

The above definitions have been visualised in Figure 3.1 by means of **Venn diagrams**.

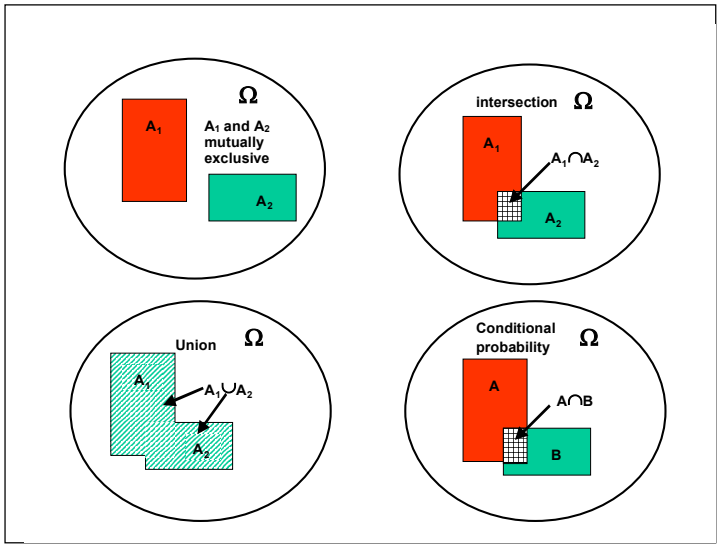

Figure 3.1: Definition sketch by Venn diagrams

The definitions are illustrated in the following example:

Example 3.1 Events in a sample space.

Sample space and events representing rainy days (i) and total rainfall (p) at a rainfall station during the period 1-10 July are given in Figure 3.2:

The sample space reads: $\Omega \equiv \{(i,p)$: i = 0, 1, 2,…,10; and $0 \leq p\}$

Event $A_1 \equiv \{(i,p)$: $i > 3$, and $p > 50\}$
Event $A_2 \equiv \{(i,p)$: $3 \leq i < 5$, and $p > 20\}$
Event $A_3 \equiv \{(i,p)$: $1 \leq i < 3$, and $2 \leq p < 30\}$

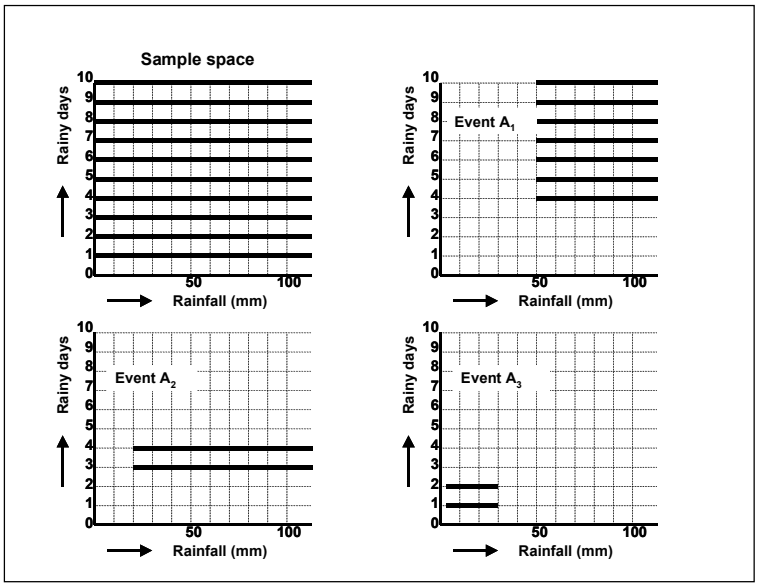

Figure 3.2:
Presentation of sample space Ω events A_1, A_2 and A_3

The union and intersection of A_1 and A_2 and of A_2 and A_3 are presented in Figure 3.3.
Event $A_1 + A_2 \equiv \{(i,p)$: $3 \leq i < 5$, and $p > 20$; $i \geq 5$, and $p > 50\}$
Event $A_1A_2 \equiv \{(i,p)$: i = 4 and $p > 50\}$
Event $A_2 + A_3 \equiv \{(i,p)$: $1 \leq i < 3$, and $2 \leq p < 30$; $3 \leq i < 5$, and $p > 20\}$
Event $A_2A_3 = \varnothing$, since A_2 and A_3 are disjoint, having no points in common.

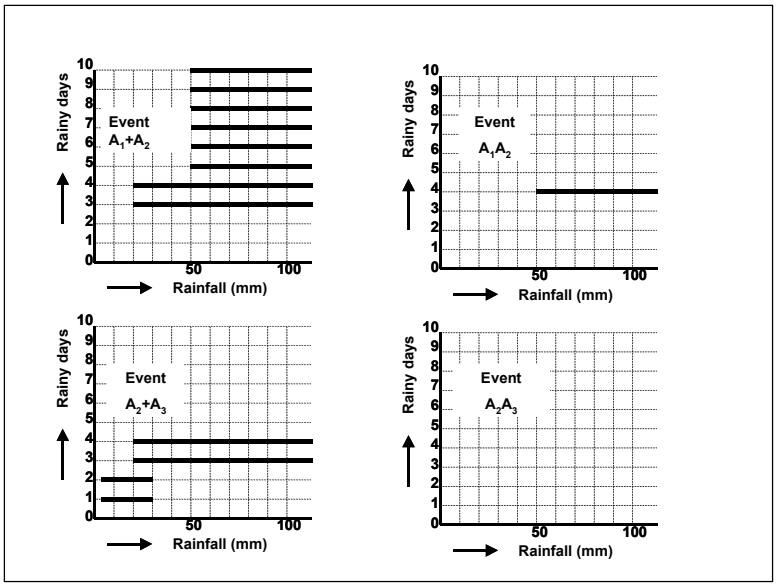

Figure 3.3:
Unions and intersections of A_I and A_2 and of A_2 and A_3

Probability axioms and theorems

Using these definitions the following axioms and theorems are discussed dealing with the probability of an event or several events in the sample space.

Definition of probability

If a random events occurs a large number of times N, of which N_A times the event A happens, then the probability of the occurrence of event A is:

$$P(A) = \lim_{N\to\infty} \frac{N_A}{N} \tag{3.1}$$

Hence, if A is any event in a sample space Ω, then:

$$0 \le P(A) \le 1 \tag{3.2}$$

The event in the sample space not contained in A is the complement of A, denoted by A^C:

$$P(A^C) = 1 - P(A) \tag{3.3}$$

If A is a certain event then:

$$P(A) = 1 \tag{3.4}$$

Probability of the union of events

For any set of arbitrary events A_1 and A_2 the probability of the **union** of the events, i.e. the probability of event A_1 **and/or** A_2 is:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \tag{3.5}$$

The last term is the intersection of A_1 and A_2, i.e. the part in the sample space they have in common. So, if A_1 and A_2 have no outcomes in common, i.e if they are **mutually exclusive**, then the intersection of the two events is a null event and then (3.5) reduces to:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) \tag{3.6}$$

For three joint events it generally follows:

$$P(A_1 + A_2 + A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1\,A_2) - P(A_1\,A_3) - P(A_2\,A_3) + P(A_1\,A_2\,A_3) \tag{3.7}$$

For any set of arbitrary events $A_1, A_2, \ldots, A_m$ the probability of the union becomes a complicated expression, (see e.g. Suhir, 1997), but if the events $A_1, A_2, \ldots, A_m$ have no outcomes or elements in common, i.e. if they are mutually exclusive, then the union of the events have the probability:

$$P\left(\sum_{j=1}^{m} A_j\right) = \sum_{j=1}^{m} P(A_j) \tag{3.8}$$

Hence, the probability of the intersection is seen to have vanished as it constitutes a null event for mutually exclusive events.

Conditional probability

The **conditional** probability P(B|A) gives the probability of event B given that A has occurred. Here A serves as a new (reduced) sample space (see Figure 3.1) and P(B|A) is that fraction of P(A) which corresponds to A∩B, hence:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{3.9}$$

Denoting $P(A \cap B) \equiv P(AB)$ it follows:

$$P(AB) = P(B|A) \,.\, P(A) \tag{3.10}$$

Independence

If A and B are **independent** events, i.e. the occurrence of B is not affected by the occurrence of A, then:

$$P(B|A) = P(B) \tag{3.11}$$

and hence:

$$P(AB) = P(B) \,.\, P(A) \tag{3.12}$$

It states that if the events A and B are independent, the probability of the occurrence of event A **and** B equals the product of the **marginal** probabilities of the individual events.

Total probability

Consider an event B in Ω with $P(B) \neq 0$ and the **mutually exclusive** events $A_1, A_2, \ldots, A_m$, which are **collectively exhaustive**, i.e. $A_1 + A_2 + \ldots + A_m = \Omega$. Then the events $BA_1, BA_2, \ldots, BA_m$ are also mutually exclusive and $BA_1 + BA_2 + \ldots + BA_m = B(A_1 + A_2 + \ldots + A_m) = B\Omega = B$. Hence:

$$P(B) = \sum_{j=1}^{m} P(BA_j) = \sum_{j=1}^{m} P(B \mid A_j).P(A_j) \tag{3.13}$$

This is called the theorem of **total probability**, which is visualised in Figure 3.4.

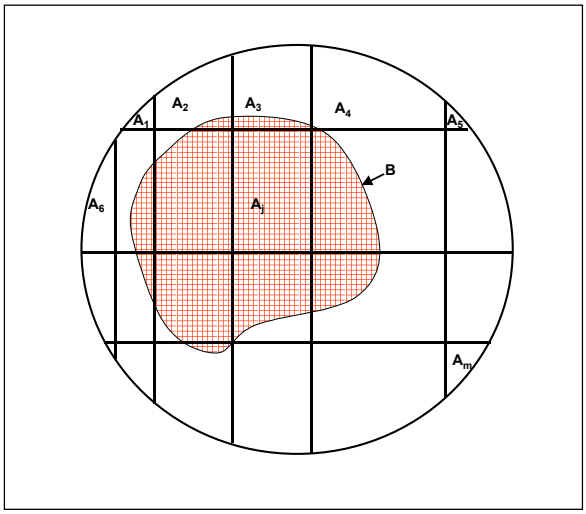

Figure 3.4:
Concept of total probability

Bayes theorem

Observe now the following conditional probability:

$$P(A_i \mid B) = \frac{P(BA_i)}{P(B)}$$

The numerator reads according to (3.10) $P(BA_i) = P(B|A_i) \; . \; P(A_i)$. The denominator is given by (3.13). It then follows for $P(A_i|B)$, Bayes rule:

$$P(A_i \mid B) = \frac{P(B \mid A_i).P(A_i)}{\sum_{j=1}^{m} P(B \mid A_j).P(A_j)} \qquad (3.14)$$

Bayes rule provides a method to update the probabilities about the true state of a system (A), by sampling (B) in stages. The probabilities $P(A_i)$'s on the right hand side of (3.14) are the probabilities about the state of the system before the sample is taken (**prior** probabilities). After each sampling the **prior** probabilities $P(A_i)$'s are updated, by replacing them with the **posterior** probability (= left hand side of the equation), found through the outcome of the sampling: B. The conditional probabilities $P(B|A_j)$ represent basically the quality of the sampling method or equipment: the probability of getting a particular sample B given that the true state of the system is A_i. Bayes rule can therefore be interpreted as follows:

$$P(state_{posterior} \mid sample) = \frac{P(sample \mid state).P(state_{prior})}{\sum_{all\,states} P(sample \mid state).P(state_{prior})} \qquad (3.15)$$

To illustrate the above axioms and theorems the following examples are given.

Example 3.2 Annual monthly maximum rainfall

The annual monthly maximum rainfall for station Chaskman is presented in Table 3.2 and Figure 3.5.

Year	P_{max} (mm)	Year	P_{max} (mm)	Year	P_{max} (mm)
1968	162.1	1978	154.4	1988	282.3
1969	320.2	1979	252.8	1989	227.6
1970	162.6	1980	325.8	1990	212.0
1971	212.6	1981	258.0	1991	404.6
1972	229.7	1982	144.0	1992	235.2
1973	312.6	1983	418.4	1993	304.0
1974	206.2	1984	225.5	1994	285.8
1975	191.8	1985	105.7	1995	262.0
1976	494.8	1986	229.0	1996	221.2
1977	207.0	1987	148.0	1997	342.6

Table 3.1:
Annual monthly maximum rainfall for Chaksman, period 1968-1997

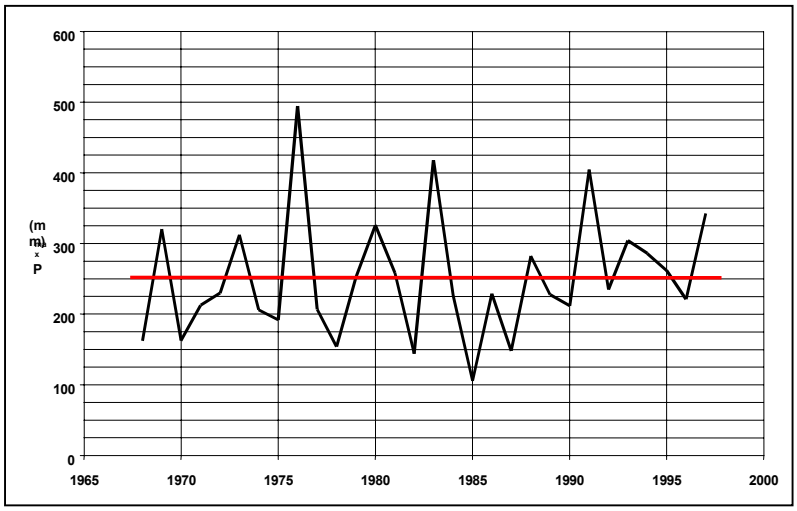

Figure 3.5:
Annual monthly maximum rainfall for Chaksman, period 1968-1997

From the table and figure it is observed that a monthly maximum > 260 mm has occurred 11 times in a period of 30 years, hence P_{max} > 260 mm = 11/30 = 0.367 in any one year. Assuming that the elements of the annual monthly maximum series are independent, it follows that the probability of having two annual maximum values in sequence > 260 mm = 0.367 x 0.367 = 0.135. From the series one observes that this event happened only 2 times in 30 years, that is 2 out of 29, i.e. having a probability of 2/29 = 0.069. If event A is the occurrence that P_{max} > 260 mm and B is the event that P_{max} > 260 mm in a second successive year then: $P(B|A) = P(A \cap B)/P(A)=(2/29)/(11/30) = 0.19$.

Example 3.3 Daily rainfall Balasinor (Gujarat)

Based on daily rainfall data of station Balasinor for the month of July in the period 1961 to 1970, the following probabilities have been determined:

Probability of a rainy day following a rainy day = 0.34
Probability of a rainy day following a dry day = 0.17
Probability of a dry day following a rainy day = 0.16
Probability of a dry day following a dry day = 0.33

Given that a particular day is dry, what is the probability the next two days are (1) dry and (2) wet?

(1) Call event A = dry day 1 after a dry day and event B = dry day 2 after a dry day. Hence required is $P(A \cap B) = P(B|A) \cdot P(A)$. The probability of having a dry day after a dry day is P(A) = 0.33 and the probability of a dry day given that the previous day was dry P(B|A) = 0.33. So, $P(A \cap B) = P(B|A) \cdot P(A) = 0.33 \cdot 0.33 = 0.11$.

(2) Call event A = wet day 1 after a dry day and event B = wet day 2 after a dry day. Now we require again $P(A \cap B) = P(B|A) \,.\, P(A)$. The probability of a wet day after a dry day is P(A) = 0.17 and the probability of a wet day given that the previous day was also wet = P(B|A) = 0.34. Hence, $P(A \cap B) = P(B|A).\ P(A) = 0.34 \,.\, 0.17 = 0.06$. This probability is seen to be about half the probability of having two dry days in a row after a dry day. This is due to the fact that for Balasinor the probability of having a wet day followed by a dry day or vice versa is about half the probability of having two wet or two dry days sequentially.

Example 3.4 Prior and posterior probabilities, using Bayes rule

In a basin for a considerable period of time rainfall was measured using a dense network. Based on these values for the month July the following classification is used for the basin rainfall.

Class	Rainfall (mm)	Probability
Dry	$P < 50$	$P[A_1] = 0.15$
Moderate	$50 \le P < 200$	$P[A_2] = 0.50$
Wet	$200 \le P < 400$	$P[A_3] = 0.30$
Extremely wet	$P \ge 400$	$P[A_4] = 0.05$

Table 3.2: Rainfall classes and probability.

The probabilities presented in Table 3.2 refer to prior probabilities. Furthermore, from the historical record it has been deduced that the percentage of gauges, which gave a rainfall amount in a certain class given that the basin rainfall felt in a certain class is given in Table 3.3.

Basin rainfall	Percentage of gauges			
	$P < 50$	$50 \le P < 200$	$200 \le P < 400$	$P \ge 400$
$P < 50$	80	15	5	0
$50 \le P < 200$	25	65	8	2
$200 \le P < 400$	5	20	60	15
$P \ge 400$	0	10	25	65

Table 3.3: Conditional probabilities for gauge value given the basin rainfall

Note that the conditional probabilities in the **rows** add up to 100%.

For a particular year a gauge gives a rainfall amount for July of 230 mm. Given that sample value of 230 mm, what is the class of the basin rainfall in July for that year.

Note that the point rainfall falls in class III. The posterior probability of the actual basin rainfall in July of that year becomes:

$$P[A_i \mid \text{sample } 1] = \frac{P[\text{sample } 1 \mid A_i].P[A_i]}{\sum_{i=1}^{4} P[\text{sample } 1 \mid A_i].P[A_i]}$$

The denominator becomes:

$$\sum_{i=1}^{4} [\text{sample}1 \mid A_i] = 0.05x0.15 + 0.20x0.50 + 0.60x0.30 + 0.15x0.05 = 0.295$$

The denominator expresses the probability of getting sample 1 when the prior probabilities are as given in Table 3.2, which is of course very low.

Hence,

$$P[A_1 \mid \text{sample1}] = \frac{0.05x0.15}{0.295} = 0.025$$

$$P[A_2 \mid \text{sample1}] = \frac{0.20x0.50}{0.295} = 0.340$$

$$P[A_3 \mid \text{sample1}] = \frac{0.60x0.30}{0.295} = 0.610$$

$$P[A_4 \mid \text{sample1}] = \frac{0.15x0.05}{0.295} = 0.025$$

Note that the sum of posterior probabilities adds up to 1.

Now, for the same month in the same year from another gauge a rainfall of 280 mm is obtained. Based on this second sample the posterior probability of the actual July basin rainfall in that particular year can be obtained by using the above posterior probabilities as revised prior probabilities for the July rainfall:

$$\sum_{i=1}^{4}[\text{sample } 2 \mid A_i] = 0.05x0.025 + 0.20x0.340 + 0.60x0.610 + 0.15x0.025 = 0.439$$

$$P[A_1 \mid \text{sample2}] = \frac{0.05x0.025}{0.439} = 0.003$$

$$P[A_2 \mid \text{sample2}] = \frac{0.20x0.340}{0.439} = 0.155$$

$$P[A_3 \mid \text{sample2}] = \frac{0.60x0.610}{0.439} = 0.834$$

$$P[A_4 \mid \text{sample2}] = \frac{0.15x0.025}{0.439} = 0.008$$

Note that the denominator has increased from 0.240 to 0.478.

Again note that the posterior probabilities add up to 1. From the above it is seen how the probability on the state of July rainfall changes with the two sample values:

Class	Prior probability	After sample 1	After sample 2
I	0.15	0.025	0.003
II	0.50	0.340	0.155
III	0.30	0.610	0.834
IV	0.05	0.025	0.008

Table 3.4: Updating of state probabilities by sampling

Given the two samples, the probability that the rainfall in July for that year is of class III has increased from 0.30 to 0.834.

Question: What will be the change in the last column of Table 3.4 if the third sample gives a value of 180 mm?

3.2 Frequency distributions

3.2.1 Univariate distributions

Discrete random variables

Formally, given a data set x_1, x_2, …., x_N of a stochatic variable X, the **probability mass function** (**pmf**) $p_X(x)$ expresses:

$$p_X(x) = P(X = x) \tag{3.16}$$

and the **cumulative distribution function** (**cdf**) $F_X(x)$ gives the probability of occurrence $X \leq x$:

$$F_X(X) = P(X \leq x) = \sum_{\text{all} x_i \text{s} x} px(x_i) \qquad \text{and} \qquad \sum_{\text{all} x_i} px(x_i) = 1 \tag{3.17}$$

Continuous random variables

In terms of continuous random variables, the continuous equivalent of the pmf is the **probability density function** (**pdf**), $f_X(x)$. The probability that X takes on values in the interval (x, x + dx) then reads $f_X(x).dx$:

$$f_X(x).dx = P(x \leq X < x + dx) \tag{3.18}$$

The **cumulative probability density function** (**cdf**) $F_X(x)$ is now defined as:

$$F_x(x) = P(X \leq x) = \int_{-\infty}^{x} f_x(y)dy \qquad \text{and} \qquad \int_{-\infty}^{\infty} f_x(y)dy = 1 \tag{3.19}$$

The functions are displayed in Figure 3.6.

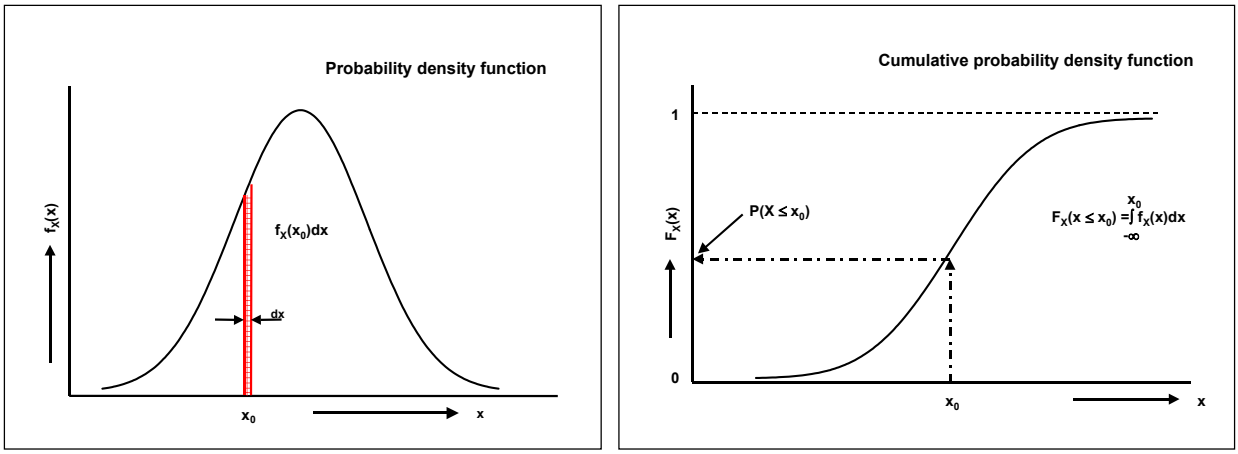

Figure 3.6: Probability density cumulative probability density function

$F_X(x)$ has the following properties:

- $F_X(-\infty) = 0$
- If $x_1 < x_2$ then $F_X(x_1) < F_X(x_2)$ ($F_X(x)$ is monotonous increasing)
- lim $F_X(x + h) = F_X(x)$ for $h \downarrow 0$ ($F_X(x)$ is right continuous)

For the pdf it follows:

$$f_X(x) = \frac{dF_X(x)}{dx} \qquad (3.20)$$

Example 3.5 Exponential pdf and cdf

The exponential pdf reads:

$f_X(x) = \lambda\exp(-\lambda x)$ for $x \geq 0$

Hence, the exponential cdf becomes with (3.19):

$$F_X(x) = \int_0^x \lambda\exp(-\lambda z)dz = -\exp(-\lambda z)\Big|_0^x = 1 - \exp(-\lambda x)$$

The exponential pdf and cdf for λ = 0.2 is shown in Figure 3.7. For example ($P(X \leq 7) = F_x(7)$ = 1 – exp(-0.2 x 7) = 0.75 as shown in Figure 3.7.

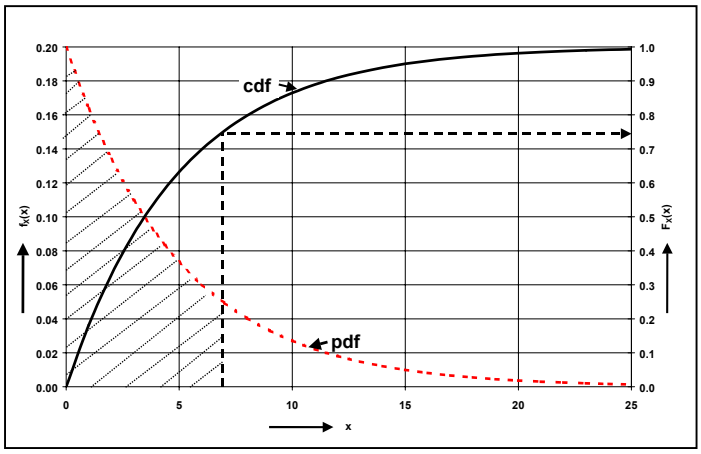

Figure 3.7:
Exponential pdf and cdf for λ=0.2

3.2.2 Features of distributions

In Chapter 2 some features of relative distribution functions were discussed. Here in a similar fashion this will be done for the pdf and the cdf. The following features of distributions are discussed:

- parameters
- return period
- mathematical expectation
- moments

Parameters

The distribution functions commonly used in hydrology are not specified uniquely by the functional form; the parameters together with the functional form describe the distribution. The parameters determine the **location**, **scale** and **shape** of the distribution.

Return period

The cdf gives the non-exceedance probability $P(X \leq x)$. Hence, the exceedance probability follows from: $P(X > x) = 1 - F_X(x)$ is. Its reciproke is called the return period. So if T is the return period and x_T is its corresponding quantile, then:

$$T = \frac{1}{P(X > x_T)} = \frac{1}{1 - P(X \le x_T)} = \frac{1}{1 - F_X(x_T)} \tag{3.21}$$

Note that in the above the notation for the quantile x_T or x(T) is used. Others use the notation x_p for quantile where $p = F_x(x_p)$, i.e. non-exceedance probability.

Mathematical expectation

If X is any continuous random variable with pdf $f_X(x)$, and if g(X) is any real-valued function, defined for all real x for which $f_X(x)$ is not zero, then the **mathematical expectation** of the function g(X) reads:

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx \tag{3.22}$$

Moments

If one chooses $g(X) = X^k$, where k = 1, 2, …. Then the k^{th} moment of X **about the origin** is defined by:

$$\mu_k' = E[X^k] = \int_{-\infty}^{+\infty} x^k f_X(x) dx \tag{3.23}$$

Note that an (') is used to indicate moments about the origin. Of special interest is the first moment about the origin, i.e. the mean:

$$\mu_1' = \mu_x = E[X] = \int_{-\infty}^{+\infty} x f_x(x) dx \tag{3.24}$$

If instead of the origin, the moment is taken around the mean, then the central moment follows (μ_k). Note that the accent (') is omitted here to denote a central moment. The second central moment is the variance:

$$\mu_2 = \mathrm{Var}(X) = E[(X - E[X])^2] = E[(X - \mu_X)^2] = \int_{-\infty}^{+\infty} (x - \mu_X)^2 f_X(x) dx \tag{3.25}$$

With the above one defines:

- the standard deviation σ_X, which expresses the spread around the mean in the same dimension as the original variate:

$$\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{\mu_2} \tag{3.26}$$

- the coefficient of variation C_v:

$$C_v = \frac{\sqrt{\mu_2}}{\mu_1'} = \frac{\sigma_X}{\mu_X} \tag{3.27}$$

- the skewness coefficient $\gamma_{1,x}$ of the distribution is defined by:

$$\gamma_{1,X} = \frac{\mu_3}{\sigma_X^3} = \frac{1}{\sigma_X^3} \int_{-\infty}^{+\infty} (x - \mu_X)^3 f_X(x) dx \tag{3.28}$$

- the peakedness of the distribution, expressed by the kurtosis $\gamma_{2,X}$:

$$\gamma_{2,X} = \frac{\mu_4}{\sigma_X^4} = \frac{1}{\sigma_X^4} \int_{-\infty}^{+\infty} (x - \mu_X)^4 f_X(x) dx \tag{3.29}$$

The parameter μ_X is a **location** parameter, σ_X a **scale** parameter, while $\gamma_{1,X}$ and $\gamma_{2,X}$ are **shape** parameters. The central moments μ_k are related to the moments about the origin μ_k' as follows:

$$\mu_1 = 0$$
$$\mu_2 = \mu_2' - (\mu_1')^2$$
$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 \qquad (3.30)$$

Example 3.7 Moments of the exponential distribution

Since the exponential pdf reads:

$f_X(x) = \lambda\exp(-\lambda x)$ for $x \geq 0$

its first moment about the origin is:

$$\mu_1' = \mu_X = \int_0^\infty x\lambda\exp(-\lambda x)dx = \lambda\left\{\frac{\exp(-\lambda x)}{-\lambda}\left(x + \frac{1}{\lambda}\right)\right\}\Bigg|_0^\infty = \lambda\left\{0 - \frac{1}{-\lambda}\left(0 + \frac{1}{\lambda}\right)\right\} = \frac{1}{\lambda}$$

It shows that the parameter λ is the reciproke of the mean value. The exponential distribution is well suited to model inter-arrival times, for example of flood occurrences. Then x has the dimension of time, and λ 1/time. If a flood of say 1,000 m^3/s is on average exceeded once every 5 years, and the exponential distribution applies, then μ_X = 5 years and hence λ = 1/5 = 0.2.

In extension to the above derivation, one can easily show, that the k^{th} order moments about the origin of the exponential distribution read:

$$\mu_k' = \frac{k!}{\lambda^k} \text{ hence : } \mu_1' = \frac{1}{\lambda};\ \mu_2' = \frac{2}{\lambda^2};\ \mu_3' = \frac{6}{\lambda^3};\ \mu_4' = \frac{24}{\lambda^4}$$

Then from (3.30) it follows for the central moments:

$$\mu_2 = \sigma_X^2 = \mu_2' - (\mu_1')^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

$$\mu_3 = \mu_3' - 3\mu_2'.\mu_1' + 2(\mu_1')^3 = \frac{6}{\lambda^3} - 3\frac{2}{\lambda^2}.\frac{1}{\lambda} + 2\left(\frac{1}{\lambda}\right)^3 = \frac{2}{\lambda^3}$$

$$\mu_4 = \mu_4' - 4\mu_3'.\mu_1' + 6\mu_2'.(\mu_1')^2 - 3(\mu_1')^4 = \frac{24}{\lambda^4} - 4\frac{6}{\lambda^3}.\frac{1}{\lambda} + 6\frac{2}{\lambda^2}.\left(\frac{1}{\lambda}\right)^2 - 3\left(\frac{1}{\lambda}\right)^4 = \frac{9}{\lambda^4}$$

And for the standard deviation, skewness and kutosis with (3.26), (3.28) and (3.29):

$$\sigma_X = \frac{1}{\lambda}$$

$$\gamma_{1,X} = \frac{\mu_3}{\sigma_X^3} = \frac{2/\lambda^3}{1/\lambda^3} = 2$$

$$\gamma_{2,X} = \frac{\mu_4}{\sigma_X^4} = \frac{9/\lambda^4}{1/\lambda^4} = 9$$

It is observed from the above that for the exponential distribution the mean and the standard deviation are the same. The distribution has a fixed positive skewness and a kurtosis of 9, which implies that the probability density of an exponential distribution is more closely concentrated around the mean than for a normal distribution.

3.2.3 Multivariate distribution functions

Occasionally, statistics about the joint occurrence of stochastic variables is of concern. In this subsection we discuss:

- Joint cdf and pdf
- Marginal cdf and pdf
- Conditional distribution function
- Moments
- Covariance and correlation

Joint distributions

The probability of joint events (i.e. intersections in the sample space) is given by the joint k-dimensional cdf $F_{X1,\ X2,\ \ldots,\ Xk}(x_1, x_2, \ldots, x_k)$.

In case of two stochastic variables X and Y the joint 2-dimensional cdf $F_{XY}(x,y)$ reads:

$$F_{XY}(x,y) = P(X \le x \cap Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(s,t)\,ds\,dt \tag{3.31}$$

where $f_{XY}(x,y)$ is the **joint 2-dimensional pdf**:

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} \tag{3.32}$$

Marginal distributions

The **marginal cdf** $F_X(x)$ of X only, gives the non-exceedance probability of X irrespective of the value of Y, hence

$$F_X(x) = P(X \le x \cap -\infty < Y < \infty) = F_{XY}(x,\infty) \tag{3.33}$$

and similarly the **marginal pdf** $f_X(x)$ reads:

$$f_X(x) = \frac{dF_X(x)}{dx} = \frac{d}{dx}F_{XY}(x,\infty) = \int_{-\infty}^{\infty} f_{XY}(x,t)\,dt \tag{3.34}$$

Conditional distribution

Analogous to (3.5) the **conditional distribution function** can be defined:

$$F_{X|Y}(x,y) = P(X \le x \mid Y \le y) = \frac{F_{XY}(x,y)}{F_Y(y)} \tag{3.35}$$

and the conditional pdf:

$$f_{X|Y}(x,y) = \frac{f_{XY}(x,y)}{f_Y(y)} \tag{3.36}$$

Independent variables

Equivalently to (3.8), if X and Y are **independent** stochastic variables, the distribution function can be written as:

$$F_{XY} = P(X \le x \cap Y \le y) = P(X \le x).P(Y \le y) = F_X(x).F_Y(y) \tag{3.37}$$

and similarly for the density function:

$$f_{XY}(x,y) = f_X(x).f_Y(y) \tag{3.38}$$

Moments

In addition to the moments for univariate distributions the moments for bivariate distributions are defined as follows:

$$\mu'_{k,m} = E\left[X^k Y^m\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^m f_{XY}(x,y)dxdy \tag{3.39}$$

Covariance and correlation function

Of special interest is the central moment expressing the linear dependency between X and Y, i.e. the **covariance**:

$$C_{XY} = E\left[(X - E[X])(Y - E[Y])\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f_{XY}(x,y)dxdy \tag{3.40}$$

Note that if X is independent of Y, then with (3.38) it follows:

$$C_{XY} = \int_{-\infty}^{\infty}(x-\mu_X)f_X(x)dx \int_{-\infty}^{\infty}(y-\mu_Y)f_Y(y)dy = 0 \tag{3.41}$$

As discussed in Chapter 2, a standardised representation of the covariance is given by the correlation coefficient ρ_{XY}:

$$\rho_{XY} = \frac{C_{XY}}{\sqrt{C_{XX}C_{YY}}} = \frac{C_{XY}}{\sigma_X\sigma_Y} \tag{3.42}$$

In Chapter 2 it was shown that ρ_{XY} varies between +1 (positive correlation) and –1 (negative correlation). If X and Y are independent, then with (3.41) it follows $\rho_{XY} = 0$.

Example 3.6: Bivariate exponential and normal distribution

Assume that storm duration and intensity, (X and Y), are both distributed according to an exponential distribution (see Kottegoda and Rosso, 1997):

$$F_X(x) = 1 - \exp(-ax), \quad x \ge 0;\ a > 0 \qquad F_Y(y) = 1 - \exp(-by), \quad y \ge 0;\ b > 0 \tag{3.43}$$

Their **joint cdf** given as a bivariate exponential distribution reads:

$$\begin{aligned} &F_{XY}(x,y) = 1 - \exp(-ax) - \exp(-by) + \exp(-ax - by - cxy) \\ &\text{with}: x,y \ge 0;\ a > 0, b > 0 \text{ and } 0 \le c \le ab \end{aligned} \tag{3.44}$$

Hence, with (3.32), the **joint pdf** becomes:

$$\begin{aligned} f_{XY}(x,y) &= \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} = \frac{\partial}{\partial y}\left(\frac{\partial F_{XY}(x,y)}{\partial x}\right) = \\ &= \frac{\partial}{\partial y}\{a\exp(-ax) - (a+cy)\exp(-ax-by-cxy)\} = \\ &= \{(a+cy)(b+cx) - c\}\exp(-ax-by-cxy) \end{aligned} \tag{3.45}$$

The joint exponential probability density function with a = 0.05 h^{-1}, b = 0.4 h/mm and c = 0.01 $mm^{-1}$is shown in Figure 3.8.

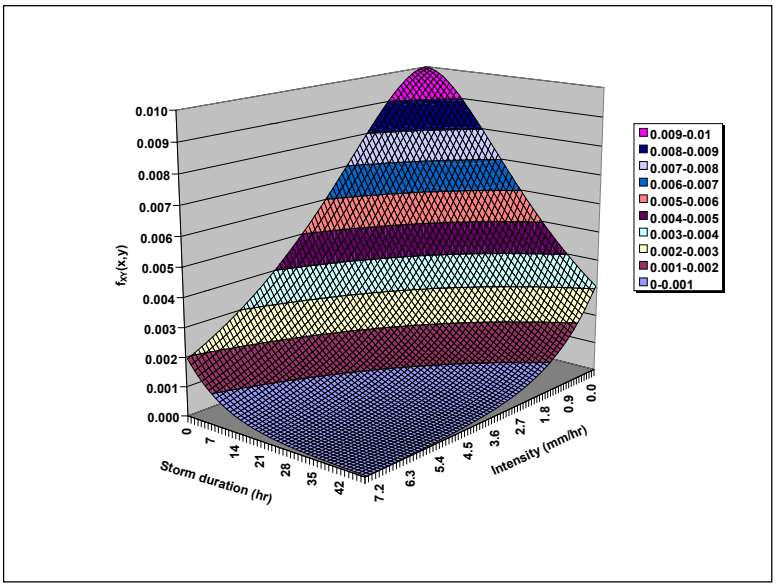

Figure 3.8:
Joint probability density function of storm duration and rainfall intensity

The **conditional pdf** of storm intensity given rain duration is:

$$f_{Y|X}(x,y) = \frac{f_{XY}(x,y)}{f_X(x)} \frac{\{(a+cy)(b+cx)-c\}\exp(-ax-by-cxy)}{a\exp(-ax)} =$$

$$= \frac{\{(a+cy)(b+cx)-c\}}{a}\exp(-y(b+cx)) \qquad (3.46)$$

The **conditional cdf** of a storm of given duration not exceeding a certain intensity reads:

$$F_{Y|X}(x,y) = \int_0^y f_{Y|X}(x,t)dt = \int_0^y \frac{\{a+ct)(b+cx)-c\}}{a}\exp(-t(b+cx))dt =$$

$$= 1 - \frac{a+cy}{a}\exp(-(b+cx)y) \qquad (3.47)$$

With a = 0.05 h^{-1}, b = 0.4 h/mm and c = 0.01 mm^{-1}, the conditional probability that a storm lasting 8 hours will exceed an average intensity of 4 mm/h becomes:

$$P(Y > 4 \mid X = 8) = 1 - F_{Y|X}(8,4) =$$

$$= 1 - 1 + \frac{0.05 + 0.01x4}{0.05}\exp(-(0.4 + 0.01x8)4) =$$

$$= 0.26$$

The **marginal distributions** follow from:

$$\left.\begin{aligned} f_X(x) &= \int_0^\infty f_{XY}(x,y)dy = \int_0^\infty \{(a+cy)(b+cx)-c\}\exp(-ax-by-cxy)dy = a\exp(-ax) \\ f_Y(y) &= \int_0^\infty f_{XY}(x,y)dx = \int_0^\infty \{(a+cy)(b+cx)-c\}\exp(-ax-by-cxy)dx = b\exp(-by) \end{aligned}\right\} \qquad (3.49)$$

If X and Y are **independent**, then c = 0 and it follows from (3.45):

$$f_{XY}(x,y) = ab\exp(-ax-by) = a\exp(-ax).b\exp(-by) = f_X(x).f_Y(y) \qquad (3.50)$$

Other examples of joint probability density functions are given in Figures 3.9 and 3.10, with the effect of correlation. In Figure 3.9 the joint standard normal pdf is given when the variables are independent, whereas in Figure 3.10 the variables are positively correlated (ρ = 0.8)

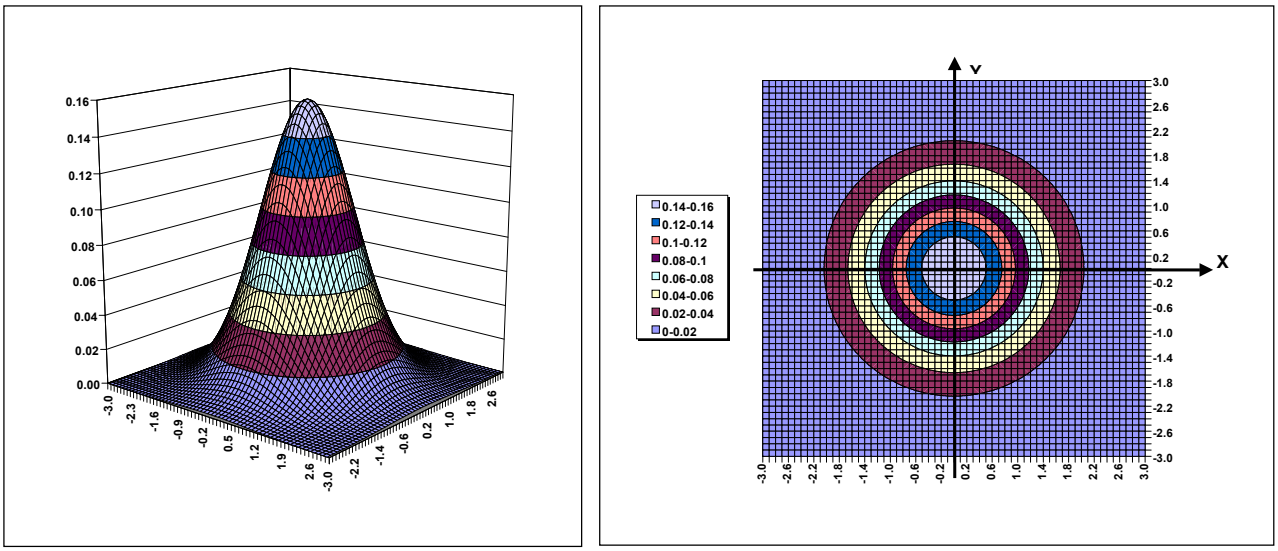

Figure 3.9: Bivariate standard normal distribution (ρ=0)

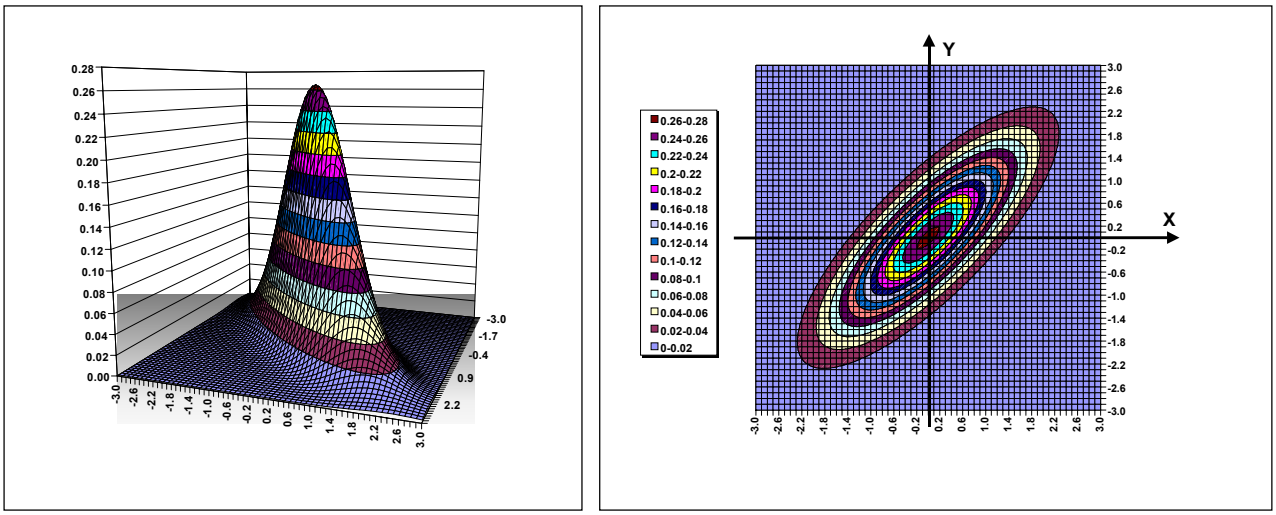

Figure 3.10: Bivariate standard normal distribution (ρ=08)

The effect of correlation on the probability density function is clearly observed from the density contours in the right hand side representations of the joint pdf's.

3.2.4 Moment generating function

In some cases the moments as discussed before, cannot be computed in a simple manner. Then, often, use can be made of an auxiliary function, called the **moment generating**

function G(s), which is the expectation of exp(sX): G(s) = E[exp(sX)]. In case of a continuous distribution:

$$G(s) = E[\exp(sX)] = \int_{-\infty}^{\infty} \exp(sx) f_X(x)dx \tag{3.50}$$

Assuming that differentiation under the integral sign is permitted one obtains:

$$\frac{d^k G(s)}{ds^k} = \int_{-\infty}^{\infty} x^k \exp(sx) f_X(x)dx \tag{3.51}$$

For s = 0 it follows: exp(sx) = 1, and the right hand side of (3.51) is seen to equal the k^{th} moment about the origin:

$$E[X^k] = G^{(k)}(0) \text{ where : } G^{(k)}(0) = \left.\frac{d^k G}{ds^k}\right|_{s=0} \tag{3.52}$$

Of course this method can only be applied to distributions for which the integral exists. Similar to the one-dimensional case, the moment generating function for bivariate distributions is defined by:

$$H(s,t) = E[\exp(sx + ty] = \int\int \exp((sx + ty) f_{XY}(x,y)dxdy \tag{3.53}$$

of which by partial differentiation to s and t the moments are found.

Example 3.7: Moment generating function for exponential distribution

The moment generating function for an exponential distribution and the k-th moments are according to (3.50) and (3.52):

$$G(s) = \int_0^{\infty} \exp(sx)\lambda \exp(-\lambda x)dx = \frac{\lambda}{\lambda - s}$$

$$E[X] = \left.\frac{dG}{ds}\right|_{s=0} = \left.\frac{\lambda}{(\lambda - s)^2}\right|_{s=0} = \frac{1}{\lambda}$$

$$E[X^2] = \left.\frac{d^2G}{ds^2}\right|_{s=0} = \left.\frac{2\lambda}{(\lambda - s)^3}\right|_{s=0} = \frac{2}{\lambda^2}$$

$$E[X^3] = \left.\frac{d^3G}{ds^3}\right|_{s=0} = \left.\frac{2x3\lambda}{(\lambda - s)^4}\right|_{s=0} = \frac{2x3}{\lambda^3} = \frac{6}{\lambda^3}$$

.....

$$E[X^k] = \left.\frac{d^kG}{ds^k}\right|_{s=0} = \left.\frac{2x3x...xk\lambda}{(\lambda - s)^{k+1}}\right|_{s=0} = \frac{k!}{\lambda^k} \tag{3.54}$$

3.2.5 Derived distributions

Consider the variables X and Y and their one to one relationship Y = h(X). Let the pdf of X be $f_X(x)$, then what is the pdf of Y? For this, consider Figure 3.11. It is observed that the probability that X falls in the interval x, x + dx equals the probability that Y falls in the interval y, y + dy. Hence,

$$f_Y(y)dy = f_X(x)dx \tag{3.55}$$

Since fY(y) cannot be negative, it follows:

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| \text{ or } f_Y(y) = |J| f_X(x) \qquad (3.56)$$

where the first derivative is called the **Jacobian** of the transformation, denoted by J.

In a similar manner bivariate distributions can be transformed.

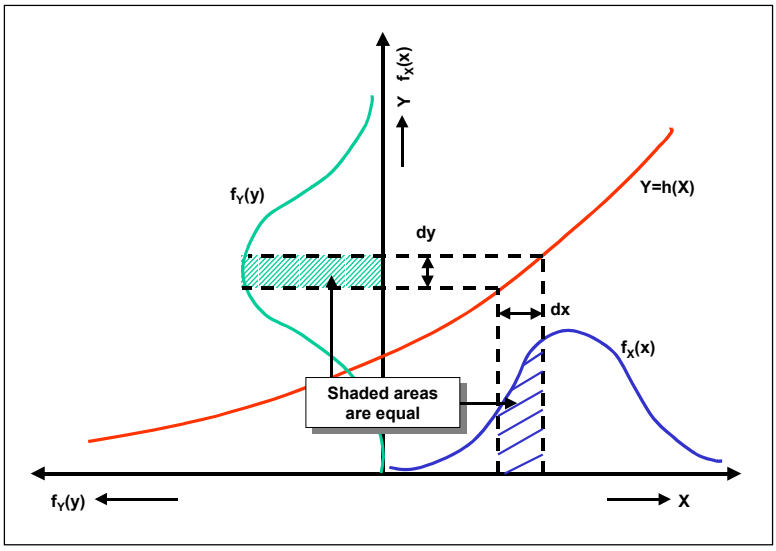

Figure 3.11: Definition sketch for derived distributions

Example 3.8: Transformation of normal to lognormal pdf

A variable Y is said to have a logarithmic normal or shortly log-normal distribution if its logarithm is normally distributed, hence ln(Y) = X. So:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right) \quad -\infty < X < \infty$$

$$X = \ln(Y)$$

$$\left|\frac{dx}{dy}\right| = \frac{1}{y}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}y\sigma_{\ln(Y)}}\exp\left(-\frac{1}{2}\left(\frac{\ln y-\mu_{\ln Y}}{\sigma_{\ln Y}}\right)^2\right) \quad 0 < Y < \infty$$

3.2.6 Transformation of stochastic variables

Consider the function Z = a + bX + cY, where X, Y and Z are stochastic variables and a, b and c are coefficients. Then for the mean and the variance of Z it follows:

$$E[Z] = E[a + bX + cY] = a + bE[X] + cE[Y] \qquad (3.57)$$

$$E[(Z - E[Z])^2] = E[(a + bX + cY - a - bE[X] - cE[Y])^2] =$$
$$= E[b^2(X - E[X])^2 + c^2(Y - E[Y])^2 + 2bc(X - E[X])(Y - E[Y])] =$$
$$= b^2E[(X - E[X])^2] + c^2E[(Y - E[Y])^2] + 2bcE[(X - E[X])(Y - E[Y])]$$

or:

$$Var(Z) = b^2Var(X) + c^2Var(Y) + 2bcCov(X,Y) \quad (3.58)$$

Equations (3.57) and (3.58) are easily extendible for any linear function Z of n-random variables:

$$Z = \sum_{i=1}^{n} a_i X_i$$

$$E[Z] = \sum_{i=1}^{n} a_i E[X_i] = \sum_{i=1}^{n} a_i \mu_i \quad (3.59)$$

$$Var(Z) = Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(X_i X_j) \quad (3.60)$$

Or in matrix notation by considering the vectors:

$$[\mathbf{a}] = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ . \\ . \\ a_n \end{bmatrix} \quad [\mathbf{X}] = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ . \\ . \\ X_n \end{bmatrix}$$

$$E[\mathbf{Z}] = E([\mathbf{a}]^T[\mathbf{X}]) = [\mathbf{a}]^T E([\mathbf{X}]) = [\mathbf{a}]^T[\mu]$$
$$\text{where : } E([\mathbf{X}]) = [\mu] \quad (3.61)$$

$$Var(\mathbf{Z}) = Var([\mathbf{a}]^T[\mathbf{X}]) = E([\mathbf{a}]^T([\mathbf{X}] - [\mu])([\mathbf{X}] - [\mu])^T[\mathbf{a}]) = [\mathbf{a}]^T[\mathbf{V}][\mathbf{a}]$$
$$\text{where : } [\mathbf{V}] = E(([\mathbf{X}] - [\mu])([\mathbf{X}] - [\mu])^T) \quad (3.62)$$

The matrix [**V**] contains the following elements:

$$[\mathbf{V}] = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \ldots & Cov(X_1,X_n) \\ Cov(X_1,X_2) & Var(X_2) & \ldots & Cov(X_2,X_n) \\ . & & & \\ . & & & \\ . & & & \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \ldots & Var(X_n) \end{bmatrix} \quad (3.63)$$

This matrix is seen to be symmetric, since $Cov(X_i,X_j) = Cov(X_j,X_i)$. This implies $[\mathbf{V}] = [\mathbf{V}]^T$. Furthermore, since the variance of a random variable is always positive, so is $Var([\mathbf{a}]^T[\mathbf{X}])$.

Taylor's series expansion

For non-linear relationships it is generally difficult to derive the moments of the dependent variable. In such cases with the aid of Taylor's series expansion approximate expressions for the mean and the variance can be obtained. If Z = g(X,Y), then (see e.g. Kottegoda and Rosso (1997)):

$$\left.\begin{aligned} E[Z] &\approx g(\mu_X,\mu_Y)+\frac{1}{2}\frac{\partial^2 g}{\partial x^2}\mathrm{Var}(X)+\frac{1}{2}\frac{\partial^2 g}{\partial y^2}\mathrm{Var}(Y)+\frac{\partial^2 g}{\partial x\partial y}\mathrm{Cov}(X,Y) \\ \mathrm{Var}(Z) &\approx \left(\frac{\partial g}{\partial x}\right)^2\mathrm{Var}(X)+\left(\frac{\partial g}{\partial y}\right)^2\mathrm{Var}(Y)+2\left(\frac{\partial g}{\partial x}\frac{\partial g}{\partial y}\right)\mathrm{Cov}(X,Y) \end{aligned}\right\} \qquad (3.64)$$

Above expressions are easily extendable to more variables. Often the variables in g(..) can be considered to be independent, i.e. Cov(..) = 0. Then (3.64) reduces to:

$$\left.\begin{aligned} E[Z] &\approx g(\mu_X,\mu_Y)+\frac{1}{2}\frac{\partial^2 g}{\partial x^2}\mathrm{Var}(X)+\frac{1}{2}\frac{\partial^2 g}{\partial y^2}\mathrm{Var}(Y) \\ \mathrm{Var}(Z) &\approx \left(\frac{\partial g}{\partial x}\right)^2\mathrm{Var}(X)+\left(\frac{\partial g}{\partial y}\right)^2\mathrm{Var}(Y) \end{aligned}\right\} \qquad (3.65)$$

Example 3.9

Given a function Z = X/Y, where X and Y are independent. Required are the mean and the variance of Z.

Use is made of equation (3.65). The coefficients read:

$$\left.\begin{aligned} \frac{\partial g}{\partial x} &= \frac{1}{y}\frac{\partial^2 g}{\partial x^2}=0 \\ \frac{\partial g}{\partial y} &= -\frac{x}{y^2}\frac{\partial^2 g}{\partial y^2}=\frac{2x}{y^3} \end{aligned}\right\} \qquad (3.66)$$

Hence:

$$\left.\begin{aligned} E[Z] &\approx \frac{\mu_X}{\mu_Y}+\frac{\mu_X}{\mu_Y^3}\sigma_Y^2=\frac{\mu_X}{\mu_Y}\left(1+C_{vY}^2\right) \\ \mathrm{Var}(Z) &\approx \left(\frac{1}{\mu_Y}\right)^2\sigma_X^2+\left(\frac{\mu_X}{\mu_Y^2}\right)^2\sigma_Y^2=\left(\frac{\mu_X}{\mu_Y}\right)^2\left(\frac{\sigma_X^2}{\mu_X^2}+\frac{\sigma_Y^2}{\mu_Y^2}\right)=\left(\frac{\mu_X}{\mu_Y}\right)^2\left(C_{vX}^2+C_{vY}^2\right) \end{aligned}\right\} \qquad (3.67)$$

Example 3.10: Joint cumulative distribution function

The joint pdf of X and Y reads:

$f_{XY}(x,y)=\exp(-x-y/2)$ for : $x>0, y>0$

$f_{XY}(x,y)=0$ for : $x\le 0, y\le 0$

Q: determine the probability that 2<X<5 and 1<Y<7

A: the requested probability is obtained from:

$$P(2<X<5\cap 1<Y<7)=\int_2^5\int_1^7 f_{XY}(x,y)\,dx\,dy=\int_2^5\exp(-x)dx\int_1^7\exp(-y/2)dy=(-\exp(-x)\big|_2^5)(-2\exp(-y/2)\big|_1^7)=$$

$$=(-0.0067-(-0.1353))(-0.0302-(-0.6065))=0.0741$$

Example 3.11: Marginal distributions and independence (from: Reddy, 1997)

Given is the joint pdf of the variables X and Y:

$$f_{XY}(x,y) = \frac{2}{3}(x+2y) \quad \text{for:} \quad 0 < x < 1, 0 < y < 1$$

$$f_{XY}(x,y) = 0 \quad \text{for:} \quad x \le 0, x \ge 1, y \le 0, y \ge 1$$

Q: a. find the marginal distributions of X and Y and
b. are X and Y independent?

A: a. the marginal distributions are obtained from:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy = \int_0^1 \frac{2}{3}(x+2y)dy = \frac{2}{3}(xy + 2\frac{y^2}{2})\Big|_0^1 = \frac{2}{3}\{(x+1)-(0)\} = \frac{2}{3}(x+1) \quad \text{for:} \quad 0 < x < 1$$

$$f_X(x) = 0 \quad \text{for:} \quad x \le 0, x \ge 1$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx = \int_0^1 \frac{2}{3}(x+2y)dx = \frac{2}{3}(\frac{x^2}{2} + 2xy)\Big|_0^1 = \frac{2}{3}\{(\frac{1}{2}+2y)-(0)\} = \frac{1}{3}(1+4y) \text{ for:} \quad 0 < y < 1$$

$$f_Y(y) = 0 \quad \text{for:} \quad y \le 0, y \ge 1$$

b. if X and Y are independent, then their conditional distributions should be equal to their marginal distributions. Hence is $f_{X|Y}(x,y) = f_X(x)$ or is $f_{Y|X}(x,y) = f_Y(y)$?

$$f_{X|Y}(x,y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{\frac{2}{3}(x+y)}{\frac{1}{3}(1+4y)} = 2\frac{(x+y)}{(1+4y)} \neq \frac{2}{3}(x+1)$$

So: $f_{X|Y}(x,y) \neq f_X(x)$, i.e. X and Y are not independent. A similar answer would of course have been obtained while examining $f_{Y|X}(x,y)$ relative to $f_Y(y)$.

Example 3.12: Joint pdf and independence (adapted from: Reddy, 1997)

Given are two variables X and Y who's marginal distributions read:

$$f_X(x) = 2a\ \exp(-bx) \quad \text{for:} \quad 0 \le x < \infty$$

$$f_Y(y) = 2a\ \exp(-by) \quad \text{for:} \quad 0 \le y < \infty$$

Q: a. find the joint pdf of X and Y if X and Y are independent
b. find the probability that X is always larger than Y

A: a. If X and Y are independent then their joint pdf is the product of their marginal distributions:

$$f_{XY}(x,y) = f_X(x).f_Y(y) = 4a^2\ \exp(-b(x+y)$$

b. the probability that X is always larger than Y can be obtained from the answer under a:

$$P(0 \le X < \infty \cap 0 \le Y < x) = \int_0^{\infty}\int_0^{x} f_{XY}(x,y)dydx = 4a^2 \int_0^{\infty} \exp(-bx)\left\{\int_0^{x} \exp(-by)dy\right\}dx =$$

$$4a^2\int_0^\infty \exp(-bx)\left\{\frac{\exp(-bx)}{-b}\Big|_0^x\right\}dx = 4a^2\int_0^\infty \exp(-bx)\left\{(-\frac{1}{b}\exp(-bx)) - (-\frac{1}{b})\right\}dx =$$

$$\frac{4a^2}{b}\int_0^\infty \exp(-bx)\{1-\exp(-bx)\}dx = \frac{4a^2}{b}\left\{\int_0^\infty \exp(-bx)dx - \int_0^\infty \exp(-2bx)dx\right\} =$$

$$\frac{4a^2}{b}\left\{\frac{\exp(-bx)}{-b}\Big|_0^\infty - \frac{\exp(-2bx)}{-2b}\Big|_0^\infty\right\} = \frac{4a^2}{b}\left\{\left(0-\frac{1}{-b}\right)-\left(0-\frac{1}{-2b}\right)\right\} = \frac{4a^2}{b}\left(\frac{1}{b}-\frac{1}{2b}\right) = 2\left(\frac{a}{b}\right)^2$$

4 Theoretical Distribution Functions

4.1 General

A number of theoretical (analytical) frequency distributions has been developed to model or represent the relative frequency distributions found in practice. In this chapter a summary is given of the distribution functions commonly used in hydrology and included in HYMOS.

A distinction is made between:

- Discrete distributions, and
- Continuous distributions.

A discrete distribution is used to model a random variable that has integer-valued outcomes, like the number of times an event occurs (successes) out of a number of trials. In contrast to this are the continuous distributions where the random variable is real-valued.

The **discrete** distributions (Section 4.2), which will be discussed, include:
- Binomial distribution
- Poisson distribution

The **continuous** distribution models comprise:
- Uniform distribution (Section 4.3),
- Distributions related to the normal distribution(Section 4.4), including:
  - Normal distribution
  - Log-normal distribution
  - Box-Cox transformations to normality
- Distributions related to Gamma or Pearson distribution, (Section 4.5), including:
  - Exponential distribution
  - Gamma distribution
  - Pearson Type 3 distribution or 3 parameter gamma distribution
  - Log-Pearson Type 3 distribution
  - Weibull distribution
  - Rayleigh distribution
- Distributions for extreme values(Section 4.6), including:
  - Generalised Extreme Value distributions, including the EV-1, EV-2 and EV-3 distributions for largest and smallest value
  - Generalised Pareto distributions, including Pareto Type 1, 2 and 3 distributions
- Sampling distributions(Section 4.7),:
  - Chi-square distribution
  - Student's t-distribution
  - Fisher F-distribution

It is stressed here that none of the theoretical distributions do have a physical background. They do not explain the physical phenomenon behind a population, but rather describe the behaviour of its frequency distribution. In this sub-section a short description of the various distributions is given.

Binomial distribution

The binomial distribution applies to a series of Bernoulli trials. In a Bernoulli trial there are two possible outcomes, that is an event occurs or does not occur. If the event occurs one speaks of a success (probability p) and if it does not occur it is a failure (probability 1 - p). If the probability of a success in each trial is constant, then the binomial distribution gives the distribution of the number successes in a series of independent trials. For example, the trial outcome could be that the water level in the river exceeds the crest of the embankment in a year and the other possible outcome that it does not. Let's call the event of an exceedance (how unfortunate for the designers) a "success". If the climatic conditions and the drainage characteristics in the basin do not vary one can assume that the success probability is constant from year to year. Knowing this success probability, then the Bernoulli distribution can be used to determine the probability of having exactly 0, 1, 2,…, or $\leq$1, $\leq$2, $\leq$…exceedances ("successes") during the next say 75 years (or any other number of years = number of trials). The distribution is therefore of extreme importance in risk analysis.

Poisson distribution

The Poisson distribution is a limiting case of the binomial distribution when the number of trials becomes large and the probability of success small, but their product finite. The distribution describes the number of occurrences of an event (a success) in a period of time (or space). Occurrences in a period of time (space) form a Poisson process if they are random, independent, and occur at some constant average rate. Essential is that the time (space) interval between the last occurrence and the next one is independent of past occurrences; a Poisson process, therefore, is **memory-less**.

Uniform distribution

The uniform distribution describes a random variable having equal probability density in a given interval. The distribution is particularly of importance for data generation, where the non-exceedance probability is a random variable with constant probability density in the interval 0,1.

Normal distribution

The normal distribution has a bell shaped probability density function, which is an appropriate model for a random variable being the sum of a large number of smaller components. Apart from being used as a sampling distribution or error model, the distribution applies particularly to the modelling of the frequency of aggregated data like monthly and annual rainfall or runoff. Direct application to model hydrological measurements is limited in view of its range from $-\infty$ to $+\infty$.

Lognormal distribution

If Y = ln X has normal distribution, then X is said to have a 2-parameter lognormal distribution. In view of its definition and with reference to the normal distribution, X can be seen as the product of a large number of small components. Its range from 0 to $+\infty$ is more appropriate to model hydrological series, whereas the logarithmic transformation reduces the positive skewness often found in hydrological data sets. Its applicability in hydrology is

further enhanced by introducing a shift parameter x_0 to X to allow a data range from x_0 to + ∞. Then, if $Y = \ln(X - x_0)$ has normal distribution it follows that X has a 3-parameter lognormal distribution

Box-Cox transformation

The Box-Cox transformation is a suitable, effective two-parameter transformation to data sets to normality. Such transformations may be desired in view of the extensive tabulation of the normal distribution.

Exponential distribution

The time interval between occurrences of events in a Poisson process or inter-arrival time is described by the exponential distribution, where the distribution parameter represents the average occurrence rate of the events.

Gamma distribution

The distribution of the time until the γth occurrence in a Poisson process has a gamma distribution. In view of the definition of the exponential distribution the gamma distribution models the sum of γ independent, identical exponentially distributed random variables. Note that γ may be a non-integer positive value. The gamma distribution is capable of modelling skewed hydrological data series as well as the lognormal distribution is capable of. The gamma distribution has a zero lower bound and is therefore not applicable to phenomena with a non-zero lower bound, unless a shift parameter is introduced.

Pearson Type 3 or 3-parameter gamma distribution

The gamma distribution with a shift parameter to increase the flexibility on the lower bound is called the Pearson Type 3 distribution. Sometimes it is also called 3-parameter gamma distribution, though in literature the name gamma distribution is generally related to the 2-parameter case. The distribution can take on variety of shapes like the 3-parameter lognormal distribution and is therefore often used to model the distribution of hydrological variables. A large number of distributions are related to the Pearson Type 3 distribution. For this, consider the standard incomplete gamma function ratio:

$$F(z) = \frac{1}{\Gamma(\gamma)} \int_0^z s^{\gamma-1} \exp(-s) ds \text{ where : } Z = \left(\frac{X - x_0}{\beta} \right)^k$$

Note that the distribution reduces to an exponential function when γ = 1. In the above distribution x0 = location parameter, β = scale parameter and γ and k are shape parameters. The following distributions are included:

- k = 1, γ = 1: exponential distribution
- k = 1, x_0 = 0: gamma distribution
- k = 1, x_0 = 0, β = 2, $\gamma = \nu/2$: chi-squared distribution
- k = 1: 3-parameter gamma or Pearson Type 3 distribution
- k = 1, $Z = (\ln(X - x_0) - y_0)/\beta)^k$: log-Pearson Type 3 distribution
- k = -1: Pearson Type 5 distribution
- k = 2, γ = 1: Rayleigh distribution
- k = 2, γ = 3/2: Maxwell distribution
- γ = 1: Weibull distribution

Log-Pearson Type 3 distribution

If $X = \ln(Y - y_0)$ has a Pearson Type 3 distribution, then Y follows a log-Pearson Type 3 distribution. The distribution is often used to model annual maximum floods when the skewness is high.

Weibull distribution

The Weibull distribution is a special type of exponential or Pearson Type 3 distribution. The Weibull distribution is often used to model distributions of annual minimum values and as such it equals the Extreme Value Type III distribution for smallest values.

Rayleigh distribution

The Rayleigh distribution is a special case of the Weibull distribution. By comparison with the definition of the chi-squared distribution it is observed that a random variable is Rayleigh distributed if it is the root of the sum of two squared normal random variables. The distribution is often used to model distributions of maximum wind speed but also for annual maximum flows, if the skewness is limited.

Generalised Extreme Value distributions

Three types of Extreme Value distributions have been developed as asymptotic distributions for the largest or the smallest values. It depends on the parent distribution which type applies. The distributions are often called Fisher-Tippett Type I, II and III or shortly EV-1, EV-2 and EV-3 distributions for largest and smallest value. EV-1 for largest is known as the Gumbel distribution, EV-2 for largest as Fréchet distribution and EV-3 for smallest value as Weibull or Goodrich distribution. Above models apply typically to annual maximum or minimum series. Despite the fact that these distributions have particularly been derived for extreme values, it does not mean that one of the types always applies. Often the lognormal, Pearson and log-Pearson Type 3, Weibull or Rayleigh distributions may provide a good fit.

Generalised Pareto distributions

The Pareto distributions are particularly suited to model the distribution of partial duration series or annual exceedance series. The Extreme Value distributions for the annual maximum value can be shown to be related to the Pareto distributions with an appropriate model for the number of exceedances. Consequently as for the Extreme Value distributions also for the generalised Pareto distributions three types are distinguished: Pareto Type 1, 2 and 3 distributions.

Sampling distributions

An estimate is thought of as a single value from the imaginary distribution of all possible estimates, called the sampling distribution. Sampling distributions are introduced to be able to give the likely range of the true value of a parameter for which an estimate is made.

Chi-squared distribution

The sum of ν squared normally distributed random variables has a chi-squared distribution, where ν is the number of degrees of freedom. The distribution is a special case of the gamma distribution. The distribution is used to describe the sampling distribution of the variance; also, it finds application in goodness of fit tests for frequency distributions.

Student's t-distribution

The sampling distribution of many statistics is approximately standard normal if the statistic is scaled by its standard deviation. If the latter is replaced by its sample estimate with ν degrees of freedom then the sampling distribution of the statistic becomes a Student's t-distribution with the same number of degrees of freedom. When the number of degrees of freedom is sufficiently large, the Student distribution can be replaced by the normal distribution. The t variable is the ratio of a normal and the root of a chi-distributed variable divided by the number of degrees of freedom.

Fisher F-distribution

The ratio of two chi-squared variables divided by their degrees of freedom has a Fisher F-distribution. The distribution is used in significance tests on difference between variances of two series.

4.2 Discrete distribution functions

4.2.1 Binomial distribution

Distribution and cumulative distribution function

A **Bernoulli trial** is defined as a trial with only two possible outcomes: a **success** or a **failure**, with constant probability p and (1-p) respectively. The outcomes of a series of such trials are independent. Let X be the random variable for the number of successes out of n trials. Its probability distribution $p_X(x)$ is then given by the **binomial distribution**:

$$p_X(x) \equiv P(X = x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \text{ with : } x = 0,1,2,...,n \text{ and : } \binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{4.1}$$

The cdf reads:

$$F_X(x) \equiv P(X \le x; n; p) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} \tag{4.2}$$

Moment related distribution parameters

The mean, variance and skewness are given by:

$$\mu_X = np$$

$$\sigma_X^2 = np(1-p) \tag{4.3a}$$

$$\gamma_{1,X} = \frac{(1-2p)}{\sqrt{np(1-p)}}$$

$$\gamma_{2,X} = 3 + \frac{1-6p(1-p)}{np(1-p)} \tag{4.3b}$$

From the skewness it is observed that only for p = 0.5 a symmetrical distribution function is obtained. For p < 0.5 the distribution is skewed to the right and for p > 0.5 skewed to the left. A few examples are given in Figure 4.1

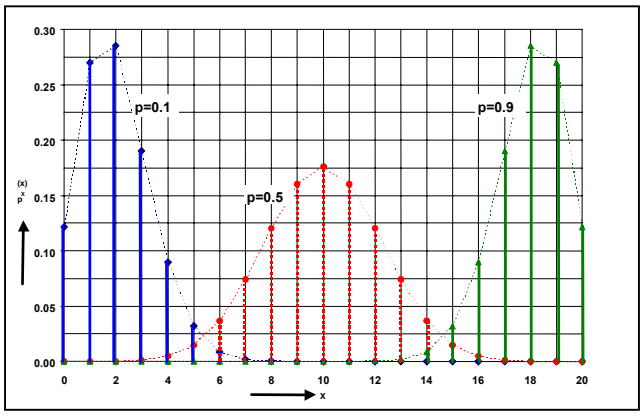

Figure 4.1:
Binomial distributions for n = 20 and p = 0.1, 0.5 and 0.9

From (4.3b) and Figure 4.2 it is observed that for large n, the skewness $\gamma_{1,X}$ gradually tends to 0 and the kurtosis $\gamma_{2,X}$ becomes close to 3. Then, the distribution approaches the **normal** distribution with same mean and variance (see Subsection 4.3.2).

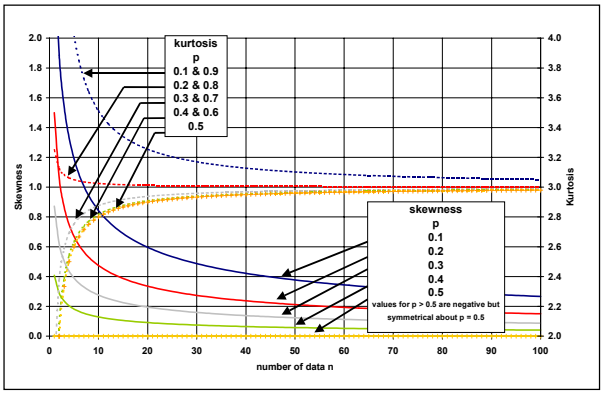

Figure 4.2:
Skewness and kurtosis of binomial distribution as function of n and p

Example 4.1 Number of rainy days in a week

Let the probability of a rainy day in a particular week in the year be 0.3, then:

- what is the probability of having exactly 4 rainy days in that week, and
- what is the probability of having at least 4 rainy days in that week?

Assuming that the occurrence of rainy days are independent, then the random variable X being the number of rainy days in that week follows a binomial distribution with n = 7 and p = 0.3. From (4.1) it then follows:

$$p_X(x) \equiv P(X = 4;7,0.3) = \binom{7}{4} 0.3^4 (1-0.3)^{7-4} = \frac{7!}{4!\,x3!} 0.3^4 0.7^3 = 0.097$$

Note that this is different from the probability of having 4 successive rainy days, which probability is 0.3 x 0.3 x 0.3 x 0.3 = 0.008, which is of course much less.

The probability of having at least 4 rainy days in that week of the year should be larger than 0.097, because also the probabilities of having 5, 6 or 7 days of rain should be included. The solution is obtained from (4.2):

$$F_X(X \geq 4) = 1 - F_X(X \leq 3) = 1 - \sum_{k=0}^{3} \binom{7}{k} 0.3^k (1-0.3)^{7-k} = 1 - (0.082 + 0.247 + 0.318 + 0.226) = 0.127$$

From the above it is observed that in case n and X are big numbers the elaboration of the sum will require some effort. In such cases the normal approximation is a better less cumbersome approach.

Related distributions

If the number of trials n = 1 then the binomial distribution is called **Bernoulli distribution** with mean p and variance p(1-p). The **geometric distribution** describes the probability that the first success takes place on the N^{th} trial. This distribution can be derived from (4.1) by noting that the N^{th} trial is preceded by (N – 1) trials without success, followed by a successful one. The probability of having first (N-1) failures is $(1-p)^{N-1}$ (from (3.12) or (4.1) with n = N-1 and x = 0) and the successful one has probability p, hence the probability of the first success in the N^{th} trial is $p(1-p)^{N-1}$ for N = 1,2,3,… In a similar manner the distribution function for the **negative binomial distribution** can be derived. This distribution describes the probability that the k^{th} exceedance takes place in the N^{th} trial. Hence, the N^{th} trial was preceded by (k-1) successes in (N-1) trials, which is given by (4.1) (with: n = N-1 and x = k-1), followed by a success with probability p.

4.2.2 Risk and return period

Consider a series of annual maximum discharges Q_{max} (t): t = 1,…, n. If a discharge Q_d is exceeded during these n years k-times then Q_d has in any one year an average probability of being exceeded of p_E = k/n and the average interval between the exceedances is n/k = $1/p_E$. The latter is called the **return period** T = $1/p_E$, as discussed in Sub-section 3.2.2, equation (3.21).

More generally, instead of Q_{max}, if we denote the random variable by Q, then the relation between $F_Q(q)$, T and p is:

$$F_Q(q) = P(Q \le q) = 1 - P(Q > q) = 1 - p_E = 1 - \frac{1}{T} \tag{4.4}$$

If one states that an embankment has been designed for a discharge with a return period of T years it means that **on average** only **once** during T years the river will overtop the embankment. But **each** year there is a probability p = 1/T that the river overtops the embankment. Consequently, each year the probability that the river does not overtop the embankment is $(1 - p_E) = F_Q(q)$. Since the outcomes in any one-year are independent, the probability of not being exceeded in N consecutive years is given by:

$$P(\text{no exceedances of q in N years}) = (F_Q(q))^N = (1-p_E)^N = \left(1 - \frac{1}{T}\right)^N \tag{4.5}$$

Note that this result is directly obtained from (4.1) with the number of successes x = 0. If q is the design level (storm, flow, stage, etc.), then the probability that this level q will be exceeded one or more times during the lifetime N of a structure (i.e. the probability of one or more failures), is simply the complement of the probability of no failures in N years. The probability of failure is called the **risk** r, hence:

$$r = 1 - (F_Q(q))^N = 1 - (1-p_E)^N = 1 - \left(1 - \frac{1}{T}\right)^N \tag{4.6}$$

It is noted that the above definition of risk is basically incomplete. The consequence of failure should also be taken into account. Risk is therefore often defined as the probability of failure times the consequence of failure.

Example 4.2 Risk of failure

A culvert has been designed to convey a discharge with a return period of 100 years. The lifetime of the structure is 50 years. What is the probability of failure during the lifetime of the structure?

$$r = 1 - \left(1 - \frac{1}{100}\right)^{50} = 1 - 0.605 = 0.395 \approx 40\%$$

Example 4.3 Return period and risk

To be 90% sure that a design discharge is not exceeded in an 80-year period, what should be the return period of the design discharge?

If we want to be 90% sure, then we take a risk of failure of 10%. From (4.6) it follows:

$$T = \frac{1}{1-(1-r)^{1/N}} = \frac{1}{1-(1-0.10)^{1/80}} = 760 \text{ years}$$

Hence for an event with an average return period of 760 years there is a 10% chance that in a period of 80 years such an event will happen.

4.2.3 Poisson distribution

Distribution and cumulative distribution function

If in the binomial distribution n becomes large and p very small, then (4.1) can be approximated by the **Poisson distribution**. Let the average number of successes in a series of n Bernoulli trials be ν = np, then the distribution of the number of successes X in n trials, with probability of occurrence in each trial of p, becomes, see also Figure 4.3:

$$p_X(x) \equiv P(X = x;\nu) = \frac{\nu^x \exp(-\nu)}{x!} \text{ for : } x = 0,1,2,....,n \qquad (4.7)$$

The cdf of the Poisson distribution reads:

$$F_X(x) \equiv P(X \le x;\nu) = \sum_{k=0}^{x} \frac{\nu^k \exp(-\nu)}{k!} \qquad (4.8)$$

Moment related distribution parameters

The mean, variance, skewness and kurtosis are:

$$\mu_X = \nu$$
$$\sigma_X^2 = \nu \qquad (4.9a)$$

$$\gamma_{1,X} = \frac{1}{\sqrt{\nu}}$$
$$\gamma_{2,X} = 3 + \frac{1}{\nu} \qquad (4.9b)$$

For $\nu \to \infty$ the skewness becomes 0 and the kurtosis 3, and the Poisson distribution converges to a normal pdf.

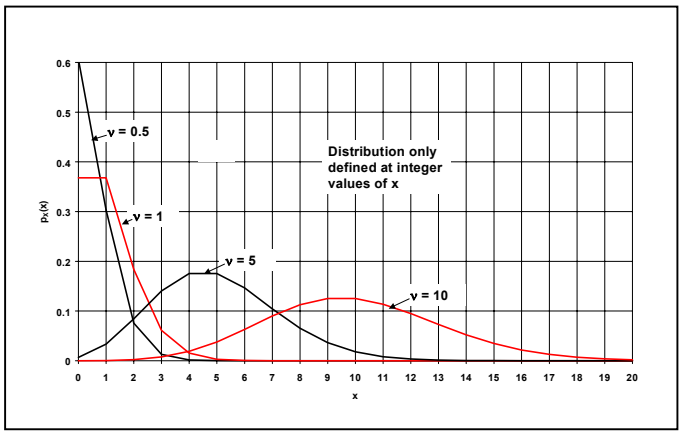

Figure 4.3:
Poisson distribution for different values of ν

Example 4.4: Drought

From a statistical analysis it was deduced that the monsoon rainfall at a location falls below 200 mm on average once in 100 years. What is the probability that the monsoon rainfall will fall below 200 mm less than twice in a 75-year period?

In this case n = 75 and the 'success' probability (falling below 200 mm) p = 1/100 = 0.01, hence n is large and p is small, which fulfils the condition for the applicability of the Poisson distribution. With $\nu = np = 75 \times 0.01 = 0.75$ it follows from (4.8):

$$F_X(x) = P(X \le 1;0.75) = \sum_{k=0}^{1} \frac{0.75^k \exp(-0.75)}{k!} =$$

$$= \left(\frac{0.75^0}{0!} + \frac{0.75^1}{1!} \right) \exp(-0.75) = (1 + 0.75)\exp(-0.75) = 0.8266$$

With the binomial cdf (4.2) we would have obtained:

$$F_X(x) = P(X \le 1\,;\,75\,;\,0.01) = \sum_{k=0}^{1} \binom{75}{k} 0.01^k (1-0.01)^{75-k} =$$

$$= 1x1x0.99^{75} + 75x0.01x0.99^{74} = 0.4706 + 0.3565 = 0.8271$$

$$P(T_a > t) = P(X = 0;\lambda t) = \frac{(\lambda t)^0 \exp(-\lambda t)}{0!} = \exp(-\lambda t) \qquad (4.10)$$

Hence the cumulative probability distribution of the time between arrivals becomes with (4.10):

$$F_{T_a}(t) = P(T_a \le t) = 1 - \exp(-\lambda t) \qquad (4.11)$$

It shows that the **waiting time** between successive events of a Poisson process follows an **exponential distribution**. Instead of time, the Poisson process can also be defined for space, length, etc. Essential for a Poisson process is that the "period" can be divided in subintervals Δt so small, that the probability of an arrival in Δt tends to $\lambda \Delta t$, while the probability of more than one arrival in Δt is zero and an occurrence in one subinterval is independent of the occurrence in any other, (Kottegoda and Rosso, 1997). This makes the process memory-less.

Example 4.2 continued Risk of failure

The average waiting time for the design event was 100 years. The structure will fail in the 50 year period, if the waiting time between the design events is less or equal to 50 years, which was defined as risk. From (4.11) with $\lambda = 1/T = 1/100$ and $t = N = 50$ we obtain:

$$r = F(T_a \le 50) = 1 - \exp(-\frac{1}{100}50) = 1 - 0.607 = 0.393$$

This result is seen to be close to the outcome of (4.6), which was r = 0.395.

4.3 Uniform distribution

Probability density and cumulative frequency distribution

The uniform or rectangular distribution describes the probability distribution of a random variable X, which has equal non-zero density in an interval 'ab' and zero density outside. Since the area under the pdf should equal 1, the pdf of X is given by:

$$\left.\begin{aligned} f_X(x) &= \frac{1}{b-a} \text{for}: a \le x \le b \\ f_X(x) &= 0 \ \text{ for}: x < a \,;\, x > b \end{aligned}\right\} \qquad (4.12)$$

The cdf of the uniform distribution reads:

$$\left.\begin{aligned} F_X(x) &= 0 \quad \text{for}: x < a \\ F_X(x) &= \int_a^x \frac{1}{b-a} ds = \frac{x-a}{b-a} \ \text{ for}: a \le x \le b \\ F_X(x) &= 1 \quad \text{for}: x > b \end{aligned}\right\} \qquad (4.13)$$

The pdf and cdf of the uniform distribution are shown in Figure 4.4.

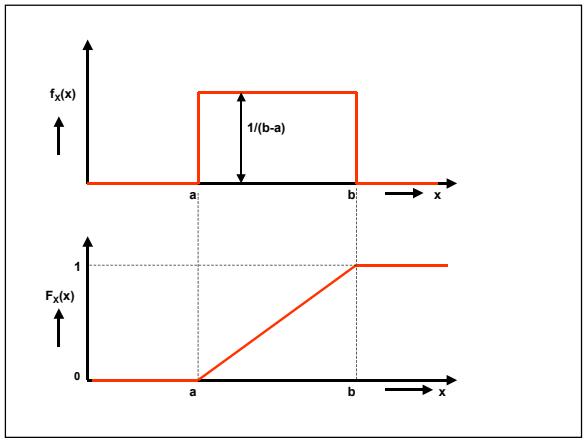

Figure 4.4:
Pdf and cdf of uniform distribution

Moment related distribution parameters

The mean and the variance simply follow from the definition of the moments:

$$\left.\begin{aligned} \mu_X &= \frac{a+b}{2} \\ \sigma_X^2 &= \frac{(b-a)^2}{12} \end{aligned}\right\} \qquad (4.14)$$

The uniform distribution is of particular importance for data generation, where with a = 0 and b = 1 the density function provides a means to generate the non-exceedance probabilities. It provides also a means to assess the error in measurements due to limitations in the scale. If the scale interval is c, it implies that an indicated value is ± ½ c and the standard deviation of the measurement error is $\sigma = \sqrt{(c^2/12)} \approx 0.3c$.

4.4 Normal distribution related distributions

4.4.1 Normal Distribution

Four conditions are necessary for a random variable to have a **normal** or **Gaussian distribution** (Yevjevich, 1972):

- A very large number of causative factors affect the outcome
- Each factor taken separately has a relatively small influence on the outcome
- The effect of each factor is independent of the effect of all other factors
- The effect of various factors on the outcome is additive.

Probability density and cumulative frequency distribution

The pdf and cdf of the normal distribution read:

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right) \quad \text{with} \quad -\infty < x < \infty, \quad -\infty < \mu_X < \infty \quad \text{and} \quad \sigma_X > 0 \tag{4.15}$$

$$F_X(x) \equiv P[X \le x] = \int_{-\infty}^{x} \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{s-\mu_X}{\sigma_X}\right)^2\right) ds \tag{4.16}$$

where: x = normal random variable
μ_X, σ_X = parameters of the distribution, respectively the mean and the standard deviation of X.

The pdf and cdf are displayed in Figure 4.5.

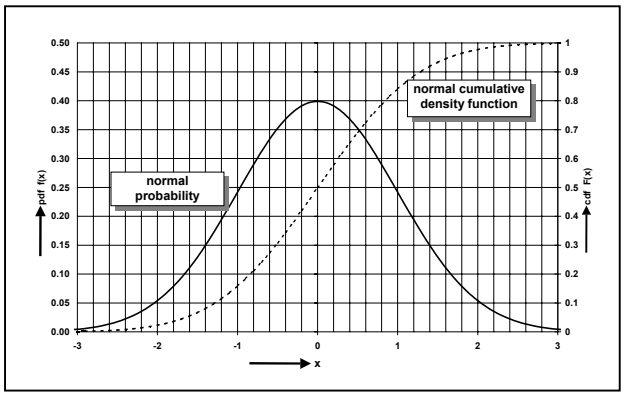

Figure 4.5:
Normal probability density and cumulative density functions for $\mu = 0$ and $\sigma = 1$

The normal pdf is seen to be a bell-shaped symmetric distribution, fully defined by the two parameters μ_X and σ_X. The coefficient $(\sigma_X\sqrt{(2\pi)})^{-1}$ in Equation (4.15) is introduced to ensure that the area under the pdf-curve equals unity, because the integral:

$$\int_{-\infty}^{\infty} \exp(-ax^2)dx = 2\int_{0}^{\infty} \exp(-ax^2)dx = \sqrt{\frac{\pi}{a}}$$

With a = $1/(2\sigma_X^2)$ the integral becomes $\sigma_X\sqrt{(2\pi)}$, so dividing the integral by the same makes the area under the pdf equal to 1.

The notation $N(\mu_X, \sigma_X^2)$ is a shorthand for the normal distribution. The normal pdf for different values of μ_X and of σ_X are shown in Figures 4.6 and 4.7. Clearly, μ_X is a **location** parameter; it shifts the distribution along the x-axis, but does not change the shape or scale of the distribution as is shown in Figure 4.6. The parameter σ_X is a **scale** parameter; it stretches or reduces the scale of the horizontal axis, see Figure 4.7, but it has no effect on the shape of the distribution.

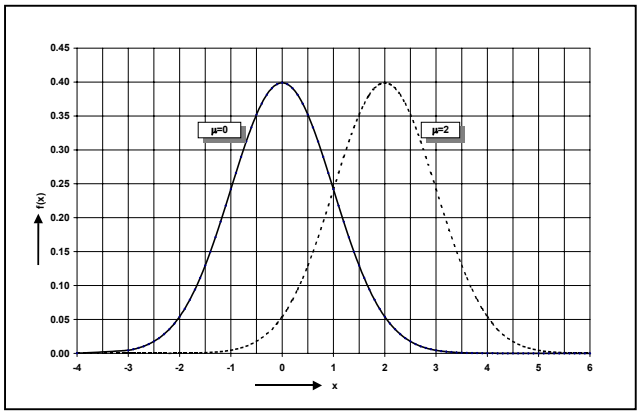

Figure 4.6:
Normal probability density functions for different values of μ_x' (σ_x=1)

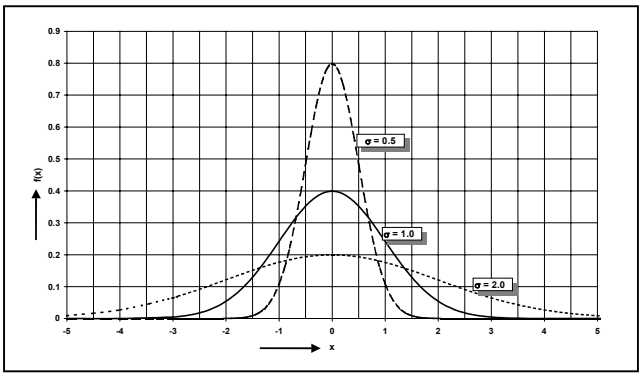

Figure 4.7:
Normal probability density functions for different values of σ_X, (μ_X = 0).

Moment related parameters of the distribution

The characteristics of the distribution are as follows:

Mean = median = mode:	μ_X	(4.17a)
Variance:	σ_X^2	(4.17b)
Standard deviation:	σ_X	(4.17c)
Coefficient of variation:	$C_{v,X} = \sigma_X/\mu_X$	(4.17d)
Skewness:	$\gamma_{1,X} = 0$	(4.17e)
Kurtosis:	$\gamma_{2,X} = 3$	(4.17f)

Standard normal distribution

The location and scale parameters μ_X and σ_X are used to define the **standard normal variate** or **reduced variate** Z:

$$Z = \frac{X - \mu_X}{\sigma_X} \qquad (4.18)$$

It is observed that Z = X for μ_X =0 and σ_X = 1, hence Z is an N(0,1) variate with pdf and cdf respectively:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} z^2) \tag{4.19}$$

$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-\frac{1}{2} s^2) ds = \frac{1}{2} + \frac{1}{2} \text{erf}(\frac{z}{\sqrt{2}}) \tag{4.20}$$

Equations (4.19) and (4.20) describe the **standard normal** probability density and cumulative density function, see Figure 4.5. From (4.18) it follows:

$$dz = \frac{1}{\sigma_X} dx$$

Substitution of this expression in (4.20) with (4.18) results in equation (4.16) and by taking the derivative with respect to X one obtains (4.15). The procedures used in HYMOS to solve (4.20) given Z and to calculate the inverse (i.e. the value of Z given $F_Z(z)$) are presented in Annex 4.1.

The standard normal distribution is generally tabulated in statistical textbooks. Such tables generally only address the positive arguments. To apply these tables for negative arguments as well, note that because of the symmetry of the pdf it follows:

$$f_Z(-z) = f_Z(z) \tag{4.21}$$

and

$$F_Z(-z) = 1 - F_Z(z) \tag{4.22}$$

Quantiles

Values of x_T and z_T for which $F_X(x_T) = F_Z(z_T) = 1 - 1/T$ are related by (4.18) and by its inverse:

$$x_T = \mu_X + \sigma_X z_T \tag{4.23}$$

Z_T is obtained as the inverse of the standard normal distribution.

Example 4.5 Tables of the normal distribution

For z = 2, $f_Z(2) = 0.0540$, hence $f_Z(-2) = 0.0540$

For z = 1.96 $F_Z(1.96) = 0.9750$,

Hence: $F_Z(-1.96) = 1 - 0.9750 = 0.0250$

It implies that the area under the pdf between z = -1.96 and z = 1.96 (see Figure 4.8) amounts 0.9750 – 0.0250 = 0.95 or 95%.

Given that the mean of a random variable is 100 and its standard deviation is 50, the quantile for T = 100 is derived as follows:

For T = 100, $F_Z(z) = 1 - 1/100 = 0.99$. From the table of the normal distribution this non-exceedance probability corresponds with a reduced variate $z_T = 2.33$. Hence, using (4.23):

$x_T = \mu_X + \sigma_X z_T = 100 + 50 \times 2.33 = 216.5$

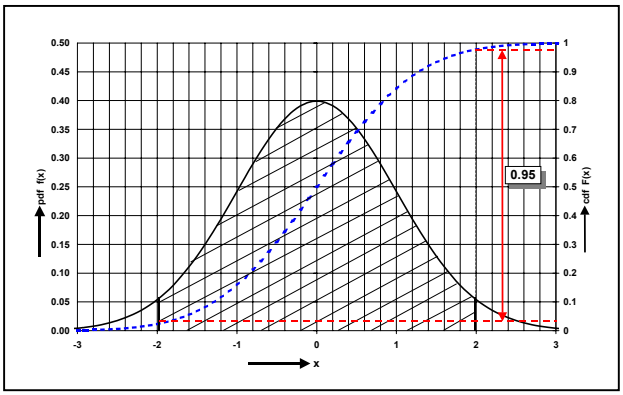

Figure 4.8:
Use of symmetry of standard normal pdf around 0 to find non-exceedance probabilities

Some Properties of the Normal Distribution

1. A linear transformation Y = a + bX of an N(μ_X, σ_X^2) random variable X makes Y an N(a + bμ_X, $b^2\sigma_X^2$) random variable.
2. If S_n is the sum of n independent and identically distributed random variables X_i each having a mean μ_X and variance σ_X^2, then in the limit as n approaches infinity, the distribution of S_n approaches a normal distribution with mean $n\mu_X$ and variance $n\sigma_X^2$.
3. Combining 1 and 2, for the mean X_m of X_i it follows, using the statement under 1 with a = 0 and b = 1/n, that X_m tends to have an N(μ_X, σ_X^2/n) distribution as n approaches infinity:

$$m_x = \frac{1}{n}\sum_{i=1} x_i = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \cdots + \frac{1}{n}x_n \text{ so: } E[m_x] = \frac{1}{n}E[E_{x_i}] = \frac{1}{n}\,E\left[\sum x_i\right] = \frac{1}{n}.n\mu_x = \mu_x$$

$$Var(m_x) = \frac{1}{n^2}Var(x_i) + \frac{1}{n^2}Var(x_i) + \cdots + \frac{1}{n^2}Var(x_n) = \frac{1}{n^2}\sum Var(x_i) = \frac{1}{n^2}\cdot n\,Var(x) = \frac{\sigma_x^2}{n}$$

If X_i is from an N(μ_X, σ_X^2) population, then the result for the sum and the mean holds regardless of the sample size n. The **Central Limit Theorem**, though, states that **irrespective of the distribution of X_i** the sum S_n and the mean X_m will tend to normality asymptotically. According to Haan (1979) if interest is in the main bulk of the distribution of S_n or X_m then n as small as 5 or 6 will suffice for approximate normality, whereas larger n is required for the tails of the distribution of S_n or X_m. It can also be shown that even if the X_i's have different means and variances the distribution of S_n will tend to be normal for large n with N($\Sigma\mu_{Xi}$; $\Sigma\sigma_{Xi}^2$), provided that each X_i has a negligible effect on the distribution of S_n, i.e. there are no few dominating X_i's.

An important outcome of the Central Limit Theorem is that if a hydrological variable is the outcome of n independent effects and n is relatively large, the distribution of the variable is approximately normal.

Application in hydrology

The normal distribution function is generally appropriate to fit annual rainfall and annual runoff series, whereas quite often also monthly rainfall series can be modelled by the normal distribution. The distribution also plays an important role in modelling random errors in measurements.

4.4.2 Lognormal Distribution

Definition

In the previous section it was reasoned, that the addition of a large number of small random effects will tend to make the distribution of the aggregate approximately normal. Similarly, a phenomenon, which arises from the **multiplicative** effect of a large number of uncorrelated factors, the distribution tends to be lognormal (or logarithmic normal); that is, the logarithm of the variable becomes normally distributed (because if $X = X_1X_2X_3\ldots$. Then $\ln(X) = \ln(X_1) + \ln(X_2) + \ln(X_3) + \ldots$).
Let X be a random variable such that $X - x_0 > 0$ and define

$$Y = \ln(X - x_0) \tag{4.23}$$

If Y has a normal distribution $N(\mu_Y, \sigma_Y^2)$, then X is said to have a 3-parameter log-normal distribution $LN(x_0, \mu_Y, \sigma_Y)$ or shortly LN-3. If x_0 is zero (or given) then the distribution of X is called a 2-parameter log-normal distribution $LN(\mu_Y, \sigma_Y)$ or LN-2.

Probability density and cumulative frequency distribution

The pdf of the normal random variable Y is given by:

$$f_Y(y) = \frac{1}{\sigma_Y\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right) \tag{4.24}$$

The pdf of X is obtained from the general transformation relation (3.56):

$$f_X(x) = f_Y(y)\left|\frac{dy}{dx}\right|$$

Since $Y = \ln(X - x_0)$ so: $|dy/dx| = 1/(X - x_0)$ for $X > x_0$, it follows from (4.24) for the pdf of X:

$$f_X(x) = \frac{1}{(x-x_0)\sigma_Y\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\ln(x-x_0)-\mu_Y}{\sigma_Y}\right)^2\right) \quad \text{with :} \quad x > x_0 \tag{4.25}$$

Equation (4.25) is the LN-3 pdf. The LN-2 pdf follows from (4.25) with $x_0 = 0$:

$$f_X(x) = \frac{1}{x\sigma_Y\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\ln x-\mu_Y}{\sigma_Y}\right)^2\right) \quad \text{with :} \quad x > 0 \tag{4.26}$$

To appreciate the parameters of the distribution, note the relation between the moment related parameters of the distribution and the parameters x_0, μ_Y and σ_Y:

Moment related parameters

$$\left.\begin{array}{ll} \text{Mean :} & \mu_X = x_0 + \exp(\mu_Y + \frac{1}{2}\sigma_Y^2) \\ \text{Median :} & M_X = x_0 + \exp(\mu_Y) \\ \text{Mode :} & m_X = x_0 + \exp(\mu_Y - \sigma_Y^2) \end{array}\right\} \tag{4.27a}$$

$$
\left.
\begin{aligned}
&\text{Variance :} && \sigma_X^2 = \left(\exp(\mu_Y + \frac{1}{2}\sigma_Y^2)\right)^2 \left(\exp(\sigma_Y^2) - 1\right) \\
&\text{Stdv :} && \sigma_X = \exp(\mu_Y + \frac{1}{2}\sigma_Y^2)\sqrt{\exp(\sigma_Y^2) - 1} \\
&\text{Parameter } \eta : && \eta = \frac{\sigma_X}{\mu_X - x_0} = \sqrt{\exp(\sigma_Y^2) - 1} \\
&\text{Skewness :} && \gamma_{1,X} = \eta^3 + 3\eta \\
&\text{Kurtosis :} && \gamma_{2,X} = 3 + 16\eta^2 + 15\eta^4 + 6\eta^6 + \eta^8
\end{aligned}
\right\} \qquad (4.27b)
$$

It is observed from the above equations that the first moment parameters are dependent on x_0, μ_Y and σ_Y. The variance depends on μ_Y and σ_Y, whereas the skewness and kurtosis are only dependent on σ_Y. This is also illustrated in the Figures 4.9 to 4.11. Clearly, x_0 is a ***location*** parameter (see Figure 4.9); it shifts only the distribution function, whereas μ_Y is a ***scale*** parameter, as the latter does not affect the skewness (see Figure 4.10). The parameter σ_Y is a ***shape*** parameter, since it affects the shape of the pdf as is deduced from (4.27) and Figure 4.11).

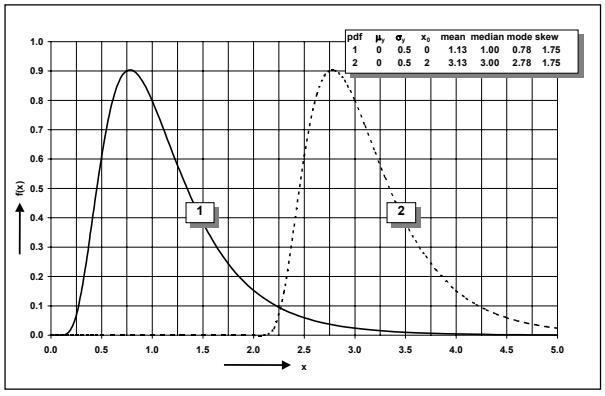

Figure 4.9:
Effect of location parameter x_0 on lognormal distribution

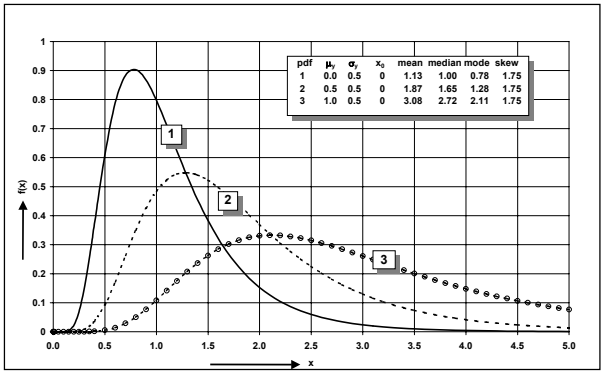

Figure 4.10:
Effect of scale parameter μ_y on lognormal distribution

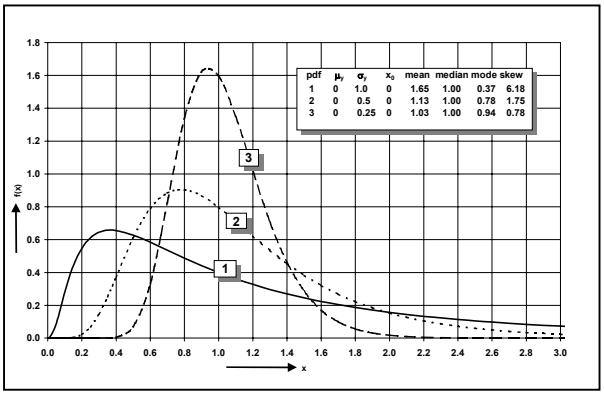

Figure 4.11:
Effect of is a shape parameter σ_y on lognormal distribution

Equation (4.27a) shows that for a lognormal distribution the following inequality holds:

x_0 < mode < median < mean

From (4.27b) it is observed that $\eta > 0$ hence $\gamma_1 > 0$ and $\gamma_2 > 3$; so the skewness is always positive and since the kurtosis is greater than 3 the lognormal distribution has a relatively greater concentration of probability near the mean than a normal distribution. The relation between γ_1 and η is displayed in Figure 4.12.To cope with negative skewness and distributions of smallest values, the sign of X or $(X-x_0)$ has to be changed, see Sub-section 4.3.13.

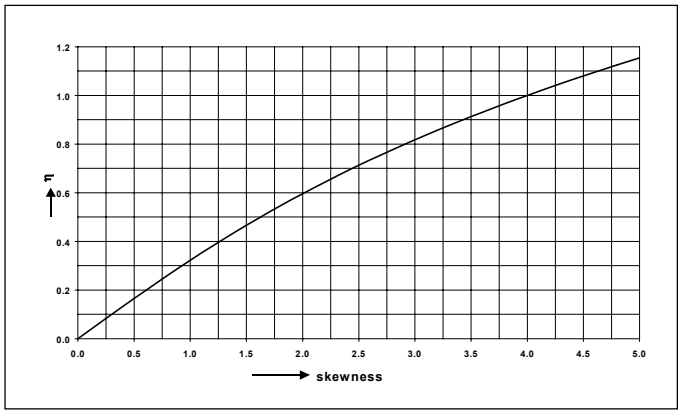

Figure 4.12:
η as function of skewness γ_1

Distribution parameters expressed in moment related parameters

The distinction between LN-2 and LN-3 is important. From equation (4.27) it is observed that when $x_0 = 0$ the parameters μ_Y and σ_Y are fully determined by the first two moments μ_X and σ_X which then also determine the skewness and kurtosis through their fixed relation with the coefficient of variation η.

For LN-2 the following inverse relations can be derived:

$$\mu_Y = \ln(\mu_X) - \frac{1}{2}\ln\left(\left(\frac{\sigma_X}{\mu_X}\right)^2 + 1\right) = \ln(\mu_X) - \frac{1}{2}\ln(C_{v,X}^2 + 1) \tag{4.28}$$

$$\sigma_Y = \sqrt{\ln\left(\left(\frac{\sigma_X}{\mu_X}\right)^2 + 1\right)} = \sqrt{\ln(C_{v,X}^2 + 1)} \tag{4.29}$$

The mean and the coefficient of variation of X are seen to describe the LN-2 pdf.

For **LN-3** the inverse relations are more complex as the starting point is the cubic equation in η relating η and $\gamma_{1,X}$, from (4.27b):

$$\eta^3 + 3\eta - \gamma_{1,X} = 0 \tag{4.30}$$

The parameters of the LN-3 distribution can be expressed in η (i.e. $\gamma_{1,X}$), μ_X and σ_X :

$$\eta = \left(\frac{\gamma_{1,X}}{2} + \sqrt{1 + \left(\frac{\gamma_{1,X}}{2} \right)^2} \right)^{1/3} - \left(-\frac{\gamma_{1,X}}{2} + \sqrt{1 + \left(\frac{\gamma_{1,X}}{2} \right)^2} \right)^{1/3} \tag{4.31}$$

The parameters of the LN-3 distribution can be expressed in η (i.e. $\gamma_{1,X}$), μ_X and σ_X:

$$x_0 = \mu_X - \frac{\sigma_X}{\eta} \tag{4.32}$$

$$\sigma_Y = \ln(\eta^2 + 1) \tag{4.33}$$

$$\mu_Y = \ln(\mu_X - x_0) - \frac{1}{2}\sigma_Y^2 = \frac{1}{2}\ln\left(\frac{\sigma_X^2}{\eta^2(\eta^2+1)} \right) \tag{4.34}$$

If the parameters would be determined according to equations (4.32) to (4.34) one observes that the **shape** parameter σ_Y is solely determined by the skewness, the **scale** parameter μ_Y by the variance and the skewness and the **location** parameter x_0 by the first three moments.

Moment generating function

The expressions presented in (4.27a/b) can be derived by observing that:

$$E[(X - x_0)^k] = \int_{-\infty}^{\infty} (x - x_0)^k f_X(x)dx = \int_{-\infty}^{\infty} \exp(ky) \frac{1}{\sigma_Y \sqrt{2\pi}} \exp(-\frac{1}{2}\left(\frac{y - \mu_Y}{\sigma_Y} \right)^2 dy = E[\exp(kY)] =$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(ky - \frac{1}{2}\frac{(y-\mu_y)^2}{\sigma_y} \right) dy$$

$$\text{with : } u = \frac{y - \mu_Y}{\sigma_Y} - k\sigma_Y \text{ it follows :}$$

$$u^2 = \left(\frac{y-\mu_y}{\sigma_y} - k\sigma_y \right)^2 = \left(\frac{y-\mu_y}{\sigma_y} \right)^2 - 2k\left(y - \mu_y\right) + k^2\sigma_y^2$$

$$-\frac{1}{2}u^2 = -\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y} \right)^2 + ky - k\mu_y - \frac{1}{2}k^2\sigma_y^2 \quad \text{or :} \quad ky - \frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y} \right)^2 = k\mu_y + \frac{1}{2}k^2\sigma_y^2 - \frac{1}{2}u^2$$

Hence, the power of the exponential can be replaced by:

$$k\mu_y + \frac{1}{2}k^2\sigma_y^2 - \frac{1}{2}u^2 \quad \text{and with :} \quad du = \frac{1}{\sigma_y}dy \text{ or : } dy = \sigma_y du \text{ one gets :}$$

$$E[\exp(kY)] = \exp(k\mu_Y + \frac{1}{2}k^2\sigma_Y^2) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)du$$

The last integral is seen to be 1, hence it follows for $E[(X-x_0)^k] = E[\exp(kY)]$:

$$E[(X - x_0)^k] = \exp(k\mu_Y + \frac{1}{2}k^2\sigma_Y^2) \tag{4.35}$$

Quantiles

The non-exceedance probability of the lognormally distributed variable X is derived through the standard normal distribution by inserting the standard normal variate Z derived as follows:

$$Z = \frac{Y - \mu_Y}{\sigma_Y} \text{ and : } Y = \ln(X - x_0) \tag{4.36}$$

The computation of the standard normal distribution is presented in Annex A4.1 or is obtained from tables in statistical textbooks.

The reverse, given a return period T or non-exceedance probability p, the quantile x_T or x_p is obtained from the standard normal distribution presented in Annex A4.2 or from tables through the standard normal deviate Z as follows:

$$x_T = x_0 + \exp(\mu_Y + z_T \sigma_Y) \tag{4.37}$$

Example 4.6 Lognormal distribution

Given is a LN-3 distributed variate X with mean 20, standard deviation 6 and skewness 1.5. Derive:

- the quantile for T=10.
- Return period of x = 35

To solve the first problem use is made of equation (4.37). The reduced variate z_T is obtained as the inverse of the standard normal distribution for a non-exceedance probability of $F_Z(z_T)$ =1-1/10 = 0.9. From the tables of the standard normal distribution one obtains:

z_T = 1.282

Next application of (4.37) requires values for the parameters x_0, σ_Y and μ_Y. These are determined using equations (4.31) to (4.34). The parameter η as a function of the skewness follows from (4.31), which gives with $\gamma_{1,X}$ = 1.5:

$$\eta = \left(\frac{\gamma_{1,X}}{2} + \sqrt{1 + \left(\frac{\gamma_{1,X}}{2} \right)^2} \right)^{1/3} - \left(-\frac{\gamma_{1,X}}{2} + \sqrt{1 + \left(\frac{\gamma_{1,X}}{2} \right)^2} \right)^{1/3} = 1.260 - 0.794 = 0.466$$

Then for x_0, σ_Y and μ_Y it follows from (4.32) to (4.34) respectively:

$$x_0 = \mu_X - \frac{\sigma_X}{\eta} = 20 - \frac{6}{0.466} = 7.130$$

$$\sigma_Y^2 = \ln(\eta^2 + 1) = 0.197 \text{ so : } \sigma_Y = 0.444$$

$$\mu_Y = \ln(\mu_X - x_0) - \frac{1}{2}\sigma_Y^2 = 2.456$$

Hence with (4.37) one obtains for the quantile x_T:

$$x_T = x_0 + \exp(\mu_Y + z_T \sigma_Y) = 7.13 + \exp(2.456 + 1.282 \text{x} 0.444) = 7.13 + 20.60 = 27.7$$

To solve the second problem, use is made of equation (4.36). The normal variate y is derived from the LN-3 variate x = 35 and x_0:

$$y = \ln(x - x_0) = \ln(35 - 7.130) = 3.328$$

$$z = \frac{y - \mu_Y}{\sigma_Y} = \frac{3.328 - 2.456}{0.444} = 1.963$$

Since z is a standard normal variate, the non-exceedance probability attached to Z is found from the standard normal distribution:

$$F_Z(z) = P(Z \le 1.963) = 0.975$$

$$P(Z > 1.963) = 1 - 0.975 = 0.025$$

$$T = \frac{1}{P(Z > 1.963)} = \frac{1}{0.025} = 40$$

Application in hydrology

The lognormal distribution function finds wide application in hydrology. It is generally appropriate to fit monthly rainfall and runoff series, whereas quite often also annual maximum discharge series can be modelled by the lognormal distribution.

4.4.3 Box-Cox transformation

Transformation equations

Box and Cox (1964) describe a general transformation of the following form:

$$\begin{aligned} Y &= \frac{(X - x_0)^{\lambda} - 1}{\lambda} \text{for} : \lambda \neq 0 \\ Y &= \ln(X - x_0) \text{for} : \lambda = 0 \end{aligned} \qquad (4.38)$$

The transformed variable Y has, by approximation, a normal distribution $N(\mu_Y, \sigma_Y)$. The transformation is seen to have two parameters, a **location** or shift parameter x_0 and the **power** and **scale** parameter λ.

The **reduced variate** Z, defined by:

$$Z = \frac{Y - \mu_Y}{\sigma_Y} \qquad (4.39)$$

with Y defined by (4.38) has a standard normal distribution. Once x_0 and λ are known, with the inverse of (4.39) and (4.38) the quantiles can be derived from the standard normal distribution.

Quantiles

For a particular return period T it follows for **quantile** x_T:

$$\left.\begin{aligned} x_T &= x_0 + \left(1 + \lambda(\mu_Y + z_T \sigma_Y)\right)^{1/\lambda} \quad \text{for} : \lambda \neq 0 \\ x_T &= x_0 + \exp(\mu_Y + z_T \sigma_Y) \qquad \text{for} : \lambda = 0 \end{aligned}\right\} \qquad (4.40)$$

It is noted that for very extreme values this transformation should not be used in view of the normality by approximation. In HYMOS flexibility is added by considering $|X-x_0|$ instead of $(X-x_0)$.

Application of the transformation shows that it returns a transformed series Y with a skewness close to zero and a kurtosis near 3.

Example 4.7 Box-Cox transformation

An example of its application is given below for annual maximum rainfall for Denee (Belgium), period 1882-1993.

Statistics before Box-Cox transformation	
Number of data	112
Mean	37.0
Standard deviation	11.8
Skewness	1.23
Kurtosis	4.56
Statistics after Box-Cox transformation with x_0 = 15.0 and λ = 0.142	
Number of data	112
Mean	3.70
Standard deviation	0.81
Skewness	0.00
Kurtosis	3.05

Table 4.1: Results of Box-Cox transformation on annual maximum rainfall

From the result it is observed that the skewness and kurtosis of the transformed variable are indeed close to 0 and 3. On the other hand λ is seen to be very small. It implies that the normal variates will be raised to a very high power to arrive at the quantiles, which is rather unfortunate. In such a case a lognormal distribution would be more appropriate.

4.5 Gamma or Pearson related distributions

4.5.1 Exponential distribution

Probability density and cumulative frequency distribution

In Sub-section 4.2.2 the exponential distribution was derived from the Poisson distribution. The exponential distribution models the distribution of the waiting time between successive events of a Poisson process. The exponential distribution is a special case of the gamma or Pearson Type 3 distribution (see next sub-sections). The general form of the exponential distribution is given by:

$$f_X(x) = \frac{1}{\beta}\exp(-\frac{x - x_0}{\beta}) \quad \text{for : } x > x_0 \tag{4.41}$$

and the cdf reads:

$$F_X(x) = \frac{1}{\beta}\int_{x_0}^{x} \exp(-\frac{s - x_0}{\beta})ds = 1 - \exp(-\frac{x - x_0}{\beta}) \tag{4.42}$$

The distribution is seen to have 2 parameters x_0 and β and will therefore be denoted by **E-2**. With x_0 = 0 it reduces to 1-parameter exponential distribution **E-1**.

Standardised distribution

Introducing the **reduced variate** Z:

$$Z = \frac{X - x_0}{\beta} \tag{4.43}$$

it is observed that Z = X if x_0 = 0 and β = 1, hence the **standardised exponential pdf** becomes:

$$f_Z(z) = \exp(-z) \tag{4.44}$$

and the **standardised exponential cdf** is given by:

$$F_Z(z) = 1 - \exp(-z) \tag{4.45}$$

Replacing Z in (4.45) by (4.43) equation (4.42) is seen to be obtained, and differentiating the cdf with respect to X gives pdf (4.41).

Moment related distribution parameters

The moment related parameters are given by:

$$\left.\begin{aligned} \mu_X &= x_0 + \beta \\ \sigma_X^2 &= \beta^2 \\ \gamma_{1,X} &= 2 \end{aligned}\right\} \tag{4.46}$$

It is observed that the distribution parameter x_0 is a **location** parameter as it affects only the first moment of the distribution. The parameter β is a **scale** parameter as it scales variate X. The skewness of the distribution is fixed. The distribution is shown in Figure 4.13.

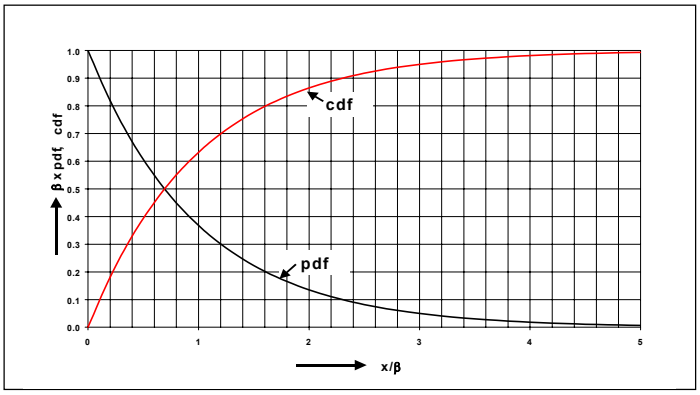

Figure 4.13:
Exponential distribution as function of the reduced variate (x-x_0)

From (4.46) it follows for the mean, variance and skewness of the standardised gamma function (x_0 = 0, β = 1) respectively 1, 1 and 2.

Distribution parameters expressed in moment related parameters

From (4.46) it follows for the distribution parameters as function of the moments:

$$\beta = \sigma_X \tag{4.47}$$

$$x_0 = \mu_X - \sigma_X \tag{4.48}$$

If $x_0 = 0$ the distribution reduces to 1-parameter exponential distribution **E-1**. Then the mean and the standard deviation are seen to be identical. Note also that with $x_0 = 0$ and $\lambda = 1/\beta$ substituted in (4.42) equation (4.11) is obtained.

Quantiles

The values of X and Z for which $F_X(x) = F_Z(z)$ are related by (4.43). Using the inverse the quantiles x_T are obtained from the reduced variate z_T for a specified return period T:

$$x_T = x_0 + \beta z_T \tag{4.49}$$

The quantile x_T can also directly be obtained from the first two moments and T:

$$x_T = \mu_X + \sigma_X(\ln(T) - 1) \tag{4.50}$$

Example 4.8: Exponential distribution

A variate X is exponentially distributed with mean 50 and standard deviation 20. Determine:

- the value of X, which corresponds with a non-exceedance probability of 0.95.
- the probability that $50 \le X \le 75$.

Note that since $\mu_X \neq \sigma_X$ the exponential distribution is E-2.The non-exceedance probability implies an exceedance probability of 1 – 0.95 = 0.05, hence the return period T is 1/0.05 = 20. From (4.50) the variate value for this return period becomes:

$X_T = 50 + 20 \times \{\ln(20) - 1\} = 50 + 20 \times (3.0 - 1) = 90$

To solve the second problem equation (4.42) is used, which requires the parameters x_0 and β to be available. From (4.47) one gets $\beta = \sigma_X = 20$ and from (4.48) $x_0 = \mu_X - \beta = 30$, hence:

$$P\{50 \le X \le 75\} = F_X(75) - F_X(50) =$$
$$= 1 - \exp(-\frac{75-30}{20}) - (1 - \exp(-\frac{50-30}{20}) = 0.895 - 0.632 = 0.263$$

Application in hydrology

The exponential distribution finds wide application. In engineering one applies the distribution to model time to failure, inter-arrival time, etc. In hydrology the distribution is a.o. applied to model time between flood peaks exceeding a threshold value. Furthermore, the distribution models a process, where the outcomes are independent of past occurrences, i.e. the process is **memory-less**.

4.5.2 Gamma distribution

Definition

The distribution of the sum of k exponentially distributed random variables each with parameter β (equation (4.41) with $x_0 = 0$) results in a gamma distribution with parameter k and β. The gamma distribution describes the waiting time till the k^{th} exceedance and is readily derived from the Poisson distribution (like the exponential) by multiplying the probability of having (k-1) arrivals till t, described by equation (4.7), and the arrival rate

($\lambda=1/\beta$) at t, leading to the Erlang distribution. Since k does not need to be an integer it is replaced by the positive real γ, and a gamma distribution with two parameters γ and β is obtained, shortly denoted by G-2.

Probability density and distribution function

The **gamma** pdf has the following form:

$$f_X(x) = \frac{\left(\frac{x}{\beta}\right)^{\gamma-1} \exp(-\left(\frac{x}{\beta}\right))}{\beta\Gamma(\gamma)} \text{ with : } x > 0; \beta > 0 ; \gamma > 0 \tag{4.51}$$

and the cdf reads:

$$F_X(x) = \frac{1}{\beta\Gamma(\gamma)} \int_0^x \left(\frac{s}{\beta}\right)^{\gamma-1} \exp(-\left(\frac{s}{\beta}\right))ds \text{ for : } \beta > 0; \gamma > 0 \tag{4.52}$$

Standardised gamma distribution

Introducing the **reduced gamma variate** Z, defined by:

$$Z = \frac{X}{\beta} \tag{4.53}$$

it is observed that Z = X for β = 1 and the pdf and cdf of the **standardised gamma distribution** then read:

$$f_Z(z) = \frac{z^{\gamma-1}\exp(-z)}{\Gamma(\gamma)} \tag{4.54}$$

$$F_Z(z) = \frac{1}{\Gamma(\gamma)} \int_0^z t^{\gamma-1}\exp(-t)dt \tag{4.55}$$

Note that by substituting (4.53) in (4.55) and with dx = β dz equation (4.52) is obtained, and by differentiating the cdf with respect to X the pdf equation (4.51) follows.

Gamma function

Equation (4.55) is called the **incomplete gamma function ratio**. The **complete** (standard) gamma function $\Gamma(\gamma)$, needed to get area = 1 under the pdf curve, is defined by:

$$\Gamma(\gamma) = \int_0^\infty t^{\gamma-1}\exp(-t)dt \tag{4.56}$$

The **gamma function** provides a continuous alternative for discrete factorials. The function has the following properties:

$$\Gamma(n+1) = n! \tag{4.57}$$

And hence:

$$\Gamma(n+1) = n\Gamma(n) \text{ for : } n = 0,1,2,... \text{ with : } 0! = 1 \tag{4.58}$$

Furthermore:

$$\begin{aligned}&\Gamma(0)=\infty\\&\Gamma(1/2)=\sqrt{\pi}\\&\Gamma(1)=\Gamma(2)=1\\&0.88560\le\Gamma(y)\le 1 \text{ for : } 1\le\gamma\le 2\end{aligned} \qquad (4.59)$$

The gamma function is tabulated for values of γ: $1 \le \gamma \le 2$. In HYMOS the complete gamma function is computed in two steps:

- first γ is reduced to a value between 1 and 2 using the recursive equation (4.58):

$\Gamma(\gamma - 1) = \Gamma(\gamma)/\gamma$ for $\gamma < 1$ or: $\Gamma(\gamma + 1) = \gamma\Gamma(\gamma)$ for $\gamma > 2$, and then

- secondly, a third order interpolation procedure is used to obtain a value from the basic gamma function table.

Example 4.9 Gamma function

Derive the gamma function values for γ = 3.2 and 0.6.
Procedure:

γ = 3.2, then $\Gamma(3.2) = 2.2\Gamma(2.2) = 1.2x2.2\Gamma(1.2) = 1.2x2.2x0.9182 = 2.424$
γ = 0.6, then $\Gamma(0.6) = \Gamma(1.6)/0.6 = 0.8935/0.6 = 1.489$

Note that the values for $\Gamma(1.2)$ and $\Gamma(1.6)$ are obtained from the basic gamma function table.
The computational procedure for the **incomplete** gamma function as used in HYMOS is presented in Annex A4.3 and A4.4 for its inverse.

Moment related parameters of the distribution

The mean, mode, variance, skewness and kurtosis of the gamma distribution read:

$$\left.\begin{aligned}&\mu_X=\beta\gamma\\&m_X=\beta(\gamma-1)\\&\sigma_X^2=\beta^2\gamma\end{aligned}\right\} \qquad (4.60a)$$

$$\left.\begin{aligned}&\gamma_{1,X}=\frac{2}{\sqrt{\gamma}}\\&\gamma_{2,X}=\frac{3(\gamma+2)}{\gamma}\end{aligned}\right\} \qquad (4.60b)$$

From (4.53) it is observed that β is a **scale** parameter and from (4.60b) γ is a **shape** parameter. This is also seen from Figures 4.14 to 4.16. Comparison of (4.60a) with (4.46) with $x_0 = 0$ shows that the mean and the variance of the gamma distribution is indeed γ-times the mean and the variance of the exponential distribution. This supports the statement that the gamma distribution is the distribution of the sum of γ exponentially distributed random variables. Note that for large γ the skewness tends to zero and kurtosis to 3 and hence the gamma distribution approaches the normal distribution. Note that the mode $m_X > 0$ for $\gamma > 1$ and the distribution is single peaked. If $\gamma \le 1$ the pdf has a reversed J-shape.

From (4.60a) it is also observed that with $\beta = 1$ the mean and the variance of the standardised gamma distribution are both equal to γ; the skewness and kurtosis are as in (4.60b).

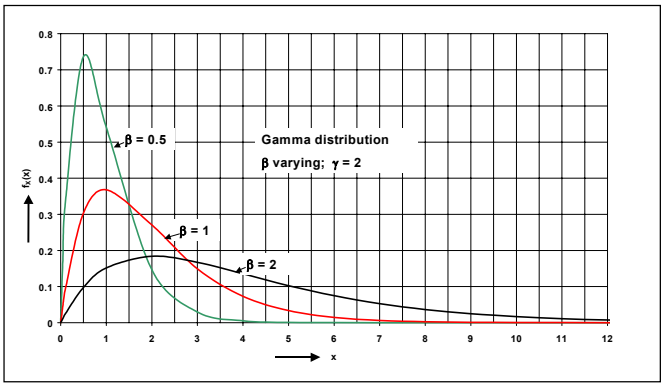

Figure 4.14:
Gamma distribution effect of scale parameter β

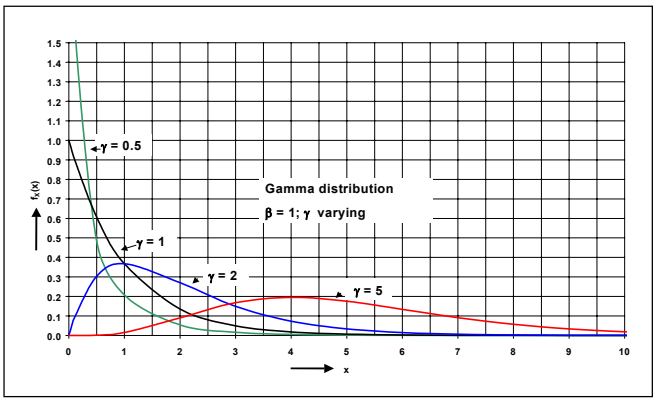

Figure 4.15:
Gamma distribution effect of shape parameter γ

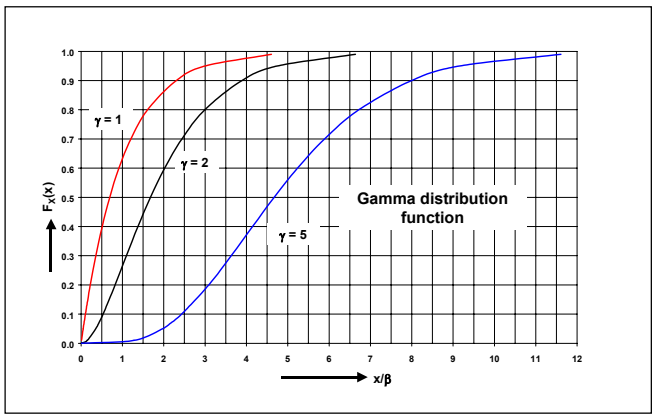

Figure 4.16:
Gamma cdf's

Distribution parameters expressed in moment related parameters

From (4.60a) it follows for the gamma parameters β and γ:

$$\beta = \frac{\sigma_X^2}{\mu_X} \tag{4.61}$$

$$\gamma = \left(\frac{\mu_X}{\sigma_X}\right)^2 = \frac{1}{C_{v,X}^2} \tag{4.62}$$

Hence, by the mean and the standard deviation the distribution parameters are fully determined. From a comparison of (4.62) with (4.60b) it is observed that for the gamma

distribution there is a fixed relation between the coefficient of variation and the skewness. It follows:

$$\gamma_{1,X} = 2C_{v,X} \tag{4.63}$$

It implies that from a simple comparison of the coefficient of variation with the skewness a first impression can be obtained about the suitability of the 2-parameter gamma distribution to model the observed frequency distribution. As will be shown in the next sub-section more flexibility is obtained by adding a location parameter to the distribution.

Quantiles of the gamma distribution

The quantiles x_T of the gamma distribution are derived from the inverse of the standard incomplete gamma function and the reduced variate z_T:

$$x_T = \beta z_T \tag{4.64}$$

The required parameters γ for the standard incomplete gamma function and β to transform the standardised variate z_T into x_T can be obtained from equations (4.61) and (4.62) or some other parameter estimation method.

4.5.3 Chi-squared and gamma distribution

Probability density and cumulative distribution function

By putting $\beta = 2$ and $\gamma = \nu/2$ the gamma distribution becomes the Chi-squared distribution:

$$f_X(x) = \frac{1}{2\Gamma(\nu/2)}\left(\frac{x}{2}\right)^{\nu/2-1} \exp(-\frac{x}{2}) \text{ for : } x \geq 0, \nu > 0 \tag{4.65}$$

$$F_X(x) = \frac{1}{2\Gamma(\nu/2)}\int_0^x \left(\frac{s}{2}\right)^{\nu/2-1} \exp(-\frac{s}{2})ds \tag{4.66}$$

The parameter ν is the number of degrees of freedom. The chi-square distribution is the distribution of the sum of ν squared normally distributed random variables N(0, 1) and find wide application in variance testing and goodness of fit testing of observed to theoretical distributions. It also follows, that the sum of 2 squared standard normal variables has an exponential distribution.

4.5.4 Pearson type 3 distribution

Probability density and cumulative distribution function

By introducing a **location** parameter x_0 in the gamma distribution, discussed in the previous sub-section, a **Pearson type 3** distribution is obtained, shortly denoted by **P-3**. This distribution is sometimes also called a **3-parameter gamma** distribution or **G-3**. Its pdf has the following form:

$$f_X(x) = \frac{\left(\frac{x-x_0}{\beta}\right)^{\gamma-1} \exp\left(-\left(\frac{x-x_0}{\beta}\right)\right)}{\beta\Gamma(\gamma)} \text{ with : } x > x_0 \text{ ; } \beta > 0 \text{ ; } \gamma > 0 \tag{4.67}$$

and the cdf reads:

$$F_X(x) = \frac{1}{\beta\Gamma(\gamma)}\int_{x_0}^{x}\left(\frac{s-x_0}{\beta}\right)^{\gamma-1} \exp\left(-\left(\frac{s-x_0}{\beta}\right)\right)ds \text{ for : } \beta > 0 \text{ ; } \gamma > 0$$

(4.68)

The reduced Pearson Type 3 variate Z, is defined by:

$$Z = \frac{X - x_0}{\beta} \tag{4.69}$$

It is observed that Z = X for x_0 = 0 and β = 1. Introducing this into (4.67) and (4.68) leads to the standardised gamma distributions presented in equations (4.54) and (4.55).

Moment related parameters of the distribution

The mean, mode, variance, skewness and kurtosis of the P-3 distribution read:

$$\left.\begin{aligned} \mu_X &= x_0 + \beta\gamma \\ m_X &= x_0 + \beta(\gamma - 1) \\ \sigma_X^2 &= \beta^2\gamma \end{aligned}\right\} \tag{4.70a}$$

$$\left.\begin{aligned} \gamma_{1,X} &= \frac{2}{\sqrt{\gamma}} \\ \gamma_{2,X} &= \frac{3(\gamma + 2)}{\gamma} \end{aligned}\right\} \tag{4.70b}$$

It is observed that x_0 is a **location** parameter as it affects only the first moment of the distribution about the origin. This is also seen from Figures 4.17. As for the (2-parameter) gamma distribution β is a **scale** parameter and γ is a **shape** parameter. Also, for large γ the distribution becomes normal. The mode of the distribution is at $x_0+\beta(\gamma-1)$, for $\gamma > 1$ and the distribution is unimodal. For $\gamma \leq 1$ the distribution is J-shaped similar to the gamma distribution, with its maximum at x_0.

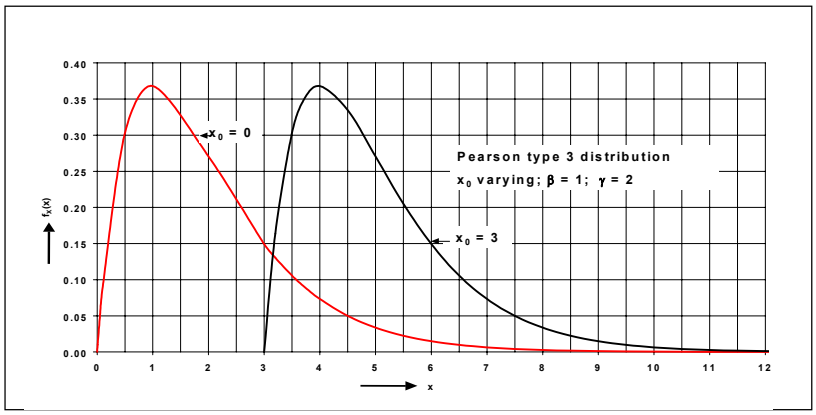

Figure 4.17: Pearson Type 3 distribution effect of location parameter x_0

Distribution parameters expressed in moment related parameters

The parameters of the Pearson Type 3 distribution can be expressed in the mean, standard deviation and skewness as follows:

$$\gamma = \left(\frac{2}{\gamma_{1,X}}\right)^2 \tag{4.71}$$

$$\beta = \frac{1}{2}\sigma_X\gamma_{1,X}$$

(4.72)

$$x_0 = \mu_X - 2\frac{\sigma_X}{\gamma_{1,X}} \tag{4.73}$$

From the last expression it is observed that:

$$\gamma_{1,X} = 2\left(\frac{\sigma_X}{\mu_X - x_0}\right) \tag{4.74}$$

The term within brackets can be seen as an adjusted coefficient of variation, and then the similarity with Equation (4.63) is observed.

Moment generating function

The moments of the distribution are easily obtained from the moment generating function:

$$G(s) = E[\exp((sx)] = \int_{x_0}^{\infty} \exp(sx)\frac{\left(\frac{x-x_0}{\beta}\right)^{\gamma-1}\exp\left(-\left(\frac{x-x_0}{\beta}\right)\right)}{\beta\Gamma(\gamma)}dx \tag{4.75}$$

Or introducing the reduced variate $Z = (x-x_0)/\beta$, and $dx = \beta\, dz$:

$$G(s) = \exp(sx_0)\int_0^{\infty}\frac{z^{\gamma-1}\exp(-z(1-s\beta)}{\Gamma(\gamma)}dz$$

Introducing further: $u = z(1-s\beta)$, or $z = u/(1-s\beta)$ and $dz = 1/(1-s\beta)du$, it follows:

$$\begin{aligned} G(s) &= \exp(sx_0)(1-s\beta)^{-\gamma}\int_0^{\infty}\frac{u^{\gamma-1}\exp(-u)}{\Gamma(\gamma)}du = \\ &= \exp(sx_0)(1-s\beta)^{-\gamma} \end{aligned} \tag{4.76}$$

By taking the derivatives of G(s) with respect to s at s = 0 the moments about the origin can be obtained:

$$\frac{dG(0)}{ds} = \mu_1^{'} = \exp(sx_0)\{x_0(1-s\beta)^{-\gamma} + \beta\gamma(1-s\beta)^{-(\gamma+1)}\}\Big|_{s=0}$$

$$\text{so}: \quad \mu_1^{'} = \mu_X = x_0 + \beta\gamma$$

Since for the computation of the central moments the location parameter is of no importance, the moment generating function can be simplified with $x_0 = 0$ to:

$$\left.\begin{aligned} &G(s) = (1-s\beta)^{-\lambda} \\ &\text{hence}: \\ &\frac{dG(0)}{ds} = \beta\gamma(1-s\beta)^{-(\gamma+1)}\Big|_{s=0} \rightarrow \mu_1^{'}\Big|_{x_0=0} = \beta\gamma \\ &\frac{d^2G(0)}{ds^2} = \beta^2\gamma(\gamma+1)(1-s\beta)^{-(\gamma+2)}\Big|_{s=0} \rightarrow \mu_2^{'} = \beta^2\gamma(\gamma+1) \\ &\text{etc.} \end{aligned}\right\} \tag{4.78}$$

Using equation (3.30) the central moments can be derived from the above moments about the origin.

Quantiles

The quantile x_T of the gamma distribution follows from the inverse of the standard incomplete gamma function z_T and (4.67):

$$x_T = x_0 + \beta z_T \tag{4.79}$$

Example 4.10: Gamma distribution

The mean, standard deviation and skewness of a P-3 variate are respectively 50, 20 and 1.2. Required is the variate value at a return period of 100.

First, the parameters of the P-3 distribution are determined from (4.71) – (4.73). It follows:

$$\gamma = \left(\frac{2}{\gamma_{1,X}}\right)^2 = \left(\frac{2}{1.2}\right)^2 = 2.78$$

$$\beta = \frac{\sigma_X}{\sqrt{\gamma}} = \frac{\gamma_{1,X}\sigma_X}{2} = \frac{1.2x20}{2} = 12$$

$$x_0 = \mu_X - \beta\gamma = 50 - 12x2.78 = 16.67$$

From the standard incomplete gamma function with γ = 2.78 it follows that $z_T = z_{100}$ = 8.03. Then from (4.77) it follows for $x_T = x_{100}$:

$$x_T = x_0 + \beta z_T = 16.67 + 12x8.03 = 113$$

Note that the standardised gamma variate can also be obtained from the tables of the chi-squared distribution for distinct non-exceedance probabilities. Since $\gamma = \nu/2$ it follows $\nu = 2\gamma$ = 2 x 2.78 = 5.56. From the χ^2 - tables one gets for T = 100 or p = 0.99 a χ^2 – value by interpolation between ν = 5 and ν = 6 of 16.052. For the chi-squared distribution β = 2, so: $\chi_T^2 = \beta z_T$ or z_T or $z_T = \chi_T^2/\beta$ = 16.052/2 = 8.03. The values can of course also directly be obtained via the "Statistical Tables" option in HYMOS under "Analysis".

Related distributions

For specific choices of the parameters x_0, β and γ, a number of distribution functions are included in the Pearson Type 3 or 3-parameter gamma distribution, see Tables 4.2 and 4.3.

The moment related parameters of these distributions are summarised in Table 4.3. By considering the logarithm of the variate or by raising the reduced variate Z of (4.69) to a power k further distributions like Weibull and Rayleigh distributions can be defined as presented in Sub-section 4.1 Those are discussed in the next sub-sections.

Pearson Type 3 or 3-parameter gamma (x_0,β,γ)	γ = 1: exponential	x_0 = 0: 1-par. exponential
		$x_0 \neq 0$: 2-par. exponential
	x_0 = 0: gamma	β = 1: 1-par gamma
		$\beta \neq 1$: 2-par gamma
		β = 2, $\gamma = \nu/2$: chi-squared

Table 4.2: Summary of related distributions

distribution	mean	mode	Variance	Skewness	kurtosis	Standardised variate z
1-par. exponential	β	-	β^2	2	9	$z=x/\beta$
2-par. exponential	$x_0+\beta$	-	β^2	2	9	$z=(x-x_0)/\beta$
1-par. gamma	γ	γ-1, γ>1	γ	$2/\sqrt{\gamma}$	$3(\gamma+2)/\gamma$	z=x
2-par. gamma	$\beta\gamma$	$\beta(\gamma-1)$	$\beta^2\gamma$	$2/\sqrt{\gamma}$	$3(\gamma+2)/\gamma$	$z=x/\beta$
3-par. Gamma or P-3	$x_0+\beta\gamma$	$x_0+\beta(\gamma-1)$	$\beta^2\gamma$	$2/\sqrt{\gamma}$	$3(\gamma+2)/\gamma$	$z=(x-x_0)/\beta$
Chi-squared	ν	ν-2, ν>2	2ν	$2^{3/2}/\sqrt{\nu}$	$3(\nu+4)/\nu$	z=x/2

Table 4.3: Moment related parameters of the exponential and gamma family of distributions

4.5.5 Log-Pearson Type 3 distribution

Probability density function

When $Y = \ln(X - x_0)$ follows a Pearson Type 3 distribution then $(X - x_0)$ is log-Pearson Type 3 distributed. Its pdf is given by:

$$f_X(x) = \frac{1}{\beta(x-x_0)\Gamma(\gamma)}\left(\frac{\ln(x-x_0)-y_0}{\beta}\right)^{\gamma-1}\exp\left(-\left(\frac{\ln(x-x_0)-y_0}{\beta}\right)\right) \text{ for : } \ln(x-x_0) > y_0 \quad (4.82)$$

The log-Pearson Type 3distribution finds application in hydrology particularly for strongly positively skewed annual flood peaks. The skewness is reduced by a logarithmic transformation, to arrive at a Pearson type III distribution. In the USA the log-Pearson type III is the standard for modelling annual maximum floods (Water Resources Council, 1967). All relations presented in the previous sub-section are valid for $\ln(X-x_0)$.

Quantiles of LP-3

The quantiles x_T of the LP-3 distribution are obtained from the inverse of the standard incomplete gamma function leading to z_T and (4.81):

$$x_T = x_0 + \exp(y_0 + \beta z_T) \quad (4.81)$$

4.5.6 Weibull distribution

Probability density and cumulative distribution function

With γ = 1 equation (4.55) reduces to:

$$F(z) = \int_0^z \exp(-s)ds \text{ with } :z = \left(\frac{x-x_0}{\beta}\right)^k \text{ so : } dz = \frac{k}{\beta}\left(\frac{x-x_0}{\beta}\right)^{k-1} dx$$

it follows for the pdf and cdf of the Weibull distribution:

$$f_X(x) = \frac{dF_X(x)}{dx} = \frac{k}{\beta}\left(\frac{x-x_0}{\beta}\right)^{k-1}\exp\left(-\left(\frac{x-x_0}{\beta}\right)^k\right) \text{ for : } x \ge x_0, k > 0, \beta > 0 \quad (4.82)$$

$$F_X(x) = 1 - \exp\left(-\left(\frac{x-x_0}{\beta}\right)^k\right) \quad (4.83)$$

Note that for k = 1 the Weibull distribution reduces to an exponential distribution.

Moment related parameters of the distribution

The mean, mode, variance and skewness of the Weibull distribution read:

$$\left.\begin{aligned} \mu_X &= x_0 + \beta\Gamma\left(1+\frac{1}{k}\right) \\ M_X &= x_0 + \beta(\ln 2)^{1/k} \\ m_X &= x_0 + \beta\left(\frac{k-1}{k}\right)^{1/k} \end{aligned}\right\} \qquad (4.84a)$$

$$\sigma_X^2 = \beta^2\left\{\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right\}$$

$$\gamma_{1,X} = \frac{\Gamma(1+\frac{3}{k}) - 3\Gamma(1+\frac{2}{k})\Gamma(1+\frac{1}{k}) + 2\Gamma^3(1+\frac{1}{k})}{\left(\Gamma(1+\frac{2}{k}) - \Gamma^2(1+\frac{1}{k})\right)^{3/2}} \qquad (4.84b)$$

The distribution is seen to have 3 parameters: x_0 is a **location** parameter, β a **scale** parameter and k is a **shape** parameter. For k > 1 the pdf is seen to be unimodal, see also Figure 4.19.

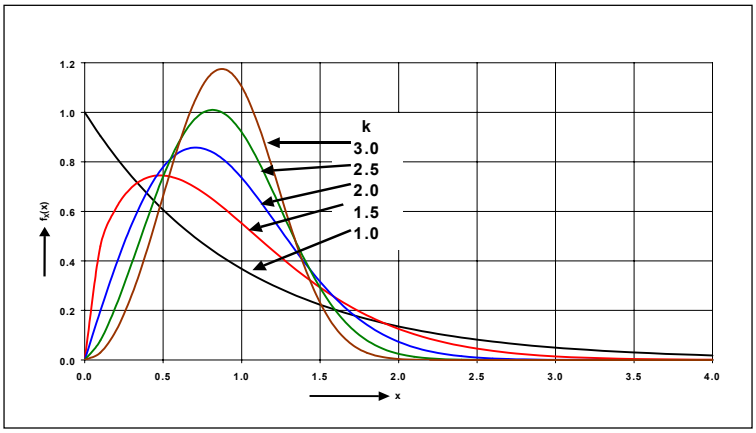

Figure 4.19a:

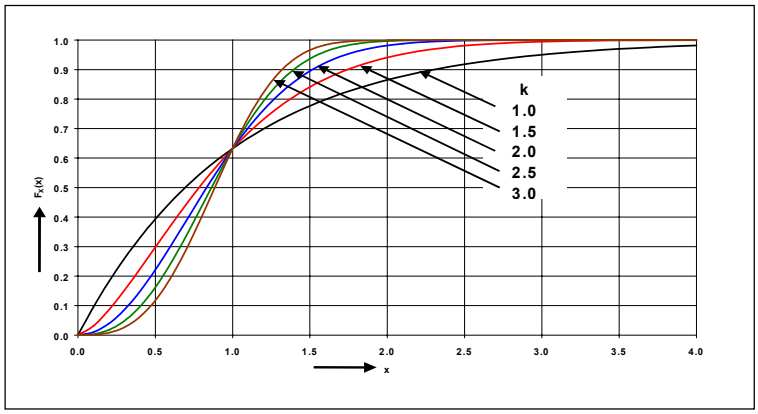

Figure 4.19b:

Figure 4.19a and 4.19b: Weibull distribution for various values of k(x_0 = 0 and β = 1)

The expression for the skewness as a function of k is rather complicated and has therefore been visualised in Figure 4.20. From the Figure it is observed that for k < 1 the skewness

increases rapidly to very high values. In practice the region 1< k < 3 is mostly of interest. Note that for k > 3.5 the skewness becomes slightly negative.

Note also that above expressions for the mean, variance and skewness can easily be derived from the moment generating function. For x_0 = 0 the r^{th} moment about the origin becomes:

$$\mu_r' = \beta^r \Gamma(1 + \frac{r}{k}) \tag{4.85}$$

Subsequently, equation (3.30) is used to obtain the central moments. For the mean x_0 has to be added.

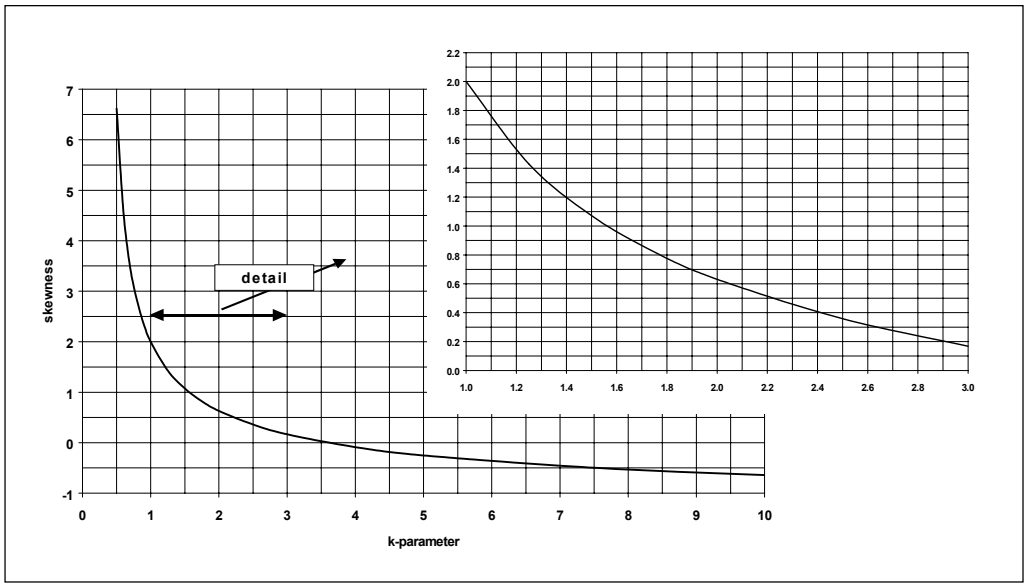

Figure 4.20: Skewness of W-3 as function of k

Quantiles of W-3

From (4.83) the quantile of the Weibull distribution is easily determined. For a given return period T it follows for x_T:

$$x_T = x_0 + \beta(\ln T)^{1/k} \tag{4.86}$$

From (4.86) it is observed that for given x_0, β and T values x_T decreases with increasing k.

The Weibull distribution is often used to model the frequency distribution of wind speed and flow extremes (minimum and maximum). It is one of the asymptotic distributions of the general extreme value theory, to be discussed in the next sub-section.

4.5.7 Rayleigh distribution

Probability density and cumulative distribution function

From the Weibull distribution with k = 2 the Rayleigh distribution is obtained. Its pdf and cdf read:

$$f_X(x) = \frac{2}{\beta}\left(\frac{x - x_0}{\beta}\right)\exp\left(-\left(\frac{x - x_0}{\beta}\right)^2\right) \tag{4.87}$$

$$F_X(x) = 1 - \exp\left(-\left(\frac{x - x_0}{\beta} \right)^2 \right) \tag{4.88}$$

Moment related parameters of the distribution

From (4.84) the mean, mode, variance and skewness are given by:

$$\left.\begin{aligned} \mu_X &= x_0 + \Gamma(1.5)\beta = x_0 + 0.88623\beta \\ m_X &= x_0 + \frac{1}{2}\sqrt{2}\beta = x_0 + 0.70711\beta \\ \sigma_X^2 &= (1 - \Gamma^2(1.5))\beta = 0.21460\beta^2 \\ \gamma_{1,X} &= \frac{\Gamma(1.5)\left\{2\Gamma^2(1.5) - 1.5\right\}}{\left\{1 - \Gamma^2(1.5)\right\}^{3/2}} = 0.631 \end{aligned}\right\} \tag{4.89}$$

The distribution is seen to have **location** parameter x_0 and a **scale** parameter β. The skewness of the distribution is fixed. The pdf and cdf of the Rayleigh distribution are shown in Figure 4.21.

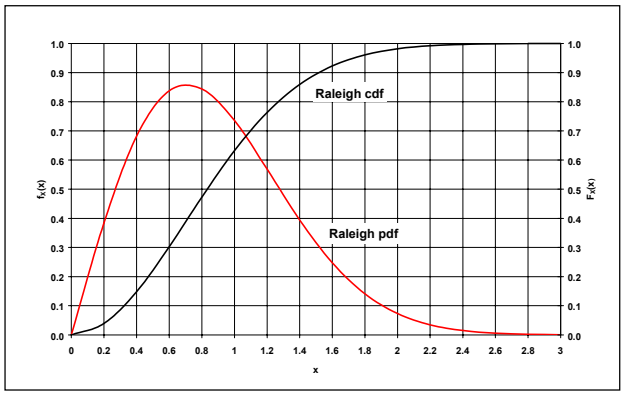

Figure 4.21:
Rayleigh distribution

The distribution parameters are easily related to the mean and standard deviation of the Rayleigh variate X:

$$\beta = 2.15866\sigma_X \tag{4.90}$$

$$x_0 = \mu_X - 1.91307\sigma_X \tag{4.91}$$

Quantiles of* R-*2

The quantiles x_T of the Rayleigh distribution for a return period T follow from (4.88):

$$x_T = x_0 + \beta\sqrt{\ln T} \tag{4.92}$$

The Rayleigh distribution is suitable to model frequency distributions of wind speed and of annual flood peaks in particular.

4.6 Extreme value distributions

4.6.1 Introduction

A number of distribution functions are available specially suited to model frequency distributions of extreme values, i.e. either largest values or smallest values. These can be divided in two groups:

General extreme value distributions GEV, or EV-1, EV-2 and EV-3, and

1. Generalised Pareto distributions, also with 3 types, P-1, P-2 and P-3.

The GEV distributions and the generalised Pareto distributions are related. The first group is generally applicable to annual maximum or annual minimum series, whereas the Pareto distributions are often used to model exceedance series, i.e. peaks exceeding a threshold value. Though any of the distributions may be applied to any of the series of extremes. There is however a distinct difference in the interpretation of the return period between extremes in a fixed interval and extremes exceeding a threshold, though both methods are related.

It is noted that instead of the extreme value distributions also the distributions dealt with in the previous sections may be applied to model the distribution to extremes.

Note further that statistical distributions are generally used far beyond the observed frequency range. It is noted, though, that the use of statistical distributions for extrapolation purposes is strongly limited by physical features and limitations in sources and basins, neither included in the distribution or in the data used to fit the distribution. The main difficulty is with the assumption of the independent identically distributed random variable ('iidrv') and the invariability of the distribution with time. In this respect, you are strongly advised to read the paper by V. Klemes entitled: 'Tall tales about tails of hydrological distributions' in Journal of Hydrologic Engineering, Vol 5, No 3, July 2000, pages 227 – 239. As an example consider the routing of a design storm through a channel reach. The design storms for different return periods are determined using the procedures proposed by NERC (1975). The design storms are routed through a channel reach with an inbank capacity of 350 m^3/s. Beyond that discharge level part of the flow is transferred through the floodplain. The exceedance of the inbank capacity occurs on average once in 30 years. Two types of flood plains are considered: a narrow one and a wide one. The effect of the two types of flood plains on the behaviour of the distribution function of the flood peaks, observed at the downstream end of the reach, is shown in Figure 4.22.

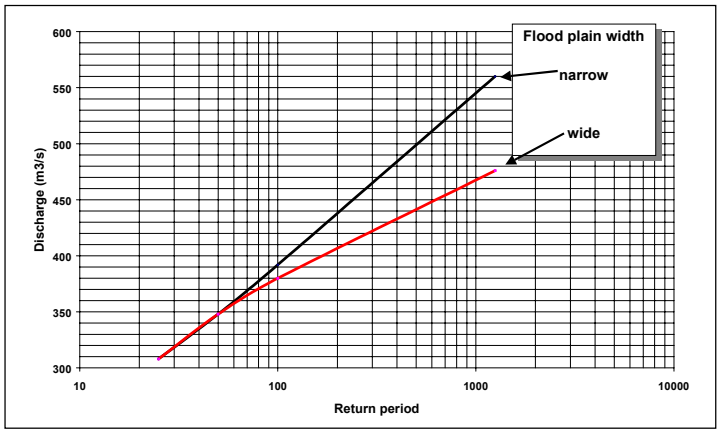

Figure 4.22:
Extreme value distribution of routed design storms

From Figure 4.22 it is observed that the frequency distribution is strongly affected by physical features of the river, which affect discharges of various magnitudes differently. It implies that data points gathered for the more frequent extreme events may include no information for the rare extreme events. Hence the validity of extrapolation beyond the measured range, no matter how scientific and/or complex the mathematical expressions may be, remains highly questionable. It should always be verified whether physical limitations and behaviour under very wet or very dry conditions may affect the extreme events. Blind application of extreme value distributions is always wrong.

The use of confidence bands about the frequency distribution will not help you much, as those are based on the assumption that the used distribution is applicable to the considered case. If the distribution is not applicable, the confidence limits will give a completely false picture of the uncertainty in the extreme value for a particular return period. Also, the use of goodness of fit tests will not help you in this respect and may lead you to an unjustified believe in the applicability of the distribution.

4.6.2 General extreme value distributions

The general extreme value distributions are applicable to series with a fixed interval like annual maximum or annual minimum series; i.e. one value per interval. Consider the extreme values (largest X_{max} and smallest X_{min}) of a sample of size n. Hence, $X_{max} = \max(X_1, X_2, \ldots, X_n)$ and let the X_i's be **independent** and **identically distributed**, then:

$$F_{Xmax}(x) = P\left(X_1 \le x \cap X_2 \le x \cap \ldots\ldots \cap X_n \le x\right) = \prod_{i=1}^{n} F_{X_i}(x) = \left(F_X(x)\right)^n \tag{4.95}$$

Note that the third expression stems from the independence of the X_i's, whereas the fourth expression is due to the identical distribution of the X_i's. The pdf of X_{max} reads:

$$f_{Xmax}(x) = n\left(F_X(x)\right)^{n-1} f_X(x) \tag{4.96}$$

Similarly for $X_{min} = \min(X_1, X_2, \ldots, X_n)$ it follows under the same assumptions of independence and identical distribution:

$$F_{Xmin}(x) = 1 - P\left(X_1 > x \cap X_2 > x \cap \ldots\ldots \cap X_n > x\right) = 1 - \prod_{i=1}^{n}\left(1 - F_{X_i}(x)\right) = 1 - \left(1 - F_X(x)\right)^n \tag{4.97}$$

and the pdf of X_{min}:

$$f_{Xmin}(x) = n\left(1 - F_X(x)\right)^{n-1} f_X(x) \tag{4.98}$$

Above expressions for X_{max} and X_{min} show that their distributions depend on sample size and the parent distribution from which the sample is taken. However, it can be shown, that full details about the parent distribution are not required to arrive at the distribution of extremes. For large n and limited assumptions about the parent distributions three types of asymptotic distributions for extreme values have been developed:

1. **Type I**: parent distribution is unbounded in the direction of the extreme and all moments of the distribution exist (exponential type distributions), like
   - Largest: normal, lognormal, exponential, gamma, Weibull
   - Smallest: normal
2. **Type II**: parent distribution is unbounded in the direction of the extreme but not all moments exist (Pareto type distributions):
   - Largest: Cauchy, Pareto, log-gamma, Student's t
   - Smallest: Cauchy distribution
3. **Type III**: parent distribution is bounded in the direction of the extreme (limited distributions):
   - Largest: beta
   - Smallest: beta, lognormal, gamma, exponential.

The above types of extreme value distributions are often indicated as Fisher-Tippett Type I, II and III distributions or shortly as EV-1, EV-2 and EV-3 respectively.

Asymptotic distributions for X_{max}

The distributions for X_{max} of the 3 distinguished types have the following forms:

- **Type I distribution, largest value,** for $-\infty < x < \infty$ and $\beta > 0$:

$$F_{Xmax}(x) = \exp(-\exp(-\left(\frac{x - x_0}{\beta}\right)) \tag{4.99}$$

- **Type II distribution, largest value,** for $x \geq x_0$, $k < 0$ and $\beta > 0$

$$F_{Xmax}(x) = \exp(-\left(\frac{x - x_0}{\beta}\right)^{1/k}) \tag{4.100}$$

- **Type III distribution, largest value,** for $x \leq x_0$, $k > 0$ and $\beta > 0$

$$F_{Xmax}(x) = \exp(-\left(-\frac{x - x_0}{\beta}\right)^{1/k}) \tag{4.101}$$

It is observed that the forms of the Type II and Type III distributions are similar, apart from sign differences and location of boundaries relative to the variable. All above asymptotic distributions for the largest value can be represented by the following general form of the extreme value distribution or shortly GEV distribution (Jenkinson, 1969):

$$F_{Xmax}(x) = \exp(-\left(1 - k\left(\frac{x - b}{a}\right)\right)^{1/k}) \tag{4.102}$$

Dependent on the sign of k the following cases are distinguished:

- k = 0: extreme value distribution Type I, EV-1
- k < 0: extreme value distribution Type II, EV-2
- k > 0: extreme value distribution Type III, EV-3

To arrive at the Type I distribution from (4.102) consider the Taylor series expansion of the argument of the exponential function in the limit for $k \rightarrow 0$:

$$\lim_{k\rightarrow 0}\left(1 - k\left(\frac{x - b}{a}\right)\right)^{1/k} = \exp\left(-\frac{x - b}{a}\right)$$

Hence, for k = 0 with $b = x_0$ and $a = \beta$ equation (4.99) is obtained from (4.102). Equivalently, with $b + a/k = x_0$ and $\pm a/k = \beta$ equations (4.100) and (4.101) for the Type II and Type III distributions follow from (4.102). The GEV-form is sometimes used in literature on extreme value distributions to describe the Type II and Type III distributions, like in the Flood Studies Report (NERC, 1975). The different type of distributions for X_{max} are presented in Figure 4.23. It is observed that there is an upper limit to X_{max} in case of EV-3.

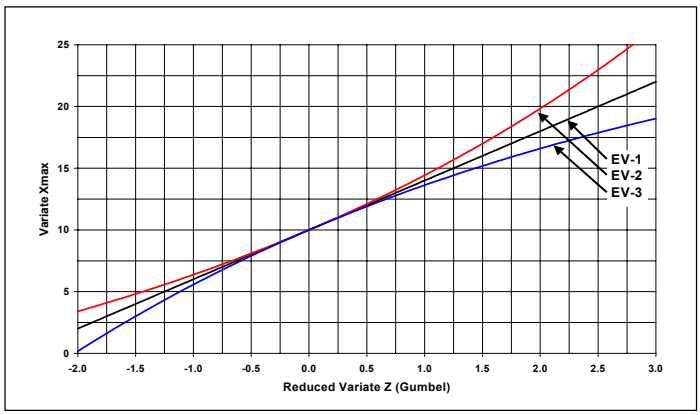

Figure 4.23: Presentation of EV-1, EV-2 and EV-3 as function of reduced EV-1 variate

As shown in Figure 4.24, there is a distinct difference in the skewness of the X_{max} series suitable to be modelled by one of the EV-distributions. EV-1 has a fixed skewness (= 1.14), whereas EV-2 has a skewness > 1.14 and EV-3 a skewness < 1.14. Hence, a simple investigation of the skewness of a series of X_{max} will give a first indication of the suitability of a distribution.

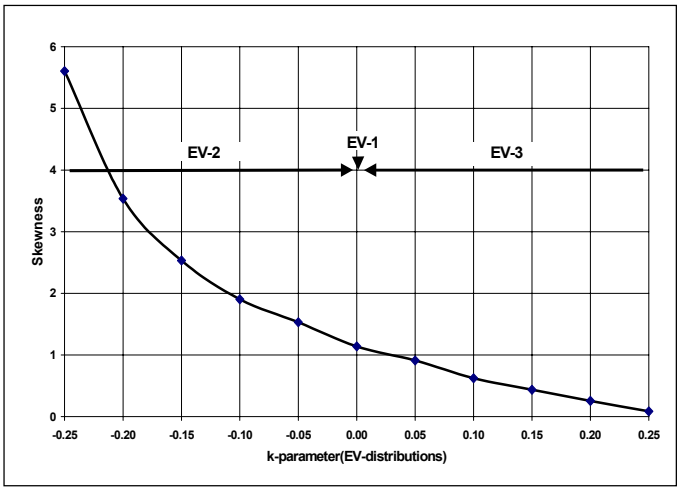

Figure 4.24:
Skewness as function of EV-parameter k

Asymptotic distributions for X_{min}

From the principle of symmetry (see e.g. Kottegoda and Rosso, 1997), the asymptotic distributions for the smallest value can be derived from the distribution of the largest value by **reversing the sign** and taking the **complementary probabilities**. Let X denote a variate with pdf $f_X(x)$ and X* a variate whose pdf is the mirror image of $f_X(x)$, it then follows: $f_X(x) = f_{X^*}(-x)$ and therefore: $1 - F_X(x) = F_{X^*}(-x)$. So for the distributions of X_{min} as a function of those of X_{max} it follows:

$$F_{Xmin}(x) = 1 - F_{Xmax}(-x) \tag{4.103}$$

Hence, the asymptotic distributions of X_{min} for the 3 distinguished types read:

- **Type I distribution, smallest value,** for $-\infty < x < \infty$ and $\beta > 0$:

$$F_{Xmin}(x) = 1 - \exp(-\exp\left(\frac{x - x_0}{\beta}\right) \tag{4.104}$$

- **Type II distribution, smallest value,** for $x \le x_0$, $k < 0$ and $\beta > 0$:

$$F_{Xmin}(x) = 1 - \exp\left(-\left(-\frac{x - x_0}{\beta}\right)^{1/k}\right) \tag{4.105}$$

- **Type III distribution, smallest value,** for $x \ge x_0$, $k > 0$ and $\beta > 0$

$$F_{Xmin}(x) = 1 - \exp\left(-\left(\frac{x - x_0}{\beta}\right)^{1/k}\right) \tag{4.106}$$

In hydrology, particularly Type I for largest value and Type III for smallest value are frequently used. In the next sub-sections all types are discussed.

4.6.3 Extreme value Type 1 or Gumbel distribution

EV-1 for largest value

The Extreme Value Type I distribution for the largest value was given by equation (4.99):

$$F_{X\max}(x) = \exp\left\{-\exp\left(-\left(\frac{x - x_0}{\beta}\right)\right)\right\} \text{ for : } -\infty < x < \infty \text{ and } \beta > 0 \qquad (4.99)$$

The pdf is obtained by differentiating (4.99) with respect to x and reads:

$$f_{X\max}(x) = \frac{1}{\beta}\exp\left\{-\left(\frac{x - x_0}{\beta}\right) - \exp\left(-\left(\frac{x - x_0}{\beta}\right)\right)\right\} \qquad (4.107)$$

In view of the form, equation (4.99) is called the **double exponential** distribution or in honour to its promoter the **Gumbel** distribution. Introducing the reduced or standardised variate Z, defined by:

$$Z = \frac{X_{\max} - x_0}{\beta} \qquad (4.108)$$

The standardised Gumbel distribution is obtained by observing that Z = X for x_0 = 0 and β = 1:

$$F_Z(z) = \exp(-\exp(-z)) \qquad (4.109)$$

$$f_Z(z) = \exp(-z - \exp(-z)) \qquad (4.110)$$

The standardised pdf and cdf are shown in Figure 4.25

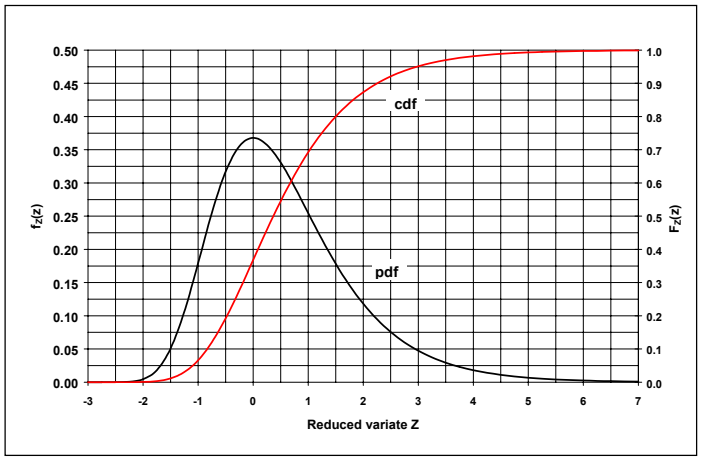

Figure 4.25:
Standardised Gumbel pdf and cdf

The moment related parameters of the distribution, the mean, median, mode, variance skewness and kurtosis are given by:

$$\left.\begin{aligned} \mu_{X\max} &= x_0 + \gamma_E\beta = x_0 + 0.5772\beta \\ M_{X\max} &= x_0 + \beta\ln(\ln 2) = x_0 + 0.3665\beta \\ m_{X\max} &= x_0 \end{aligned}\right\} \qquad (4.111a)$$

$$\left.\begin{aligned} \sigma^2_{X\max} &= \frac{\pi^2\beta^2}{6} \\ \gamma_{1,X\max} &\approx 1.1396 \\ \gamma_{2,X\max} &= 5.4 \end{aligned}\right\} \qquad (4.111b)$$

The constant γ_E = 0.577216 is called Euler's constant and can be read from mathematical tables. The parameter x_0 is seen to be a **location** parameter and β is a **scale** parameter. The skewness is fixed at 1.14 and the kurtosis is > 3, hence the pdf is more peaked than the normal distribution.

The moments of the distribution and its related parameters can be obtained from the moment generating function:

$$G_{Xmax}(s) = \exp(x_0 s)\Gamma(1-\beta s) \tag{4.112}$$

More easily the moment related parameters for the Gumbel distribution can be obtained from the cumulants κ_n of the distribution (see e.g. Abramowitz and Stegun, 1970):

$$\left.\begin{aligned} &\kappa_1 = x_{0+}\gamma_E\beta \\ &\kappa_n = \beta^n\Gamma(n)\varsigma(n) \\ &\text{where : } \varsigma(n) = \sum_{r=1}^{\infty}\frac{1}{r^n} \text{ specifically : } \varsigma(2) = \frac{\pi^2}{6};\ \varsigma(4) = \frac{\pi^4}{90} \end{aligned}\right\} \tag{4.113}$$

The function $\zeta(n)$ is the Riemann Zeta Function and is tabulated in mathematical tables. The relation between the cumulants and the moments are:

$$\kappa_1 = \mu_1';\ \kappa_2 = \mu_2;\kappa_3 = \mu_3;\ \kappa_4 = \mu_4 - 3\mu_2^2 \tag{4.114}$$

Hence:

$$\sigma^2 = \kappa_2 = \beta^2\Gamma(2)\varsigma(2) = \beta^2\frac{\pi^2}{6}$$

$$\gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}} = \frac{\beta^3\Gamma(3)\varsigma(3)}{\left(\beta^2\varsigma(2)\right)^{3/2}} = \frac{\Gamma(3)\varsigma(3)}{\left(\frac{\pi^2}{6}\right)^{3/2}} = \frac{2x1.20205}{2.10971} = 1.139541$$

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} + 3 = \frac{\beta^4\Gamma(4)\varsigma(4)}{\left(\beta^2\Gamma(2)\varsigma(2)\right)^2} + 3 = \frac{\Gamma(4)\frac{\pi^4}{90}}{\left(\Gamma(2)\frac{\pi^2}{6}\right)^2} + 3 = \frac{2x3x36}{90} + 3 = \frac{12}{5} + 3 = 5.4$$

Distribution parameters expressed in moment related parameters

From (4.111) the following relations between x_0, β and μ and σ are obtained:

$$\beta = \frac{\sqrt{6}}{\pi}\sigma \tag{4.115}$$

$$x_0 = \mu - \gamma_E\frac{\sqrt{6}}{\pi}\sigma = \mu - 0.45\sigma \tag{4.116}$$

Quantiles of EV-1 for X_{max}

The value for X_{max} for a specified return period T, $x_{max}(T)$, can be derived from (4.108) and (4.109):

$$x_{max}(T) = x_0 - \beta\ln\left(\ln\left(\frac{T}{T-1}\right)\right) = \mu_{Xmax} - \sigma_{Xmax}\frac{\sqrt{6}}{\pi}\left\{\gamma_E + \ln\left(\ln\left(\frac{T}{T-1}\right)\right)\right\} \tag{4.117}$$

In some textbooks the quantiles are determined with the aid of a frequency factor K(T):

$$x_{max}(T) = \mu_{Xmax} + K(T)\sigma_{Xmax} \tag{4.118}$$

Hence:

$$K(T) = -\frac{\sqrt{6}}{\pi}\left\{\gamma_E + \ln\left(\ln\left(\frac{T}{T-1}\right)\right)\right\} \tag{4.119}$$

Values for K(T) for selected return periods are presented in table below:

T	K(T)	T	K(T)
2	-0.1643	100	3.1367
5	0.7195	250	3.8535
10	1.3046	500	4.3947
25	2.0438	1000	4.9355
50	2.5923	1250	5.1096

From (4.118) it is observed that if to a given set of extremes some very low values are added the quantile for high return periods may increase!! This stems from the fact that though μ_{xmax} may reduce some what, σ_{xmax} will increase, since the overall variance increases. Because for large T, K(T) becomes large, it follows that $x_{max}(T)$ may be larger than before. This is a "lever" effect.

Application of EV-1 for largest value

The Gumbel distribution appears to be a suitable model for annual maximum rainfall and runoff in a number of cases, though many a times it does not apply. A first rapid indication about the applicability of the Gumbel distribution can be obtained from the skewness of the data set of maximum values. If this deviates substantially from 1.14, the distribution is not suitable to model the extremes.

EV-1 for smallest value

The cdf of the EV-1 distribution for the smallest value is given by (4.104):

$$F_{Xmin}(x) = 1 - \exp\left(-\exp\left(\frac{x - x_0}{\beta}\right)\right) \tag{4.104}$$

and the pdf then reads:

$$f_{Xmin}(x) = \frac{1}{\beta}\exp\left\{\left(\frac{x - x_0}{\beta}\right) - \exp\left(\frac{x - x_0}{\beta}\right)\right\} \tag{4.120}$$

Introducing the reduced variate Z defined by:

$$Z = \frac{X_{min} - x_0}{\beta} \tag{4.121}$$

then the standardised cdf and pdf read:

$$F_Z(z) = 1 - \exp(-\exp(z)) \tag{4.122}$$

$$f_Z(z) = \exp(z - \exp(z)) \tag{4.123}$$

The standardised distribution is shown in Figure 4.26. From this figure it is observed that the pdf for the smallest value is the mirror image of the pdf of the largest value around z = 0.

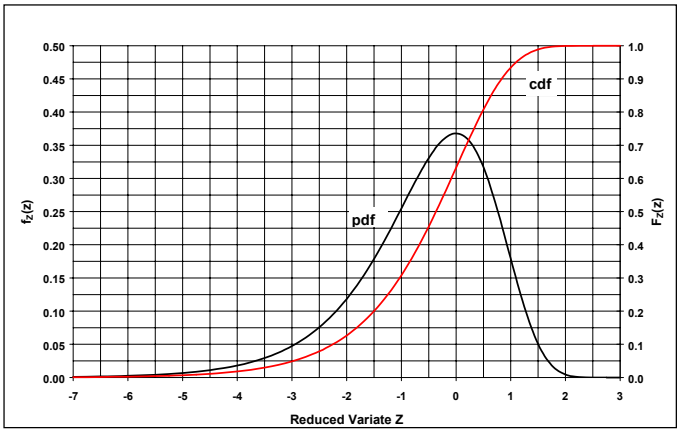

Figure 4.26:
Standardise EV-1 pdf and cdf for smallest value

The moment related parameters of the distribution, the mean, median, mode, variance skewness and kurtosis are given by:

$$\left.\begin{aligned} \mu_{X\min} &= x_0 - \gamma_E\beta = x_0 - 0.5772\beta \\ M_{X\min} &= x_0 - \beta\ln(\ln 2) = x_0 - 0.3665\beta \\ m_{X\min} &= x_0 \end{aligned}\right\} \qquad (4.124a)$$

$$\left.\begin{aligned} \sigma^2_{X\min} &= \frac{\pi^2\beta^2}{6} \\ \gamma_{1,X\min} &\approx -1.1396 \\ \gamma_{2,X\min} &= 5.4 \end{aligned}\right\} \qquad (4.124b)$$

Comparing these results with (4.111) it is observed that, apart from some changes in sign, the components of the above formulae are similar. For the distribution parameters expressed in the moment related parameters it now follows:

$$\beta = \frac{\sqrt{6}}{\pi}\sigma \qquad (4.125)$$

$$x_0 = \mu + \gamma_E\frac{\sqrt{6}}{\pi}\sigma = \mu + 0.45\sigma \qquad (4.126)$$

Quantiles of EV-1 for X_{min}

In case of the smallest value we are interested in non-exceedance probability of X_{min}. Let this non-exceedance probability pbe denoted by p then the value of X_{min} for a specified non-exceedance probability p can be derived from (4.121) and (4.122):

$$x_{\min}(p) = x_0 + \beta\ln(-\ln(1-p)) = \mu_{X\min} + \sigma_{X\min}\frac{\sqrt{6}}{\pi}\{\gamma_E + \ln(-\ln(1-p))\} \qquad (4.125)$$

Example 4.11 EV-1 for smallest value

Annual minimum flow series of a river have a mean and standard deviation of 500 m^3/s and 200 m^3/s. Assuming that the frequency distribution of the minimum flows is EV-1, what is the probability of zero flow?

The problem can be solved by equation (4.104), which requires values for x_0 and β. From (4.125) and (4.126) it follows for x_0 and β:

$$\beta = \frac{\sqrt{6}}{\pi}\sigma = 0.7797 x 200 = 155.9$$

$$x_0 = \mu + 0.45\sigma = 500 + 0.45 x 200 = 590.0$$

Substituting the parameter values in equation (4.104) gives:

$$F_{X\min}(0) = 1 - \exp\left(-\exp\left(\frac{0 - x_0}{\beta}\right)\right) = 1 - \exp\left(-\exp\left(-\frac{590.0}{155.9}\right)\right) = 1 - 0.9775 = 0.0225 \approx \frac{1}{45}$$

Hence, on average once every 45 years the river will run dry according to the EV-1 distribution

4.6.4 Extreme value Type 2 or Fréchet distribution

EV-2 for largest value

The cdf of the Extreme Value Type II distribution for largest value for is given by (4.100):

$$F_{X\max}(x) = \exp\left(-\left(\frac{x - x_0}{\beta}\right)^{1/k}\right) \text{ for : } x \geq x_0 \; ; \; k < 0 \; ; \; \beta > 0 \qquad (4.100)$$

The pdf is obtained by differentiation:

$$f_{X\max}(x) = -\frac{1}{k\beta}\left(\frac{x - x_0}{\beta}\right)^{1/k-1} \exp\left(-\left(\frac{x - x_0}{\beta}\right)^{1/k}\right) \qquad (4.126)$$

Introducing the reduced variate Z according to (4.108), the following standardised forms are obtained for the cdf and the pdf:

$$F_Z(z) = \exp(-z^{1/k}) \qquad (4.127)$$

$$f_Z(z) = -\frac{1}{k} z^{1/k-1} \exp(-z^{1/k}) \qquad (4.128)$$

In Figures 4.27 and 4.28 the pdf and cdf of the EV-2 distribution are presented for different values of k.

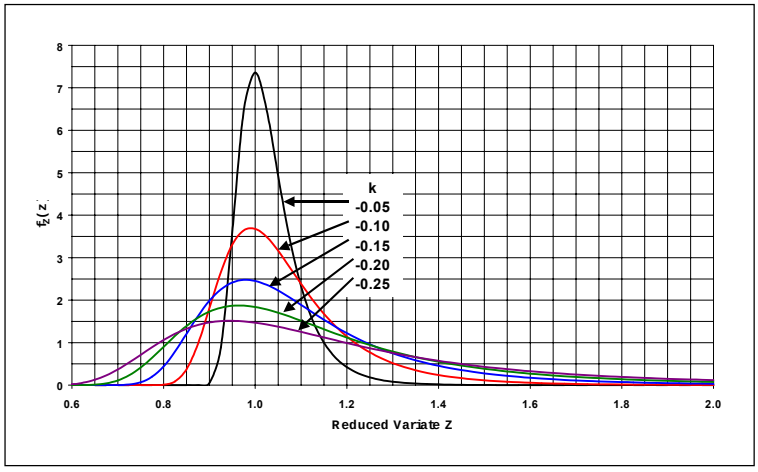

Figure 4.27:
Pdf of EV-2 distribution for different k values

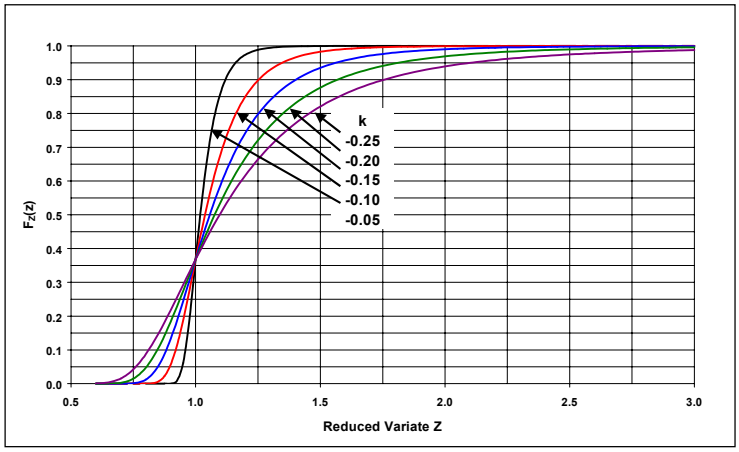

Figure 4.28:
Cdf of EV-2 distribution for different k-values

The moment related parameters of the distribution read:

$$\left.\begin{aligned} \mu_{Xmax} &= x_0 + \beta\Gamma(1+k) \\ M_{Xmax} &= x_0 + \beta(\ln 2)^k \\ m_{Xmax} &= x_0 + \beta(1-k)^k \end{aligned}\right\} \qquad (4.129a)$$

$$\left.\begin{aligned} \sigma^2_{Xmax} &= \beta^2\left\{\Gamma(1+2k) - \Gamma^2(1+k)\right\} \\ \gamma_{1,Xmax} &= \frac{\Gamma(1+3k) - 3\Gamma(1+k)\Gamma(1+2k) + 2\Gamma^3(1+k)}{\left(\Gamma(1+2k) - \Gamma^2(1+k)\right)^{3/2}} \end{aligned}\right\} \qquad (4.129b)$$

Above expressions show that x_0 is a **location** parameter, β a **scale** parameter and k a **shape** parameter as the latter is the sole parameter affecting skewness. From the above figures it is observed that the skewness decreases with increasing k.

The moment related parameters (4.129 a and b) can easily be derived from the following expression for the r^{th} moment about the origin in case $x_0 = 0$ substituted in (3.30):

$$\mu_r' = \beta^r\Gamma(1+rk) \qquad (4.130)$$

From (4.129) it is observed that the distribution parameters cannot analytically be expressed in the moments of the distribution; an iterative procedure is required for this.

Quantiles of EV-2 for X_{max}

The quantile $x_{max}(T)$ for a given return period T follows from (4.100):

$$x_{max}(T) = x_0 + \beta\left(\ln\left(\frac{T}{T-1}\right)\right)^k \qquad (4.131)$$

Fréchet and log-Gumbel distributions

EV-2 for the largest value is also indicated as **Fréchet** distribution or **log-Gumbel** distribution. With respect to the latter it can be shown that if $(x_{max}-x_0)$ has a EV-2 distribution, its logarithm $Y= \ln(x_{max}-x_0)$ has a Gumbel distribution with parameters *a* and *b*, as follows:

$$F_Y(y) = \exp\left\{-\exp\left(-\left(\frac{y-b}{a}\right)\right)\right\}$$

$$F_{Xmax}(x) = \exp\left\{-\exp\left(-\left(\frac{\ln(x-x_0)-b}{a}\right)\right)\right\}$$

Since:

$$\exp\left(-\left(\frac{\ln(x-x_0)}{a}\right)\right) = (x-x_0)^{-1/a}$$

it follows:

$$F_{Xmax}(x) = \exp\left\{-(x-x_0)^{-1/a}e^{b/a}\right\} = \exp\left\{-\left(\frac{x-x_0}{e^b}\right)^{-1/a}\right\}$$

It is observed that above expression equals (4.100) for

$$\left.\begin{aligned} a &= -k \\ b &= \ln(\beta) \end{aligned}\right\} \qquad (4.132)$$

EV-2 for smallest value

The Extreme Value Type II distribution for the smallest value is given by (4.105)

$$F_{Xmin}(x) = 1 - \exp\left(-\left(-\frac{x-x_0}{\beta}\right)^{1/k}\right) \text{ for : } x \le x_0 \text{ ; } k < 0 \text{ ; } \beta > 0 \qquad (4.105)$$

The pdf can be derived by taking the derivative of (4.105) with respect to x:

$$f_{Xmin}(x) = -\frac{1}{k\beta}\left(-\frac{x-x_0}{\beta}\right)^{1/k-1}\exp\left(-\left(-\frac{x-x_0}{\beta}\right)^{1/k}\right) \qquad (4.133)$$

The moment related parameters of the distribution can easily be obtained from (4.129a and b) knowing that the pdf is the mirror image of the pdf for the largest value:

$$\left.\begin{aligned} \mu_{Xmin} &= x_0 - \beta\Gamma(1+k) \\ M_{Xmin} &= x_0 - \beta(\ln 2)^k \\ m_{Xmin} &= x_0 - \beta(1-k)^k \end{aligned}\right\} \qquad (4.134a)$$

$$\left.\begin{aligned} \sigma^2_{Xmin} &= \beta^2\left\{\Gamma(1+2k) - \Gamma^2(1+k)\right\} \\ \gamma_{1,Xmin} &= -\frac{\Gamma(1+3k) - 3\Gamma(1+k)\Gamma(1+2k) + 2\Gamma^3(1+k)}{\left(\Gamma(1+2k) - \Gamma^2(1+k)\right)^{3/2}} \end{aligned}\right\} \qquad (4.134b)$$

It appears that the EV-2 for the smallest value finds little application in hydrology and will therefore not be discussed any further.

4.6.5 Extreme value Type 3 distribution

EV-3 for largest value

The Extreme Value Type III distribution for largest value is given by (4.101) and is defined for $x \leq x_0$, $k > 0$ and $\beta > 0$

$$F_{Xmax}(x) = \exp\left(-\left(-\frac{x - x_0}{\beta}\right)^{1/k}\right) \tag{4.101}$$

The pdf reads:

$$f_{Xmax}(x) = \frac{1}{k\beta}\left(-\frac{x - x_0}{\beta}\right)^{1/k-1} \exp\left(-\left(-\frac{x - x_0}{\beta}\right)^{1/k}\right) \tag{4.135}$$

The mean, median, mode, variance and skewness are given by:

$$\left.\begin{aligned} \mu_{Xmax} &= x_0 - \beta\Gamma(1+k) \\ M_{Xmax} &= x_0 - \beta(\ln 2)^k \\ m_{Xmax} &= x_0 - \beta(1-k)^k \end{aligned}\right\} \tag{4.136a}$$

$$\left.\begin{aligned} \sigma^2_{Xmax} &= \beta^2\left\{\Gamma(1+2k) - \Gamma^2(1+k)\right\} \\ \gamma_{1,Xmax} &= -\frac{\Gamma(1+3k) - 3\Gamma(1+k)\Gamma(1+2k) + 2\Gamma^3(1+k)}{\left(\Gamma(1+2k) - \Gamma^2(1+k)\right)^{3/2}} \end{aligned}\right\} \tag{4.136b}$$

Note that these expressions are similar to those of the smallest value modelled as EV-2. Above moment related parameters are easily obtained from the r^{th} moment of $(x_0 - X_{max})$ which can shown to be:

$$E\left[(x_0 - X_{max})^r\right] = \beta^r\Gamma(1+rk) \tag{4.137}$$

To simplify the computation, note that for the higher moments x_0 can be omitted, so for $r > 1$ one can put $x_0 = 0$ and use (3.30). Equation (4.137) then simplifies to:

$$\mu'_r = (-1)^r\beta^r\Gamma(1+rk)$$

So:

$$\mu'_2 = \beta^2\Gamma(1+2k)$$
$$\mu'_3 = -\beta^3\Gamma(1+3k)$$

The fact that X_{max} is bounded by x_0 makes that EV-3 is seldom used in hydrology for modelling the distribution of X_{max}. Its application only make sense, if there is a physical reason that limits X_{max} to x_0.

EV-3 for smallest value

The extreme value Type III distribution for the smallest value, for $x \geq x_0$, $k > 0$ and $\beta > 0$, has the following form:

$$F_{Xmin}(x) = 1 - \exp\left(-\left(\frac{x - x_0}{\beta}\right)^{1/k}\right) \tag{4.106}$$

and the pdf reads:

$$f_{Xmin}(x) = \frac{1}{k\beta}\left(\frac{x - x_0}{\beta}\right)^{1/k-1} \exp(-\left(\frac{x - x_0}{\beta}\right)^{1/k}) \qquad (4.138)$$

In above equations, x_0 is a location parameter, β a scale parameter and k a shape parameter.

This distribution is seen to be identical to the **Weibull** distribution, equation (4.84) and (4.85), by putting 1/k = k*, where k* is the shape parameter of the Weibull distribution. Hence reference is made to Sub-section 4.3.11 for further elaboration of this distribution. Above distribution is also called **Goodrich** distribution.

The moment related parameters according to the above definition are shown here, as it corresponds to the parameter definition adopted in HYMOS. The mean, median, mode, variance and skewness read:

$$\left.\begin{aligned} \mu_{Xmin} &= x_0 + \beta\Gamma(1+k) \\ M_{Xmin} &= x_0 + \beta(\ln 2)^k \\ m_{Xmin} &= x_0 + \beta(1-k)^k \end{aligned}\right\} \qquad (4.139a)$$

$$\left.\begin{aligned} \sigma^2_{Xmin} &= \beta^2\left\{\Gamma(1+2k) - \Gamma^2(1+k)\right\} \\ \gamma_{1,Xmin} &= \frac{\Gamma(1+3k) - 3\Gamma(1+k)\Gamma(1+2k) + 2\Gamma^3(1+k)}{\left(\Gamma(1+2k) - \Gamma^2(1+k)\right)^{3/2}} \end{aligned}\right\} \qquad (4.139b)$$

The location parameter x_0 is seen to be the lower bound of the distribution. Often, the parent distribution will have a lower bound equal to zero and so will have the EV-3 for the smallest value. Above form with x_0 is therefore often indicated as the **shifted** Weibull distribution.

In literature the shifted Weibull distribution is often presented as:

$$F_{Xmin}(x) = 1 - \exp\left(-\left(\frac{x-b}{a-b}\right)^c\right) \quad \text{with : } x > b\,;\, a > b \qquad (4.140)$$

where the resemblance with the above parameter definition is seen for: $x_0 = b$, $\beta = a - b$ and $k = 1/c$.

Quantiles of EV-3 for X_{min}

Since one is dealing with the smallest value, interest is in the non-exceedance probability of X_{min}. If this non-exceedance probability is denoted by p then the value of x_{min} for a specified non-exceedance probability p can be derived from (4.106):

$$x_{min}(p) = x_0 + \beta\left\{-\ln(1-p)\right\}^k \qquad (4.141)$$

Example 4.11 (continued.) EV-3 for smallest value.

Annual minimum flow series of a river have a mean and standard deviation of 500 m^3/s and 200 m^3/s. Assuming that the frequency distribution of the minimum flows is EV-3, with $x_0 = 0$, what low flow value will not be exceeded on average once in 100 years?

The non-exceedance probability q = 0.01. To apply (4.141) k and β have to be known. The parameters k and β are obtained as follows. Note that for x_0, the coefficient of variation becomes:

$$C_{v,Xmin}^2 = \left(\frac{\sigma_{Xmin}}{\mu_{Xmin}}\right)^2 = \frac{\Gamma(1+2k)}{\Gamma^2(1+k)} - 1 = \left(\frac{200}{500}\right)^2 = 0.16$$

From above equation it is observed that the coefficient of variation is only a function of k when $x_0 = 0$. By iteration one finds k = 0.37. From (4.139b) it follows for β:

$$\beta = \frac{\sigma_{Xmin}}{\left(\Gamma(1+2k) - \Gamma^2(1+k)\right)^{1/2}} = \frac{200}{0.3549} = 564$$

With β = 564 and k = 0.37 one finds with (4.141) for the 100 year low flow:

$$X_{min}(0.01) = 0 + 564x\{-\ln(1-0.01)\}^{0.37} = 564x0.182 = 103 m^3/s$$

According to the EV-1 distribution for the smallest value, which was applied to the same series in Sub-section 4.4.3, Q = 103 m^3/s has a return period of about 23 years. It follows that the two distributions lead to very different results. In practice, the EV-3 for smallest value finds widest application.

4.6.6 Generalised Pareto distribution

For modelling frequency distributions of extremes, particularly of partial duration series, the Pareto distribution is often used.The cdf of the generalised Pareto distribution has the following form:

$$F_X(x) = 1 - \left(1 - \theta\left(\frac{x - x_0}{\sigma}\right)\right)^{1/\theta} \qquad (4.142)$$

Like for the Extreme Value distributions as discussed in the previous sub-sections, three types of Pareto distributions are distinguished, which are directly related to EV-1, 2 and 3 (see next sub-chapter):

- **Type I distribution, P-1:**

$$F_X(x) = 1 - \exp\left(-\left(\frac{x - x_0}{\sigma}\right)\right) \text{ for : } x_0 \le x < \infty \text{ when : } \theta = 0 \qquad (4.143)$$

- **Type II distribution, P-2:**

$$F_X(x) = 1 - \left(1 - \theta\left(\frac{x - x_0}{\sigma}\right)\right)^{1/\theta} \text{ for : } x_0 \le x < \infty \text{ when : } \theta < 0 \qquad (4.144)$$

- **Type III distribution, P-3:**

$$F_X(x) = 1 - \left(1 - \theta\left(\frac{x - x_0}{\sigma}\right)\right)^{1/\theta} \text{ for : } x_0 \le x \le x_0 + \frac{\sigma}{\theta} \text{ when : } \theta > 0 \qquad (4.145)$$

The pdf's of the Pareto distributions are respectively with the validity range as defined for the cdf's above, for **P-1**:

$$f_X(x) = \frac{1}{\sigma}\exp\left(-\left(\frac{x - x_0}{\sigma}\right)\right) \qquad (4.146)$$

and for **P-2** and **P-3**:

$$f_X(x) = \frac{1}{\sigma}\left(1-\theta\left(\frac{x-x_0}{\sigma}\right)\right)^{1/\theta-1} \qquad (4.147)$$

Note that the P-1 distribution results as a special for $\theta = 0$ from P-2 or P-3 similar to the EV-1 distribution resulting from GEV, see Sub-section 4.4.2. In the above distributions, x_0 is a **location** parameter, σ is a **scale** parameter and θ is a **shape** parameter. The mean, variance, skewness and kurtosis of the distributions are given by:

$$\left.\begin{aligned}
\mu_X &= x_0 + \frac{\sigma}{1+\theta} \\
\sigma_X^2 &= \frac{\sigma^2}{(1+\theta^2)(1+2\theta)} \\
\gamma_{1,X} &= \frac{2(1-\theta)\sqrt{1+2\theta}}{1+3\theta} \\
\gamma_{2,X} &= \frac{3(1+2\theta)(3-\theta+2\theta^2)}{(1+3\theta)(1+4\theta)}
\end{aligned}\right\} \qquad (4.148)$$

Above expressions can be derived by noticing (Metcalfe, 1997):

$$E\left[(1-\theta\frac{X}{\sigma})^r\right] = \frac{1}{1+r\theta} \text{ for : } \theta > -\frac{1}{r} \qquad (4.149)$$

For $\theta < -1/r$ the r^{th} moment does not exist.

The generalised Pareto distribution in a standardised form ($x_0 = 0$ and $\sigma = 1$) for various values of θ are given in Figures 4.29 and 4.30.

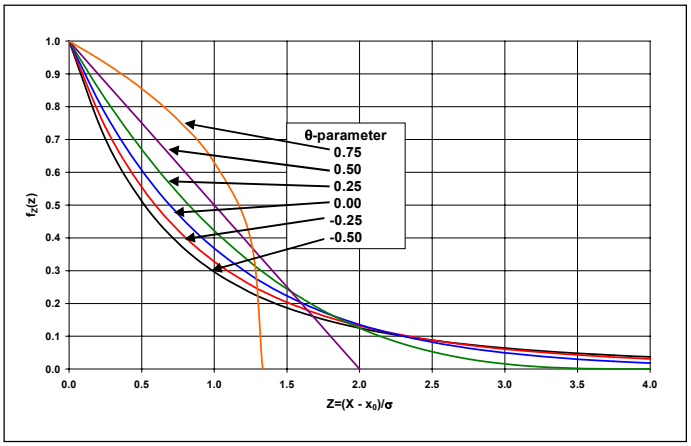

Figure 4.29:
Pdf of Pareto distribution for various values of shape parameter

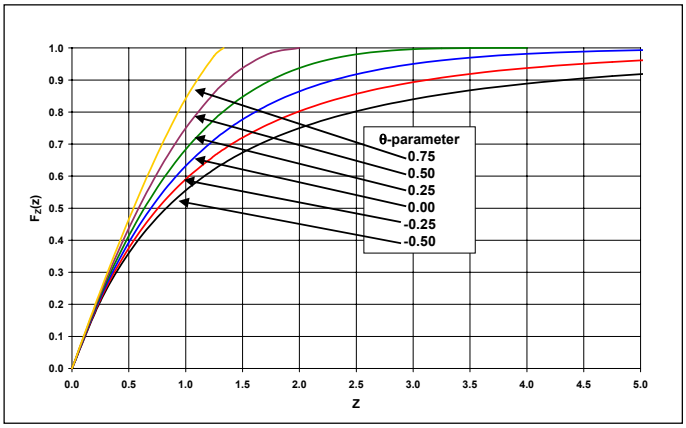

Figure 4.30:
Cdf of Pareto distribution for various of shape parameter

Quantiles

The quantiles, referring to a return period of T years, follow from (4.143) to (1.145) and read:

- For Type I distribution P-1:

$$x_T = x_0 + \sigma \ln T \qquad (4.150)$$

- For Type II and III distributions, P-2, P-3:

$$x_T = x_0 + \frac{\sigma}{\theta}(1 - T^{-\theta}) \qquad (4.151)$$

Note that above two expressions should not directly be applied to exceedance series unless the number of data points coincide with the number of years, see next sub-section.

4.6.7 Relation between maximum and exceedance series

The GEV distributions are applicable to series with a fixed interval, e.g. a year: series of the largest or smallest value of a variable each year, like annual maximum or minimum flows. If one considers largest values, such a series is called an **annual maximum series.** Similarly, **annual minimum series** can be defined.

In contrast to this, one can also consider series of extreme values above or below a certain threshold value, i.e. the maximum value between an upcrossing and a downcrossing or the minimum between a downcrossing and an upcrossing, see Figure 4.31.

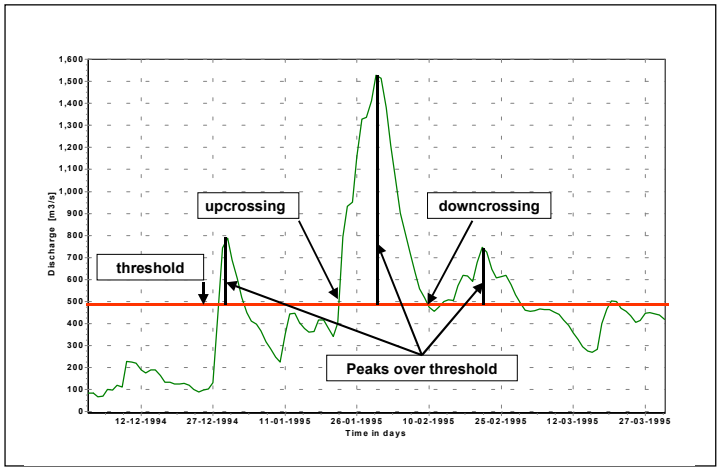

Figure 4.31:
Definition of partial duration or peaks over threshold series

The series resulting from exceedance of a base or threshold value x_0 thereby considering only the maximum between an upcrossing and a downcrossing is called a partial duration series (PDS) or peaks over threshold series, POT-series. The statistics may be developed for the exceedance of the value relative to the base only or for the value as from zero. The latter approach will be followed here. In a similar manner partial duration series for non-exceedance of a threshold value can be defined. When considering largest values, if the threshold is chosen such that the number of exceedances N of the threshold value equals the number of years n, the series is called annual exceedance series. So, if there are n years of data, in the annual exceedance series the n largest independent peaks out of $N \geq n$ are considered. To arrive at independent peaks, there should be sufficient time between successive peaks. The physics of the process determines what is a sufficient time interval between peaks to be independent; for flood peaks a hydrograph analysis should be carried out. The generalised Pareto distribution is particularly suited to model the exceedance series.

Note that there is a distinct difference between annual maximum and annual exceedance series. In an annual maximum series, for each year the maximum value is taken, no matter how low the value is compared to the rest of the series. Therefore, the maximum in a particular year may be less than the second or the third largest in another year, which values are considered in the annual exceedance series if the ranking so permits. Hence the lowest ranked annual maximums are less than (or at the most equal to) the tail values of the ranked annual exceedance series values.

The procedure to arrive at the annual exceedance series via a partial duration series and its comparison with the annual maximum series is shown in the following figures, from a record of station Chooz on Meuse river in northern France (data 1968-1997). The original discharge series is shown in Figure 4.32. Next a threshold level of 400 m^3/s has been assumed. The maximum values between each upcrossing and the next downcrossing are considered. In this particular case, peaks which are distanced $\geq$ 14 days apart are expected to be independent and are included in the partial duration series, shown in Figure 4.33. This results in 72 peaks. Since there are 30 years of record, the partial duration series has to be reduced to the 30 largest values. For this the series values are ranked in descending order and the first 30 values are taken to form the annual exceedance series. The threshold value for the annual exceedance series appears to 620 m^3/s. The annual exceedance series is shown in Figure 4.34. It is observed that some years do not contribute to the series, as their peak values were less than 620 m^3/s, whereas other years contribute with 2 or some even with 3 peaks. The annual maximum series is presented in Figure 4.35, together with the threshold for the annual exceedance series. It is observed that indeed for a number of years that threshold level was not reached. A comparison of the two series is depicted in Figure 4.36.

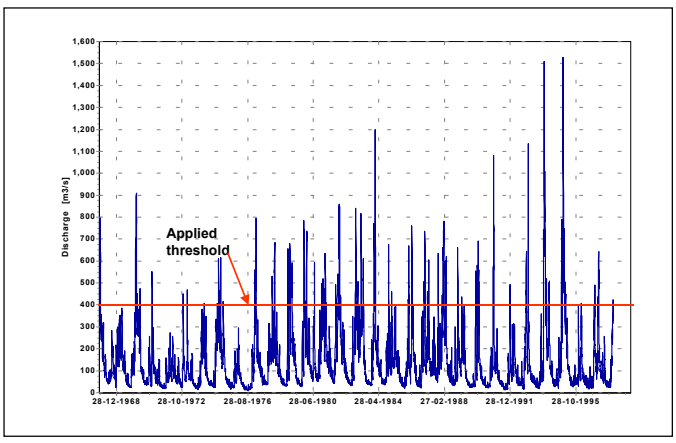

Figure 4.32:
Discharge series of station chooz on Meuse river with applied threshold Q = 400 m^3/s

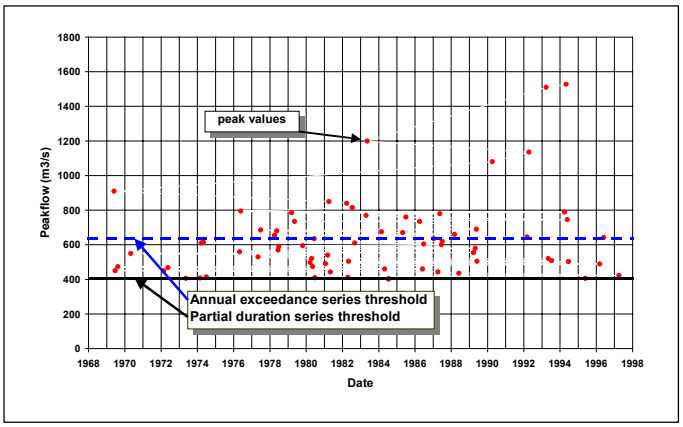

Figure 4.33:
Partial duration series of peaks over 400 m^3/s

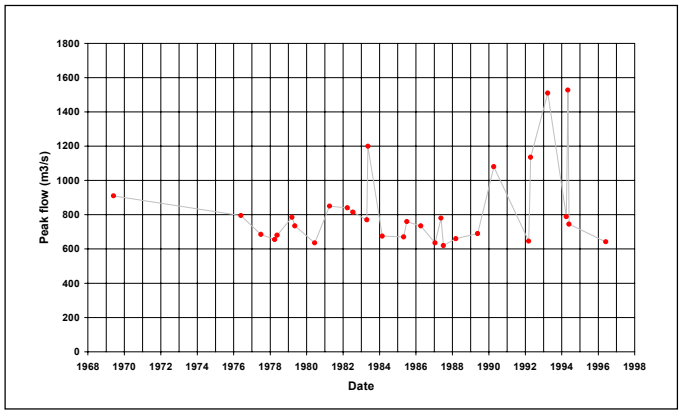

Figure 4.34:
Annual exceedance series $Q \geq 620$ m^3/s

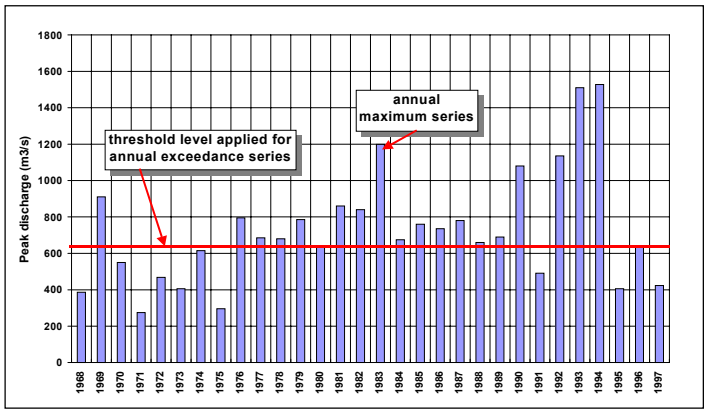

Figure 4.35:
Annual maximum series

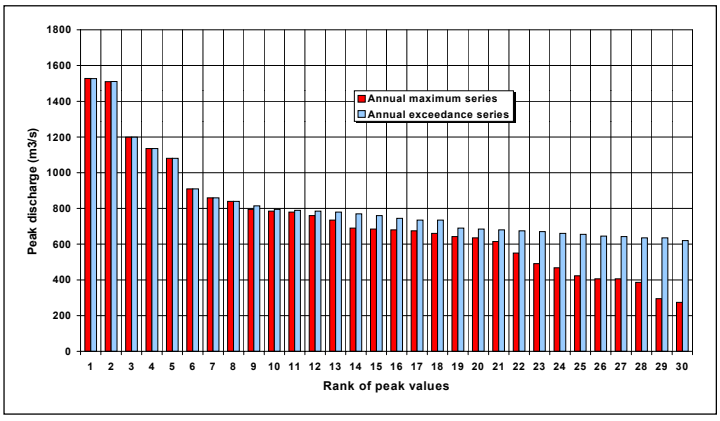

Figure 4.36:
Comparison of annual maximum series and annual exceedance series

From Figure 4.36 it is observed that the largest values in both series are the same, but the lower tail is quite different. It follows that the annual maximum series will produce lower extremes for low return periods, say up to T = 5 or T = 10 years return period.

Conditional exceedance probabilities

It is noted that straightforward application of fitting a frequency distribution to a partial duration or peak over threshold series (i.e. an **exceedance** series) involves a conditional distribution, i.e. the probability of an exceedance given that a threshold level x_0 has been exceeded. Let this distribution of peaks over a threshold x_0 be denoted by $F_{POT}(x)$. If there are N_e exceedances of x_0 during N_y years, then the average number of exceedances of x_0 in one year is $\lambda = N_e / N_y$, and the average number of peaks $X>x|x>x_0$ per year becomes $\lambda(1- F_{POT}(x))$. The average number of peaks $X>x|x>x_0$ in T years then is $\lambda T(1- F_{POT}(x))$. To arrive at the T year flood the average number of peaks in T year should be 1, i.e.

$$\lambda T\left(1-F_{POT}(x)\right)=1$$

or:

$$F_{POT}(x)=1-\frac{1}{\lambda T} \qquad (4.152)$$

Substitution of a suitable model for $F_{POT}(x)$ in (4.152) like the P-1 distribution gives for the quantile x_T:

$$x_T = x_0 + \sigma \ln(\lambda T) \qquad (4.153)$$

It is observed that (4.153) is identical to (4.150) for λ = 1, i.e. when the number of exceedances is equal to the number of years and then the peak over threshold series becomes the annual exceedance series.

From exceedances to maximum

Consider again the distribution of the peaks over threshold: $F_{POT}(x)$. The number of exceedances N of the threshold in a fixed time period is a random variable, having a certain probability mass function $p_N(n)$. It can be shown (see e.g. Kottegoda and Rosso, 1997) that the cdf of X_{max} (i.e. the largest of the exceedances) can be derived from the conditional frequency distribution $F_{POT}(x)$ and $p_N(n)$ as follows:

$$F_{Xmax}(x)=\sum_{n=0}^{\infty}\{F_{POT}(x)\}^n p_N(n) \qquad (4.154)$$

If $p_N(n)$, i.e. the number of exceedances, is modelled by a **Poisson distribution**, which is equivalent to stating that the intervals between exceedances is exponentially distributed, then (4.154) simplifies to:

$$F_{Xmax}(x)=\exp\{-\lambda(1-F_{POT}(x))\} \qquad (4.155)$$

where: λ = average number of exceedances (e.g. per year).

Equation (4.155) gives a relation between the **conditional exceedance** distribution $F_{POT}(x)$ and the **unconditional (annual) maximum** distribution. If **annual exceedance** series are considered (i.e. on average one exceedance per year: λ = 1) with distribution function $F_{AE}(x)$ it follows from (4.155):

$$F_{Xmax}(x)=\exp\{-(1-F_{AE}(x)\} \qquad (4.156)$$

Equation (4.156) gives the relation between the annual maximum distribution $F_{xmax}(x)=F_{AM}(x)$ and the frequency distribution of the annual exceedance series $F_{AE}(x)$. For the relation between the return period of the annual exceedance series T_{AE} and the annual maximum series T_{AM} it follows:

$$1-\frac{1}{T_{AM}}=\exp\left(-\frac{1}{T_{AE}}\right) \text{ or : } T_{AM}=\left\{1-\exp\left(-\frac{1}{T_{AE}}\right)\right\}^{-1} \tag{4.157}$$

Equivalently

$$T_E=\left\{\ln\left(\frac{T_M}{T_M-1}\right)\right\}^{-1} \tag{4.158}$$

From Pareto to GEV

If one substitutes in equation (4.156) for the distribution of the exceedances $F_{AE}(x)$ the generalised Pareto distribution as discussed in the previous sub-section, then the distribution of X_{max} will be a GEV distribution with the same **shape** parameter. The cdf of the generalised Pareto distribution was given by (4.142):

$$F_X(x)=1-\left(1-\theta\left(\frac{x-x_0}{\sigma}\right)\right)^{1/\theta} \tag{4.142}$$

Substitution in (4.156) gives:

$$F_{Xmax}(x)=\exp(-\lambda\left\{1-\left[1-\left(1-\theta\left(\frac{x-x_0}{\sigma}\right)\right)^{1/\theta}\right]\right\})=\exp(-\left(\frac{\frac{\sigma}{\theta}-(x-x_0)}{\frac{\lambda^{-\theta}\sigma}{\theta}}\right)^{1/\theta})$$

To prove the resemblance with the GEV distribution given by equation (4.102), note that:

$$F_{Xmax}(x)=\exp(-\left(1-k\left(\frac{x-b}{a}\right)\right)^{1/k})=\exp(-\left(\frac{\frac{a}{k}-(x-b)}{\frac{a}{k}}\right)^{1/k}) \tag{4.102}$$

It follows that (4.142) and (4.102) are equivalent if:

$$\left.\begin{aligned} k&=\theta\\ a&=\lambda^{-\theta}\sigma\\ b&=x_0+\frac{\sigma}{\theta}(1-\lambda^{-\theta}) \end{aligned}\right\} \tag{4.159}$$

It shows that the generalised Pareto distribution and the GEV distribution are directly related, provided that the number of exceedances per fixed period of time can be modelled by a Poisson distribution.

Example 12: Annual exceedances and annual maxima

As an example consider the exceedances shown above for Chooz on Meuse river. Since there are 72 exceedances in 30 years, the average number of exceedances per year is 72/30 = 2.4, hence λ = 2.4. The comparison of the Poisson distribution with the observed distribution of exceedances N is presented in Figure 4.37.

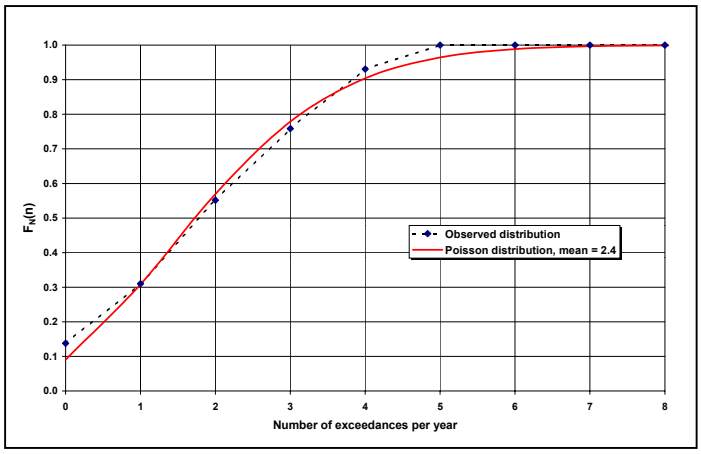

Figure 4.37:
Modelling of number of Q = 400 m^3/s threshold; Meuse river at Chooz

From Figure 4.37 it is observed that in the example case the Poisson distribution is a suitable model for the frequency distribution of the number of exceedances per year.

Summing up

To model the distributions of exceedances, apart from Pareto type distributions, basically any other distribution may be used, provided a proper fit is obtained. Then equation (4.155) or (4.156) is used to compute from such a fit the return period referring to the annual maximum value, consistent with annual maximum series. It follows:

$$T_{AM}(x) = \frac{1}{1-\exp\{-\lambda(1-F_{POT}(x))\}} = \frac{1}{1-\exp\{-(1-F_{AE}(x))\}} \qquad (4.160)$$

Example 12 (continued)

To show the procedure let's follow the Meuse example presented above. The average number of exceedances per year was λ = 2.4. The exceedances are fitted by an exponential distribution. The average discharge of the recorded peak flows exceeding 400 m^3/s is 232.5 m^3/s, hence x_0 = 400 and β = 233, see Sub-section 4.5.1 Hence $F_{POT}(x)$ reads:

$$F_{POT}(x) = 1-\exp\left(-\left(\frac{x-400}{233}\right)\right) \qquad (4.161)$$

The fit of the exponential distribution to the observed frequencies is shown in Figure 4.38.

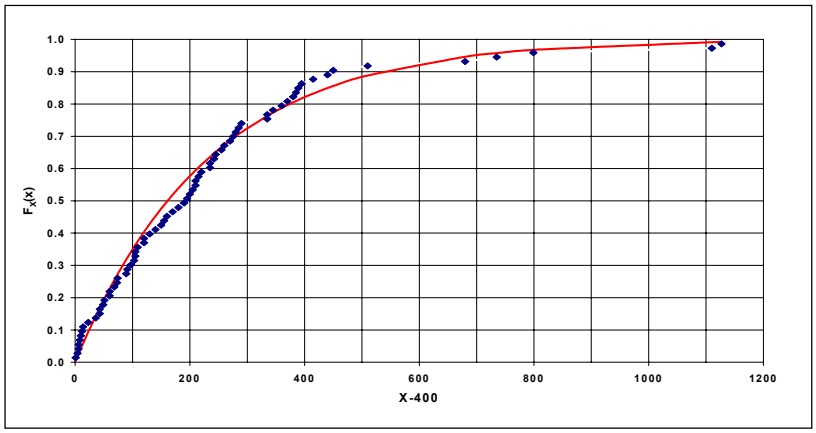

Figure 4.38:
Fir of exponential distribution to Meuse flow at Chooz exceeding threshold of 400 m^3/s

From equation (4.155) the cdf of the annual flood discharge then reads:

$$F_{X\max}(x) = \exp(-\lambda(1-F_X(x))) = \exp(-2.4(1-(1-\exp(-\left(\frac{x-400}{233}\right))) =$$

$$= \exp(-\exp(-\left(\frac{x-604}{233}\right))) \qquad (4.157)$$

The distribution of the annual maximum is seen to have a Gumbel distribution, and for the return period it follows:

$$T(x) = \frac{1}{1-\exp(-\exp(-\left(\frac{x-604}{233}\right)))} \qquad (4.158)$$

If the procedure is carried out by applying the Gumbel distribution on annual maximum series for the same period, the parameter values are instead of 604 and 233, respectively 591 and 238. A comparison between both approaches is shown in Figure 4.39. It is observed that both procedures give very similar results (differences <1% for 2<T≤100).

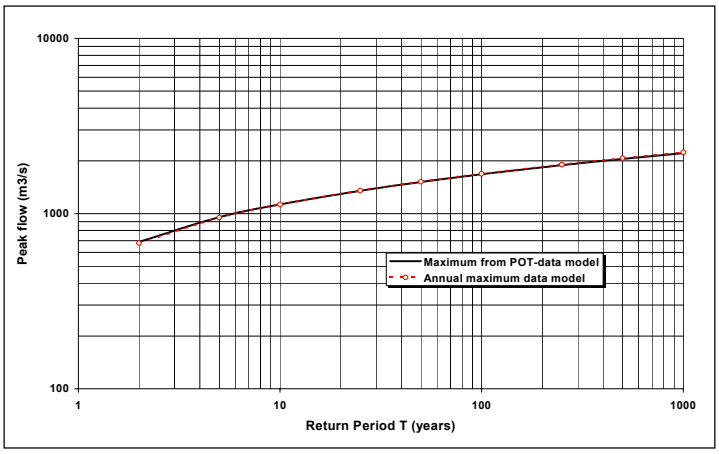

Figure 4.39:
Flow extremes as function of return period derived from POT-series transferred to maximum and directly from annual maximum series.

From (4.158) it follows for the quantile x_T:

$$x_T = 604 - 233\ln(\ln\left(\frac{T}{T-1}\right)) \qquad (4.159)$$

According to the conditional distribution it follows from (4.153) and (4.161) with λ = 2.4 for the quantile x_T:

$$x_T = 400 + 233\ln(2.4T) \qquad (4.160)$$

A comparison between both approaches is seen in Figure 4.40:

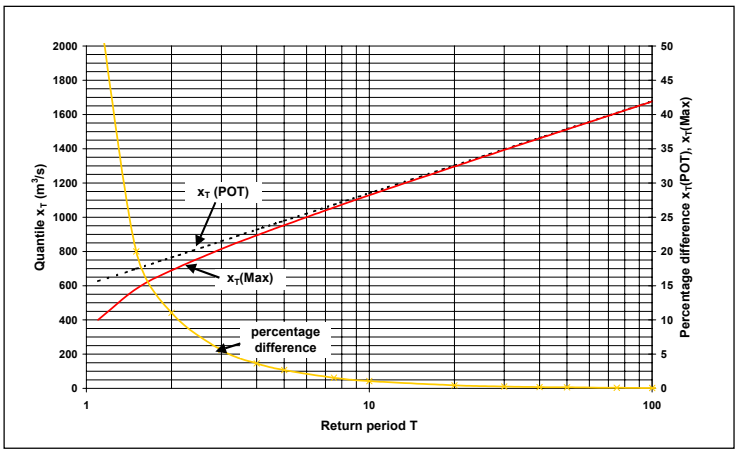

Figure 4.40:
Quantiles according to POT and Maximum, both from exceedance series

From Figure (4.40) it is observed that there is a distinct difference between the two approaches for return periods up to about 3 (diff > 5%), at a return periods of 10 the difference in only 1% and reduces thereafter to insignificant differences.

4.7 Sampling distributions

4.7.1 General

A distribution parameter can be estimated from a particular sample in a number of ways. The rule or method used to estimate a parameter is called an **estimator**; the value that the estimator gets, when applied, is called an **estimate**. An estimate of a distribution parameter of a particular series will assume a number of values dependent on the sample taken from the entire population. It is a random variable itself with a particular frequency distribution. Hence, one can only speak about the true value of a parameter in probabilistic terms. Consequently, also the quantiles computed from the frequency distributions are random variables with a particular distribution. Many of the estimated distribution parameters and quantiles are asymptotically normally distributed. This implies that for large sample sizes N the estimate and the standard error fully describe the probability distribution of the statistic. For small sample sizes the sampling distributions may, however, deviate significantly from normality. In addition to the normal distribution important sampling distributions are the Chi-square distribution, the Student-t distribution and the Fisher F-distribution. The normal distribution was described in detail in Sub-section 4.4.1. The latter 3 distributions will be described in the next sub-sections.

4.7.2 Chi-squared distribution

Let Z_1, Z_2, Z_3, …, Z_ν be ν independent standard normal random variables, then the Chi-squared variable χ_ν^2 with ν degrees of freedom is defined as:

$$\chi_\nu^2 = Z_1^2 + Z_2^2 + Z_3^2 + \ldots + Z_\nu^2 \qquad (4.161)$$

The number of degrees of freedom ν represents the number of independent or 'free' squares entering into the expression. The pdf and cdf are given by (4.65) and (4.66) respectively, which with X replaced by χ^2 read:

$$f_{\chi^2}(x) = \frac{1}{2\Gamma(\nu/2)}\left(\frac{x}{2}\right)^{\nu/2-1} \exp(-\frac{x}{2}) \text{ for : } x \geq 0\,,\ \nu > 0 \qquad (4.162)$$

$$F_{\chi^2}(x) = \frac{1}{2\Gamma(\nu/2)} \int_0^x \left(\frac{s}{2}\right)^{\nu/2-1} \exp(-\frac{s}{2})ds \qquad (4.163)$$

The χ^2 -distribution is a particular case of the gamma distribution by putting $\beta = 2$ and $\gamma = \nu/2$ in equations (4.51) and (4.52).The function $f_{\chi 2}(x)$ for different degrees of freedom is depicted in Figure 4.41.

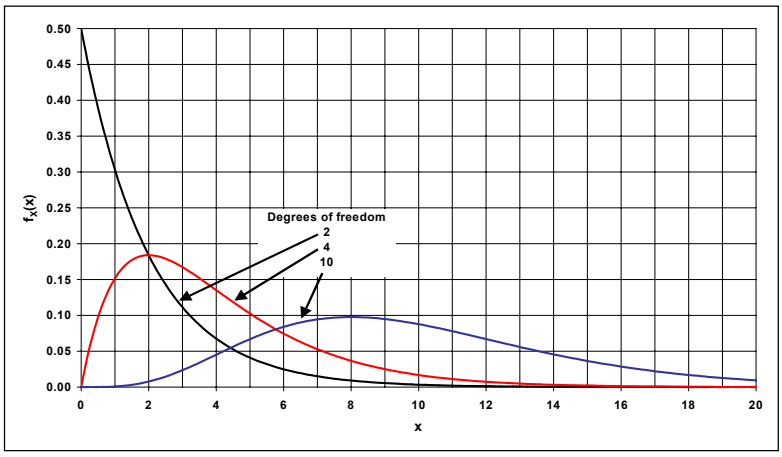

Figure 4.41: χ_ν^2 -probability density function for ν = 2, 4 and 10 degrees of freedom

Moment related parameters of the distribution

The mean, mode, variance, skewness and kurtosis of the distribution of χ_ν^2 are:

$$\left.\begin{aligned} \mu_{\chi_\nu^2} &= \nu \\ m_{\chi_\nu^2} &= \nu - 2 \text{ for : } \nu \geq 2 \\ \sigma^2_{\chi_\nu^2} &= 2\nu \end{aligned}\right\} \qquad (4.164a)$$

$$\left.\begin{aligned} \gamma_{1,\chi_\nu^2} &= \sqrt{\frac{8}{\nu}} \\ \gamma_{2,\chi_\nu^2} &= 3\left(\frac{\nu+4}{\nu}\right) \end{aligned}\right\} \qquad (4.164b)$$

From (164b) it is observed that for large ν the skewness tends to 0 and the kurtosis becomes 3, and the χ^2 -distribution approaches the normal distribution, with N(ν, 2ν).

It is noted that the addition theorem is valid for the χ^2 –distribution. This implies that a new variable formed by $\chi_\nu^2 = \chi_{\nu 1}^2 + \chi_{\nu 2}^2$ has $\nu = \nu_1 + \nu_2$ degrees of freedom as is simple seen from (4.161). The χ^2 –distribution is often used for making statistical inference about the variance. An unbiased estimator for the variance reads, see (2.5), with the mean estimated by (2.3):

$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)^2 \qquad (2.5)$$

The sum term can be written as follows:

$$\sum_{i=1}^{N}(x_i - m_X)^2 = \sum_{i=1}^{N}\{(x_i - \mu_X) - (m_X - \mu_X)\}^2 = \sum_{i=1}^{N}(x_i - \mu_X)^2 - N(m_X - \mu_X)^2 \qquad (4.165)$$

When the first term of the last right-hand part is divided by σ_X, then one gets a sum of N standard normal variates; if one divides the second part by the standard deviation of the mean, which is $\sigma_X/\sqrt{N}$ then one standard normal variate is obtained. Hence it follows:

$$\sum_{i=1}^{N}(x_i - m_X)^2 = \sigma_X^2 \sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}\right)^2 - \sigma_X^2\left(\frac{m_X - \mu_X}{\sigma_X/\sqrt{N}}\right)^2 = \sigma_X^2(\chi_N^2 - \chi_1^2) = \sigma_X^2\chi_{N-1}^2 \qquad (4.166)$$

Substitution of (4.166) into (2.5) gives:

$$\frac{(N-1)s_X^2}{\sigma_X^2} = \chi_{N-1}^2 \text{ or } :\frac{\nu s_X^2}{\sigma_X^2} = \chi_\nu^2 \text{ with: } \nu = N-1 \qquad (4.167)$$

Hence the random variable $\nu s_X^2/\sigma_X^2$ has a χ^2 –distribution with ν = N-1 degrees of freedom. So, the distribution can be used to make statistical inference about the variance. The χ^2 – distribution is also used for statistical tests on the goodness of fit of a theoretical distribution function to an observed one. This will be discussed in Chapter 6.

4.7.3 Student t distribution

The Student t-distribution results from a combination of a normal and a chi-square random variable. Let Y and Z be independent random variables, such that Y has a χ_ν^2 -distribution and Z a standard normal distribution then the variable T_ν is the Student t variable with ν degrees of freedom when defined by:

$$T_\nu = \frac{Z}{\sqrt{Y/\nu}} \qquad (4.168)$$

The probability density function of T_ν it follows:

$$f_T(t) = \frac{\Gamma\{(\nu+1)/2\}}{\Gamma(\nu/2)}\frac{1}{\sqrt{\pi\nu}}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2} \qquad (4.169)$$

The function $f_T(t)$ for different degrees of freedom is shown in Figure 4.42.

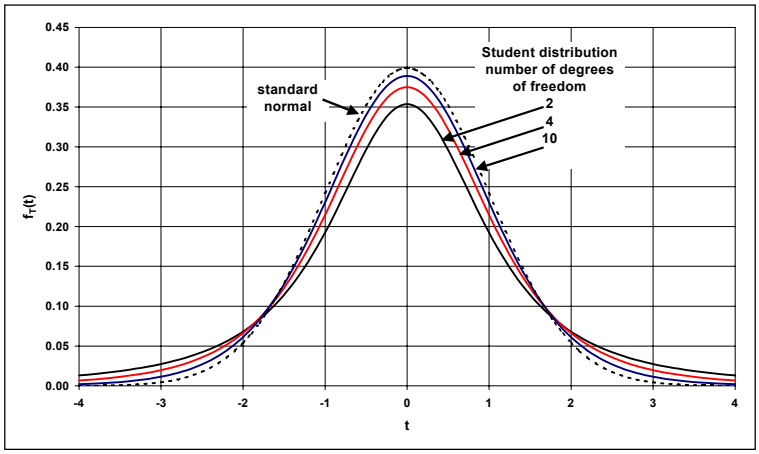

Figure 4.42: Student t-distribution for ν = 2, 4 and 10 degrees of freedom

Moment related parameters of the distribution

The mean and the variance of the variable T_ν are respectively:

$$\left.\begin{aligned} \mu_T &= 0 \ \text{ for : } \nu > 1 \\ \sigma_T^2 &= \frac{\nu}{\nu-2} \ \text{ for : } \nu > 2 \end{aligned}\right\} \qquad (4.170)$$

The Student t-distribution approaches a standard normal distribution when the number of degrees of freedom becomes large. From (4.170) it is observed that the standard deviation is slightly larger than 1 particularly for small ν. Hence, the dispersion about the mean is somewhat larger than in the standard normal case.

The sampling distribution of the sample mean when the standard deviation is estimated by (2.5) can shown to be a t-distribution as follows. Consider the random variable:

$$\frac{m_X - \mu_X}{s_X/\sqrt{N}} = \left(\frac{m_X - \mu_X}{\sigma_X/\sqrt{N}}\right)\frac{\sigma_X}{s_X} = \left(\frac{m_X - \mu_X}{\sigma_X/\sqrt{N}}\right)\frac{1}{\sqrt{\frac{\chi_\nu^2}{\nu}}} \ \text{ with : } \nu = N-1 \qquad (4.171)$$

The first part of the last term is a standard normal variate, whereas the second part, which followed from (4.167), is the root of a χ^2-variate with ν = N-1 divided by ν. Hence the expression is a T_ν –variate with ν = N-1 degrees of freedom:

$$\frac{m_X - \mu_X}{s_X/\sqrt{N}} = T_\nu \ \text{ with : } \nu = N-1 \qquad (4.172)$$

It will be shown in the next sub-section that the t-distribution is related to the Fisher F-distribution.

4.7.4 Fisher's F-distribution

Let X and Y de independent random variables, both distributed as χ^2 with respectively ν_1 and ν_2 degrees of freedom, then the random variable F defined by:

$$F = \frac{X/\nu_1}{Y/\nu_2} \qquad (4.173)$$

has a so called F-distribution, which probability density function reads:

$$h_F(f) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}\frac{f^{(\nu_1/2)-1}}{(1+\frac{\nu_1}{\nu_2}f)^{(\nu_1+\nu_2)/2}} \qquad (4.174)$$

With the definition of the beta function $B(\alpha,\beta)$:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad (4.175)$$

equation (4.174) may also be written as:

$$h_F(f) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}}{B(\frac{\nu_1}{2},\frac{\nu_2}{2})}\frac{f^{(\nu_1/2)-1}}{(1+\frac{\nu_1}{\nu_2}f)^{(\nu_1+\nu_2)/2}} \qquad (4.176)$$

The pdf is shown in Figure 4.43:

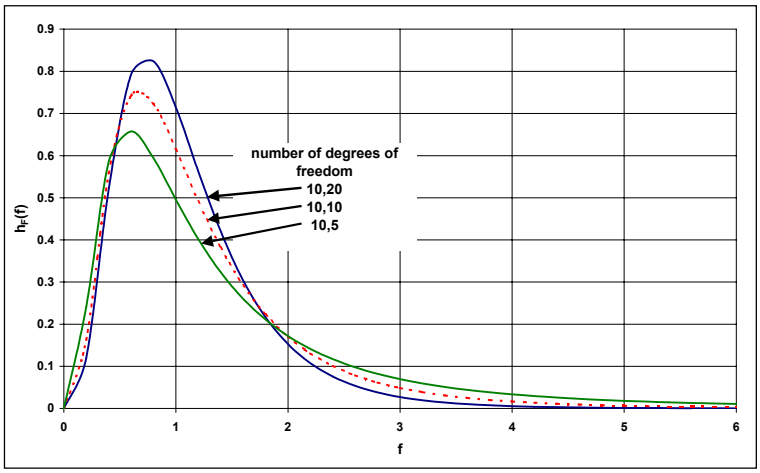

Figure 4.43:
Fisher F-probability density function for various degrees of freedom

The F-distribution is also called the variance-ratio distribution as from the definition of the F-variable (4.173) combined with (4.167) can be observed. Hence, if we consider m respectively n observations from two standard normal random variables Z_1 and Z_2 with variances σ_1^2 and σ_2^2 estimated according to (2.5) by s_1^2 and s_2^2 then the ratio:

$$\frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} = F_{m-1,n-1} \qquad (4.177)$$

has an F-distribution with (m-1,n-1) degrees of freedom. The F-distribution is thus particularly suited for variance ratio tests. From a comparison of (4.173) with (4.167) it is observed that the root of an F-variate with $(1,\nu)$ degrees of freedom has a Student t-distribution

5 Estimation of Statistical Parameters

5.1 General

To apply the theoretical distribution functions dealt with in the previous chapter the following steps are required:

1. Investigate the homogeneity of the data series, subjected to frequency analysis
2. Estimate the parameters of the postulated theoretical frequency distribution
3. Test the goodness of fit of the theoretical to the observed frequency distribution

In this chapter the second step will be dealt with. The objective of representing the observed frequency distribution by a theoretical one is to increase its mathematical tractability, and to facilitate extrapolation. The procedure in itself is no more than curve fitting. It involves the estimation of the parameters of a theoretical distribution function based on a sample from the population. It implies that the sample values of the parameters are stochastic variables themselves with a frequency distribution, called the sampling distribution as discussed in Chapter 4. The parameters can be estimated in various ways including:

1. Graphical method, and
2. Analytical methods, like:
    - Method of moments

- Maximum likelihood method
- Method of least squares
- Mixed moment-maximum likelihood method, etc.

Estimation error

The parameters estimated with the above methods differ. To compare the quality of different estimators of a parameter, some measure of accuracy is required. The following measures are in use:

- mean square error and root mean square error
- error variance and standard error
- bias
- efficiency
- consistency

Mean square error

A measure for the quality of an estimator is the **mean square error,** mse. It is defined by:

$$\text{mse} = E[(\phi - \Phi)^2] \tag{5.1}$$

where ϕ is an estimator for Φ.

Hence, the mse is the average of the squared differences between the sample value and the **true** value. Equation (5.1) can be expanded to the following expression:

$$\text{mse} = E[(\phi - E[\phi])^2] + E[(E[\phi] - \Phi)^2] \tag{5.2}$$

Since:

$$E\left[(\phi - E[\phi])^2\right] = \sigma_\phi^2 \tag{5.3}$$

and:

$$E[(E[\phi] - \Phi)^2] = b_\phi^2 \tag{5.4}$$

it follows that:

$$\text{mse} = \sigma_\phi^2 + b_\phi^2 \tag{5.5}$$

The mean square error is seen to be the sum of two parts:

- the first term is the **variance** of ϕ, equation (5.3), i.e. the average of the squared differences between the sample value and the **expected** mean value of ϕ based on the sample values, which represents the **random** portion of the error, and
- the second term of (5.5) is the square of the **bias** of ϕ, equation (5.4), describing the systematic deviation of expected mean value of ϕ from its true value Φ, i.e. the **systematic** portion of the error.

Note that if the bias in ϕ is zero, then mse = σ_ϕ^2. Hence, for **unbiased** estimators, i.e. if systematic errors are absent, the mean square error and the variance are equivalent. If mse(ϕ_1) < mse (ϕ_2) then ϕ_1 is said to be **more efficient** than ϕ_2 with respect to Φ.

Root mean square error

Instead of using the mse it is customary to work with its square root to arrive at an error measure, which is expressed in the same units as Φ, leading to the **root mean square (rms) error**:

$$\text{rmse} = \sqrt{E[(\phi - \Phi)^2]} = \sqrt{\sigma_\phi^2 + b_\phi^2} \tag{5.6}$$

Standard error

When discussing the frequency distribution of statistics like of the mean or the standard deviation, for the standard deviation σ_ϕ the term **standard error** is used, e.g. standard error of the mean and standard error of the standard deviation, etc.

$$\sigma_\phi = \sqrt{E[(\phi - E[\phi])^2]} \tag{5.7}$$

In Table 5.1, a summary of unbiased estimators for moment parameters is given, together with their standard error. With respect to the latter it is assumed that the sample elements are **serially uncorrelated**. If the sample elements are serially correlated a so-called **effective number of data** N_{eff} has to be applied in the expressions for the standard error in Table 5.1

Consistency

If the probability that ϕ approaches Φ becomes unity if the sample becomes large then the estimator is said to be consistent or asymptotically unbiased:

$$\lim_{n\to\infty} \text{Prob}(|\varphi - \Phi| > \varepsilon) = 0 \text{ for any } \varepsilon \tag{5.8}$$

To meet this requirement it is sufficient to have a zero mean square error in the limit for $n \to \infty$.

5.2 Graphical estimation

In graphical estimations, the variate under consideration is regarded as a function of the standardised or reduced variate with known distribution. With a properly chosen probability scale a linear relationship can be obtained between the variate and the reduced variate representing the transformed probability of non-exceedance. Consider for this the Gumbel distribution. From (4.108) it follows:

$$x = x_0 + \beta z \tag{5.9}$$

According to the Gumbel distribution the reduced variate z is related to the non-exceedance probability by:

$$z = -\ln(-\ln(F_X(x)) \tag{5.10}$$

To arrive at an estimate for x_0 and β we plot the ranked observations x_i against z_i by estimating the non-exceedance probability of x_i, i.e. F_i. The latter is called the plotting position of x_i, i.e. the probability to be assigned to each data point to be plotted on probability paper. Basically, appropriate plotting positions depend on the distribution function one wants to fit the observed distribution function to. A number of plotting positions has been proposed, which is summarised in Table 5.4.To arrive at an unbiased plotting position for the Gumbel distribution Gringorten's plotting position has to be applied, which reads:

$$F_i = \frac{i - 0.44}{N + 0.12} \qquad (5.11)$$

This non-exceedance frequency is transformed into the reduced variate z_i by using (5.10). If the data x_i are from a Gumbel distribution then the plot of x_i versus z_i will produce approximately a straight line. The slope of the line gives an estimate for the parameter β and the intercept is x_0. Hence the steps involved are as follows:

1. Rank the observations in ascending order, i = 1 is the smallest and i = N the largest
2. Compute the non-exceedance frequency F_i of x_i using (5.11)
3. Transform F_i into z_i using equation (5.10)
4. Plot x_i versus z_i and draw a straight line through the points
5. Estimate the slope of the line and the intercept at z=0 to get estimates for β and the intercept is x_0

The same steps apply to other frequency distributions, though with different plotting positions.

Parameter	Estimator	Standard error	Remarks
Mean	$m_Y = \frac{1}{N}\sum_{i=1}^{N} y_i$	$\sigma_{m_Y} = \frac{\sigma_Y}{\sqrt{N}}$	The sampling distribution of m_Y is very nearly normal for N>30, even when the population is non-normal. In practice σ_Y is not known and is estimated by s_Y. Then the sampling distribution of m_Y has a Student distribution, with N-1 degrees of freedom
Variance	$s_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - m_Y)^2$	$\sigma_{s_Y^2} = \sqrt{\frac{2}{N}}\sigma_Y^2$	Expression applies if the distribution of Y is approximately normal. The sampling distribution of s_Y^2 is nearly normal for N>100. For small N the distribution of s_Y^2 is chi-square (χ^2), with N-1 degrees of freedom
Standard deviation	$s_Y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - m_Y)^2}$	$\sigma_{s_Y} = \frac{\sigma_Y}{\sqrt{2N}}$	The remarks made for the standard error of the variance apply here as well
Coefficient of variation	$\hat{CV}_Y = \frac{s_Y}{m_Y}$ Sample value of CV_Y limited to: $CV_Y < \sqrt{(N-1)}$	$\sigma_{\hat{CV}} = \frac{\frac{\sigma_Y}{\mu_Y}}{\sqrt{2N}}\sqrt{1+2\left(\frac{\sigma_Y}{\mu_Y}\right)^2}$	This result holds if Y being normally or nearly normally distributed and N>100.
Covariance	$\hat{C}_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)(y_i - m_Y)$		
Correlation coefficient	$r_{XY} = \frac{C_{XY}}{s_X s_Y}$	$\sigma_W = \frac{1}{\sqrt{N-3}}$ where $W = \frac{1}{2}\ln\left(\frac{1+r_{XY}}{1-r_{XY}}\right)$	Rather then the standard error of r_{XY} the standard error of the transformed variable W is considered. The quantity W is approximately normally distributed for N>25.
Lag one auto-correlation coefficient	$r_{YY}(1) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}(y_i - m_Y)(y_{i+1} - m_Y)}{s_Y^2}$	as for r_{XY} above	
Skewness	$g_Y = \frac{\frac{N}{(N-1)(N-2)}\sum_{i=1}^{N}(y_i - m_Y)^3}{s_Y^3}$ Skewness limited to: $g_Y < \frac{N-2}{\sqrt{N-1}}$	$\sigma_{g_Y} = \sqrt{\frac{6}{N}}$	A reasonably reliable estimate of the skewness requires a large sample size. Standard error applies if Y is normally distributed.
Quantiles	1. first rank the sample values in ascending order: $y_{(i)} < y_{(i+1)}$ 2. next assign to each ranked value a non-exceedance probability i/(N+1) 3. then interpolate between the probabilities to arrive at the quantile value $\hat{y}_p$ of the required non-exceedance level	$\sigma_{\hat{y}_p} = \frac{1}{f_Y(y_p)}\sqrt{\frac{p(1-p)}{N}}$ $\sigma_{\hat{y}_p} = \frac{\beta}{\sqrt{N}}\sigma_Y$	The denominator is derived from the pdf of Y. If Y is normally distributed then the standard error of the quantile is determined by the second expression. The coefficient β depends on the non-exceedance probability p. For various values of p the value of β can be obtained from Table 5.2.

Table 5.1: Estimators of sample parameters with their standard error

p	0.5	0.4/ 0.6	0.3/ 0.7	0.25/0.75	0.2/ 0.8	0.15/0.85	0.1/0.9	0.05/0.95
β	1.253	1.268	1.318	1.362	1.428	1.531	1.709	2.114

Table 5.2: β(p) for computation of $\sigma_{\hat{y}_p}$ of quantiles if Y is normally distributed

Example 5.1: Graphical estimation of distribution parameters

Above procedure is shown for annual maximum river flows of the Meuse river at Chooz for the period 1968-1997 presented in Example 4.12. In Table 5.3 the peak flows are presented in Column 2. In Column 4 the ranked discharges are presented in ascending order.

Subsequently the non-exceedance frequency F_i of x_i is presented in Column 5, derived from equation (5.11), whereas in the last column the reduced variate z_i referring to the non-exceedance frequency F_i.

Year	Q_{max}	Rank	x_i	Freq	z_i	Year	Q_{max}	Rank	x_i	Freq	z_i
1	2	3	4	5	6	1	2	3	4	5	6
1968	386	1	274	0.019	-1.383	1983	1199	16	685	0.517	0.415
1969	910	2	295	0.052	-1.085	1984	675	17	690	0.550	0.514
1970	550	3	386	0.085	-0.902	1985	760	18	735	0.583	0.617
1971	274	4	406	0.118	-0.759	1986	735	19	760	0.616	0.725
1972	468	5	406	0.151	-0.635	1987	780	20	780	0.649	0.840
1973	406	6	423	0.185	-0.524	1988	660	21	785	0.683	0.963
1974	615	7	468	0.218	-0.421	1989	690	22	795	0.716	1.096
1975	295	8	491	0.251	-0.324	1990	1080	23	840	0.749	1.241
1976	795	9	550	0.284	-0.230	1991	491	24	860	0.782	1.404
1977	685	10	615	0.317	-0.138	1992	1135	25	910	0.815	1.589
1978	680	11	635	0.351	-0.047	1993	1510	26	1080	0.849	1.807
1979	785	12	642	0.384	0.043	1994	1527	27	1135	0.882	2.073
1980	635	13	660	0.417	0.134	1995	406	28	1199	0.915	2.421
1981	860	14	675	0.450	0.226	1996	642	29	1510	0.948	2.934
1982	840	15	680	0.483	0.319	1997	423	30	1527	0.981	3.976

Table 5.3: Annual maximum river flows of Meuse river at Chooz, period 1968-1997

The Columns 6 and 4 are plotted in Figure 5.1. It is observed that the points are located on a straight line, which indicates that the Gumbel distribution is applicable to data set of annual maximum riverflows in this case. The slope of the line is estimated at 1200/4.85 = 247 and the intercept at z = 0 is about 590 m^3/s, which are the estimates for β and x_0 respectively.

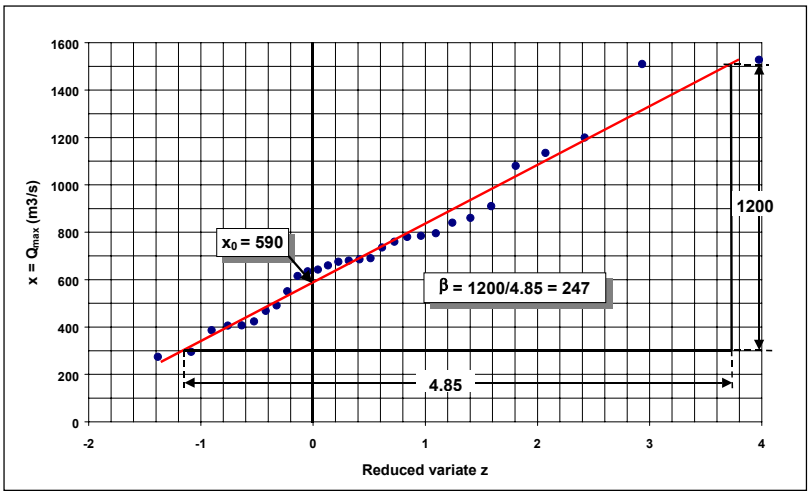

Figure 5.1: Application of graphical estimation method to annual maximum river flows of Meuse river at Chooz, period 1968-1997

In Chapter 4 Example 4.12 the parameters were estimated using the maximum likelihood method (MLM), which gave estimates for β and x_0 respectively of 238 and 591 m^3/s. For a 100 year return period flood (T = 100 years, i.e. $F_X(x) = 1 - 1/100 = 0.99$ or $z = -\ln(-\ln(0.99))=4.60$) the quantile $x_{T=100}$ becomes with the two methods using (5.9):

Graphical method: $x_{100} = 590 + 247 \times 4.60 = 1726$ m^3/s

MLM: $x_{100} = 591 + 238 \times 4.60 = 1686$ m^3/s

It is observed that the difference between the methods in this case is very small.

There is in the graphical method, however, a strong subjective element. Different analysts may obtain different results. This method is therefore not suitable for final design calculations. Plotting of the observed frequency distribution with the fitted one is extremely important. Before accepting a theoretical frequency distribution to be applicable to an observed frequency distribution inspection of the frequency plot is a must. Such a comparison gives you a visual impression about the goodness of fit particularly at the lower and upper end of the curve, something a statistical test does not give. In this respect it is of importance to apply the appropriate plotting position for each of the frequency distributions to arrive at an unbiased plotting position.

Plotting positions

Defining the plotting position for each data point, when put in ascending order, by:

$$F_i = \frac{i - b}{N - 2b + 1} \tag{5.12}$$

where: F_i = non-exceedance frequency of x_i ranked in ascending order
$i = i^{th}$ element in ranked sequence in ascending order
N = number of data in series
b = parameter dependent on type of distribution

Cunnane (1978) investigated various plotting positions that can be derived from (5.12) by assuming an appropriate value for b. Two criteria were used:

- unbiasedness, which implies that for a large number of equally sized samples the average of the plotted points for each i will fall on the theoretical line
- minimum variance, i.e. the variance of the plotted point about the theoretical line is minimum.

It appears that the often-used Weibull plotting position with b = 0 gives a biased result, plotting the largest values at a too low return period. Some of his results and those of NERC (1975) are summarised in Table 5.4.

Name of formula	b	distribution	remarks
Hazen	0.5	-	For i = N: T = 2N
Weibull	0	-	biased
Blom	3/8	N, LN-2, LN-3, G-2 for large γ	LP-3: for γ_1>0 b>3/8 and γ_1<0 b<3/8
Chegodayev	0.3	various	Overall compromise
Gringorten	0.44	EV-1, E-1, E-2, G-2	
NERC	2/5	G-2, P-3	Compromise plotting position
Tukey	1/3	-	

Table 5.4: Plotting position formula (Cunnane, 1978; NERC, 1975)
In HYMOS the parameter b can be set to the requirement; the default value is b = 0.3.

5.3 Parameter estimation by method of moments

The method of moments makes use of the fact that if all the moments of a frequency distribution are known, then everything about the distribution is known. As many moments as there are parameters are needed to define the distribution. The frequency distributions discussed in Chapter 4 contain at maximum four parameters, hence the first four moments,

generally represented by the mean, variance, skewness and kurtosis, are at maximum required to specify the distribution and to derive the distribution parameters. Most distributions, however, need only one, two or three parameters to be estimated. It is to be understood that the higher the order of the moment the larger the standard error will be.

In HYMOS the unbiased estimators for the mean, variance, skewness and kurtosis as presented by equations (2.3), (2.5) or (2.6), (2.8) and (2.9) are used, see also Table 5.1. Substitution of the required moments in the relations between the distribution parameters and the moments will provide the moment estimators:

- Normal distribution: the two parameters are the mean and the standard deviation, which follow from (2.3) and (2.6) immediately
- LN-2: equations (2.3) and (2.6) substituted in (4.28) and (4.29)
- LN-3: equations (2.3), (2.6) and (2.8) substituted in (4.31) to (4.34)
- G-2: equations (2.3) and (2.6) substituted in (4.61) and (4.62)
- P-3: equations (2.3), (2.6) and (2.8) substituted in (4.71) to (4.73)
- EV-1:equations (2.3) and (2.6) substituted in (4.115) and (4.116)

For all other distributions the method of moments is not applied in HYMOS.

Biased-unbiased

From (2.5) it is observed that the variance is estimated from:

$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)^2 \qquad (2.5)$$

The denominator (N-1) is introduced to obtain an unbiased estimator. A straightforward estimator for the variance would have been:

$$\hat{s}_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m_X)^2 \qquad (5.13)$$

The expected value of this estimator, in case the x_i's are *independent*, is:

$$E[\hat{s}_X^2] = \frac{1}{N}\sum_{i=1}^{N}E[(x_i - m_X)^2] = E[((x_i - \mu_X) - (m_X - \mu_X))^2] = \sigma_X^2 - \frac{\sigma_X^2}{N} = \left(\frac{N-1}{N}\right)\sigma_X^2 \qquad (5.14)$$

From equation (5.14) it is observed that although the estimator is consistent, it is biased. Hence, to get an unbiased estimator for σ_X^2 the moment estimator should be multiplied by N/(N-1), which leads to (2.5)

Remark

The method of moments provides a simple procedure to estimate distribution parameters. For small sample sizes, say N < 30, the sample moments may differ substantially from the population values. Particularly if third order moments are being used to estimate the parameters, the quality of the parameters will be poor if the sample size is small. In such cases single outliers will have a strong effect on the parameter estimates.

Probability weighted moments and L-moments

The above method of moments is called Product Moments. The negative effects the use of higher moments have on the parameter estimation is eliminated by making use of L-moments, which are linear functions of **probability weighted moments** (PWM's). Probability weighted moments are generally defined by:

$$M_{p,r,s} = E\left[X^p \{F_X(x)\}^r \{1 - F_X(x)\}^s\right]$$

(5.15)

By choosing p=1 and s=0 in (5.15) one obtains the r^{th} PWM, which reads:

$$\beta_r = E\left[X\{F_X(x)\}^r\right] = \int_{-\infty}^{\infty} x\{F_X(x)\}^r f_X(x)dx \tag{5.16}$$

Comparing this expression with the definition of moments in (3.23) it is observed that instead of raising the variable to a power ≥ 1 now the cdf is raised to a power ≥ 1. Since the latter has values < 1, it is observed that these moments are much less sensitive for outliers, which in the case of product moments strongly affect the moments and hence the parameters to be estimated.

L-moments are developed for order statistics. Let the X_i's be independent random variables out of a series of sample of size N, which are put in ascending order:

$X_{1:N} < X_{2:N} < \ldots . < X_{N:N}$

then $X_{i:N}$ is the i^{th} largest in a random sample of N, and is known as the **i^{th} order statistic**. L-moments are used to characterize the distribution of order statistics. In practice the first four L-moments are of importance:

$$\begin{aligned}
\lambda_1 &= E[X] \\
\lambda_2 &= \frac{1}{2}\{E[X_{2:2}] - E[X_{1:2}]\} \\
\lambda_3 &= \frac{1}{3}\{E[X_{3:3}] - 2E[X_{2:3}] + E[X_{1:3}]\} \\
\lambda_4 &= \frac{1}{4}\{E[X_{4:4}] - 3E[X_{3:4}] + 3E[X_{2:4}] - E[X_{1:4}]\}
\end{aligned} \tag{5.17}$$

The first moment is seen to be the mean, the second a measure of the spread or scale of the distribution, the third a measure of asymmetry and the fourth a measure of peakedness. Dimensionless analogues to the skewness and kurtosis are (Metcalfe, 1997):

$$\begin{aligned}
&L-\text{skewness}: \tau_3 = \frac{\lambda_3}{\lambda_2} \text{ with}: -1 < \tau_3 < 1 \\
&L-\text{kurtosis}: \tau_4 = \frac{\lambda_4}{\lambda_2} \text{ with}: \frac{1}{4}(5\tau_3^2 - 1) \le \tau_4 < 1
\end{aligned} \tag{5.18}$$

The relation between the L-moments and parameters of a large number of distributions are presented in a number of statistical textbooks. For some distributions they are given below (taken from Dingman, 2002):

- Uniform distribution

$$\begin{aligned}
\lambda_1 &= \frac{a+b}{2} \\
\lambda_2 &= \frac{b-a}{6} \\
\tau_3 &= 0 \\
\tau_4 &= 0
\end{aligned} \tag{5.19}$$

- Normal distribution

$$\begin{aligned}
\lambda_1 &= \mu_X \\
\lambda_2 &= \frac{\sigma_X}{\sqrt{\pi}} \\
\tau_3 &= 0 \\
\tau_4 &= 0.1226
\end{aligned}$$

(5.20)

- Gumbel

$$\begin{aligned}\lambda_1 &= x_0 + 0.5772\,\beta \\ \lambda_2 &= 0.6931\,\beta \\ \tau_3 &= 0.1699 \\ \tau_4 &= 0.1504\end{aligned} \tag{5.21}$$

So to estimate the parameters of a distribution estimates of L-moments are required. From (5.17) it is observed that to estimate the L-moments all possible combinations of samples of size 2, 3 and 4 have to be selected to arrive at the expected value of the various order statistics. This is a rather cumbersome exercise. However, the L-moments can be related to the probability weighted moments as follows:

$$\begin{aligned}\lambda_1 &= \beta_0 \\ \lambda_2 &= 2\beta_1 - \beta_0 \\ \lambda_3 &= 6\beta_2 - 6\beta_1 + \beta_0 \\ \lambda_4 &= 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0\end{aligned} \tag{5.22}$$

The sample estimates of the probability weighted moments follow from the ordered set of data:

$$\begin{aligned}b_0 &= \frac{1}{N}\sum_{i=1}^{N} x_{i:N} = m_X \\ b_1 &= \frac{1}{N(N-1)}\sum_{i=2}^{N}(i-1)x_{i:N} \\ b_2 &= \frac{1}{N(N-1)(N-2}\sum_{i=3}^{N}(i-1)(i-2)x_{i:N} \\ b_3 &= \frac{1}{N(N-1)(N-2)(N-3)}\sum_{i=4}^{N}(i-1)(i-2)(i-3)x_{i:N}\end{aligned} \tag{5.23}$$

Example 5.1: continued

The L-moments method is applied to the annual maximum river flows of Meuse river at Chooz. The computation of the probability weighted moments is presented in Table 5.5. Note that first the data are ordered. The ordered series is presented in Column 2. In Column 3 the numerical values of $(i-1)x_{i:N}$ is presented, which is the sum term in the derivation of b_1; similarly the columns 4 and 5 contain the sum-terms for the derivation of b_2 and b_3. The values in the columns are summed and subsequently divided by N, N(N-1), N(N-1)(N-2) and N(N-1)(N-2)(N-3) respectively to arrive at the estimates for the probability weighted moments b_0, b_1, b_2 and b_3, according to equation (5.23).

Rank	Q-max	C-b1	C-b2	C-b3
1	274			
2	295	295		
3	386	772	772	
4	406	1217	2435	2435
5	406	1624	4872	9744
6	423	2117	8467	25402
7	468	2808	14040	56160
8	491	3437	20622	103110
9	550	4400	30800	184800
10	615	5535	44280	309960
11	635	6350	57150	457200
12	642	7066	70656	635908
13	660	7920	87120	871200
14	675	8775	105300	1158300
15	680	9520	123760	1485120
16	685	10275	143850	1870050
17	690	11040	165600	2318400
18	735	12495	199920	2998800
19	760	13680	232560	3720960
20	780	14820	266760	4534920
21	785	15700	298300	5369400
22	795	16695	333900	6344100
23	840	18480	388080	7761600
24	860	19780	435160	9138360
25	910	21840	502320	11051040
26	1080	27000	648000	14904000
27	1135	29511	737776	17706624
28	1199	32373	841698	21042450
29	1510	42270	1141295	29673680
30	1527	44295	1240273	33487375
Sum	21898	392090	8145767	177221098
Parameters	b_0	b_1	b_2	b_3
	729.92	450.68	334.39	269.45

Table 5.5: Annual maximum river flows of Meuse river at Chooz, period 1968-1997

From the probability weighted moments one can derive the L-moments, with the aid of equation (5.22) as follows. If the estimates for λ are indicated by L then:

$L_1 = b_0 = 729.92$
$L_2 = 2b_1 – b_0 = 2x450.68 – 729.92 = 171.44$
$L_3 = 6b_2 – 6b_1 + b_0 = 6x334.39 – 6x450.68 + 729.92 = 32.18$
$L_4 = 20b_3 - 30b_2 + 12b_1 – b_0 = 20x269.45 – 30x334.39 + 12x450.68 – 729.92 = 35.54$

The parameters of the Gumbel distribution can be obtained through equation (5.21):

$$\hat{\beta} = \frac{L_2}{0.6931} = \frac{171.44}{0.6931} = 247$$

$$\hat{x}_0 = L_1 - 0.5772\hat{\beta} = 729.92 - 0.5772x247.35 = 587$$

$$\hat{\tau}_3 = \frac{L_3}{L_2} = \frac{32.18}{171.44} = 0.19$$

$$\hat{\tau}_4 = \frac{L_4}{L_2} = \frac{35.54}{171.44} = 0.21$$

With the product moment method one obtains for the two parameters respectively 244 and 589 and with the MLM-method 238 and 591. Hence the 100-year flood derived with the various methods becomes:

Product moments: 589 + 244x4.6 = 1711 m^3/s

L-moments: 587 + 247x4.6 = 1723 m^3/s
MLM-method: 591 + 238x4.6 = 1686 m^3/s

The 100-year flood values are seen to be very close to each other. The values for the L-skewness and L-kurtosis of 0.19 and 0.21, respectively, are close to their theoretical values of 0.17 and 0.15 for the Gumbel distribution, which shows that the distribution is an appropriate model for the data set. Charts have been designed where L-skewness and L-kurtosis are plotted against each other for various distributions to guide the selection of a distribution, see also Figure 5.2.

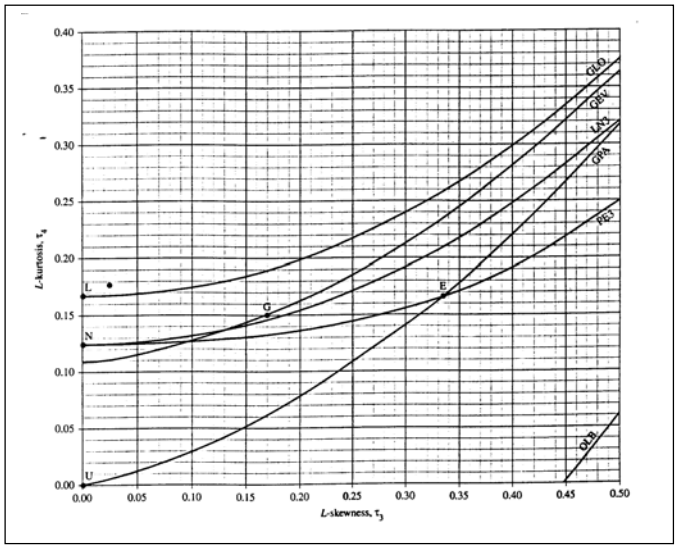

Figure 5.2:
L-moment diagram

(source: Dingman, 2002)

Note

By definition of the probability weighted moments and by close observation of Table 5.5 it is noticed that in the estimation of the probability weighted moments larger weight is given to the higher ranked values in the data set. Hence, the method is biased towards the larger values, particularly when more than 2 parameters have to be estimated. So, though the method is less sensitive to outliers than the product moment method, its application also has its drawbacks.

5.4 Parameter estimation by maximum likelihood method

The Maximum Likelihood method (MLM) was developed by R.A. Fisher in 1922. It is based on the idea that the best estimators for a (set of) parameter(s) are those, which give the greatest probability that precisely the sample series is obtained with the set of parameters. Let X be a random variable with pdf $f_X(x)$, with parameters α_1, α_2, …, α_k. The sample taken out of X is x_i, i=1, 2, …, N. Making the basic assumption that the sample values are **independent and identically distributed**, then with the parameter set α the probability that the random variable will fall in the interval including x_i is $f_X(x_I|\alpha)dx$. So, the joint probability of the occurrence of the sample set x_i, i=1, 2, …, N is, in view of their independence, equal to the product:

$$f_X(x_1 \mid \alpha)dx.f_X(x_2 \mid \alpha)dx.....f_X(x_N \mid \alpha)dx = \left(\prod_{i=1}^{N} f_X(x_i \mid \alpha)\right)dx^N$$

Since all dx are equal, maximising the joint probability simply implies the maximisation of the product:

$$L(x \mid \alpha) = \prod_{i=1}^{n} f_X(x_i \mid \alpha) \tag{5.24}$$

L is called the likelihood function. Then the best set of parameters α are those which maximise L. Hence the estimators for the parameters α_1, α_2, …, α_k are found from:

$$\frac{\partial L(x \mid \alpha)}{\partial \alpha_i} = 0 \text{ for : } i = 1, 2, 3, ..., k \tag{5.25}$$

The estimators obtained in this way are called Maximum Likelihood estimators. Instead of using the likelihood function itself it is usually more convenient to maximise its logarithm in view of the many distributions of the exponential type. Therefore instead of (5.25) the log-likelihood function lnL is usually maximised:

$$\frac{\partial \ln L(x \mid \alpha)}{\partial \alpha_i} = 0 \text{ for : } i = 1, 2, 3, ..., k \tag{5.26}$$

This has the advantage of replacing the products by sum-terms.

Application to lognormal distribution

The procedure will be shown for getting estimators for the lognormal-2 distribution, LN-2. From (4.26) the likelihood function for a sample x_i, i=1, 2, …, N reads:

$$L(x \mid \mu_Y, \sigma_Y) = \prod_{i=1}^{N} \frac{1}{x_i \sigma_Y \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln x_i - \mu_Y}{\sigma_Y} \right)^2 \right) \tag{5.27}$$

Hence, the log-likelihood function reads:

$$\ln L(x \mid \mu_Y, \sigma_Y) = -\sum_{i=1}^{N} (\ln x_i) - \frac{N}{2} \ln \sigma_Y^2 - \frac{N}{2} \ln(2\pi) - \frac{1}{2\sigma_Y^2} \sum_{i=1}^{N} (\ln x_i - \mu_Y)^2 \tag{5.28}$$

$$\frac{\partial \ln L}{\partial \mu_Y} = -\frac{1}{2\sigma_Y^2} \sum_{i=1}^{N} 2(\ln(x_i - \mu_Y)(-1) = \frac{1}{\sigma_Y^2}\left(\sum_{i=1}^{N} (\ln x_i) - N\mu_Y \right) = 0 \tag{5.29}$$

$$\frac{\partial \ln L}{\partial \sigma_Y^2} = -\frac{N}{2}\frac{1}{\sigma_Y^2} - \frac{1}{2}\left(\sum_{i=1}^{N} (\ln x_i - \mu_Y \right)^2 \left(-\frac{1}{(\sigma_Y^2)^2} \right) = 0$$

From above equations the MLM estimators for μ_Y and σ_Y^2 become respectively:

$$\hat{\mu}_Y = \frac{1}{N} \sum_{i=1}^{N} \ln x_i \tag{5.30}$$

$$\hat{\sigma}_Y^2 = \frac{1}{N} \sum_{i=1}^{N} (\ln x_i - \hat{\mu}_Y)^2 \tag{5.31}$$

From (5.30) and (5.31) it is observed that the MLM estimators are equivalent to the first moment about the origin and the second moment about the mean of ln(x). In a similar manner for the 3-parameter lognormal distribution the estimators for the distribution parameters can be derived, however, at the expense of more complicated equations. As is discussed in Sub-section 5.6 mixed moment-maximum likelihood estimators are preferred when a third parameter (generally the shift or location parameter) is to be estimated particularly when the sample sizes are small.

For the other distribution functions the MLM procedure can also easily be developed along the same lines as discussed for the lognormal distribution, though their solutions are sometimes cumbersome. Reference is made to the HYMOS manual for a description of the formulas used.

5.5 Parameter estimation by method of least squares

The graphical estimation procedure explained in Subsection 5.3 by drawing a line through the data points of the variable x and the reduced variate z can also be done applying linear regression, with z the independent variable and x the dependent variable. The parameters then follow from a minimisation of the sum of squared differences. Such a procedure does not suffer from subjectivity as the graphical method does. The procedures for regression analysis are dealt with in detail in Module 37.

Example 5.1: continued

The annual maximum flows presented in column 4 of Table 5.3 are regressed against the reduced variate z shown in column 6. From linear regression the following estimates for the parameters are obtained (with standard error): x_0 = 589 $\pm$ 10.8 and β = 250 $\pm$ 8.0, values which are very close to those obtained from the graphical method.

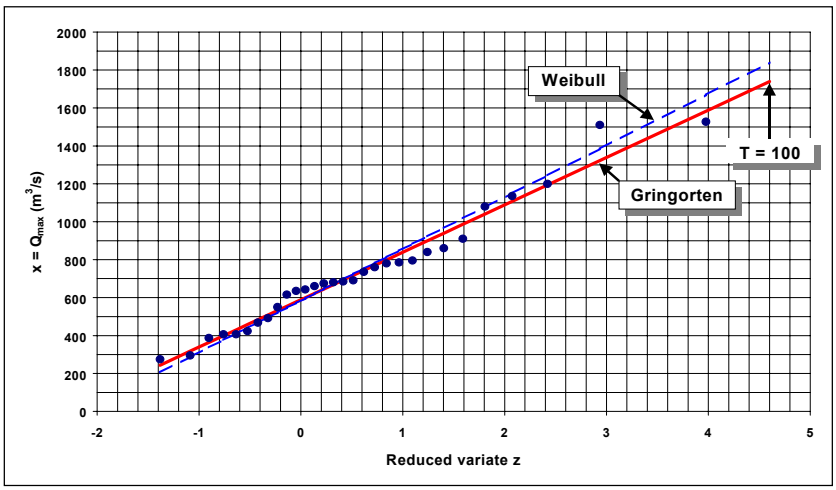

Figure 5.3: Fitting annual maximum flows by regression on reduced variate

If instead of Gringorten's plotting position Weibull's plotting position would have been used, the result would have been: x_0 = 584 $\pm$ 11.3 and β = 273 $\pm$ 9.2. The T=100 year floods from these procedures would have been for:

- Gringorten: x_{100} = 1739 m^3/s and
- Weibull: x_{100} = 1840 m^3/s

The difference with the MLM estimate are respectively: 3% and 9%. It is observed that the Weibull procedure leads to considerably higher quantile values. This is due to the fact that this method assigns a relatively low return period to the largest values. As a result, the slope of the regression line (i.e. β) will be larger, and so will be the quantiles.

5.6 Parameter estimation by mixed moment-maximum likelihood method

For frequency distributions with a location parameter often the MLM method performs poorly particularly when the sample series is small, like for LN-3 and P-3. In such cases estimating one parameter from a moment relation and the rest with the MLM procedure provides much better parameter estimators, as can be shown by means of Monte Carlo simulations.

The procedure will be shown for LN-3. For given location parameters the MLM estimators for μ_Y and σ_Y^2 become similar to (5.30) and (5.31) with x replaced by x-x_0 respectively:

$$\hat{\mu}_Y = \frac{1}{N}\sum_{i=1}^{N} \ln(x_i - x_0) \tag{5.32}$$

$$\hat{\sigma}_Y^2 = \frac{1}{N}\sum_{i=1}^{N} \left(\ln(x_i - x_0) - \hat{\mu}_Y\right)^2 \tag{5.33}$$

Next, the first moment relation for the lognormal distribution is taken, (4.27a), to arrive at a value for x_0:

$$x_0 = \hat{\mu}_X - \exp\left(\hat{\mu}_Y + \frac{\hat{\sigma}_Y^2}{2}\right) \tag{5.34}$$

The location parameter x_0 is solved iteratively from a modified form of (5.34) as follows:

$$g(x_0) = \hat{\mu}_Y + \frac{\hat{\sigma}_Y^2}{2} - \ln(\hat{\mu}_X - x_0) = 0 \tag{5.35}$$

For each value of x_0 the parameters μ_Y and σ_Y^2 are estimated by (5.32) and (5.33). Given an initial estimate of x_0, an improved estimate is obtained by means of the Newton-Raphson method:

$$x_{0,new} = x_{0,old} - \frac{g(x_{0,old})}{g'(x_{0,old})} \tag{5.36}$$

Since μ_Y and σ_Y^2 are also a function of x_0 it follows for the derivative $g'(x_{0,old})$:

$$g'(x_0) = \frac{dg}{dx_0} = (\hat{\mu}_Y - 1)\frac{1}{N}\sum_{i=1}^{N}(x_i - x_0)^{-1} + (\hat{\mu}_X - x_0)^{-1} \tag{5.37}$$

To speed up the computations, in HYMOS the expected value of $g'(x_{0,old})$ is calculated rather than computing $g'(x_{0,old})$ for each x_0:

$$E[g'(x_0)] = \frac{(\hat{\sigma}_Y^2 - 1)\exp(\hat{\sigma}_Y^2) + 1}{\hat{\mu}_X - x_0} \tag{5.38}$$

By substitution of (5.37) in (5.36) it follows for the improved estimate of x_0:

$$x_{0,new} = x_{0,old} - \frac{\left(\hat{\mu}_Y + \frac{\hat{\sigma}_Y^2}{2} - \ln(\hat{\mu}_X - x_0)\right).(\hat{\mu}_X - x_0)}{1 + (\hat{\sigma}_Y^2 - 1).\exp(\hat{\sigma}_Y^2)} \tag{5.39}$$

The iteration is continued till:

$$\left|x_{0,new} - x_{0,old}\right| < \varepsilon \text{ with : } \varepsilon = \frac{\mu_X - x_{min}}{1000} \tag{5.40}$$

The initial value of x_0 is taken as:

$$x_0 = x_{min} - 0.1(\mu_X - x_{min}) \tag{5.41}$$

Similar to this mixture of moment and MLM procedures, HYMOS provides mixed moment MLM estimators for the Pearson Type distributions. Reference is made to the HYMOS manual for the details.

5.7 Censoring of data

In some cases one wants to eliminate data from frequency analysis either at the upper end or at the lower end. Eliminating data from the frequency analysis at the upper end is called **right censoring** and eliminating data at the lower end is called **left censoring**. This is illustrated in Figure 5.3.

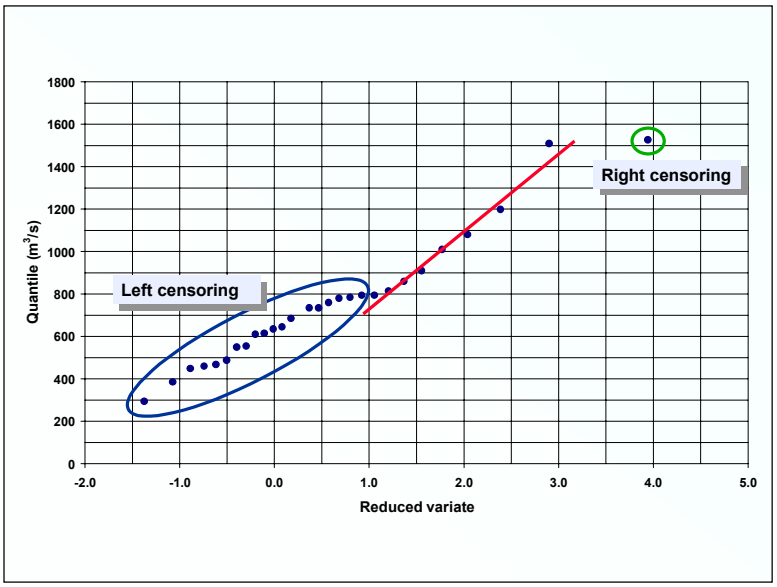

Figure 5.4:
Left and right censoring

With censoring, the **relative frequencies** attached to the remaining data is left **unchanged**. Hence, one performs frequency analysis on a reduced data set, but with frequency information from the **original** set. So the procedure is not the same as simply eliminating data from the data set and working with a reduced set, where the relative frequencies are determined based on the reduced series.

Right censoring may be required when there is evidence that the highest or a few of the highest values are unreliable (poorly measured extremes) or do have a return period which is believed to be much higher than one would expect based on the ordered data set. Left censoring may be required if the lower part of the ordered data set is not representative for the physics of the phenomena, which govern the higher part. Then, if one wants to extrapolate based on the higher values, the lower part can be censored, thereby leaving the relative frequencies of the higher ones intact. This procedure is often applied for analysis of river flow extremes, where the flow extremes refer to situations when the river stays inbank for the low peaks (lower part) and enters the flood plain with strong attenuation of the flood peaks (higher part). In such case the lower part will be steeper than the higher part (opposite to what is shown in Figure 5.3 !!).

In HYMOS censoring is possible for the Gumbel distribution. Great care is needed in applying censoring: there should be clear evidence that censoring is required.

5.8 Quantile uncertainty and confidence limits

Quantile uncertainty

The estimates for the distribution parameters involve estimation errors, and hence the same applies for the quantiles derived from it. The parameter uncertainties have to be translated to the uncertainty in the estimate of the quantile. The estimation error is used to draw the confidence limits about the estimated quantiles to indicate the likely range of the true value of the quantile. The procedure to derive the confidence limits will be illustrated for the quantile of a normally distributed random variable. From (4.23) the quantile x_p is given by:

$$x_p = \mu_X + \sigma_X.z_p \tag{5.42}$$

where: z_p = standard normal deviate corresponding to a non-exceedance probability p. The quantile is estimated by:

$$x_p = m_X + s_X.z_p \tag{5.43}$$

The parameters m and s are estimated by (2.3) and (2.6) respectively. The estimation variance of the quantile follows from:

$$\text{var}(x_{e,p}) = \text{var}\,(m_X + s_X.z_p) = \text{var}(m_X) + z_P^2\ \text{var}\,(s_X) + 2\ \text{cov}(m_X,s_X) \tag{5.44}$$

Since var $(m_X) = \sigma_X{}^2/N$, var $(s_X) \approx \sigma_X{}^2/(2N)$ (see Table 5.1) and for a normally distributed variable cov$(m_X, s_X) = 0$, the variance of x_p becomes approximately:

$$\text{var}(x_p) \approx \text{var}(m_x) + z_p^2\ \text{var}(s_x) = \frac{\sigma_x^2}{N}\left(1+\frac{1}{2}z_p^2\right) \tag{5.45}$$

Hence with σ_X replaced by s_X, the standard error of the quantile follows from:

$$s_{x_p} \approx s_x\sqrt{\frac{1}{N}\left(1+\frac{1}{2}z_p^2\right)} \tag{5.46}$$

The 100(1-α)% confidence limits for x_p then read:

$$x_{p,LCL} = x_p - z_{1-\alpha/2}s_X\sqrt{\frac{1}{N}\left(1+\frac{1}{2}z_p^2\right)} \quad x_{p,UCL} = x_p + z_{1-\alpha/2}s_X\sqrt{\frac{1}{N}\left(1+\frac{1}{2}z_p^2\right)} \tag{5.47}$$

The confidence limits express that the true quantile x_p falls within the interval $x_{p,LCL}$ and $x_{p,UCL}$ with a confidence of 100(1-α)%. The quantity 100(1-α)% is the **confidence level** and α is the **significance level**. From the limits shown in (5.47) it is observed that the confidence band about the quantile increases with z_p, i.e. the further away from the mean of the distribution the larger the uncertainty of the quantile becomes. Also the effect of the number of data is apparent from (5.47); a small number of data results in a large uncertainty for the quantile.

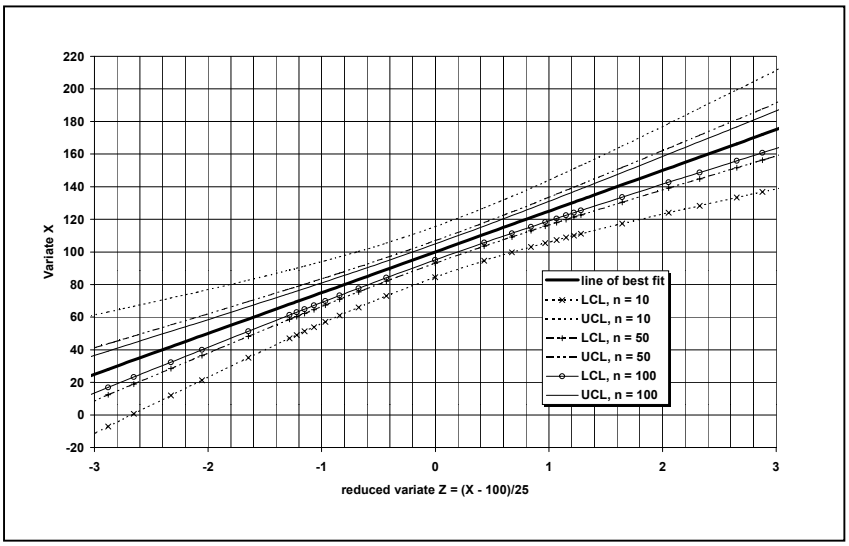

Figure 5.5: Fit by normal distribution (m_X= 100, s_X = 25) with 95% confidence limits for different length of data series

Uncertainty in the probability of the quantile

In the above we were looking at the standard error of the quantile for a given non-exceedance probability. One can also look at the uncertainty in the non-exceedance probability for a fixed value of x_p. From (5.42) it follows:

$$z_p = \frac{x_p - \mu_X}{\sigma_X} \tag{5.48}$$

Hence, the standard error of the reduced variate z_p becomes:

$$\sigma_{z_p} = \frac{\sigma_{x_p}}{\sigma_X} \text{ estimated by } s_{z_p} = \frac{s_{x_p}}{s_X} \tag{5.49}$$

The reduced variate z_p is approximately normally distributed with $N(z_p, \sigma_{zp})$. Hence, the confidence interval for p at a significance level α becomes $P_{LCL} = F_N(z_p - z_{1-\alpha/2}.\sigma_{zp})$ and $P_{UCL} = F_N(z_p - z_{1-\alpha/2}.\sigma_{zp})$, where F_N is the standard normal distribution function. The standard error σ_p of p for fixed x_p then becomes:

$$\sigma_p \approx f_N(z_p)\sigma_{z_p} \approx \left(\frac{1}{\sqrt{2\pi}}\exp(-\frac{z_p^2}{2})\right)s_{z_p} = \left(\frac{1}{\sqrt{2\pi}}\exp(-\frac{z_p^2}{2})\right)\frac{s_{x_p}}{s_X} \tag{5.50}$$

Example 5.2: Annual rainfall Vagharoli

Annual rainfall of station Vagharoli for the period 1978 –1997 is considered. After having tested the homogeneity of the series, the observed frequency distribution was fitted by the normal distribution, which should be applicable on basis of the conditions needed for a Gaussian distribution.

The result with HYMOS is presented in the table below. In the result first the basic statistics are presented. From the skewness and kurtosis being close to 0 and 3 respectively it is observed that the data are approximately normally distributed.

In the next part of the result a summary is presented of the ranked observations, including:

- In the 1st column the year number as from 1978 onward is presented for each ranked observation; e.g. the first row has year number 10 which means that this represents the

value of year (1978 –1) +10 i.e. 1987. The observation for the year 1978 is seen to be ranked as one but highest value.

- The 2nd column shows the ranked observations.
- The 3rd column gives the non-exceedance probability of the observation according to the observed frequency distribution, using the plotting position most appropriate for the normal distribution. According to Table 5.4, Blom's formula gives an unbiased plotting position for the normal distribution. For the first row (rank 1) the following value will then be obtained:

$$F_i = \frac{i - 3/8}{N + 1/4} \text{ becomes : } F_1 = \frac{1 - 3/8}{20 + 1/4} = \frac{0.625}{20.25} = 0.0309$$

- The 4th column gives the theoretical non-exceedance probability accepting the normal distribution with mean m = 877.3 and standard deviation 357.5. The reduced variate then reads:

$$z = \frac{x - m_X}{s_X} = \frac{x - 877.3}{357.5}$$

For the lowest ranked value (on the first row) it then follows:

$$z_1 = \frac{x_1 - 877}{357} = \frac{232 - 877.3}{357.5} = -1.805$$

From tables of the normal distribution one reads for z = 1.805 a non-exceedance probability of p = 0.9645. Hence the non-exceedance probability for z = -1.805 is in view of the symmetry of the normal distribution p_1 = 1 - 0.9645 = 0.0355. Using HYMOS it is not necessary to consult a statistical textbooks for the table of the normal distribution as it is included in the software under the option 'Statistical Tables'.

- The 5th column gives the return period, which is derived from the non-exceedance probability by:

$$T = \frac{1}{1 - F(x)} \text{ hence : } T_{x_1} = \frac{1}{1 - p_1} = \frac{1}{1 - 0.0355} = 1.037 \approx 1.04$$

The 6th column presents the standard error of the quantile x_p, derived from (5.46). Since we are discussing here observations, hence, there is no statistical uncertainty in it as such (apart from measurement errors). But the standard error mentioned here refers to the standard error one would have obtained for a quantile with the same value as the observation when derived from the normal distribution. It is a necessary step to derive the uncertainty in the non-exceedance probability presented in column 7. For the first row e.g. it then follows with (5.46):

$$s_{x_p} = s_X \sqrt{\frac{1}{N}\left(1 + \frac{z_p^2}{2}\right)} \text{ hence : } s_{x_{0.0355}} = 357 \sqrt{\frac{1}{20}\left(1 + \frac{(-1.805)^2}{2}\right)} = 357 \text{x} 0.363 = 129.6$$

- The standard error of the non-exceedance probability follows from (5.50):

$$s_p \approx \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_p^2}{2} \right) \right) \frac{s_{x_p}}{s_x} \text{ hence for the first row :}$$

$$s_{p_1} \approx \left(\frac{1}{\sqrt{2x3.14}} \exp \frac{(-1.805)^2}{2} \right) \frac{129.6}{357.5} = 0.0283$$

In the third part of the results the output from the goodness of fit tests are presented. This will be discussed in the next chapter.
In the last part of the results for distinct return periods and non-exceedance probabilities the quantiles are presented with their standard error and 100(1-α) = 95% confidence limits, which are also shown in the plot of the observed distribution fitted by the normal one in Figure 5.4. The values are obtained as follows:

- The 1st column presents the return period
- The 2nd column gives the non-exceedance probability associated with the return period in column 1
- In the 3rd column the quantile is given, which is derived from (5.43) for the reduced variate corresponding with the non-exceedance probability; this is derived from the inverse of the standard normal distribution. E.g. for T=100, p = 0.99, z_p =2.33 and the quantile follows from:

$$x_p = m_X + s_X z_p \text{ hence : } x_p = 877.3 + 357.5x2.33 = 1709.0 \text{ mm}$$

- In the 4th column the standard error of x_p is given which is obtained from (5.46)., e.g. for the T= 100 year event:

$$s_{x_p} \approx s_X \sqrt{\frac{1}{N}\left(1 + \frac{z_p^2}{2}\right)} = 357.5 \sqrt{\frac{1}{20}\left(1 + \frac{(2.33)^2}{2}\right)} = 153.9 \text{ mm}$$

- In the 5th and 6th column the lower and upper confidence limits for the quantile are given, which are derived from (5.47) in case of 95% confidence limits. In case e.g. 90% limits are used (hence α = 0.10 instead of 0.05) then in equation (5.47) the value 1.96 (p=1-α/2 =0.975) has to be replaced by 1.64 (p=1-α/2=0.95), values which can be obtained from the tables of the normal distribution or from the Statistical Tables option in HYMOS. It follows for the 100 year event:

$$x_{p,LCL} = x_p - 1.96 s_{x_p} = 1709 - 1.96x153.9 = 1407.3 \text{ mm}$$

$$x_{p,UCL} = x_p + 1.96 s_{x_p} = 1709 + 1.96x153.9 = 2010.7 \text{ mm}$$

Results by HYMOS:

```
Annual rainfall Vagharoli

Period 1978 - 1997

 Fitting the normal distribution function

 Number of data     =    20
 Mean               =    877.283
 Standard deviation =    357.474
 Skewness           =      -.088
 Kurtosis           =      2.617

 Nr./year observation obs.freq. theor.freq.p theo.ret-per.  st.dev.xp st.dev.p
     10     232.000       .0309       .0355        1.04    129.6295       .0283
      5     267.000       .0802       .0439        1.05    125.3182       .0325
      9     505.000       .1296       .1488        1.17     99.2686       .0644
     18     525.000       .1790       .1622        1.19     97.4253       .0669
```

```
     15     606.000       .2284       .2240        1.29     90.7089       .0759
     14     628.000       .2778       .2428        1.32     89.1161       .0780
      7     649.580       .3272       .2621        1.36     87.6599       .0799
      4     722.000       .3765       .3320        1.50     83.6122       .0849
     11     849.400       .4259       .4689        1.88     80.0545       .0891
      3     892.000       .4753       .5164        2.07     79.9673       .0892
     16     924.000       .5247       .5520        2.23     80.2727       .0888
     20     950.000       .5741       .5806        2.38     80.7532       .0883
     19    1050.000       .6235       .6855        3.18     84.4622       .0839
      6    1110.000       .6728       .7425        3.88     87.9885       .0795
     12    1167.684       .7222       .7917        4.80     92.1776       .0740
      8    1173.000       .7716       .7959        4.90     92.5994       .0734
     13    1174.000       .8210       .7967        4.92     92.6794       .0733
      2    1197.000       .8704       .8144        5.39     94.5736       .0708
      1    1347.000       .9198       .9056       10.59    109.1187       .0513
     17    1577.000       .9691       .9748       39.76    136.5096       .0224

Results of Binomial goodness of fit test
 variate dn = max(|Fobs-Fest|)/sd=      .7833 at Fest= .7917
 prob. of exceedance P(DN>dn)    =      .4335
 number of observations          =    20

Results of Kolmogorov-Smirnov test
 variate dn = max(|Fobs-Fest|)   =      .0925
 prob. of exceedance P(DN>dn)    =      .9955

 Results of Chi-Square test
 variate = chi-square           =     1.2000
 prob. of exceedance of variate =      .2733
 number of classes              =     4
 number of observations         =    20
 degrees of freedom             =     1

  Values for distinct return periods
  Return per.  prob(xi<x) p     value x  st. dev. x   confidence intervals
                                                        lower       upper
           2      .50000     877.283      79.934     720.582    1033.985
           5      .80000    1178.082      93.013     995.740    1360.424
          10      .90000    1335.468     107.878    1123.984    1546.952
          25      .96000    1503.247     127.221    1253.844    1752.650
          50      .98000    1611.602     140.961    1335.263    1887.941
         100      .99000    1709.048     153.900    1407.343    2010.753
         250      .99600    1825.469     169.899    1492.399    2158.539
         500      .99800    1906.275     181.273    1550.908    2261.643
        1000      .99900    1982.065     192.101    1605.471    2358.660
        1250      .99920    2005.533     195.482    1622.312    2388.754
        2500      .99960    2075.895     205.685    1672.672    2479.118
        5000      .99980    2142.841     215.477    1720.421    2565.260
       10000      .99990    2206.758     224.893    1765.878    2647.638
```

The fit of the normal distribution to the observed frequency distribution is shown in Figure 5.6. The Blom plotting position has been used to assign non-exceedance frequencies to the ranked observations.

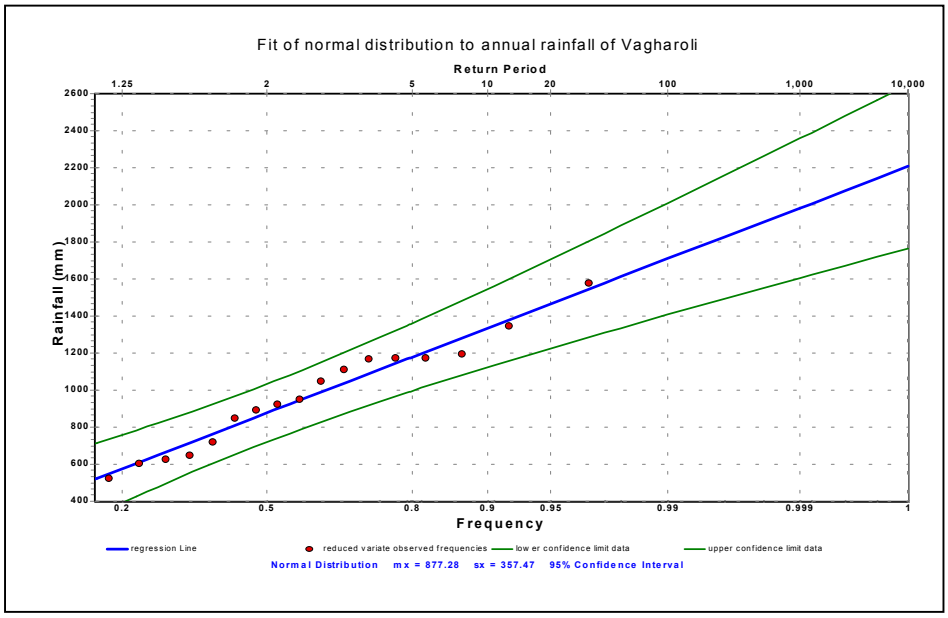

Figure 5.6: Fit of normal distribution to annual rainfall at Vaharoli, period 1978-1997

6 Hypothesis Testing

6.1 General

To apply the theoretical distribution functions dealt with in Chapter 5 the following steps are required:

1. Investigate the homogeneity of the data series, subjected to frequency analysis
2. Estimate the parameters of the postulated theoretical frequency distribution
3. Test the goodness of fit of the theoretical to the observed frequency distribution

In this chapter attention will be given to series homogeneity tests and goodness of fit tests. First an overview is given of the principles of hypothesis testing.

6.2 Principles

A statistical hypothesis is an assumption about the distribution of a statistical parameter. The assumption is stated in the **null-hypothesis H_0** and is tested against one or more alternatives formulated in the **alternative hypothesis H_1**. For easy reference the parameter under investigation is usually presented as a ***standardised*** variate, called ***test statistic***. Under the null-hypothesis the test statistic has some standardised sampling distribution, e.g. a standard normal, a Student t-distribution, etc. as discussed in Chapter 4. For the null-hypothesis to be true the value of the test statistic should be within the acceptance region of the sampling distribution of the parameter under the null-hypothesis. If the test statistic does not lie in the acceptance region, the null-hypothesis is rejected and the alternative is assumed to be true. Some risk, however, is involved that we make the wrong decision about the test:

- **Type I error**, i.e. rejecting H_0 when it is true, and
- **Type II error**, i.e. accepting H_0 when it is false.

The probability of making a Type I error is equal to the significance level of the test α. When a test is performed at a 0.05 or 5% level of significance it means that there is about 5% chance that the null-hypothesis will be rejected when it should have been accepted. This

probability represents the critical region at the extreme end(s) of the sampling distribution under H0. Note, however, the smaller the significance level is taken, the larger becomes the risk of making Type II error and the less is the discriminative power of the test.

Choosing the significance level α

Consider the following hypothesis. Let Φ denote the parameter under investigation and let:

H_0: $\Phi = \Phi_0$, and
H_1: $\Phi = \Phi_1$, with $\Phi_1 > \Phi_0$

The estimate of Φ is ϕ. The hypothesis is tested by means of a one-tailed test. The decision rule of acceptance is stated as follows:

Accept H_0 if: $\phi \leq c$
Reject H_0 and accept H_1 if: $\phi > c$

where c is a constant, for the time being chosen arbitrarily between Φ_0 and Φ_1. To specify c the relative positions of the pdf's of ϕ are considered $f_0(\phi|H_0)$ and $f_1((\phi|H_1)$ are, see Figure 6.1.

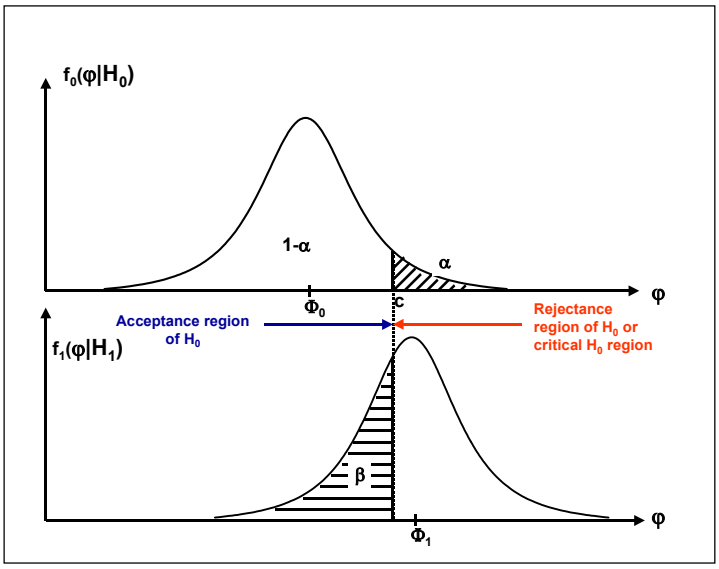

Figure 6.1:
Definition sketch for hypothesis testing

The region $\phi \leq c$ is called the **acceptance region** for H_0 and, reversely, the region $\phi > c$ is called the **rejectance or critical H_0 region**. If H_0 is true and $\phi \leq c$, then the right decision is made. However, if H_0 is true and $\phi > c$ then the wrong decision is made, i.e. an error of Type I. Formally:

$$P(\text{error of Type I}) = P(\phi > c \mid H_0 \text{ is true}) = \int_c^{\infty} f_0(\phi \mid H_0) d\phi = \alpha \qquad (6.1)$$

On the other hand, if H_1 is true and $\phi \leq c$, or equivalently, accepting H_0 when it is false, then a Type II error is made. It has a probability of occurrence defined by:

$$P(\text{error of Type II}) = P(\phi \leq c \mid H_1 \text{ is true}) = \int_{-\infty}^{c} f_1(\phi \mid H_1) d\phi = \beta \qquad (6.2)$$

In production processes, the risk associated with Type I errors is called the **producer's risk** and the Type II risk the **consumer's risk**. Now basically c has to be chosen such that the total loss associated with making errors of Type I and of Type II are minimised. Hence, if L_α

and L_β are the losses associated with errors of Type I and Type II respectively, and L is the total loss, with:

$$L = \alpha(c)\, L_\alpha + \beta(c) L_\beta \qquad (6.3)$$

Then c follows from the minimum of L. In practice, however, the loss functions L_α and L_β are usually unknown and the **significance level** α is chosen **arbitrarily** small like 0.1 or 0.05. From Figure 6.1 it is observed that a low value of α implies a very high value of β. The test then is seen to have a very low **discriminative power**; the likelihood of accepting H_0, when it is false, is becoming very large. By definition, the **power of a test** = 1 - β, i.e. the complement of β and it expresses the probability of rejecting H_0 when it is false, or the probability of avoiding Type II errors. In this case:

$$1-\beta = \int_c^\infty f_1(\phi \mid H_1)\, d\phi \qquad (6.4)$$

If the test is two-sided with acceptance region for H_0: $d \le \phi \le c$, the power of the test is given by:

$$1-\beta = \int_{-\infty}^{d} f_1(\phi \mid H_1)\, d\phi + \int_c^\infty f_1(\phi \mid H_1)\, d\phi \qquad (6.5)$$

If the alternative is not a single number, but can take on different values, then β becomes a function of ϕ. This function $\beta(\phi)$ is called the **operating characteristic** (OC) of the test and its curve the OC-curve. Similarly, $\eta(\phi) = 1-\beta(\phi)$ is called the **power function** of the test.

In summary: Type I and Type II errors in testing a hypothesis $\Phi = \Phi_0$ against an alternative $\Phi = \Phi_1$ read:

		Test hypothesis H_0: $\Phi = \Phi_0$	
		Accepted	**Rejected**
True state	$\Phi = \Phi_0$	**Correct decision** P = 1 - α	**Type I error** P = α
	$\Phi = \Phi_1$	**Type II error** P = β	**Correct decision** P = 1 - β

Table 6.1: Overview of hypothesis test results

Test procedure

Generally, the following procedure is used in making statistical tests (Haan, 1977):

1. Formulate the hypothesis to be tested
2. Formulate an alternative hypothesis
3. Determine a test statistic
4. Determine the distribution of the test statistic
5. Collect data needed to calculate the test statistic
6. Determine if the calculated value of the test statistic falls in the rejection region of the distribution of the test statistic.

Depending on the type of alternative hypothesis H_1 one- or two-tailed tests are considered. This is explained by the following example. To test the significance of serial correlation the value of the serial correlation coefficient r is considered. The null-hypothesis reads H_0: $\rho = 0$ against one of the following alternatives:

1. $H_1 : \rho > 0$, i.e. a right-sided test
2. $H_1 : \rho < 0$, i.e. a left-sided test
3. $H_1 : \rho \neq 0$, i.e. a two-sided test

The serial correlation coefficient is estimated from:

$$r = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}(x_i - m_X)(x_{i+1} - m_X)}{\frac{1}{N}\sum_{i=1}^{N}(x_i - m_X)^2} \qquad (6.6)$$

The test statistic to measure the significance of r is:

$$T_r = r\sqrt{\frac{N-3}{1-r^2}} \qquad (6.7)$$

Under the null-hypothesis the test statistic T_r has a Student t-distribution with ν = N-3 degrees of freedom. Let the tests be performed at a significance level α, then H_0 will not be rejected in:

1. a right-sided test, if: $T_r \leq t_{\nu,1-\alpha}$
2. a left-sided test, if: $T_r \geq t_{\nu,\alpha}$
3. a two-sided test, if: $t_{\nu,\alpha/2} \leq T_r \leq t_{\nu,1-\alpha/2}$

Since the Student distribution is symmetrical the last expression may be replaced by:

$$|T_r| \leq t_{\nu,1-\alpha/2} \qquad (6.8)$$

The latter condition is investigated when testing randomness of a series. The various options are displayed in Figure 6.2.

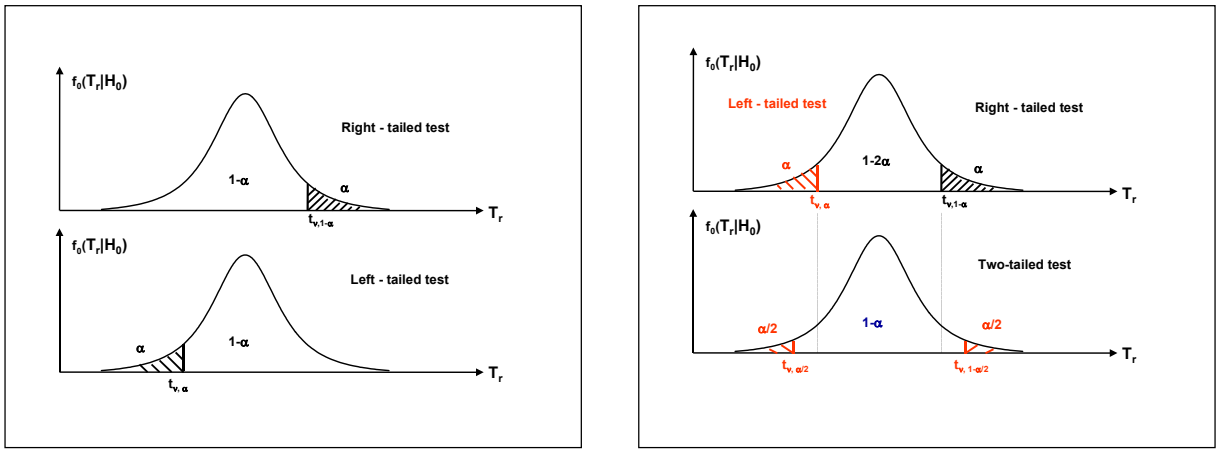

Figure 6.2: Right-tailed, left-tailed and two-tailed tests

From Figure 6.2 it is observed that for the same significance level the critical values differ in a one-tailed or a two-tailed test.

6.3 Investigating homogeneity

Prior to fitting of theoretical distributions to observed ones, the sample series should fulfil the following conditions:

stationarity: i.e. the properties or characteristics of the series do not vary with time;
homogeneity: i.e. all elements of a series belong to the same population;
randomness: i.e. series elements are independent.

The first two conditions are transparent and obvious. Violating the last one, while the series were tested homogeneous, means that the effective number of data is reduced and hence the power of the tests and the quality of the estimates. Lack of randomness may, however, have several causes; in case of a trend there will also be serial correlation.

HYMOS includes numerous statistical tests to investigate the stationarity, homogeneity or randomness. A number of them are **parametric** tests, which assume that the sample is taken from a population with an approximately normal distribution. **Non-parametric** or distribution-free do not set conditions to the distribution of the sample. Generally, this freedom affects the discriminative power of the test negatively.

Tests included in HYMOS suitable for series inspection prior to frequency analysis comprise a.o.:

On randomness:

1. **Median run test**: a test for randomness by calculating the number of runs above and below the median;
2. **Turning point test**: a test for randomness by calculating the number of turning points;
3. **Difference sign test**: a test for randomness by calculating the number of positive and negative differences;

On correlation and trend:

1. **Spearman rank correlation test**: the Spearman rank correlation coefficient is computed to test serial correlation or significance of a trend;
2. **Spearman rank trend test**
3. **Arithmetic serial correlation coefficient**: a test for serial correlation;
4. **Linear trend test**: a test on significance of linear trend by statistical inference on slope of trend line;

On homogeneity:

1. **Wilcoxon-Mann-Whitney U-test**: a test to investigate whether two series are from the same population;
2. **Student t-test**: a test on difference in the mean between two series;
3. **Wilcoxon W-test**: a test on difference in the mean between two series;
4. **Rescaled adjusted range test**: a test for series homogeneity by the rescaled adjusted range.

From each group an example will be given.

Difference sign test

The difference-sign test counts the number of positive differences n_p and of negative differences n_n between successive values of series x_i,(i = 1,N): $x_{(i+1)} - x_{(i)}$. Let the maximum of the two be given by N_{ds}:

$$N_{ds} = \mathrm{Max}(n_p, n_n) \qquad (6.9)$$

For an independent stationary series of length N_{eff} (N_{eff} = N - zero differences) the number of negative or positive differences is asymptotically *normally* distributed with $N(\mu_{ds},\sigma_{ds})$:

$$\left.\begin{aligned}\mu_{ds} &= \frac{1}{2}(N_{eff}-1)\\ \sigma_{ds}^2 &= \frac{1}{12}(N_{eff}+1)\end{aligned}\right\} \qquad (6.10)$$

The following hypothesis is considered:

H_0: series x_i is random, and
H_1: series is not random, with no direction for the deviation of randomness; hence, a two-tailed test is performed

The following standardised test statistic is considered:

$$|n_{ds}| = \frac{|N_{ds} - \mu_{ds}|}{\sigma_{ds}} \qquad (6.11)$$

The null-hypothesis will not be rejected at a α level of significance if:

$$|n_{ds}| < z_{1-\alpha/2} \qquad (6.12)$$

where z1-α/2 is the standard normal deviate with F(z < z1-α/2) = 1-α/2. A requirement is that the sample size has to be N $\geq$ 10.

Linear trend test

The slope of the trend line of series x_i, (i=1,N) with time or sequence is investigated. The linear trend equation reads:

$$x_i = b_1 + b_2 i + \varepsilon_i \quad \text{with}: \varepsilon_i \approx N(0,\sigma_\varepsilon) \qquad (6.13)$$

The trend parameters are given by:

$$b_2 = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)(i - m_i)}{\sigma_i^2} \quad \text{with}: m_i = \frac{N+1}{2} \quad \text{and}: \sigma_i^2 = \frac{1}{12}N(N+1) \qquad (6.14)$$

$$b_1 = m_X - b_2 m_i$$

where: m_X = mean of x_i, i = 1, N

The following hypothesis is made:

H_0: no trend, i.e. the slope of the trend line should be zero: $\mu b2 = 0$, and
H_1: significant trend, i.e. $\mu b2 \neq 0$, hence a two-tailed test is performed

The absolute value of the following standardised test statistic is computed:

$$|T_t| = \frac{|b_2|}{s_{b_2}} \quad \text{with}: s_{b_2}^2 = \frac{1}{N-1}\frac{\sigma_n^2}{\sigma_i^2} \quad \text{and}: \sigma_n^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(x_i - (b_1 + b_2 i)\right)^2 \qquad (6.15)$$

Under the null-hypothesis of no trend, the test statistic Tt has a Student t-distribution with ν=N-2 degrees of freedom for N $\geq$ 10. The null-hypothesis of zero trend will not be rejected at a significance level α, if:

$$|T_t| < t_{\nu,1-\alpha/2} \tag{6.16}$$

where $t\nu,1-\alpha/2$ is the Student-t variate defined by: $F(t<t\nu,1-\alpha/2) = 1-\alpha/2$

Student t-test and Fisher F-test

A good indicator for stationarity and homogeneity of a series is the behaviour of the mean value, for which the t-test is appropriate. With the Student t-test differences in mean values of two series yi,(i=1,m) and zi,(i=1,n) are investigated. In this case of frequency analysis the test is used as a split-sample test as it will be applied to the data from the same data set xi, I = 1, N. The series X is split in two parts Y and Z. The series Y and Z are chosen such that the first m represent Y and the last N-m are represented by Z. Let m_Y and m_Z denote the sample values of population means of Y and Z: μ_Y and μ_Z.

The following hypothesis is now tested:

H_0: $\mu_Y = \mu_Z$, and
H_1: $\mu_Y \neq \mu_Z$, hence a two-tailed test is performed

The absolute value of the following standardised test statistic is therefore investigated:

$$|T_S| = \frac{|m_Y - m_Z|}{s_{YZ}} \tag{6.17}$$

Under the null-hypothesis of equal population means the test statistic T_S has a *Student t*-distribution with ν = m+n-2 degrees of freedom for N = m + n > 10. The null-hypothesis $\mu_Y = \mu_Z$ will not be rejected at a significance level α, if:

$$|T_S| < t_{\nu,1-\alpha/2} \tag{6.18}$$

where $t_{\nu,1-\alpha/2}$ is the Student-t variate defined by: $F(t<t_{\nu,1-\alpha/2}) = 1-\alpha/2$

The way the standard deviation s_{YZ} is computed depends on whether the series Y and Z have the same population variance. For this a Fisher F-test is performed on the ratio of the variances.

The following hypothesis is made:

H_0: $\sigma_Y^2 = \sigma_Z^2$, and
H_1: $\sigma_Y^2 \neq \sigma_Z^2$, by putting the largest one on top a one-tailed test is performed.

Following test statistic is considered:

$$F_S = \frac{s_Y^2}{s_Z^2} \text{ if : } s_Y^2 > s_Z^2 \text{ else : } F_S = \frac{s_Z^2}{s_Y^2} \tag{6.19}$$

Under the null-hypothesis the test statistic FS has a Fisher F-distribution with (m-1, n-1) degrees of freedom if sY2 > sZ2, otherwise the number of degrees of freedom is (n-1, m-1). The null-hypothesis σY2 = σZ2 will not be rejected at a significance level α, if:

$$F_S < f_{m-1,n-1,1-\alpha} \tag{6.20}$$

where fm-1,n-1,1-α is the Fisher-F variate defined by: F(f < fm-1,n-1,1-α) = 1-α.

For fitting distributions to the sample series X it is essential that the hypothesis on the mean and the variance are both not rejected. If one of the hypotheses is rejected, the series should not be applied.

The outcome of the variance test determines in which way the standard deviation s_{YZ} is being estimated (Hald, 1952). The standard deviation s_{YZ} is computed from:

1. in case of equal variances:

$$s_{YZ} = \sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)\frac{(m-1)s_Y^2+(n-1)s_Z^2}{m+n-2}} \tag{6.21}$$

2. in case of unequal variances:

$$s_{YZ} = \sqrt{\frac{s_Y^2}{m}+\frac{s_Z^2}{n}} \text{ and : } \nu = \left(\frac{\psi^2}{m-1}+\frac{(1-\psi)^2}{n-1}\right)^{-1} \text{ and : } \psi = \frac{\frac{s_Y^2}{m}}{\frac{s_Y^2}{m}+\frac{s_Z^2}{n}} \tag{6.22}$$

Practically, it implies that in the latter case the number of degrees of freedom ν becomes less than in the equal variance case, so the discriminative power of the test diminishes somewhat. With respect to the sample size it is noted that the following conditions apply: N≥10, m≥5 and n≥5.

Example 5.2: continued: Annual rainfall Vagharoli.

The above-discussed tests have been applied to the annual rainfall series of Vagharoli available for the period 1978-1997. In the split-sample test on the mean and the variance the series have been split in equal parts. It is noted though, that in practice one should first inspect the time series plot of the series to determine where the boundary between the two parts is to be put. The time series of the annual rainfall is shown in Figure 6.3.

Results of tests

Difference Sign Test

Number of difference signs (=N_{ds})= 11
Mean of N_{ds} = 9.500
Standard deviation of N_{ds} = 1.323
Test statistic [n_{ds}](abs.value) = 1.134
Prob(n_{ds}.le. $n_{ds,obs}$ = .872
Hypothesis: H_0: Series is random
H_1: Series is not random
A two-tailed test is performed
Level of significance is α 5.00 percent
Critical value for test statistic $z_{1-\alpha/2}$ = 1.960
Result: H_0 not rejected

Test for Significance of Linear Trend

Intercept parameter (=b_1) = 871.612
Slope parameter (=b_2) = .5401E+00
St.dev. of b2 (=s_{b2}) = .1424E+02
St.dev. of residual (=s_e) = .3673E+03
Test statistic [T_t] (abs.value) = .038
Degrees of freedom □ = 18
Prob(T_t.le $T_{t,obs}$ = .515
Hypothesis: H_0: Series is random

H_1: Series is not random
A two-tailed test is performed
Level of significance α is 5.00 percent
Critical value for test statistic $t_{□,1-\alpha/2}$ = 2.101

Result: H_0 not rejected

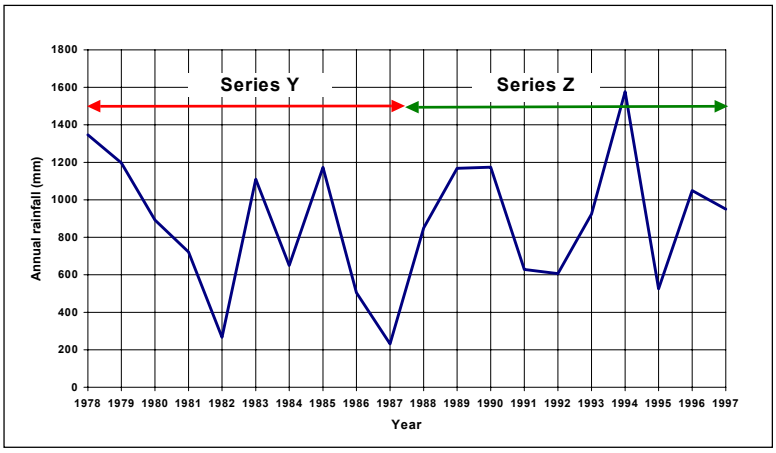

Figure 6.3:
Annual rainfall at Vagharoli, period 1978-1997, with division for split sample test

Student t-Test with Welch modification

Number of data in first set = 10
Number of data in second set = 10
Test statistic [T_S] (abs.value) = .842
Degrees of freedom = 18
Prob(t.<.[T_S]) = .795
Mean of first set (m_Y) = 809.458
St.dev. of first set (s_Y) = 397.501
Mean of second set (m_Z) = 945.108
St.dev. of second set (s_Z) = 318.659
Var. test stat. $F_S = s_Y^2/s_Z^2$) = 1.556
Prob($F \leq F_S$) = .740
Hypothesis: H_0: Series is homogeneous
H_1: Series is not homogeneous
A two-tailed test is performed
Level of significance is α = 5.00 percent
Critical value for test statistic mean $t_{□,1-\alpha/2}$ = 2.101
Critical value for test statistic variance $F_{m-1,n-1,1-\alpha}$ = 3.18

Result: H_0 not rejected

6.4 Goodness of fit tests

To investigate the goodness of fit of theoretical frequency distribution to the observed one three tests are discussed, which are standard output in the results of frequency analysis when using HYMOS, viz:

- Chi-square goodness of fit test
- Kolmogorov-Smirnov test, and
- Binomial goodness of fit test.

Chi-square goodness of fit test

The hypothesis is that F(x) is the distribution function of a population from which the sample x_i, i = 1,…,N is taken. The hypothesis is tested by comparing the actual to the theoretical

number of occurrences within given class intervals. The following procedure is followed in the test

First, the data set is divided in k class intervals such that each class contains at least 5 values. The class limits are selected such that all classes have equal probability $p_j = 1/k = F(z_j)-F(z_{j-1})$. For example if there are 5 classes, the upper class limits will be derived from the variate corresponding with the non-exceedance frequencies p = 0.20, 0.40, 0.60, 0.80 and 1.0. The interval j contains all x_i with: $U_c(j-1) < x_i \leq U_c(j)$, where $U_c(j)$ is the upper class limit of class j, see Figure 6.4. The number of sample values falling in class j is denoted by b_j.

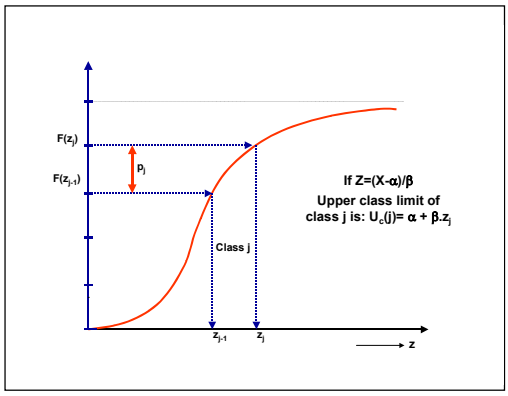

Figure 6.4:
Definition sketch for class selection in Chi-square goodness of fit test

Next, the number of values expected in class j according to the theoretical distribution is determined, which number is denoted by e_j. The theoretical number of values expected in any class is N/k, since all classes have equal probability.

The following test statistic is considered:

$$\chi_c^2 = \sum_{j=1}^{k} \frac{(b_j - e_j)^2}{e_j} \qquad (6.23)$$

This test statistic has under the assumption of the null-hypothesis a chi-squared distribution with ν = k-1-m degrees of freedom, where k = number of classes and m = number of parameters in the theoretical distribution. Because of the choice of equal probabilities (6.23) can be simplified as follows:

$$\chi_c^2 = \sum_{j=1}^{k} \frac{(b_j - N/k)^2}{N/k} = \frac{k}{N}\sum_{j=1}^{k} b_j^2 - N \qquad (6.24)$$

The null-hypothesis will not be rejected at a significance level α if:

$$\chi_c^2 < \chi_{\nu,1-\alpha}^2 \quad \text{with}: \nu = k-1-m \qquad (6.25)$$

The following number of class intervals k given N are suggested, see Table 6.2

N	k	N	k	N	k
20-29	5	100-199	13	800-999	27
30-39	7	200-399	16	1000-1499	30
40-49	9	400-599	20	1500-1999	35
50-99	10	600-799	24	≥ 2000	39

Table 6.2: Recommended number of class intervals for Ch-square goodness of fit test

Example 5.2: continued:

Annual rainfall Vagharoli. It is investigated if the null-hypothesis that the sample series of annual rainfall fits to the normal distribution. It is observed from the results in Chapter 5 that HYMOS has selected 4 class intervals, hence k = 4 and the upper class levels are obtained at non-exceedance probabilities 0.25, 0.50, 0.75 and 1.00. The reduced variates for these probabilities can be obtained from tables of the normal distribution or with the Statistical Tables option in HYMOS. The reduced variates are respectively –0.674, 0.000, 0.674 and ∞, hence with mean = 877 and standard deviation = 357 the class limits become 877-0.674x357, 877, 877+0.674x357 and ∞, i.e. 636, 877, 1118 and ∞. The number of occurrences in each class is subsequently easily obtained from the ranked rainfall values presented in Chapter 5, Example 5.2. The results are presented in Table 6.3

Non-exc. probability of upper class limits	Reduced variate of upper class limits	Class intervals expressed in mm	Number of occurrences b_i	b_i^2
0.25	-0.67	0- 636	6	36
0.50	0.00	637-877	3	9
0.75	0.67	878-1118	5	25
1.00	∞	1119-∞	6	36
			sum	106

Table 6.3: Number of occurrences in classes

From Table 6.3 it follows for the test statistic (6.24):

$$\chi_c^2 = \frac{4}{20} x106 - 20 = 12$$

The critical value at a 5% significance level, according to the chi-squared distribution for ν = 4-1-2 = 1 degrees of freedom, is 3.84. Hence the computed value is less than the critical value. Consequently, the null-hypothesis is not rejected at the assumed significance level, as can be observed from the HYMOS results as well.

Kolmogorov-Smirnov test

In the Kolmogorov-Smirnov test the differences between the theoretical and observed frequency distribution is analysed and when the difference at a particular non-exceedance frequency exceeds a critical limit then the null-hypothesis that the sample is from the assumed theoretical distribution is rejected.

Let the observed frequency distribution be denoted by $S_N(x)$ and is defined by:

$$S_N(x) = \begin{cases} 0 & \text{for : } x < x_1 \\ \frac{i}{N} & \text{for : } x_i \le x < x_{i+1} \\ 1 & \text{for : } x_N \le x \end{cases} \qquad (6.26)$$

where x1 and xN are respectively the smallest and largest elements of the sample. Now, at each observed value xi, I = 1,N the difference between F(x), i.e. the theoretical distribution, and SN(x) is determined. The difference has two values as SN(x) changes at each step. If these two differences are denoted by ∂i+ and ∂i-, (see Figure 6.5) then the test statistic DN is developed as follows:

$$\partial_i^+ = \frac{i}{N} - F(x) \quad and: \quad \partial_i^- = F(x) - \frac{(i-1)}{N}$$
$$d_i = Max(\partial_i^+, \partial_i^-) \qquad (6.27)$$
$$D_N = Max(d_1, d_2, ..., d_N)$$

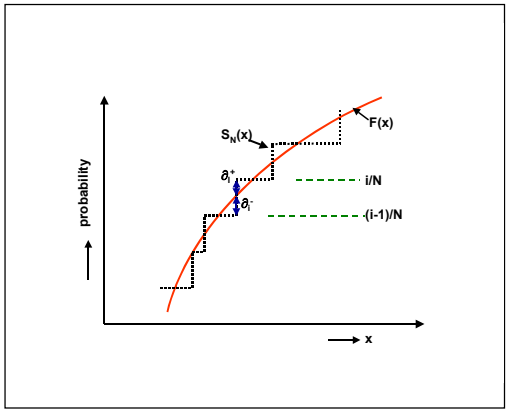

Figure 6.5:
Definition sketch Kolmogorov-Smirnov test (adapted from NERC, 1975)

The null hypothesis is not rejected at a significance level α if D_N does not exceed the critical values Δ read from Kolmogorov-Smirnov's table:

$$D_N < \Delta_\alpha \qquad (6.28)$$

Critical values at the 10, 5 and 1% significance level for $N \geq 35$ are respectively $1.22/\sqrt{N}$, $1.36/\sqrt{N}$, and $1.63/\sqrt{N}$.

Example 5.2: continued: annual rainfall Vagharoli.

The results of the application of the Kolmogorov-Smirnov test to the annual rainfall series of Vagharoli are presented in the table below.

Year nr	Rainfall	Blom	i/N	(i-1)/N	F(x)	d+	d-	max(d+,d-)
10	232	0.031	0.05	0.00	0.0355	0.0145	0.0355	0.0355
5	267	0.080	0.10	0.05	0.0439	0.0561	-0.0061	0.0561
9	505	0.130	0.15	0.10	0.1488	0.0012	0.0488	0.0488
18	525	0.179	0.20	0.15	0.1622	0.0378	0.0122	0.0378
15	606	0.228	0.25	0.20	0.2240	0.0260	0.0240	0.0260
14	628	0.278	0.30	0.25	0.2428	0.0572	0.0072	0.0572
7	650	0.327	0.35	0.30	0.2621	0.0879	0.0379	0.0879
4	722	0.377	0.40	0.35	0.3320	0.0680	-0.0180	0.0680
11	849	0.426	0.45	0.40	0.4689	-0.0189	0.0689	0.0689
3	892	0.475	0.50	0.45	0.5164	-0.0164	0.0664	0.0664
16	924	0.525	0.55	0.50	0.5520	-0.0020	0.0520	0.0520
20	950	0.574	0.60	0.55	0.5806	0.0194	0.0306	0.0306
19	1050	0.624	0.65	0.60	0.6855	-0.0355	0.0855	0.0855
6	1110	0.673	0.70	0.65	0.7425	-0.0425	0.0925	0.0925
12	1168	0.722	0.75	0.70	0.7917	-0.0417	0.0917	0.0917
8	1173	0.772	0.80	0.75	0.7959	0.0041	0.0459	0.0459
13	1174	0.821	0.85	0.80	0.7967	0.0533	-0.0033	0.0533
2	1197	0.870	0.90	0.85	0.8144	0.0856	-0.0356	0.0856
1	1347	0.920	0.95	0.90	0.9056	0.0444	0.0056	0.0444
17	1577	0.969	1.00	0.95	0.9748	0.0252	0.0248	0.0252
							Max	0.0925

Table 6.4: Kolmogorov-Smirnov test on annual rainfall

It is observed from Table 6.4 that the test statistic D_N = 0.0925. According to the Statistical Tables of the Kolmogorov-Smirnov test the critical value at a 5% confidence level for N = 20 is: Δ_5 = 0.29. Hence, the observed D_N is less than the critical value, so the null hypothesis that the observations are drawn from a normal distribution with mean 877 mm and standard deviation 357 mm is not rejected.

Binomial goodness of fit test

A third goodness of fit test is based on the fact that, when the observed and the theoretical distribution functions, respectively F1(x) and F2(x), are from the same distribution, then the standardised variate DB, defined by:

$$D_B = \frac{|F_1(x) - F_2(x)|}{s_B} \quad \text{with:} \quad s_B = \sqrt{\frac{F_2(x)(1-F_2(x))}{N}} \tag{6.29}$$

is approximately normally distributed with N(0,1). Hence, the null-hypothesis is not rejected at a α % significance level if:

$$D_B < z_{1-\alpha/2} \tag{6.30}$$

The test is used in the range where:

$$N\ F2(x)\{1- F2(x)\} > 1 \tag{6.31}$$

This criterion generally means that the tails of the frequency distribution are not subjected to the test.

Example 5.2 continued: annual rainfall Vagharoli. The results of the test are displayed in Table 6.5

Nr./year	observation	$F_1(x)$	$F_2(x)$	s_B	D_B	criterion
10	232	0.0343	0.0355	0.0414	0.0290	0.6848
5	267	0.0833	0.0439	0.0458	0.8601	0.8395
9	505	0.1324	0.1488	0.0796	0.2061	2.5332
18	525	0.1814	0.1622	0.0824	0.2329	2.7178
15	606	0.2304	0.2240	0.0932	0.0686	3.4765
14	628	0.2794	0.2428	0.0959	0.3817	3.6770
7	650	0.3284	0.2621	0.0983	0.6742	3.8681
4	722	0.3775	0.3320	0.1053	0.4321	4.4355
11	849	0.4265	0.4689	0.1116	0.3800	4.9807
3	892	0.4755	0.5164	0.1117	0.3660	4.9946
16	924	0.5245	0.5520	0.1112	0.2473	4.9459
20	950	0.5735	0.5806	0.1103	0.0643	4.8701
19	1050	0.6225	0.6855	0.1038	0.6068	4.3118
6	1110	0.6716	0.7425	0.0978	0.7251	3.8239
12	1168	0.7206	0.7917	0.0908	0.7830	3.2982
8	1173	0.7696	0.7959	0.0901	0.2918	3.2489
13	1174	0.8186	0.7967	0.0900	0.2434	3.2394
2	1197	0.8676	0.8144	0.0869	0.6120	3.0231
1	1347	0.9167	0.9056	0.0654	0.1698	1.7098
17	1577	0.9657	0.9748	0.0350	0.2597	0.4913
				Max	**0.7830**	

Table 6.5: Results of binomial goodness of fit test, annual rainfall Vagharoli

In HYMOS, the observed non-exceedance frequency distribution $F_1(x)$ is obtained from Chegodayev plotting position, see Table 5.4. From Table 6.5 it is observed that the maximum value for D_B = 0.8601 at a non-exceedance frequency = 0.0439. However, criterion (6.31), which is presented in the last column, is not fulfilled for that non-exceedance frequency (criterion is less than 1). For the range of data for which this criterion is fulfilled, the maximum value for D_B = 0.7830 at $F_2(x)$ = 0.7917. The critical value for D_B at a 5% confidence level is 1.96, hence, according to (6.30), the null-hypothesis that both $F_1(x)$ and $F_2(x)$ are from the same distribution is not rejected.

ANNEX 4.1 Standard normal distribution

The standard normal distribution function reads:

$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-s^2) ds = \frac{1}{2} + \frac{1}{2} erf\left(\frac{z}{\sqrt{2}}\right) \qquad \text{(A4.1.1)}$$

The following approximation is used in HYMOS to solve $F_Z(z)$ for a given value of the standard normal variate z:

$$F = \exp\left(-\frac{z^2}{2}\right)((((a_1 T + a_2)T + a_3)T + a_4)T + a_5)T \quad with: \quad T = \frac{1}{1 + b|z|}$$

$$For \quad z \le 0: \quad F_Z(z) = F \qquad \text{(A4.1.2)}$$

$$For \quad z > 0: \quad F_Z(z) = 1 - F$$

The coefficients in (A4.2) read:

a_1 = 0.530702715
a_2 = -0.726576014
a_3 = 0.71070687
a_4 = -0.142248368
a_5 = 0.127414796
b = 0.2316419

The absolute error in above approximation is < 7.5 x 10^{-8}.
The equation in a slightly different form can be found in Ambramowitz et al (1970) equation 26.2.17

ANNEX 4.2 Inverse of the standard normal distribution

The standard normal distribution function is given by (A4.1.1). The inverse of the standard normal distribution is found from:

$$y = T - \frac{a_1 + a_2 T + a_3 T^2}{1 + a_4 T + a_5 T^2 + a_6 T^3}$$

$$for \quad F_Z(z) < 0.5: \quad z = -y$$

$$for \quad F_Z(z) \ge 0.5: \quad z = y \qquad \text{(A4.2.1)}$$

$$with: \quad T = \sqrt{-2 \ln P}$$

$$where: \quad P = F_Z(z) \quad for \quad F_Z(z) \le 0.5$$

$$and \quad P = 1 - F_Z(z) \quad for \quad F_Z(z) > 0.5$$

The coefficients in (A4.2.1) read:
a_1 = 2.515517
a_2 = 0.802853
a_3 = 0.010328
a_4 = 1.432788
a_5 = 0.189269
a_6 = 0.001308

The absolute error in above approximation is < 4.5x10^{-4}.

The equation can be found in Ambramowitz et al (1970) equation 26.2.23.

ANNEX 4.3 Incomplete gamma function

The incomplete gamma function is defined by:

$$F_Z(z) = \frac{1}{\Gamma(\gamma)} \int_0^z t^{\gamma-1} \exp(-t) dt \qquad \text{(A4.3.1)}$$

To determine the non-exceedance probability for any value of z > 0 the following procedure is used. Three options are considered dependent on the value of γ and z:

- If $\gamma \geq 500$: then the Wilson-Hilverty transformation:

  $$y = 3\sqrt{\gamma}\left[\left(\frac{z}{\gamma}\right)^{1/3} - 1 + \frac{1}{9\gamma}\right] \qquad \text{(A4.3.2)}$$

  The variable y has a standard normal distribution.
- If $z \leq \gamma$ or $z \leq 1$ a rapidly converging series development is used:

  $$F_Z(z) = \exp(-z)\, z^{\gamma} \sum_{j=1}^{\infty} \frac{z^{j-1}}{\Gamma(\gamma + j)} \qquad \text{(A4.3.3)}$$

  The algorithm is taken to have converged when the summation S fulfils:
  $\frac{S_n - S_{n-1}}{S_n} \leq 10^{-6}$
- If $z > \gamma$ and $z > 1$ a rapidly converging continued fraction development is used:

$$F_Z(z) = 1 - \frac{\exp(-z) z^{\gamma}}{\Gamma(\gamma)} \left(\cfrac{1}{z + \cfrac{1-\gamma}{1 + \cfrac{1}{z + \cfrac{2-\gamma}{1 + \cfrac{2}{z + \cfrac{3-\gamma}{1 + \cfrac{3}{z + \ldots.}}}}}}} \right)$$

or shortly written as:

$$F_Z(z) = 1 - \frac{\exp(-z) z^{\gamma}}{\Gamma(\gamma)} \left(\frac{1}{z+} \frac{1-\gamma}{1+} \frac{1}{z+} \frac{2-\gamma}{1+} \frac{2}{z+} \frac{3-\gamma}{1+} \frac{3}{z+} \ldots\ldots\ldots \right) \qquad \text{(A4.3.4)}$$

The continued fraction S can be rewritten as:

$$S = \frac{1}{z}\left(1 + \frac{\gamma - 1}{(2 - \gamma + z) +} \frac{\gamma - 2}{(4 - \gamma + z) +} \frac{2(\gamma - 3)}{(6 - \gamma + z) +} \ldots\ldots\right)$$

The n^{th} convergent of S reads:

$$S_n = \frac{A_n}{B_n} = \frac{1}{z}\left(1 + \frac{a_1}{b_{1+}} \frac{a_2}{b_2 +} \frac{a_3}{b_3 +} \cdots\cdots\cdots \frac{a_n}{b_n}\right) \qquad \text{(A4.3.5)}$$

which is calculated using recursively:

$A_0 = 1$ $\quad B_0 = z$
$A_1 = z + 1$ $\quad B_1 = z(2 - \gamma + z)$
$a_j = (j - 1)(\gamma - j)$ $\quad b_j = 2j - g + z$
$A_j = b_jA_{j-1} + a_jA_{j-2}$ $\quad B_j = b_jB_{j-1} + a_jB_{j-2}$ for: j = 2, ……, n

The iteration is taken to have converged when:

$$\frac{S_n - S_{n-1}}{S_n} \leq 10^{-6}$$

ANNEX 4.4 Inverse of incomplete gamma function

The above procedure is also used to arrive at the inverse of the incomplete gamma function. For this the routine to compute the incomplete gamma function is seeded with a variate $z = 2^k$, for k = 1, 2, …, 50. The function returns the non-exceedance probability $F_Z(z)$ for each z.

Let the required exceedance probability be denoted by P. If for a particular value of $z = 2^k$ the function return be an $F_Z(z) > P$, then the computation is stopped and an interpolation is made between $z = 2^{k-1}$ and 2^k such that $F_Z(z) - P = 0$. The interpolation is repeated to arrive at a required accuracy.