Chapter 5

Typical univariate statistical analysis in geophysical processes

Demetris Koutsoyiannis
Department of Water Resources and Environmental Engineering
Faculty of Civil Engineering, National Technical University of Athens, Greece

Summary

Assuming that a certain geophysical process on a particular time scale (typically annual) can be represented by a single random variable (rather than a stochastic process, in which time dependence cannot be neglected), we can use classical statistical analysis to carry out several statistical tasks, such as:

1. *Sample description by summary statistics.* This is done either numerically, using some representative statistical indicators, or graphically, using box plots, histograms and empirical distribution plots.
2. *Fitting of a theoretical model*. This comprises the selection of an appropriate model (distribution function), the estimation of its parameters and the statistical testing of the fitting.
3. *Statistical prediction*. This aims to estimate the value of the variable (on a point or an interval basis) that corresponds to a certain return period.

The first task belongs to the so-called *descriptive statistics*, whereas the other two tasks are part of the *inferential statistics* or *statistical induction*. Although such statistical analyses are applicable for any type of theoretical model, in the discourse of this chapter we merely use the normal distribution, which is simple and best for illustration purposes.

5.1 Summary statistics

Summary statistics or *statistical characteristics* are various statistical indicators that enable description of the most characteristic properties of an observed sample (or even of a population) using a few numbers. The most common statistical characteristics can be classified into two categories. The first comprises the *sample moments* and their derivative characteristics. In particular, it involves: (a) the average, which, as we have seen, is a location measure; (b) the sample variance and the derivative indicators of dispersion (standard deviation and coefficient of variation); (c) the third central moment and the coefficient of skewness. The second category includes simpler statistical indicators, whose computation requires the sorting of the sample in descending or ascending order. Here these are referred to as *summary statistics of sorted sample* and include the minimum and maximum value of the sample, the median (location parameter), the upper and lower quartiles and the interquartile range (dispersion parameter).

The sample moments and their derivative characteristics are calculated by applying the related estimators that have been discussed in chapter 3 and are also summarized in Table 5.1, in a form convenient for calculations. Furthermore, Table 5.1 includes coefficients of bias correction (column 3), by which the simple estimates (column 2) must be multiplied to find unbiased estimates. Table 5.1 gives also instructions to find the summary statistics of the sorted sample.

Table 5.1 Typical summary statistics and formulae for their calculation

Statistical indicator	Simple estimate	Coefficient for bias correction
1. Sample moments and derivative characteristics		
Mean value	$\bar{x} = \frac{1}{n}\sum x_i$	—
Variance	$s_X^2 = \frac{1}{n}\sum x_i^2 - \frac{1}{n^2}\left(\sum x_i\right)^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$	$\frac{n}{n-1}$
Standard deviation	s_X	$\approx\sqrt{\frac{n}{n-1}}$
Coefficient of variability	$\hat{C}_{v_X} = \frac{s_X}{\bar{x}}$	$\approx\sqrt{\frac{n}{n-1}}$
Third central moment	$\hat{\mu}_X^{(3)} = \frac{1}{n}\sum x_i^3 - \frac{3}{n^2}\left(\sum x_i\right)\left(\sum x_i^2\right) + \frac{2}{n^3}\left(\sum x_i\right)^3 = \frac{1}{n}\sum x_i^3 - 3\bar{x}s_X^2 - \bar{x}^3$	$\frac{n^2}{(n-1)(n-2)}$
Coefficient of skewness	$\hat{C}_{s_X} = \frac{\hat{\mu}_X^{(3)}}{s_X^3}$	See section 3.3.4
2. Summary characteristics of sorted sample		
Minimum value	$\hat{x}_{\min} = \min(x_1, x_2, \ldots, x_n)$	—
Maximum value	$\hat{x}_{\max} = \max(x_1, x_2, \ldots, x_n)$	—
Median	$\hat{x}_{0.5}$: The middle term of the sorted sample or, for even number of observations, the mean of the two middle values.	—
Lower quartile	$\hat{x}_{0.25}$: The median of the part of the sample containing the values $x_i \le \hat{x}_{0.5}$.	—
Upper quartile	$\hat{x}_{0.75}$: The median of the part of the sample containing the values $x_i \ge \hat{x}_{0.5}$.	—
Interquartile range	$\hat{\delta}_X = \hat{x}_{0.75} - \hat{x}_{0.25}$	—

The summary statistics of the sorted sample can be also visualized by means of a simple diagram, the so-called *box plot* (see an example in Fig. 5.1, p.6). This diagram contains a central orthogonal "box" and two vertical "whiskers", up and down of it. All these elements are plotted in an appropriate scale. This is constructed according to the following guidelines (Hirsch et al., 1993, p. 17.10):

1. The middle horizontal line of the box represents the median of the sample.
2. The bottom line of the box represents the lower quartile of the sample.
3. The top line of the box represents the upper quartile of the sample.
4. An auxiliary quantity, the step, is defined as 1.5 times the interquartile range.
5. The lower whisker extends from the bottom line of the box to the smallest value of the sample that is one step away from this line.
6. The upper whisker extends from the top line of the box to the largest value of the sample that is one step away from this line.
7. Sample values lying 1-2 steps away of the box are called *outside values* and are marked with a ×.
8. Sample values lying more that 2 steps away of the box are called *far-outside values* and are marked with a ○.

According to the above, the minimum and the maximum values of the sample are indicated in the box plot either as the whiskers' ends, if they are less than one step away from the box edges, or as the farthermost outside or far-outside values. The box plot provides thus a simple and general statistical depiction of the sample, illustrating simultaneously the characteristics of location (median), dispersion (interquartile range), and asymmetry. The symmetry or asymmetry of the sample is recognized from the position of the middle line in comparison to the bottom and top lines of the box, as well as from comparison of the lengths of the whiskers. Furthermore, the diagram informs us about how close to the normal distribution a sample is. For a normal distribution a symmetric picture of the diagram is expected and no outside or far-outside values are expected, except with frequencies 1 in 100 and 1 in 300 000 points, respectively.

5.1.1 Demonstration of summary statistics via a numerical example

Table 5.2 lists the observations of annual runoff of the Evinos river basin, central-western Greece, upstream of the hydrometric gauge at Poros Reganiou.* We wish to extract the summary statistics of the sample and draw its box plot.

a. Sample moments and derivative characteristics

Nowadays the computation of moments is easily performed by computers tools.† For completeness we present here the manual computations.

---

* Evinos river is part of the hydrosystem for the water supply of Athens. Poros Reganiou is located at a considerable distance downstream of the Aghios Demetrios dam, which enables diversion of Evinos to Athens.

† See for instance the Excel functions Average, Var, StDev, VarP, StDevP etc.

Table 5.2 Annual runoff volume (in hm^3)* of river Evinos, at Poros Reganiou gauge.

Hydrological year	Runoff volume	Hydrological year	Runoff volume	Hydrological year	Runoff volume
1970-71	807	1977-78	715	1984-85	588
1971-72	695	1978-79	1064	1985-86	874
1972-73	788	1979-80	942	1986-87	552
1973-74	705	1980-81	1042	1987-88	529
1974-75	462	1981-82	1037	1988-89	469
1975-76	580	1982-83	674	1989-90	217
1976-77	807	1983-84	906	1990-91	772

Table 5.3 Traditional calculations of sample moments.

i	x_i	x_i^2	x_i^3
1	807	651 249	525 557 943
2	695	483 025	335 702 375
3	788	620 944	489 303 872
4	705	497 025	350 402 625
5	462	213 444	98 611 128
6	580	336 400	195 112 000
7	807	651 249	525 557 943
8	715	511 225	365 525 875
9	1064	1 132 096	1 204 550 144
10	942	887 364	835 896 888
11	1042	1 085 764	1 131 366 088
12	1037	1 075 369	1 115 157 653
13	674	454 276	306 182 024
14	906	820 836	743 677 416
15	588	345 744	203 297 472
16	874	763 876	667 627 624
17	552	304 704	168 196 608
18	529	279 841	148 035 889
19	469	219 961	103 161 709
20	217	47 089	10 218 313
21	772	595 984	460 099 648
Sum	15 225	11 977 465	9 983 241 237

The calculation of sums $\sum x$, $\sum x^2$ and $\sum x^3$ is done in Table 5.3; their values are $\sum x$ = 15 225, $\sum x^2$ = 11 977 465 and $\sum x^3$ = 9 983 241 237. The average is

$$\bar{x} = \sum x / n = 15\,225 / 21 = 725.0 \text{ hm}^3$$

The sample variance is

* We remind that the unit hm^3 represents cubic hectometers (1 hm^3 = $(100 \text{ m})^3$ = 1 000 000 m^3).

$$s_X^2 = \sum x^2 / n - \bar{x}^2 = 11\,977\,465 / 21 - 725.0^2 = 44\,730.5\ (\text{hm}^3)^2$$

the sample standard deviation

$$s_X = \sqrt{44\,730.5} = 211.5\ \text{hm}^3$$

and the sample coefficient of variation

$$\hat{C}_{v_X} = s_X / \bar{x} = 211.5 / 725.0 = 0.29$$

The third central moment is

$$\hat{\mu}_X^{(3)} = \sum x^3 / n - 3\,\bar{x}\,s_X^2 - \bar{x}^3 = 9\,983\,241\,237 / 21 - 3 \times 725.0 \times 44\,730.5 - 725.0^3$$
$$= -2\,974\,523\ (\text{hm}^3)^3$$

and the coefficient of skewness

$$\hat{C}_{s_X} = \hat{\mu}_X^{(3)} / s_X^3 = -2\,974\,523 / 211.5^3 = -0.31$$

Table 5.4 Statistical characteristics (moments and derivative characteristics) of annual runoff (in hm^3) of the Evinos river basin at Poros Reganiou.

Statistical indicator	Simple estimation	Coefficient of bias correction	Unbiased estimation
Mean	$\bar{x} = \sum x / n = 725.0$	—	$\bar{x} = 725.0$
Variance	$s_X^2 = \sum x^2 / n - \bar{x}^2 = 44\,730$	$\frac{n}{n-1} = 1.05$	$s_X^{*2} = 46\,967$
Standard deviation	$s_X = 211.5$	$\approx\sqrt{\frac{n}{n-1}} = 1.025$	$s_X^* \approx 216.7$
Coefficient of variation	$\hat{C}_{v_X} = s_X / \bar{x} = 0.29$	$\approx\sqrt{\frac{n}{n-1}} = 1.025$	$\hat{C}_{v_X}^* \approx 0.29$
Third central moment	$\hat{\mu}_X^{(3)} = \sum x^3/n - 3\bar{x}s_X^2 - \bar{x}^3 = -2\,974\,523$	$\frac{n^2}{(n-1)(n-2)} = 1.16$	$\hat{\mu}_X^{*(3)} = -3\,542\,012$
Coeficient of skewness	$\hat{C}_{s_X} = \hat{\mu}_X^{(3)} / s_X^3 = -0.31.$	$\approx\frac{n^2}{(n-1)(n-2)} = 1.16$	$\hat{C}_{s_X}^* \approx -0.36$

The coefficients for correction of bias are: (i) for the variance

$$n / (n - 1) = 21 / 20 = 1.05.$$

(ii) for the standard deviation and the coefficient of variation (approximately)

$$\sqrt{n / (n - 1)} = \sqrt{1.05} = 1.025$$

and (iii) for the third central moment (and, approximately, for the coefficient of skewness)

$$n^2 / [(n-1)\,(n-2)] = 21^2 / (20 \times 19) = 1.16$$

Table 5.5 Sorted (is descending order) sample of annual runoff (in hm^3) of the Evinos river basin at Poros Reganiou.

Rank	Runoff volume	Rank	Runoff volume	Rank	Runoff volume
1	1064	8	807	15	588
2	1042	9	788	16	580
3	1037	10	772	17	552
4	942	11	715	18	529
5	906	12	705	19	469
6	874	13	695	20	462
7	807	14	674	21	217

Table 5.6 Summary characteristics of the sorted sample of annual runoff (in hm^3) of the Evinos river basin at Poros Reganiou

Statistical indicator	Estimate
Minimum value	$\hat{x}_{\max} = \min(x_1, \ldots, x_n) = 217$
Maximum value	$\hat{x}_{\max} = \max(x_1, \ldots, x_n) = 1064$
Median	$\hat{x}_{0.5} = x_{(11)} = 715$
Lower quartile	$\hat{x}_{0.25} = x_{(16)} = 580$
Upper quartile	$\hat{x}_{0.75} = x_{(6)} = 874$
Interquartile range	$\hat{d}_X = \hat{x}_{0.75} - \hat{x}_{0.25} = 294$

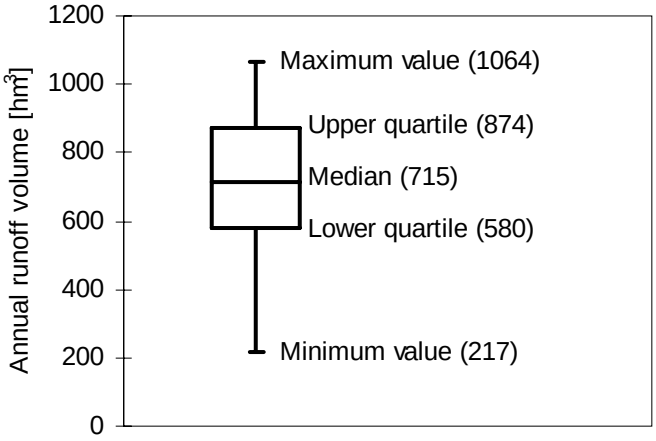

Fig. 5.1 Box plot of the annual runoff of the Evinos river basin at Poros Reganiou.

The results are summarized in Table 5.4.

b. Summary characteristics of the sorted sample

The observed sample, sorted in descending order*, is shown in Table 5.5. From this table we have calculated directly the summary characteristics of the sorted sample shown in Table 5.6.

* This sorting can be done in Excel using the function Large.

The median is the rank 11 (middle) value of the sorted sample, whereas the lower and upper quartiles are the rank 16 and 6 values, respectively.

c. Box plot

Following the procedure in section 5.1 and using the summary statistical characteristics of the sorted sample in Table 5.6, we easily construct the diagram of Fig. 5.1. The step size is 1.5 × 294 = 441 hm^3 and therefore the maximum ordinate of the upper whisker is 874 + 441 = 1315 hm^3. Given, however, that the maximum value of the sample is 1064 hm^3, the upper whisker should end up in this value. Likewise, the minimum ordinate of the lower whisker is 580 − 441 = 139 hm^3. Given, however, that the minimum value of the sample is 217 hm^3, the lower whisker should end up in this value.

5.2 Histograms

Histograms provide another graphical display of a sample, whose construction requires counting the sample values lying in k intervals, each of length Δ.* If the ith interval is $c_i \le x < c_{i+1}$ (where $c_{i+1} = c_i + \Delta$) and the number of the sample values lying within it is n_i, then the histogram is the function

$$(x) = \frac{n_i}{n\Delta}, \quad c_i \le x < c_{i+1}, \quad i = 1, \ldots, k \tag{5.1}$$

An example is depicted in Fig. 5.2. Often, the histogram is defined in a simpler manner, such as $\varphi(x) = n_i/n$, or $\varphi(x) = n_i$. For these two forms we use the terms *relative frequency histogram* and (absolute) *frequency histogram*, respectively. To avoid confusion, the histogram defined by (5.1) can be termed *frequency density histogram*.

To construct the histogram, we first select the number of intervals k. As a rule, we take k = ln n / ln 2 and the resulting value is rounded up. The length Δ is taken equal for all intervals (although for the density frequency histogram irregular intervals are also allowed).

5.2.1 Demonstration of histogram

We will construct a histogram for the sample of section 5.1.1. The number of intervals should be taken k = ln 21 / ln 2 = 4.4. By rounding up, we choose 5 intervals. The range of the sample values is [217, 1064]. After rounding, we get the range [200, 1100] with Δ = (1100 − 200) / 5 = 180. The rest of calculations are given in tabular form in Table 5.7 and the histogram is illustrated in Fig. 5.2. For comparison, we also plot the theoretical probability density function of the normal distribution (see section 5.4).

* In Excel this can be done by the function CountIf.

Table **5.**7 Calculations for the histogram of the sample of Table 5.2.

Class rank	Class limits	Absolute frequency n_i	Relative frequency n_i / n	Frequency density $\varphi = n_i / (n\,\Delta)$
	200			
1		1	0.048	0.00026
	380			
2		4	0.190	0.00106
	560			
3		6	0.286	0.00159
	740			
4		6	0.286	0.00159
	920			
5		4	0.190	0.00106
	1100			

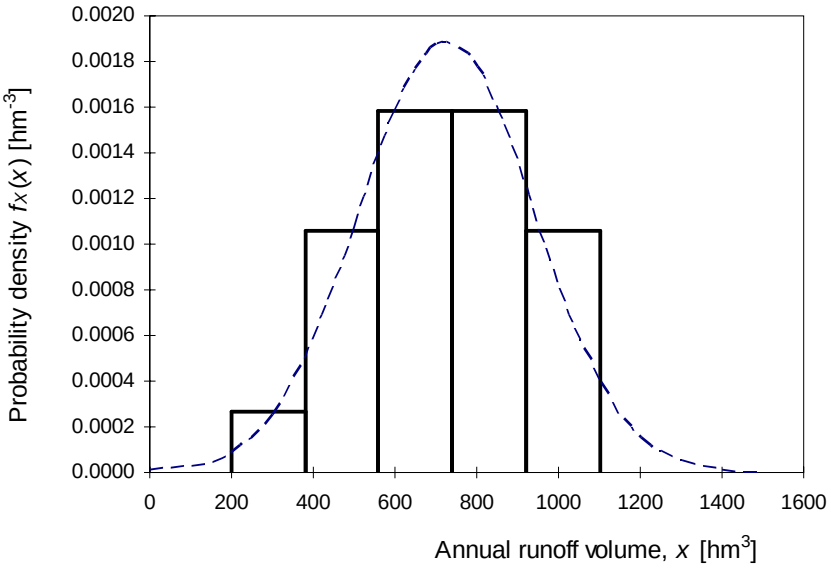

Fig. **5.**2 Histogram of the sample of Table 5.2. For comparison the probability density function of the normal distribution $N(725, 211.5)$ is plotted (dotted line).

5.3 Empirical distribution function

The histogram is the empirical equivalent of the probability density function; likewise the empirical equivalent of the distribution function is the *empirical distribution function*. In principle, such an empirical function may be constructed from the histogram, by integrating with respect to x, hence getting an increasing broken line that corresponds to some type of a distribution function. However, the introduction of the empirical distribution function may be done in a more direct and objective manner, bypassing histogram, which has some degree of subjectivity, due to the arbitrary selection of the intervals and their limits.

5.3.1 Order statistics

Let X be a random variable with distribution function $F(x)$ and $X_1, X_2, \ldots, X_n$ a sample of it. From realizations $x_1, x_2, \ldots, x_n$ of the variables $X_1, X_2, \ldots, X_n$, we take the maximum value $x_{(1)}$

:= max(x_1, x_2, …, x_n).* This can be thought of as a realization of a variable $X_{(1)}$. Likewise, we can construct the variables $X_{(2)}$ (corresponding to $x_{(2)}$, the second largest value), $X_{(3)}$, …, $X_{(n)}$. The random variables $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$ are called *order statistics*. Obviously, for each realization the values $x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(n)}$ represent the observed sample sorted in decreasing order.

5.3.2 Classical empirical distribution function

The classical empirical distribution function is a staircase-like function defined by

$$\hat{F}(x) = \frac{n_x}{n} \tag{5.2}$$

where n_x is the number of sample values that do not exceed the value x. $\hat{F}(x)$ is a point estimate of the unknown distribution function of the population $F(x)$.

5.3.3 Plotting position

Plotting position q_i of the value $x_{(i)}$ of the sorted sample is the empirical exceedence probability of this value. Based on the classical definition of the empirical distribution, for $x = x_{(1)}$ we will have $n_x = n$, and generally for $x = x_{(i)}$ we will have $n_x = n + 1 - i$. Therefore, the empirical distribution function is

$$\hat{F}\left(x_{(i)}\right) = \frac{n+1-i}{n}, \quad i = 1, \ldots, n \tag{5.3}$$

Thus the plotting position, i.e. the empirical exceedance probability is

$$q_i = \hat{F}^*\left(x_{(i)}\right) = 1 - \hat{F}\left(x_{(i)}\right) = \frac{i-1}{n}, \quad i = 1, \ldots, n \tag{5.4}$$

We observe that for $n = 1$ the above equation assumes zero exceedance probability. Thus, for example from an annual rainfall sample with maximum value $x_{(1)}$ = 1800 mm, we would conclude that the probability of an annual rainfall more than 1800 mm is zero. Evidently, this is a wrong conclusion; rainfall depths more than those observed are always possible.

To avoid the above problem we use the random variable

$$U_i = F^*\left(X_{(i)}\right) = 1 - F\left(X_{(i)}\right) \tag{5.5}$$

A point estimate† of this variable is, simultaneously, an estimate of q_i. From first glance, it seems impossible to calculate values of U_i from the sample, given that $F(x)$ is an unknown function. However, it can be shown‡ that (for random samples) the distribution of U_i is independent of $F(x)$ and has mean§

---

* Notice the difference in notation: x_1 is the value first in time and $x_{(1)}$ is the (first) largest of all x_i.

† More precisely, and according to the terminology of chapter 3, this is a prediction of the variable, since U_i is a random variable and not a parameter.

‡ This results from the distribution function of the order statistics (see e.g. Papoulis, 1990, p. 207-208) after appropriate substitution of variables.

§ More precisely, U_i has beta distribution function (see chapter 6), with parameters i and $n - i + 1$.

$$E[U_i] = \frac{i}{n+1} \tag{5.6}$$

and variance

$$\mathrm{Var}[U_i] = \frac{i(n-i+1)}{(n+1)^2(n+2)} \tag{5.7}$$

The simplest estimate of U_i is its mean, namely

$$q_i = \frac{i}{n+1} \tag{5.8}$$

which is known in literature as *Weibull plotting position*. This is an unbiased estimate of the exceedance probability, because $q_i = E[U_i] = E[F^*(X_{(i)})]$. We observe that with this estimation method we have eliminated the problem of a zero q_i for $i = 1$. Indeed, for $i = 1$ we obtain $q_i = 1 / (n + 1)$ and for $i = n$, $q_i = n / (n + 1)$.

Table 5.8 Alternative formulae for empirical exceedance probabilities (plotting positions)*

Name	Formula $q_i =$	Constant $a =$	Return period of maximum value $T_1 =$	Applicability
Weibull	$\frac{i}{n+1}$	0	$n + 1$	All distributions, unbiased estimation of exceedance probability
Blom	$\frac{i-0.375}{n+0.25}$	0.375	$1.6\,n + 0.4$	Normal distribution, unbiased estimation of quantiles
Cunnane	$\frac{i-0.4}{n+0.2}$	0.4	$1.667\,n + 0.33$	Broad range of distributions, approx. unbiased estimation of quantiles
Gringorten	$\frac{i-0.44}{n+0.12}$	0.44	$1.786\,n + 0.21$	Gumbel distribution†
Hazen	$\frac{i-0.5}{n}$	0.5	$2\,n$	The oldest proposed estimate; today it tends to be abandoned

Equation (5.8) is the most popular for the estimation of exceedance probabilities in engineering applications, but not the only one. Other similar equations have been developed in order to provide unbiased estimations of quantiles, namely to satisfy (approximately) the condition

$$F^{-1}(q_i) = E[X_{(i)}] = E[F^{-1}(U_i)] \tag{5.9}$$

In that case, this estimation, as opposed to (5.8), does depend on the distribution function $F(x)$. The various equations that have been developed are expressed by the general formula

$$q_i = \frac{i-a}{n+1-2a} \tag{5.10}$$

* See also Stedinger et al. (1993) where additional formulae are also given.
† See chapter 6.

where a is a constant (< 1). This equation is antisymmetric, since $q_i = 1 - q_{n+1-i}$ and also incorporates (5.8) as a special case ($a = 0$). Table 5.8 lists the most frequently used formulae for calculating the plotting position along with the corresponding values of constant a. Application of the different formulae results in very similar values, except for very low values of i and mainly for $i = 1$, where the differences are appreciable (see col. 4 in Table 5.8). The value for $i = 1$ is of great importance in engineering applications, because it gives the empirical exceedance probability of the maximum observed value, i.e. $T_1 = 1 / q_1$.

5.3.4 Probability plots

Estimating the plotting position for each value of the sample using one of the above formulae, we construct a set of n points $(x_{(i)}, q_i)$ or $(x_{(i)}, 1 - q_i)$, which can be presented graphically to provide an overview of the distribution function. Initially, this could be done on a regular decimal plot, thus resulting in a graph similar to Fig. 2.1 or Fig. 2.3b, except that, instead of a staircase-like or a continuous line, we will get just a set of points. However, in engineering applications, since the information obtained by such a graph is very essential, we wish to be more systematic in plotting. In particular, we wish to obtain a linear arrangement of the points through appropriate transformations of the axes. This facilitates several purposes, such as easier drawing, more precise comparison of theoretical and empirical distribution, easier graphical extrapolation beyond sample limits etc. Plots on which the axes are designed via appropriate transformations, to represent the graphs of specific distribution functions as straight lines, are called *probability plots*. There exist commercial papers (like the logarithmic paper) constructed so as to incorporate the appropriate transformation for a specific distribution (e.g. the normal distribution) which can be readily used to make a probability plot. However, it is easy to construct such plots using computer tools.

Let us take, for instance, the normal distribution $N(\mu, \sigma)$. If we represent graphically the function $F(x)$ with horizontal axis $h = F$ and vertical $v = x$, we will obtain a shape like $\curvearrowright$. On the other hand, we know that $x = \mu + \sigma z_F$, where z_F the F-quartile of the standard normal distribution $N(0, 1)$. Hence, if we set the horizontal axis as $h = z_F$, then the equation to plot will be $v = \mu + \sigma h$, which is a straight line. This is equivalent to transforming the horizontal axis as $h = z_F = F_0^{-1}(F)$, where $F_0^{-1}(\,)$ is the inverse of the standard normal distribution. Through appropriate transformations of the horizontal or/and the vertical axis, we may achieve linearization of other distribution functions, as we will see in more detail in chapter 6.

Since there is one-to-one correspondence between the quantities F and z_F, the marking of the horizontal axis may be done is units of F instead of z_F, which facilitates the interpretation of the graph. Moreover, the marking of the horizontal axis may be done in terms of the exceedence probability $F^* = 1 - F$ or versus the return period $T = 1 / F^*$. An example of a normal distribution plot is shown in Fig. 5.3, where two different markings of the horizontal axis (z_F and F^*) are simultaneously illustrated.

The graphical representation of the set of points $(x_{(i)}, q_i)$ in a normal distribution plot (namely $h = z_{1-q_i}, v = x_{(i)}$) will give an almost linear arrangement of points, provided that the distribution of X is normal. Hence, this plot provides a graphical way for checking the normality of the distribution of a sample. The above are clarified via the following example.

5.3.5 Demonstration of numerical probability plot

We will construct a normal probability plot of the sample of section 5.1.1. For the calculation of the empirical exceedence probabilities we use the formulae of Weibull (unbiased estimation of exceedence probability) and Blom (unbiased estimation of the normal distribution quantiles, see Table 5.8). The calculations are very simple, given that the sample has been put already in descending order (Table 5.5) and are shown in Table 5.9. For a manual plot on normal probability paper, the last two columns are not necessary. Otherwise, they are necessary because the normal probability plot (shown in Fig. 5.3) is a plot of observed values x_i against values z_{1-q_i} of the standard normal distribution. The latter either are taken from the normal distribution table (Table A1, Appendix), or are calculated using numerical methods*. The empirical exceedence probabilities for this sample are shown in Fig. 5.3, where for comparison, the theoretical normal distribution function it is also plotted (see section 5.5.4).

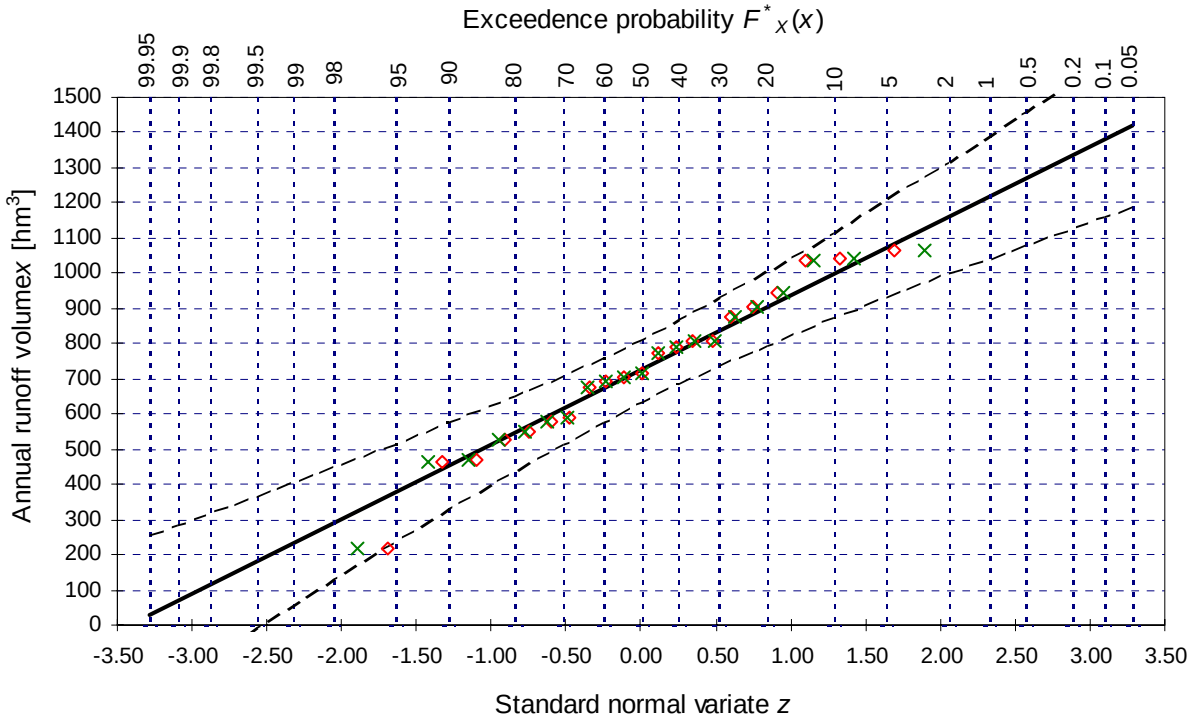

Fig. 5.3 Example of normal probability plot of the empirical distribution function using Weibull (diamonds) and Blom (symbols ×) plotting positions. For comparison, the theoretical normal distribution function $N(725, 211.5)$ (see section 5.4.2) is also plotted (continuous line) along the corresponding 95% confidence curves (dashed curves, see section 5.6.1).

* In Excel the function to calculate z_{1-q} from $1 - q$ is NormSInv.

Table 5.9 Demonstration of calculation of empirical exceedence probabilities.

Rank	Value	Empirical exceedence probability		Value of standardized normal variable	
		Weibull	Blom	Weibull	Blom
i	x_i	$q_i = \frac{i}{n+1}$	$q_i = \frac{i-0.375}{n+0.25}$	z_{1-q_i}	z_{1-q_i}
1	1064	0.045	0.029	1.691	1.890
2	1042	0.091	0.076	1.335	1.429
3	1037	0.136	0.124	1.097	1.158
4	942	0.182	0.171	0.908	0.952
5	906	0.227	0.218	0.748	0.780
6	874	0.273	0.265	0.605	0.629
7	807	0.318	0.312	0.473	0.491
8	807	0.364	0.359	0.349	0.362
9	788	0.409	0.406	0.230	0.238
10	772	0.455	0.453	0.114	0.118
11	715	0.500	0.500	0.000	0.000
12	705	0.545	0.547	-0.114	-0.118
13	695	0.591	0.594	-0.230	-0.238
14	674	0.636	0.641	-0.349	-0.362
15	588	0.682	0.688	-0.473	-0.491
16	580	0.727	0.735	-0.605	-0.629
17	552	0.773	0.782	-0.748	-0.780
18	529	0.818	0.829	-0.908	-0.952
19	469	0.864	0.876	-1.097	-1.158
20	462	0.909	0.924	-1.335	-1.429
$n = 21$	217	0.955	0.971	-1.691	-1.890

5.4 Selection and fitting of the theoretical distribution function

In sections 5.1 and 5.2 the aim was to summarize a sample, which is part of descriptive statistics. Section 5.3, in addition to summarizing a sample, dealt also with statistical estimation of population properties, specifically the distribution function. However, we were able to make such estimations for a few values of the random variable only, those that were values of the sample. This could be combined with some empirical techniques, for instance interpolation, to make inferences for other values of the random variable. Thus, we could make an empirical interpolation of any value provided that it lies within the range defined by the minimum and maximum values in the observed sample. The range of such estimations would be limited. In engineering design, we usually have to deal with values far beyond the observed range (e.g. to estimate design quantities for return periods 100, 1000 or 10 000 years based on a sample of, say, 20-50 years), i.e. to make extrapolations. To this aim, we should follow a different path, which should be also able to provide interval estimates of the quantities of interest.

This would be easy if we knew the distribution function of the population. Generally, the distribution of the population could be any function with the properties described in section 2.4. Its precise knowledge would require to have measured the entire population, or, at least, to have a sample much longer than the return period for which an estimation is sought. Apparently, this is infeasible and thus the remaining solution is to *hypothesize* a *probability model* for the population. The term probability model refers to one of the typical distribution functions of the probability theory that have a specific, relatively simple, mathematical expression. The most typical case is the normal distribution discussed in section 2.10.2. Other examples will be provided in chapter 6. Certainly, the use of a probability model is always an approximation of the reality. The distributions of geophysical variables are not identical to the simple models of the probability theory.

The selection of an appropriate model is guided by the following:

1. *The probability theory*. In some cases, there are theoretical reasons because of which a particular hydrological or geophysical variable is expected to have a particular distribution type. For instance, according to the central limit theorem, the annual rainfall in a wet area is expected to follow a normal distribution (see section 2.10.1). Another principle that can provide theoretical justification of a probability model is the principle of maximum entropy (Koutsoyiannis, 2005).
2. *The general empirical experience*. In many cases, accumulated hydrological or geophysical experience indicates that specific variables tend to follow particular distribution types, even if there are not apparent theoretical reasons pointing to the latter. For instance, the monthly runoff has been very often modelled using gamma or log-normal distributions (see chapter 6).
3. *The properties of the specific sample*. The statistical characteristics of the observed sample help us to choose or exclude a particular distribution type. For instance, if the sample coefficient of skewness has a value close to zero, then we can choose the normal (or another symmetric) distribution. Conversely, if the coefficient of skewness differs substantially from zero, we should exclude the normal distribution.

Certainly, the suitability of a specific distribution type is not ensured by the above criteria, which are just indications of suitability. The testing of the suitability of the distribution is done a posteriori. After estimating its parameters, we examine the goodness of its fit to the empirical distribution function. Initially, this may be done empirically, on the basis of the graphical representation of the empirical and the theoretical distribution functions on an suitable probability paper. More objective results are achieved by means of formal statistical tests, as described in section 5.5.

5.4.1 Indications of suitability of the normal distribution for geophysical variables

So far, we have referred many times to indications of the suitability of the normal distribution for describing geophysical variables. Next, we list all these indications of suitability.

1. *Theoretical criterion based on the central limit theorem*. We examine whether the variable under study is a sum of various natural components, which should obey (even approximately) the assumptions of the central limit theorem. This criterion is theoretical and does not require numerical calculations. A similar theoretical criterion is provided by the principle of maximum entropy, independently of the central limit theorem (Koutsoyiannis, 2005).
2. *Numerical criterion based on the coefficient of skewness*. A sample coefficient of skewness that is almost zero is a strong indication of the suitability for the normal distribution.
3. *Numerical criterion based on the coefficient of variation*. Let X be random variable representing a physical quantity. In most cases, X can take only positive or zero values, whereas negative ones have no physical meaning. However, the normal distribution allows negative values of X. Thus, in theory, the normal distribution cannot represent physically nonnegative variables, except approximately. To ensure a satisfactory approximation, the probability $P\{X < 0\}$ must be very low, so to be ignored, namely $P\{X < 0\} \le \varepsilon$ where ε an acceptably low probability, e.g. $\varepsilon < 0.02$. If $Z = (X - \mu_X) / \sigma_X$ is the corresponding standard normal variable, then $P\{Z < -\mu_X / \sigma_X\} \le \varepsilon$. If z_ε is the ε-quantile of the standard normal distribution, then, equivalently, $C_{vX} = \sigma_X/\mu_X \le -1/z_\varepsilon$. For $\varepsilon = 0.02$ we get $z_\varepsilon \approx -2$, so $C_{vX} \le 0.5$. Likewise, for $\varepsilon = 0.00005$ we get $z_\varepsilon \approx -4$, so $C_{vX} \le 0.25$. Hence, we conclude that if $C_{vX} \le 0.25$ we have a very strong indication of suitability of the normal distribution. If $C_{vX} > 0.5$, the use of the normal distribution should be excluded. For intermediate values of the coefficient, the normal distribution may be acceptable but with lower degree of approximation.
4. *Graphical criterion based on the synoptic depiction of the sample*. As referred in section 5.1 a symmetric box plot of the sample, without unjustifiably large number of outside points, is an indication of the suitability of the normal distribution.
5. *Graphical criterion based on the empirical distribution function*. The linear arrangement of the series of points of the empirical distribution function, in a normal probability plot, is a strong indication of the suitability of the normal distribution.

The above criteria are simple indications and cannot be thought of as statistical proofs of the suitability of the normal distribution.

5.4.2 Demonstration of fitting the normal distribution

The fitting of the normal distribution on the sample of section 5.1.1 is very simple. The parameters of the distribution are $\mu = \overline{x} = 725.0$ hm^3, $\sigma = s_X = 211.5$ hm^3 (the value $\sigma = s_X^* = 216.7$ hm^3 is also acceptable). The normal distribution function with these parameters has been plotted in Fig. 5.3 and the corresponding probability density function in Fig. 5.2. The reader can confirm that in the example under study all indications of suitability of the normal distribution listed in section 5.4.1 are validated. In section 5.5.2 we will provide a statistical test of suitability of the normal distribution.

5.5 Testing the goodness of fit of a probability model

After adopting a certain distribution function to model a physical variable and estimating its parameters, the next step is to test the fitting of this distribution to the observed sample. The test is based on the statistical theory of hypothesis testing that was summarized in section 3.6. Various statistical tests have been developed, which can be applied for testing the goodness of fit of a distribution function. We present the most classical of them, the *χ^2 (chi-square) test*. Other statistical tests often used in engineering applications are the *Kolmogorov-Smirnov test* (see e.g. Benjamin and Cornell, 1970, p. 466; Kottegoda, 1980, p. 89) and the more recent *probability plot correlation coefficient test* (see e.g. Stedinger et al., 1993, p. 18.27).

5.5.1 The χ^2 test

The χ^2 test is based on comparing the theoretical distribution function to the empirical one. The comparison is made on a finite set of selected points x_j of the domain of the random variable, and not on the observed values x_i of the sample. The null hypothesis H_0 and its alternative H_1 are

$$H_0: F(x_j) = F_0(x_j) \text{ for all } j, \quad H_1: F(x_j) \neq F_0(x_j) \text{ for some } j \tag{5.11}$$

where $F(x)$ the unknown true distribution function and $F_0(x)$ the hypothesized distribution. $F_0(x)$ may be completely known, in terms of its mathematical expression as well as its parameters values, prior to the examination of the specific sample. In this case, the null hypothesis is named *perfect*. However, the parameters values are most usually calculated from the sample and so we speak about an *imperfect null hypothesis*.

The control points x_j, $j = 0, \ldots, k$ partition the domain of the random variable in k classes, namely intervals of the form $(x_0, x_1]$, $(x_1, x_2]$, …, $(x_{k-1}, x_k]$. For the hypothesized distribution function $F_0(x)$, the probability of finding a randomly selected point in $(x_{j-1}, x_j]$ is obviously

$$p_j = F_0(x_j) - F_0(x_{j-1}) \tag{5.12}$$

and therefore the expected number of sample points that would be located within this class is $l_j = n\, p_j$, where n is the sample size. Apparently, a small departure between n_j and l_j, namely a small $|n_j - n\, p_j|$, is in favour of the suitability of the distribution $F_0(x_j)$ and hence of the non-rejection of the null hypothesis. The *Pearson's test statistic* defined by

$$Q := \sum_{j=1}^{k} \frac{(N_j - n\, p_j)^2}{n\, p_j} \tag{5.13}$$

where N_j is the random variable whose realization is n_j, is an aggregated measure of the differences between the actual and the theoretical number of points in all classes. If the null hypothesis is perfect, the distribution of Q is χ^2 with $k - 1$ degrees of freedom. In the most

usual case of imperfect null hypothesis, the number of degrees of freedom is $k - r - 1$, where r is the number of parameters that are estimated from the sample.*

In the most common version of the χ^2 test the classes are chosen so that the probabilities p_j are equal for all classes j. In this case, equation (5.13) simplifies to

$$Q := \frac{k}{n} \sum_{j=1}^{k} N_j^2 - n \tag{5.14}$$

The advantage of this version is that it specifies the class limits for a given number of classes k and thus it is more objective For choosing the number of classes k, the following two conflicting rules are followed:

- Necessarily, it must be $k \geq r + 2$, where r is the number of parameters of the distribution that are estimated from the sample.
- Generally, it is suggested (see e.g. Benjamin and Cornell, 1970, p. 465; Kottegoda, 1980, p. 88) that the theoretical number of points in each class must be grater than 5, which results in $k \leq n / 5$.

For small samples, these two rules may be not satisfied simultaneously, hence we satisfy the first one only.

The algorithm for applying the χ^2 test is described in the following steps:

1. We choose the number of classes k, according to the above rules.†
2. We divide the probability interval [0, 1] in k equal sub-intervals with limits $u_j = j / k$ ($j = 0, \ldots, k$).
3. We calculate the class limits x_j (the value x_j is the u_j-quantile of the variable).
4. We count the number of points n_j in each class (this step is simplified if the sample is already sorted in descending or ascending order).
5. From (5.14) (or (5.13)), we calculate the value q of the Pearson statistic.
6. For a chosen significance level α, we calculate the critical value of the test statistic $q_c = q_{1-\alpha}$. For this purpose, we use the χ^2 distribution with $k - r - 1$ degrees of freedom, where r is the number of distributional parameters estimated from the sample (see Table A2 in Appendix).
7. We reject the null hypothesis if $q > q_c$.

The algorithm is clarified in the following example.

---

* Theoretical consistency demands that the maximum likelihood method is used for parameter estimation; however, this is often neglected in applications.

† The choice of the number of classes can be made using the formula (see Kottegoda, 1980, p. 88):

$$k = 2^{1.2} [(n - 1) / z_{1-\alpha})]^{0.4}$$

where $z_{1-\alpha}$ the $(1 - \alpha)$-quantile of the normal distribution and α the significance level of the test. Kendall and Stuart (1973, p. 455) provide a more analytical method for choosing the number of classes, which however is for large samples that are rarely available in practice.

5.5.2 Demonstration of testing the goodness of fit

Continuing the numerical example started in section 5.1.1, we will test the suitability of the normal distribution that has been already fitted (section 5.4.2), with parameters $\mu = \overline{x} = 725.0$ hm^3, $\sigma = s_X = 211.5$ hm^3.

The number of the parameters of the distribution is $r = 2$ and the sample size is $n = 21$. According to the above discussion, the number of classes k must satisfy the relationships

$$k \geq 2 + 2 = 4, \quad k \leq 21 / 5 = 4.2$$

that hold for $k = 4$. Therefore, we take $k = 4$.

The calculations for steps 2-4 of the above algorithm are summarized in Table 5.10. The calculation of the limits of the variable is done as usual; for instance, the upper limit of the first class is

$$x_1 = 725.0 - 0.675 \times 211.5 = 528.3$$

Table 5.10 Elementary calculations demonstrating the χ^2 test.

Class	1	2	3	4	
Probability limits	0	0.25	0.5	0.75	1.0
Variable limits	-∞	582.3	725.0	867.7	+∞
Actual number of points	6	5	4	6	

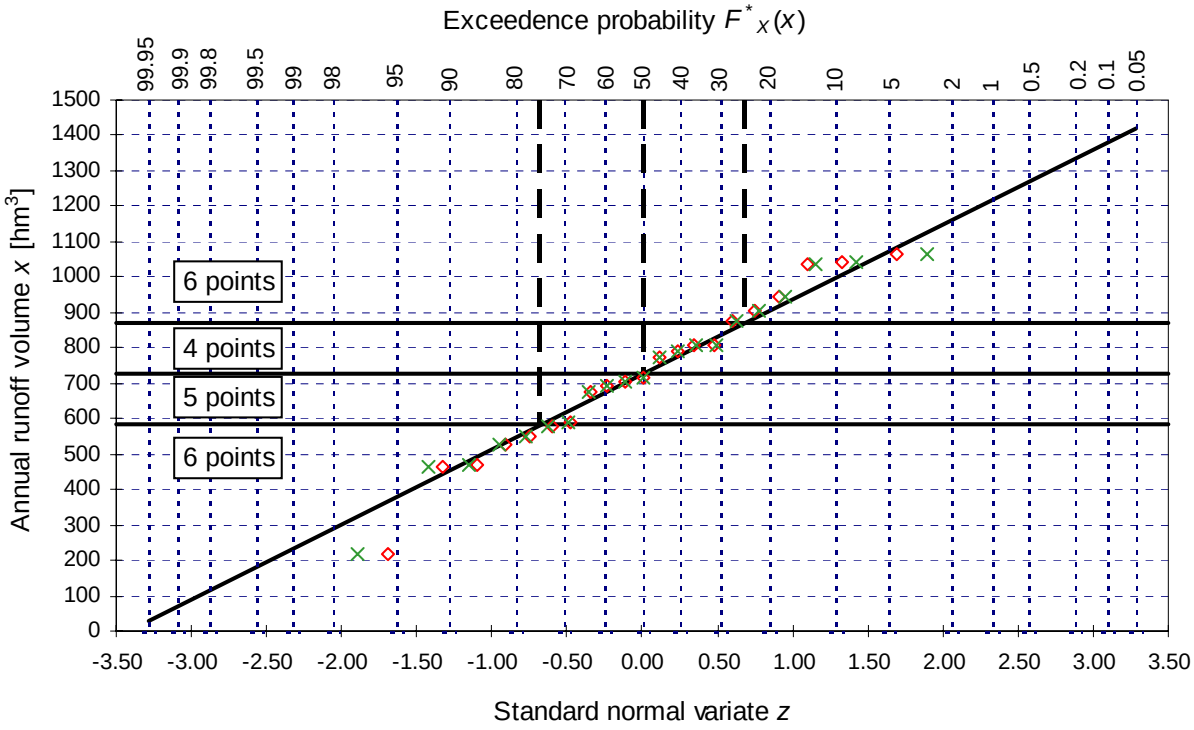

Fig. 5.4 Explanatory sketch for the numerical example of section 5.5.2.

For the sake of demonstration (as it is not part of the test), we provide in Fig. 5.4 graphical depiction of the classes and their actual number of points on a normal probability plot.

From (5.14) we obtain

$$q = (4/21) \times (6^2 + 5^2 + 4^2 + 6^2) - 21 = 0.52$$

For significance level $\alpha = 0.05$ the critical value of the variable is

$$q_c = \chi^2_{0.95}(1) = 3.84$$

(as derived from Table A2 in Appendix for $u = 1 - \alpha = 0.95$ and number of degrees of freedom = 4 – 2 – 1 = 1). Hence, $q < q_c$ and the normal distribution is accepted.

5.6 Statistical prediction

Statistical prediction in engineering applications aims at estimating the value of a physical quantity that corresponds to a given exceedence probability (or return period). Provided that a specific probability model is already set up and fitted to the sample under interest, this prediction is computationally done applying the methods described in chapter 3. The prediction may be either point or interval, as demonstrated in the following example.

5.6.1 Demonstration of statistical prediction

Completing the numerical example started in section 5.1.1, we wish to estimate the 100-year maximum and minimum annual runoff volume of the Evinos river basin upstream of Poros Reganiou, as well as its 95% confidence limits. We apply the same procedure as in section 3.4.7. As the sample size is very small in comparison to the return period of 100 years, we expect that the confidence intervals will be wide (high uncertainty).

We calculate first the point estimates. For the 100-year maximum runoff volume the probability of non-exceedence is $u = 1 - 1/100 = 0.99$ and $z_u = 2.326$ (e.g. from Table A1 in the Appendix). Thus, the point estimate is

$$x_u = 725.0 + 2.326 \times 211.5 = 1216.9 \text{ hm}^3$$

Likewise, for the 100-year minimum runoff volume, the probability of non-exceedence is $u = 1 / 100 = 0.01$ and $z_u = -2.326$, so

$$x_u = 725.0 - 2.326 \times 211.5 = 233.1 \text{ hm}^3$$

We proceed with the calculation of confidence limits. For $\gamma = 95\%$ and $z_{(1+\gamma)/2} = 1.96$, the limits for the 100-year maximum runoff volume are (equation (3.46)) :

$$\hat{x}_{u1} = 1216.9 - 1.96\sqrt{1 + \frac{2.326^2}{2}}\frac{211.5}{\sqrt{21}} = 1042.8$$

$$\hat{x}_{u2} = 1216.9 + 1.96\sqrt{1 + \frac{2.326^2}{2}}\frac{211.5}{\sqrt{21}} = 1391.0$$

Likewise, the limits for the 100-year minimum runoff volume are:

$$\hat{x}_{u1} = 233.1 - 1.96\sqrt{1+\frac{2.326^2}{2}}\frac{211.5}{\sqrt{21}} = 59.0$$

$$\hat{x}_{u2} = 233.1 + 1.96\sqrt{1+\frac{2.326^2}{2}}\frac{211.5}{\sqrt{21}} = 407.2$$

Repeating these calculations for several other return periods we have determined a series of point estimates and confidence limits which we have plotted in Fig. 5.3. More specifically, connecting the points of the confidence limits in the graph we have obtained the 95% confidence curves of the distribution. We observe the all points of the observed sample lie within these confidence curves; particularly the lowest observed value (217 hm^3 for the year 1989-90) is just on the border, which reflects the severity of the drought of 1989-90.

Acknowledgement I thank Andreas Efstratiadis for his help in translating Greek texts into English.

References

Benjamin, J.R., and C.A. Cornell, *Probability, Statistics and Decision for Civil Engineers*, McGraw-Hill, 1970.

Hirsch, R. M., D.R. Helsel, T.A. Cohn, and E.J. Gilroy, Statistical analysis of hydrologic data, Chapter 17 in *Handbook of Hydrology*, edited by D. R. Maidment, McGraw-Hill, 1993.

Kendall, M.G., and A. Stuart, *The Advanced Theory of Statistics*, Vol. 2, Inference and relationship, Third edition, Charles Griffin & Co., London, 1973.

Kottegoda, N.T., *Stochastic Water Resources Technology*, Macmillan Press, London, 1980.

Koutsoyiannis, D., Uncertainty, entropy, scaling and hydrological stochastics, 1, Marginal distributional properties of hydrological processes and state scaling, *Hydrological Sciences Journal*, 50(3), 381-404, 2005.

Papoulis, A., *Probability and Statistics*, Prentice-Hall, New Jersey, 1990.

Stedinger, J.R., R.M. Vogel, and E. Foufoula-Georgiou, Frequency analysis of extreme events, Chapter 18 in *Handbook of Hydrology*, edited by D. R. Maidment, McGraw-Hill, 1993.