3 BASIC DATA FOR CALCULATING THE STORAGE FUNCTION OF RESERVOIRS

The basic data for the calculation of the storage function of reservoirs consist of
- hydrological records
- the needs of users and consumers.

3.1 HYDROLOGICAL RECORDS AND THEIR PROCESSING

The basic hydrological records for the mathematical solution of the reservoir function are the flow series of the stream cross-sections. To what extent and how these hydrological records are processed depends on
- the aim and significance of the reservoir,
- the stage of the design,
- the character and range of the hydrological data.

3.1.1 Selection and evaluation of hydrological series

Longer hydrological series are usually a more reliable basis than shorter ones. However, the so-called *representative character of the series* is important. A series represents the hydrological conditions in the watershed well if it includes data leading to reliable results in the solution of flow-regime problems.

Other observations (outflow from adjacent watersheds, precipitation, etc.) can increase the representative value of a given shorter series within the framework of a long series, if
- they cover a long period
- they include the years of the studied short-term hydrological series
- their relationship to the data in the hydrological series is close enough.

If these conditions are fulfilled, the given hydrological series can be prolonged. Sufficiently reliable parameters can be expected in the prolonged series, if the series in question is not shorter than 15 years and if the correlation coefficient for the synchronous values of the series and their analogues does not amount to less than 0.90. A period of about 20 years is considered to be long enough for the calculation of seasonal control. For over-year flow periods we need at least 30 years of data which naturally have to be assessed as to their representativeness and suitability.

In Czechoslovakia better results are usually achieved by hydrological analogies of the discharges from two analogous watersheds, since Czechoslovakia has a dense network of watergauges with comparatively long series: the relationship between precipitation and discharge will usually be less close.

Reservoirs on streams usually have more complete and reliable hydrological records than those on small rivers.

The disadvantage of real years is that they introduce a random element into the results of the calculations. Statistical methods are used to adjust the measured data.

The adjusted measured data, especially for the decisive years or periods, result in hypothetical hydrological records which are more reliable than the measured ones.

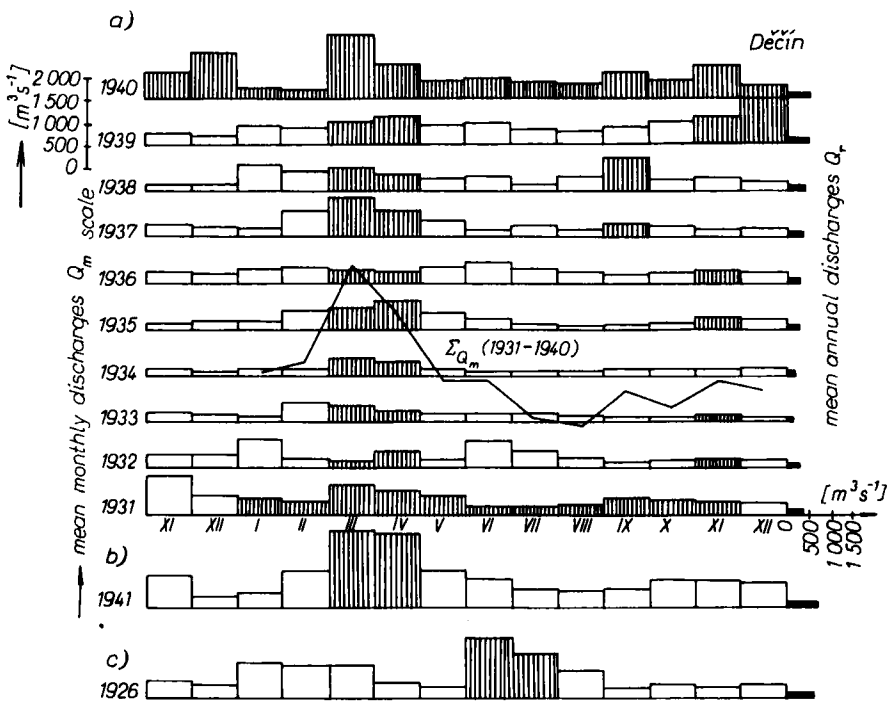

Fig. 3.4 Hydrograph of mean monthly discharges of the river Labe at Děčín

In the past, a hypothetical year was considered to be favourable basis for calculating reservoirs with annual flow control. It was worked out by selecting in the hydrological series of the flow, in individual decades or months, those with the same required probability of exceedance. This is how a hypothetical year is created, and it usually has a smoother pattern than any real year. Its demands regarding the reservoir volume are lower than for a real year with the same exceedance probability; it cannot therefore be recommended as a basis for the calculation of reservoir volumes.

More promising is the method of hydrological phase characteristics (Potapov, 1951) which is based on the assumption that a number of characteristic phases (periods) of the annual discharge regime, repeated every year and differing only in intensity or in the date of their beginning or end, can be found for each type of river.

The flow conditions of Middle European rivers are much more complicated and heterogeneous (as shown in Fig. 3.4, depicting the average monthly Q_m discharges of the river Labe in Děčín for the decade 1931–1940). The sum of the monthly discharges for the identical months of the entire decade (Q_m curve for 1931–1940) has a higher discharge in March and April and a minimum discharge in July and August, as well as higher flows in autumn and winter. In winter the flow is not very low, as even in winter there is rainfall and ice melts in the catchment area of the Labe.

Comparing the flow in the respective years to this cumulative curve for the whole decade, a good similarity can only be seen for the years 1931, 1935, 1937 and 1940; in the other years the flow is distributed differently over the year. The spring flow from snow water is not always very high (in 1936, 1938); the flow is also strongly influenced by summer rainfall (in 1932, 1936).

A good example is the comparison of the two years with the highest flow in the entire 100-year period, i.e., 1941 (Fig. 3.4b) and 1926 (Fig. 3.4c). In 1941, the highest flow period occurred in March and April, while in 1926 it occurred in the summer months of June and July.

This unsteady nature of the flow has to be taken into account when designing reservoirs in such hydrological conditions, and only after careful deliberation can methods which proved successful on rivers with a simpler flow regime be used. For the design—but mainly for the operation of these reservoirs—it is important to know that one cannot rely on regular filling with the spring floods; these often have values of a long-term average Q_a or even less (1932, 1933, 1934, 1936).

Often, not only the annual discharge regime, but also the water demand (withdrawal from the reservoirs) changes considerably during the year. Water is withdrawn in the vegetation period for irrigation purposes and in the winter mainly for power production. The relationship between the flow in the rivers and the withdrawal from the reservoirs—their size and their timing (see Fig. 1.19b)—always has to be taken into consideration.

The variability of the beginning and end of the individual flow periods (phases) and the flow in the individual periods must be analysed. A hydrograph adjusted according to such analyses does not lead to an unrealistic equalization of the flow, but preserves the character of the annual discharge regime.

A design period of several years is also selected according to the analysis of the entire series. The chosen period should: 1. have a mean discharge close to the long-term average, 2. include years with various stream flows in characteristic groups, 3. have a variation coefficient of the annual discharges close to that for the entire series.

The fulfillment of these conditions is obvious from the mass curve of the deviation from the mean of the module coefficients $(k_i - 1)$ and the mass curve of the values $(k_i - 1)^2$ where

$$k_i = \frac{Q_{r,i}}{Q_a} \tag{3.1}$$

$Q_{r,i}$ is the mean annual flow in an optional ith year,
Q_a is the long-term mean annual flow.

The sum of all n values $(k_i - 1)$ in an n-year long hydrological series is zero. The mass curve $\sum(k_i - 1) = f(t)$ plotted in right angle coordinates therefore terminates at the end of the series in the horizontal axis of the abscissae. Any other horizontal straight line intersecting the mass curve of the module coefficients is always traced between two points of intersection of the period in which $\sum(k - 1) = 0$, or the period with the same average flow as the whole of the series. A period selected in this way therefore fulfills the first condition for a correct determination of the design period.

Obviously the mass curve of the mean annual flow $\sum O_{r,i} = f(t)$ can serve the same purpose. If drawn from the pole at the height Q_a, it terminates at the end of the series in the horizontal straight line drawn through the beginning of the mass curve. Any other horizontal straight line intersecting this mass curve also traces, always between two points of intersection, periods with a flow identical with the entire series. If, generally speaking, the pole is not at height Q_a, all data given so far about the horizontal straight line are true for the line connecting the beginning and the end of the mass curve, and for all lines parallel to it, intersecting the mass curve.

The mass curve $\sum(k_i - 1)^2 = f(t)$ (Andreyanov, 1957) also helps to determine the period in which the variation coefficient is close to the value C_v of the entire n-year series. It is not difficult to draw this curve, so that at the end of the n-year series, it ends in a horizontal line drawn from the starting point of the mass curve. We either first calculate the sum $\sum_{i=1}^{n}(k_i - 1)^2$ and draw the base of the oblique coordinates of the mass curve to intersect its starting point at the end of the n-year series, at a perpendicular distance of $-\sum_{i=1}^{n}(k_i - 1)^2$ from the horizontal line, or we draw the mass curve $\sum(k_i - 1)^2$ from the pole at the point of the mean value

$$(k - 1)^2 = \frac{\sum_{i=1}^{n}(k_i - 1)^2}{n} \tag{3.2}$$

Thus each horizontal line intersecting the mass curve $\sum(k - 1)^2$, always traces

an m-year period between two points of intersection. Its mean value for $(k-1)^2$ corresponds to the quantity for the entire period, i.e.

$$\frac{\sum_{i=1}^{n}(k_i-1)^2}{n}=\frac{\sum_{i=x}^{x+m}(k_i-1)^2}{m} \tag{3.3}$$

If we multiply both sides of the equation by the fraction

$$\frac{1}{1-\frac{1}{n}}$$

and extract the root, we obtain

$$\sqrt{\frac{\sum_{i=1}^{n}(k_i-1)^2}{n-1}}=\sqrt{\frac{\sum_{i=x}^{x+m}(k_i-1)^2}{m-\frac{m}{n}}} \tag{3.4}$$

or

$$C_v^{(n)} \doteq C_v^{(m)} \tag{3.5}$$

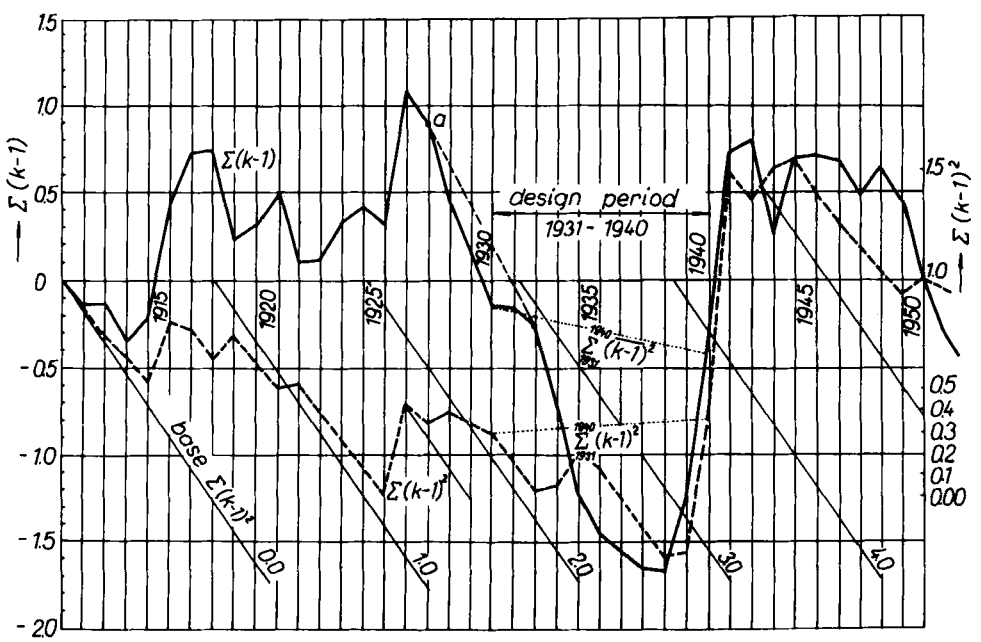

Fig. 3.5 Selection of a design period from a 40-year series (Slapy on the Vltava)

as the value of $\sqrt{m - (m/n)}$ does not differ greatly from the value of $\sqrt{m - 1}$ if m has a high value.

Instead of the mass curve of the $(k_i - 1)^2$ values, the mass curve for $(Q_{r,i} - Q_a)^2$ can be used, since these are values proportional to the proportion coefficient $1/Q_a^2$. This method is advantageous if the module coefficients $(k - 1)$ need not be calculated for other purposes.

A 40-year long hydrological series of the river Vltava in Slapy, with data for the mean annual flow recorded from 1911 to 1950 (Fig. 3.5), was chosen as an example.

From the entire series, a common period of several years was sought in which the connecting lines between the beginning and the end of the respective sections of the two mass curves would be approximately horizontal. The period 1931 to 1940 fulfilled both conditions.

The selected decade contains elements with various water yields in suitable groupings; years with a water yield very similar to the average years (1931 and 1938), a one very high-flow year (1940) the water yield of which was only surpassed in 1941, a year with an extraordinarily high-flow.

The selected decade suits all the requirements of the period to solve the problem of within-year control. However, the discharge distribution in the given years will have to be checked.

This favourable conclusion was reached by comparison of a 10-year series with a 40-year series (1911 to 1950). The question remains: to what extent are the above-mentioned forty years representative. Thus, comparisons for other cross-sections with longer hydrological series were carried out; the results are recorded in Table 3.1.

Table 3.1 Comparison of statistical characteristics for various periods

Profile	1931–1940		1931–1960		Long-term period	Q_a	C_v	Note
	Q_a [m³ s⁻¹]	C_v	Q_a [m³ s⁻¹]	C_v		[m³ s⁻¹]		
Labe–Děčín	305.5	0.35	305.0	0.36	1851–1960	300	0.29	year
					1851–1900	293	0.27	
Berounka–Křivoklát	32.7	0.47	31.8	0.46	1887–1960	30.1	0.36	water years
	water year XII–XI		water year XI–X		water year XII–XI			
Vltava–Štěchovice	86	0.41	discharge influenced by reservoirs		1891–1940	84	0.34	water years XII–XI
Mean annual precipitations (in Bohemia)	682 [mm]	0.152	669 [mm]	0.157	1876–1960	679 [mm]	0.133	calendar years

The table shows that the ten years 1931 to 1940 and the thirty years 1931 to 1960 have similar average flows as well as a similar variation coefficient; but both these values are much higher than the respective values for long-term periods reaching further back into the past.

The table also indicates that the river Berounka has a much higher C_v than the rivers Labe and Vltava, i.e., their flow variability is lower.

The last line of the table lists the average annual precipitations in Bohemia for these periods, as well as their variation coefficients. We can see that the precipitation variability is much smaller than the flow variability, but even the precipitation values indicate that their variability was much higher during the last thirty years than in a longer period going far back to the past. That the variability of the annual flow is larger than that of the annual precipitations is also evident from the relationship between the maximum and minimum values.

As indicator for the degree of representativeness of the selected period is also the distance between the exceedance curves: the empirical and the theoretical curves for the period.

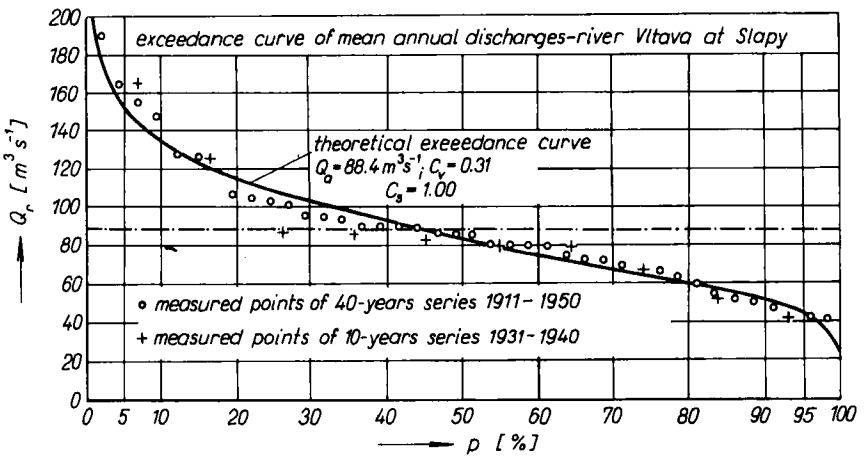

Fig. 3.6 Fitting of theoretical exceedance curve of mean monthly flows with empirical curves (Slapy on the river Vltava)

In Fig. 3.6, the theoretical exceedance curves are plotted according to Rybkin's table. The empirical points of the 40-year series are denoted by circles and the 10-year series (1931–1940) by crosses. The exceedance probability of the individual members of the empirical series was calculated from Chegodayev's formula

$$p\,[\%] = \frac{m - 0.3}{n + 0.4} \cdot 100 \tag{3.6}$$

where m is the ordinal number of the member in the degressive sequence, and n the number of members in the whole sequence (here 40 and 10).

The empirical points of the 40-year series can easily be plotted along the theoretical exceedance curve. The deviations of the 10-year series from the theoretical curve are the same as the respective deviations of the 40-year series. This means that according to this indicator it can be stated that the 10-year series (1931 to 1940) has a good representative character within the 40-year series (1911–1950).

As it is usually not possible to find a given period with an average flow Q'_a equalling the long-term average Q_a, this drawback has either to be eliminated by multiplying the flow Q_r by Q_a/Q'_a or by correcting the results of the solution.

The compilation of synthetic hydrological series by modelling is a further stage in hydrological data processing (see Section 3.2).

3.1.2 Solutions where there is a lack of direct hydrological observations

For reservoirs on streams, the flow of which has not been measured or has only short hydrological series, hydrological data have to be procured in an indirect manner (Dub, 1963; Ogievski, 1952).

If there are no direct observations, we can work in the following way:

(a) The long-term mean flow, as well as the statistical parameters (C_v, C_s) determining the flow distribution, can be ascertained from the *empirical dependence* of the mean annual flow on the climatic conditions. The results of such calculations are often unsatisfactory and cannot be recommended generally. An analogy with the nearest observed catchment area usually supplies better data. If the flow parameters are to be studied by the most reliable analogy, a careful analysis of all circumstances which might influence it is necessary, both in the catchment area where there are no direct measurements, as well as in the watershed of the similar stream which is furnishing the information. First of all, we have to compare the annual and seasonal precipitations, but also the slope conditions in the catchment area, the geological conditions and soil characteristics, the vegetation, air temperature, saturation deficit, etc. In the case of streams in mountainous areas, anyone using the analogy must be aware of the fact that discharges depend greatly on the height of the waterhed above sea-level.

(b) The flow distribution during the year is also determined according to a suitable, *analogous stream*; the lack of directly measured data is balanced by the selection of a most unfavourable flow distribution. In this way a fictive hydrograph is compiled – but, of course, based on hydrological analogy. According to this analogy we also determine the discharge of the required exceedance probability.

If the direct observations supply a *short hydrological series* (4 to 10 years), their representative nature has to be tested by comparison with a related profile and if it is not representative, it has to be prolonged according to hydrological analogy with that profile. The size of the catchment area also has to be taken into consideration, as the correlation between small and much larger watersheds is often not very close.

Correlation methods serve to express approximate relationships between two quantities x and y, as compared to a functional relationship in which a certain exact value y corresponds to a certain x, i.e., y is a "function" of x.

In a graphic presentation of such relationships the points found by measurement do not form a continuous line, they are scattered around the curve. Correlation

analyses are used for a quantitative expression of a close relationship between two quantities. In hydrology, a *linear correlation relationship* is most frequently used, with the regression equation

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \tag{3.7}$$

and with the correlation coefficient

$$r = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(x_i - \bar{x})^2}} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{n\sigma_x\sigma_y} \tag{3.8}$$

where σ_x, σ_y are the standard deviations of x and y from their mean values $\bar{x}$ and $\bar{y}$:

$$\sigma_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \qquad \sigma_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} \tag{3.9}$$

Equation (3.7) is the equation of a regression of y according to x; for the same values we can write the equation of regression of x according to y:

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \tag{3.10}$$

Both equations express straight lines with the slopes

$$b_y = r\frac{\sigma_y}{\sigma_x} \qquad b_x = r\frac{\sigma_x}{\sigma_y} \tag{3.11}$$

and the relationship between the two slopes must be

$$b_x b_y = r^2 \tag{3.12}$$

The probable error ε of the correlation coefficient r is calculated from

$$\varepsilon = \pm 0.674\frac{1 - r^2}{\sqrt{n}} \tag{3.13}$$

and the extreme error is usually assumed to be four times the probable error. Therefore, the complete expression for the correlation coefficient is

$$r \pm 4\varepsilon$$

In the case of a definite functional dependence, both regression lines are identical and $r = 1$. If $r < 1$ the two lines intersect in a point, with the coordinates $\bar{x}$ and $\bar{y}$ at an angle—the larger the angle, the smaller the r—i.e., the less close is the relationship.

If we search for the respective y values with respect to the selected x values according to the regression lines, we use the line (3.7); if we search for the respective x for a selected y, we use line (3.10). The line going through the point of intersection $(\bar{x}, \bar{y})$

and halving the angle of the regression lines, expresses the approximate relationship between x and y, but does not express how close that relationship is.

The correlation method can also be applied for the dependence of one variable y on two variables x and z, i.e., for three variables (Ogievski, 1952, p. 273).

There are also curvilinear correlations in hydrology and mathematical statistics can also solve these. Such calculations are complicated and laborious, thus *curvilinear relationships* are usually defined graphically.

A curved relationship can frequently be expressed satisfactorily by the exponential equation $y = ax^n$. The relationship between the logarithms of the coordinates of such curves, the so-called *logarithmic transformations*, are then expressed in a straight line. This is then no longer a correlation between x and y, but between their logarithms. The results are equations of straight lines of the regression of the logarithmic transformations, from which we pass on to the curves themselves.

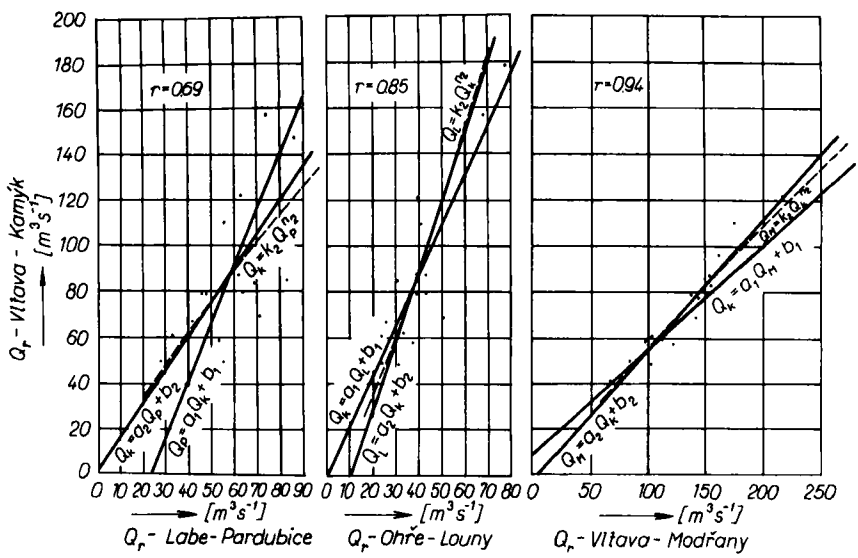

Fig. 3.7 Linear and non-linear correlation between the flows of the Vltava, Labe and Ohře rivers (1925–1946)

When applying correlation methods, we first have to predetermine graphically the general character of the correlation and then afterwards to perform the calculation.

The correlation methods give an accurate description of the hydrological relations and offer a more objective criterion for judging how close the relationship between the variable quantities is; these methods are rather laborious and that is why often only the graphical presentation of these relationships, which is fully satisfactory, is used.

If the variation coefficient is known in both profiles, the correlation between the annual flows can be determined by the Pearson curves or Rybkin tables for the theoretical exceedance curves.

Figure 3.7 shows the correlation between certain profiles on Czech rivers. A linear correlation was used

$$Q_y = aQ_x + b \tag{3.14}$$

and a curvilinear correlation with equation

$$Q_y = kQ_x^n \tag{3.15}$$

in which Q_y, is the value of the required flow, Q_x is the known flow of the site used as a basis and a, b, k, n are the constant parameters for a certain correlation.

Results of the calculation of some correlations for some rivers in the catchment area of the Labe can be found in the book by Votruba and Broža (1966).

3.1.3 Statistical and probability processing of hydrological series

The aim of the processing of measured hydrological quantities is to gain more useful information for the calculation of a reservoir than the original series could supply. This is why the probability theory, mathematical statistics and the theory of random processes, are used. The results are

- general statistical characteristics of the series,
- the laws of distribution
- synthetic (modelled) pseudo-chronological series.

In water management the above-mentioned methods are used to study random phenomena, i.e., phenomena which cannot be forecast for each individual case generally, although a certain complex of conditions is adhered to. For instance, we cannot forecast the spring flow from the snow cover of the same watershed, although we know the water content of the snow cover in the watershed and other genetic elements at a certain date; we cannot determine the flow process exactly, although we know the precipitation rate.

One or more variables or even entire processes might be random. We shall first of all analyse the properties of the random variable and afterwards those of the random process. In both cases there might be one or more random phenomena involved.

A frequently applied characteristic of the random variable is the probability of exceedance (reliability) expressed for the flow Q by $P(Q \geqq Q_m)$, in other words: the probability that the flow Q_m in the series $Q_1 \geqq Q_2 \geqq \ldots \geqq Q_m \geqq \ldots \geqq Q_n$ has been exceeded. This probability can be expressed by a formula derived by Chegodayev (equation (3.6)).

In the probability theory, it is usual to characterize a random variable by its *cumulative distribution function.* Expressed in (%), the relationship between the exceedance probabilities curve $P(x)$ and the cumulative distribution function $F(x)$ is

$$P(x) = 100 - F(x) \tag{3.16}$$

or for a general random variable ξ

$$P(x) = P(\xi \geqq x) \tag{3.17}$$

$$F(x) = P(\xi \leqq x) \tag{3.18}$$

A random variable can be

– *discrete*, if all its possible, but countable values can be expressed in integer numbers, such as 0, 1, 2, ...

– *continuous*, if its set of possible values is uncountable and continuously fills at least a part of the axis of real numbers. The random variable, therefore, here represents the value of the *random time function*, i.e., of a *random process* (in the case of a continuous variable) or a *random sequence* (in the case of a discrete variable).

The probability of each of the possible values of a continuous variable equals zero, because the number of possible values is infinite. We therefore divide the entire range of possible values of continuous random variable into a finite number of intervals and introduce the concept of the probability that the value of the continuous random variable will lie within the range of some interval, analogous to the work with the discrete random variable.

In hydrology and water management most phenomena take a continuous character: water levels, flows, reservoir fillings, precipitations, etc. For calculations or graphical presentations we transform them into "discrete" variables in the sense of the previous paragraph. Even when measuring in intervals (e.g., once in 24 hours) and not practising continuous measurings (e.g., on the recording gauge) we have to bear in mind that these are "discrete" expressions of the continuous variable within the range of the respective interval. In this respect, mean monthly flows are, e.g., discrete variables.

From the nature of the *distribution function*, according to equation (3.18), it follows that for an increasing x it has to rise continuously and when reaching $x = \xi_{max}$ it must be $F(x \geqq \xi_{max}) = 1$ (or 100%).

More generally: from the fact that the values of the distribution function express a *probability* and, therefore, are in the range from 0 to 1, it follows that for an impossible phenomenon the value $F(x)$ is:

$$F(-\infty) = P(\xi \leqq -\infty) = 0 \tag{3.19}$$

for a certain phenomenon

$$F(\infty) = P(\xi \leqq \infty) = 1 \tag{3.20}$$

The derivative of the distribution function is the *probability density* $f(x)$ expressing the probability that the random variable ξ will lie in a certain interval. Between the distribution function and the probability density the relation exists

$$P(\xi \leqq x) = F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x \tag{3.21}$$

or

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} \tag{3.22}$$

As the distribution function does not decrease, its derivative (probability density) cannot be negative.

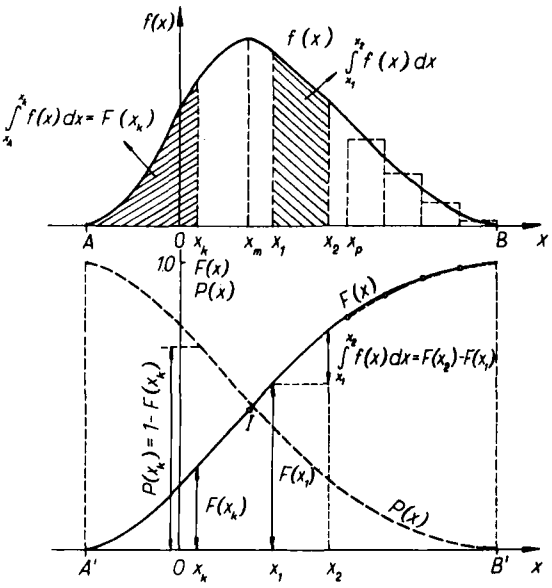

Fig. 3.8 Relationship between probability density $f(x)$, distribution function $F(x)$ and exceedance probability curve $P(x)$

The relationship between the probability density, the distribution function and the exceedance probabilities curve of the continuous random variable is graphically presented in Fig. 3.8. The diagram of the distribution shows the mass curve corresponding to the probability density curve; all the relationships valid between mass curve and their basic curves (see Section 2.1.1) are valid for them. From these relationships and equation (3.21), it follows that $F(x_k)$ is determined by the ordinate in the diagram of the distribution function vs. the abscissa x_k, or in the basic curve (e.g., the probability density diagram) by an area limited by this curve above the horizontal axis to the ordinate of the respective abscissa x_k.

The relation in Fig. 3.8 can be written as

$$P(x_1 < \xi \leqq x_2) = F(x_2) - F(x_1) \tag{3.23}$$

or using equation (2.7)

$$P(x_1 < \xi \leqq x_2) = \int_{x_1}^{x_2} f(x)\,dx \tag{3.24}$$

This means that the probability that the value of the continuous random variable will lie in the interval (x_1, x_2) equals the area below the probability density curve above the respective interval, or the difference between the ordinates of the distribution function belonging to the abscissas x_1 and x_2. The inflect point I on the $F(x)$ curve belongs to the coordinate x_m with a maximum $f(x)$.

Figure 3.8 shows—from the coordinate x_p to B—the division of the probability density of the continuous random variable into intervals which are supposed to have the mean value of the random variable within the range of the respective interval. The probability is marked as a step-shaped line. Compared to this is the distribution curve (broken line) with peaks on the distribution function curve to the continuous random variable.

A similar case is that in which a very large set of discrete random variables (e.g., mean daily flows) is divided in equal intervals into groups (e.g., for every 5 days—pentads, 10 days—decades) denoted classes. The interval centre represents all the values of the symbol of this class interval. The number of values of the symbols in each interval is the frequency, and the graphical presentation of such a group division is called a histogram (frequency distribution).

The *relative frequency* of the ith class is the ratio n_i/n (usually in %), where n_i is the number of values of the symbol (frequency) in the ith class interval and n number of all values of the symbol.

The *cumulative frequency* in relation to the ith class is defined by the sum $\sum_{j=1}^{i} n_j$.

The *cumulative relative frequency* is denoted by

$$\sum_{j=1}^{i} \frac{n_j}{n} = \frac{1}{n} \sum_{j=1}^{i} n_j$$

The diagram of the exceedance curve $P(x)$ is bound to the diagram of the distribution function $F(x)$ according to the relation in equation (3.16), as shown in Fig. 3.8 for the coordinate x_k.

From equation (3.24), it follows that for $x_1 = -\infty$ and $x_2 = \infty$ (or according to Fig. 3.8 for $x_1 = A$ and $x_2 = B$), with regard to equations (3.19) and (3.20)

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \tag{3.25}$$

The part of this area between A and x_k (Fig. 3.8) then corresponds to probability p that the density of the random variable is smaller than, or equal to, the selected number x_k. In such a case, the number x_k is called a *quantile* corresponding to a certain probability p.

A sample quantile is the characteristic which divides the set of sample values, ranked in ascending order of magnitude, into two appropriate parts. For example, the quintile divides the sample into $\frac{1}{5}$ and $\frac{4}{5}$ of all its values ranked in order of magnitude, the quartile into $\frac{1}{4}$ and $\frac{3}{4}$. The lower quartile x_1 separates $\frac{1}{4}$ of the smallest sample values, the upper quartile x_2 separates $\frac{1}{4}$ of the greatest values.

A special case of a quantile is the *median* $\tilde{x}$ (the sample centre), dividing the set of sample values ranked in order of magnitude into two parts containing the same number of values. Median is denoted $\tilde{x}$ (in English literature also *Me*), the notation being read "x tilda". If the number of sample values is odd and equals $2m + 1$, the median is

$$\tilde{x} = x_{m+1} \tag{3.26}$$

If the number is even and equals $2m$, the median is

$$\tilde{x} = \frac{1}{2}(x_m + x_{m+1}) \tag{3.27}$$

The median of the distribution of a continuous random variable is determined, according to (3.25), from the equation

$$\int_{-\infty}^{x} f(x)\,\mathrm{d}x = \int_{x}^{\infty} f(x)\,\mathrm{d}x = \frac{1}{2} \tag{3.28}$$

The value of the random variable $\xi = x_m$, maximizing the probability density $f(x)$ is called the *mode*. Hence, in order to determine the mode, the problem of finding an extreme is solved with the help of the first derivative of the function $f(x)$ with respect to $x(\mathrm{d}f(x)/\mathrm{d}x = 0)$. The mode is denoted by $\hat{x}$ (in English literature also *Mo*), the notation being read "x hat".

3.1.4 Statistical parameters and characteristics of a random variable

Complete information about a random variable is given by its frequency function (probability density function) $f(x)$ or cumulative distribution function $F(x)$. However, a certain amount of information can also be obtained from numerical characteristics, derived from the frequency function. The distribution characteristics (mean value, variance) are called *parameters* and denoted by Greek letters (μ, σ^2, etc.), while the sample characteristics are denoted by italic letters ($\bar{x}$, s^2, etc.). The sample characteristics obtained by calculation serve to estimate the corresponding parameters. The

calculation formulae for parameters of a discrete, uniformly distributed random variable are identical with those for the sample characteristics.

The moments (of various orders) of probability distribution are the most important distribution parameters. The moments may be of the following two types:

- *general moments* m,
- *central moments* M (moments around the mean).

The *general moment* of the kth order is denoted by $m_k(\xi)$ and for a continuous random variable is given by

$$m_k(\xi) = \int_{-\infty}^{\infty} x^k f(x)\,\mathrm{d}x \tag{3.29}$$

From the geometrical viewpoint it is the kth order moment with respect to the axis of ordinates of an area bounded by the curve $y = f(x)$ and the axis of abscissae, i.e., the so-called *initial moment*.

In the case of a discrete random variable, assuming values $x_1, x_2, \ldots, x_n$ with the probabilities $p_1, p_2, \ldots, p_n$, the general moment of the kth order is given by

$$m_k(\xi) = \sum_{i=1}^{n} x_i^k p_i \tag{3.30}$$

or with $p_1 = p_2 = \ldots = p_n = p = 1/n$

$$m_k(\xi) = \frac{1}{n} \sum_{i=1}^{n} x_i^k \tag{3.31}$$

The right-hand sides of the equations (3.29) and (3.30) are assumed to be absolutely convergent.

The *general moment of the first order* of a probability distribution represents the *mean value of the random variable* ξ, i.e., for a continuous random variable according to equation (3.29)

$$m_1(\xi) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \mu(x) \tag{3.32}$$

The *central moments* of a probability distribution $M_k(\xi)$ are the moments with respect to the axis containing the centre of gravity of the distribution.

For a continuous random variable the central moment of the kth order can be derived from equation (3.29), the following expression being obtained

$$M_k(\xi) = m_k(\xi - \mu) = \int_{-\infty}^{\infty} (x - \mu)^k f(x)\,\mathrm{d}x \tag{3.33}$$

for a discrete random variable the following holds, according to equation (3.30),

$$M_k(\xi) = \sum_{i=1}^{n} (x_i - \mu)^k p_i \tag{3.34}$$

and particularly, for $p_1 = p_2 = \ldots = p_n = p = 1/n$ the equation (3.31) yields the expression

$$M_k(\xi) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k \tag{3.35}$$

The *central moment of the first order* is always equal to zero. For example, using the equation (3.35) we obtain

$$M_1(\xi) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}n\bar{x} = \bar{x} - \bar{x} = 0 \tag{3.36}$$

The *central moment of the second order* is called *variance of the random variable* and its expression follows from (3.33), (3.34) and (3.35) with $k = 2$. The square root of the variance $\sqrt{M_2}$ is called *standard deviation of the random variable* and is denoted by σ, while the sample standard deviation is denoted by s. In this case

$$s = \sqrt{M_2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{3.37}$$

From the qualitative viewpoint the variance M_2 and consequently the standard deviation s are regarded as the measures of dispersion (or spread) of the values of the random variable about the mean value.

Frequently the dispersion of a random variable about the mean is characterized by the ratio of the standard deviation and the mean, the quantity

$$C_v = \frac{\sqrt{M_2}}{\bar{x}} \tag{3.38}$$

is called the *coefficient of variation*. Using equation (3.37) its value may be written as follows

$$C_v = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n\bar{x}^2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{x_i}{\bar{x}} - 1\right)^2}{n}} \tag{3.39}$$

With a very small number (n) of values the dispersion is described by the *range* R, i.e., the difference between the greatest and smallest value of the variable

$$R = x_n - x_1 \tag{3.40}$$

Furthermore, the question is whether or not the distribution is symmetric with respect to the vertical axis containing the centre of gravity. With a symmetric dis-

tribution, every central moment of an odd order is evidently equal to zero, which follows from equation (3.35) because in such a case

$$\sum_{i=1}^{n} (x_i - \bar{x})^k = 0$$

The *central moment of the third order*, e.g., for a discrete random variable given by the formula

$$M_3 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^3 \tag{3.41}$$

is therefore used to characterize the skewness of the distribution of a random variable.

As a rule, however, the skewness of a probability distribution is characterized by the zero dimension at ratio

$$C_s = \frac{M_3}{\sqrt{M_2^3}} \tag{3.42}$$

which is called the *coefficient of skewness* and is equal, for a continuous random variable, to

$$C_s = \frac{1}{\sigma^3(x)} \int_{-\infty}^{\infty} (x - \mu)^3 f(x)\, dx \tag{3.43}$$

for a discrete random variable to

$$C_s = \frac{1}{\sigma^3(x)} \sum_{i=1}^{n} (x_i - \mu)^3 p_i \tag{3.44}$$

or, in particular, for $p_1 = p_2 = \ldots = p_n = p = 1/n$ it holds that

$$C_s = \frac{\sum_{i=1}^{n} (x_i - \mu)^3}{n\, \sigma^3(x)} \tag{3.45}$$

The *central moment of the fourth order* is used to characterize the measure of flattening of a frequency curve near its centre. The form reduced to zero dimension

$$\gamma = \frac{M_4}{M_2^2} - 3 = \frac{M_4}{\sigma^4(x)} - 3 \tag{3.46}$$

is used, called the *coefficient of excess* (abbreviated to *excess*). As far as the normal Gauss–Laplace distribution $M_4/M_2^2 = 3$ is concerned its excess is equal to zero.

Positive values of excess γ indicate that the frequency curve is, in the neighbourhood of the mode, taller and slimmer than the normal curve with the same mean value and variance; and conversely for the negative values.

In hydrology and water management the following characteristics are most frequently used:

(a) the sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.47}$$

(b) the sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \tag{3.48}$$

(c) the sample coefficient of variation

$$C_v = \frac{s}{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n} \left(\frac{x_i}{\bar{x}} - 1\right)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} (k_i - 1)^2}{n-1}} \tag{3.49}$$

where $k_i = x_i/\bar{x}$ is the modulus coefficient;

(d) the sample skewness

$$C_s = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3} = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n} (k_i - 1)^3}{C_v^3} \tag{3.50}$$

As a rule, the expression

$$C_s = \frac{\sum_{i=1}^{n} (k_i - 1)^3}{(n-1) C_v^3} \quad \text{or} \quad C_s = \frac{\sum_{i=1}^{n} (k_i - 1)^3}{n C_v^3} \tag{3.51}$$

may be used, at least for a sufficiently large n $(n > 50)$ $n/(n-1) \doteq 1$.

The numerical sample characteristics in water management needed to construct the theoretical exceedance curve can also be obtained with the aid of quantiles (see further).

3.1.5 Theoretical probability distributions used in water management

The exceedance probabilities (distribution functions) are preferred to density functions and histograms for problems in hydrology and control of release from reservoirs. The empirical exceedance probabilities are not suitable to be employed, even for long series of observations because they are not sensitive enough for the extreme values where the probability of exceedance is near to either zero or one. For this, and other reasons, *theoretical exceedance probabilities* are used.

In the investigation of the theoretical probability distribution, the following problems arise:

(a) to determine the type of theoretical distribution by means of analysing the properties of the random variable, its boundary conditions, skewness, etc.;

(b) to estimate the parameters of the theoretical distribution on the basis of the characteristics of the empirical distribution (curve fitting);

(c) to evaluate the quality of the approximation to the chosen theoretical distribution by the empirical distribution.

We shall now introduce the probability distributions and evaluate their suitability for the tasks of water management.

The normal, log-normal, Pearson's, and exponential distributions, as well as the extreme-values distributions and the three-parameter distribution of Kritsky–Menkel, belong to the continuous theoretical distributions.

The binomial and Poisson distributions are discrete theoretical distributions.

More detailed explanation can be found in the literature on probability theory and mathematical statistics.

Normal probability distribution (Gauss–Laplace)

Normal probability distribution holds a prominent place in mathematical statistics as it is well suited for the distribution of many random variables, and for the approximation of some other continuous as well as discrete distributions. For its existence, the value of the investigated quantity should be the sum of many independent effects, each of them having only a small influence.

One must distinguish quite clearly the role of the large number of sample values and the large number of independent influences creating the value of a random variable in each particular case. The large number of independent effects (influences) yields the normal character of the theoretical distribution. The large number of sample values ensures goodness of fit of the empirical distribution with the theoretical distribution, whatever type of distribution might be concerned.

The probability density of a continuous normally distributed random variable x depends on the two parameters (constants): the mean value μ and the standard deviation σ. The density is given by

$$f(x) = c \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{3.52}$$

The value of constant c follows from the condition (3.25)

$$c \int_{-\infty}^{\infty} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx = 1 \tag{3.53}$$

By introducing a *standardized variable* t instead of variable x, we can compare the curves of distributions with different parameters, μ and σ:

$$t = \frac{x - \mu}{\sigma(x)} \qquad (\sigma\, dt = dx) \tag{3.54}$$

and we obtain

$$c\sigma \int_{-\infty}^{\infty} e^{-t^2/2}\, dt = c\sigma\sqrt{2\pi} = 1$$

hence

$$c = \frac{1}{\sigma\sqrt{2\pi}} \tag{3.55}$$

Thus the probability density of the normal distribution is given by the expression

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{3.56}$$

or using the standardized variable according to eqn. (3.54)

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \tag{3.57}$$

(the standardized variable has the zero mean $\bar{t} = 0$ and the unit standard deviation $\sigma = 1$).

The probability density or the normal distribution with the parameters $\mu = 0$ and $\sigma = 1$ is called the *standardized probability density.*

From eqns. (3.23) and (3.56), the expression for the distribution function of the normal distribution is

$$F(x) = \int_{-\infty}^{x} f(x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx \tag{3.58}$$

With the standardized variable t we obtain from equation (3.57) the distribution function of the standardized normal distribution in the form of

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-t^2/2}\, dt \tag{3.59}$$

which represents the part of the total unit area, bounded by the normal curve and the horizontal axis in the interval $(-\infty, t)$.

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2}\, dt = 1 \tag{3.60}$$

The graphs of the probability density and the distribution function of normal distribution for parameters $\sigma = 0.4$, 1.0 and 2.5 are shown in Fig. 3.9. In addition, it shows the cumulative normal distribution function on the probability net, the so-called Henry's line. In hydrology a skew distribution, i.e., limited in one direction or another, usually occurs. These conditions are not fulfilled for the normal distribution which, nevertheless, serves to derive suitable distributions.

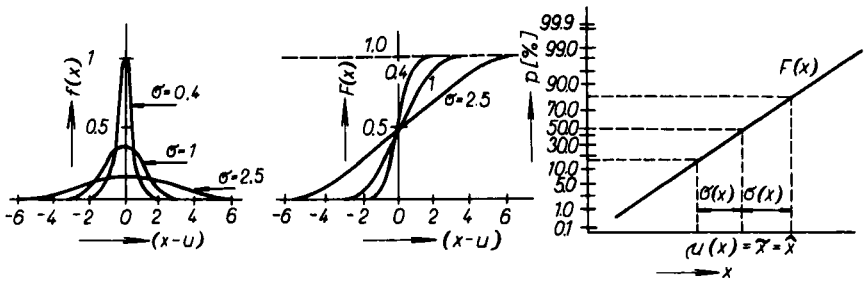

Fig. 3.9 Normal probability distribution for the values of the standard deviation $\sigma = 0.4$; 1; 2.5 (*a*) probability density; (*b*) distribution functions; (*c*) Henry's line

Log-normal probability distribution

The log-normal distribution is obtained by the logarithmic transformation of the normal distribution. Then the logarithm lg of the random variable is normally distributed, not the random variable itself. The distribution of given quantities, being asymmetric (which in hydrology is quite usual) is thus transformed to symmetric which makes the solution of the water-management problem easier. The log-normal distribution is specially suited for considerably asymmetrically distributed culmination flood flows where $C_s > 3C_v$.

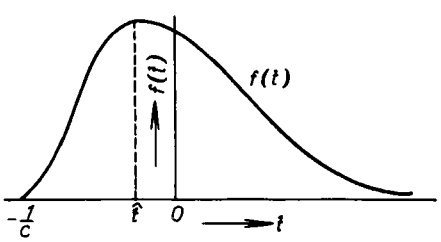

Fig. 3.10 Probability density of log-normal distribution

If the random variable

$$y = \lg(x - x_0) \tag{3.61}$$

is normally distributed with parameters $\mu(y)$ and $\sigma(y)$, then the random variable x has a log-normal distribution determined by three (constant) parameters: $\mu(y)$, $\sigma(y)$

and x_0. The range of the random variable x is (x_0, ∞) in the case of positive skewness and $(-\infty, x_0)$ in the case of negative skewness.

The graph of the probability density of the log-normal distribution with positive skewness is shown in Fig. 3.10.

For $x_0 = 0$ we obtained the log-normal distribution with the origin in zero determined by only two parameters $\mu(y)$ and $\sigma(y)$, its skewness is always positive and it holds that

$$C_s = C_v^3 + 3C_v \tag{3.62}$$

$$\mu(y) = \lg \mu(x) - \frac{1}{2} \lg (1 + C_v^2) \tag{3.63}$$

$$\sigma^2(y) = \lg (1 + C_v^3) \tag{3.64}$$

It follows from the last equation that for lower values of the coefficient of variance C_v (up to 0.50) it holds that $C_v \doteq \sigma(y)$; e.g., for $C_v = 0.500$ it is $\sigma(y) = 0.472$.

Gauss–Gibrat log-normal probability distribution

Processing data of 13.878 daily flows Q, measured in the years 1893 to 1930 on the Truyère river in France, Gibrat (1951) used the probability law, which he called the *law of proportional effect* ("Loi de l'effet proportionnel") and defined by the two equations

$$F(x) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{y} e^{-y^2} \, dy \tag{3.65}$$

$$y = a \log (x - x_0) + b \tag{3.66}$$

The symbol log denotes the decadic logarithm $\log_{10}$.

According to the law of proportional effect, the relative differential $dx/(x - x_0)$ is to be considered instead of the differential dx, and by integrating we obtain the normally distributed variable

$$\ln (x - x_0) + b = \ln 10 \log (x - x_0) + b \equiv a \log (x - x_0) + b$$

In Gibrat's case, the random variable x stood for the daily flow Q_d. J. Procházka succeeded in applying this distribution in Czechoslovakia for the annual flow maxima on the Ondava in Horovce in the years 1920–1962 (Votruba – Broža, 1966).

The Gauss–Gibrat distribution depends on the three constants a, x_0, b, which are determined by a suitable graphical or numerical approximation to the empirical distribution.

In investigating a random variable with the Gauss–Gibrat law, for convenience one may use the probability paper in which the scale of F is normal [the curve

$F(y) = 1/\sqrt{\pi} \int_{-\infty}^{y} e^{-y^2} dy$ is a straight line in the coordinate system $(F, y)]$ and the scale of $(x - x_0)$ is logarithmic. Thus the line approximated by the empirical distribution will have the gradient a. The value x_0 close to the minimum of the sample values is guessed at in flood-flow analyses, however, its influence is not great. The coefficient a characterizes the variability of the flow.

Galton log-normal probability distribution

F. Galton was the first who studied the log-normal distribution (1875, Galton's law). He considered the variable y as a function of the logarithm of the random variable x in the form

$$y = \alpha(\lg x - \beta) \tag{3.67}$$

where the values of parameters α and β can be empirically estimated

$$\alpha = \frac{1}{s\sqrt{2}}, \qquad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\lg x_i - \beta)^2} \tag{3.68}$$

$$\beta = \overline{\lg x} = \frac{1}{n} \sum_{i=1}^{n} \lg x_i \tag{3.69}$$

The logarithm lg may be taken to an arbitrary base.

Ven-Te-Chow log-normal probability distribution

Chow (1964) writes the random variable y as

$$y = \ln x \tag{3.70}$$

and the probability density is expressed by

$$f(x) = \frac{1}{\sigma(y)\, e^{y} \sqrt{2\pi}} \exp\left\{ - \frac{[y - \mu(y)]^2}{2\sigma^2(y)} \right\} \tag{3.71}$$

The statistical parameters of the distribution were obtained in the form

$$\mu(x) = \exp\left[\mu(y) + \tfrac{1}{2}\sigma^2(y)\right] \tag{3.72}$$

$$\sigma(x) = \mu(x)\,(e^{\sigma^2(y)} - 1)^{1/2} \tag{3.73}$$

$$C_v(x) = (e^{\sigma^2(y)} - 1)^{1/2} \tag{3.74}$$

$$C_s(x) = 3C_v(x) + C_v^3(x) \tag{3.75}$$

Chow also showed that Gumbel's extreme-values distribution is essentially a special case of log-normal distribution, providing

$$C_v = 0.364 \quad \text{and} \quad C_s = 1.139$$

Johnson log-normal probability distribution

N. L. Johnson (1949) presented three types of log-normal distributions, the second one having four parameters and being limited from both sides*).

The probability density of the distribution can be written as equation (3.71), substituting for y

$$y = \ln \frac{x - a}{b - x} \tag{3.76}$$

where b and a are the upper and lower bounds, respectively, of the random variable x; finally we obtain

$$f(x) = \frac{(b-a)(x-a)^{-1}(b-x)^{-1}}{\sigma(y)\sqrt{2}} \exp\left\{-\frac{1}{2}\left|\frac{\ln\frac{x-a}{b-x} - \mu(y)}{\sigma(y)}\right|^2\right\} \tag{3.77}$$

For the random sample from discrete random quantities (e.g., annual modulus flow coefficients $x_i = k_i = Q_{r,i}/Q_a$), it is

$$y_i = \ln \frac{x_i - a}{b - x_i} \tag{3.78}$$

and the standardized random variable

$$t_i = \frac{\ln\frac{x_i - a}{b - x_i} - m_y}{s_y} \tag{3.79}$$

where

$$m_y = \frac{\sum_{i=1}^{n} \ln\frac{x_i - a}{b - x_i}}{n} \quad \text{and} \quad s_y = \sqrt{\frac{\sum_{i=1}^{n}\left(\ln\frac{x_i - a}{b - x_i} - m_y\right)^2}{n-1}} \tag{3.80}$$

Then the equation (3.77) can be rewritten as

$$f(x) = \frac{(b-a)(x-a)^{-1}(b-x)^{-1}}{\sigma(y)\sqrt{2\pi}} e^{-t^2/2} \tag{3.81}$$

---

*) Yevievich (1972) wrote that because of the difficulties and inaccuracies in determining the parameters, it is not used in hydrology. Svanidze (1974) considered it suitable, and very flexible, for hydrological processes, mainly for the mean annual flows.

Extreme probability distribution

For the purpose of the statistical processing of hydrological data, the sampling distribution of extreme values following from the statistical theory of extreme values can be used. Mathematical statistics have three types of extreme-values distributions (cf. Smirnov *et al.*, 1965, p. 400) two of which are suitable for our purpose:

(a) the type suited for the minimum value of a set of random variables, sometimes called Weibull's distribution;

(b) the type suited for the maximum value of the set—Gumbel distribution.

The two distribution laws can be applied well in practice. They can be used not only for the calculation of the minimum and maximum flows in rivers, but similarly also for the minimum and maximum annual temperatures, rain and snow, precipitations, atmospheric pressure, strength of wind, etc.

Weibull probability distribution

The Weibull extreme-values distribution has lower bound x_0 with no upper bound, and therefore it is used for the statistical analysis of the minimum values. Depending on the three constants x_0, α, β, the probability density of the distribution takes the form

$$f(x) = \alpha\beta(x - x_0)^{\alpha-1} \exp\left[-\beta(x - x_0)^{\alpha}\right] \tag{3.82}$$

Through integration we obtain the relationship for the distribution function

$$F(x) = \int_{x_0}^{x} f(x)\,\mathrm{d}x = 1 - \exp\left[-\beta(x - x_0)^{\alpha}\right] \tag{3.83}$$

In order to simplify the expressions for mean, variance and skewness, we introduce the function gamma Γ (Euler's function of the second kind) which is

$$\Gamma(x) = \int_0^{\infty} \mathrm{e}^{-t} t^{x-1}\,\mathrm{d}t \tag{3.84}$$

being defined by the integral (eqn. (3.84)) only for $x > 0$, as for $x \leqq 0$ the integral is divergent.

Mean value of the Weibull distribution

$$\mu(x) = x_0 + \frac{1}{\sqrt[\alpha]{\beta}}\Gamma\left(\frac{\alpha + 1}{\alpha}\right) \tag{3.85}$$

variance

$$\sigma^2(x) = \frac{1}{\sqrt[\alpha]{\beta^2}}\left[\Gamma\left(\frac{\alpha + 2}{\alpha}\right) - \Gamma^2\left(\frac{\alpha + 1}{\alpha}\right)\right] \tag{3.86}$$

skewness

$$C_s = \frac{\Gamma\left(\frac{\alpha+3}{\alpha}\right) - 3\Gamma\left(\frac{\alpha+1}{\alpha}\right)\Gamma\left(\frac{\alpha+2}{\alpha}\right) + 2\Gamma^3\left(\frac{\alpha+1}{\alpha}\right)}{\left[\Gamma\left(\frac{\alpha+2}{\alpha}\right) - \Gamma^2\left(\frac{\alpha+1}{\alpha}\right)\right]^{3/2}} \tag{3.87}$$

The skewness C_s depends on the constant α and the relationship between the two parameters is shown in Fig. 3.11.

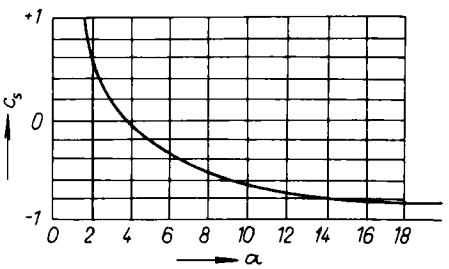

Fig. 3.11 The relation $C_s = f(\alpha)$ for Weibull's distribution

For the standardized variable $t = [x - \mu(x)]/\sigma(x)$ the Weibull distribution depends on one single constant α.

In the analysis of minimum flows we can write the basic formulae in the form of: for the frequency function

$$f(x) = \frac{\alpha}{Q' - x_0}\left(\frac{x - x_0}{Q' - x_0}\right)^{\alpha-1} \exp\left[-\left(\frac{x - x_0}{Q' - x_0}\right)^{\alpha}\right] \tag{3.88}$$

for the distribution function

$$F(x) = \int_{x_0}^{x} f(x)\,dx = 1 - \exp\left[-\left(\frac{x - x_0}{Q' - x_0}\right)^{\alpha}\right] \tag{3.89}$$

for the exceedance probabilities

$$P(x) = \exp\left[-\left(\frac{x - x_0}{Q' - x_0}\right)^{\alpha}\right] \tag{3.90}$$

Using the form as above, the distribution depends on three constants:

Q' – the so-called characteristic of "minimal" flow corresponding to the flow with the exceedance probability $P(Q') = 0.368 = e^{-1}$;

x_0 – the minimum possible value of x;

α – the slope of the exceedance probabilities curve in straight-line transformation, depending on the variance.

Gumbel probability distribution

E. J. Gumbel introduced his probability distribution law while investigating maximum age and called it the "law of maximum value". In the U.S.A., the Gumbel distribution is commonly used for statistical analysis of flood flows. Sometimes, it is, according to its mathematical description, called the *double exponential distribution.*

The double exponential distribution is unlimited from either side and its probability density is given by

$$f(x) = e^{-z} e^{-e^{-z}} \tag{3.91}$$

where z denotes the standardized deviation from the mode and depends on the random variable x according to the linear relation

$$z = \frac{1}{0.780\sigma}(x - \mu(x) + 0.450\sigma) \tag{3.92}$$

or with the aid of general parameters it may be written as

$$z = \alpha(x - \beta) \tag{3.93}$$

Using sample characteristics the parameters can be estimated by the terms

$$\alpha = \frac{\pi}{s(x)\sqrt{6}} \quad \text{and} \quad \beta = \bar{x} - \frac{\gamma}{\alpha} \tag{3.94}$$

where $\pi = 3.14$ (Ludolph's number)
$\gamma = 0.577$ (Euler's constant).

The preceeding formulae remain valid only in the asymptotic sense. For any finite sample they must be considered as approximations (Dub and Němec, 1969).

The distribution function of the Gumbel distribution is

$$F(x) = e^{-e^{-z}} \tag{3.95}$$

hence the exceedance probability is

$$P(x) = 1 - e^{-e^{-z}} \tag{3.96}$$

A disadvantage of the Gumbel distribution is that there is only one constant value of the coefficient of skewness, namely $C_s = 1.139$, however, the value is close to the coefficients of skewness for a large class of rivers. The exceedance probability for mode is $P = 63.2\%$ and that for the mean is $P = 42.97\%$.

In spite of having been recommended for analysing maximum peak discharges, the Gumbel distribution was found to be less suitable in some cases. Nevertheless, it is often used with probabilistic analysis of maximum precipitations (of n days, of n hours).

Pearson probability distribution

K. Pearson considered a very general differential equation

$$\frac{dy}{dx} = \frac{x + a}{b_0 + b_1 x + b_2 x^2} y \tag{3.97}$$

where a, b_0, b_1, b_2 are real (constant) numbers. According to the explicit values of these parameters and the domain of the variable x, the Pearson curves are of twelve different types.

The basic relation, defining the probability density of Pearson distribution, assumes the form

$$f(x) = \exp\left(\int_{-\infty}^{x} \frac{x + a}{b + b_1 x + b_2 x^2} dx\right) \tag{3.98}$$

The type of distribution is determined by the values of the following quantities

$$\begin{aligned} \beta_1 &= \frac{M_3^2}{M_2^3}; \qquad \beta_2 = \frac{M_4}{M_2^2}; \\ k &= \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)} \end{aligned} \tag{3.99}$$

where M_2, M_3 and M_4 are the second, the third and the fourth central moments (around the mean). For $\beta_1 = 0, \beta_2 = 3$ and $k = 0$ the Pearson distribution is identical with the normal distribution. The first and the third type curves are often used in the probabilistic analysis of hydrological data.

Pearson distribution of the first type

The first type is determined by the condition $k < 0$. The distribution is asymmetric, unbounded at either side, and usually bell-shaped.

The first-type Pearson distribution density curve, also called beta distribution, may be expressed by the following analytic formula

$$f(x) = \frac{[\alpha\gamma\,\sigma(x) - \mu(x) + x]^{\alpha-1}\,[\beta\gamma\,\sigma(x) + \mu(x) - x]^{\beta-1}}{B(\alpha, \beta)\,[(\alpha + \beta)\,\gamma\,\sigma(x)]^{\alpha+\beta-1}} \tag{3.100}$$

Symbol B is used to denote the beta function (first kind Euler's integral) defined by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1 - x)^{\beta-1}\,dx = \int_0^\infty \frac{x^\alpha\,dx}{(1 + x)^{\alpha+\beta}} \qquad (\alpha > 0;\ \beta > 0) \tag{3.101}$$

Beta function, values of which are tabulated, is related to the function $\Gamma(x)$ by the formula

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{3.102}$$

The mean value $\mu(x)$, the standard deviation $\sigma(x)$ and two independent constants α, β can be found in eqn. (3.100) while the auxiliary constant γ simplifying the notation is connected with the constants α and β by the relation

$$\gamma = \sqrt{\frac{\alpha + \beta + 1}{\alpha\beta}} \tag{3.103}$$

With the aid of α and β we can also express both the skewness C_s and the excess E:

$$C_s = 2\gamma \frac{\beta - \alpha}{\alpha + \beta + 2} \tag{3.104}$$

$$E = 3\gamma^2 \frac{2(\alpha + \beta)^2 + \alpha\beta(\alpha + \beta - 6)}{(\alpha + \beta + 2)(\alpha + \beta + 3)} - 3 \tag{3.105}$$

The domain of the variable x is bounded on both sides, the lower boundary being $[\mu(x) - \alpha\gamma\,\sigma(x)]$, and the upper boundary being $[\mu(x) + \alpha\gamma\,\sigma(x)]$. For $\alpha > 1$, $\beta > 1$ which is the most frequent case, the probability density curve $f(x)$ is bell-shaped. With $\alpha = \beta$ the curve is symmetric with respect to the ordinate of the corresponding mean value $\mu(x)$. For $\beta \to \infty$ the first type curve turns into the third-type Pearson curve.

The first-type Pearson curves may successfully be applied whenever the domain of variable is bounded on both sides. The advantage consists in the fact that we can choose the beginning of the curve, often choosing zero, and yet there are three independent parameters $\mu(x)$, $\sigma(x)$ and S_s left.

Pearson distribution of the third type

The third type is recognizable from $k = \infty$ or $2\beta_2 = 3\beta_1 + 6$. The distribution is asymmetric, bounded on one side, and usually bell-shaped.

The general analytical formula for the third-type Pearson curve is

$$f(x) = \frac{\alpha^{\alpha^2} \exp\left[\frac{\alpha\,\mu(x)}{\sigma(x)} - \alpha^2\right]}{\Gamma(\alpha^2)\,\sigma^{\alpha^2}(x)} [\alpha\,\sigma(x) - \mu(x) + x]^{\alpha^2 - 1} \exp\left[-\frac{\alpha}{\sigma(x)} x\right] \tag{3.106}$$

where Γ denotes the gamma function.

The curve depends on three parameters: the mean value $\mu(x)$ the variance $\sigma(x)$ a constant α defining the skewness by the relation

$$C_s = \frac{2}{\alpha} \tag{3.107}$$

Under positive skewness the domain of the variable x is from $[\mu(x) - \alpha\,\sigma(x)]$ to $+\infty$, under negative skewness from $-\infty$ to $[\mu(x) - \alpha\,\sigma(x)]$. For α tending to

infinity it holds that $C_s \to 0$, and the curve becomes symmetric and turns to normal distribution with parameters $\mu(x)$ and $\sigma(x)$, hence normal distribution is a special case of the Pearson distribution of the third type. The curve is bell-shaped whenever $\alpha^2 > 1$, i.e., $C_s < 2$.

If $C_s > 0$ the curve is decreasing, if $C_s = 2C_v$ the curve is increasing.

If it holds that $C_s > 0$ under $C_s = 2C_v$, the curve has its origin in zero. Hence the variable assumes positive values providing $C_s \geqq 2C_v$.

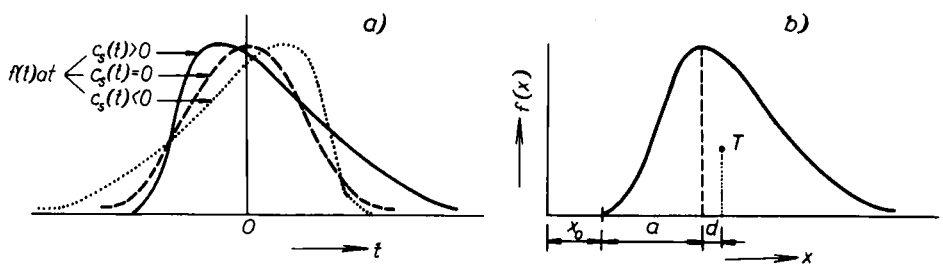

Fig. 3.12 Pearson curve of the third type
(*a*) for standardized variable t; (*b*) for $x_0 > 0$

Excess E depends on C_s according to the relation

$$E = \frac{3}{2} C_s^2 \tag{3.108}$$

The third-type curve for standardized variable (Fig. 3.12a) is

$$f(t) = \frac{\alpha^{\alpha^2} e^{-\alpha^2}}{\Gamma(\alpha^2)} (\alpha + t)^{\alpha^2 - 1} e^{-\alpha t} \tag{3.109}$$

The domain of the standardized variable is from $-\alpha$ to $+\infty$ under $\alpha > 0$, from $-\infty$ to $-\alpha$ under $\alpha < 0$.

Mode of the curve is

$$\hat{t} = -\frac{1}{\alpha}$$

The inflexion points exist only under $\alpha^2 > 1$, their distance from the mode being $\pm\sqrt{1 - (1/\alpha^2)}$ in both directions.

The third-type curve is very general because it may be used to describe both asymmetric distributions with either positive or negative skewness and symmetric distributions. When $|C_s|$ is close to zero, the third-type curve is close to the normal distribution. Both the frequency and distribution functions are tabulated in detail, hence they are easy to deal with.

The probability density of the Pearson distribution of the third type may, for convenience, be expressed by

$$f(x) = \frac{1}{d\left(\frac{a}{d}+1\right)\Gamma\left(\frac{a}{d}+1\right)}(x - x_0)^{a/d}\exp\left(-\frac{x - x_0}{d}\right) \tag{3.110}$$

where a is the distance between the mode and the origin of the curve, d is the distance between the ordinate of the centre of gravity and the mode, and x_0 is the minimum value of x (Fig. 3.12b).

Moments of the distribution may be expressed by means of the parameters from equation (3.110)

$$\mu(x) = x_0 + a + d \tag{3.111}$$

$$C_v = \frac{1}{\mu(x)}[d(a + d)]^{1/2} \tag{3.112}$$

$$C_s = 2\left(\frac{d}{a + d}\right)^{1/2} \tag{3.113}$$

From these relations it follows that

$$x_0 = \mu(x)\left(1 - \frac{2C_v}{C_s}\right)$$

In almost all hydrological events it must hold that $x_0 \geqq 0$ and therefore $C_s \geqq 2C_v$. Hence, the whole curve is situated in the domain of positive x's; with $C_s = 2C_v$ the random variable could assume negative values as well, which is an impossible rule in hydrological events.

Kritsky–Menkel's three-parameter probability distribution

In order to describe annual river discharge in the U.S.S.R., a theoretical distribution is often used which was recommended by Kritsky and Menkel under the following assumptions:

(a) the distribution function differs from zero on the interval $(0, \infty)$;

(b) the distribution function depends on three parameters, i.e., an arbitrary choice of the first three moments must be possible.

The analytic formula for the Kritsky–Menkel three-parameter distribution function is (for $x \geqq 0$) in the form

$$F(x) = \left[\frac{\Gamma(\gamma + b)}{\Gamma(\gamma)}\right]^{\gamma/b}\frac{1}{\Gamma(\gamma)\bar{x}b}\int_0^x \exp\left\{-\left[\frac{\Gamma(\gamma + b)t}{\Gamma(\gamma)\bar{x}}\right]\right\}\left(\frac{t}{x}\right)^{\gamma/b - 1} dt \tag{3.114}$$

where $\bar{x}$, γ, b are the parameters.

For $b = 1$ the distribution function (3.114) is reduced to the third-type Pearson distribution with equality $C_s = 2C_v$, i.e., a two-parameter distribution. Reznikovski (1969) concluded, on the basis of investigations performed in the U.S.S.R., that for rivers with $C_v \leqq 0.5$ and $r \gtrless 0.3$ the relation $C_s = C_v^3 + 3C_v$ according to equation (3.62) is closer to reality than the relation $C_s = 2C_v$, i.e., the log-normal distribution is more suitable than the third-type Pearson distribution in such cases.

Exponential probability distribution

The exceedance probability curve of exponential probability distribution is given by the exponential function

$$P(z) = e^{-z} \tag{3.115}$$

hence the distribution function is

$$F(x) = 1 - e^{-z} \tag{3.116}$$

where $z = f(x)$.

Various expressions of the variable z yield various types of exponential probability laws. The Goodrich exponential distribution was found suitable for analysing some flood-flow samples.

Goodrich distribution is usually given by

$$P(x) = e^{-k(x-x')^{1/\alpha}} \tag{3.117}$$

or

$$P(x) = 10^{-k(x-x')^{1/\alpha}} \tag{3.118}$$

i.e.

$$z = k(x - x')^{1/\alpha} \tag{3.119}$$

The best way to estimate the constants k, x' and α, is to express the exceedance probability curve in the Goodrich probability paper. The constant x' is chosen so that the empirical distribution in the Goodrich probability paper creates approximatelly a straight line (for $P \to 1$ the deviation may be a bit larger).

By double-logarithming the equation (3.117), rewritten in the form

$$e^{k(x-x')^{1/\alpha}} = \frac{1}{P}$$

we obtain

$$\log(x - x') = \alpha\left[\log\log\frac{1}{P} - (\log k + \log\log e)\right] \tag{3.120}$$

The constant α determining the slope of the approximating line is equal to

$$\alpha = \frac{\Delta \log (x - x')}{\Delta \log \left(\log \frac{1}{P} \right)} \tag{3.121}$$

For the Goodrich probability chart, the scale of $1/P$ is double logarithmic ($\log \log 1/P$), and the scale of $(x - x')$ is logarithmic.

The ordinate of the exceedance probability curve corresponding to $P = 0.10$, i.e., with $\log (\log 1/P) = 0$, determines the value $\log (x - x') = -\alpha(\log k + \log \log e)$ whence k can be calculated. For an arbitrary P, the corresponding x_p is calculated from equation (3.120).

In Czechoslovakia, V. Broža succeeded in applying the Goodrich probability distribution to analyses of samples of all maximum peak discharges (not the annual maximum discharges for which some other distributions are more suitable).

Binomial probability distribution

An asymmetric binomial distribution is frequently used in hydrological data processing. Explaining the nature of the binomial distribution, we must keep in mind that it is a distribution of a discrete random variable.

In a sequence of independent trials performed under stable conditions (Bernoulli trials), we observe whether or not an even A occurs, the probability of which being p in every trial. Thus the probability

$$p(n, x) = \binom{n}{x} p^x q^{n-x} = \binom{n}{x} p^x (1 - p)^{n-x} \tag{3.122}$$

here has the meaning of probability density, hence

$$f(x) = \binom{n}{x} p^x q^{n-x} \tag{3.123}$$

where x denotes the number of the occurrences of event A.

The constants n, p are called parameters of the binomial distribution.

The set of probabilities $p(n, x)$ for $x = 0, 1, 2, \ldots, n$, i.e., $p(n, 0), p(n, 1), p(n, 2), \ldots$ $\ldots, p(n, n)$, is called binomial probability distribution.

As these probabilities correspond to disjointed events, forming the complete system of events, it follows that

$$\sum_{x=0}^{n} p(n, x) = 1 \tag{3.124}$$

which may easily be checked because the values $p(n, x)$ defined by eqn. (3.122) are the terms of the binomial expression $(q + p)^n$ from which their name was derived.

The mode of the binomial distribution is the value $\hat{x}$ corresponding to the greatest probability. As a rule, the binomial distribution has only one mode. The maximum value may, of course, correspond to either 0 or n, the probabilities $p(n, x)$ either decrease or increase, respectively. Such cases may occur with small n and p close to either zero or one. With a large n, the mode is always somewhere in the central part of the distribution and the probabilities decrease in both directions from the mode.

In order to facilitate the calculation of the moments of the binomial distribution, the so-called *moment-generating function* $\varphi(\varepsilon)$ is introduced. This function is used to calculate the moments especially of a discrete random variable, and is defined as follows:

$$\varphi(\varepsilon) = \sum_x e^{\varepsilon x} p(x) \tag{3.125}$$

Summation concerns all possible values $x = 0, 1, 2, \ldots, n$.

Let us suppose the sum to converge at least for sufficiently small ε. Then it holds that

$$\frac{d^k \varphi(\varepsilon)}{d\varepsilon^k} = \sum_x x^k e^{\varepsilon x} p(x) \tag{3.126}$$

The formula for calculating the kth order general moment is

$$m_k(\xi) = \sum_{i=1}^{n} x_i^k p_i \qquad \text{[see equation (3.30)]}$$

From eqn. (3.126) with $\varepsilon = 0$ we obtain, substituting from eqn. (3.30), an important formula for calculating the general moments

$$\left.\frac{d^k \varphi(\varepsilon)}{d\varepsilon^k}\right|_{\varepsilon=0} = \sum_x x^k p(x) = m_k \tag{3.127}$$

Now we shall calculate the mean value and variance of the binomial distribution with the aid of the moment-generating function. Substituting the formula (3.122) for the probability $p(x)$ into (3.125) we obtain

$$\varphi(\varepsilon) = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} e^{\varepsilon x} = (pe^{\varepsilon} + q)^n \tag{3.128}$$

The first two moments follow from relations (3.127) and (3.128)

$$\mu(x) = m_1(x) = \varphi'(\varepsilon)|_{\varepsilon=0} = n(pe^{\varepsilon} + q)^{n-1} pe^{\varepsilon}|_{\varepsilon=0} = np \tag{3.129}$$

$$m_2(x) = \varphi''(\varepsilon)|_{\varepsilon=0} = n(n-1)(pe^{\varepsilon} + q)^{n-2} p^2 e^{2\varepsilon} + np(pe^{\varepsilon} + q)^{n-1} e^{\varepsilon}|_{\varepsilon=0} =$$
$$= n^2p^2 - np^2 + np = n^2p^2 + npq \tag{3.130}$$

The variance $\sigma^2(x)$ is the second-order central moment M_2 which is given by $M_2 = m_2 - \mu^2(x)$. Hence it holds that

$$\sigma^2(x) = n^2p^2 + npq - n^2p^2 = npq \tag{3.131}$$

The distribution function is

$$F(x_b) = \sum_{x_i \leqq x_b} p(n, x_i) \tag{3.132}$$

For $x_b = n$ it follows that:

$$F(n) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + 1 - p)^n = 1$$

[see equation (3.124)].

Equations (3.122) and (3.132) show that the distribution function $F(x_b)$ depends on the three parameters x_b, n, p. Therefore, the construction of a table for $F(x_b)$ is very complicated and the calculation for great n and x_b is difficult.

We shall, however, make use of the fact that for $n \to \infty$ the binomial distribution tends to the normal distribution. Therefore we shall determine $F(x_b)$ for large n approximately from the normal distribution.

Poisson probability distribution

To explain the Poisson distribution law we again use the Bernoulli trial as it was introduced in connection with the binomial distribution. The Poisson distribution, also called the distribution of rare events, is well suited for a large number of independent trials, the probability of occurrence of some event in each of them being relatively small. The distribution is tabulated in detail, because it depends on only one parameter. Its usefulness arises in cases where neither n nor p are known but their product, $n \cdot p$, is known, or may be estimated.

Suppose the number of trials is growing and simultaneously the probability of occurrence of an event A is decreasing so that the mean number of occurrences is constant, i.e., $n \cdot p = \lambda$, where λ is positive; then

$$p = \frac{\lambda}{n} \tag{3.133}$$

Let us first calculate the probability that, in n trials, the event will not occur at all. Using (3.133) we can then rewrite eqn. (3.122) as

$$p(n, 0) = (1 - p)^n = \left(1 - \frac{\lambda}{n}\right)^n$$

whence

$$\ln p(n, 0) = n \ln\left(1 - \frac{\lambda}{n}\right) = -\lambda - \frac{\lambda^2}{2n} - \ldots \tag{3.134}$$

If

$$\frac{\lambda^2}{n} \ll 1 \quad \text{or} \quad p \ll \frac{1}{\sqrt{n}}$$

holds, we may reduce eqn. (3.134) to the first term and write

$$p(n, 0) \doteq \mathrm{e}^{-\lambda} \tag{3.135}$$

Similarly it follows that

$$p(n, 1) = np(1-p)^{n-1} = \frac{np}{1-p}(1-p)^n = \frac{\lambda}{1-\frac{\lambda}{n}} p(n, 0) \doteq \lambda \mathrm{e}^{-\lambda} \tag{3.136}$$

and generally

$$p(n, x) = \frac{n(n-1)\ldots(n-x+1)}{x!} p^x (1-p)^{n-x} = \frac{\lambda^x}{x!} \mathrm{e}^{-\lambda} \tag{3.137}$$

The result may be formulated as follows: Providing that number n is very large, the occurrence probability p of an event is very small so that the mean number of occurrences remains equal to a constant positive number λ, the probability $p(n, x)$ for binomial distribution tends for every $x = 0, 1, \ldots$, to the limit value

$$p(\lambda, x) = \frac{\lambda^x}{x!} \mathrm{e}^{-\lambda} = f(x) \tag{3.138}$$

This asymptotic expression is sometimes called Poisson's formula. It holds that

$$\sum_{x=0}^{\infty} p(\lambda, x) = \sum_{x=0}^{\infty} \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} = 1 \tag{3.139}$$

The parameters are calculated again with the aid of the moment-generating function. Combining eqn. (3.138) with eqn. (3.125) we obtain

$$\varphi(\varepsilon) = \sum_{x=0}^{\infty} \frac{\mathrm{e}^{\varepsilon x} \lambda^x \mathrm{e}^{-\lambda}}{x!} = \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{\mathrm{e}^{\varepsilon x} \lambda^x}{x!} = \mathrm{e}^{-\lambda} \left(1 + \lambda \mathrm{e}^{\varepsilon} + \frac{\lambda^2}{2!} \mathrm{e}^{2\varepsilon} + \ldots\right) =$$

$$= \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda \mathrm{e}^{\varepsilon}}$$

$$\left(\text{for } \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = 1 + \lambda + \frac{\lambda^2}{2!} + \ldots = \mathrm{e}^{\lambda}\right)$$

whence according to eqns. (3.127) and (3.128), the general moments are given by

$$\begin{aligned} m_1(x) &= \mu(x) = \varphi'(\varepsilon)\big|_{\varepsilon=0} = \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda \mathrm{e}^{\varepsilon}} \lambda \mathrm{e}^{\varepsilon}\big|_{\varepsilon=0} = \lambda \\ m_2(x) &= \varphi''(\varepsilon)\big|_{\varepsilon=0} = \lambda \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda \mathrm{e}^{\varepsilon}} \mathrm{e}^{\varepsilon} (\lambda \mathrm{e}^{\varepsilon} + 1)\big|_{\varepsilon=0} = \lambda + \lambda^2 \end{aligned} \tag{3.140}$$

hence

$$\sigma^2(x) = M_2 = m_2 - \mu^2 = \lambda + \lambda^2 - \lambda^2 = \lambda \tag{3.141}$$

From equation (3.140) and (3.141) we get the obvious identity between the mean value $\mu(x)$ and the variance $\sigma^2(x)$, which is characteristic for the Poisson probability distribution.

The coefficient of skewness

$$C_s = \frac{M_3}{\sqrt{M_2^3}} = \frac{\lambda}{\sqrt{\lambda^3}} = \frac{1}{\sqrt{\lambda}} \tag{3.142}$$

is always positive because λ is always positive.

The coefficient of excess

$$\gamma = \frac{M_4}{\sigma^4(x)} - 3 = \frac{3\lambda^2 + \lambda}{\lambda^2} - 3 = \frac{1}{\lambda} \tag{3.143}$$

of the Poisson distribution is always positive, too.

For calculating the probabilities for the Poisson distribution, as well as its distribution function (i.e., the probability of a maximum of n occurrences)

$$F(\lambda, n) = \sum_{x=0}^{n} \frac{\lambda^x}{x!} e^{-\lambda} \tag{3.144}$$

tables (Reisenauer, 1970) have been elaborated.

3.1.6 Estimates of exceedance probability curve parameters by means of quantiles

Up to now, the method of moments has been used to determine the values of parameters $\mu(x)$, $\sigma^2(x)$, C_v, C_s, γ in the appropriate types of distributions.

Lately, using many random samples, the fit has been checked between the sample characteristics C_v and C_s and the respective values of parameters given by eqns. (3.49) and (3.50).

The optimal estimates based on the random samples are greater than the corresponding parameters. This is why it is sometimes recommended that the estimates of C_v and C_s should be calculated with the aid of quantiles. A quantile denotes the point chosen so that the distribution function assumes a prescribed value, e.g., 5%, 25%, etc.

The method is based on several values of quantiles obtained from the empirical curve (approximately constructed), and on the standardized deviations of exceedance probability curves $\Phi(P_i; C_s)$ for the corresponding distribution, found in tables. The index of Φ denotes the type of distribution (Φ_B – binomial, Φ_{LN} – log-normal, Φ_G – Gumbel). The quantiles for fixed probabilities P_i are used taking into account the

relationship between the cumulative distribution function and the exceedance probability curve – see Fig. 3.8.

An approach essentially simplifying the calculation as compared with the complicated calculation based on equations (3.49) and (3.50), was elaborated by Alexeyev (1960). Only three points of the empirical curve, corresponding to $P_1 = 5\%$, $P_2 = 50\%$ and $P_3 = 95\%$, are considered.

Binomial distribution

The calculation is described here in detail:

1. The set of sample values (e.g., the average annual flows) are ranked in descending order of magnitude and the probabilities P are calculated according to Chegodayev.

2. The empirical curve in the probability paper is approximated by a straight line or a continuous curve (curve fitting). From this curve we obtain three values for fixed characteristic probabilities, namely $P = 5\%$, 50% and 95%, the values being denoted x_5, x_{50} and x_{95}.

3. We calculate the auxiliary quantity S defined by

$$S = \frac{x_5 + x_{95} - 2x_{50}}{x_5 - x_{95}} \tag{3.145}$$

which is, according to Alexeyev, the function of sample skewness (Fig. 3.13).

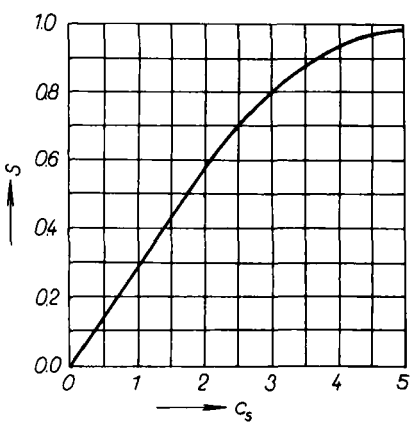

Fig. 3.13 The relation $S = f(C_s)$ for binomial distribution

With binomial distribution it also holds that

$$S = \frac{\Phi_5 + \Phi_{95} - 2\Phi_{50}}{\Phi_5 - \Phi_{95}} \tag{3.146}$$

where Φ_5, Φ_{50} and Φ_{95} are values of $\Phi_B(P, C_s)$ obtained from Foster–Rybkin tables for $P = 5\%$, 50% and 95%.

Table 3.2 Auxiliary values to determine parameters of binomial distribution according to G. A. Alexeyev

C_s	$\Phi(p, C_s)$		S	C_s	$\Phi(p, C_s)$		S
	$\Phi_5 - \Phi_{95}$	Φ_{50}			$\Phi_5 - \Phi_{95}$	Φ_{50}	
0.0	3.28	0.00	0.00	2.7	2.74	−0.38	0.74
0.1	3.28	−0.02	0.03	2.8	2.71	−0.39	0.76
0.2	3.28	−0.03	0.06	2.9	2.68	−0.39	0.78
0.3	3.27	−0.05	0.08	3.0	2.64	−0.40	0.80
0.4	3.27	−0.07	0.11	3.1	2.62	−0.40	0.81
0.5	3.26	−0.08	0.14	3.2	2.59	−0.41	0.83
0.6	3.25	−0.10	0.17	3.3	2.56	−0.41	0.85
0.7	3.24	−0.12	0.20	3.4	2.53	−0.41	0.86
0.8	3.22	−0.13	0.22	3.5	2.50	−0.41	0.87
0.9	3.21	−0.15	0.25	3.6	2.48	−0.42	0.89
1.0	3.20	−0.16	0.28	3.7	2.45	−0.42	0.90
1.1	3.17	−0.18	0.31	3.8	2.43	−0.42	0.91
1.2	3.16	−0.19	0.34	3.9	2.41	−0.41	0.92
1.3	3.14	−0.21	0.37	4.0	2.40	−0.41	0.92
1.4	3.12	−0.22	0.39	4.1	2.38	−0.41	0.93
1.5	3.09	−0.24	0.42	4.2	2.36	−0.41	0.94
1.6	3.07	−0.25	0.45	4.3	2.34	−0.40	0.94
1.7	3.04	−0.27	0.48	4.4	2.32	−0.40	0.95
1.8	3.01	−0.28	0.51	4.5	2.30	−0.40	0.96
1.9	2.98	−0.29	0.54	4.6	2.28	−0.40	0.97
2.0	2.95	−0.31	0.57	4.7	2.26	−0.40	0.97
2.1	2.92	−0.32	0.59	4.8	2.23	−0.39	0.98
2.2	2.89	−0.33	0.63	4.9	2.21	−0.39	0.98
2.3	2.86	−0.34	0.64	5.0	2.18	−0.38	0.98
2.4	2.82	−0.35	0.67	5.1	2.15	−0.38	0.98
2.5	2.79	−0.36	0.69	5.2	2.15	−0.37	0.98
2.6	2.76	−0.37	0.72				

S is obtained from Table 3.2 [according to equation (3.145)], along with the value C_s and other auxiliary values $(\Phi_5 - \Phi_{95})$, and Φ_{50}, which enable us to calculate other characteristics of the binomial distribution.

4. We calculate the following characteristics:

The standard deviation

$$s_x = \frac{x_5 - x_{95}}{\Phi_5 - \Phi_{95}} \tag{3.147}$$

the mean value

$$\bar{x} = x_{50} - s_x \Phi_{50} \tag{3.148}$$

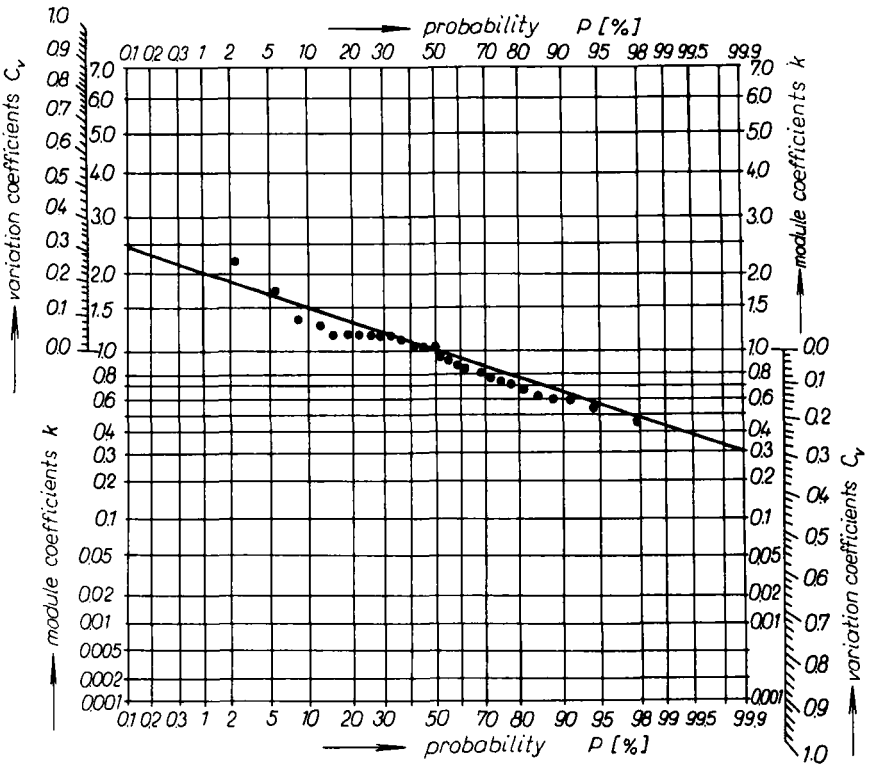

Fig. 3.14 Brovkovich's probability paper (Pearson $C_s = 2C_v$)

the coefficient of variation

$$C_v = \frac{s_x}{\bar{x}} \tag{3.149}$$

We compare C_s and C_v; for theoretical correctness of the following stepwise relation, $C_s \geqq 2C_v$ must be valid.

5. From the general equation

$$x_p = \bar{x} + s_x \Phi(P, C_s) = \bar{x}[1 + C_v \Phi(P, C_s)] \tag{3.150}$$

we calculate values of the random variable x for any exceedance probability P. The function $\Phi_B(P, C_s)$ is tabulated in detail.

The problem may be solved also with the aid of the graphico-numerical method using the Brovkovich probability paper (Fig. 3.14).

Log-normal distribution

The approach is again based on quantiles corresponding to the probabilities $P_1 = 5\%$, $P_2 = 50\%$ and $P_3 = 95\%$. Steps 1 and 2 are the same as for the binomial distribution.

3. We calculate an auxiliary quantity x_0 given by

$$x_0 = \frac{x_5 x_{95} - x_{50}^2}{x_5 + x_{95} - 2x_{50}} \tag{3.151}$$

and the standard deviation of a new variable y according to equation (3.61)

$$s_y = 0.305 \log \frac{x_5 - x_0}{x_{95} - x_0} \tag{3.152}$$

4. We calculate values x_p corresponding to various probability P's from the equation

$$\log (x_p - x_0) = \log (x_{50} - x_0) + s_y \Phi_B(P, C_s = 0) \tag{3.153}$$

Values of the function $\Phi_B(P, 0)$ can be found in the tables mentioned above.

The fit of the log-normal distribution to the investigated sample is checked by representing the exceedance probabilities curve in the logarithmic probability paper. If the sample points with coordinates $(P_m, x_m - x_0)$ are near a line, the log-normal distribution is well-suited.

Gumbel distribution

The Gumbel distribution depends on two parameters α, β according to equation (3.93), and therefore only two quantiles are used to estimate the parameters, x_5 and x_{95}, corresponding to probabilities $P_1 = 5\%$ and $P_2 = 95\%$, respectively.

Following steps 1 and 2, we determine the values x_5 and x_{95}. The parameters α and β are calculated from the relations

$$\alpha = \frac{4.067}{x_5 + x_{95}} \tag{3.154}$$

$$\beta = 0.27x_5 + 0.73x_{95} \tag{3.155}$$

The ordinate x_p corresponding to various exceedance probabilities P may be calculated either from the relation

$$x_p = \beta + \frac{1}{\alpha} z_P \tag{3.156}$$

where z_P is the standardized deviation from the mode according to the table, or from a general equation of type (3.150), substituting the values $\Phi_G(P, C_s)$ also according to the table (e.g., Votruba and Nacházel, 1971).

The other characteristics of equation (3.150) are given by

$$\bar{x} = 0.412x_5 + 0.588x_{95} \tag{3.157}$$

$$s_x = 0.315(x_5 - x_{95}) \tag{3.158}$$

Exponential distribution

Following steps 1 and 2, the values x_5, x_{50} and x_{95} are obtained.

Step 3. We calculate the auxiliary quantity

$$s_z = \frac{2 \log x_{50} - \log x_5 - \log x_{95}}{\log x_5 - \log x_{95}} \tag{3.159}$$

and the sample coefficient of skewness

$$C_{sz} = 7.15 s_z \tag{3.160}$$

The calculation continues according to whether $C_{sz} > 0$ or $C_{sz} < 0$ (see Votruba and Nacházel, 1971, p. 64).

3.1.7 Tables and probability papers for constructing the exceedance probability curves

An analytical formula for a probability distribution is usually very difficult to obtain, therefore numerical and graphical aids, such as tables and probability charts, are used for constructing exceedance probability curves.

For the third type of Pearson distribution tables were constructed by E. E. Foster and S. J. Rybkin. Rybkin's table, completed by B. I. Serpik, is very comprehensive (Alexeyev, 1960, pp. 136, 137). The values are calculated for $p = 0.01\%$ up to 100% and $C_s = 0.0$ up to 5.2 under $C_v = 1$ (Table 3.3). The table serves to determine the ordinates of the theoretical exceedance probabilities curve with given values $\bar{x}$, C_v, C_s.

In the row corresponding to a given value C_s, the values $\Phi(p, C_s)$ for various exceedance probabilities are found. The modulus factors $k_p = x_p/\bar{x}$ or (directly) the ordinates x_p of the exceeding probability curve are calculated from the relations

$$k_p = \Phi(p, C_s)\, C_v + 1 \tag{3.161}$$

$$x_p = [\Phi(p, C_s)\, C_v + 1]\, \bar{x} \tag{3.162}$$

Example: The value of average annual flow corresponding to $p = 95\%$ is to be calculated, providing that the Pearson probability law holds, with $Q_a = 300\ \mathrm{m^3\ s^{-1}}$, $C_v = 0.29$, $C_s = 0.8$.

For $C_s = 0.8$ and $p = 95\%$, the value $\Phi(p, C_s) = -1.38$ is found in the table; $k_{95} = -1.38 \cdot 0.29 + 1 = 0.60$ or $Q_{r,95} = [-1.38 \cdot 0.29 + 1] \cdot 300 = 180\ \mathrm{m^3\ s^{-1}}$ is calculated.

Probability papers

Exceedance probabilities curves represented in a coordinate system with a linear scale (especially for the horizontal axis P) are not suitable because of inaccuracies which may occur for the extreme values of probability ($P \to 0$, $P \to 1$). Therefore, for a graphical description of both empirical and theoretical exceedance probability

Table 3.3 Deviations of ordinates for the binomial curve of probability exceedance $\frac{x_i - \bar{x}}{\sigma} = \Phi(p, C_s)$ (according to B. L. Serpik)

C_s	Reliability										
	0.01	0.1	0.5	1	2	3	5	10	20	25	30
0.0	3.72	3.09	2.58	2.33	2.02	1.88	1.64	1.28	0.84	0.67	0.52
0.1	3.94	3.23	2.67	2.40	2.11	1.92	1.67	1.29	0.84	0.66	0.51
0.2	4.16	3.38	2.76	2.47	2.16	1.96	1.70	1.30	0.83	0.65	0.50
0.3	4.38	3.52	2.86	2.54	2.21	2.00	1.72	1.31	0.82	0.64	0.48
0.4	4.61	3.66	2.95	2.61	2.26	2.04	1.75	1.32	0.82	0.63	0.47
0.5	4.83	3.81	3.04	2.68	2.31	2.08	1.77	1.32	0.81	0.62	0.46
0.6	5.05	3.96	3.13	2.75	2.35	2.12	1.80	1.33	0.80	0.61	0.44
0.7	5.28	4.10	3.22	2.82	2.40	2.15	1.82	1.33	0.79	0.59	0.43
0.8	5.50	4.24	3.31	2.89	2.45	2.18	1.84	1.34	0.78	0.58	0.41
0.9	5.73	4.38	3.40	2.96	2.50	2.22	1.86	1.34	0.77	0.57	0.40
1.0	5.96	4.53	3.49	3.02	2.54	2.25	1.88	1.34	0.76	0.55	0.38
1.1	6.18	4.67	3.58	3.09	2.58	2.28	1.89	1.34	0.74	0.54	0.36
1.2	6.41	4.81	3.66	3.15	2.62	2.31	1.92	1.34	0.73	0.52	0.35
1.3	6.64	4.95	3.74	3.21	2.67	2.34	1.94	1.34	0.72	0.51	0.33
1.4	6.87	5.09	3.83	3.27	2.71	2.37	1.95	1.34	0.71	0.49	0.31
1.5	7.09	5.28	3.91	3.33	2.74	2.39	1.96	1.33	0.69	0.47	0.30
1.6	7.31	5.37	3.99	3.39	2.78	2.42	1.97	1.33	0.68	0.46	0.28
1.7	7.54	5.50	4.07	3.44	2.82	2.44	1.98	1.32	0.66	0.44	0.26
1.8	7.76	5.64	4.15	3.50	2.85	2.46	1.99	1.32	0.64	0.42	0.24
1.9	7.98	5.77	4.23	3.55	2.88	2.49	2.00	1.31	0.63	0.40	0.22
2.0	8.21	5.91	4.30	3.60	2.91	2.51	2.00	1.30	0.61	0.39	0.20
2.1	–	6.04	4.38	3.65	2.94	2.53	2.01	1.29	0.59	0.37	0.18
2.2		6.14	4.46	3.68	2.95	2.54	2.02	1.27	0.57	0.35	0.16
2.3		6.26	4.52	3.73	2.98	2.57	2.01	1.26	0.55	0.32	0.14
2.4		6.37	4.59	3.78	3.02	2.60	2.00	1.25	0.52	0.29	0.12
2.5		6.50	4.66	3.82	3.05	2.62	2.00	1.23	0.50	0.27	0.10
2.6		6.54	4.71	3.86	3.08	2.63	2.00	1.21	0.48	0.25	0.085
2.7		6.75	4.80	3.92	3.10	2.64	2.00	1.19	0.46	0.24	0.070
2.8		6.86	4.86	3.96	3.12	2.65	2.00	1.18	0.44	0.22	0.057
2.9		7.00	4.91	4.01	3.12	2.66	1.99	1.15	0.41	0.20	0.041
3.0		7.10	4.95	4.05	3.14	2.66	1.97	1.13	0.39	0.19	0.027
3.1		7.23	5.01	4.09	3.14	2.66	1.97	1.11	0.37	0.17	0.010
3.2		7.35	5.08	4.11	3.14	2.66	1.96	1.09	0.35	0.15	−0.006
3.3		7.44	5.14	4.15	3.14	2.66	1.95	1.08	0.33	0.13	−0.022
3.4		7.54	5.19	4.18	3.15	2.66	1.94	1.06	0.31	0.11	−0.036
3.5		7.64	5.25	4.21	3.16	2.66	1.93	1.04	0.29	0.085	−0.049

40	50	60	70	75	80	90	95	97	99	99.9	100
0.25	0.00	−0.25	−0.52	−0.67	−0.84	−1.28	−1.64	−1.88	−2.33	−3.09	$-\infty$
0.24	−0.02	−0.27	−0.53	−0.68	−0.85	−1.27	−1.61	−1.84	−2.25	−2.95	−20.0
0.22	−0.03	−0.28	−0.55	−0.69	−0.85	−1.26	−1.58	−1.79	−2.18	−2.81	−10.0
0.20	−0.05	−0.30	−0.56	−0.70	−0.85	−1.24	−1.55	−1.75	−2.10	−2.67	−6.67
0.19	−0.07	−0.31	−0.57	−0.71	−0.85	−1.23	−1.52	−1.70	−2.03	−2.54	−5.00
0.17	−0.08	−0.33	−0.58	−0.71	−0.85	−1.22	−1.49	−1.66	−1.96	−2.40	−4.00
0.16	−0.10	−0.34	−0.59	−0.72	−0.85	−1.20	−1.45	−1.61	−1.88	−2.27	−3.33
0.14	−0.12	−0.36	−0.60	−0.72	−0.85	−1.18	−1.42	−1.57	−1.81	−2.14	−2.86
0.12	−0.13	−0.37	−0.60	−0.73	−0.86	−1.17	−1.38	−1.52	−1.74	−2.02	−2.50
0.11	−0.15	−0.38	−0.61	−0.73	−0.85	−1.15	−1.35	−1.47	−1.66	−1.90	−2.22
0.09	−0.16	−0.39	−0.62	−0.73	−0.85	−1.13	−1.32	−1.42	−1.59	−1.79	−2.00
0.07	−0.18	−0.41	−0.62	−0.74	−0.85	−1.10	−1.28	−1.38	−1.52	−1.68	−1.82
0.05	−0.19	−0.42	−0.63	−0.74	−0.84	−1.08	−1.24	−1.33	−1.45	−1.58	−1.67
0.04	−0.21	−0.43	−0.63	−0.74	−0.84	−1.06	−1.20	−1.28	−1.38	−1.48	−1.54
0.02	−0.22	−0.44	−0.64	−0.73	−0.83	−1.04	−1.17	−1.23	−1.32	−1.39	−1.43
0.00	−0.24	−0.45	−0.64	−0.73	−0.82	−1.02	−1.13	−1.19	−1.26	−1.31	−1.33
−0.02	−0.25	−0.46	−0.64	−0.73	−0.81	−0.99	−1.10	−1.14	−1.20	−1.24	−1.25
−0.03	−0.27	−0.47	−0.64	−0.72	−0.81	−0.97	−1.06	−1.10	−1.14	−1.17	−1.18
−0.05	−0.28	−0.48	−0.64	−0.72	−0.80	−0.94	−1.02	−1.06	−1.09	−1.11	−1.11
−0.07	−0.29	−0.48	−0.64	−0.72	−0.79	−0.92	−0.98	−1.01	−1.04	−1.05	−1.05
−0.08	−0.31	−0.49	−0.64	−0.71	−0.78	−0.90	−0.95	−0.97	−0.99	−1.00	−1.00
−0.10	−0.32	−0.50	−0.64	−0.70	−0.76	−0.886	−0.914	−0.930	−0.945	−0.952	−0.952
−0.12	−0.33	−0.50	−0.64	−0.69	−0.75	−0.842	−0.882	−0.895	−0.905	−0.910	−0.910
−0.13	−0.34	−0.50	−0.63	−0.68	−0.74	−0.815	−0.850	−0.860	−0.867	−0.870	−0.870
−0.14	−0.35	−0.51	−0.62	−0.67	−0.72	−0.792	−0.820	−0.826	−0.830	−0.833	−0.833
−0.16	−0.36	−0.51	−0.62	−0.66	−0.71	−0.768	−0.790	−0.795	−0.800	−0.800	−0.800
−0.17	−0.37	−0.51	−0.61	−0.66	−0.70	−0.746	−0.764	−0.766	−0.770	−0.770	−0.770
−0.18	−0.38	−0.51	−0.61	−0.65	−0.68	−0.724	−0.736	−0.739	−0.740	−0.740	−0.740
−0.20	−0.39	−0.51	−0.60	−0.64	−0.67	−0.703	−0.711	−0.714	−0.715	−0.715	−0.715
−0.21	−0.39	−0.51	−0.60	−0.63	−0.65	−0.681	−0.689	−0.690	−0.690	−0.690	−0.690
−0.22	−0.40	−0.51	−0.59	−0.62	−0.64	−0.661	−0.665	−0.666	−0.666	−0.667	−0.667
−0.23	−0.40	−0.51	−0.58	−0.60	−0.62	−0.641	−0.645	−0.646	−0.646	−0.646	−0.646
−0.25	−0.41	−0.51	−0.57	−0.59	−0.61	−0.621	−0.625	−0.625	−0.625	−0.625	−0.625
−0.26	−0.41	−0.50	−0.56	−0.58	−0.59	−0.605	−0.606	−0.606	−0.606	−0.607	−0.607
−0.27	−0.41	−0.50	−0.55	−0.57	−0.58	−0.586	−0.588	−0.588	−0.588	−0.588	−0.588
−0.28	−0.41	−0.50	−0.54	−0.55	−0.56	−0.570	−0.571	−0.571	−0.571	−0.572	−0.572

Table 3.3 (*continued*)

C_s	Reliability										
	0.01	0.1	0.5	1	2	3	5	10	20	25	30
3.6		7.72	5.30	4.24	3.17	2.66	1.93	1.03	0.28	0.064	−0.072
3.7		7.86	5.35	4.26	3.18	2.66	1.91	1.01	0.26	0.048	−0.084
3.8		7.97	5.40	4.29	3.18	2.65	1.90	1.00	0.24	0.032	−0.095
3.9		8.08	5.45	4.32	3.20	2.65	1.90	0.98	0.23	0.020	−0.11
4.0		8.17	5.50	4.34	3.20	2.65	1.90	0.96	0.21	0.010	−0.12
4.1		8.29	5.55	4.36	3.22	2.65	1.89	0.95	0.20	0.000	−0.13
4.2		8.38	5.60	4.39	3.24	2.64	1.88	0.93	0.09	−0.010	−0.13
4.3		8.49	5.65	4.40	3.24	2.64	1.87	0.92	0.17	−0.021	−0.14
4.4		8.60	5.69	4.42	3.25	2.63	1.86	0.91	0.15	−0.032	−0.15
4.5		8.69	5.74	4.44	3.26	2.62	1.85	0.89	0.14	−0.042	−0.16
4.6		8.79	5.79	4.46	3.27	2.62	1.84	0.87	0.13	−0.052	−0.17
4.7		8.89	5.84	4.49	3.28	2.61	1.83	0.85	0.11	−0.064	−0.18
4.8		8.96	5.89	4.50	3.29	2.60	1.81	0.82	0.10	−0.075	−0.19
4.9		9.04	5.90	4.51	3.30	2.60	1.80	0.80	0.084	−0.087	−0.19
5.0		9.12	5.94	4.54	3.32	2.60	1.78	0.78	0.068	−0.099	−0.20
5.1		9.20	5.98	4.57	3.32	2.60	1.76	0.76	0.051	−0.11	−0.21
5.2		9.27	6.02	4.59	3.33	2.60	1.74	0.73	0.035	−0.12	−0.21

curves (or distribution functions), and for graphical statistical analysis (extrapolation, graphico-numerical estimations of distribution parameters, etc.), we use papers with a special scale typical to the increasing scale for the probabilities P tending to the extreme values. Thus the exceedance probabilities curve is represented by a straight line, or by a line with low curvature. In the probability paper, corresponding to a probability distribution, the theoretical exceedance probabilities curve of the distribution is represented by a straight line. Therefore the probability paper is chosen according to which distribution we presume that the random sample is from.

The probability paper of the normal distribution, having a linear vertical scale, is also called normal probability paper.

The probability paper of the log-normal distribution differs from the preceding one by a logarithmic vertical scale. The log-normal curves are represented as straight lines providing the relation $C_s = 3C_v + C_v^3$ holds.

The same vertical scale, but a different horizontal scale for the relative values of the variable $k_i = x_i/\bar{x}$, correspond to Brovkovich's probability paper. Here, straight lines represent all binomial distribution curves, providing the relation $C_s = 2C_v$ holds.

40	50	60	70	75	80	90	95	97	99	99.9	100
−0.28	−0.42	−0.49	−0.54	−0.54	−0.55	−0.555	−0.556	−0.556	−0.556	−0.556	−0.556
−0.29	−0.42	−0.48	−0.52	−0.53	−0.54	−0.541	−0.541	−0.541	−0.541	−0.541	−0.541
−0.30	−0.42	−0.48	−0.51	−0.52	−0.52	−0.526	−0.526	−0.526	−0.526	−0.527	−0.527
−0.30	−0.41	−0.47	−0.50	−0.51	−0.51	−0.513	−0.513	−0.513	−0.513	−0.513	−0.513
−0.31	−0.41	−0.46	−0.49	−0.49	−0.50	−0.500	−0.500	−0.500	−0.500	−0.500	−0.500
−0.31	−0.41	−0.46	−0.48	−0.484	−0.486	−0.487	−0.487	−0.487	−0.488	−0.488	−0.488
−0.31	−0.41	−0.45	−0.47	−0.473	−0.475	−0.476	−0.476	−0.476	−0.477	−0.477	−0.477
−0.32	−0.40	−0.44	−0.46	−0.462	−0.465	−0.465	−0.465	−0.465	−0.465	−0.465	−0.465
−0.32	−0.40	−0.44	−0.451	−0.454	−0.455	−0.455	−0.455	−0.455	−0.455	−0.455	−0.455
−0.32	−0.40	−0.43	−0.441	−0.444	−0.445	−0.445	−0.445	−0.445	−0.445	−0.445	−0.445
−0.32	−0.40	−0.42	−0.432	−0.434	−0.435	−0.435	−0.435	−0.435	−0.435	−0.435	−0.435
−0.32	−0.40	−0.42	−0.424	−0.425	−0.426	−0.426	−0.426	−0.426	−0.426	−0.426	−0.426
−0.32	−0.39	−0.41	−0.416	−0.416	−0.416	−0.416	−0.416	−0.416	−0.417	−0.417	−0.417
−0.33	−0.386	−0.401	−0.407	−0.407	−0.407	−0.408	−0.408	−0.408	−0.408	−0.408	−0.408
−0.33	−0.380	−0.395	−0.399	−0.400	−0.400	−0.400	−0.400	−0.400	−0.400	−0.400	−0.400
−0.33	−0.376	−0.388	−0.391	−0.392	−0.392	−0.392	−0.392	−0.392	−0.392	−0.393	−0.393
−0.33	−0.370	−0.382	−0.384	−0.385	−0.385	−0.385	−0.385	−0.385	−0.385	−0.385	−0.385

Using the horizontal scale linearizing the distribution of the standardized z, we obtain the Gumbel probability paper, the vertical scale of which is linear.

The Fréchet probability paper differs from the preceding one by a logarithmic vertical scale.

The horizontal scale of the Goodrich probability paper is $\log\log 1/P'$ and the vertical is $\log x$.

The probability papers of the normal Gauss–Laplace distribution (and hence that of the log-normal one which has the same horizontal scale), the Gumbel distribution (hence the Fréchet distribution), and the Goodrich distribution may be constructed for probabilities in the range from 0.01% to 99.99%.

Brovkovich's paper serves to illustrate the application of probability papers. The empirical distribution is approximated by a straight line tending to the same value C_v in the appropriate scales both on the left and right sides of the chart.

Brovkovich's paper may be used to check the correct calculation of C_v. If the straight line, corresponding to the estimate of C_v, approximates the empirical distribution well, then the calculation of C_v is correct, if not, it is incorrect.

Table 3.4 Probability distributions used in water management

Type of distribution	Density $f(x)$ Distribution function $F(x)$	Domain of variable	Parameters	Mean value μ
normal (Gauss–Laplace)	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	$-\infty$ to $+\infty$	μ, σ	μ
standardized normal	$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$	$-\infty$ to $+\infty$	–	$\emptyset$
3parameters log-normal (Gauss–Gibrat)	$F(x) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{y} e^{-y^2}\, dy$		a, b, x_0	
2parameters log-normal (Ven-Te-Chow)	$f(x) = \frac{1}{x \cdot \sigma(y)\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu(y))^2}{2\sigma^2(y)}\right]$	$\emptyset$ to $+\infty$	$\mu(y), \sigma(y)$	$\exp[\mu(y) + \frac{1}{2}\sigma^2(y)]$
4parameters log-normal (Johnson)	$f(x) = \frac{1}{\sigma(y)\exp\left(\frac{x-a}{b-x}\right)\sqrt{2\pi}} \exp - \left\{\frac{1}{2}\left[\frac{\ln\frac{x-a}{b-x} - \mu(y)}{\sigma(y)}\right]^2\right\}$	a to b	$\mu(y), \sigma(y), a, b$	
3parameters exponential (Weibull)	$f(x) = \alpha\beta(x - x_0)^{\alpha-1} \exp[-\beta(x-x_0)^\alpha]$ $F(x) = 1 - \exp[-\beta(x-x_0)^\alpha]$	x_0 to $+\infty$	α, β, x_0	$x_0 + \frac{1}{\sqrt[\alpha]{\beta}}\Gamma\left(\frac{\alpha+1}{\alpha}\right)$
2parameters double exponential (Gumbel)	$f(x) = e^{-y}\exp(-e^{-y})$ $y = \alpha(x-\beta)$ $F(x) = \exp(-e^{-y})$	$-\infty$ to $+\infty$	α, β	$\beta + \frac{0.577}{\alpha}$

2parameters beta-distribution (Pearson I)	$f(z)$	$= \dfrac{z^{\alpha-1}(1-z)^{\beta-1}}{B(\alpha,\beta)}$	$\emptyset$ to 1	α, β $\alpha > \emptyset;\ \beta > \emptyset$	$\dfrac{\alpha}{\alpha+\beta}$
gamma-distribution Pearson III	$f(x)$	$= \dfrac{1}{d\left(\frac{a}{d}+1\right)\Gamma\left(\frac{a}{d}+1\right)}(x-x_0)^{a/d}\exp\left(-\dfrac{x-x_0}{d}\right)$	α_0 to $+\infty$	a, d, x_0	$x_0 + a + d$
binomial	$f(x)$	$= \binom{n}{x} p^x q^{n-x}$	$\emptyset \leqq x < n$	n, p	np
	q	$= 1 - p$			
Poisson	$f(x)$	$= \dfrac{\lambda^x}{x!}e^{-\lambda}$	$\emptyset \leqq x \leqq \infty$	λ	λ
χ^2	$f_m(\chi^2)$	$= \dfrac{1}{2^{m/2}\Gamma\left(\frac{m}{2}\right)}(\chi^2)^{(m/2)-1}e^{-1/2\chi^2}$	$\emptyset$ to $+\infty$	m (number of degrees of freedom)	m
Student	$f_m(t)$	$= \dfrac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\,\Gamma\left(\frac{m}{2}\right)}\left(1+\dfrac{t^2}{m}\right)^{-(m+1)/2}$	$-\infty$ to $+\infty$	m (number of degrees of freedom)	$\emptyset$
Snedecor (F)	$f_{m_1,m_2}(F)$	$= \dfrac{m_1}{m_2}\dfrac{\Gamma\frac{(m_1+m_2)}{2}}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)}\left(\dfrac{m_1}{m_2}F\right)^{(m_1/2)-1}\left(1+\dfrac{m_1}{m_2}F\right)^{-(m_1+m_2)/2}$	0 to $+\infty$	m_1, m_2 (degrees of freedom)	$\dfrac{m_2}{m_2-2}$
Fisher (z)	logarithmic transformation of F-distribution		$-\infty$ to $+\infty$	m_1, m_2 (degrees of freedom)	$\dfrac{m_2}{m_2-2}$ for $m_2 > 2$
	z	$= \frac{1}{2}\lg F$			

Table 3.4 (continued)

Type of distribution	Variance σ^2	Skewness C_s	Note
normal (Gauss–Laplace)	σ^2	$\emptyset$	–
standardized normal	1	$\emptyset$	$t = \dfrac{x-\mu}{\sigma}$
3parameters log-normal (Gauss–Gibrat)			$y = a\log(x - x_0) + b$ x_0 – close to min. observed value
2parameters log-normal (Ven-Te-Chow)	$\mu^2(y)\,(e^{\sigma^2(y)} - 1)$	$3C_v(x) + C_v^3(x)$	$y = \ln x$
4parameters log-normal (Johnson)			$y = \ln\dfrac{x-a}{b-x}$
3parameters exponential (Weibull)	$\dfrac{1}{\sqrt[\alpha]{\beta^2}}\left[\Gamma\left(\dfrac{\alpha+2}{\alpha}\right) - \Gamma^2\left(\dfrac{\alpha+1}{\alpha}\right)\right]$	$\dfrac{\Gamma\left(\dfrac{\alpha+3}{\alpha}\right) - 3\Gamma\left(\dfrac{\alpha+1}{\alpha}\right)\Gamma\left(\dfrac{\alpha+2}{\alpha}\right) + 2\Gamma^3\left(\dfrac{\alpha+1}{\alpha}\right)}{\left[\Gamma\left(\dfrac{\alpha+2}{\alpha}\right) - \Gamma^2\left(\dfrac{\alpha+1}{\alpha}\right)\right]^{3/2}}$	
2parameters double exponential (Gumbel)	$\dfrac{\pi^2}{6\alpha^2}$	~ 1.14 (const.)	$\alpha = \dfrac{\pi}{\sigma(x)\sqrt{6}}$; $\beta = \mu(x) - \dfrac{0.577}{\alpha}$
2parameters beta-distribution (Pearson I)	$\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$2\sqrt{\dfrac{\alpha+\beta+1}{\alpha\beta}}\,\dfrac{\beta-\alpha}{\alpha+\beta+2}$	$B(\alpha,\beta) = \displaystyle\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = \dfrac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

gamma distribution (Pearson III)	$d(a+d)$	$2\sqrt{\frac{d}{a+d}}$	x_0 – min. value of x a – distance from the mode to the origin d – distance from the mode to the centre of gravity
binomial	npq	$\frac{1-2p}{npq}$	$\binom{n}{x} = \frac{n!}{(n-x)!\,x!}$
Poisson	λ	$\frac{1}{\sqrt{\lambda}}$	$\lambda = pn, \quad \lambda > 0$
χ^2	$2m$	$2\sqrt{\frac{2}{m}}$	$\lambda^2 = \sum_{i=1}^{n} \frac{(x_i - \mu(x))^2}{\sigma^2}$
Student	$\frac{m}{m-2}$	0	for $m \to \infty$ close to normal
Snedecor (F)	$\frac{m_2^2}{(m_2-2)^2}\,\frac{2(m_1+m_2-2)}{m_1(m_2-4)}$	asymmetric	$F = \frac{m_2 u}{m_1 v}$; independent variables u, v are distributed χ^2
Fisher (z)	$\frac{m_2^2}{(m_2-2)^2}\,\frac{2(m_1+m_2-2)}{m_1(m_2-4)}$ for $m_2 > 4$	asymmetric	more symmetric than F-distribution

3.1.8 *Applicability of various probability distributions*

The suitability of applying certain probability distributions (summary in Table 3.4) to a given sample is most effectively checked by marking several points on a probability paper; the paper is chosen so that the points are close to a straight line.

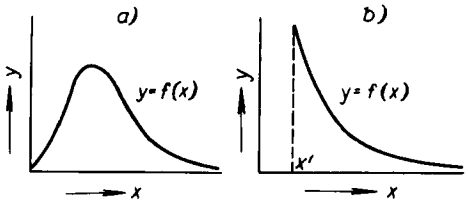

Fig. 3.15 Typical shape of probability density of
(*a*) Pearson, three parameter, log-normal, Gumbel and Fréchet distribution; (*b*) exponential distribution

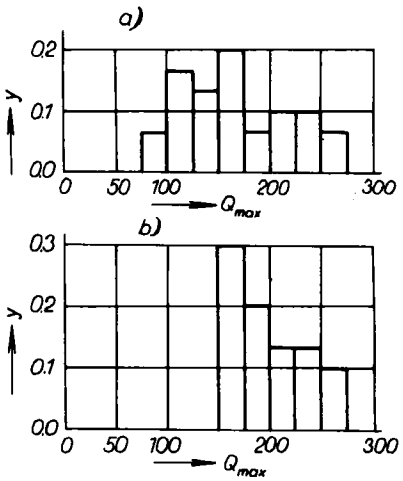

Fig. 3.16 Frequency distribution (histograms) of flood maxima on the Otava river at Písek (1931–1960)
(*a*) annual; (*b*) thirty largest floods

The exponential distribution differs essentially from the other mentioned distributions by the course of the probability density $y = f(x)$. This curve assumes its maximum (mode) at the point x' to which the exceedance probability $P = 1$ corresponds. (Fig. 3.15b); the probability density of the Pearson, log-normal, Gumbel and other distributions at first increases up to the mode then decreases (Fig. 3.15a). For example, the sample of observations of maximum flood discharges represented by annual maxima on the Otava river in Písek from 1931 to 1960 yields a frequency distribution (histogram) (Fig. 3.16a) corresponding to the theoretical distribution in Fig. 3.15a. The samples of the thirty greatest observed flow maxima in the same period, however, have a distribution (Fig. 3.16b) corresponding to exponential law (Fig. 3.15).

In spite of being similar in form, the probability density functions $y = f(x)$ of various distribution laws differ essentially, namely in values of the radius of asymmetry $b = \bar{x} - \hat{x}$, the frequency of the mode $y_{max} = f(\hat{x})$, and the asymptotic behaviour for $P(x) \to 0$. The latter has a great influence on the extrapolation of the exceedance probability curve towards the extreme values. The convergence is very

slow in case of the Fréchet distribution, as well as in the case of the Galton distribution.

In water management the average annual flows and maximum flood discharges are most frequently analysed by means of statistics. Alexeyev (1960) stated that in the case of average annual flows, the use of the binomial, the three-parameter (Kritsky and Menkel), and the log-normal distributions yield approximately the same results. For the maximum annual flood discharges the three parameter and the log-normal distributions are recommended.

In Czechoslovakia, particularly, distributions of all the maxima correspond, due to their nature, to the exponential distribution laws; the more general Goodrich exponential law is the most suitable.

The theoretical distribution law may be chosen from among those mentioned above, to fit the sample values. Extrapolation of the curves out of the range of the observed values towards the extreme values of probability $P = 0$ and $P = 1$, presumes an agreement between the theoretical and the empirical laws. The uncertainty decreases with the growing number of observations, but cannot be avoided completely because it is impossible to obtain sufficiently long series of observations to determine the probabilities of extreme values reliably. Therefore, the extrapolated values have only a conventional meaning; the advantage is that they are unambiguous for the chosen conditions.

3.1.9 Accuracy of statistical characteristics in water management

Let us evaluate the accuracy of the most frequently used statistical characteristics in hydrology and water management which are

(a) long-term average (modulus), equation (3.47)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(b) sample standard deviation, equation (3.48)

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

(c) sample coefficient of variation, equation (3.49)

$$C_v = \sqrt{\frac{\sum_{i=1}^{n} (k_i - 1)^2}{n - 1}}$$

(d) sample coefficient of skewness, equation (3.51)

$$C_s = \frac{\sum_{i=1}^{n} (k_i - 1)^3}{(n-1)C_v^3}$$

(e) sample coefficient of excess, equation (3.46)

$$E = \frac{M_4}{s^4} - 3$$

The form of the exceedance probabilities curve is influenced the most by C_v, less by C_s and even less by the excess E (Fig. 3.17). The sample coefficient of variation $C_v = 0$ corresponds to a sample consisting of values mutually identical; increasing C_v means that the sample values differ essentially from the average. Figure 3.17d

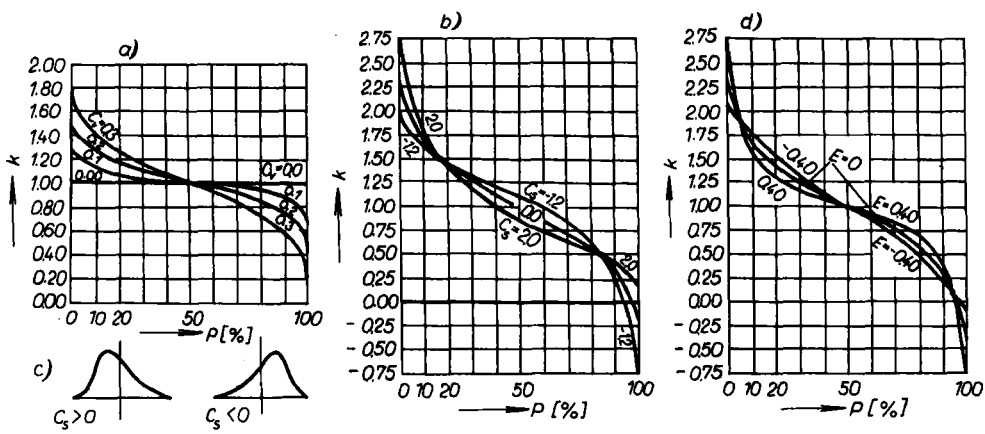

Fig. 3.17 Influence of C_v, C_s and E on the shape of the Pearson type III exceedance probability curve (*a*) influence of C_v ($C_s = 0$); (*b*) C_s ($C_v = 0.5$); (*c*) graphic illustration of positive and negative skewness; (*d*) influence of E ($C_v = 0.5$, $C_s = 0$)

shows both positive skewness ($C_s > 0$, the mode to the left of the mean) and negative skewness ($C_s < 0$, the mode to the right of the mean). On the diagram (Fig. 3.17b), a graphical descriptions of the curves with $C_v = 0.5$ is given for three different values of C_s. The course of the curves is less dissimilar than in Fig. 3.17a, especially in the middle part. The greatest difference can be observed between the respective ordinates corresponding to the extreme values of P. Greater values of C_s yield greater values of the end ordinates. The curves corresponding to the smaller values of C_s are more suitable from the low flow period viewpoint. Therefore, as a rule, the minimum admissible value of C_s is chosen if the series of observations is not long enough to provide a reliable calculation of C_s.

From Foster's expression $C_s = 2b/C_v$ (Morozov, vol. I, 1954), it follows that for the flows the lower boundary of C_s is

$$C_s \geqq 2C_v \tag{3.163}$$

The upper boundary follows from the equation of the Pearson type III curve

$$C_s = \frac{2C_v}{1 - k_{min}} \tag{3.164}$$

The relation $C_s = 2C_v$ is often used; it results from the theory of binomial distribution curves. The values of C_s might, however, be smaller, close to zero or even negative, or on the other hand much greater. It is recommended not to insist strictly on the Foster relation $C_s \geqq 2C_v$ and to choose according to circumstances, even $C_s < 2C_v$.

With $C_s < 2C_v$ the binomial distribution curve may, for a greater dependability of P, in case the of flows, have unrealistic negative values of k.

The accuracy of statistical characteristics is influenced not only by the observation time, but also by the variability of flows.

The coefficient of variation of annual runoff for most rivers is in the range of 0.10 to 1.20.

The influence of the number of observations n and the coefficient of variation C_v on the accuracy of the characteristics of runoff ($\bar{x}$, s, C_v, C_s and E) is given by estimating their standard errors (Votruba – Broža, 1966, p. 82).

In order to illustrate the above, Table 3.5 shows the standard errors of all the statistical characteristics under $C_v = 0.1$ and 1.0, providing that $n = 50$, $C_s = 2C_v$ and $E = \frac{3}{2}C_s^2 = 6C_v^2$.

Table 3.5 Standard errors of statistical characteristics $\bar{x}$, a, C_v, C_s, E for number of observations $n = 50$

C_v	$\sigma_{\bar{x}}$ [%]	σ_s, σ_{C_v} [%]	σ_{C_s} [%]	σ_E [%]
0.1	1.4	10	149	1400
1.0	14.0	20	60	210

The real hydrological series of many Czechoslovak rivers are sufficiently accurate to determine the characteristics $\bar{x}$, s and C_v; the errors σ_{C_s} and especially σ_E are, however, too great. The calculations above are based on the assumption that the observed random variables are mutually independent. If the assumption is not fulfilled the errors are even greater.

In dry countries there are rivers that in some years have zero runoff, hence corresponding to an exceedance probability of less than 100%. It therefore holds that $C_s < 2C_v$; according to the requirement of agreement between the theoretical and the empirical distributions, we choose as a rule $C_s = 1.5C_v$ or $C_s = C_v$ (Voskresenski, 1956).

The value C_v may be influenced, even for very long series, by one highly extreme year, the theoretical probability of its occurrence being too small in a series of a given length (e.g., W_r of the extraordinarily high water year 1941 in Czech and Moravian rivers). Svoboda (1964) used the 110 year runoff series obtained in Děčín on the Labe to calculate the long-term values of the coefficient of variation of annual runoff of Czech and Moravian rivers on the basis of shorter series of observations.

For C_v and C_s empirical formulae were derived, applicable for preliminary calculations or for very short series. For Czechoslovak rivers the derivations were made by A. Bratránek, Dub (1953), Svoboda (1963); see also Dub and Němec (1969).

3.1.10 Evaluation of goodness-of-fit between empirical and theoretical distributions

When investigating the same quantity, but under different conditions (e.g., average annual flows, maximum peak discharges, flood volumes, etc.), by means of the probability theory, we cannot estimate the type of theoretical probability distribution unambiguously. Thus, the uncertainty contained in the estimate gives rise to the question as to what extent the chosen theoretical distribution agrees with the empirically obtained quantities.

Using the same coordinate system for a graphical description of both the histogram of the sample and the theoretical frequency function, we obtain the most illustrative idea, at least qualitatively, about the agreement between the theoretical and the empirical distributions; less illustrative is the conception of the table of numerical empirical values of frequency n_i of a respective class interval and theoretical values nf_i where n is the number of elements of the sample, f_i is a part of the area limited by the theoretical distribution curve within the limits of the respective interval.

We now proceed to solve the problem of testing the hypothesis that data form a sample of a random variable ξ with the distribution $P(x)$ of a given type. Such tests are called *tests of goodness of fit*, and are based on a choice of a certain measure of deviation between the theoretical (hypothetical) and the empirical distributions. If in the case investigated the deviation exceeds a fixed level, the hypothesis is rejected, and vice-versa. The tests most frequently used are the χ^2-test and the Kolmogorov–Smirnov test.

Applying the χ^2 *test* introduced by Pearson, we calculate the quantity

$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - nf_i)}{nf_i} \tag{3.165}$$

and compare it with the value χ_p^2 for m degrees of freedom and small level p (usually 5% or 1%). The number of degrees of freedom is

$$m = l - c - 1 \tag{3.166}$$

where l is the number of classes into which the space of the variable is divided, and c is the number of sample characteristics used to estimate the unknown parameters of the distribution $f(x)$.

With $\chi^2 < \chi_p^2$ the agreement between the distributions is considered to be sufficient, with $\chi^2 > \chi_p^2$ it is not. Using this test we must bear in mind that it should hold if $nf_i \geqq 5$. As this condition might not be fulfilled, especially for the outer classes, we pool two or three of them if necessary.

The agreement between the distributions is tested very simply with the aid of the *Kolmogorov–Smirnov test*, based on the probability

$$P\{\max |F_n(x_i) - F(x_i)| > D_\alpha(n)\} = \alpha \tag{3.167}$$

Table 3.6 Critical values $D_\alpha(n)$ of the Kolmogorov–Smirnov test

n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$
1	0.975	0.995	21	0.287	0.344
2	0.842	0.929	22	0.281	0.337
3	0.708	0.829	23	0.275	0.330
4	0.624	0.734	24	0.269	0.323
5	0.563	0.669	25	0.264	0.317
6	0.519	0.617	26	0.259	0.311
7	0.483	0.576	27	0.254	0.305
8	0.454	0.542	28	0.250	0.300
9	0.430	0.513	29	0.246	0.295
10	0.409	0.489	30	0.242	0.290
11	0.391	0.468	31	0.238	0.285
12	0.375	0.449	32	0.234	0.281
13	0.361	0.432	33	0.231	0.277
14	0.349	0.418	34	0.227	0.273
15	0.338	0.404	35	0.224	0.269
16	0.327	0.392	36	0.221	0.265
17	0.318	0.381	37	0.218	0.262
18	0.309	0.371	38	0.215	0.258
19	0.301	0.361	39	0.213	0.255
20	0.294	0.352	40	0.210	0.252
			50	0.190	0.228
			60	0.175	0.210
			over 60	$\frac{1.358}{\sqrt{n}}$	$\frac{1.628}{\sqrt{n}}$

where the expression $\max |F_n(x_i) - F(x_i)|$ means the maximum value of the difference between the theoretical and the empirical distribution functions; the test criterion $D_\alpha(n)$ depends on the level α (usually $\alpha = 5\%$ or 1%) and the number of observations. With $n > 60$ it holds that

$$D_\alpha(n) = \frac{1.358}{\sqrt{n}} \quad \text{for} \quad \alpha = 5\% \tag{3.168}$$

$$D_\alpha(n) = \frac{1.628}{\sqrt{n}} \quad \text{for} \quad \alpha = 1\% \tag{3.169}$$

If the calculated value $D = \max |F_n(x_i) - F(x_i)|$ exceeds the tabulated value $D_\alpha(n)$ (Table 3.6), the hypothesis about the agreement between the theoretical and the empirical distributions is rejected on the given level of significance (the deviation between the theoretical and the empirical distributions is statistically significant).

Statistical *estimation of probability distribution parameters* is one of the most important parts of mathematical statistics.

So-called *sampling distributions* of the sample characteristics are used for the estimation, namely

- χ^2 distribution
- Student distribution
- F-distribution (Snedecor)
- Fisher distribution.

Either point or interval estimates are used to estimate the mean and the variance. The interval estimate is more suitable as it also provides the accuracy of the result. The accuracy depends on the number n of the sample values. A frequent problem is what the number n should be for the unknown parameter to be estimated with a certain (chosen in advance) maximum permissible error or accuracy. The estimated values of parameters can be checked (tested) if the distribution law is considered as known.

3.1.11 Random processes and sequences in hydrology and water management

In a random (stochastic) process, the random variable X depends on a parameter which, as a rule, is the time t. If the variable X is continuous in time we use the term "random process", if the parameter assumes only integers "random sequence" is used. The following considerations for random processes are also valid for random sequences.

Just as a random variable is characterized by its distribution function $F(x)$ and frequency function $f(x)$, a random process can similarly be characterized just by adding the variable parameter as another argument, hence, e.g., $F(x, t)$.

Figure 3.18 gives a graphical description of several realizations $x^{(1)}(t), x^{(2)}(t), \ldots$ $\ldots, x^{(k)}(t)$ of a random process $X(t)$. For a fixed value of parameter $t = t_1$ these realizations can be considered as a sample from a random variable which will be denoted by x_1 in agreement with the index of t_1. The sample values $x^{(1)}(t_1), x^{(2)}(t_1), \ldots$ $\ldots, x^{(k)}(t_i)$ are denoted in the figure by $1, 2, \ldots, k$. We calculate the number m of values less than or equal to, a given value x_1. The relative frequency $P = m/k$ is a function

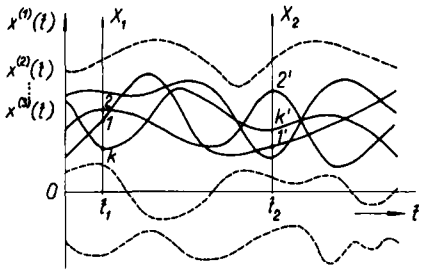

Fig. 3.18 Graph of several realizations of a random process

of x_1 and estimates the probability of the random variable X_1 being less than or equal to x_1. Hence, the distribution function of the random variable X_1 may be written in a form similar to equation (3.18):

$$F(x_1) = P(X_1 \leqq x_1) \tag{3.170}$$

This distribution function characterizing the random process is written as

$$F_1(x_1, t_1) = P(X_1 \leqq x_1) \tag{3.171}$$

and is called the *first (one-dimensional) distribution function of a random process*; it depends on the given parameter (time) t_1 and level x_1.

The *first probability density of a random process* is

$$f_1(x_1, t_1) = \frac{\partial F_1(x_1, t_1)}{\partial x_1} \tag{3.172}$$

The random process is characterized at a fixed time (in static way) by (3.171) and (3.172). We may, however, simultaneously fix another value of parameter t_2.

The *second (two-dimensional) distribution function of a random process* is given by

$$F_2(x_1, t_1; x_2, t_2) = P(X_1 \leqq x_1; X_2 \leqq x_2) \tag{3.173}$$

and denotes the probability of the random process not exceeding the value x_1 at time $t = t_1$ and simultaneously not exceeding the value x_2 at time $t = t_2$.

The *second probability density of the random process* is

$$f_2(x_1, t_1; x_2, t_2) = \frac{\partial^2 F_2(x_1, t_1; x_2, t_2)}{\partial x_1 \, \partial x_2} \tag{3.174}$$

Analogously, the n-dimensional distribution function may be derived, expressing the probability of the random process $X(t)$ not exceeding the levels $x_1, x_2, \ldots, x_n$ in the n values of parameter $t_1, t_2, \ldots, t_n$, respectively.

In problems of water management the first and second-order distributions are usually sufficient. Similarly to the case of random variables, we use certain *numerical characteristics of random process*, namely

(a) the first-order general moment determining the mean value of the random process

$$m_1[X(t)] = \int_{-\infty}^{\infty} x f_1(x, t)\,\mathrm{d}x = \mu(t); \tag{3.175}$$

(b) the second-order central moment determining the variance of the random process

$$M_2[X(t)] = \int_{-\infty}^{\infty} (x - \mu)^2 f_1(x, t)\,\mathrm{d}x = m_2 - m_1^2 = \sigma^2(t) \tag{3.176}$$

(c) the first-order product general moment determining the correlation function of the random process

$$m_1[X(t_1)\,X(t_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_2(x_1, t_1; x_2, t_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = R(t_1, t_2) \tag{3.177}$$

The investigation of the random processes is much simpler if they are stationary and ergodic.

A random process is *stationary* if all the finite-dimensional densities remain the same whenever an arbitrary τ is added to each parameter, hence

$$f_n(x_1, t_1; x_2, t_2; \ldots; x_n t_n) = f_n(x_1, t_1 + \tau; x_2, t_2 + \tau; \ldots; x_n, t_n + \tau) \tag{3.178}$$

In our investigations it will be sufficient to check the first- and the second-order densities.

Similarly, only the invariance of the first and the second distribution functions is usually checked, and consequently that of their numerical characteristics – the mean value eqn. (3.175) and the variance eqn. (3.176) which are constant in the case of a stationary process, and the correlation function eqn. (3.177) which turns to a function of a unique time parameter. All random processes shown in Fig. 3.18 are stationary.

Non-stationary random processes show a certain trend, e.g., either decreasing or increasing the mean value or variance. As an example of a non-stationary random process let us mention the time series of flows in a river site where water is withdrawn for irrigation, with a successive growth of the irrigated area.

Since about 1975 we have been studying theoretically the non-stationary phenomena in hydrology and their consequences for the design of reservoirs (Nacházel and

Patera, 1975; Nacházel, 1976). Since 1976 the non-stationarity has been examined in the Ohře basin. The non-stationarity may either grow continuously due to man's interference in nature, or it may arise suddenly, e.g., as a consequence of building a large new reservoir on a river.

The most significant non-stationarity lies in water demands which are quickly and progressively growing; in water management plans and projects this must not be overlooked. In the future, the non-stationarity of the hydrological quantities will also have to be taken into account.

A random process is *ergodic* if the average of sufficiently many values in the realization of the process is close to the population mean value, with a probability close to one. In Fig. 3.18, the processes represented by the full line are stationary ergodic, those represented by hatched lines are stationary non-ergodic.

In the case of an ergodic process, a large number of realizations need not be investigated, a single long one is sufficient. The assumptions of stationarity and ergodicity enable us to consider one realization as representative for many others. The fulfilment of the assumptions may be checked only with a very long series which may be divided into several parts and their characteristics compared.

The first-order probability densities are mutually the same for a stationary random process, hence we may write

$$f(x_1, t_1) = f_1(x) \tag{3.179}$$

Thus, the calculation of the first-order density of a random process is transformed to the calculation of the density of the random variable X, the realizations of which are given by the values of the realizations of the random process in any fixed time. The mean value $m_1(X)$ of the random variable X is, therefore, the mean value $m_1[X(t)]$ of the random process $X(t)$.

3.1.12 *Correlation (autocorrelation) functions of random processes*

Figure 3.19 shows two random processes $X_1(t)$ and $X_2(t)$ which can have the same mean value and variance, and yet differ essentially: in Fig. 3.19a the value of each realization in the process $X_1(t)$ is decreasing, while in Fig. 3.19b the realizations show

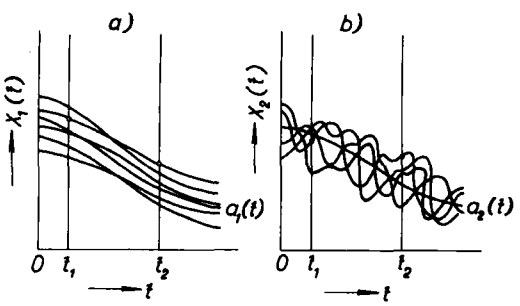

Fig. 3.19 Inner structure of several realizations of a random process

considerable pulsations. This difference is described by the correlation function which is another characteristic of a random process, expressing its inner structure, i.e., the correlation dependence between the values of the process in various moments of time. For a pair with a time lag the value of the correlation function is equal to the coefficient of correlation of the corresponding random variables.

According to equation (3.177), the correlation function $R(t_1, t_2)$ is given by the mean value of the product $[X(t_1) \cdot X(t_2)]$, hence

$$R(t_1, t_2) = m_1[X(t_1)\, X(t_2)] \qquad \text{(equation 3.177')}$$

Sometimes it is useful to express the deviations of the random process from the mean $a(t)$. Such a random process is called centred and is given by

$$\dot{X}(t) = X(t) - a(t) \tag{3.180}$$

Its correlation function is

$$\dot{R}(t_1, t_2) = m_1[\dot{X}(t_1)\, \dot{X}(t_2)] = m_1\{[X(t_1) - a(t_1)]\, [X(t_2) - a(t_2)]\} \tag{3.181}$$

and is sometimes called the correlation function of the pulsations of a random process.

Between the correlation function of a random process [equation (3.178)] and the correlation function of the centred process [equation (3.181)] is the relation

$$\dot{R}(t_1, t_2) = R(t_1, t_2) - a(t_1)\, a(t_2) \tag{3.182}$$

Dividing $\dot{R}(t_1, t_2)$ by the product of the standard deviations $\sigma[X(t_1)]\, \sigma[X(t_2)]$ of the corresponding random variables $X(t_1)$ and $X(t_2)$, we obtain the standardized correlation function of the random process

$$r(t_1, t_2) = \frac{R(t_1, t_2) - a(t_1)\, a(t_2)}{\sigma[X(t_1)]\, \sigma[X(t_2)]} \tag{3.183}$$

It follows from equation (3.183) that the standardized correlation function is equal to the coefficient of correlation of the random variables $X(t_1)$ and $X(t_2)$. Thus for the normalized correlation function it holds that

$$|r| \leqq 1 \tag{3.184}$$

If the random process is stationary, the respective correlation functions of the process and the centred process depend on the difference $\tau = t_2 - t_1$, only, hence

$$R(t_1, t_2) = R(\tau) \quad \text{and} \quad \dot{R}(t_1, t_2) = \dot{R}(\tau) \tag{3.185}$$

The relationship between them is

$$\dot{R}(\tau) = R(\tau) - a^2 \tag{3.186}$$

The standardized correlation function of a stationary random process is

$$r(\tau) = \frac{R(\tau) - a^2}{\sigma_x^2} = \frac{\dot{R}(\tau)}{\sigma_x^2} \tag{3.187}$$

For example of correlation function see Fig. 3.20.

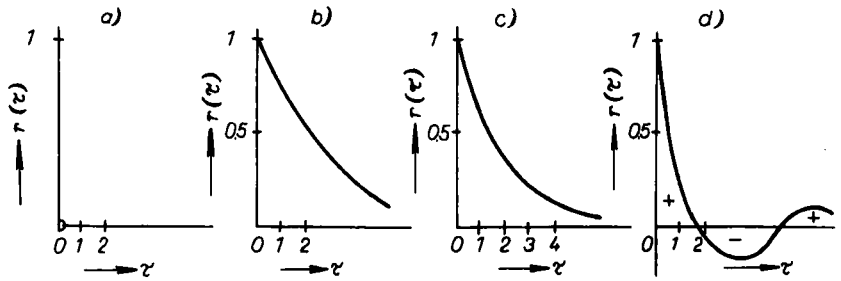

Fig. 3.20 Examples of random-process correlation functions
(*a*) absolutely random process ("white noise"); (*b*) exponential correlation function; (*c*) power correlation function; (*d*) correlation function in the form of a damped harmonic motion

The correlation function in Fig. 3.20a may be described by the relation

$$r(\tau) = \begin{cases} 1 & \text{for } \tau = 0 \\ 0 & \text{for } \tau \neq 0 \end{cases} \tag{3.188}$$

i.e., the values $X(t)$ are uncorrelated for any two moments in time. Such processes are called absolutely random (white noise) and differ from each other only by either mean values and distribution functions. In spite of not being really possible in hydrology, they are used when the problem of storage reservoirs is solved with the aid of general statistical characteristics.

The correlation function in Fig. 3.20b may be expressed by the exponential function

$$r(\tau) = e^{-\alpha(\tau)} \qquad (\alpha > 0) \tag{3.189}$$

The curve rapidly tends to zero. The constant α determines the rate of disappearance of the correlation.

The correlation function in Fig. 3.20c is given by the power function

$$r(\tau) = r(1)^{\tau} \tag{3.190}$$

where $r(1) = r(\tau)_{\tau=1}$.

This curve also decreases rapidly; e.g., with $r(1) = 0.3$ it holds that $r(3) = 0.027$ for $\tau = 3$.

The power correlation function is typical for the simple Markov process which is used to design the storage reservoir volume.

The correlation function in Fig. 3.20d may be expressed as the damped harmonic motion

$$r(\tau) = e^{-\alpha(\tau)} \cos \beta\tau \tag{3.191}$$

For small τ the curve decreases, as in Fig. 3.20b, c; then it turns to negative values, and oscillates around zero, the amplitudes becoming smaller and smaller. It was proved that the empirical correlation functions of rivers in Czechoslovakia are of a harmonic nature; therefore they may be better approximated by the function (3.191), in spite of not exhibiting any perpetual damping. The possibility of negative correlation of more distant quantities substantially influences the reservoir volume under higher values of the coefficient of safe yield.

Figure 3.21 shows three empirical correlation functions of a 30-year sample moving-average series (Votruba–Nachází, 1971):

(a) that of annual relative Wolf's numbers,

(b) annual precipitations,

(c) average annual flows.

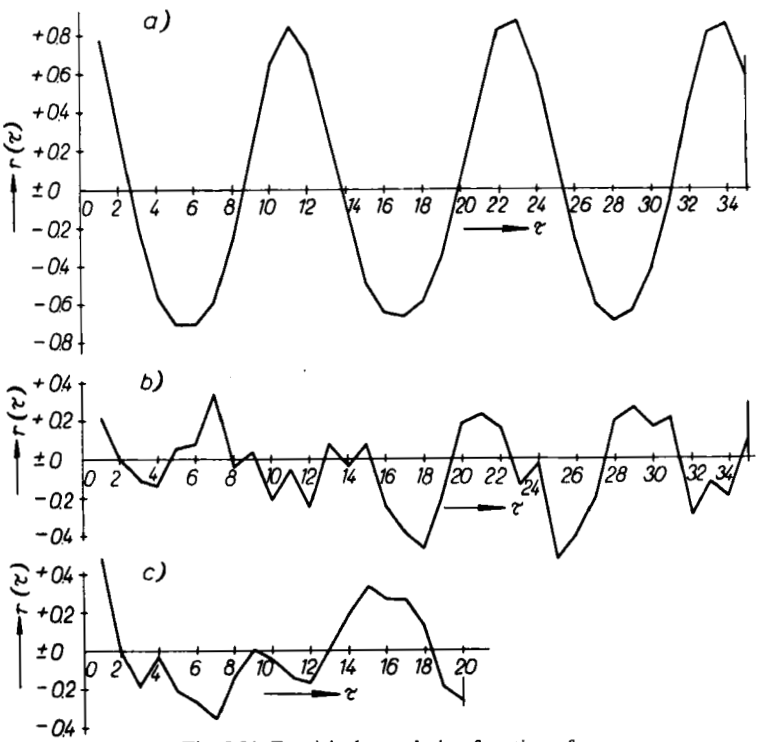

Fig. 3.21 Empirical correlation function of
(*a*) average annual relative Wolf's numbers (1831–1894); (*b*) annual total of precipitations in Prague–Klementinum (1851–1914); (*c*) average annual flows of the river Berounka at Křivoklát (1911–1960)

The correlation function of the average annual relative Wolf's numbers (Fig. 3.21a) has a very regular harmonic behaviour with the values of the positive maxima being mutually similar. The 11-year period of solar activity is very significant.

The correlation function of precipitations (Fig. 3.21b) is also periodic, although not as regular as in Fig. 3.21a. The negative maxima are even higher than the positive ones.

The course of the correlation function of the annual flows (Fig. 3.21c) is similar to that in Fig. 3.21b which proves the similar behaviour of the long-time fluctuation of precipitation totals and runoffs.

Calculation of sample correlation function

In the case of an ergodic random sequence, two approaches may be applied (Fig. 3.22).

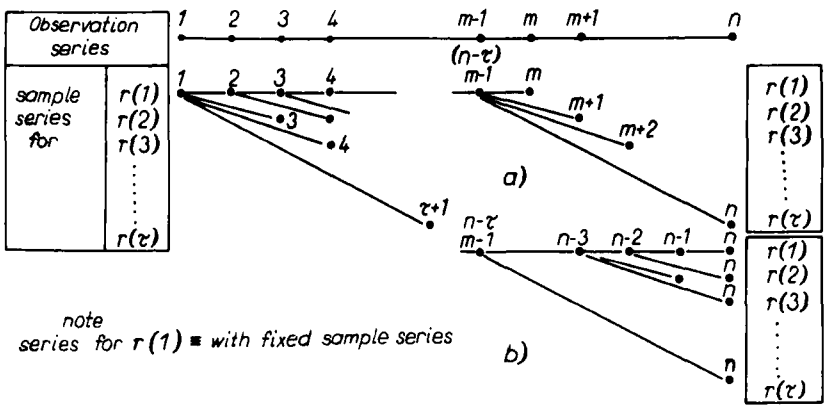

Fig. 3.22 Diagram of calculations of the correlation function of an ergodic random sequence (*a*) for every τ the same number of sample value pairs; (*b*) for every τ the complete sample range

(a) From all the n values of the observation series a shorter sequence containing m members is taken and correlated to the sequence shifted by $1, 2, \ldots, \tau$ (Fig. 3.22a). Hence, according to equation (3.8), it follows that

$$r(\tau) = \frac{\sum_{i=1}^{m} (x_i - \bar{x}_i)(x_{i+\tau} - \bar{x}_{i+\tau})}{\sigma_i \sigma_{i+\tau}(m-1)} \tag{3.192}$$

where x_i are the sample values from x_1 to x_m,

$x_{i+\tau}$ are the sample values from $x_{1+\tau}$ to $x_{m+\tau}$,

τ is the time parameter of the correlation function (time lag), varying from $\tau = 0$ to $\tau_{max} = n - m$ (the complete range $\tau\,(0, \tau_{max})$ need not be exhausted),

$\bar{x}_i, (\bar{x}_{i+\tau})$ is the sample mean of the values $x_i, (x_{i+\tau})$,
$\sigma_i, (\sigma_{i+\tau})$ is sample standard deviation of the values $x_i, (x_{i+\tau})$.

The advantage of this calculation method consists in assigning the same weight to all the $r(1), r(2), \ldots, r(\tau)$.

All the empirical correlation functions in Fig. 3.21 were calculated in this way.

(b) The complete range of the observation series (Fig. 3.22b) is used to calculate the correlation function $r(\tau)$ for every $\tau = 0, 1, 2, \ldots$. Using the same notation as above we can write

$$r(\tau) = \frac{\sum_{i=1}^{n-\tau} (x_i - \bar{x}_i)(x_{i+\tau} - \bar{x}_{i+\tau})}{\sigma_i \sigma_{i+\tau}(n - \tau - 1)} \tag{3.193}$$

The number of correlated pairs is greater than in case (a), but with a growing τ it decreases to $(n - \tau)$ pairs at the end. For a growing τ, the reliability of the calculation of $r(\tau)$ is, therefore, reduced.

For method (a) the confidence limits are (according to R. L. Anderson):

$$r_{t_\alpha} = \frac{-1 \pm t_\alpha \sqrt{m - 2}}{m - 1} \tag{3.194}$$

where t_α is a standardized random variable quantile corresponding to the significance level $(1 - \alpha)$.

For the usual significance levels of 95% and 99%, the quantile values are $t_{\alpha,95} = 1.645$ and $t_{\alpha,99} = 2.326$, respectively.

For the method (b) the confidence limits are

$$r_{t_\alpha}(\tau) = \frac{-1 \pm t_\alpha \sqrt{m - \tau - 2}}{m - \tau - 1} \tag{3.195}$$

where t_α is the same as in equation (3.194).

3.1.13 Spectral densities of random processes

The spectral density $S(\omega)$, being the Fourier transformation of the correlation function, is a very important characteristic used to investigate the periodicity properties of a random process. In the case of a centred random process

$$S_x(\omega) = \frac{2}{\pi} \int_0^\infty \dot{R}(\tau) \cos \omega\tau \, d\tau \tag{3.196}$$

where ω is the circular frequency

$$\omega = \frac{2\pi}{T} \tag{3.197}$$

providing the period is denoted by T (e.g., number of years).

According to N. Wiener and A. J. Chinchin the correlation function of a stationary random process may be expressed by means of the spectral density as follows

$$r(\tau) = \int_0^\infty S(\omega) \cos \omega\tau \, d\omega \tag{3.198}$$

Hence, with the aid of equations (3.197) and (3.198), it is possible to transform $S(\omega)$ into $r(\tau)$, and vice-versa. If the complex representation

$$\cos \omega\tau = \frac{e^{i\omega\tau} + e^{-i\omega\tau}}{2} \tag{3.199}$$

is used, we can derive (substituting $S_x(\omega) = 2S_x^*(\omega)$ and integrating from $-\infty$ to ∞) the expressions

$$S_x^*(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} r(\tau) e^{-i\omega\tau} \, d\tau \tag{3.200}$$

$$r(\tau) = \int_{-\infty}^{\infty} S_x^*(\omega) e^{i\omega\tau} \, d\omega \tag{3.201}$$

The spectral densities $S_x(\omega)$ and $S_x^*(\omega)$ differ mutually by the scale and the range of frequency ω (only non-negative ω for $S_x(\omega)$).

Similarly to equation (3.200), we may write for a non-stationary random process

$$S_x^*(\omega_1, \omega_2) = \frac{1}{4\pi^2} \iint_{-\infty}^{\infty} r(t_1, t_2) \exp\left[-i(\omega_1 t_1 + \omega_2 t_2)\right] dt_1 \, dt_2 \tag{3.202}$$

With hydrological series mostly stationary spectral densities have been investigated. Here, because of short observation series the spectral density is calculated on the basis of a single realization.

Nacházel and Patera (1975) investigated the correlation functions and spectral densities of 17 river points, 13 in the catchment of the river Labe, 4 in the catchment of the rivers Morava and the Danube. They used the approximation expression of A. S. Monin

$$S(T) = \frac{1}{2\pi}\left[1 + 2 \sum_{\tau=1}^{m} \left(1 - \frac{\tau}{m+1}\right) r(\tau) \cos \frac{2\pi\tau}{T}\right] \tag{3.203}$$

where T is the length of the period in years,

m is the length of the interval of the correlation function argument τ values,

$r(\tau)$ is the value of the correlation function according to equation (3.193) where the values x_i are substituted by annual discharges Q_i.

Figure 3.23 shows the correlation function m (Fig. 3.23a) calculated by method (b) (see Section 3.1.12), and the spectral density (Fig. 3.23b) of the flow series of the river Berounka at Křivoklát (1911–1960). The course of the correlation function differs slightly from that in Fig. 3.21c which was calculated by means of method (a) in Section 3.1.12. The confidence limits corresponding to the levels of significance 95% and 99%, respectively, are the dashed lines. The curve of the spectral density assumes two strict maxima for $T = 3$ and 6 years, respectively, and another maximum for $T = 13$ years. The river Vltava at Kamýk and the river Sázava at Poříčí show a similar course in their series.

Reznikovski (1969) introduced analyses of 83 rivers in the U.S.S.R. In 25 of them periodic components appear, while for the others simple Markov chains without periodic components are better suited.

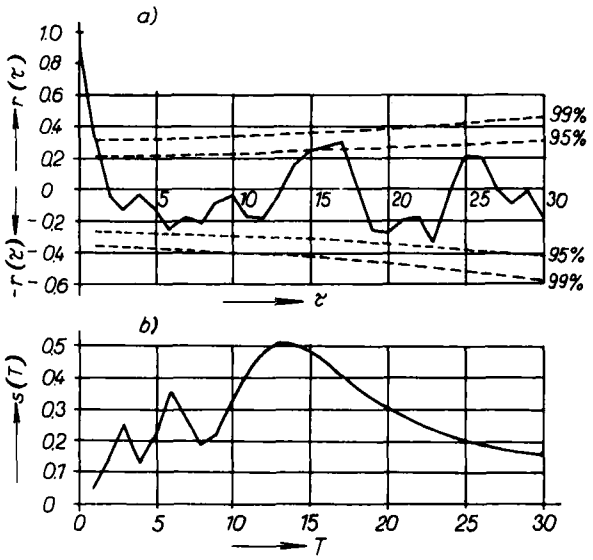

Fig. 3.23 Correlation function (*a*) and spectral density (*b*) of annual flows series of the river Berounka at Křivoklát (1911–1960)

The *methods of filtration* are supplementary in statistical analysis of hydrological series. Their essence consists in separating the random influences from the random sequence under investigation, e.g., the average annual flows; after filtration the properties (correlation, periodicity, etc.) briefly intervene and a more reliable basis for the mathematical model of the random sequence is obtained.

An analysis of correlation functions and spectral densities of filtered series was performed by Nacházel and Patera (1974) for annual flow series, precipitation series, temperatures, and relative Wolf's numbers.

Thus, a filter used in the filtration methods transforms the given random process. This leads to another process with the desired probability properties. The theory of transformation of random processes was first elaborated in radio technology (see e.g., Levin, 1965, p. 252).

The function of the filter is expressed by the *transfer response function* $h(\tau)$. A linear system input random process $\xi_1(t)$ is transformed to another random process $\xi_2(t)$ which may be expressed by the convolution integral

$$\xi_2(t) = \int_{-\infty}^{\infty} \xi_1(t - \tau)\, h(\tau)\, \mathrm{d}\tau \tag{3.204}$$

In this way we generate continuous random processes which have not yet been used in hydrology and water management, rather they are approximated by discrete sequences. The advantage consists in the possibility of using digital instead of analogue computers.

3.2 MODELLING FLOW SERIES

Models of hydrological series are supposed to give a better basis for solving problems in water management than observed series. The uncertainty contained in the observed series is replaced by a more complete description of the possible realization of the hydrological quantity (flow) sequence to be expected. Hence, moreover, a better expression for the exceedance probabilities of the quantity is obtained, especially in the neighbourhood of the extreme values.

The modelled flow sequences serve mainly to solve the over-year control of a reservoir discharge. For concrete hydrological problems, a 500-year long sequence is sufficient; for generalizing the results, at least a 1000-year long series is necessary. The modelled synthetic sequence must, as a rule, contain even more extreme values than the observed one. The statistical characteristics of a correct model, however, must agree with the characteristics of the original empirical series. Therefore, in this sense, we cannot expect the model to be more representative than the original series. On the contrary, we must check the fit by a test for any modelled random sequence.

3.2.1 Modelling annual flow sequences

The model of a series of *independent average annual flows* is the simplest. On the basis of the observed values, we can calculate the long-term average Q_a, relative annual flows $k_i = Q_{r,i}/Q_a$, the coefficient of variation C_v, and the coefficient of skewness C_s. The theoretical exceedance probabilities curve is constructed presuming a suitable probability distribution, e.g., Pearson type III. Random numbers (Gaussian noise or Pearson type III noise) obtained from the tables (Dupač and Hájek, 1962; Reisenauer, 1970) are considered to be values of the exceedance probabilities p. We deduct the corresponding values k_i from the probability curve, and the arbitrary long synthetic series which has arisen, is used to solve the problem. As the table of random numbers and the values of Q_a, C_v and C_s are sufficient to construct the model, the method may be used even when we have no time series of observations.

Modelling an *annual discharge sequence with correlated values* is more complicated. Svanidze (1961) presented two methods with correlation between values of any two neighbouring years, namely the *method of continuous functions* and the *method of non-continuous functions* (Votruba and Broža, 1966, pp. 141–143).

Modelling any probability distribution with any correlation structure is simplified by dividing the calculation into two steps: in the first step we form a sequence with a given correlation structure, but normally distributed; in the second step the sequence is transformed into a sequence with a given distribution. The advantage consists in employing profound knowledge of the normal distribution and having many aids in its application.

Before proper modelling we must transform the original real series so that we obtain a normally distributed sequence. In the case of the log-normal distribution we choose, e.g., the transformation

$$y = \lg(x - x_0) \qquad \text{[equation (3.61)]}$$

or

$$y = \ln(x) \qquad \text{[equation (3.70)]}$$

so that the variable y is normally distributed.

The transformation for the Pearson type III distribution (Beard, 1967) is known:

$$y = \frac{6}{C_s}\left[\left(\frac{C_s}{2}x + 1\right)^{1/3} - 1\right] + \frac{C_s}{6} \tag{3.205}$$

The model of the series quantity y with normal distribution must be inversely transformed to the model of the quantity x (e.g., flows Q_t) with a given probability distribution.

The universal *method of the linear regression model* (Kos, 1969) has been used to model river flows, for which computer programs have been prepared. The method assumes that there is a linear relationship between the flow at time t and the flows in the preceding years $(t-1), (t-2), \ldots, (t-k)$, given by the properties of the real series; the number of preceding years corresponds to the order of the Markov chain. The mathematical model is expressed by the basic formula called the linear regression stochastic model*):

$$x_t = b_1 x_{t-1} + b_2 x_{t-2} + \ldots + b_k x_{t-k} + e_t \tag{3.206}$$
$$k + 1 \leqq t \leqq T$$

where T is the number of terms of the real series,

$x_{t-1}, \ldots, x_{t-k}$ – average annual flows in years $(t-1), \ldots, (t-k)$,

$b_1, \ldots, b_k$ – regression coefficients depending on the correlation function,

e_t – random deviation.

The coefficients $b_1, \ldots, b_k$ are determined by the least squares method. They follow from the system of k linear equations obtained by zero value of the partial derivations of the sum $\sum e_t^2$ with respect to individual values $b_1, \ldots, b_k$. The equations are

$$\begin{aligned}
\sum_{t=k+1}^{T} x_{t-1}x_t &= b_1 \sum_{t=k+1}^{T} x_{t-1}x_{t-1} + b_2 \sum_{t=k+1}^{T} x_{t-1}x_{t-2} + \ldots + b_k \sum_{t=k+1}^{T} x_{t-1}x_{t-k} \\
&\vdots \\
\sum_{t=k+1}^{T} x_{t-k}x_t &= b_1 \sum_{t=k+1}^{T} x_{t-k}x_{t-1} + b_2 \sum_{t=k+1}^{T} x_{t-k}x_{t-2} + \ldots + b_k \sum_{t=k+1}^{T} x_{t-k}x_{t-k}
\end{aligned} \tag{3.207}$$

*) The model is not connected with the previous transformation and the standardization of the random variable, but it is more easily investigated for a standardized and centred process.

Dividing the equations (3.207) by the term $[T - (k + 1)]$, we obtain

$$\begin{aligned} m_1(x_{t-1}x_t) &= b_1 m_1(x_{t-1}x_{t-1}) + b_2 m_1(x_{t-1}x_{t-2}) + \ldots + b_k m_1(x_{t-1}x_{t-k}) \\ &\vdots \\ m_1(x_{t-k}x_t) &= b_1 m_1(x_{t-k}x_{t-1}) + b_2 m_1(x_{t-k}x_{t-2}) + \ldots + b_k m_1(x_{t-k}x_{t-k}) \end{aligned} \tag{3.208}$$

where $m_1(\ldots)$ denotes, according to equation (3.177), the product general moment of the first order with the meaning of the random process correlation function.

In the case of a normalized process (the variable z_t), the expressions $m_1(\ldots)$ mean the correlation coefficients of the correlation function (see Section 3.1.12). Thus equation (3.208) may be rewritten in the form

$$\begin{aligned} r(1) &= b_1 \cdot 1 + b_2 r(1) + \ldots + b_k r(k-1) \\ &\vdots \\ r(k) &= b_1 r(k-1) + b_2 r(k-2) + \ldots + b_k \cdot 1 \end{aligned} \tag{3.209}$$

$$b_1 = \frac{D_1}{D},\ b_2 = \frac{D_2}{D},\ \ldots,\ b_k = \frac{D_k}{D}$$

where D is the determinant of the matrix containing the correlation coefficients,

D_i is the determinant constructed by substituting the left-hand side vector for the ith column of the matrix.

The value of the random deviation e_t in equation (3.206) is determined by its relationship to the residual variation s^2, from which it follows that

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n} = m_1(e_t^2) = m_1(z_t z_t) - b_1 m_1(z_t z_{t-1}) - b_2 m_1(z_t z_{t-2}) - \ldots - b_k m_1(z_t z_{t-k}) \tag{3.210}$$

and after reducing, with respect to the limited length of the series, we obtain

$$\bar{s}^2 = \frac{m_1(e_t^2)\, T}{T - (k + 1)} \tag{3.211}$$

Hence

$$e_t = \bar{s} d_t \tag{3.212}$$

where d_t is the standard normal random variable.

Calculation

1. We analyse the set of basic statistical characteristics of given real series of annual flows x_t which, if necessary, we transform to a series y_t with normal probability distribution.

2. We calculate the arithmetical mean and the standard deviation of the transformed flows y_1, and the series is standardized to the series z_t.

3. We choose the maximum length of the Markov chain M which will be considered further, and we calculate the auxiliary value $pT = T - M - 1$, where T is the number of years of observation.

4. We calculate the coefficients of the correlation function for various time shifts up to the value M, according to the formula

$$r(i) = \frac{1}{p} \sum_{t=M+1}^{T} z_t z_{t-i} \quad \text{for} \quad i = 1, 2, \ldots, M \tag{3.213}$$

5. The system of linear equations (3.208) is solved with the aid of a known method (e.g., by means of determinants), the coefficients b_i being obtained.

6. We calculate the residual standard deviation according to eqn. (3.210), i.e.

$$s = \sqrt{1 - \sum_{i=1}^{M} b_i r(i)} \tag{3.214}$$

7. The lengths of the Markov chain are chosen successively shorter by one; steps 5–6 are repeated until the minimum of s is reached.

8. Random flows z_t are modelled according to eqn. (3.206).

9. In the case where the relation $y = \lg(x - x_0)$ in the first (direct) transformation is used, then the random flows z_t obtained are inversely transformed to

$$x_t = \exp\left[z_t s_y + \bar{y}\right] + x_0 \tag{3.215}$$

Reznikovski (1969) introduced a method of modelling, assuming either a simple or a high-order Markov chain, and a Pearson type III annual-flow probability distribution with $C_s = 2C_v$.

3.2.2 Modelling monthly flow sequences

In water management we must usually consider not only the long-term fluctuation, expressed by the sequence of average annual flows, but also the cyclic variability during a year.

In such a case we may follow two ways:

(a) the annual and the seasonal functions are treated separately; the annual functions on the basis of the sequence of average annual flows and the seasonal ones on the basis of the distribution of the flows in one year;

(b) the complete function of a reservoir is solved directly, as a rule on the basis, of the average monthly flows.

Method (b) is more advantageous and more exact from the methodological point

of view. Constructing the model of a sequence of average monthly flows is more complicated than that of the annual flows and two methods were developed

- the method of the linear regression model,
- the method of fragments.

Method of linear regression models

This method is similar to that in Section 3.2.1, but is more complicated, because it deals with statistical characteristics and correlations of individual calendar months and this has the following consequences:

(a) The probability distribution of flows in particular months (i.e., all January flows, February, ..., December) exhibits greater skewness than that of annual flows; therefore a transformation to the normal distribution is necessary, as a rule.

(b) The cyclical nature of the flow fluctuations during any one year means the non-stationarity of the process; therefore the random variable must usually be standardized in order to obtain a model with a comparable distribution function.

(c) Annual hydrographs are considered as various realizations of a random process and serve to calculate the statistical characteristics; the coefficients of correlation form the correlation matrix, the elements of which express the correlation between the flows in any two months.

Similarly to equation (3.206), for modelling a synthetic monthly flow (with a transformed and standardized z) in any month m, we can write

$$z_{c,m} = b_{1,m}z_{c,m-1} + b_{2,m}z_{c,m-2} + \ldots + b_{k,m}z_{c,m-k} + e_m \tag{3.216}$$

where $m = 1, 2, \ldots, 12$ (the corresponding month),
$c = 1, 2, \ldots, T/12$ (number of cycles—years),
T = total number of months,
e_m = the standard normal random deviation in month m.

Modelling according to equation (3.216)

The first value, $z_{c,m}$, is calculated on the basis of the preceding (arbitrarily chosen) values $z_{c,m-i}$, whose number corresponds to the order of the Markov chain.

The next value of the flow, $z_{c,m+1}$, in the following calendar month of the same year, is calculated from the value $z_{c,m}$ and the (chosen) preceding values, but with the aid of the corresponding equation (3.216) for $z_{c,m+1}$. Then the value $z_{c,m+1}$ gives the value $z_{c,m+2}$, etc.

For the regression model it is necessary to solve twelve equations of the type in eqn. (3.216). The modelling thus continues with step 12 (annual cycle). After calculating the twelve synthetic flows in the first year, we continue by calculating the flows $z_{2,m}$ in the second year, etc., thus obtaining a pseudo-chronological series of monthly values z_m.

The coefficients $b_{i,m}$ $(i = 1, 2, \ldots, k,\ m = 1, 2, \ldots, 12)$ in every equation (3.216) are determined as in the model of the annual flows by means of the least squares method. From the sequence of real flows we obtain a system of k-linear equations for every month m:

$$\begin{aligned}
\sum_{c=\bar{c}}^{c} z_{c,m-1}z_{c,m} &= b_{1,m}\sum_{\bar{c}}^{c} z_{c,m-1}z_{c,m-1} + b_{2,m}\sum_{\bar{c}}^{c} z_{c,m-1}z_{c,m-2} + \ldots + b_{k,m}\sum_{\bar{c}}^{c} z_{c,m-1}z_{c,m-k} \\
&\vdots \\
\sum_{\bar{c}}^{c} z_{c,m-k}z_{c,m} &= b_{1,m}\sum_{\bar{c}}^{c} z_{c,m-k}z_{c,m-1} + b_{2,m}\sum_{\bar{c}}^{c} z_{c,m-k}z_{c,m-2} + \ldots + b_{k,m}\sum_{\bar{c}}^{c} z_{c,m-k}z_{c,m-k}
\end{aligned} \tag{3.217}$$

where $c = T/12$,

$$\bar{c} = \text{entire}\left(\frac{k-1}{12}\right) + 2 \text{ (in years)}.$$

After dividing each equation of the system [eqn. (3.217)] by the number of correlated pairs reduced by one, i.e., by $c - \bar{c}$, it follows that

$$\begin{aligned}
m_1(z_{m-1}z_m) &= b_{1,m}m_1(z_{m-1}z_{m-1}) + b_{2,m}m_1(z_{m-1}z_{m-2}) + \ldots + b_{k,m}m_1(z_{m-1}z_{m-k}) \\
&\vdots \\
m_1(z_{m-k}z_m) &= b_{1,m}m_1(z_{m-k}z_{m-1}) + b_{2,m}m_1(z_{m-k}z_{m-2}) + \ldots + b_{k,m}m_1(z_{m-k}z_{m-k})
\end{aligned} \tag{3.218}$$

The expressions m_1 mean the elements of the correlation matrix which are calculated from the given real flow series. The unknown regression coefficients $b_{i,m}$ are determined by solving the system in eqn. (3.218). Thus,

$$\begin{aligned}
r(1,m) &= b_{1,m}\cdot 1 + b_{2,m}r(1,m) + \ldots + b_{k,m}r(k-1,m) \\
&\vdots \\
r(k,m) &= b_{1,m}r(k-1,m) + b_{2,m}r(k-2,m) + \ldots + b_{k,m}\cdot 1
\end{aligned} \tag{3.219}$$

$$b_{1,m} = \frac{D_{1,m}}{D_m},\ b_{2,m} = \frac{D_{2,m}}{D_m},\ \ldots,\ b_{k,m} = \frac{D_{k,m}}{D_m}$$

The meaning of the determinants D_m and $D_{i,m}$ $(i = 1, 2, \ldots, k)$ is the same as in the model of the annual flows; however, they depend on every particular month m.

The residual variance in every month m follows from the equation

$$s_m^2 = m_1(e_m^2) = m_1(z_m^2) - b_{1,m}m_1(z_m z_{m-1}) - b_{2,m}m_1(z_m z_{m-2}) - \ldots - b_{k,m}m_1(z_m z_{m-k}) \tag{3.220}$$

According to the limited length of the series, it holds that

$$\bar{s}_m^2 = \frac{m_1(e_m^2)\,c}{c-k-1} \tag{3.221}$$

The previously calculated values of coefficients $b_{i,m}$ and $m_1(z_m z_{m-i})$ may be used advantageously; thus only $m(z_m^2)$ remains to be determined.

When testing for the proper length of the Markov chain, we follow the same approach as in the case of the model of annual flows, i.e., the residual variance is determined in dependence on the length of the Markov chain until the minimum is reached.

The particular steps in the calculation with the aid of a computer are analogous to those of the annual flow modelling, the only difference is that the appropriate procedures are repeated for every month. The last step of the calculation is again the inverse transformation of the artificial flows; if the log-normal distribution is used for the direct transformation, then for the inverse transformation it holds that

$$x_{c,m} = \exp\left[z_{c,m}s_m + \bar{y}_m\right] + x_{m,0} \tag{3.222}$$

Method of fragments

The method of fragments (Svanidze, 1963) is based on double random sampling. With the aid of the method described above, the first random sampling of the average annual flows is performed (Fig. 3.24a). Then the second random sampling of fragments

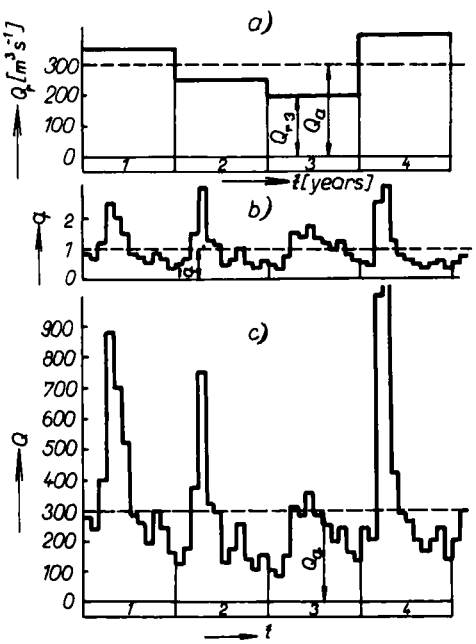

Fig. 3.24 Modelling of monthly flow series by means of the fragment method
(*a*) first random sample $Q_{r,i}$; (*b*) second random sample (of fragments $q(t)$); (*c*) section of modelled series of average monthly flows

follows, corresponding to the real relative flows $Q/Q_{r,i} = q = q(t)$ in individual years of observation; thus in every year it holds that $\bar{q} = 1$ (Fig. 3.24b). The fragments are numbered and with the aid of the random-numbers table are assigned to the individual years of the synthetic series. The ordinates of the fragments are multiplied by the

corresponding $Q_{r,i}$ of the synthetic series and an arbitrarily long, pseudo-chronological synthetic series with distribution of the flow within the years is obtained (Fig. 3.24c).

Applying the method in this form, when no relation between Q_r and the distribution of the flow within the year is supposed, may lead to unrealistic results*). By investigating relations between Q_r and variability of flow during a year, the unrealistic randomness may be excluded. The method might be improved by increasing the number of fragments to be comparable with the number of years of observation by means of modelling. Thereby, however, the method loses its main advantage, i.e., simplicity, whereby it does not need modern computers. Since the introduction of numerical computers, the method of fragments has lost its importance as compared with the method of the linear regression model.

There is, however, one advantage: it maintains the correlations between the annual flows Q_r, whereas with the regression method the correlations between monthly flows Q_m are retained, but the correlations between values of Q_r might not be.

For a long-term annual cycle, the method of fragments should give more exact results, because of the importance of a proper correlation between Q_r values, whereas for a cycle of a few years, more exact results should be obtained from a series modelled by the regression method, because the correctness of the correlation between Q_m values is ensured.

3.2.3 Modelling sequences of monthly flows in a system of stations

Between the empirical flow series at points on a river a significant correlation often exists (Fig. 3.7, Table 3.7). The table shows that the coefficient of correlation $r = 0.69$ is high, even between distant catchments (the Labe to Josefov and the Vltava to Kamýk). In two cases the coefficient is as high as, $r = 0.94$.

If reservoirs work in a system, their collective efficiency is generally higher than the total of their isolated effects. The rate of the increased effect of the system, as compared with the isolated effects of individual reservoirs in the system, depends on the synchronization of their hydrological regimes and their relative volume. The more asynchronous the regimes, the higher the effect of the system.

If flow series of a system of river profiles are modelled, a double correlation relationship has to be considered:

- correlation (auto-correlation) within each series,
- cross-correlation between synchronized flows of various stations of the system.

---

*) If from the 30-year series from 1931 to 1960 for the river Berounka at Křivoklát a fragment from 1954 is assigned to the highest-flow year, 1941, then in July we obtain an average flow equal to 2.5 times the greatest observed monthly maximum and greater than the maximum peak flow of a two-year flood. On the contrary by assigning a fragment of 1947 to the year 1934, the minimum monthly mean value is smaller by one third than the minimum daily flow of the whole thirty years (Votruba and Broža, 1966, p. 145).

Table 3.7 Correlation between the annual flows of some gauging stations in Bohemia (1926–1946)

Stream gauging site	Catchment area [km²]	Correlation coefficient r						
		the Labe–Josefov	the Labe–Pardubice	the Vltava–Kamýk	the Berounka–Beroun	the Vltava–Modřany	the Ohře–Louny	the Labe–Mělník
the Labe–Josefov	1 840.2	–	0.92	0.69	0.71	0.71	0.75	0.79
the Labe–Pardubice	5 987.3	–	–	0.76	0.78	0.79	0.80	0.86
the Vltava–Kamýk	12 213.0			–	0.89	0.94	0.85	0.92
the Berounka–Beroun	8 287.9				–	0.93	0.89	0.92
the Vltava–Modřany	26 706.1					–	0.87	0.94
the Ohře–Louny	4 981.4						–	0.89
the Labe–Mělník	41 798.3							–

Two basic methods are developed for modelling flow series of a system of stations:
- method of central and satellite stations,
- method of orthogonal transformation.

(a) Method of central and satellite stations

By this method we model first the flows in the central station (e.g., by means of the method of the linear regression stochastic model according to Sections 3.2.1 or 3.2.2) and then the temporally corresponding (coinciding) flows in the other satellite stations, according to the linear regression:

$$y_{i,j,k} = \bar{y}_{j,k} + b_{j,k}(x_{i,j} - \bar{x}_j) + u_i d_{j,k} \tag{3.223}$$

$$d_{j,k} = s_{j,k}\sqrt{(1 - r_{j,k}^2)} \tag{3.224}$$

where $y_{i,j,k}$ is the flow with the ordinal number i in month j, in station k,
$\bar{y}_{j,k}$ – average monthly flow in month j, in station k,
$b_{j,k}$ – regression coefficient in month j for the relationship between the flows of the central station and of the satellite stations,
$x_{i,j}$ – flow of the central station, with the ordinal number i, in month j,
$\bar{x}_j$ – average monthly flow from the central station in month j,
u_i – standardized random variable,

$s_{j,k}$ – standard deviation of flows in month j, in station k,
$r_{j,k}$ – coefficient of correlation between flows in the central and the satellite station k, in month j.

For any pair of the central and satellite stations we prepare twelve equations of type (3.223) and (3.224) and the calculation is carried out for the individual calendar months.

The method is suitable mainly for relatively synchronic regimes; otherwise the auto-correlation of the series of the satellite stations may be considerably disrupted. In the case of a looser bond between the two stations, equation (3.223) can contain not only the contemporary flows in the central station x_i, but also the preceding flows x_{j-1}, and also, if need be, the other flows y_{j-1} in the satellite station.

(b) Method of orthogonal transformation

This method is more general and its principle consists in the transformation of the linearly dependent flows in any calendar month at various stations into independent (orthogonal) flows. The given real flows $x_{i,j,k}$ in the same month j, but at various stations k, form correlated random vectors which are transformed to independent vectors.

First the real flow sequence in one calendar month and at all stations is transformed to a normally distributed sequence which, moreover, is standardized. Then the orthogonal transformation is performed, and a random series at hypothetical stations is modelled. By the inverse transformation we obtain the correlated random sequences in the real stations and by another inverse transformation we finally obtain the desired synthetic sequences with a given mean value and standard deviation.

In the case of a large system of stations (reservoirs), we may be limited by the computer capacity. Then we cannot model the whole system at once, but parts (sub-systems) with closer connections have to be modelled separately. Hence, the same problem arises as with the independent modelling of isolated series: the possibility of deranging the real correlations between the individual sub-systems. The most important things are to analyse the problem, to prepare a program for the computer, and finally to interpret the results properly.

3.3 INFLOW TO A RESERVOIR

In designing a reservoir, it is absolutely imperative to know the law of inflow. Inflow can be

- natural, influenced by man, or completely controlled by human intervention,
- deterministic or stochastic,
- constant or variable.

The characteristics of inflow to a reservoir depend on its function (flood control or direct supply) and on the duration of the control cycle (Table 3.8). For short-term control, a *deterministic inflow* can be considered; for over-year control, a *stochastic* one in necessary.

Table 3.8 Inflow to a reservoir and withdrawal with various release control cycles

Control cycle	daily	weekly	annual	over-year
Inflow	deterministic	deterministic	deterministic or stochastic	stochastic
	steady or variable	steady as a rule	variable	variable
	frequently controlled by man	natural or controlled	natural or changed	natural or changed
Withdrawal and controlled release	deterministic or stochastic	deterministic or stochastic	stochastic	stochastic
	variable or steady	variable as a rule	variable	variable
	controlled	controlled	controlled	controlled

A *constant inflow* to a reservoir is introduced only for short-term release control by distribution reservoirs.

Controlled inflow to a reservoir is frequently introduced for short-term discharge control (pumping of water to a water tank, to the reservoir of pumped storage hydropower plants, etc.). Lateral reservoirs have a more or less controlled inflow as the water is brought from the main stream, or generally speaking, reservoirs that receive their water from other catchments.

The *flood inflow* to a reservoir must also be taken into consideration in the determination of its flood-control effect. This inflow is short, but very variable. To solve the problem, the real pattern of a flood wave is sometimes used. However, more frequently, models of floods are constructed with various probabilities of occurrence, both as to their maximum peak discharge and to the form this takes, and thus the flood volume. This is usually a natural or influenced inflow, and is included in the design as deterministic, with a certain probability of being exceeded.

For the inflow to a reservoir it is not sufficient to consider the statistical data of the past and the present states, but also a *forecast* of the development of inflow in the future has to be elaborated with regard to the long physical life span of a reservoir and of the more progressive facilities that control the inflow.

More and more frequently, water for reservoirs will be diverted from one catchment area to another. Economic analysis optimizes the size of the transported flow and

determines the dimensions and costs of the conduits. For preliminary calculations it is possible to determine the amount of water diverted with a certain chosen reliability; this can be done either numerically or graphically by a summation curve, which is especially suited for an analysis of alternative conduit capacities. The diagram in Fig. 3.25 illustrates that

- with a minimum maintained discharge in the river $Q_{min} = 0.25Q_r$, the amount of water diverted in one year decreases by 25% of the annual discharge W_r,
- if the sum of the conduit capacity and the maintained minimum flow $Q_{con,c} + Q_{min} = 1.5Q_r$, the amount of water diverted per year decreases by the excess discharge W_i, which is 18% W_r,
- if the conduit is $Q_{con,c} = 1.25Q_r$, the annual diversion has the value of $W_{div} = 0.57W_r$ of the design year,
- the total water diversion W_{div} is decreased to a greater extent by maintaining a minimum flow in the stream than by a small capacity of the conduit, as peak flows last only a short time so that the volume of the excess discharge W_i is relatively small.

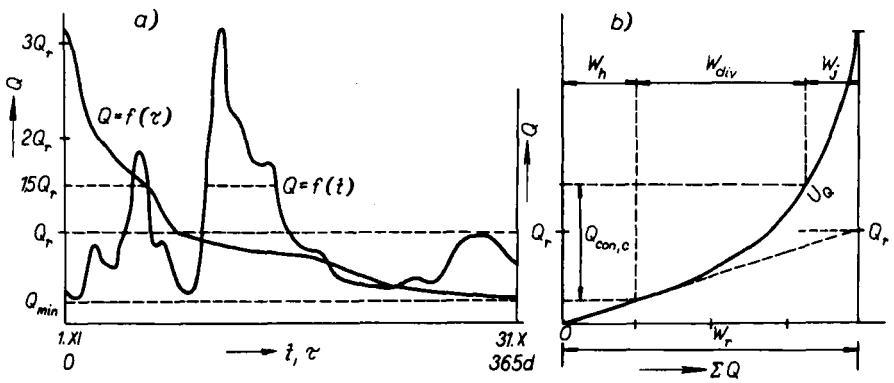

Fig. 3.25 Analysis of water diversion by means of the summation curve
(*a*) time curve (hydrograph) and exceedance curve of discharges in the river from where the water is diverted; (*b*) summation curve corresponding to curves (*a*)

When *diverting water directly from a river* (without storage) it is usually not sufficient to work with monthly or ten-day discharges, but the continuous curve of discharges or at least the average daily discharges have to be considered. The calculations could be incorrect if the average discharges within the month are used. This error need not be significant if the conduit has a large capacity and if the flood peaks exceeding this capacity are only very short.

It is more difficult to construct a chronological or pseudo-chronological series for the balance solution of a reservoir, if we do not want to limit ourselves to a direct solution on the basis of observed series. Essentially there are two such cases:

– the natural inflow to a reservoir is negligibly small (as for a lateral reservoir) so that the inflow is only the diverted water,

– the design has to consider the natural inflow together with the diverted water.

For the over-year reservoir, a long series of discharges in the river site from where the water is withdrawn and diverted to the reservoir will have to be modelled. Let us presume that mean daily discharges will be considered. No reliable method for direct modelling of the chronology of mean daily discharges with the help of a linear regression model has yet been elaborated. These are actually not natural daily discharges, but discharges from which the maintained minimum discharge

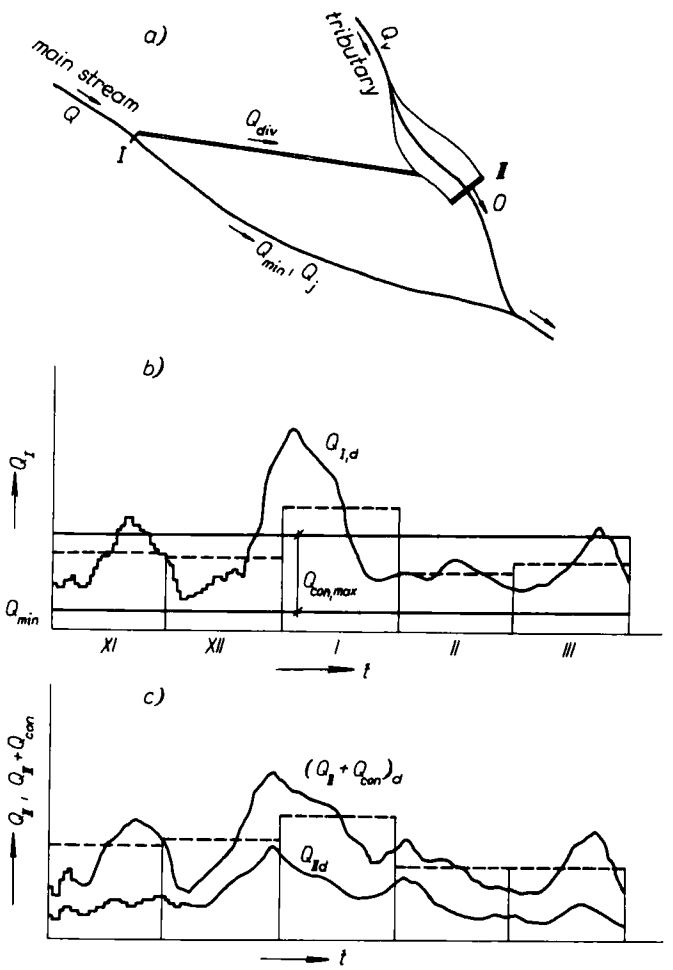

Fig. 3.26 Reservoir with water diversion

(*a*) layout; (*b*) adjustment of hydrological input data at site I; (*c*) construction of inflow to reservoir ($Q_{II} + Q_{con}$)

and the excess discharge, exceeding the conduit capacity, will have to be deducted. Naturally, such a reduced discharge will have different characteristics from a natural discharge.

For the fragment method, the method explained in Section 3.2.2 is applied. Only the pattern of the fragments would change: instead of monthly (or ten-day) discharges, the fragments consist of daily discharges.

A model of the inflow to a reservoir, i.e., a series reduced by the minimum maintained and excess discharges, can be constructed as follows (Fig. 3.26), e.g.:

1. observed series of mean daily discharges Q_I in the withdrawal profile I are reduced by the minimum maintained discharge Q_{min} and the excess discharge Q_i (Fig. 3.26b), whereby a series of daily discharges through the conduit Q_{con} is obtained;
2. the synchronous daily discharges in profile II and the reduced discharges through the conduit Q_{con} are added; from these the series of the mean monthly discharges to a reservoir $(Q_{II} + Q_{con})$ are calculated; this is the input set for the modelling of the series;
3. from this initial series $(Q_{II} + Q_{con})$, e.g., a linear regression model of a series of any length of average monthly flows to a reservoir is modelled;
4. the modelled series serves as a basis for the design.

It is also possible to model the series of the mean monthly discharges in profiles I and II separately, bearing in mind the correlation relationship between the two profiles according to Section 3.2.3; but as the series in profile I was deformed by the reduction, it does not seem that this more complicated method can have more reliable and accurate results.

If the discharge Q_{II} is negligible, modelling is limited to the model of the reduced series Q_{con} using the above method, but substituting $Q_{II} = 0$.